Computation foundations of spatial transcriptomics

Thesis by

Lambda Moses

In Partial Fulfillment of the Requirements for the
degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2023
Defended May 24, 2023

Lambda Moses
ORCID: 0000-0002-7092-9427

ACKNOWLEDGEMENTS

"So which of the favors of your Sustainer will you deny?" Quran 55:13

I have come a long way in my life. This thesis signifies the end of one stage, but it's only a beginning. There's a lot to thank, given that it took a lot of blessings and privileges for me to make it this far. I don't have to travel far to see people without such privileges. I wish that my research will benefit people of all classes.

Here I list people and institutions I thank chronologically. I thank my parents for moving numerous times when I was young to seek out new opportunities, and taking my education very seriously, so I could attend a prestigious international high school, which helped to get me into UCLA. In order to make sure that I focus on my study rather than worrying about money, they paid for my tuition and living expenses at UCLA in full, so I didn't having to work as a lab technician to earn a wage, nor did I fall into debt, and I could double major and work on my own research project, without which I wouldn't have gotten into Caltech. I also thank them for allowing me to pursue my extracurricular interests in computer graphics and art while they were irrelevant to standardized exams. From these interests I developed my ability to learn new things on my own, which is crucial to research because so much is added to our knowledge every day that textbooks and classes don't yet exist for many important things to learn.

I thank Edip Yuksel, philosophy instructor at Pima Community College in Tuscon, Arizona, author of Quran: a Reformist Translation, and international peace activist, whom I met online back in high school, who taught me critical thinking, introduced me to philosophy, and instilled a strong sense of justice. I had never heard of critical thinking before then. He brought me an Islam that is liberating.

I thank Jinzhu Duan, then at Arjun Deb lab at UCLA, who introduced me to hematoxylin and eosin (H&E) and Masson's trichrome staining, which later sparked my interest in spatial -omics, the theme of this thesis.

I thank Aldons Jake Lusis, in whose lab I found a supportive environment to work on my undergrad research project on the intersections between lipid and iron metabolism, postdoc Brid Fuqua who mentored me through the project and proofread my graduate school statements of purpose, bioinformatician Calvin Pan who introduced me to computational biology which is now my specialty, and the then postdocs Marcus Seldin, Forde Norheim, lab manager Nam Che, lab technician

Sarada Charugundla, and Eleazar Eskin who supported the research project. I did not complete the project after I graduated, but Brie, Calvin, Beyza Ozdemir who took over the project after I left, and many new lab members I have never met continued working on the project. This project helped to get me to Caltech. Although I'm doing a very different kind of computational biology at present, this project made me fluent in R and taught me computational methods I'm still using today. My interest in metabolism continues to this day, which I would like to revisit after the software in this thesis matures.

I thank my friends and roommates at UCLA, Gunilla, Momoko, and Samantha, who along with Jake's lab made me enjoy my time at UCLA. Furthermore, my time at UCLA wouldn't have been great if not for my secondhand commuter bike which I retroactively name United Space Ship (USS) Resonance, with which I explored coastal regions of LA. With Resonance, I frequented LA County Museum of Art (LACMA), whose exhibitions made me think.

At Caltech, I thank Lior Pachter, my advisor, who has been patient with me when I struggled with mental health problems and unsuccessful projects, and who has created a friendly and supportive lab culture. I especially thank him for being supportive when I worked on the unconventional Museum of Spatial Transcriptomics project (Chapter 2), in which I ended up writing a book. Thanks to his support, many people have messaged me saying that the book is helpful. He also helped me a lot in connecting with collaborators and in public relations, and in suggesting cool research ideas and proofreading papers. I greatly thank him for the supportive environment, which has helped several of my colleagues who moved from other labs which they found toxic. Accordingly, I also thank friends, lab mates, and collaborators who had interesting discussions with me on spatial -omics, especially Sina Booeshaghi, Kayla Jackson, and Laura Luebbert, and the Pall Melsted lab in Iceland, who contributed to my current project. I also thank my committee members, Matt Thomson, David Van Valen, and Katie Bouman for giving constructive suggestions in committee meetings, and examination committee members Barbara Wold and Harold Pimentel for reading this thesis.

I thank the community of free and open source developers. I learnt to write comprehensive and reproducible documentation and to emphasize user friendliness in my packages. My participation in this community taught me the value of sharing and mutual aid, that an egalitarian, decentralized, cooperative, and non-capitalist community that can uphold high standards is possible.

Because everything is connected, I don't compartmentalize my life. I'm having the best time of my life in LA, by which I mean something more like Tovaangar, territory of the Tongva people going beyond modern colonial boundaries. It is this land, and water imported from Owens Valley and Colorado River, that nourished me in the best time of my life, as if I'm a cell in a tissue and the rivers are blood vessels. Modern LA was initially built by enslaved Tongva labor, without which we would not have the vibrant and diverse, albeit segregated and unjust metropolis today. It is the vibrancy and diversity that gives LA a thought provoking and edgy culture, which has inspired many new ideas in research and otherwise, and which gives me hope in repairing the harm done by the injustices of colonialism and racism.

I must thank those who grew and sold my food, especially Urban Homestead and re_ grocery that try to be kind to the environment. Also, I must not forget the working class, including the dining hall workers, janitors, delivery workers, restaurant cooks and waiters, transit operators, cashiers, farmers, and factory workers, whose labor I benefit from. Especially considering that given the amount of privilege I needed to get to Caltech, their children might find it much more difficult to get to somewhere like Caltech than I did.

I thank the cycling community, which helped me to overcome an age old social anxiety, so I no longer fear giving presentations and communicating with collaborators. USS Resonance was stolen in 2016, when I no longer had much time to ride as I double majored. But its spirit lives on in my current tradition of the "voyages" of USS Voyager, a road racing bike with which I explored far and wide in Tovaangar and beyond. It is no coincidence that the R package this thesis culminated into is also named Voyager (Chapter 4); the bike was initially in part named after the R package, but now it goes both ways, though USS Voyager in Star Trek is another inspiration. Captain Janeway of Voyager in Star Trek is my role model of a woman who is brave and adventurous yet kind. The voyages made me more connected to the land and community, and inspired ideas central to this thesis that brings decades of geospatial research to the new rising field of spatial -omics.

I thank my beautiful and inclusive faith communities, the Threshold Society which helped me to climb out of the mental health problems in 2020, Muslims for Progressive Values and Women's Mosque of America in LA, and El-Tawhid Jumah Circle in Toronto. Last but not least, I thank Khaled Abou El-Fadl, UCLA law professor and Islamic scholar, from whom I learnt wisdom and courage.

ABSTRACT

Single-cell and spatial transcriptomics have come of age in the past few years; datasets and data analysis software packages have proliferated. With the increasing sizes of datasets, proliferating new data collection technologies, and mainstreaming of high-throughput technologies, the software can be improved for better speed and memory efficiency, standardized and consistent user interface for multiple technologies, and in documentation to onboard new users. First, I collected a database of spatial transcriptomics literature and analyzed the data on trends and sociology in this field. Based on the database and data analyses, I wrote a comprehensive book both qualitatively and quantitatively documenting the history of the field since the 1960s and reviewing more recent developments, which informed the software and methods I later developed. Then, to address the challenges with the pre-processing large datasets, we developed `kallisto` `bustools` for fast and modular pseudoalignment of sequencing reads to the transcriptome in single-cell RNA-seq (scRNA-seq), giving consistent results with the established and much more computationally demanding alignment method Cell Ranger. Briefly summarized are my attempt to map dissociated cells in scRNA-seq to a spatial gene expression reference and to build a image processing pipeline for image based spatial transcriptomics data analysis. Finally, to address the challenges in downstream analyses of spatial -omics data, I first wrote the new `SpatialFeatureExperiment` (SFE) data structure to represent and operate on geometries in spatial transcriptomics data and to organize results from spatial analyses. Based on SFE, I wrote Voyager, which brings decades of research in geospatial data analysis to spatial transcriptomics, to better utilize the opportunities from spatial information to gain novel biological insights. To reduce user learning curve, Voyager conforms to SCE styles and conventions and has a comprehensive documentation website and consistent user interface to many geospatial methods.

PUBLISHED CONTENT AND CONTRIBUTIONS

1. Melsted P, Booeshaghi AS, Liu L, Gao F, Lu L, Min KH(, Veiga Beltrame E da, Hjörleifsson KE, Gehring J, and Pachter L. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nature Biotechnology 2021. I (Lu, L.) used kallisto bustools to generate gene count matrices from FASTQ files of example 10X v3 datasets from 10X Genomics' website, comparing the kallisto bustools output with Cell Ranger output. I focused on a mouse 10k neuron dataset to show concordance of downstream results between kallisto bustools and Cell Ranger. I also wrote the R Colab notebooks on the documentation website. DOI: 10.1038/s41587-021-00870-2

2. Moses L and Pachter L. Museum of spatial transcriptomics. Nature Methods 2022; 19. I (Moses, L) curated the database, performed the analyses of the metadata, and wrote the manuscript and the supplement. DOI: 10.1038/s41592-022-01409-2

3. Moses L, Jackson K, Luebbert L, Einarsson PH, Booeshaghi S, Antonsson S, Melsted P, and Pachter L. Voyager: exploratory spatial data analysis from geospatial to spatial -omics. To be submitted 2023. I (L.M.) conceived the idea and wrote the SpatialFeatureExperiment, Voyager, and SFEData R packages and most of the manuscript. I also wrote most of the vignettes on the documentation website.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

Single-cell and then spatial transcriptomics have come of age in the past few years; datasets and data analysis software packages have proliferated. As these high-throughput technologies have become mainstream, studies profiling gene expression in millions of cells have been produced, and many software packages have been written for a variety of data analysis tasks from upstream sequence alignment to downstream data visualization. Spatial transcriptomics data analysis largely inherits from the single-cell tradition. While many software packages have been written for spatial analyses, many opportunities from the spatial information have not been utilized.

Part 1 is a comprehensive review of spatial transcriptomics. Chapter 2 is adapted from my review paper on spatial transcriptomics, and the following chapters until Chapter 10 are the book which is the supplementary material of the paper with more details about this field. This part is based on a database of literature on spatial transcriptomics. First the history of this field is surveyed, including predecessors to current technologies dating back to the 1960s and early attempts of high-throughput gene expression profiling in space from the 1980s to the 2000s. Then current data collection technologies and data analysis methods are reviewed. Metadata of the database is analyzed, including the number of publications on each type of data collection or analysis, institutions where the studies were performed, species and tissues where the data was collected, number of genes and cells profiled, and text mining abstracts. This gives both a qualitative and a quantitative overview of the history and sociology of the field. While the text is about developments up until 2022 when the paper was published, the figures shown in this thesis have been updated as the database is continuously updated. The figures should reflect the database as of March 2023. The text is also updated if it's inconsistent with the updated figures or no longer true given the rapid development of this field.

Part 2 concerns my contributions to single-cell and spatial transcriptomics. Chapter 11 summarizes my contributions to non-spatial single-cell RNA sequencing (scRNA-seq) and unpublished contributions to spatial transcriptomics. First is an attempt to map dissociated cells from scRNA-seq datasets that profile the whole

transcriptome to a spatial references that profiles a much smaller number of genes.

Next, I describe my contributions to the `kallisto bustools` project. As of writing, 10X Genomics is the company that sells the most popular scRNA-seq and spatial transcriptomics technologies. Our lab developed `kallisto bustools` for fast and modular pseudoalignment of sequencing reads to the transcriptome, much faster and memory efficient than Cell Ranger, the standard 10X read alignment software. We also wrote comprehensive documentation for `kallisto bustools` with tutorials on downstream analysis after getting the gene count matrix from `kallisto bustools` that can be run reproducibly on Google Colab. While Cell Ranger is specific to 10X data, `kallisto bustools` can be used for a variety of single-cell and spatial sequencing data. Using example datasets from the 10X website, I showed that the output gene count matrices of `kallisto bustools` give consistent downstream analysis results as those of Cell Ranger.

Next, to address the problem that image processing software is very specific to technology, I built a pipeline that can apply across technologies going from stitching multiple fields of views to segmenting cells and transcript spots, but found the problem much deeper as there is no standard in file formats in the field. One would need to devise a standard file format suitable to the field in order to solve the problem. While I haven't devised such a format, I point to literature on this issue.

In Chapter 12, I describe the `SpatialFeatureExperiment` (SFE) data structure to represent processed spatial -omics data for downstream spatial analyses. Existing data structures for spatial -omics data don't fully take advantage of the spatial information in cell morphology and geometric relations between cells and other entities such as pathologist annotations. SFE brings Simple Features to the existing single-cell data structure `SingleCellExperiment` (SCE). Simple Features is a standard format to represent vector geometries in the geospatial field. In SFE, Simple Features is used for efficient representation of and operations on geometries such as cell segmentation polygons and histological regions, allowing for studies of cell morphology and geometric relations with other geometries. In addition, SFE organizes various spatial analysis results and link them to the genes or features for which they were computed. SFE follows the styles and conventions of SCE, easing adoption by users already familiar with SCE.

Finally, in Chapter 13, I describe the Voyager project that centers on my R package `Voyager`, which fills a gap in exploratory spatial data analysis (ESDA). ESDA is exploratory data analysis (EDA) specifically for spatial aspects of the data. Voy-

ager performs the spatial analyses on SFE objects, and brings decades of geospatial research to spatial transcriptomics. Because the original implementations of some ESDA methods were written for much smaller datasets, they are not scalable to spatial -omics data. I have reimplemented some of these methods and performed benchmarks to show that my implementations are many times faster and more memory efficient than the original implementations, so can be used on larger datasets. Visualization is essential to EDA, and Voyager implements elegant plotting functions for the data and spatial analysis results, with colorblind friendly default palettes. I show examples of novel biological insights gained from ESDA, including the presence of negative spatial autocorrelation in the tissue and that the library size, commonly treated as a technical artifact, can be biologically relevant. These show that bringing in decades of research in ESDA has the potential for more novel biological discoveries. In addition, Voyager addresses the following challenges in spatial transcriptomics data analysis:

First, as shown by my database, spatial transcriptomics data analysis is largely split between programming languages R and Python, and which language to choose is often quite personal. For both single-cell and spatial transcriptomics, the *de facto* standard EDA package in R is `Seurat`, and the *de facto* standard in Python is `scanpy`. However, they give different results for some ostensibly the same tasks, such as principal component analysis (PCA) and log-fold change of gene expression between clusters, because of defaults most users may be unaware of or different implementation details that are not documented, causing a personal preference to lead to different biological conclusions. To cater to a wider community of users, our collaborators wrote a Python implementation of Voyager. To address this problem of inconsistency, we wrote "compatibility tests" to make sure that the two implementations of Voyager give the same results for core functionalities, and document defaults and implementation details even if the reason behind them is simply convention in the field.

Second, new packages performing specific tasks often rely on syntax or data structures that are very different from other packages in the field, forcing users to learn new syntax or to convert between data structures in order to perform additional analyses in a workflow. The learning can be difficult when as shown in my database, many packages have no documentation. Most are not on a standard public repository so can be more difficult to install. The `Voyager` R and Python packages reuse existing standard data structures and conform to styles and conventions in the

ecosystem around these structures, to reduce user learning curve. The R packages Voyager and SFE are on Bioconductor, which requires packages to pass an initial manual review, have unit tests and comprehensive documentation, and pass a daily automated check that runs the unit tests and examples and checks for problems in the code. The Python package is on PyPI. While PyPI does not check the packages, with the compatibility tests, the Python package is indirectly held to Bioconductor standards.

Stemming from the previous point, we have written a comprehensive and reproducible documentation website, with tutorials using data from several spatial transcriptomics technologies and introducing various ESDA methods. This goes beyond transcriptomics, as we have included a spatial proteomics tutorial as well. To ensure reproducibility and scalability, the website, including all the tutorials, is built from scratch on a fresh machine on GitHub Actions with limited computational resources.

Part I

Museum of Spatial Transcriptomics

Chapter 2

THE MUSEUM OF SPATIAL TRANSCRIPTOMICS PAPER

1. Moses L and Pachter L. Museum of spatial transcriptomics. Nature Methods 2022; 19. DOI: 10.1038/s41592-022-01409-2

2.1 Introduction

It has long been recognized that in biological systems ranging from the Drosophila embryo to the hepatic lobule, many genes need to be properly regulated in space for the system to function. In order to study the spatial patterns of gene expression, many different spatial transcriptomics methods, which produce spatially localized quantification of mRNA transcripts as proxies for gene expression, have been developed. Thanks to growing interest in the field, several reviews have been written in the past 5 years, providing overviews of experimental techniques for data collection [1, 2], and describing how such techniques can be applied to specific biological systems, e.g. tumors [3], brain [4], and liver [5]. These reviews typically begin with either laser capture microdissection (LCM) or single molecular fluorescent in situ hybridization (smFISH) in the late 1990s, although the quest to profile the transcriptome in space is much older.

Unlike the previous reviews, this paper presents a database of literature dating back to 1987 comprehensively documenting the historical evolution and current development in data collection and analysis in spatial transcriptomics. In addition, we have analyzed the literature metadata from the database to show trends in the field. Key highlights from the database and analyses are presented in this paper, and more details are presented in our book length supplement: https://pachterlab.github.io/LP_2021/

Section and figure numbers of the supplement in this paper refer to those in the DOI PDF version, while those in the online HTML version are subject to change as it is continuously updated to reflect changes in the field. This database was curated by searching keywords such as "spatial transcriptomics" and "Visium" on PubMed and BioRxiv and manually screening literature citing influential papers in the field. Literature metadata collected include date published or posted and institution of the first author. In addition, metadata for publications concerning new datasets include species and tissue where the data was collected, experimental techniques used to

collect the data, and programming languages used to analyze the data. Metadata for publications concerning new data analysis methods include programming languages used in the implementation, code repository of the implementation, and whether the code is packaged and documented. The database is continuously updated by manually screening RSS feeds from PubMed and BioRxiv for relevant keywords, or by submission via a Google Form.

2.2 Prequel era

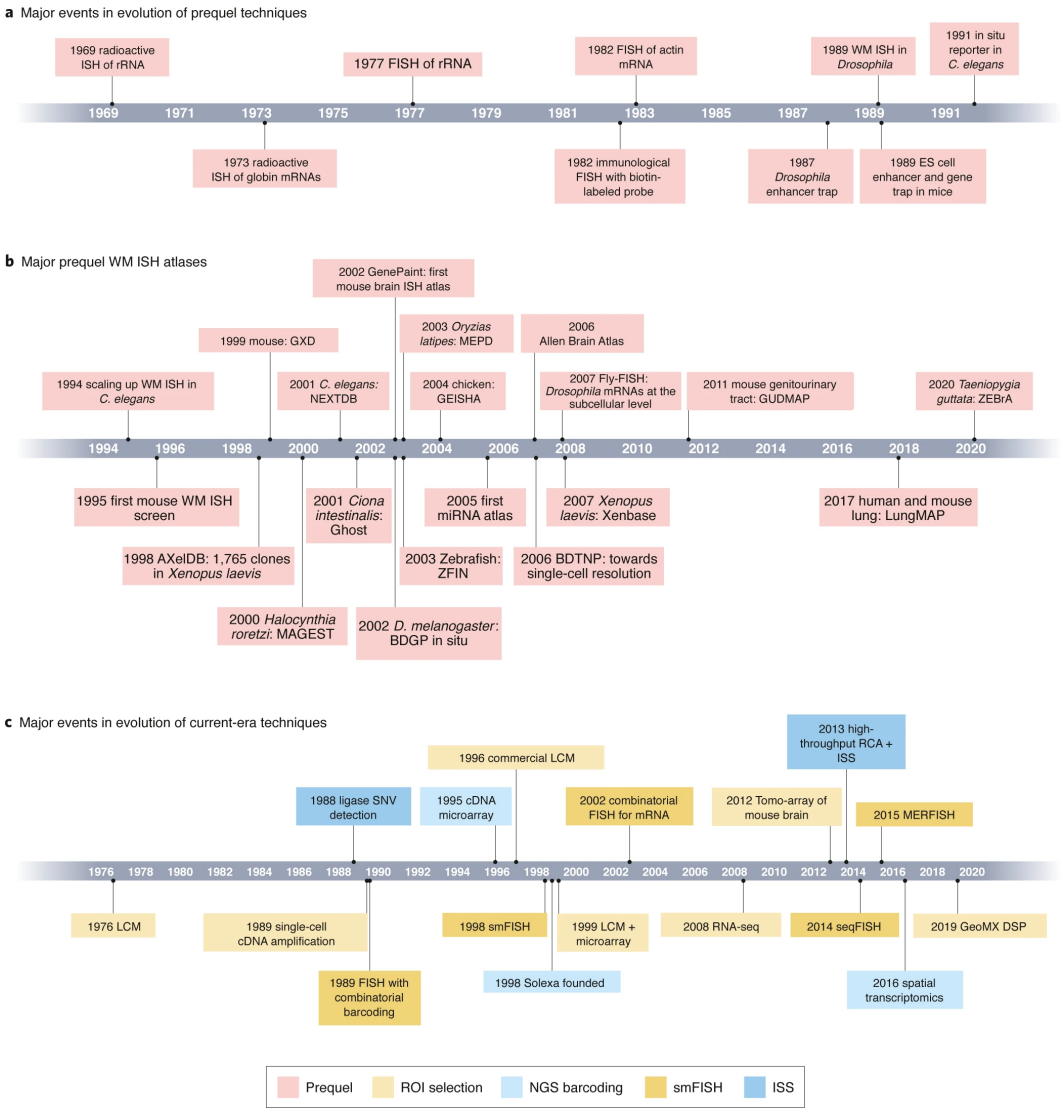

Figure 2.1: See Section 2.11 for caption.

By "spatial transcriptomics", we mean attempts to quantify mRNA expression of large numbers of genes within the spatial context of tissues and cells. Some important technologies enabling spatial transcriptomics date back to the 1970s (Chapter

4). Various forms of in situ hybridization (ISH) have been used for a long time to visualize gene expression in space. Radioactive ISH was first introduced in 1969, visualizing ribosomal RNA [6] and DNA [7] in Xenopus laevis oocytes, and was first used to visualize transcripts of specific genes (globin) in 1973 [8] (Figure 2.1A). Non-radioactive fluorescent or colorimetric ISH was developed in the 1970s and the early 1980s, improving spatial resolution, enabling 3D staining, and shortening required exposure times [9, 10] (Fig. 2.1a). Early ISH was performed in tissue sections, making it challenging to apply to blastulas and to reconstruct 3D tissue structures; whole mount ISH (WMISH) was first introduced in Drosophila in 1989 [11] and was soon adapted to other species such as mice in the early 1990s [12].

Another strand of development in early spatial transcriptomics was the enhancer and gene trap screen, which was developed in the 1980s when DNA sequencing throughput was increasing [13] and metazoan genomes were newly opened frontiers. The first screens in Drosophila [14] and mice [15] were performed in the late 1980s in order to visualize expression of untargeted, and often previously unknown, genes. With increasing throughput, enhancer and gene traps became the technology of choice for spatial transcriptomics in the 1990s, until the rise of (WM)ISH in the late 1990s which leveraged automation. WM(ISH) also avoided the need for transgenic lines, and was facilitated by the availability of reference genomes in the early 2000s for computational probe design. Although now eclipsed by newer methods, enhancer trap, gene trap, and in situ reporter methods have been used to build reference databases of gene expression and enhancer usage patterns in transgenic lines throughout the 2000s and 2010s [16].

The foundation for many current era technologies was built in the decades between the 1970s and the 2000s (Fig. 2.1c). For example, UV laser was first used to cut tissue in 1976 [17]. Popular IR and UV LCM systems were first reported in 1996 [18, 19] and were soon commercialized. Some highly multiplexed smFISH technologies such as seqFISH [20] rely on combinatorial barcoding, i.e. encoding each gene with a combination of colors so transcripts of more genes than easily discernible colors (up to 5) can be quantified simultaneously. Combinatorial barcoding was first reported in immunological DNA FISH in 1989 [21] and was first used for transcripts in 2002 [22]. The first unequivocal demonstration of smFISH showing each mRNA molecule as a spot was reported in 1998 [23]. Highly multiplexed smFISH would not have been possible without the development of these technologies.

(WM)ISH was the technology of choice in the late 1990s and the 2000s before the

rise of highly multiplexed, high resolution, and more quantitative technologies, and has been used to create gene expression atlases in embryos of several species such as Drosophila melanogaster [24], Mus musculus, and Gallus gallus [25], in various mouse organs such as the brain [26], genitourinary tract [27], and lung [28], and for specific types of genes such as miRNAs [29] (Fig. 2.1b). For many species other than mice and humans and miRNAs, the only spatial transcriptomics resources currently available may still be (WM)ISH atlases. Model organism databases collecting the proliferating gene expression patterns from various sources were also established in this period, such as gene expression database (GXD) [30] and Zebrafish Information Network (ZFIN) [31] (Fig. 2.1b). The golden age of (WM)ISH seems to have ended in the 2010s (Fig. 2.1b), perhaps due to some of the disadvantages of (WM)ISH, such as requiring stereotypical tissue structure, the need for thousands of animals to generate an atlas, and the largely qualitative nature of results.

Early motivating applications for spatial transcriptomics included identification of genes with restricted patterns which indicated function in development, identification of novel cell type markers, and identification of novel cell types not evident from tissue morphology [14, 15]. In the 1980s and 1990s, analyses were typically done manually, although more recently automated methods have been developed (Chapter 5). Convergence of strands of technologies including more powerful computing infrastructure, decreasing cost of sequencing, and the generation of more quantitative data, have mainstreamed and revolutionized spatial transcriptomics and opened up new possibilities. However, the legacy of the prequel era still lives on, in usage of prequel resources such as referencing the Allen Brain Atlas (ABA) [32] and the Allen mouse Common Coordinate Framework (CCF) [33], and in institutional legacy such as the Allen Brain Institute and the Jackson Laboratory which are contributing to the current era [34, 35].

2.3 Data collection

Current era technologies broadly fall into five categories in terms of how spatial information is acquired: region of interest (ROI) selection (Section 7.1), smFISH (Section 7.2), in situ sequencing (ISS) (Section 7.3), next generation sequencing (NGS) with spatial barcoding (Section 7.7), and methods not requiring *a priori* spatial locations (Section 7.6). Developers of such technologies often seek to enable a trifecta of transcriptome wide profiling, single-cell resolution, and high gene detection efficiency. While this achievement appears to be increasingly within reach, current era technologies are characterized by trade-offs between these goals.

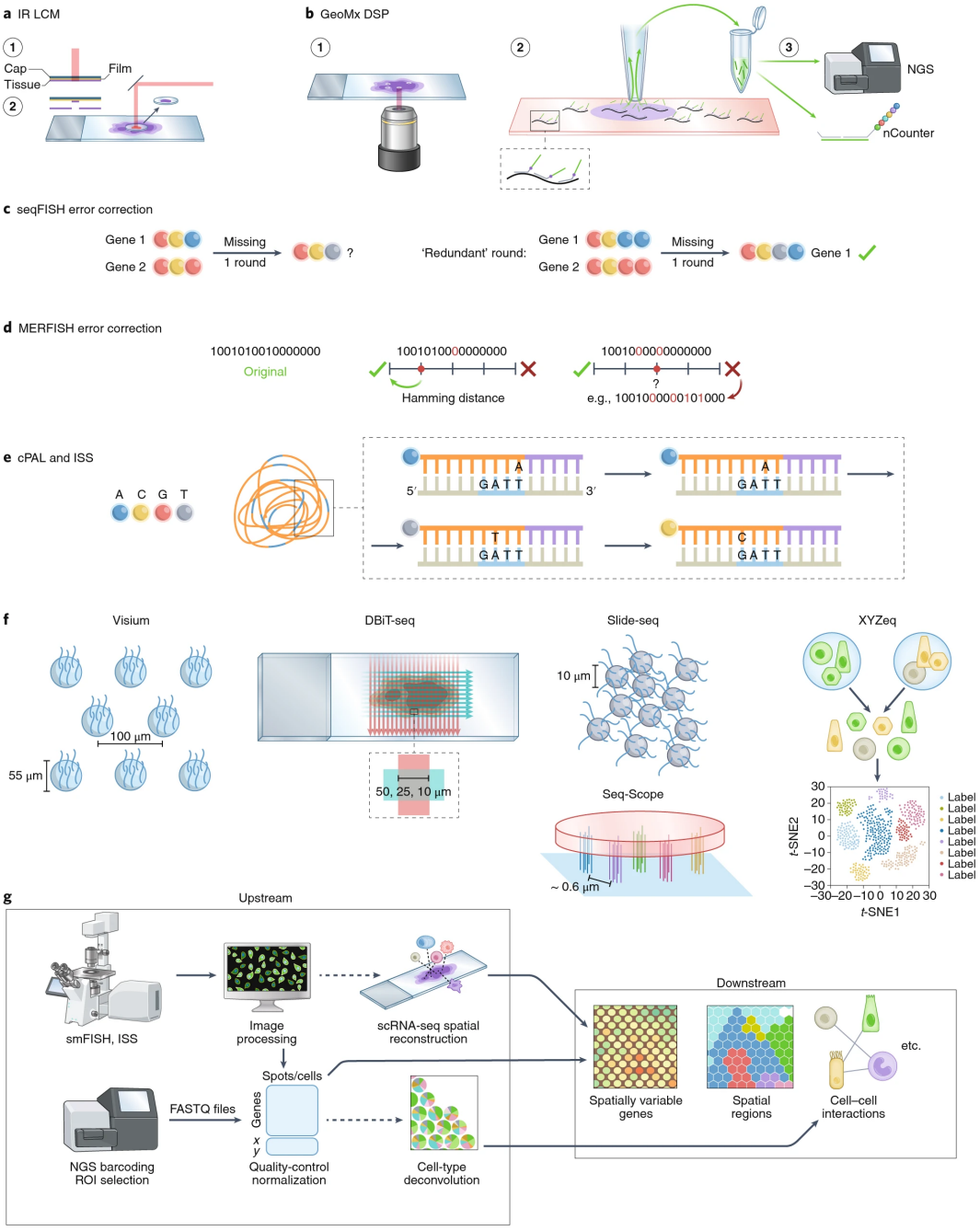

Figure 2.2: See Section 2.11 for caption.

ROI selection

Spatial locations can be obtained by selection and isolation of ROIs of known locations and shapes, which can be performed by physical (Section 7.1) and optical marking of ROIs for isolation (Section 7.1). The isolated ROIs can then be analyzed with cDNA microarray or RNA-seq, or dissociated into single-cells for scRNA-seq.

Physical microdissection includes LCM, 2000s voxelation [36], and Tomo-seq [37], which sections a tissue with a cryotome along an axis of interest, followed by RNA-seq on each section. Since 1999, by far the most widely used microdissection technology is LCM, which has been applied to various biological fields such as oncology, neuroscience, immunology, developmental biology, and botany (see Chapter 8 for topic modeling of PubMed and BioRxiv LCM literature). In LCM, ROIs in the tissue section are dissected by either UV laser cutting (Zeiss and Leica) or fusion of tissue with a membrane by IR laser (Arcturus, Fig. 2.2a); the two are combined in recent versions of Arcturus where IR fusion removes the ROI cut by UV. Combining LCM and Tomo-seq, spatial transcriptome in 3D can be profiled as in Geo-seq [38], albeit with limited spatial resolution. An innovative physical microdissection method is STRP-seq [39], which slices adjacent tissue sections into stripes at different angles and reconstructs gene expression patterns in 3D with an algorithm inspired by ray-based computerized tomography. On the other hand, manual dissection is commonly used to profile gene expression along one spatial axis of interest in plants [40].

Optical marking of ROIs includes Niche-seq [41], which uses two photon irradiation to mark ROIs in tissue from transgenic mice expressing photoactivable GFP (PA-GFP) and then uses fluorescence activated cell sorting (FACS) to isolate cells with activated PA-GFP for scRNA-seq. Similar to Niche-seq but without transgenic mice is SPACECAT [42], which stains cultured live cells or organoids with photocaged fluorophores and photoactivates ROIs for FACS and scRNA-seq. Also using photocaging, ZipSeq [43] attaches anchor oligonucleotides with photocaged overhangs to tissue with antibodies or lipid insertion, and adds spatial "zipcodes" to photoactivated ROIs hybridizing to the overhangs. A more popular commercial optical ROI selection technique is GeoMX digital spatial profiler (DSP) [44] and whole transcriptome atlas (WTA) [45] of Nanostring (Fig. 2.2b), which shines UV light on ROIs to release photo-cleavable gene barcodes for quantification with either nCounter or with NGS. As GeoMX uses pre-defined gene panels rather than poly-A capture, Nanostring provides the Cancer Transcriptome Atlas (CTA) gene panel with over 1800 genes as well as human and mouse whole transcriptome panels with over 18,000 genes.

Single Molecule FISH

Chronologically, the next technology developed in the current era is highly multiplexed single-molecule FISH (smFISH), which began with a 2012 prototype (seq-

FISH) that relied on super-resolution microscopy (SRM) to simultaneously profile 32 genes in yeast by hybridizing probes with different colors to transcripts, and then deducing relative locations of the colors present [46]. SRM is no longer needed; in 2014 seqFISH [20] was published, in which one color per gene is visualized per round of hybridization and the probes are stripped before the next round for the next color in the barcode. All transcripts of the same gene have the same barcode. With 4 colors, 8 rounds of hybridization $4^8 = 65536$ are more than enough to encode all genes in the human or mouse genome. In practice, an error correcting round of hybridization is performed, so that genes can still be distinguished if signal from one round of hybridization is missing [47] (Fig. 2.2c). More recently in a version of seqFISH based on RNA SPOTS [48], the "colors" themselves are one hot encoded by a sequence of hybridizations, expanding the palette to 20 "colors" per channel and enabling the profiling of 10,000 genes [49].

Another smFISH technique is multiplexed error-robust FISH (MERFISH) [50], which uses a different barcoding strategy, in which each gene is encoded by a binary code. The color codes in each experiment must be separated by a Hamming distance (HD) of 4 to allow for correction of missing signal in one round, and by 2 to identify error without the facility for correcting it (Fig. 2.2d). The length of barcodes can be increased to encode 10,000 genes [51]. As only the fluorophores are removed but the probes are not stripped, numerous rounds of hybridization in MERFISH are less time consuming than those in seqFISH. Most other smFISH-based techniques, such as HybISS [52] and split-FISH [53], use either seqFISH-like or MERFISH-like barcoding.

SmFISH faces a number of challenges, which have been addressed by various methods: signal-to-noise ratio can be improved with rolling circle amplification (RCA) [52], branched DNA (bDNA) [54], hybridization chain reaction (HCR) [47], primer exchange reaction [55], and tissue clearing [56]. With an increasing number of genes profiled, the transcript spots are increasingly likely to overlap, causing optical crowding. This can be mitigated by expansion microscopy (ExM) [57], only imaging a subset of probes at a time and using computational super-resolution [49], imaging highly expressed genes without combinatorial barcoding [50], and computationally resolving overlapping spots [58].

In Situ Sequencing

ISS methods yield spatial transcriptome information by sequencing, typically by ligation (SBL), gene barcodes (targeted), or short fragments of cDNAs (untargeted) *in situ*. Such methods rely on ligase only joining two pieces of DNA—a primer with known sequence and a probe—if they match the template, with non-matching probes washed away. The probes used are degenerate except for one or two query bases encoded by a color. RCA is commonly used for signal amplification. The 2013 ISS [59], later commercialized by Cartana, and BOLORAMIS [60] use one query base per probe as in cPAL [61] to sequence gene barcodes (Fig. 2.2e). FISSEQ [62] and a later adaptation with ExM called ExSeq [63] use SOLiD, which uses two query bases per probe to sequence circularized and RCA amplified cDNAs. In STARmap [64], gene barcodes are sequenced by SEDAL, in which SOLiD-like two query bases are used to reject error, but one base encoding can also be used. BARseq also RCA amplifies probes with gene barcodes, but uses sequencing by synthesis (SBS) instead of SBL to sequence the barcodes [65].

NGS with spatial barcoding

Spatial locations of transcripts can also be preserved by capturing the transcripts from tissue sections on in situ arrays. Such arrays can be manufactured by printing spot barcodes, UMIs, and poly-T oligos on commercial microarray slides to capture polyadenylated transcripts, as in the Spatial Transcriptomics (ST) and Visium technologies (Fig. 2.2f). They can also be Drop-seq-like beads [66] with split pool barcodes, UMIs, and poly-T oligos spread on slides in a single layer (e.g. Slide-Seq [67]) or confined in wells etched on the slides (e.g. HDST [68]), with bead barcodes subsequently located using in situ SBL. Alternatively, in DBiT-seq [69], an array is generated by microfluidic channels, which are used to deposit one type of barcode in one direction, and then another in a perpendicular direction, with the orthogonal barcodes ligated so each spot can be identified with a unique pairwise combination. While NGS barcoding techniques are typically designed for 3' end Illumina sequencing, Visium has been adapted to Nanopore long read sequencing [70].

NGS barcoding techniques have been applied to large areas of tissue [33], and their use is increasing (Fig. 2.4b). Nevertheless, they do not have single-cell spatial resolution. The commonly used Visium has spots in a hexagonal array with diameter 55 μm 100 μm center to center (Fig. 2.2f). Bead diameter is 10 μm in Slide-seq, and 2 μm in HDST (Fig. 2.2f). Slide-seq and HDST use bead sizes smaller than single-cells, but they may not always provide single-cell resolution because one

bead can span two or more cells. Resolution of DBiT-seq is determined by channel width (either 50, 25, or 10 µm, Fig. 2.2f). More recently, the spot size can be reduced to below 1 µm, with RCA amplified DNA nanoballs as small as 0.22 µm across with spot barcodes deposited in wells 0.5 or 0.715 µm apart in Stereo-seq [71], and in Seq-Scope polonies with spatial barcodes 0.6 µm center to center on an Illumina flow cell re-purposed to capture transcripts from tissue sections71 (Fig. 2.2f). Another polony based method PIXEL-seq achieves a spot diameter of about 1.22 µm but unlike in the flow cell, PIXEL-seq does not have much spacing around each polony [72]. Techniques such as XYZeq [73] and sci-Space [74] have been developed to dissociate the single-cell or nuclei in spatially barcoded spots for scRNA-seq, so the data has single-cell transcriptomic but not spatial resolution (Fig. 2.2f).

De novo reconstruction of spatial information

Some technologies have been developed to preserve information necessary to computationally reconstruct spatial gene expression patterns without knowing or collecting spatial locations. One such technology is DNA microscopy [75, 76], which records proximity between cDNAs. This information can be used to reconstruct relative locations of transcripts. At the cellular level, gene expression in rare cell types can be reconstructed by deliberately assaying multiplets, and then mapping them to locations in a spatial reference based on gene expression of cells from common cell types attached to cells from the rare cell types [77]. Variants of the term "spatial transcriptomics" have also been used to describe techniques localizing transcripts to organelles (e.g., APEX-seq [78]), although no spatial coordinates are recorded.

Multi-omics

The transcriptome is only one aspect of cell function. Other aspects, such as the proteome, neuronal connectome, and 3D chromatin conformation are also important to cell function, and some methods have been developed to profile them along with the transcriptome in the same cells (Section 7.8). For the proteome, oligo-tagged antibodies are used to detect proteins of interest, and the oligonucleotide signifying the protein species can be detected with smFISH-based methods. Such antibody panels have been combined with a variant of ST as DBiT-seq [69], SM-Omics [79], GeoMX DSP [44], and MERFISH [56]. With the oligonucleotide barcode, over 100 antibodies can be used, such as when using all available antibody panels for GeoMX DSP. For 3D chromatin conformation, MERFISH and seqFISH+ have been adapted

to visualize chromatin structure, by targeting DNA genomic loci [80] or introns of nascent transcripts [80, 81]. For the neuronal connectome, multiplexed transcript quantification can also be combined with neuron projection tracing. For instance, cholera toxin subunit b (CTb) retrograde tracing has been used in conjunction with MERFISH to visualize axons [82]. Also, BARseq was originally designed to use ISS for axon tracing by sequencing neuron specific barcodes introduced by a virus injected into the brain, but was later adapted to sequence gene barcodes [65] as well. In addition, while not an -ome per se, electrophysiology has been recorded prior to transcriptome profiling in the same cells, such as with patch-clamp in explanted human neurons followed by HCR-smFISH [83] and with extracellular electrodes in cultured cardiomyocytes followed by STARmap in electro-seq [84].

2.4 Comparison across categories

In this section we discuss trade-offs made by different types of technologies among high detection efficiency, transcriptome wide profiling, high spatial resolution, and sometimes larger tissue area, as well as practical factors relevant to selection of technology such as FFPE compatibility and cost/usability.

Detection efficiency

Detection efficiency is commonly estimated by performing non-barcoded smFISH with near 100% sensitivity for select marker genes on the same cell type and comparing the average number of transcripts detected for each gene per cell for techniques where cells can be segmented, or per unit tissue area for techniques without single-cell resolution. For NGS based techniques with UMI, sometimes the number of UMIs and genes detected per cell or unit area is compared with that of other techniques with UMI. Note that comparisons of efficiencies are confounded by different tissues and methods used to estimate efficiencies in different studies and by different sequencing depths in NGS.

Highly multiplexed smFISH techniques tend to excel in this area, with 95% for Hamming distance 4 MERFISH [85] compared to non-barcoded smFISH; multiple rounds of hybridization tend to decrease the efficiency in part because barcodes with incorrigible errors are discarded. NGS barcoding techniques tend to have lower efficiency. The efficiency of ST is estimated to be 6.9% per area compared to smFISH for select genes in the same tissue type [86], comparable to that of scRNA-seq. Visium's efficiency seems to be moderately higher than that of ST, and DBiT-seq's is even higher, at 15.5% per area compared to smFISH [69]. Efficiencies

of the submicron techniques, in the number of UMIs per unit area in the same tissue, might be comparable to that of Visium [72]. ISS tends to be less efficient, in part because of inefficiency of reverse transcription (RT) and SBL. Whereas detection efficiency of scRNA-seq techniques is between 3% - 25%65 [66, 87, 88, 89, 90], the detection efficiencies of Cartana ISS and FISSEQ [91] are 5% and 0.005% respectively, with STARmap only marginally better than scRNA-seq. However, ExSeq claims up to 62% efficiency compared to smFISH per cell for genes tested [63]. Recent development tends to skip RT and make ligation of padlock probe on RNA template more efficient, such as in BOLORAMIS and HybRISS [73], or substitute SBL with seqFISH-like barcoding as in HybISS, to improve detection efficiency.

Transcriptome wide profiling

Techniques not targeting specific gene with a panel of known probes are transcriptome wide, such as ROI selection followed by NGS and NGS barcoding where NGS is performed on poly-A captured transcripts, as well as untargeted ISS such as FISSEQ and untargeted ExSeq. However, these transcriptome wide techniques tend to have lower detection efficiency. It is possible to use certain techniques that require gene probe panels to quantify transcripts of over 10,000 genes, such as seqFISH+, MERFISH, and GeoMX WTA, though unlike in NGS, novel transcripts not targeted by the probes cannot be detected. While GeoMX WTA has been used in some studies outside Nanostring[92], overall the number of genes profiled with smFISH-based techniques per dataset has not increased over time (Fig. 2.3g). Instead, in studies using smFISH and ISS based techniques, a smaller number of genes are profiled and the smFISH or ISS dataset is complementary to a transcriptome wide scRNA-seq dataset [93]. The number of genes that can be detected by highly multiplexed smFISH is limited by optical crowding, and expansion microscopy was used to address this issue in MERFISH and ExSeq. However, expansion reduces the amount of tissue covered per field of view, thus limiting imaging throughput.

Spatial resolution

SmFISH and ISS based techniques have single-cell and single molecule resolution, although cell segmentation can be challenging. In addition, smFISH and ISS based techniques can be applied to cleared thick tissue sections [56], though the number of genes profiled in these cases are much smaller than in most 2D highly multiplexed smFISH studies. All other types of techniques require tissue sections and are thus

limited to 2D, or 3D with z resolution limited to section thickness which is usually at least 10 µm for frozen sections. While there are submicron resolution NGS barcoding techniques and the ROIs of LCM and GeoMX can in principle be single-cell or smaller, the most common usage of these types of techniques tend to have lower spatial resolution, such as 55 µm for Visium and several hundred microns across for GeoMX (e.g. 700 × 800 µm in [92] due to insufficient sensitivity of transcript detection at the single-cell or subcellular resolution [94].

Tissue area

Overall, techniques with lower detection efficiencies tend to be better at profiling larger tissue area, and for smFISH, there seems to be a trade-off between number of cells and number of genes. In current era spatial transcriptomics, a tissue section several millimeters across, such as a substantial portion of a mouse brain coronal section, which can fit into a Visium or ST tissue capture area, is considered large, and increasing tissue area and sequencing depth for sensitivity would increase sequencing cost. Cartana ISS and HybISS have also been used to profile large areas of tissue several millimeters across but only around 100 genes [95]. An advantage of (Hyb)ISS here is the strong RCA signal and less optical crowding thanks to lower detection efficiency facilitating lower magnification (20x, while MERFISH uses 60x) and thus faster imaging. While most highly multiplexed smFISH datasets remain at 100s of genes (Fig. 2.3g), among studies that reported the number of cells, the total number of cells per study have increased (Fig. 2.3h, $p < 0.001$, two sided t-test). ROI selection techniques are generally used for small numbers of ROIs as it's labor intensive to select very large numbers of ROIs and process them separately without spatial barcoding. However, when high spatial resolution is not as crucial or practical, ROIs with very low resolution can be selected to cover more tissue, as in the LCM dataset in the Allen Human Brain Atlas [96].

Usability

While most techniques were originally developed for frozen sections, some are compatible with FFPE, which as a common tissue archive, may at times be the only type of tissue available. Among smFISH-based techniques, ACD RNAscope [97] is FFPE compatible but can only profile 12 genes at a time in FFPE as opposed to 48 in frozen sections. Among NGS barcoding techniques, Visium [98] and DBiT-seq [99] are FFPE compatible, but due to crosslinking and RNA fragmentation in archival storage, detection efficiency in FFPE tissue in UMIs and genes detected

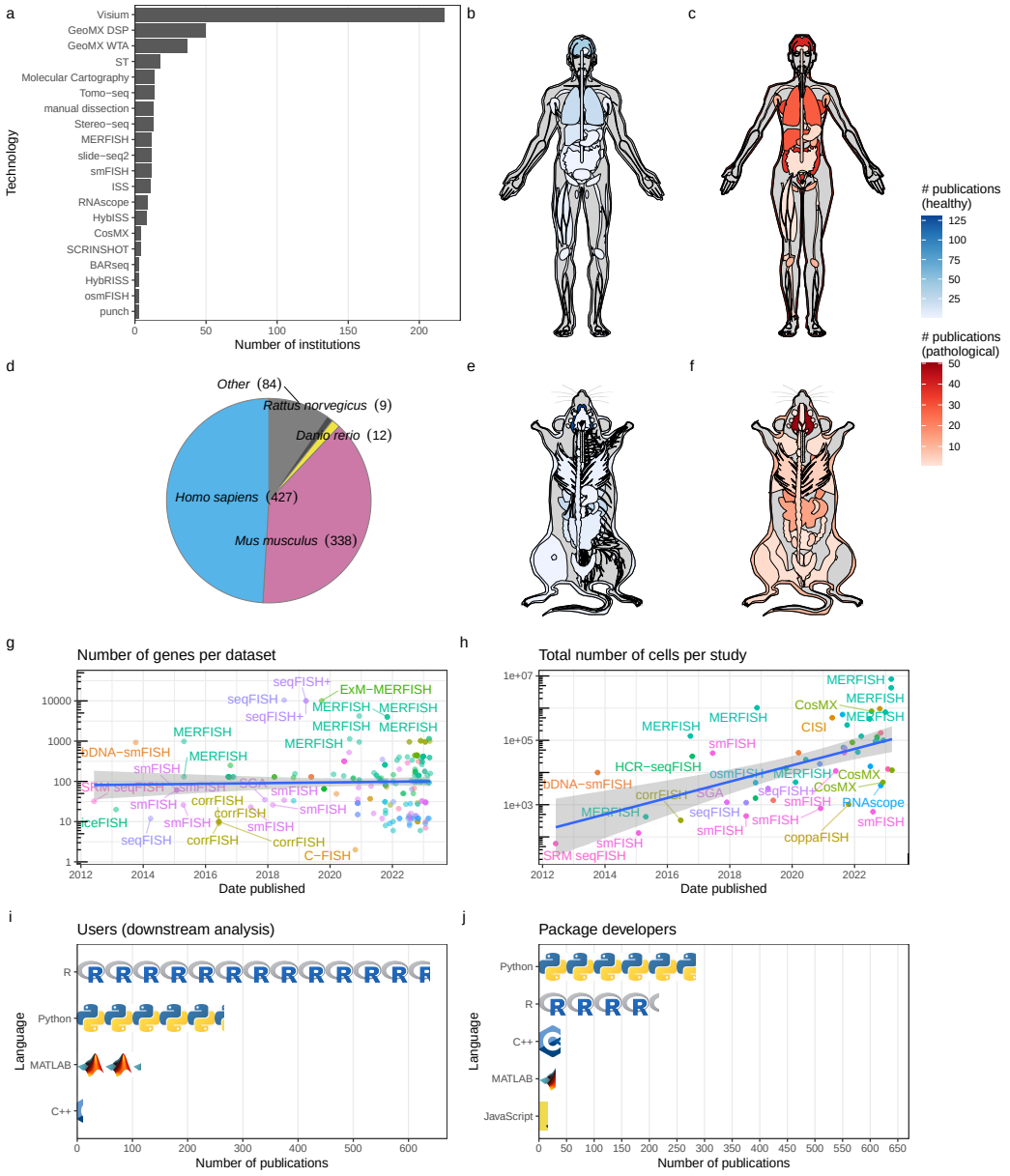

Figure 2.3: See Section 2.11 for caption.

per spot is about 5 to 10 times lower than in their frozen counterparts. LCM has long been applied to FFPE tissues, even at single-cell resolution with the sensitive SMART-3Seq [100]. GeoMX is not only FFPE compatible but also predominantly used on pathological human FFPE tissues (Fig. 7.8).

While many new techniques have been developed, most never spread beyond their institutions of origin (Fig. 6.9). Among those that did spread far and wide, the most popular ones tend to have commercial platforms, such as LCM, 10X Visium

and its precursor ST, Cartana ISS (acquired by 10X), and Nanostring GeoMX (Fig. 2.3). In addition, many major institutions have core facilities for NGS, if not LCM, Visium, and GeoMX (e.g. the TPCL at UCLA and the Advanced Genomics Core at University of Michigan, Ann Arbor), reducing cost of purchasing new equipment and training personnel in individual laboratories. Tomo-seq has also spread, perhaps due to its ease of implementation with standard equipment. In contrast, smFISH-based techniques have not spread as much thus far, perhaps due to the complicated home built fluidic system, long imaging time, terabytes of images, and expensive probes. However, some smFISH techniques are being commercialized, with automated imaging and fluidic platforms, such as MERFISH commercialized by Vizgen and another smFISH-based technique in Molecular Cartography of Resolve Biosciences. In addition, Rebus Esper can be programmed to automate different smFISH technologies and can process images online as in Illumina sequencing, and has been used to automate osmFISH [101]. With the new automated commercial platforms, popularity of smFISH-based technique might rise, especially if such platforms are adopted by core facilities.

2.5 Data analysis

The processing and analysis of high-throughput spatial transcriptomics data requires novel methods and tools, especially for problems such as image preprocessing, spatial reconstruction of scRNA-seq data, cell type deconvolution of NGS barcoding data, identification of spatially variable genes, and inference of cell-cell interactions (Fig. 2.2g).

Upstream

Upstream data analysis converts raw data into forms more amenable to biological interpretation and is dependent on the data collection technology.

For smFISH and ISS based data, the raw data consists of images of fluorescent spots, which must be processed to identify transcript spots, match spots to genes, and assign spots to cells (Section 9.1). SmFISH and ISS studies often use classical image processing tools such as top-hat filtering to remove background, translation to align images from different rounds of hybridization, and watershed for cell segmentation [47, 64, 85]. Machine learning in Ilastik, deep learning packages like DeepCell [102], and alternative tools incorporating scRNA-seq data [103], can also be used for cell segmentation. However, without visualizing the plasma membrane, accuracy of cell segmentation is limited. Some analyses, such as identification of tissue

regions, can be performed without cell segmentation [103]. Until 2019, image processing was typically performed with poorly documented and technique specific code written in the proprietary language MATLAB, but more recently such code is increasingly written in the open source language Python. The package starfish [104] was developed as an attempt to provide a unified and well-documented user interface to process images from different techniques such as seqFISH, MERFISH, and ISS, but it has not been widely adopted.

Improvements in scRNA-seq technology have inspired new methods for leveraging the complementary nature of high-resolution transcriptome quantification with spatial transcriptomics data. For smFISH and ISS data that is not transcriptome wide, expression patterns of genes not profiled in the spatial data can be imputed with scRNA-seq data, either by mapping dissociated scRNA-seq cells to the spatial reference or by directly imputing gene expression in space using expression profiles from scRNA-seq (Section 9.3). Cells can be mapped to spatial locations on an existing spatial dataset with genes shared by the two datasets, with an ad hoc score favoring similarity between cell and location [105] or via optimal transport modeling [106]. While ad hoc scoring is simple to implement, the results tend to be qualitative. Gene expression in space can also be imputed from scRNA-seq without explicitly mapping scRNA-seq cells to locations. A common approach is to project the spatial and scRNA-seq data into a shared low-dimensional and batch-free latent space, and to subsequently estimate gene expression by projecting the spatial cells into the latent space. Examples of this approach include Seurat [32] and gimVI [107]. These methods may also be used to add spatial context to single-cell multi-omics data when spatial techniques for some of the multi-omics data are not available.

In spatial data without single-cell resolution, such as those derived from ST and Visium, scRNA-seq data can inform cell type composition of the spots or voxels (Section 9.4). Negative binomial models and non-negative least squares (NNLS) are common principles underlying cell type deconvolution methods. Negative binomial models are typically parameterized with rate and dispersion, and the rate is modeled as a weighted sum of cell type signatures from scRNA-seq, with scaling factors for library size and technology sensitivity; the non-negative weights may be normalized to sum up to 1 as cell type proportions per spot. Negative binomial based methods include stereoscope [108] and cell2location [109]. Simpler than negative binomial, gene expression is modeled as Poisson instead in RCTD [110]. Cell type deconvolution can also be performed by modeling gene expression at each spot as a

weighted sum of cell type signatures outside the rate parameter of negative binomial distributions, and the weights are inferred with NNLS. For example, AdRoit [111] uses the means of negative binomial distributions fitted to spot gene expression and to scRNA-seq cell type signatures. The cell type signatures can be non-negative matrix factorization (NMF) cell factors from scRNA-seq assigned to cell types, as in NMFreg66 and SPOTlight [112]. The cell type weights can be regularized or thresholded to limit the number of cell types assigned to each spot. Parallels can also be drawn between cell type deconvolution and topic modeling in text mining; cell types are analogous to topics, and genes are analogous to words. Latent Dirichlet allocation (LDA) from topic modeling has been adapted to cell type deconvolution, such as in STRIDE [65] and STdeconvolve [113]; the latter is unsupervised and does not require a scRNA-seq reference.

Downstream

Downstream analyses most often apply to the gene count matrix and cell/spot locations, and are thus largely independent from data collection technologies.

Given the relevance of scRNA-seq to spatial data and how spatial data is often analyzed like scRNA-seq data at the exploratory data analysis (EDA), popular scRNA-seq EDA ecosystems such as Seurat [32], Scanpy (Squidpy) [114, 115], and `SingleCellExperiment` (`SpatialExperiment`) [116] have added functionalities for spatial data, such as updates to data containers and functions to facilitate visualization of gene expression and cell/spot metadata at spatial locations (Section 9.2). EDA packages dedicated to spatial data with beautiful graphics and good documentation have also been written, such as Giotto [117] and STUtility [118]. Seurat and Giotto also implement basic methods to identify spatially variable genes. In addition, Giotto implements methods to identify cell type enrichment in ST and Visium spots, identify gene coexpression and association between gene expression and cell type colocalization, and to identify spatial regions [119].

Spatially variable genes are genes whose expression is associated with spatial location (Section 9.5). Three approaches are commonly used: Gaussian process regression (GPR) [120] and its generalization to Poisson [121] and NB [122], Laplacian score [123], and Moran's I. The former models normalized gene expression or the rate parameter of Poisson or NB gene expression as a GPR and finds whether the model better describes the data with the spatial term than without. The latter approach identifies genes whose expression better reflects the structure of a spatial

neighborhood graph. The locations of cells can also be modeled as a spatial point process with gene expression as marks; spatially variable genes can be identified as marks associated with locations [124]. Fitting GPR models to numerous genes can be time consuming, especially when a Bayesian approach with Markov chain Monte Carlo is used. Permutation testing used in Laplacian score based methods can also be time consuming. As both GPR and Laplacian score based methods seek to identify spatial autocorrelation, sometimes the classic spatial autocorrelation metric Moran's I is directly used to identify spatially variable genes, as in Seurat v3 and above. MERINGUE [125] uses a local version of Moran's I. Moran's I and its significance testing are implemented in established geospatial packages and are easy and fast to run, but may have less statistical power than model based methods [121].

Spatial information also enables identification of potential cell-cell interaction (Section 9.8). This is commonly done with knowledge of ligand-receptor (L-R) pairs and testing whether L-R pairs are more likely to be expressed in neighboring cells or spots [126], or whether two cell types each expressing the ligand and the receptor are more likely to colocalize [125]. The cross-type L function from spatial point process can be used to find cell types that colocalize [127]. Expression of genes of interest can also be modeled, including a term for cell-cell co-localization; the gene is considered associated with cell-cell co-localization if the model better describes the data with this term than without [128].

There are many other types of downstream analyses that are useful for spatial transcriptomics analysis, including identification of archetypal gene patterns (Chapter 9.6), spatial regions defined by the transcriptome (Chapter 9.7), inferring gene-gene interactions (Chapter 9.9), subcellular transcript localization (Chapter 9.10), and gene expression imputation from H&E images (Chapter 9.11).

2.6 Trends in the spatial transcriptomics field

The quality vs. quantity trade-off inherent in existing technologies means that there is no single "best" solution currently available, and the difficulty in implementing methods has resulted in many technologies never spreading beyond their institutions of origin. LCM, Visium, ST, GeoMX DSP, and Tomo-seq have been the most widely adopted (Fig. 2.3a), and in most cases in the US and western Europe (Figs. 6.12). In terms of tissues analyzed, multiplexed current era techniques have been used widely to characterize human tissues [129], tumors [86] (especially breast tumors),

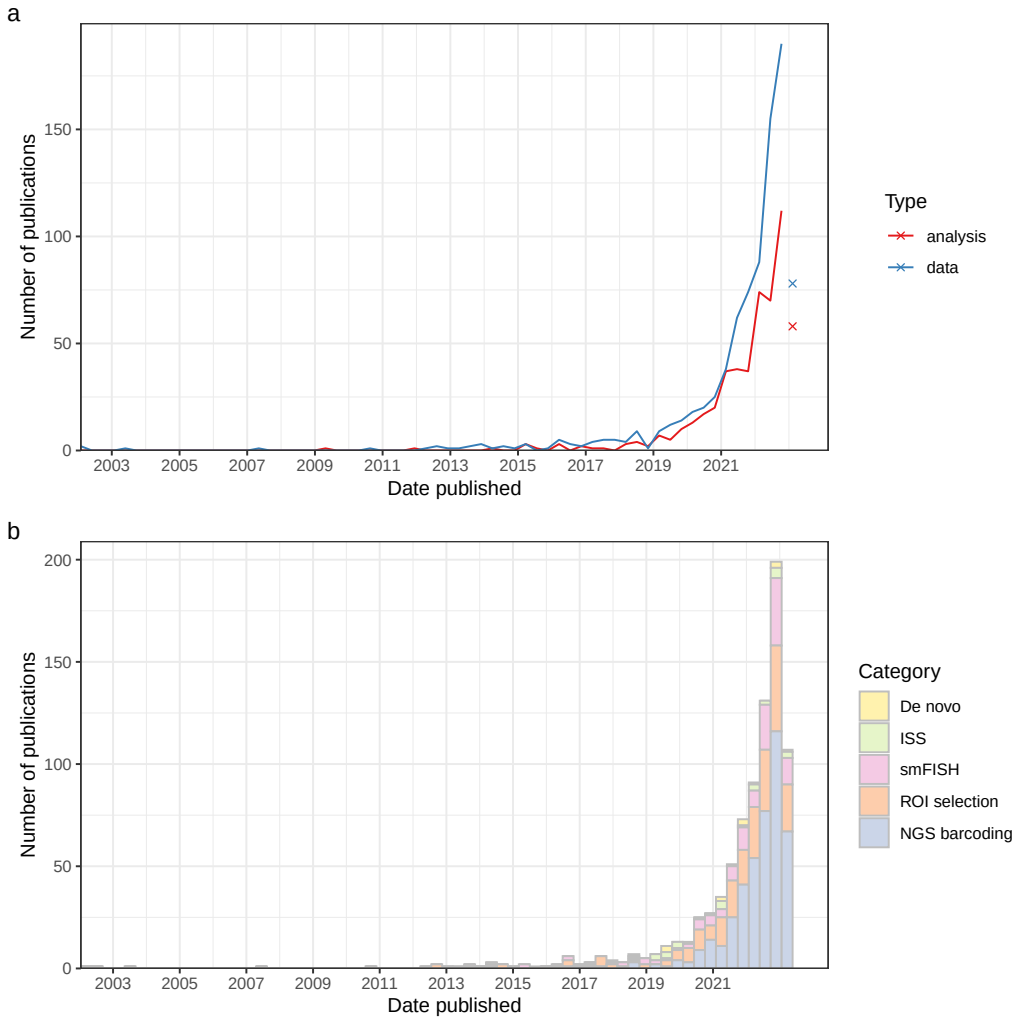

Figure 2.4: **Growth of the current era. a,** Number of publications over time for current-era data collection and data analysis. Bin width is 120 days; the curves drop because the plot was made at the beginning of a new bin. Non-curated LCM literature is excluded. **b,** The data collection curve in a, broken down by category of techniques. The colors are stacked and sorted in descending order of total number of publications using techniques in that category.

and pathological tissues that don't necessarily have a stereotypical structure [130] (Fig. 2.3b,c). In the SARS-CoV-2 pandemic, GeoMX DSP has been used for spatial transcriptomic profiling in lung autopsy of COVID victims [92].

Some of the processed data, and associated spatially variable genes, can be downloaded and visualized from SpatialDB [131]. Excluding LCM, the vast majority of current era studies were performed on either humans or mice (Fig. 2.3d), and the brain is the most studied healthy organ while the lung (COVID) and breast tumor

are also often studied in humans (Fig. 2.3b,e,f). In particular, the international project Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative - Cell Census Network (BICCN) is constructing a multi-modal atlas for the human, mouse, and non-human primate brain, including spatial data such as MERFISH and seqFISH.

All packages mentioned in the Data analysis section are open source and written in languages such as R, Python, and Julia. Downstream analyses in studies primarily concerning new data and data analysis packages predominantly use open source programming languages such as R, Python, and C++ (Fig. 2.3i,j). While MATLAB is still popular, its use does not rise as in R and Python (Fig. 9.12). While R is more popular for downstream analyses and EDA, Python and C++ are more popular for package development (Fig. 2.3i,j). Most of the packages are not hosted on standard repositories such as the Comprehensive R Archive Network (CRAN), Bioconductor, PyPI, or conda (Fig. 9.13). While most packages using R, Python, and C++ are well-documented, many MATLAB packages are not (Fig. 9.12). The standard repositories and documentation make packages more usable, and is discussed in more details in Section 9.12.

2.7 Future perspective

While technologies of the prequel are rapidly being deprecated, the ideas and methods that underlie them are fundamental to current era spatial transcriptomics. The field has dramatically expanded over the past 5 years (Fig. 2.4a), with a plethora of new techniques and popularization of Visium driving growth (Fig. 2.4b, Fig 6.9, 7.37, 10.1).

What lies ahead of the rising curves? First, more can be done to improve data collection techniques. For example, most current era techniques require tissue sections. Highly multiplexed whole mount smFISH and tissue clearing protocols, and more efficient computational tools for aligning multiple sections that may come from multiple individuals or even developmental stages, should be developed to extend current era techniques to 3D and to spatiotemporal analysis. Future techniques may also extend the current era from the scale of millimeters to centimeters and across other modalities such as epigenomics and metabolomics to give fuller pictures of cellular function. Furthermore, smFISH and ISS techniques, with signal amplification to reduce the number of probes per transcript, can be adapted to target isoform specific exons or untranslated regions rather than all transcripts of a gene.

Second, current era data has not yet been integrated into comprehensive databases. Prequel databases such as GXD and e-Mouse Atlas and Gene Expression (EMAGE) [132] include data from multiple sources and can be queried by gene symbol and developmental and spatial ontologies. In addition, ABA and EMAGE aligned ISH images to common coordinates and can be queried with expression patterns. While some current era authors provide online interactive visualization of datasets from their studies [33], comprehensive databases integrating, querying, and visualizing data from multiple sources as in the prequel era have not yet been developed. Furthermore, while prequel ontologies are still used in current era studies, such ontologies may be improved with the transcriptome wide quantitative data from the current era.

Third, outside of LCM, the current era is highly focused on humans and mice, with potential spatial transcriptomics investigations of other species such as plants and invertebrates lagging behind. Technological modernization of prequel consortia for organisms other than humans and mice holds much promise for the development of useful spatial transcriptomics atlases.

Fourth, an open source, well-documented, interoperable, and scalable workflow with an integrated easy-to-use interface would greatly simplify spatial transcriptomics data collection and analysis. At present, for tasks beyond EDA, users still often need to learn new syntax, convert object types, and even learn new languages to use some data analysis tools. Finally, our survey of methods shows that spatial transcriptomics methods need to be more open and accessible so that they become adopted around the world, and are not restricted to Western elite institutions.

2.8 Data availability

The database of spatial transcriptomics literature can be accessed here. The version used as of writing is in the metadata.xlsx file in the frozen DOI version of the GitHub repository to reproduce the figures in this paper and render the supplementary website.

2.9 Code availability

All code used to generate figures in this paper and render the supplementary website is in the GitHub repository here. The frozen DOI version of the repository as of final submission of this paper is on Zenodo.

2.10 Acknowledgements

This work was supported by a grant from the National Institute of Mental Health (NIMH), National Institute of Health (NIH), of the U.S. Department of Health & Human Services (number U19MH114830, L.P.). We thank the following people for providing feedback for earlier versions of this paper and the supplement: D. Furth from the Cold Spring Harbor Laboratories, L. Cai from the California Institute of Technology, and G. Victora from the Rockefeller University.

2.11 Figure legends

Figure 2.1: Timelines of major events a, Timeline of development of prequel era technologies. References: 1969 radioactive ISH [6, 7], 1973 goblin [8], 1977 FISH [10], 1982 immunological [9], 1982 FISH [133], 1987 enhancer trap [14], 1989 WMISH [11], 1989 ES cell [15], 1991 C. elegans [134]. **b,** Timeline of major (WM)ISH atlases and gene expression pattern databases. References: 1994 WMISH [135], 1995 mouse WMISH [136], 1998 AXelDB [137], 1999 GXD [138], 2000 Maboya Gene Expression patterns and Expression Sequence Tags (MAGEST) [139], 2001 Nematode Expression Pattern Database (NEXTDB) [140], 2001 GHOST [141], 2002 GenePaint [142], 2002 *D. melanogaster*: Berkeley *Drosophila* Genome Project (BDGP) [24], 2003 Medaka Expression Pattern Database (MEPD) [143], 2003 Zebrafish Information Network (ZFIN) [31], 2004 *Gallus* Expression *In Situ* Hybridization Analysis (GEISHA) [25], 2005 miRNA [29], 2006 Allen [26], 2006 Berkeley *Drosophila* Transcription Network Project (BDTNP) [144], 2007 FlyFISH [145], 2007 Xenbase [146], 2011 mouse Genitourinary Development Molecular Anatomy Project (GUDMAP) [27], 2017 LungMAP [28], 2020 Zebra finch Expression Brain Atlas (ZEBrA) [147]. **c,** Timeline of development of current era technologies and their notable precursors, colored by type of technology. References: 1976 LCM [17], 1988 ligase mediated single nucleotide variant (SNV) detection [148], 1989 amplification [149, 150], 1989 FISH [21], 1995 microarray [151], 1996 LCM [18, 19], 1998 smFISH [23], 1999 LCM [152], 2002 combinatorial [22], 2008 RNA-seq [153], 2012 Tomo-array [154], 2013 high-throughput RCA + ISS[59], 2014 seqFISH [20], 2015 MERFISH [50], 2016 ST [86], and 2019 GeoMX DSP [44].

Figure 2.2: Schematics of common current-era technologies. a, IR LCM. **b,** GeoMX DSP. The purple circle in step 2 is the UV-illuminated ROI. **c,** seqFISH barcoding and error correction scheme: if signal from one round of hybridization is missing, the remaining rounds can still uniquely identify the gene barcoded. **d,**

MERFISH Hamming distance 4 barcoding and error correction scheme: from the design of the barcodes, if signal from one round of hybridization is missing, the correct barcode can be recovered. If two rounds are missing, the remaining signals are equidistant to two different barcodes so the original barcode cannot be recovered. **e,** Cartana ISS with cPAL sequencing: many copies of the gene barcode are made with RCA for signal amplification, which are then sequenced in situ with cPAL. The orange line stands for the RCA amplicon. Short blue lines stand for the gene barcode. Brown stands for the probe; bases not labeled are degenerate. Gray stands for primer matching constant region. **f,** NGS barcoding techniques. In Visium, the spots are arranged in a hexagonal grid, 100 μm apart center to center and 55 μm in diameter. In DBiT-seq, positional barcodes are deposited in microfluidic channels and spatial resolution is determined by the width (down to 10 μm) and spacing of the channels. In Slide-seq, barcoded beads 10 μm in diameter are spread in a single layer on a slide. In XYZeq, spatial barcodes are conferred on multiple cells in wells 500 μm in diameter, which are then dissociated for scRNA-seq. In Seq-Scope, the tissue is mounted on a repurposed Illumina flow cell with barcoded polony spots 0.6 μm apart on average. For Visium and Slide-seq, the lines represent oligonucleotides attached to the slide or bead. For DBiT-seq, red and green lines represent the flow in microfluidic channels carrying barcoding oligonucleotides. For Seq-Scope, the tissue (pink block) is mounted on repurposed Illumina flow cell with bridge-amplified polonies each with its own spatial barcode represented by different colors. For XYZeq, different colors of the cells represent different spatial barcodes in the microwells, and the cells are dissociated for scRNA-seq. t-SNE, t-distributed stochastic neighbor embedding. **g,** Data-analysis workflow: upstream analysis is technology specific, and includes image processing for smFISH and ISS-based technologies, and FASTQ file processing, quality control of the gene count matrix, and data normalization for NGS-based technologies. Non-spatial scRNA-seq data can be integrated by mapping cells to locations with landmark genes in the smFISH or ISS data or deconvolving cell types in Visium spots. Downstream analyses tend to be technology agnostic, and include finding spatially variable genes, transcriptionally defined spatial regions, and cell–cell interactions. Created with BioRender.com.

Figure 2.3: Current-era metadata. a, Number of institutions that have published papers or preprints with each technique, excluding LCM literature too vast to be manually curated. Only techniques used by at least three institutions are shown. **b,** Number of publications for each healthy organ in humans (male shown here,

as there is no study on healthy female-specific organs in humans at present). **c,** Number of publications for pathological organs in humans (female shown here, but there are two studies on prostate cancer). **d,** Number of publications per species. **e,** Number of publications per healthy organ in mice. **f,** Number of publications for pathological organs in mice. **g,** Number of genes per dataset over time. Gray ribbon in g and h stands for 95% confidence interval. The slope is not significantly different from 0 in g (t test). In g and h, the y axis is log-transformed. **h,** Total number of cells per study profiled by smFISH-based techniques over time among studies that reported the number of cells. IceFISH, intron chromosomal expression FISH; C-FISH, consecutive FISH; MOSAICA, multi-omic single-scan assay with integrated combinatorial analysis; SGA, spatial genomic analysis; corrFISH, correlation FISH; EASI-FISH, expansion-assisted iterative FISH; par-seqFISH, parallel seqFISH; CISI, composite in situ imaging; SCRINSHOT, single-cell-resolution in situ hybridization on tissue; coppaFISH, combinatorial padlock-probe-amplified FISH. **i,** Number of publications for data collection using each of the fice most popular programming languages for downstream data analysis. **j,** Number of publications for data analysis using each of the five most popular programming languages for package development. In both i and j, each icon stands for 50 publications. Note that multiple programming languages can be used in one publication.

References

1. Liao J, Lu X, Shao X, Zhu L, and Fan X. Uncovering an Organ's Molecular Architecture at Single-Cell Resolution by Spatially Resolved Transcriptomics. Trends in Biotechnology 2020 Jun. DOI: 10.1016/j.tibtech.2020.05.006. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167779920301402

2. Asp M, Bergenstråhle J, and Lundeberg J. Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration. BioEssays 2020 Oct; 42:1900221. DOI: 10.1002/bies.201900221. Available from: https://doi.org/10.1002/bies.201900221

3. Smith EA and Hodges HC. The Spatial and Genomic Hierarchy of Tumor Ecosystems Revealed by Single-Cell Technologies. Trends in Cancer 2019 Jul; 5:411–25. DOI: 10.1016/j.trecan.2019.05.009. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2405803319301013

4. Lein E, Borm LE, and Linnarsson S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. 2017. DOI: 10.1126/science.aan6827

5. Saviano A, Henderson NC, and Baumert TF. Single-cell genomics and spatial transcriptomics: Discovery of novel cell states and cellular interactions in liver physiology and disease biology. Journal of Hepatology 2020 Nov; 73:1219–30. DOI: 10.1016/j.jhep.2020.06.004. Available from: https://linkinghub.elsevier.com/retrieve/pii/S016882782030372X

6. Gall JG, Lou M, Kline P, and Giles NH. Formation and Detecction of RNA-DNA Hybrid Molecules in Cytological Preparations. PNAS 1969; 63:378–83. DOI: 10.1073/pnas.63.2.378. Available from: https://www.pnas.org/content/63/2/378

7. John HA, Birnstiel ML, and Jones KW. RNA-DNA hybrids at the cytological level. Nature 1969; 223:582–7. DOI: 10.1038/223582a0. Available from: https://www.nature.com/articles/223582a0

8. Harrison P, Conkie D, Paul J, and Jones K. Localisation of cellular globin messenger RNA by in situ hybridisation to complementary DNA. FEBS Letters 1973 May; 32:109–12. DOI: 10.1016/0014-5793(73)80749-5. Available from: https://doi.org/10.1016/0014-5793(73)80749-5

9. Langer-Safer PR, Levine M, and Ward DC. Immunological method for mapping genes on Drosophila polytene chromosomes. Proceedings of the National Academy of Sciences 1982 Jul; 79:4381–5. DOI: 10.1073/pnas.79.14.4381. Available from: https://www.pnas.org/content/79/14/4381

10. Rudkin G and Stollar BD. High resolution detection of DNA–RNA hybrids in situ by indirect immunofluorescence. Nature 1977; 265:472–3. DOI: 10.1038/265472a0. Available from: https://doi.org/10.1038/265472a0

11. Tautz D and Pfeifle C. A non-radioactive in situ hybridization method for the localization of specific RNAs in Drosophila embryos reveals translational control of the segmentation gene hunchback. Chromosoma 1989 Aug; 98:81–5. DOI: 10.1007/BF00291041. Available from: https://link.springer.com/article/10.1007/BF00291041

12. Rosen B and Beddington RS. Whole-mount in situ hybridization in the mouse embryo: gene expression in three dimensions. Trends in Genetics 1993 May; 9:162–7. DOI: 10.1016/0168-9525(93)90162-B. Available from: https://linkinghub.elsevier.com/retrieve/pii/016895259390162B

13. Giani AM, Gallo GR, Gianfranceschi L, and Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. Computational and Structural Biotechnology Journal 2020 Jan; 18:9–19. DOI: 10.1016/j.csbj.2019.11.002. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2001037019303277

14. O'Kane CJ and Gehring WJ. Detection in situ of genomic regulatory elements in Drosophila. Proceedings of the National Academy of Sciences of the United

States of America 1987 Dec; 84:9123–7. DOI: 10.1073/pnas.84.24.9123. Available from: https://www.pnas.org/content/84/24/9123

15. Gossler A, Joyner A, Rossant J, and Skarnes W. Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. Science 1989 Apr; 244:463–5. DOI: 10.1126/science.2497519. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.2497519

16. Jenett A, Rubin GM, Ngo TT, Shepherd D, Murphy C, Dionne H, Pfeiffer BD, Cavallaro A, Hall D, Jeter J, Iyer N, Fetter D, Hausenfluck JH, Peng H, Trautman ET, Svirskas RR, Myers EW, Iwinski ZR, Aso Y, DePasquale GM, Enos A, Hulamm P, Lam SCB, Li HH, Laverty TR, Long F, Qu L, Murphy SD, Rokicki K, Safford T, Shaw K, Simpson JH, Sowell A, Tae S, Yu Y, and Zugates CT. A GAL4-Driver Line Resource for Drosophila Neurobiology. Cell Reports 2012 Oct; 2:991–1001. DOI: 10.1016/j.celrep.2012.09.011. Available from: http://dx.doi.org/10.1016/j.celrep.2012.09.011

17. Meier-Ruge W, Bielser W, Remy E, Hillenkamp F, Nitsche R, and Unsold R. The laser in the Lowry technique for microdissection of freeze-dried tissue slices. The Histochemical Journal 1976 Jul; 8:387–401. DOI: 10.1007/BF01003828. Available from: http://link.springer.com/10.1007/BF01003828

18. Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, and Liotta LA. Laser Capture Microdissection. Science 1996 Nov; 274:998–1001. DOI: 10.1126/science.274.5289.998. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.274.5289.998

19. Becker I, Becker KF, Röhrl MH, Minkus G, Schütze K, and Höfler H. Single-cell mutation analysis of tumors from stained histologic slides. Laboratory Investigation 1996 Dec; 75:801–7. Available from: https://pubmed.ncbi.nlm.nih.gov/8973475/

20. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, and Cai L. Single-cell in situ RNA profiling by sequential hybridization. 2014 Mar. DOI: 10.1038/nmeth.2892. Available from: https://www.nature.com/articles/nmeth.2892

21. Nederlof PM, Flier S van der, Wiegant J, Raap AK, Tanke HJ, Ploem JS, and Ploeg M van der. Multiple fluorescence in situ hybridization. Cytometry 1990; 11:126–31. DOI: 10.1002/cyto.990110115. Available from: http://doi.wiley.com/10.1002/cyto.990110115

22. Levsky JM. Single-Cell Gene Expression Profiling. Science 2002 Aug; 297:836–40. DOI: 10.1126/science.1072241. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1072241

23. Femino AM, Fay FS, Fogarty K, and Singer RH. Visualization of Single RNA Transcripts in Situ. Science 1998 Apr; 280:585 LP –590. DOI: 10.1126/science.280.5363.585. Available from: http://science.sciencemag.org/content/280/5363/585.abstract

24. Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu SQ, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, and Rubin GM. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome biology 2002 Dec; 3:research0088.1. DOI: 10.1186/gb-2002-3-12-research0088. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-12-research0088

25. Bell GW, Yatskievych TA, and Antin PB. GEISHA, a whole-mount in situ hybridization gene expression screen in chicken embryos. Developmental Dynamics 2004 Mar; 229:677–87. DOI: 10.1002/dvdy.10503. Available from: http://doi.wiley.com/10.1002/dvdy.10503

26. Lein ES et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature 2007 Jan; 445:168–76. DOI: 10.1038/nature05453. Available from: http://www.nature.com/articles/nature05453

27. Harding SD, Armit C, Armstrong J, Brennan J, Cheng Y, Haggarty B, Houghton D, Lloyd-MacGilp S, Pi X, Roochun Y, Sharghi M, Tindal C, McMahon AP, Gottesman B, Little MH, Georgas K, Aronow BJ, Potter SS, Brunskill EW, Southard-Smith EM, Mendelsohn C, Baldock RA, Davies JA, and Davidson D. The GUDMAP database - an online resource for genitourinary research. Development 2011 Jul; 138:2845–53. DOI: 10.1242/dev.063594. Available from: http://golgi.ana.%20http://dev.biologists.org/cgi/doi/10.1242/dev.063594

28. Ardini-Poleske ME, Clark RF, Ansong C, Carson JP, Corley RA, Deutsch GH, Hagood JS, Kaminski N, Mariani TJ, Potter SS, Pryhuber GS, Warburton D, Whitsett JA, Palmer SM, and Ambalavanan N. LungMAP: The Molecular Atlas of Lung Development Program. American Journal of Physiology-Lung Cellular and Molecular Physiology 2017 Nov; 313:L733–L740. DOI: 10.1152/ajplung.00139.2017. Available from: https://www.physiology.org/doi/10.1152/ajplung.00139.2017

29. Wienholds E. MicroRNA Expression in Zebrafish Embryonic Development. Science 2005 Jul; 309:310–1. DOI: 10.1126/science.1114519. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1114519

30. Ringwald M, Baldock R, Bard J, Kaufman M, Eppig JT, Richardson JE, Nadeau JH, and Davidson D. A database for mouse development. Science 1994 Sep; 265:2033–4. DOI: 10.1126/science.8091224. Available from: https://science.sciencemag.org/content/265/5181/2033.abstract

31. Sprague J. The Zebrafish Information Network (ZFIN): the zebrafish model organism database. Nucleic Acids Research 2003 Jan; 31:241–3. DOI: 10.1093/nar/gkg027. Available from: http://zfin.org/zf_info/%20https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg027

32. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, and Satija R. Comprehensive Integration of Single-Cell Data. Cell 2019 Jun; 177:1888–902. DOI: 10.1016/j.cell.2019.05.031. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867419305598

33. Ortiz C, Navarro JF, Jurek A, Märtin A, Lundeberg J, and Meletis K. Molecular atlas of the adult mouse brain. Science Advances 2020 Jun; 6:eabb3446. DOI: 10.1126/sciadv.abb3446. Available from: www.brain-map.org%20https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.abb3446

34. Callaway EM et al. A multimodal cell census and atlas of the mammalian primary motor cortex. Nature 2021 598:7879 2021 Oct; 598:86–102. DOI: 10.1038/s41586-021-03950-0. Available from: https://www.nature.com/articles/s41586-021-03950-0

35. Baker D, Al-Naggar IM, Sivajothi S, Flynn WF, Amiri A, Luo D, Hardy CC, Kuchel GA, Smith PP, and Robson P. A Cellular Reference Resource for the Mouse Urinary Bladder. bioRxiv 2021 Jan :2021.09.20.461121. DOI: 10.1101/2021.09.20.461121. Available from: http://biorxiv.org/content/early/2021/09/23/2021.09.20.461121.abstract

36. Brown VM, Ossadtchi A, Khan AH, Yee S, Lacan G, Melega WP, Cherry SR, Leahy RM, and Smith DJ. Multiplex Three-Dimensional Brain Gene Expression Mapping in a Mouse Model of Parkinson's Disease. Genome Research 2002 May; 12:868–84. DOI: 10.1101/gr.229002. Available from: http://www.genome.org/cgi/doi/10.1101/gr.229002

37. Junker JP, Noël ES, Guryev V, Peterson KA, Shah G, Huisken J, McMahon AP, Berezikov E, Bakkers J, and Van Oudenaarden A. Genome-wide RNA Tomography in the Zebrafish Embryo. Cell 2014. DOI: 10.1016/j.cell.2014.09.038

38. Peng G, Suo S, Chen J, Chen W, Liu C, Yu F, Wang R, Chen S, Sun N, Cui G, Song L, Tam PP, Han JDJ, and Jing N. Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. Developmental Cell 2016 Mar; 36:681–97. DOI: 10.1016/j.devcel.2016.02.020. Available from: http://dx.doi.org/10.1016/j.devcel.2016.02.020

39. Schede HH, Schneider CG, Stergiadou J, Borm LE, Ranjak A, Yamawaki TM, David FPA, Lonnerberg P, Laurent G, Tosches MA, Codeluppi S, and La Manno G. Spatial tissue profiling by imaging-free molecular tomography. bioRxiv 2020 Aug :2020.08.04.235655. DOI: 10.1101/2020.08.04.235655. Available from: http://biorxiv.org/content/early/2020/08/04/2020.08.04.235655.abstract

40. Hufnagel B, Marques A, Soriano A, Marquès L, Divol F, Doumas P, Sallet E, Mancinotti D, Carrere S, Marande W, Arribat S, Keller J, Huneau C, Blein T, Aimé D, Laguerre M, Taylor J, Schubert V, Nelson M, Geu-Flores F, Crespi M, Gallardo K, Delaux PM, Salse J, Bergès H, Guyot R, Gouzy J, and Péret B. High-quality genome sequence of white lupin provides insight into soil exploration and seed quality. Nature Communications 2020; 11:492. DOI: 10.1038/s41467-019-14197-9. Available from: https://doi.org/10.1038/s41467-019-14197-9

41. Medaglia C, Giladi A, Stoler-Barak L, De Giovanni M, Salame TM, Biram A, David E, Li H, Iannacone M, Shulman Z, and Amit I. Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. Science 2017 Dec; 358:1622–6. DOI: 10.1126/science.aao4277. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aao4277

42. Genshaft AS, Ziegler CGK, Tzouanas CN, Mead BE, Jaeger AM, Navia AW, King RP, Mana MD, Huang S, Mitsialis V, Snapper SB, Yilmaz ÖH, Jacks T, Van Humbeck JF, and Shalek AK. Live cell tagging tracking and isolation for spatial transcriptomics using photoactivatable cell dyes. Nature Communications 2021; 12:4995. DOI: 10.1038/s41467-021-25279-y. Available from: https://doi.org/10.1038/s41467-021-25279-y

43. Hu KH, Eichorst JP, McGinnis CS, Patterson DM, Chow ED, Kersten K, Jameson SC, Gartner ZJ, Rao AA, and Krummel MF. ZipSeq: barcoding for real-time mapping of single cell transcriptomes. Nature Methods 2020 Jul :1–11. DOI: 10.1038/s41592-020-0880-2. Available from: https://doi.org/10.1038/s41592-020-0880-2

44. Merritt CR, Ong GT, Church SE, Barker K, Danaher P, Geiss G, Hoang M, Jung J, Liang Y, McKay-Fleisch J, Nguyen K, Norgaard Z, Sorg K, Sprague I, Warren C, Warren S, Webster PJ, Zhou Z, Zollinger DR, Dunaway DL, Mills GB, and Beechem JM. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. Nature Biotechnology 2020 38:5 2020 May; 38:586–99. DOI: 10.1038/s41587-020-0472-9. Available from: https://www.nature.com/articles/s41587-020-0472-9

45. Roberts K, Aivazidis A, Kleshchevnikov V, Li T, Fropf R, Rhodes M, Beechem JM, Hemberg M, and Bayraktar OA. Transcriptome-wide spatial RNA profiling maps the cellular architecture of the developing human neocortex. bioRxiv 2021 Jan :2021.03.20.436265. DOI: 10.1101/2021.03.

20.436265. Available from: http://biorxiv.org/content/early/2021/03/20/2021.03.20.436265.abstract

46. Lubeck E and Cai L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. Nature Methods 2012 9:7 2012 Jun; 9:743–8. DOI: 10.1038/nmeth.2069. Available from: https://www.nature.com/articles/nmeth.2069

47. Shah S, Lubeck E, Zhou W, and Cai L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. Neuron 2016. DOI: 10.1016/j.neuron.2016.10.001

48. Eng CHL, Shah S, Thomassie J, and Cai L. Profiling the transcriptome with RNA SPOTs. Nature Methods 2017; 14:1153–5. DOI: 10.1038/nmeth.4500. Available from: https://doi.org/10.1038/nmeth.4500

49. Eng CHL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, Yun J, Cronin C, Karp C, Yuan GC, and Cai L. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature 2019 Apr; 568:235–9. DOI: 10.1038/s41586-019-1049-y. Available from: https://www.nature.com/articles/s41586-019-1049-y

50. Chen KH, Boettiger AN, Moffitt JR, Wang S, and Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 2015. DOI: 10.1126/science.aaa6090

51. Xia C, Fan J, Emanuel G, Hao J, and Zhuang X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proceedings of the National Academy of Sciences of the United States of America 2019. DOI: 10.1073/pnas.1912459116

52. Gyllborg D, Langseth CM, Qian X, Salas SM, Hilscher M, Lein E, and Nilsson M. Hybridization-based In Situ Sequencing (HybISS): spatial transcriptomic detection in human and mouse brain tissue. bioRxiv 2020 Feb :2020.02.03.931618. DOI: 10.1101/2020.02.03.931618. Available from: https://doi.org/10.1101/2020.02.03.931618

53. Goh JJL, Chou N, Seow WY, Ha N, Cheng CPP, Chang YC, Zhao ZW, and Chen KH. Highly specific multiplexed RNA imaging in tissues with split-FISH. Nature Methods 2020 Jul; 17:689–93. DOI: 10.1038/s41592-020-0858-0. Available from: https://doi.org/10.1038/s41592-020-0858-0

54. Battich N, Stoeger T, and Pelkmans L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. Nature Methods 2013 Nov; 10:1127–36. DOI: 10.1038/nmeth.2657. Available from: https://www.nature.com/articles/nmeth.2657

55. Kishi JY, Lapan SW, Beliveau BJ, West ER, Zhu A, Sasaki HM, Saka SK, Wang Y, Cepko CL, and Yin P. SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. Nature Methods 2019; 16:533–44. DOI: 10.1038/s41592-019-0404-0. Available from: https://doi.org/10.1038/s41592-019-0404-0

56. Wang G, Moffitt JR, and Zhuang X. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. Scientific Reports 2018. DOI: 10.1038/s41598-018-22297-7

57. Chen F, Tillberg PW, and Boyden ES. Expansion microscopy. Science 2015 Jan; 347:543–8. DOI: 10.1126/science.1260088. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1260088

58. Coskun AF and Cai L. Dense transcript profiling in single cells by image correlation decoding. Nature Methods 2016 Jul; 13:657–60. DOI: 10.1038/nmeth.3895. Available from: https://www.nature.com/articles/nmeth.3895

59. Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wählby C, and Nilsson M. In situ sequencing for RNA analysis in preserved tissue and cells. Nature Methods 2013 Sep; 10:857–60. DOI: 10.1038/nmeth.2563. Available from: https://www.nature.com/articles/nmeth.2563

60. Liu S, Punthambaker S, Iyer EPR, Ferrante T, Goodwin D, Fürth D, Pawlowski AC, Jindal K, Tam JM, Mifflin L, Alon S, Sinha A, Wassie AT, Chen F, Cheng A, Willocq V, Meyer K, Ling KH, Camplisson CK, Kohman RE, Aach J, Lee JH, Yankner BA, Boyden ES, and Church GM. Barcoded oligonucleotides ligated on RNA amplified for multiplexed and parallel in situ analyses. Nucleic Acids Research 2021 Jun; 49:e58–e58. DOI: 10.1093/nar/gkab120. Available from: https://doi.org/10.1093/nar/gkab120

61. Shendure J. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. Science 2005 Sep; 309:1728–32. DOI: 10.1126/science.1117389. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1117389

62. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, and Church GM. Highly Multiplexed Subcellular RNA Sequencing in Situ. Science 2014 Mar; 343:1360 LP –1363. DOI: 10.1126/science.1250212. Available from: http://science.sciencemag.org/content/343/6177/1360.abstract

63. Alon S, Goodwin DR, Sinha A, Wassie AT, Chen F, Daugharthy ER, Bando Y, Kajita A, Xue AG, Marrett K, Prior R, Cui Y, Payne AC, Yao CC, Suk HJ, Wang R, Yu CC, Tillberg P, Reginato P, Pak N, Liu S, Punthambaker S, Iyer EP, Kohman RE, Miller JA, Lein ES, Lako A, Cullen N,

Rodig S, Helvie K, Abravanel DL, Wagle N, Johnson BE, Klughammer J, Slyper M, Waldman J, Jané-Valbuena J, Rozenblatt-Rosen O, Regev A, Church GM, Marblestone AH, and Boyden ES. Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. Science 2021 Jan; 371. DOI: 10.1126/SCIENCE.AAX2656/SUPPL{_}FILE/AAX2656{_}TABLESS1-S6ANDS9-S14.XLSX. Available from: https://www.science.org/doi/10.1126/science.aax2656

64. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, Nolan GP, Bava FA, and Deisseroth K. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science 2018 Jul; 361:eaat5691. DOI: 10.1126/science.aat5691. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aat5691

65. Sun D, Xiao Y, Liu Z, Li T, Wu Q, and Wang C. STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing. bioRxiv 2021 Jan :2021.09.08.459458. DOI: 10.1101/2021.09.08.459458. Available from: http://biorxiv.org/content/early/2021/09/09/2021.09.08.459458.abstract

66. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, and McCarroll SA. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015 May; 161:1202–14. DOI: 10.1016/j.cell.2015.05.002. Available from: http://dx.doi.org/10.1016/j.cell.2015.05.002

67. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, and Macosko EZ. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science 2019. DOI: 10.1126/science.aaw1219

68. Vickovic S, Eraslan G, Salmén F, Klughammer J, Stenbeck L, Schapiro D, Äijö T, Bonneau R, Bergenstråhle L, Navarro JF, Gould J, Griffin GK, Borg Å, Ronaghi M, Frisén J, Lundeberg J, Regev A, and Ståhl PL. High-definition spatial transcriptomics for in situ tissue profiling. Nature Methods 2019. DOI: 10.1038/s41592-019-0548-y

69. Liu M, Lu Y, Yang B, Chen Y, Radda JS, Hu M, Katz SG, and Wang S. Multiplexed imaging of nucleome architectures in single cells of mammalian tissue. Nature Communications 2020 Dec; 11:1–14. DOI: 10.1038/s41467-020-16732-5. Available from: https://doi.org/10.1038/s41467-020-16732-5

70. Lebrigand K, Bergenstråhle J, Thrane K, Mollbrink A, Barbry P, Waldmann R, and Lundeberg J. The spatial landscape of gene expression isoforms in tissue sections. bioRxiv 2020 Aug :2020.08.24.252296. DOI: 10.1101/

2020.08.24.252296. Available from: https://doi.org/10.1101/2020.08.24.252296

71. Chen S, Loper J, Chen X, Vaughan A, Zador AM, and Paninski L. BARcode DEmixing through Non-negative Spatial Regression (BarDensr). PLOS Computational Biology 2021 Mar; 17:e1008256. Available from: https://doi.org/10.1371/journal.pcbi.1008256

72. Fu X, Sun L, Chen JY, Dong R, Lin Y, Palmiter RD, Lin S, and Gu L. Continuous Polony Gels for Tissue Mapping with High Resolution and RNA Capture Efficiency. bioRxiv 2021 Jan :2021.03.17.435795. DOI: 10.1101/2021.03.17.435795. Available from: http://biorxiv.org/content/early/2021/03/17/2021.03.17.435795.abstract

73. Lee H, Salas SM, Gyllborg D, and Nilsson M. Direct RNA targeted transcriptomic profiling in tissue using Hybridization-based RNA In Situ Sequencing (HybRISS). bioRxiv 2020 Jan :2020.12.02.408781. DOI: 10.1101/2020.12.02.408781. Available from: http://biorxiv.org/content/early/2020/12/02/2020.12.02.408781.abstract

74. Srivatsan SR, Regier MC, Barkan E, Franks JM, Packer JS, Grosjean P, Duran M, Saxton S, Ladd JJ, Spielmann M, Lois C, Lampe PD, Shendure J, Stevens KR, and Trapnell C. Embryo-scale, single-cell spatial transcriptomics. Science 2021; 373:111–7. DOI: 10.1126/science.abb9536

75. Weinstein JA, Regev A, and Zhang F. DNA Microscopy: Optics-free Spatiogenetic Imaging by a Stand-Alone Chemical Reaction. Cell 2019 Jun; 178:229–41. DOI: 10.1016/j.cell.2019.05.019. Available from: https://doi.org/10.1016/j.cell.2019.05.019

76. Hoffecker IT, Yang Y, Bernardinelli G, Orponen P, and Högberg B. A computational framework for DNA sequencing microscopy. Proceedings of the National Academy of Sciences of the United States of America 2019 Sep; 116:19282–7. DOI: 10.1073/pnas.1821178116. Available from: https://www.pnas.org/content/116/39/19282.abstract

77. Halpern KB, Shenhav R, Massalha H, Toth B, Egozi A, Massasa EE, Medgalia C, David E, Giladi A, Moor AE, Porat Z, Amit I, and Itzkovitz S. Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. Nature Biotechnology 2018 Nov; 36:962. DOI: 10.1038/nbt.4231. Available from: https://www.nature.com/articles/nbt.4231

78. Fazal FM, Han S, Parker KR, Kaewsapsak P, Xu J, Boettiger AN, Chang HY, and Ting AY. Atlas of Subcellular RNA Localization Revealed by APEX-Seq. Cell 2019 Jul; 178:473–90. DOI: 10.1016/j.cell.2019.05.027. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867419305550

79. Vickovic S, Lötstedt B, Klughammer J, Segerstolpe Å, Rozenblatt-Rosen O, and Regev A. SM-Omics: An automated platform for high-throughput spatial multi-omics. bioRxiv 2020 Jan :2020.10.14.338418. DOI: 10.1101/2020.10.14.338418. Available from: http://biorxiv.org/content/early/2020/10/15/2020.10.14.338418.abstract

80. Su J and Song Q. DSTG: Deconvoluting Spatial Transcriptomics Data through Graph-based Artificial Intelligence. bioRxiv 2020 Jan :2020.10.20.347195. DOI: 10.1101/2020.10.20.347195. Available from: http://biorxiv.org/content/early/2020/10/21/2020.10.20.347195.abstract

81. Shah S, Takei Y, Zhou W, Lubeck E, Yun J, Eng CHL, Koulena N, Cronin C, Karp C, Liaw EJ, Amin M, and Cai L. Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. Cell 2018. DOI: 10.1016/j.cell.2018.05.035

82. Zhang M, Eichhorn SW, Zingg B, Yao Z, Zeng H, Dong H, and Zhuang X. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics. bioRxiv 2020 Jan :2020.06.04.105700. DOI: 10.1101/2020.06.04.105700. Available from: http://biorxiv.org/content/early/2020/06/05/2020.06.04.105700.abstract

83. Kim M, Kim DM, and Kim DE. Label-free fluorometric detection of microRNA using isothermal rolling circle amplification generating tandem G-quadruplex. Analyst 2020; 145:6130–7. DOI: 10.1039/D0AN01329C. Available from: http://dx.doi.org/10.1039/D0AN01329C

84. Li Q, Lin Z, Liu R, Tang X, Huang J, He Y, Zhou H, Sheng H, Shi H, Wang X, and Liu J. In situ electro-sequencing in three-dimensional tissues. bioRxiv 2021 Jan :2021.04.22.440941. DOI: 10.1101/2021.04.22.440941. Available from: http://biorxiv.org/content/early/2021/04/23/2021.04.22.440941.abstract

85. Moffitt JR and Zhuang X. RNA Imaging with Multiplexed Error-Robust Fluorescence in Situ Hybridization (MERFISH). Methods in Enzymology 2016. DOI: 10.1016/bs.mie.2016.03.020

86. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, Mollbrink A, Linnarsson S, Codeluppi S, Borg Å, Pontén F, Costea PI, Sahlén P, Mulder J, Bergmann O, Lundeberg J, and Frisén J. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 2016 Jul; 353:78–82. DOI: 10.1126/science.aaf2403. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aaf2403

87. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt

PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, and Bielas JH. Massively parallel digital transcriptional profiling of single cells. Nature Communications 2017 Apr; 8:14049. DOI: 10.1038/ncomms14049. Available from: http://www.nature.com/articles/ncomms14049

88. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, and Kirschner MW. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell 2015 May; 161:1187–201. DOI: 10.1016/j.cell.2015.04.044. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867415005000

89. Hashimshony T, Senderovich N, Avital G, Klochendler A, Leeuw Y de, Anavy L, Gennert D, Li S, Livak KJ, Rozenblatt-Rosen O, Dor Y, Regev A, and Yanai I. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biology 2016 Dec; 17:77. DOI: 10.1186/s13059-016-0938-8. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0938-8

90. Grün D, Kester L, and Oudenaarden A van. Validation of noise models for single-cell transcriptomics. Nature Methods 2014 Jun; 11:637–40. DOI: 10.1038/nmeth.2930. Available from: https://www.nature.com/articles/nmeth.2930

91. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, Turczyk BM, Yang JL, Lee HS, Aach J, Zhang K, and Church GM. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. Nature Protocols 2015 Mar; 10:442–58. DOI: 10.1038/nprot.2014.191. Available from: http://www.nature.com/articles/nprot.2014.191

92. Delorey TM et al. COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. Nature 2021; 595:107–13. DOI: 10.1038/s41586-021-03570-8. Available from: https://doi.org/10.1038/s41586-021-03570-8

93. La Manno G, Siletti K, Furlan A, Gyllborg D, Vinsland E, Mossi Albiach A, Mattsson Langseth C, Khven I, Lederer AR, Dratva LM, Johnsson A, Nilsson M, Lönnerberg P, and Linnarsson S. Molecular architecture of the developing mouse brain. Nature 2021; 596:92–6. DOI: 10.1038/s41586-021-03775-x. Available from: https://doi.org/10.1038/s41586-021-03775-x

94. Zimmerman SM, Fropf R, Kulasekara BR, Griswold M, Appelbe O, Bahrami A, Boykin R, Buhr DL, Fuhrman K, Hoang ML, Huynh Q, Isgur L, Klock A, Kutchma A, Lasley AE, Liang Y, McKay-Fleisch J, Nguyen K, Piazza E, Rininger A, Zollinger DR, Rhodes M, and Beechem JM. Spatially resolved whole transcriptome profiling in human and mouse tissue using Digital Spatial Profiling. bioRxiv 2021 Jan :2021.09.29.462442. DOI: 10.1101/2021.

09.29.462442. Available from: http://biorxiv.org/content/early/2021/10/01/2021.09.29.462442.1.abstract

95. Qian X, Harris KD, Hauling T, Nicoloutsopoulos D, Muñoz-Manchado AB, Skene N, Hjerling-Leffler J, and Nilsson M. Probabilistic cell typing enables fine mapping of closely related cell types in situ. Nature Methods 2020 Jan; 17:101–6. DOI: 10.1038/s41592-019-0631-4. Available from: https://doi.org/10.1038/s41592-019-0631-4

96. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, Lagemaat LN van de, Smith KA, Ebbert A, Riley ZL, Abajian C, Beckmann CF, Bernard A, Bertagnolli D, Boe AF, Cartagena PM, Chakravarty MM, Chapin M, Chong J, Dalley RA, Daly BD, Dang C, Datta S, Dee N, Dolbeare TA, Faber V, Feng D, Fowler DR, Goldy J, Gregor BW, Haradon Z, Haynor DR, Hohmann JG, Horvath S, Howard RE, Jeromin A, Jochim JM, Kinnunen M, Lau C, Lazarz ET, Lee C, Lemon TA, Li L, Li Y, Morris JA, Overly CC, Parker PD, Parry SE, Reding M, Royall JJ, Schulkin J, Sequeira PA, Slaughterbeck CR, Smith SC, Sodt AJ, Sunkin SM, Swanson BE, Vawter MP, Williams D, Wohnoutka P, Zielke HR, Geschwind DH, Hof PR, Smith SM, Koch C, Grant SGN, and Jones AR. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature 2012 Sep; 489:391–9. DOI: 10.1038/nature11405. Available from: http://www.nature.com/articles/nature11405

97. Wang F, Flanagan J, Su N, Wang LC, Bui S, Nielson A, Wu X, Vo HT, Ma XJ, and Luo Y. RNAscope. The Journal of Molecular Diagnostics 2012 Jan; 14:22–9. DOI: 10.1016/j.jmoldx.2011.08.002. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1525157811002571

98. Villacampa EG, Larsson L, Kvastad L, Andersson A, Carlson J, and Lundeberg J. Genome-wide Spatial Expression Profiling in FFPE Tissues. bioRxiv 2020 Jul :2020.07.24.219758. DOI: 10.1101/2020.07.24.219758. Available from: https://doi.org/10.1101/2020.07.24.219758

99. Liu Y, Yang M, Deng Y, Su G, Enninful A, Guo CC, Tebaldi T, Zhang D, Kim D, Bai Z, Norris E, Pan A, Li J, Xiao Y, Halene S, and Fan R. High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. Cell 2020; 183:1665–81. DOI: https://doi.org/10.1016/j.cell.2020.10.026. Available from: https://www.sciencedirect.com/science/article/pii/S0092867420313908

100. Foley JW, Zhu C, Jolivet P, Zhu SX, Lu P, Meaney MJ, and West RB. Gene expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ. Genome Research 2019 Nov; 29:1816–25. DOI: 10.1101/gr.234807.118. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.234807.118

101. Bhaduri A, Sandoval-Espinosa C, Otero-Garcia M, Oh I, Yin R, Eze UC, Nowakowski TJ, and Kriegstein AR. An Atlas of Cortical Arealization Identifies Dynamic Molecular Signatures. bioRxiv 2021 Jan :2021.05.17.444528. DOI: 10.1101/2021.05.17.444528. Available from: http://biorxiv.org/content/early/2021/05/18/2021.05.17.444528.abstract

102. Van Valen DA, Kudo T, Lane KM, Macklin DN, Quach NT, DeFelice MM, Maayan I, Tanouchi Y, Ashley EA, and Covert MW. Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. PLOS Computational Biology 2016 Nov; 12:e1005177. Available from: https://doi.org/10.1371/journal.pcbi.1005177

103. Petukhov V, Soldatov RA, Khodosevich K, and Kharchenko PV. Bayesian segmentation of spatially resolved transcriptomics data. bioRxiv 2020 Jan :2020.10.05.326777. DOI: 10.1101/2020.10.05.326777. Available from: http://biorxiv.org/content/early/2020/10/06/2020.10.05.326777.abstract

104. Perkel JM. Starfish enterprise: finding RNA patterns in single cells. Nature 2019 Aug; 572:549–51. DOI: 10.1038/D41586-019-02477-9

105. Karaiskos N, Wahle P, Alles J, Boltengagen A, Ayoub S, Kipar C, Kocks C, Rajewsky N, and Zinzen RP. The Drosophila embryo at single-cell transcriptome resolution. Science 2017 Oct; 358:194–9. DOI: 10.1126/science.aan3235. Available from: https://science.sciencemag.org/content/358/6360/194

106. Nitzan M, Karaiskos N, Friedman N, and Rajewsky N. Gene expression cartography. Nature 2019 Dec; 576:132–7. DOI: 10.1038/s41586-019-1773-3. Available from: https://doi.org/10.1038/s41586-019-1773-3

107. Lopez R, Nazaret A, Langevin M, Samaran J, Regier J, Jordan MI, and Yosef N. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. 2019 May. Available from: http://arxiv.org/abs/1905.02269

108. Andersson A, Bergenstråhle J, Asp M, Bergenstråhle L, Jurek A, Fernández Navarro J, and Lundeberg J. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. Communications Biology 2020; 3:565. DOI: 10.1038/s42003-020-01247-y. Available from: https://doi.org/10.1038/s42003-020-01247-y

109. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, Lomakin A, Kedlian V, Jain MS, Park JS, Ramona L, Tuck E, Arutyunyan A, Vento-Tormo R, Gerstung M, James L, Stegle O, and Bayraktar OA. Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics. bioRxiv 2020 Jan :2020.11.15.378125. DOI:

10.1101/2020.11.15.378125. Available from: http://biorxiv.org/content/early/2020/11/17/2020.11.15.378125.abstract

110. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, and Irizarry RA. Robust decomposition of cell type mixtures in spatial transcriptomics. bioRxiv 2020 May :2020.05.07.082750. DOI: 10.1101/2020.05.07.082750. Available from: https://doi.org/10.1101/2020.05.07.082750

111. Yang Y, Shi X, Liu W, Zhou Q, Lau MC, Lim JCT, Sun L, Yeong J, and Liu J. SC-MEB: spatial clustering with hidden Markov random field using empirical Bayes. bioRxiv 2021 Jan :2021.06.05.447181. DOI: 10.1101/2021.06.05.447181. Available from: http://biorxiv.org/content/early/2021/09/07/2021.06.05.447181.abstract

112. Elosua M, Nieto P, Mereu E, Gut I, and Heyn H. SPOTlight: Seeded NMF regression to Deconvolute Spatial Transcriptomics Spots with Single-Cell Transcriptomes. bioRxiv 2020 Jun :2020.06.03.131334. DOI: 10.1101/2020.06.03.131334. Available from: https://www.biorxiv.org/content/10.1101/2020.06.03.131334v1

113. Miller BF, Atta L, Sahoo A, Huang F, and Fan J. Reference-free cell-type deconvolution of pixel-resolution spatially resolved transcriptomics data. bioRxiv 2021 Jan :2021.06.15.448381. DOI: 10.1101/2021.06.15.448381. Available from: http://biorxiv.org/content/early/2021/06/16/2021.06.15.448381.abstract

114. Wolf FA, Angerer P, and Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biology 2018; 19:15. DOI: 10.1186/s13059-017-1382-0. Available from: https://doi.org/10.1186/s13059-017-1382-0

115. Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, Rybakov S, Ibarra IL, Holmberg O, Virshup I, Lotfollahi M, Richter S, and Theis FJ. Squidpy: a scalable framework for spatial omics analysis. en. Nat. Methods 2022 Feb; 19:171–8

116. Righelli D, Weber LM, Crowell HL, Pardo B, Collado-Torres L, Ghazanfar S, Lun AT, Hicks SC, and Risso D. SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor. Bioinformatics 2022 May; 38:3128–31. DOI: 10.1093/BIOINFORMATICS/BTAC299. Available from: https://academic.oup.com/bioinformatics/article/38/11/3128/6575443

117. Dries R, Zhu Q, Dong R, Eng Chee-Huat Linus, Li H, Liu K, Fu Y, Zhao Tianxiao, Sarkar A, Bao F, George RE, Pierson N, Cai L, and Yuan GC. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. en. Genome Biol. 2021 Mar; 22:78

118. Bergenstråhle J, Larsson L, and Lundeberg J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. BMC Genomics 2020 Dec; 21:482. DOI: 10.1186/s12864-020-06832-3. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-06832-3

119. Zhu Q, Shah S, Dries R, Cai L, and Yuan GC. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nature Biotechnology 2018 Dec; 36:1183–90. DOI: 10.1038/nbt.4260. Available from: https://www.nature.com/articles/nbt.4260

120. Svensson V, Veiga Beltrame E da, and Pachter L. A curated database reveals trends in single-cell transcriptomics. Database 2020 Jan; 2020. DOI: 10.1093/database/baaa073. Available from: https://doi.org/10.1093/database/baaa073

121. Sun S, Zhu J, and Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nature Methods 2020 Feb; 17:193–200. DOI: 10.1038/s41592-019-0701-7. Available from: https://doi.org/10.1038/s41592-019-0701-7

122. BinTayyash N, Georgaka S, John ST, Ahmed S, Boukouvalas A, Hensman J, and Rattray M. Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments. bioRxiv 2020 Jan :2020.07.29.227207. DOI: 10.1101/2020.07.29.227207. Available from: http://biorxiv.org/content/early/2020/07/30/2020.07.29.227207.abstract

123. Govek KW, Yamajala VS, and Camara PG. Clustering-independent analysis of genomic data using spectral simplicial theory. PLOS Computational Biology 2019 Nov; 15:e1007509. Available from: https://doi.org/10.1371/journal.pcbi.1007509

124. Edsgärd D, Johnsson P, and Sandberg R. Identification of spatial expression trends in single-cell gene expression data. Nature Methods 2018 May; 15:339–42. DOI: 10.1038/nmeth.4634. Available from: http://www.nature.com/articles/nmeth.4634

125. Miller BF, Bambah-Mukku D, Dulac C, Zhuang X, and Fan J. Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. Genome Research 2021 May. DOI: 10.1101/gr.271288.120. Available from: http://genome.cshlp.org/content/early/2021/09/20/gr.271288.120.abstract

126. Pham VVH, Li X, Truong B, Nguyen T, Liu L, Li J, and Le TD. The winning methods for predicting cellular position in the DREAM single-cell transcriptomics challenge. Briefings in Bioinformatics 2020 Aug. DOI: 10.1093/bib/bbaa181. Available from: https://doi.org/10.1093/bib/bbaa181

127. Canete NP, Iyengar SS, Ormerod JT, Baharlou H, Harman AN, and Patrick E. spicyR: spatial analysis of in situ cytometry data in R. en. Bioinformatics 2022 May; 38:3099–105

128. Arnol D, Schapiro D, Bodenmiller B, Saez-Rodriguez J, and Stegle O. Modeling Cell-Cell Interactions from Spatial Molecular Data with Spatial Variance Component Analysis. Cell Reports 2019 Oct; 29:202–11. DOI: 10.1016/j.celrep.2019.08.077. Available from: https://doi.org/10.1016/j.celrep.2019.08.077

129. Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, Catallini JL, Tran MN, Besich Z, Tippani M, Chew J, Yin Y, Kleinman JE, Hyde TM, Rao N, Hicks SC, Martinowich K, and Jaffe AE. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nature Neuroscience 2021 24:3 2021 Feb; 24:425–36. DOI: 10.1038/s41593-020-00787-0. Available from: https://www.nature.com/articles/s41593-020-00787-0

130. Lundmark A, Gerasimcik N, Båge T, Jemt A, Mollbrink A, Salmén F, Lundeberg J, and Yucel-Lindberg T. Gene expression profiling of periodontitis-affected gingival tissue by spatial transcriptomics. Scientific Reports 2018; 8:9370. DOI: 10.1038/s41598-018-27627-3. Available from: https://doi.org/10.1038/s41598-018-27627-3

131. Fan Z, Chen R, and Chen X. SpatialDB: a database for spatially resolved transcriptomes. Nucleic Acids Research 2020 Jan; 48:D233–D237. DOI: 10.1093/nar/gkz934. Available from: https://doi.org/10.1093/nar/gkz934

132. Armit C, Richardson L, Venkataraman S, Graham L, Burton N, Hill B, Yang Y, and Baldock RA. eMouseAtlas: An atlas-based resource for understanding mammalian embryogenesis. Developmental Biology 2017 Mar; 423:1–11. DOI: 10.1016/j.ydbio.2017.01.023. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0012160616308582

133. Singer RH and Ward DC. Actin gene expression visualized in chicken muscle tissue culture by using in situ hybridization with a biotinated nucleotide analog. Proceedings of the National Academy of Sciences of the United States of America 1982 Dec; 79:7331–5. DOI: 10.1073/pnas.79.23.7331. Available from: http://www.pnas.org/content/79/23/7331.abstract

134. Hope IA. 'Promoter trapping' in Caenorhabditis elegans. Tech. rep. 1991 :399–408. Available from: https://dev.biologists.org/content/develop/113/2/399.full.pdf

135. Seydoux G and Fire A. Soma-germline asymmetry in the distributions of embryonic RNAs in Caenorhabditis elegans. Development 1994 Oct; 120:2823

LP –2834. Available from: http://dev.biologists.org/content/120/10/2823.abstract

136. Bettenhausen B and Gossler A. Efficient Isolation of Novel Mouse Genes Differentially Expressed in Early Postimplantation Embryos. Genomics 1995 Aug; 28:436–41. DOI: 10.1006/geno.1995.1172. Available from: https://linkinghub.elsevier.com/retrieve/pii/S088875438571172X

137. Gawantka V, Pollet N, Delius H, Vingron M, Pfister R, Nitsch R, Blumenstock C, and Niehrs C. Gene expression screening in Xenopus identifies molecular pathways, predicts gene function and provides a global view of embryonic patterning. Mechanisms of Development 1998 Sep; 77:95–141. DOI: 10.1016/S0925-4773(98)00115-4. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0925477398001154

138. Ringwald M, Mangan ME, Eppig JT, Kadin JA, and Richardson JE. GXD: a Gene Expression Database for the laboratory mouse. Nucleic Acids Research 1999 Jan; 27:106–12. DOI: 10.1093/nar/27.1.106. Available from: https://doi.org/10.1093/nar/27.1.106

139. Kawashima T. MAGEST: MAboya Gene Expression patterns and Sequence Tags. Nucleic Acids Research 2000 Jan; 28:133–5. DOI: 10.1093/nar/28.1.133. Available from: http://www.genome.ad.jp/magest/%20https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.133

140. Maeda I, Kohara Y, Yamamoto M, and Sugimoto A. Large-scale analysis of gene function in Caenorhabditis elegans by high-throughput RNAi. Current Biology 2001; 11:171–6. DOI: https://doi.org/10.1016/S0960-9822(01)00052-5. Available from: https://www.sciencedirect.com/science/article/pii/S0960982201000525

141. Satou Y, Takatori N, Yamada L, Mochizuki Y, Hamaguchi M, Ishikawa H, Chiba S, Imai K, Kano S, Murakami SD, Nakayama A, Nishino A, Sasakura Y, Satoh G, Shimotori T, Shin-i T, Shoguchi E, Suzuki MM, Takada N, Utsumi N, Yoshida N, Saiga H, Kohara Y, and Satoh N. Gene expression profiles in Ciona intestinalis tailbud embryos. Development 2001 Aug; 128:2893 LP –2904. Available from: http://dev.biologists.org/content/128/15/2893.abstract

142. Carson JP, Thaller C, and Eichele G. A transcriptome atlas of the mouse brain at cellular resolution. Current Opinion in Neurobiology 2002 Oct; 12:562–5. DOI: 10.1016/S0959-4388(02)00356-2. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0959438802003562

143. Henrich T. MEPD: a Medaka gene expression pattern database. Nucleic Acids Research 2003 Jan; 31:72–4. DOI: 10.1093/nar/gkg017. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg017

144. Luengo Hendriks CL, Keränen SV, Fowlkes CC, Simirenko L, Weber GH, DePace AH, Henriquez C, Kaszuba DW, Hamann B, Eisen MB, Malik J, Sudar D, Biggin MD, and Knowles DW. Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution I: Data acquisition pipeline. Genome Biology 2006 Dec; 7:R123. DOI: 10.1186/gb-2006-7-12-r123. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2006-7-12-r123

145. Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, Hughes TR, Tomancak P, and Krause HM. Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function. Cell 2007 Oct; 131:174–87. DOI: 10.1016/j.cell.2007.08.003. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867407010227

146. Bowes JB, Snyder KA, Segerdell E, Gibb R, Jarabek C, Noumen E, Pollet N, and Vize PD. Xenbase: a Xenopus biology and genomics resource. Nucleic Acids Research 2008 Jan; 36:D761–D767. DOI: 10.1093/nar/gkm826. Available from: https://doi.org/10.1093/nar/gkm826

147. Lovell PV, Wirthlin M, Kaser T, Buckner AA, Carleton JB, Snider BR, McHugh AK, Tolpygo A, Mitra PP, and Mello CV. ZEBrA: Zebra finch Expression Brain Atlas—A resource for comparative molecular neuroanatomy and brain evolution studies. Journal of Comparative Neurology 2020 Aug; 528:2099–131. DOI: 10.1002/cne.24879. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.24879

148. Landegren U, Kaiser R, Sanders J, and Hood L. A ligase-mediated gene detection technique. Science 1988 Aug; 241:1077 LP –1080. DOI: 10.1126/science.3413476. Available from: http://science.sciencemag.org/content/241/4869/1077.abstract

149. Belyavsky A, Vinogradova T, and Rajewsky K. PCR-based cDNA library construction: general cDNA libraries at the level of a few cells. eng. Nucleic acids research 1989 Apr; 17:2919–32. DOI: 10.1093/nar/17.8.2919. Available from: https://pubmed.ncbi.nlm.nih.gov/2471144%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC317702/

150. Van Gelder RN, Zastrow ME von, Yool A, Dement WC, Barchas JD, and Eberwine JH. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proceedings of the National Academy of Sciences 1990 Mar; 87:1663 LP –1667. Available from: http://www.pnas.org/content/87/5/1663.abstract

151. Schena M, Shalon D, Davis RW, and Brown PO. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science 1995 Oct; 270:467 LP –470. DOI: 10.1126/science.270.5235.467. Available from: http://science.sciencemag.org/content/270/5235/467.abstract

152. Luo L, Salunga RC, Guo H, Bittner A, Joy K, Galindo JE, Xiao H, Rogers KE, Wan JS, Jackson MR, and Erlander MG. Gene expression profiles of laser-captured adjacent neuronal subtypes. Nature Medicine 1999 Jan; 5:117–22. DOI: 10.1038/4806. Available from: http://www.nature.com/articles/nm0199_117

153. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, and Ecker JR. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. Cell 2008; 133:523–36. DOI: https://doi.org/10.1016/j.cell.2008.03.029. Available from: https://www.sciencedirect.com/science/article/pii/S0092867408004480

154. Okamura-Oho Y, Shimokawa K, Takemoto S, Hirakiyama A, Nakamura S, Tsujimura Y, Nishimura M, Kasukawa T, Masumoto Kh, Nikaido I, Shigeyoshi Y, Ueda HR, Song G, Gee J, Himeno R, and Yokota H. Transcriptome Tomography for Brain Analysis in the Web-Accessible Anatomical Space. PLoS ONE 2012 Sep; 7. Ed. by Hayasaka S:e45373. DOI: 10.1371/journal.pone.0045373. Available from: https://dx.plos.org/10.1371/journal.pone.0045373

Chapter 3

INTRODUCTION OF THE MUSEUM OF SPATIAL TRANSCRIPTOMICS BOOK

The spatial organization of the components of biological systems is crucial for their proper function. For instance, morphogen gradients in embryos are tightly regulated to ensure that the right cell types differentiate at the right place. In adults, spatial organization of cells in tissues is important to proper functions of organs. For instance, the liver lobule is divided in labor according to distance from the portal triad as such distance affects suitability of different tasks. Both oxygen level and morphogen gradient regulate zonation of metabolism [1]; there is more oxidative phosphorylation and gluconeogenesis in the more oxygenated periportal region and more glycolysis in the more deoxygenated pericentral region. How cell types and cellular functions vary in space can be measured by quantifying gene expression in space. Conversely, the expression of an unknown gene in space can give clues to its function. Gene expression is usually quantified by quantifying proteins or transcripts encoded by the gene, and high throughput spatial methods exist for both protein and transcripts. In other words, cellular function exemplifies the maxim that "the whole is greater than the sum of its parts", and in large part this follows from "location, location, location".

Here we focus on spatial transcriptomics (the field of spatial proteomics is covered elsewhere [2, 3, 4]. Even spatial transcriptomics is a vast field, and it is useful to begin by considering the scope of what it contains. Naïvely, one may say, spatial transcriptomics means quantifying the complete set of RNAs encoded by the genome in space. Usually the “in space” is at some microscopic resolution rather than geospatial as often assumed in the term “spatial statistics”; the resolution is usually cellular, though sometimes subcellular. The “spatial” is in contrast to other transcriptomics methods that by virtue of the nature of their assays, lose information of tissue structure in space. That is the case with microarray technology for bulk tissue analysis, for bulk RNA-seq, and single-cell RNA-seq (scRNA-seq) that is based on dissociation of tissue—the “spatial” usually means tissue structure in space. More broadly, the “spatial” can mean knowing spatial context of samples although the spatial context is only a label and the coordinates are not collected or not used, such as in some laser capture microdissection (LCM) literature [5, 6,

7], Niche-seq [8]), and APEX-seq [9]. The "spatial" can also mean preserving spatial coordinates of samples within tissue, though the coordinates may or may not be explicitly used in data analysis, such as in the various single molecular fluorescent in situ hybridization (smFISH) based technologies such as seqFISH [10] and MERFISH [11] and array based technologies such as Spatial Transciptomics (ST) [12].

There is more complexity in defining "transcriptomics". While some technologies usually called "spatial transcriptomics" are indeed transcriptome-wide, such as ST, Visium, and LCM followed by RNA-seq, many technologies that only profile a panel of usually a few hundred genes are nevertheless considered part of "spatial transcriptomics". Here "transcriptomics" actually means high-throughput quantification of gene expression, preferably highly multiplexed, quantifying numerous genes within the same piece of tissue at the same time. However, what counts as "high-throughput"? Is there a minimum number of genes required? Should 50 genes be enough? Or a hundred genes? The threshold number of genes required to be considered "high-throughput" is difficult to define; here, by "high-throughput", we mean the intent to quantify expression of more genes than normally done with fluorescent in situ hybridization (FISH) or immunofluorescence when only color distinguishes between genes, which can mean more than about 5 genes. There is also some complication regarding whether "highly multiplexed" should be required. Some fairly recent studies that intended to perform high-throughput gene expression profiling in space did not profile most genes at the same time (e.g. multiple rounds of smFISH hybridization, each round for a different set of genes) [13, 14], or even profiled different genes in different tissue sections [15, 16]; these papers nevertheless claimed to be spatial transcriptomic or something similar.

When terms are to be defined by how they are used, then we rely on a generic and inclusive definition of "spatial transcriptomics", which can be summarized as: Quantifying transcripts while keeping spatial context of samples within tissue or cell, with intent to quantify transcripts of more genes than normally done with one round of FISH or immunofluorescence when color is the only way to distinguish between genes. This is the criterion we used in considering what methods to include in our review.

3.1 Database

The field of spatial transcriptomics has grown drastically in the past 5 years, during which several reviews have already been written. These survey existing technologies [17, 18, 19, 20, 21] or discuss how the technologies apply to specific biological systems such as tumors [22, 23, 24]. Unlike the review papers, we aim to be more systematic and detailed in our review of spatial transcriptomics technology. In addition, we review existing data analysis methods in this field, a crucial aspect of spatial transcriptomics which has not yet been comprehensively reviewed in depth. Moreover, we present a curated database of spatial transcriptomics literature and analyses of the literature metadata to show trends in different aspects of spatial transcriptomics. This database is publicly available here. Similar databases have been curated for scRNA-seq literature [25], and for scRNA-seq data analysis tools [26], which have been analyzed to show trends in the field, although the metadata in our database and the analyses are much more extensive.

Curation of the database was performed by searching terms "spatial transcriptomics", "visium", "merfish", "seqfish", and "geomx dsp" on PubMed and in addition, the term "ISS" on bioRxiv as searching "ISS" on PubMed does not yield many relevant results. Then the search results are manually screened and publications that fit the definition of "spatial transcriptomics" as stated above are added to the database. In addition, publications citing well-known publications that are commonly recognized as "spatial transcriptomics" (e.g. the original paper for MERFISH) are screened. Such searches can find publications for spatial transcriptomics data analysis as well. Additional criteria of inclusion for data analysis publications are discussed in Chapter 9. If a method fitting the definition of "spatial transcriptomics" is mentioned anywhere outside the search results, such as a review paper, the publication of that method is also added to the database. For historical methods (i.e. prequel) loosely fitting our definition of "spatial transcriptomics" and sharing objectives with more recent spatial transcriptomics but are not highly multiplexed and don't involve cDNA microarrays or next generation sequencing (NGS), search terms such as "gene trap screen" and "in situ hybridization atlas" were used. Review papers and protocols are excluded.

Metadata of the publications collected include date published (or posted on bioRxiv for preprints), title, journal, PMID if applicable, DOI URL, species and tissue the data comes from (or the data analysis method is designed for), whether the tissue is pathological (mouse and human only), and city and institution of the first au-

thor. Such metadata allow for analyses of trends in spatial transcriptomics through time and how and where spatial transcriptomics technologies are used. In addition, for historical databases such as for *in situ* hybridization atlases, a metadata column indicates whether the database is still available. Metadata for data and code availability are also recorded. For cDNA microarray and NGS data, accessions in Gene Expression Omnibus (GEO), Short Read Archive (SRA), database of Genotypes and Phenotypes (dbGaP), European Nucleotide Archive (ENA), DNA Bank of Japan, The National Omics Data Encyclopedia (China), and BIG Sub (China) are recorded when available. For both downstream analysis and package development, the programming languages used and code repository are recorded when available. Other metadata specific to certain types of publications are collected as well, such as whether the method was used to target specific histologically defined regions of interest (ROI) or to analyze the tissue in a regular grid for microdissection based methods, and whether the implementation of a data analysis method is packaged and reasonably well-documented for data analysis publications.

There are some caveats to our review and database. First, while we narrate a history of evolution of techniques and in some cases explain how one technique influenced another, we do not present aspects of the history that are not apparent from the publications. Studying those aspects of the history of the field may require interviewing the people who developed the techniques, as well as exploration of additional unpublished material. Second, our database was originally only meant for papers, so relevant materials that are not in presented in that format are underrepresented. Examples of such materials include databases and software not presented as papers (e.g. the XDB3 database [27]. This means that the metadata analyses in this book might not be representative of all material that exists in spatial transcriptomics. Third, as the curation was done manually and the search engines are imperfect, the database might not include some relevant literature unknown to us. Please contact us or open an issue in the GitHub repo of this book if you wish to suggest new entries to the database.

The database is continuously manually updated daily by screening RSS feeds from the search terms in PubMed and bioRxiv mentioned above. New entries and the associated metadata can also be submitted via the Google Form.

3.2 Organization of the database and this book

The database is organized as several different sheets for different types of publications. Many technologies can be classified in several different ways and some ways are more useful in some contexts than others, and spatial transcriptomics is no exception. Furthermore, the line between different categories can at times be difficult to draw and there are gray areas.

Our database starts with articles published in the 1980s to provide historical context of what is now commonly known as spatial transcriptomics; this literature is summarized in Chapter 4, and historical methods of data analysis are reviewed in Chapter 5.

The literature is broken down into the following categories, corresponding to sheets in the database, to be defined and elaborated on in the subsequent chapters. Technologies to collect data (Chapter 6) can be broadly classified by mechanisms spatial contexts of samples are obtained: ROI selection (Section 7.1), next generation sequencing with spatial barcodes (abbreviated as NGS barcoding, Section 7.4), single molecular FISH (smFISH) (Section 7.2), *in situ* sequencing (ISS) (Section 7.3), and no *priori* (Section 7.6). Within some of the categories, especially microdissection and NGS barcoding, are large varieties of mechanisms and gray areas. Methods in the gray areas and don't fit nicely into any category are placed in the "Other" sheet.

These technologies can be classified in other ways, such as whether transcripts can be traced back to individual cells, and whether the spatial context takes the form of manually selected ROIs or a regular grid or both or neither. These other categories can cut across different mechanisms to acquire spatial contexts. In addition, studies using these technologies can be classified: demonstration of new data collection techniques, reference atlases intended to more comprehensively characterize the system of interest, characterization of tissues without intending to build reference atlases, and demonstration of data analysis methods. As the purpose of this database and book is to systematically document data collection and analysis methods in spatial transcriptomics, the mechanisms to acquire spatial contexts are used to structure the database and text; the other ways of categorization are mentioned in the text to give some perspectives for potential users of data collection techniques or users of existing datasets.

Data analysis methods (Chapter 9) are placed under the following categories: Preprocessing (Section 9.1), exploratory data analysis (EDA) (Section 9.2), spatial reconstruction of single-cell RNA-seq (scRNA-seq) data (Section 9.3), spatially

variable genes (Section 9.5), archetypal gene expression patterns (Section 9.6), using transcriptome to identify spatially coherent regions in tissue (Section 9.7), cell type deconvolution of non-single-cell resolution spatial data (Section 9.4), cell-cell interaction (Section 9.8), and other types of analyses. These data analysis methods can also be placed on a upstream to downstream spectrum. Upstream methods prepare the data to be more amenable to downstream analyses, and downstream methods aim to give biological relevant information and hypotheses. Then preprocessing, including cell segmentation in highly multiplexed smFISH images and obtaining a gene count matrix from fastq files, would be upstream. Quality control of the gene count matrix and EDA would be downstream from that, followed by cell type deconvolution, mapping cells to locations, and then spatially variable genes and cell-cell interactions. The types of data analysis methods are introduced roughly in the order from upstream to downstream.

In each of the following chapters, besides introducing the relevant technologies, the literature metadata is analyzed to show relevant sociological trends such as who is using each technology, usage trends of technologies, and the programming languages used. The metadata analyses can be run interactively in RStudio Cloud.

References

1. Gebhardt R. Liver zonation: Novel aspects of its regulation and its impact on homeostasis. World Journal of Gastroenterology 2014 Jul; 20:8491. DOI: 10.3748/wjg.v20.i26.8491. Available from: http://www.wjgnet.com/1007-9327/full/v20/i26/8491.htm

2. Lundberg E and Borner GHH. Spatial proteomics: a powerful discovery tool for cell biology. Nature Reviews Molecular Cell Biology 2019 May; 20:285–302. DOI: 10.1038/s41580-018-0094-y. Available from: http://www.nature.com/articles/s41580-018-0094-y

3. Baharlou H, Canete NP, Cunningham AL, Harman AN, and Patrick E. Mass Cytometry Imaging for the Study of Human Diseases—Applications and Data Analysis Strategies. Frontiers in Immunology 2019 Nov; 10:2657. DOI: 10.3389/fimmu.2019.02657. Available from: https://www.frontiersin.org/article/10.3389/fimmu.2019.02657/full

4. Buchberger AR, DeLaney K, Johnson J, and Li L. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. Analytical Chemistry 2018 Jan; 90:240–65. DOI: 10.1021/acs.analchem.7b04733. Available from: https://pubs.acs.org/doi/10.1021/acs.analchem.7b04733

5. Aguila J, Cheng S, Kee N, Cao M, Wang M, Deng Q, and Hedlund E. Spatial RNA sequencing identifies robust markers of vulnerable and resistant human midbrain dopamine neurons and their expression in Parkinson's Disease. bioRxiv 2021 Jan :334417. DOI: 10.1101/334417. Available from: http://biorxiv.org/content/early/2021/04/09/334417.abstract

6. Baccin C, Al-Sabah J, Velten L, Helbling PM, Grünschläger F, Hernández-Malmierca P, Nombela-Arrieta C, Steinmetz LM, Trumpp A, and Haas S. Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. Nature Cell Biology 2020 Jan; 22:38–48. DOI: 10.1038/s41556-019-0439-6. Available from: https://doi.org/10.1038/s41556-019-0439-6

7. Nichterwitz S, Chen G, Aguila Benitez J, Yilmaz M, Storvall H, Cao M, Sandberg R, Deng Q, and Hedlund E. Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. Nature Communications 2016 Nov; 7:12139. DOI: 10.1038/ncomms12139. Available from: http://www.nature.com/articles/ncomms12139

8. Medaglia C, Giladi A, Stoler-Barak L, De Giovanni M, Salame TM, Biram A, David E, Li H, Iannacone M, Shulman Z, and Amit I. Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. Science 2017 Dec; 358:1622–6. DOI: 10.1126/science.aao4277. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aao4277

9. Fazal FM, Han S, Parker KR, Kaewsapsak P, Xu J, Boettiger AN, Chang HY, and Ting AY. Atlas of Subcellular RNA Localization Revealed by APEX-Seq. Cell 2019 Jul; 178:473–90. DOI: 10.1016/j.cell.2019.05.027. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867419305550

10. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, and Cai L. Single-cell in situ RNA profiling by sequential hybridization. 2014 Mar. DOI: 10.1038/nmeth.2892. Available from: https://www.nature.com/articles/nmeth.2892

11. Chen KH, Boettiger AN, Moffitt JR, Wang S, and Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 2015. DOI: 10.1126/science.aaa6090

12. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, Mollbrink A, Linnarsson S, Codeluppi S, Borg Å, Pontén F, Costea PI, Sahlén P, Mulder J, Bergmann O, Lundeberg J, and Frisén J. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 2016 Jul; 353:78–82. DOI: 10.1126/science.aaf2403. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aaf2403

13. Lignell A, Kerosuo L, Streichan SJ, Cai L, and Bronner ME. Identification of a neural crest stem cell niche by Spatial Genomic Analysis. Nature Communications 2017 Dec; 8:1830. DOI: 10.1038/s41467-017-01561-w. Available from: http://www.nature.com/articles/s41467-017-01561-w

14. Wang Y, Eddison M, Fleishman G, Weigert M, Xu S, Henry FE, Wang T, Lemire AL, Schmidt U, Yang H, Rokicki K, Goina C, Svoboda K, Myers EW, Saalfeld S, Korff W, Sternson SM, and Tillberg PW. Expansion-Assisted Iterative-FISH defines lateral hypothalamus spatio-molecular organization. bioRxiv 2021 Jan :2021.03.08.434304. DOI: 10.1101/2021.03.08.434304. Available from: http://biorxiv.org/content/early/2021/03/08/2021.03.08.434304.abstract

15. Bayraktar OA, Bartels T, Holmqvist S, Kleshchevnikov V, Martirosyan A, Polioudakis D, Ben Haim L, Young AM, Batiuk MY, Prakash K, Brown A, Roberts K, Paredes MF, Kawaguchi R, Stockley JH, Sabeur K, Chang SM, Huang E, Hutchinson P, Ullian EM, Hemberg M, Coppola G, Holt MG, Geschwind DH, and Rowitch DH. Astrocyte layers in the mammalian cerebral cortex revealed by a single-cell in situ transcriptomic map. Nature Neuroscience 2020 Apr; 23:500–9. DOI: 10.1038/s41593-020-0602-1. Available from: https://doi.org/10.1038/s41593-020-0602-1

16. Battich N, Stoeger T, and Pelkmans L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. Nature Methods 2013 Nov; 10:1127–36. DOI: 10.1038/nmeth.2657. Available from: https://www.nature.com/articles/nmeth.2657

17. Crosetto N, Bienko M, and Van Oudenaarden A. Spatially resolved transcriptomics and beyond. 2015. DOI: 10.1038/nrg3832

18. Moor AE and Itzkovitz S. Spatial transcriptomics: paving the way for tissue-level systems biology. Current Opinion in Biotechnology 2017 Aug; 46:126–33. DOI: 10.1016/j.copbio.2017.02.004. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0958166916302397

19. Strell C, Hilscher MM, Laxman N, Svedlund J, Wu C, Yokota C, and Nilsson M. Placing RNA in context and space – methods for spatially resolved transcriptomics. 2019. DOI: 10.1111/febs.14435

20. Liao J, Lu X, Shao X, Zhu L, and Fan X. Uncovering an Organ's Molecular Architecture at Single-Cell Resolution by Spatially Resolved Transcriptomics. Trends in Biotechnology 2020 Jun. DOI: 10.1016/j.tibtech.2020.05.006. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167779920301402

21. Waylen LN, Nim HT, Martelotto LG, and Ramialison M. From whole-mount to single-cell spatial assessment of gene expression in 3D. Communications

Biology 2020; 3:602. DOI: 10.1038/s42003-020-01341-1. Available from: https://doi.org/10.1038/s42003-020-01341-1

22. Smith CM, Hayamizu TF, Finger JH, Bello SM, McCright IJ, Xu J, Baldarelli RM, Beal JS, Campbell J, Corbani LE, Frost PJ, Lewis JR, Giannatto SC, Miers D, Shaw DR, Kadin JA, Richardson JE, Smith CL, and Ringwald M. The mouse Gene Expression Database (GXD): 2019 update. Nucleic Acids Research 2019 Jan; 47:D774–D779. DOI: 10.1093/nar/gky922. Available from: http://www.informatics.jax.%20https://academic.oup.com/nar/article/47/D1/D774/5133672

23. Lein E, Borm LE, and Linnarsson S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. 2017. DOI: 10.1126/science.aan6827

24. Saviano A, Henderson NC, and Baumert TF. Single-cell genomics and spatial transcriptomics: Discovery of novel cell states and cellular interactions in liver physiology and disease biology. Journal of Hepatology 2020 Nov; 73:1219–30. DOI: 10.1016/j.jhep.2020.06.004. Available from: https://linkinghub.elsevier.com/retrieve/pii/S016882782030372X

25. Svensson V, Veiga Beltrame E da, and Pachter L. A curated database reveals trends in single-cell transcriptomics. Database 2020 Jan; 2020. DOI: 10.1093/database/baaa073. Available from: https://doi.org/10.1093/database/baaa073

26. Zappia L, Phipson B, and Oshlack A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLOS Computational Biology 2018 Jun; 14:e1006245. Available from: https://doi.org/10.1371/journal.pcbi.1006245

27. XDB3. 2004. Available from: http://xenopus.nibb.ac.jp/

Chapter 4

PREQUEL ERA

Some previous reviews on spatial transcriptomics start the history of spatial transcriptomics with laser capture microdissection (LCM) followed by microarray or RNA-seq and single molecular fluorescent *in situ* hybridization (smFISH) in the late 1990s [1, 2, 3]. We will discuss these later, but note that by 1999 and the early 2000s, when the earliest LCM microarray studies were published [4, 5, 6, 7], the quest to profile the transcriptome in space had already begun, with enhancer and gene trap screens, *in situ* reporter screens, and (whole mount) *in situ* hybridization ((WM)ISH) atlases. Although this early literature, dating from the late 1980s, generally does not refer to itself as "spatial transcriptomics", it fits into the definition of spatial transcriptomics as stated in Chapter 3.

We call this body of literature "prequel", because first, its origin predates LCM microarray. Second, unlike most technologies covered by existing spatial transcriptomics reviews, the techniques used were not multiplexed and were less quantitative, and as a result, they have fallen out of favor. In contrast, what comes after "prequel" will be called "current", although the prequel and current eras chronologically overlap. Given what current era spatial transcriptomics is commonly perceived to be, here "prequel" is broadly defined as methods that fulfill the more relaxed definition of "spatial transcriptomics" in this book, but do not involve cDNA microarray, next generation sequencing (NGS), or single molecular imaging.

There are 207 prequel papers in our database. Prequel literature is included in the database and covered here for the following reasons. First, the legacy of the prequel era has influenced more recent spatial transcriptomic research; the present and future are shaped by the past. For example, spatial reconstruction of scRNA-seq data in Seurat v1 [8], the Achim et al. *Platynereis* study [9], `DistMap` [10], and the Zeisel et al. Mouse Brain Atlas [11] used (WM)ISH atlases as spatial references. Recent Spatial Transcriptomics™ (ST) mouse brain data are still compared to the ISH atlas of Allen Brain Atlas (ABA) [12, 13]. A study on spatial reconstruction of scATAC-seq data compared the *in silico* reconstruction to the FlyLight *Drosophila* enhancer atlas [14, 15]. Hence prequel resources can still be useful in the current era. We expand on this in Chapter 6. Second, some features of the prequel era

may benefit future spatial transcriptomics studies; this will be discussed after more recent technologies are reviewed. Third, the various quests in the current era have already begun in the prequel era, and this history can show how the coming together of new technologies made us better at achieving the previous generation's dreams.

Fourth, as shown later in this book, existing current era spatial transcriptomics data are by and large from humans and mice, and especially the brain (Figure 6.4, Figure 4.7). For other model and non-model organisms (e.g. *Xenopus laevis* [16, 17], *Ciona intestinalis* [18], *Danio rerio* [19, 20], *Oryzias latipes* [21], *Gallus gallus* [22], *Taeniopygia guttata* [23], and to some extent, even *Drosophila melanogaster* ([24, 25], some tissues other than the brain (e.g. lung [26] (prior to the increase interest following the COVID pandemic), retina [27], genitourinary tract [28], and miRNAs [29, 30, 31, 32, 33, 34] [34], the most comprehensive spatial transcriptomic resources, if any are are available at all, are still (WM)ISH atlases. For plants, the most comprehensive resources can still be enhancer and gene trap screens [35, 36]. Hence, while current era technologies may produce more quantitative and highly multiplexed data, they have not completely superseded (WM)ISH atlases. This may be likened to the Jet Age in the history of aviation. While massive jet airliners made aviation available to the masses so when most people fly they fly with jets, jet airliners have not completely superseded airplanes with reciprocating engines and propellers; the latter are still very common in general aviation. Finally, the historical literature is curated for the same reason why museums and libraries keep historical maps and scientific works that have been superseded by more recent work; it is part of our heritage.

An overall timeline for prequel techniques is shown in Figure 4.1, which will be discussed in more detail in the rest of this chapter.

4.1 Enhancer and gene traps

Long before the advent of reference genomes for common model organisms, the quest to characterize genes based on expression pattern in space had already begun. The earliest high-throughput efforts to identify and characterize such genes were enhancer traps. To the best of our knowledge, the first use of a reporter to visualize gene expression in space was reported in 1983. It used lacZ fused to sequences upstream to the hsp70 gene encoding a heat shock protein in *Drosophila melanogaster* and inserted into the genome with P element to characterize the puffs formed in polytene chromosomes and the tissue distribution of hsp70 in response to

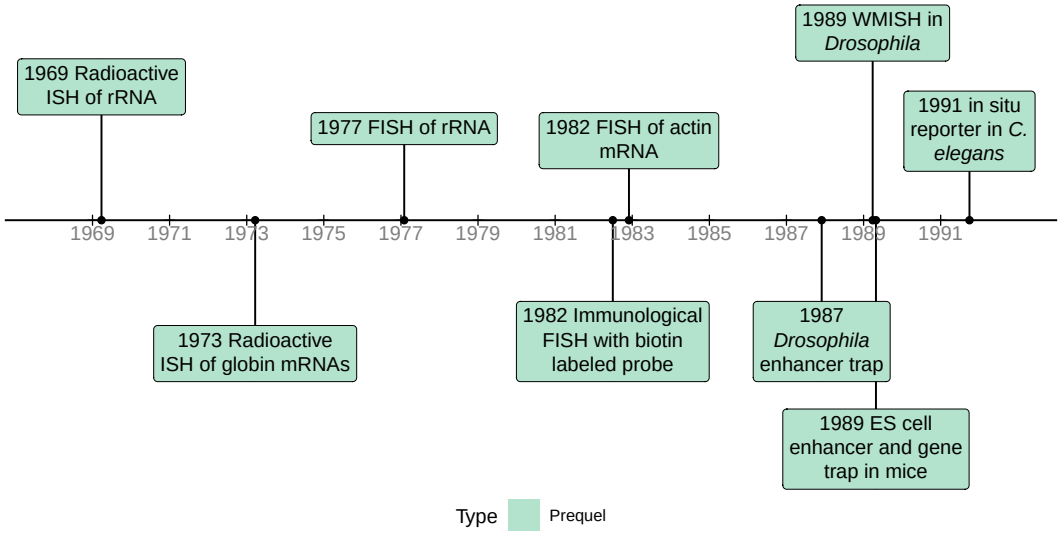

Figure 4.1: Timeline of prequel techniques.

heat shock [37].

The first enhancer trap screen in *Drosophila melanogaster* was published in 1987 [38]O'Kane and Gehring 1987. The P element is a transposable element found in *Drosophila*. In an enhancer trap vector, a reporter gene, such as lacZ, here with the polyadenylation site of the hsp70 gene, and a marker gene with its own promoter that can be used to identify individuals and their offspring with the vector integrated into the germline, such as rosy which can be used in *Drosophila* to identify the individuals with eye color, are flanked by the 5' and 3' ends of the P element necessary for transposition (Figure 4.2). The vector is injected into *Drosophila* embryos before the formation of pole cells [39]. As a transposon, the construct is randomly inserted into the genome, and since the P element promoter is so weak that an enhancer is required for the promoter to drive transcription of the reporter gene, the location of the reporter gene expression marks where the enhancer is active. As the transposon is inserted into different locations of the genome in different individuals, each individual that has the vector integrated into the germline forms a transformant line. In *Drosophila*, in many cases, expression patterns of β-galactosidase do reflect expression pattern of a nearby gene [40, 41].

Since then, different vectors have been developed for better efficiency and flexibility [43], and enhancer traps have been applied at increasing scale. The 1987 study recovered 39 lines [38], possibly characterizing 39 genes, but already in 1989, over 3000 lines were possible in one study [44]. Enhancer trapping was also adapted to

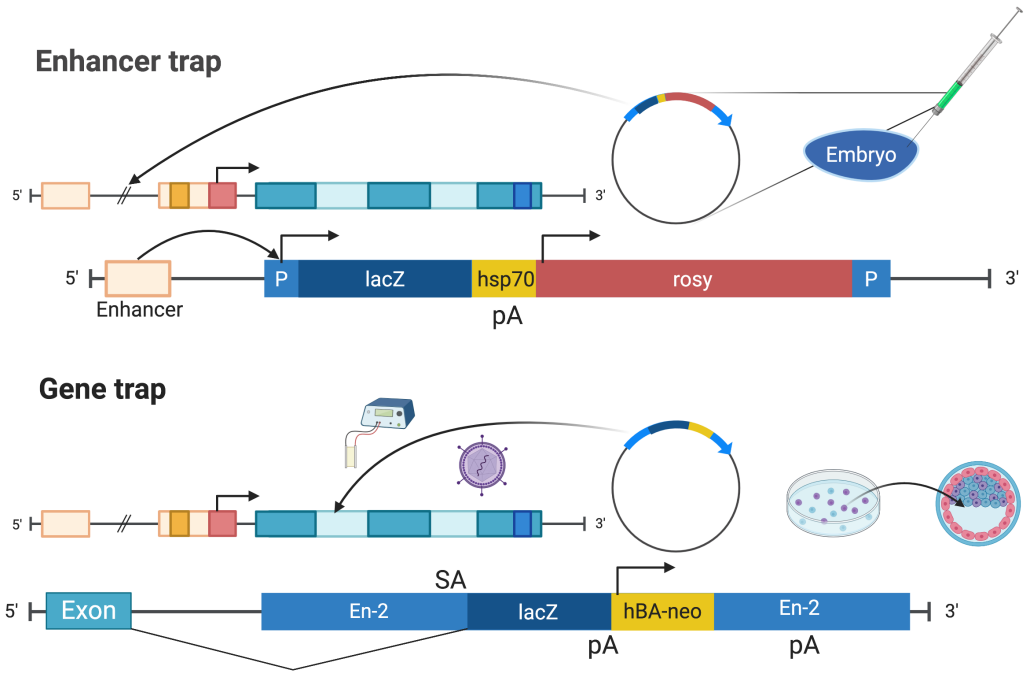

Figure 4.2: Illustrations of enhancer trap as described in ([38]O'Kane and Gehring 1987) and gene trap as described in ([42]Gossler et al. 1989) (Created with BioRender.com).

other species, such as mouse [42, 45] and *Arabidopsis thaliana* [46].

Enhancer traps were not intended to be mutagenic [38], nor is it highly mutagenic [43]. Gene trap and promoter traps were introduced to not only screen for genes with restricted expression patterns, but also to enable functional analysis of the gene from homozygote mutant phenotypes [47]. Like the typical enhancer trap vector, gene trap and promoter trap vectors contain a reporter gene, such as lacZ (β-gal), to visualize gene expression, and sometimes also a marker to screen for integration, such as the neomycin resistance gene (neo). Though often, lacZ itself, or in a fusion with neo (β-geo), was also used as the marker when screening mouse embryonic stem (ES) cells (Figure 4.2).

Unlike the enhancer trap vector, gene trap and promoter trap vectors do not have a promoter for the reporter, though the marker, if present, can have its own promoter. In a promoter trap, the construct needs to be inserted in frame and in the correct orientation into an exon of a gene to be expressed, making it very inefficient [47, 43].

In contrast, in gene traps, a splice acceptance site is added to the 5' end of the reporter, so the construct can be expressed when inserted into an intron at the right orientation; this is over 50 times more efficient than a promoter trap because introns tend to be much longer than exons and the construct does not have to be in frame to an exon [47, 43]. Gene traps and promoter traps are mutagenic as the reporter has a stop codon, thus truncating the endogenous protein.

While enhancer traps are more commonly used in *Drosophila*, gene traps are more commonly used in mice. In mice, in 1988, the enhancer trap vector was initially introduced by injection into the male pronucleus in the fertilized egg [45]. The throughput of the screen is increased by inserting the construct into genomes of ES cells by electroporation or retroviral infection [43], screening for ES cells expressing lacZ or the marker, injecting these ES cells into blastocysts to generate chimeric mice to characterize gene expression patterns; chimera are especially useful for characterizing dominant and lethal mutations [47, 42].

The first gene trap screen in mouse ES cells was reported in 1989 [42], recovering 14 lines. Again, variants of the vector emerged and gene trap screens increased in scale. In 1995, nearly 300 mouse gene trap lines were recovered from one study [48]. Later, smaller gene trap studies specific to particular types of genes made possible by additional steps to screen ES cell colonies were performed, such as genes encoding membrane and secreted proteins [49], genes responding to retinoic acid [50], and genes expressed in hematopoitic and endothelial lineages [51]. In 2001, gene trapping was used to examine not only expression pattern of genes in cell bodies of neurons in the mouse brain, but also axon guidance [52]. By 2001, a number of gene trap consortia have been established as resources of gene trap vectors and transformant mouse ES cell lines, hoping to create at least one line for each gene in the mouse genome [43].

In the 1980s and 1990s, with increasing throughput of Sanger sequencing and the advent of shotgun sequencing, the amount of sequencing data in GenBank exploded [53]. With 5' or 3' rapid amplification of cDNA ends (RACE) PCR, the fusion transcript of the reporter and an endogenous gene could be cloned [54], sequenced, and potentially aligned to the existing sequences to identify the gene of interest [51]. However, the golden age of gene trapping was soon to pass, with the rise of ISH atlases in the late 1990s and the advent of reference genomes of *Drosophila melanogaster* [55], mouse [56], and human [57, 58]Lander et al. 2001 in the early 2000s that would make it easier to design ISH probes from the reference genome

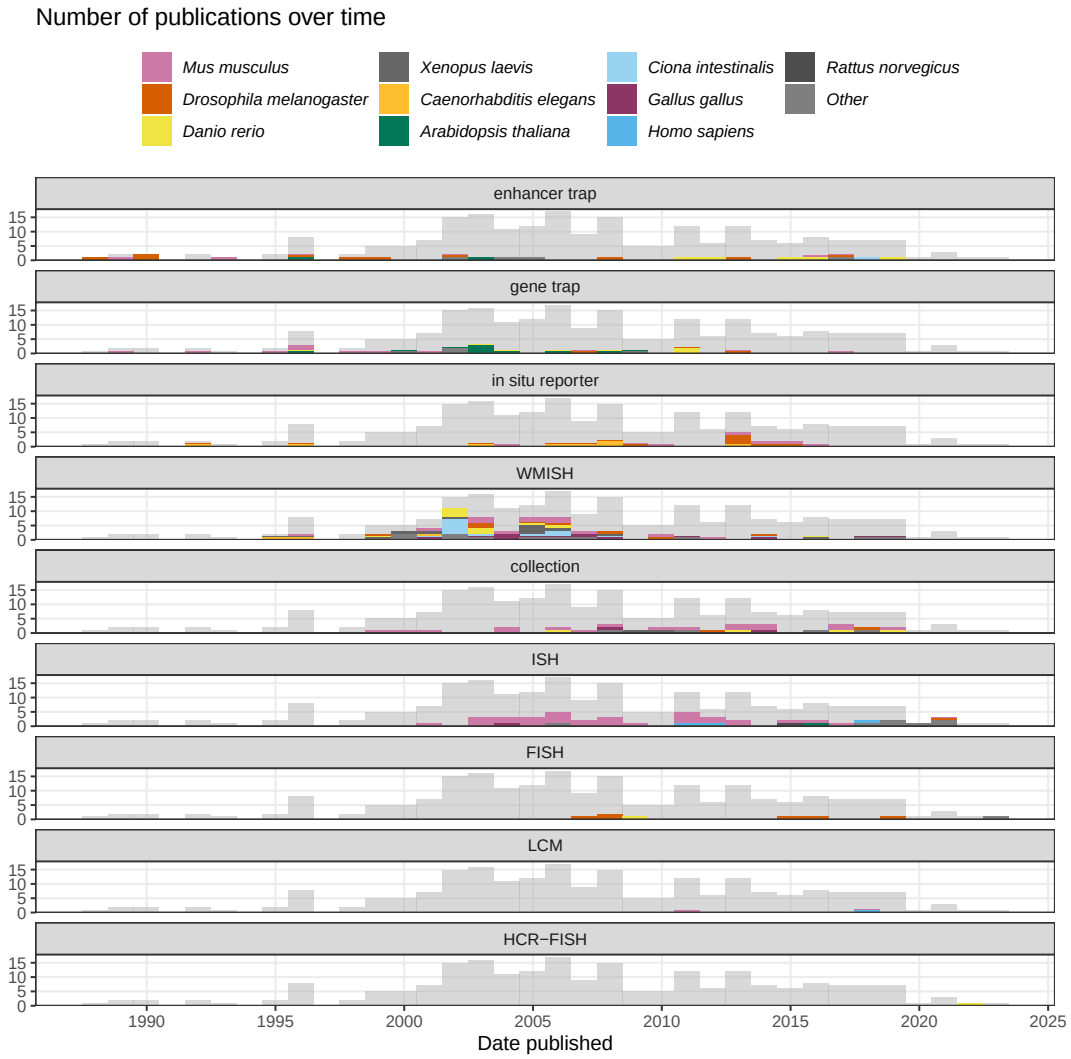

Figure 4.3: Number of publications over time in the prequel era, broken down by technique and colored by species. The gray histogram in the background is the histogram for all prequel publications over time. The bin width of this histogram is 365 days. Here WMISH and ISH exclude fluorescent ISH (FISH).

to target annotated genes, as is done today. Nevertheless, enhancer and gene traps were not rendered obsolete by these developments. They have been used in plants and zebrafish through the 2000s and 2010s, as resources of gene expression patterns [35, 36, 59, 60, 61, 62] (Figure 4.3).

4.2 In situ reporter

In enhancer, gene, or promoter trap screens, the reporter is randomly inserted into the genome, not targeting predetermined genes. In contrast, in what we call *in situ* reporter screens, the reporter is fused to predefined regulatory sequences of a gene of

interest, with the hope that expression pattern of the reporter would recapitulate that of the gene of interest. Chronologically, this is the second type of high throughput method to profile gene expression patterns (Figure 4.3).

A precursor to this type of method was used in 1991, where random genomic fragments were fused to a lacZ reporter lacking a transcription start signal and injected as plasmids, screening for fragments driving lacZ expression and characterizing the expression patterns in *C. elegans* [63]. To the best of our knowledge, the first time *in situ* reporter with predefined regulatory sequences was used to screen for gene expression patterns in a multicellular organism, was in 1995, in *C. elegans* [64]. At that time, the *C. elegans* genome sequencing project was already in progress [64, 65], and the genome sequence was declared "essentially complete" in 1998 [66]. Computationally predicted upstream regulatory sequences of 35 putative genes were fused to a promoterless lacZ as a reporter, cloned into plasmid vectors, and microinjected into *C. elegans* gonads to create transformed lines then stained with X-gal [64].

A reliable *in situ* reporter was first reported in mice in 1997. It used a recombinant bacterial artificial chromosome (BAC) with part of the full RU49 gene in the BAC replaced by a lacZ construct and showed that the construct is heritable [67]. In 2003, a similar strategy, replacing coding sequences of genes in BACs with EGFP reporter gene, was used to create a mouse brain gene expression atlas GENSAT with BAC transgenic mouse lines [68]. The GENSAT lines were used again in 2009 to create a gene expression atlas for retina [69]. Again, GENSAT benefited from the reference genome, which greatly helped with identifying BACs that include sequences flanking a gene that may contain regulatory elements that make the reporter better recapitulate expression pattern of the endogenous gene [68].

Through the 2000s and 2010s, *in situ* reporters have been used as a targeted alternative to enhancer and gene trap screens informed by the reference genomes. To address limitations of gene traps, such as inability to precisely define the allele and favoring genes expressed in ES cells when screening for transformant colonies, high-throughput mouse knock out resources with knock out alleles computationally designed according to a reference genome and annotations have been established [70, 71]. As these alleles contain a lacZ reporter, these resources have been used to characterize gene expression in over 40 tissues in mutant mice with lacZ staining [72, 73, 74]. However, for some tissues, only low resolution whole mount staining was performed. Similarly, in both mouse [75] and *Drosophila* [14, 76], transgenic

lines with genomic fragments containing putative enhancers driving expression of reporter genes were established as alternatives to enhancer traps. The enhancer candidates can be selected from sequence homology and ChIP-seq predictions [75], or from tiles of sequences flanking genes thought to have restricted expression patterns or within introns of such genes [14].

In situ reporter atlases exceeded the scale of enhancer and gene trap screens. The largest such atlas in *C. elegans*, WormAtlas, profiled 1886 genes [77]; we are unaware of enhancer and gene trap screens in *C. elegans* because *C. elegans* genome sequencing was already underway by 1992 [65], making *in situ* reporter screening feasible before it was so in mice and *Drosophila*. The largest such study in *Drosophila* profiled 7705 enhancer candidates [76], which far exceeded the 3768 enhancer trap lines in 1989 [44]. *In situ* reporters were used in mice to profile up to 536 genes [69] and 329 enhancer candidates [75], while the large scale gene trap screen in 1995 only reached 279 lines [48] and later mouse gene trap screens did not typically exceed 100 lines. However, where comparable, *in situ* reporter atlases never reached the scale of (WM)ISH atlases, perhaps because of the large number of transgenic lines required. Allen Brain Atlas (ABA) profiled over 20,000 genes in the mouse brain, and as of April 2021, the Berkeley Drosophila Genome Project (BDGP) WMISH atlas already has 8533 genes. However, *in situ* reporters might still be a good way to profile enhancer usage in space.

4.3 ISH and WMISH atlases

In situ hybridization was first used in 1969 to visualize ribosomal RNA (rRNA) [78] and ribosomal DNA (rDNA) [79] in *Xenopus laevis* oocytes with probes labeled with radioisotope ^{3}H (Figure 4.1). To the best of our knowledge, the earliest use of ISH to visualize what was thought to be a specific transcript was done in 1973, to visualize globin mRNAs in various cultured erythroid and non-erythoid cell types by hybridization of radiolabeled cDNA to the mRNA [80]. As radioactive ISH requires long exposure time (several weeks), has low spatial resolution and high background, and requires handling hazardous radioactive material, alternatives emerged in the mid 1970s and early 1980s. Among the alternatives were variants of FISH and labeled probes detected by primary and enzyme or fluorophore labeled secondary antibodies [81, 82]; the latter, immunological method is commonly used in ISH and WMISH atlases. To the best of our knowledge, the first report of using immunological fluorescent and peroxidase ISH to visualize expression of a specific gene was published in 1982, the same year such a technique was published [82],

visualizing actin transcripts in chicken muscle tissue culture; the authors reported puncta of cytoplasmic fluorescence which might be clumps of mRNAs or artefact, but could possibly be individual transcripts [83].

Non-radioactive ISH not only has shorter exposure time and higher resolution than radioactive ISH, but also made WMISH possible. WMISH was first reported in Drosophila embryos in 1989 [84], and was adapted to other model organisms such as mice, *Xenopus laevis*, and *Paracentrotus lividus* (purple sea urchin) in the early 1990s [85]. Advantages of WMISH compared to section ISH is preservation of 3D structure of the tissue, ease of interpretation in blastoderm stage embryos, and ease of performing ISH on larger number of embryos [85, 84].

Just like genome sequencing in multi-cellular organisms and *in situ* reporter screens, WMISH atlases got a head start in *C. elegans*. The first WMISH screen with higher throughput than typically used on select marker genes was reported in 1994, of 21 genes in *C. elegans* [86]. Early (WM)ISH atlases in the late 1990s typically made probes from cDNA clones from poly-A selected RNAs in tissue or developmental stage of interest without pre-selecting genes to stain for [24, 87, 88, 89]. Some early atlases were intended to be improvements to enhancer and gene trapping and *in situ* reporter screens, as a simpler and more direct alternative [89] or as a way that can better capture endogenous and dynamic spatial distribution of transcripts [88]. Since 1998, (WM)ISH has been automated, enabling staining for thousands of probes [88, 90].

The genes from which the clones come from were often unknown, so early (WM)ISH atlases referred to the entities stained for as "clones" (Figure 4.4), though the genes, homology, and putative functions of the genes can be identified by aligning sequences of the cDNA clones to existing sequences in databases [89, 88, 91]. However, again, the first WMISH screen with probes made from cloning PCR amplified pre-defined genomic sequences was performed in *C. elegans* in 1995 [92]. By the turn of the century, the entities stained for were sometimes referred to as "clusters", especially in the GHOST atlas for *Ciona intestinalis* [18] (Figure 4.4); the sequences of the probes were clustered by alignment and these probes might have come from the same gene.

The rise of (WM)ISH atlases started before the completion of genome projects in humans and common model organisms, although their later growth was transformed by the reference genome. In the 2000s, with the availability of sequenced cDNA collections covering increasing proportion of predicted genes and the consequent

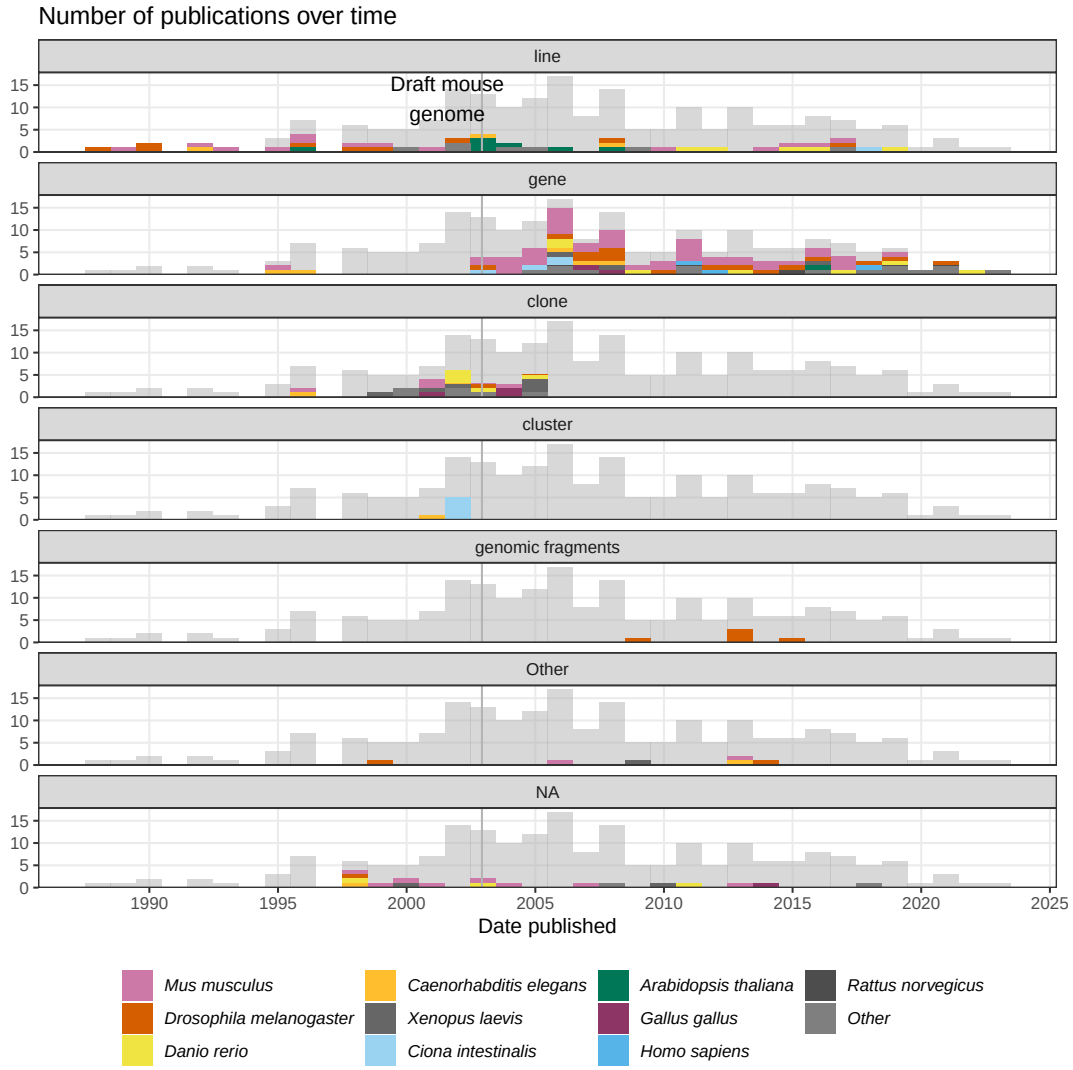

Figure 4.4: Number of prequel publications over time, broken down by what the entities stained for were called and colored by species. Bin width is 365 days. Vertical line marks the date when the draft mouse reference genome was published [56], as context of transition from “clone” and “line” to “gene”.

rise of transcriptome-wide microarray [87, 93], genes to be stained for in (WM)ISH atlases could be pre-screened based on microarray data of the tissue of interest, with probes made from cDNA clones readily available from such collections [94, 95]. In addition, probes could be computationally designed based on reference genome sequences [96]. Perhaps because of these developments, since the turn of the century, entities stained for have been predominantly referred to as "genes" (Figure 4.4). Notably, while radioactive ISH has been mostly replaced by non-radioactive ISH by the 2000s, there is a mouse hippocampus ISH atlas published in 2004 that used radioactive ISH to profile all of its 104 genes [95].

Also with the rise of cDNA microarray in the late 1990s and early 2000s, some (WM)ISH atlases were made as an improvement to microarray with bulk tissue to profile the transcriptome, not only at cellular resolution, but also preserving spatial and sometimes temporal context [96, 22], analogous to how scRNA-seq and various later forms of spatial transcriptomics were developed in response to bulk RNA-seq.

Since the 2000s, (WM)ISH atlases have been made for specific types of genes and a number of mouse tissues. In 2004, locked nucleic acid (LNA) modified oligonucleotide probes were introduced, greatly improving sensitivity of miRNA northern blot [97] and made (WM)ISH atlases for miRNAs possible. The first miRNA WMISH atlas was published in 2005, which profiled 115 miRNAs in zebrafish embryos [33]. Since then, miRNA atlases were created for mice [30, 31, 98], *Drosophila* [32], chicken [34], and *Xenopus laevis* [29].

While (WM)ISH atlases are available for several species, the mouse is by far the favored model organism (Figure 4.5). A timeline of the first (WM)ISH atlas for each of the species and some notable atlases are shown in Figure 4.6. Especially for mice, atlases for other specific types of genes were published in the late 2000s and the 2010s, such as genes coding for RNA binding proteins [99], fibroblast growth factors and their receptors [100], proteins with catalytic activities [101], transcription factors and cofactors [102], metabolic enzymes and soluble carriers [103], cholesterol biosynthetic enzymes [104], and ion channels (in rats) [105]. Among the mouse atlases, while the brain gets disproportionately strong interests, with the influential ABA [96] and GenePaint [90], ISH atlases exist for the eye [106, 27], genitourinary tract (GenitoUrinary Development Molecular Anatomy Project (GUDMAP)) [28], and lung (LungMAP) [26] (Figure 4.6, Figure 4.7).

While the vast majority of (WM)ISH atlases used bright field imaging, a few used FISH (Figure 4.3), for advantages conferred by FISH discussed below. A notable

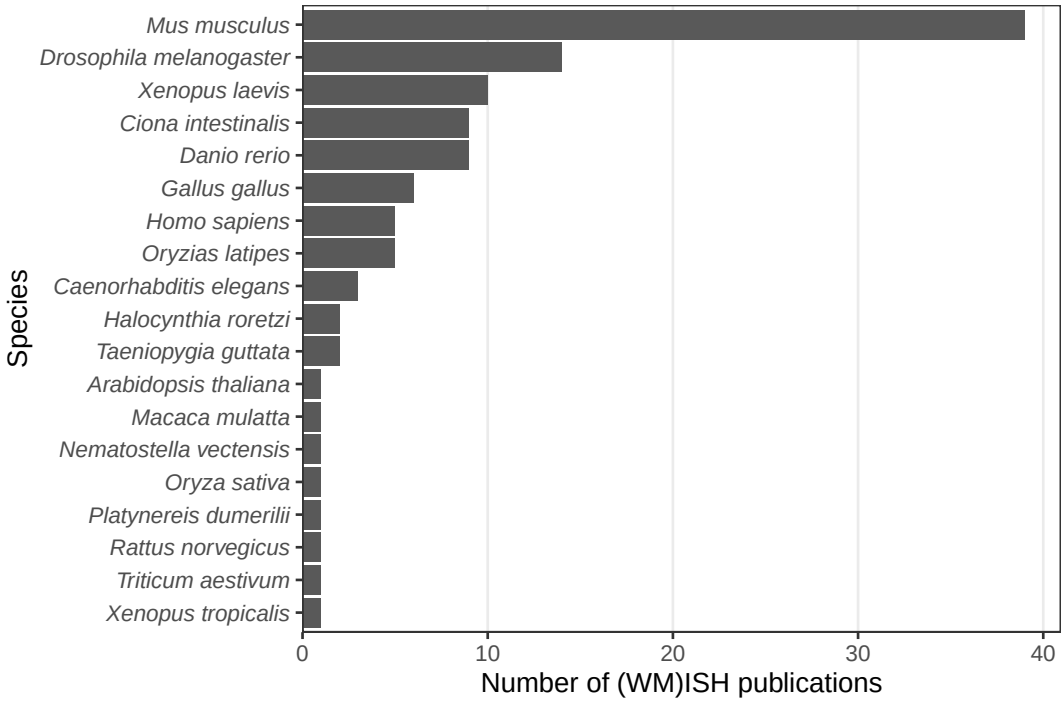

Figure 4.5: Number of (WM)ISH publications per species.

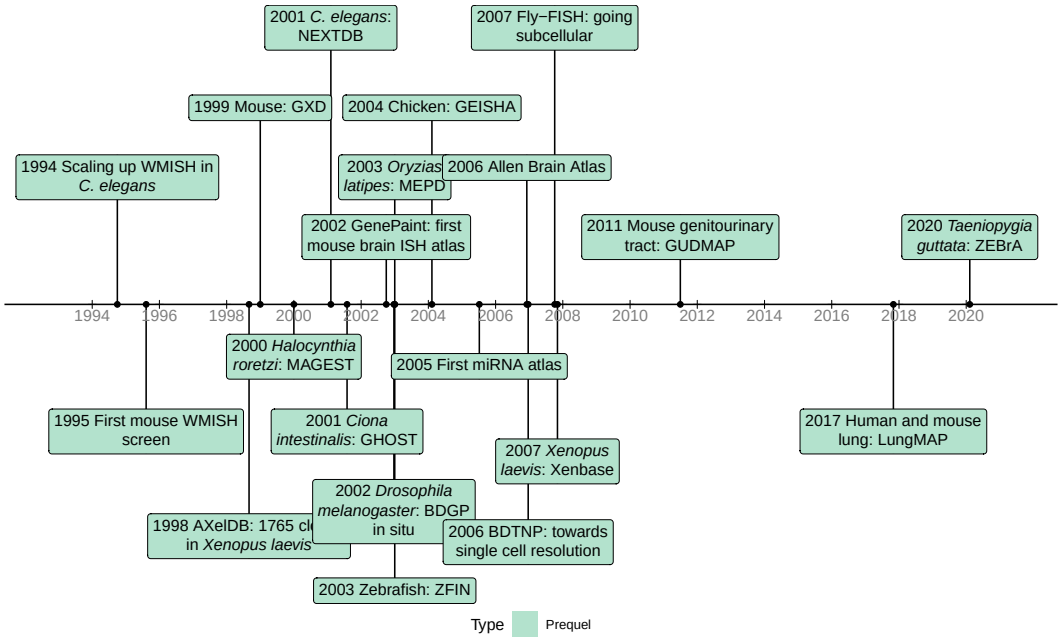

Figure 4.6: Timeline of the first (WM)ISH databases for each species for which such databases are available, as well as some notable databases.

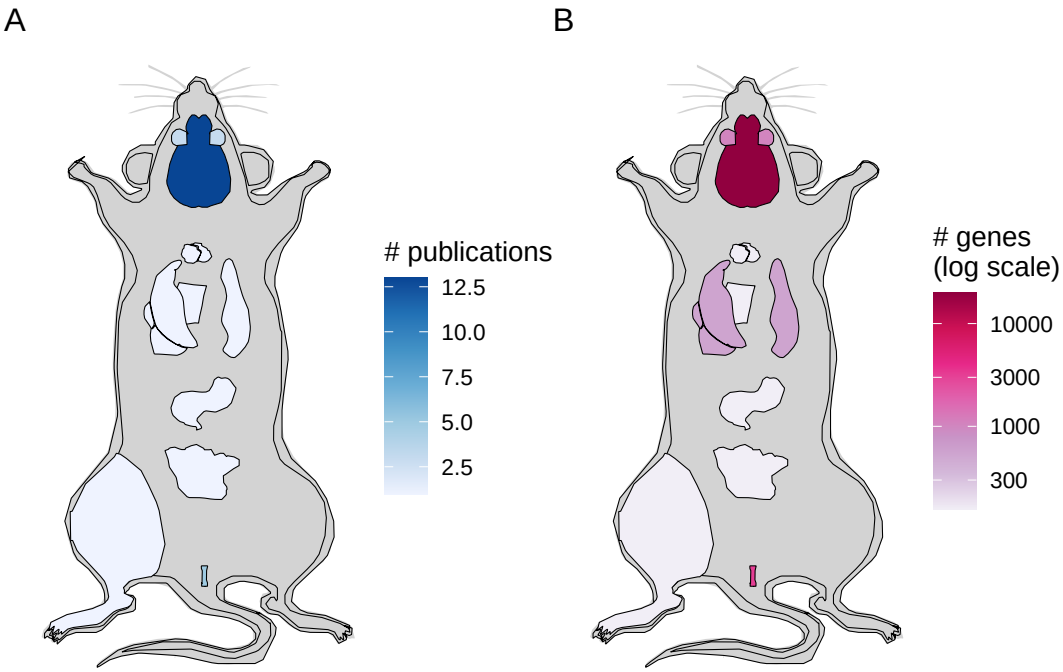

Figure 4.7: A) Number of mouse publications per organ for (WM)ISH atlases (including FISH). B) Maximum number of genes in atlases for each organ, as of publication of the paper about the atlases. The color is in log scale to improve dynamic range.

FISH atlas is the Berkeley *Drosophila* Transcription Network Project (BDTNP) from 2006 to 2008, which profiled expression patterns of 95 genes in the *Drosophila* embryo across 6 developmental stages up to the beginning of gastrulation [107, 25]. Two genes are imaged in each embryo, and the images of 1822 embryos were registered across both space and time to construct 3D virtual embryos on which patterns of different genes can be quantitatively compared [107]; the 3D imaging and penetration into the opaque yolk is made possible by two photon microscopy, in which only the fluorophores in the region of focus are excited [25]. Another notable FISH atlas is Fly-FISH from 2007, which profiled subcellular localization of transcripts of 3370 genes in *Drosophila* embryos [108]. While subcellular localization of transcripts can sometimes be discerned in bright field WMISH [24], Fly-FISH shows higher subcellular resolution thanks to a FISH protocol using tyramide signal amplification. To our best knowledge, this is the first transcriptomic atlas of a multi-cellular organism to profile subcellular transcript localization. While more recent smFISH-based methods record subcellular information, such information is typically not used in downstream analyses.

WMISH was the most commonly used technique in the prequel era, followed by

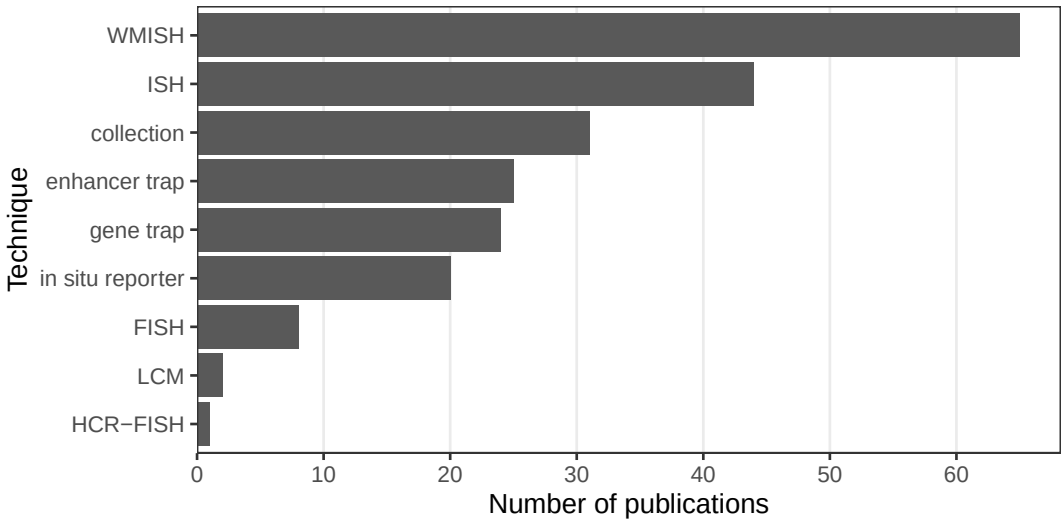

Figure 4.8: Number of prequel publications per technique.

ISH (Figure 4.8). In summary, advances of non-radioactive ISH and WMISH from radioactive ISH, limitations of enhancer and gene trap and *in situ* reporter screens, cDNA collections that cover most of predicted genes, limitations of bulk microarray, reference genomes that allow for computational probe design, and ISH robots may have been responsible for the rise of (WM)ISH atlases. Another important factor may be the rise of digital photography and the internet in the 1990s, as developing thousands of analogue photos is an arduous task. Moreover, online digital atlases have been much more accessible to the wider community. Assuming that the number of publications in a field reflects interest in that field during a period of time, and if our collection is representative of the actual body of literature, then the golden age of the prequel era was the 2000s and WMISH was responsible for that peak, while section ISH and "collection", i.e. databases of gene expression patterns curated from publications and some (WM)ISH atlases, account for much of the interest after 2010 (Figure 4.3). The websites of many of the older (WM)ISH atlases are no longer accessible. However, some of the atlases from that period of time still live on in extant curated databases, which will be discussed in the next section.

The golden age declined before the rise of current era spatial transcriptomics, which started around 2014 6.2. What contributed to the decline of the golden age? Perhaps with proliferation of such atlases, curated databases exceeding 10,000 genes, and especially with over 20,000 genes in ABA mapped to a high quality 3D mouse brain model, there are already enough gene expression pattern resources for the most commonly studied genes, tissues (especially the brain), and developmental stages

in the most common model organisms, thus making new atlases in those systems unnecessary. Moreover, in the last decade, the under-utilization of gene expression atlases [109] may have reduced motivation to build new atlases. Or perhaps, more importantly, inherent limitations of non-multiplexed (WM)ISH contributed to the decline in interest in such methods. In these atlases, typically only one gene is stained for in each individual embryo or tissue section. Gene expression patterns of different genes can only be meaningfully compared and classified in tissues with a stereotypical structure, such as wild type embryos and the brain, but not tumors and pathological tissues, even though there is intense interest in spatial transcriptomics in tumors as evidenced by the LCM and ST literature 8.3. A large number of embryos or sections are required for such atlases, thus increasing cost and making human atlases extremely difficult and costly, if ethical at all. Furthermore, since the chromogenic reaction in bright field ISH can be prolonged to increase staining intensity, the patterns are not quantitative and consequently, analyses of such patterns typically involve binarization and quantitative expression levels of genes cannot be compared. Even with a stereotypical structure, image registration can be challenging because of biological differences between individuals [107].

4.4 Databases of the prequel era

Many of the (WM)ISH atlases discussed above, such as BDGP [24], Gallus *In Situ* Hybridization Atlas (GEISHA) [22], ABA [96], BDTNP [107], GUDMAP [28], and LungMAP [26] are stored in databases that can be queried online, typically by gene symbol or by controlled anatomical or developmental vocabulary (i.e. ontology, reviewed in depth in [110]. There are additional gene expression databases for images curated from publications, some containing non-spatial data as well and some specifically for spatial data.

The rise of the curated databases started in the 1990s. Already in 1992, the challenges of managing the increasing amount of gene expression data in developmental biology emerged and a spatiotemporal database of mouse gene expression that would later become the Edinburgh Mouse Atlas of Gene Expression (EMAGE) was discussed [111]. In 1994, Jackson Laboratory proposed the Gene Expression Database (GXD) [112], in collaboration with EMAGE to build the most comprehensive mouse gene expression database. In 1997, work was already in progress to produce (WM)ISH atlases and construct the database infrastructure for mouse [113] (GXD and EMAGE), Drosophila melanogaster [114], *C. elegans* [115], and zebrafish [116]. Curated databases of mice (GXD and EMAGE), zebrafish (Zebrafish

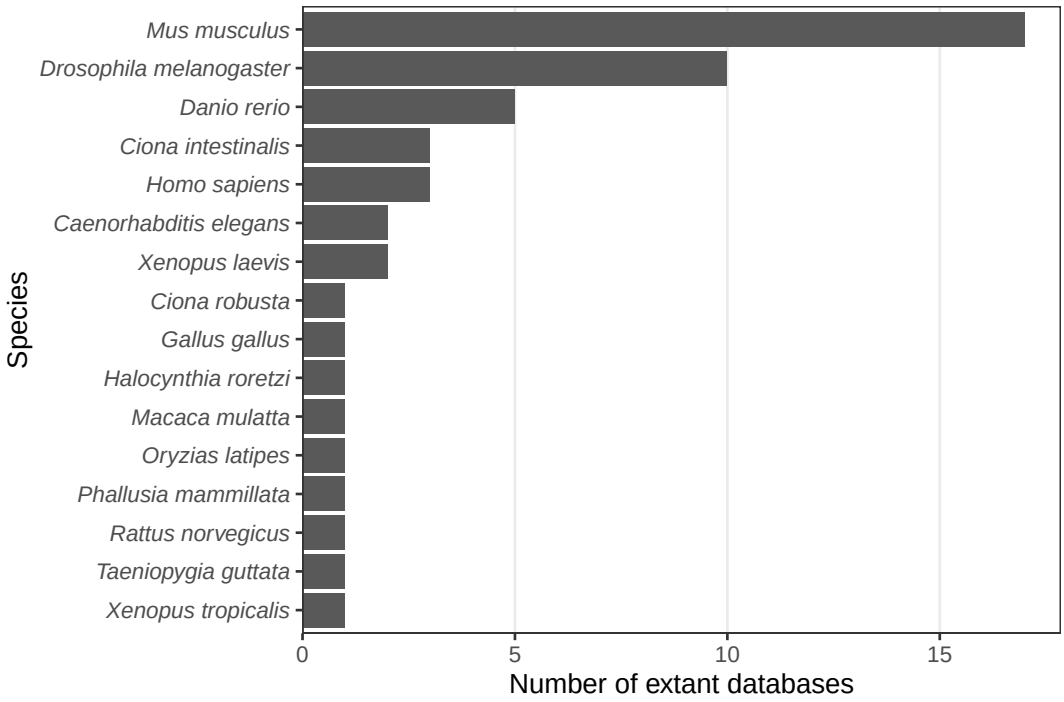

Figure 4.9: Number of extant spatial gene expression databases per species.

Information Network (ZFIN) [117]), and *Xenopus laevis* (Xenbase [16]) were released in the 2000s, within a tide of (WM)ISH atlases for new species (Figure 4.6). Some of these databases are regularly updated and the updates are responsible for many of the "collection" publications after 2010 (Figure 4.3, Figure 4.8); our historical literature collection has not only the original publications for the databases, but also publications for later updates that involve new spatial gene expression images. Examples of other extant curated databases: for *Drosophila melanogaster* FlyExpress [118], for *Xenopus laevis* XenMARK [119], and for ascidians Ascidian Network for *In Situ* Expression and Embryological Data (ANISEED) [120]. Databases, curated or not, are available for several species; mice, *Drosophila*, and zebrafish have the most extant databases (Figure 4.9).

Data can be exchanged between databases. For example, among mouse databases GenePaint [90] and EMAGE now contain data from Eurexpress [31, 109], and EMAGE uses data from GXD for the 3D gene expression models [121]. ANISEED contains data from WMISH atlases GHOST for *Ciona intestinalis* [18] and MAboya Gene Expression patterns and Sequence Tags (MAGEST) for *Halocynthia roretzi* [122]. FlyExpress contains data from *Drosophila* atlases such as BDGP and FlyFISH. Data in databases that ceased to operate may still be available in extant

databases. For instance, AXelDb WMISH atlas and database for *Xenopus laevis* [88] has been subsumed in Xenbase while AXelDb's own website has long been defunct. Likewise, as of April 2021, the MAGEST website is defunct but the data lives on in ANISEED.

Some of the databases go beyond collecting data from other databases. Databases such as EMAGE, ANISEED, and ABA registered multiple 2D section images to map gene expression patterns onto 3D anatomical models for better comparison between different genes. FlyExpress also standardized the images from the atlases and enables search for coexpressed genes by expression pattern [118]. There have also been efforts to integrate databases from multiple model organisms. In 2007, COMPARE [123] and 4DXpress [124] were developed to make gene expression patterns and developmental stages in zebrafish, mouse, and *Drosophila* (also medaka in 4DXpress) comparable. While COMPARE and 4DXpress are no longer available, interest in integrating the databases continues, so in 2016, the Alliance of Genome Resource was founded, producing a unified user interface to genome and gene expression databases for *Saccharomyces cerevisiae*, *C. elegans*, *Drosophila melanogaster*, mouse, rat, and zebrafish [125], although spatial patterns are not its focus.

4.5 Geography of the prequel era

Where were prequel era research conducted? Our database includes affiliation of the first author as of publication for all papers, and the affiliations have been geocoded to plot on maps. Around the world, most of prequel studies were performed in coastal US and Western Europe, but a some studies were performed in Asia and Oceania, but especially Japan (Figure 4.10). Not all of the top contributing institutions are readily recognizable "elite" institutions. Institutions include BDGP from UC Berkeley, ZFIN from University of Oregon (UO), ABA from Allen Brain Institute (Allen), GEISHA from University of Arizona (UofA), GXD from Jackson Laboratory (JAX), EMAGE from Western General Hospital (WGH), MEPD (for *Oryzias latipes*) from European Molecular Biology Laboratory (EMBL), and GHOST from Kyoto University (Kyodai), and mouse gene trap lines from Mount Sinai.

This can be better visualized by breaking the map down by species. Here we see locations of some model organism consortia, and that GHOST is a result of collaboration of multiple Japanese institutions (Figure 4.14).

That some institutions have disproportional contribution of one technique can also be

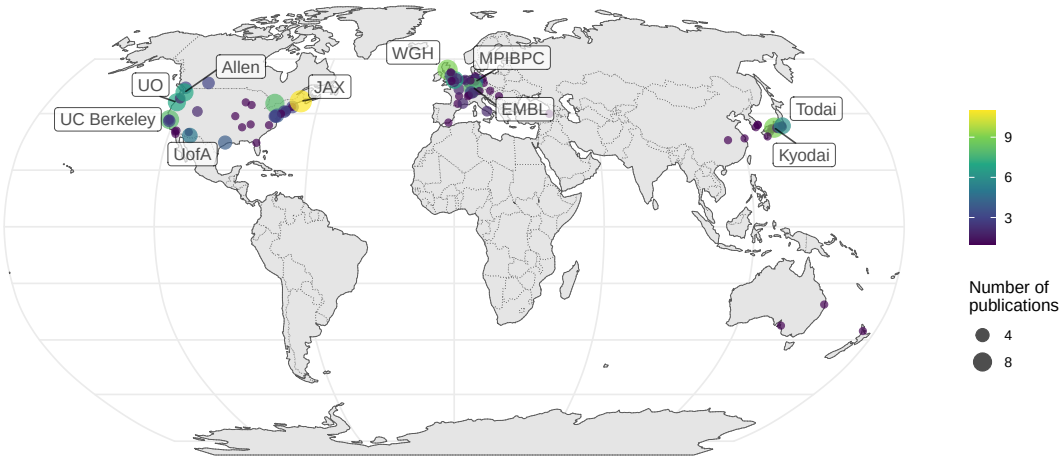

Figure 4.10: Number of prequel publications per city around the world, with top contributing institutions labeled.

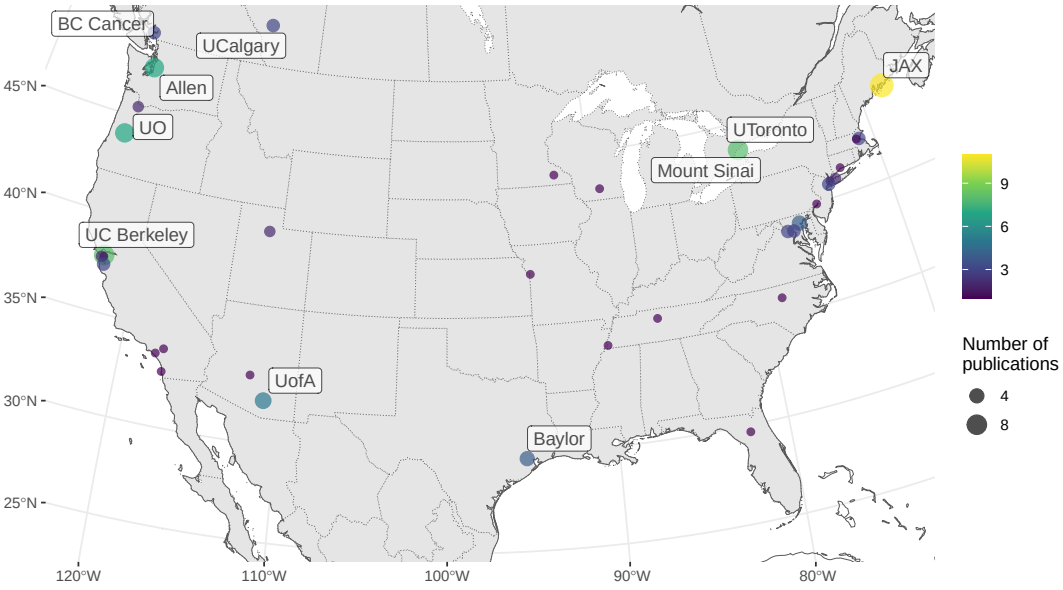

Figure 4.11: Number of prequel publications in the US and Canada, with top contributing institutions labeled.

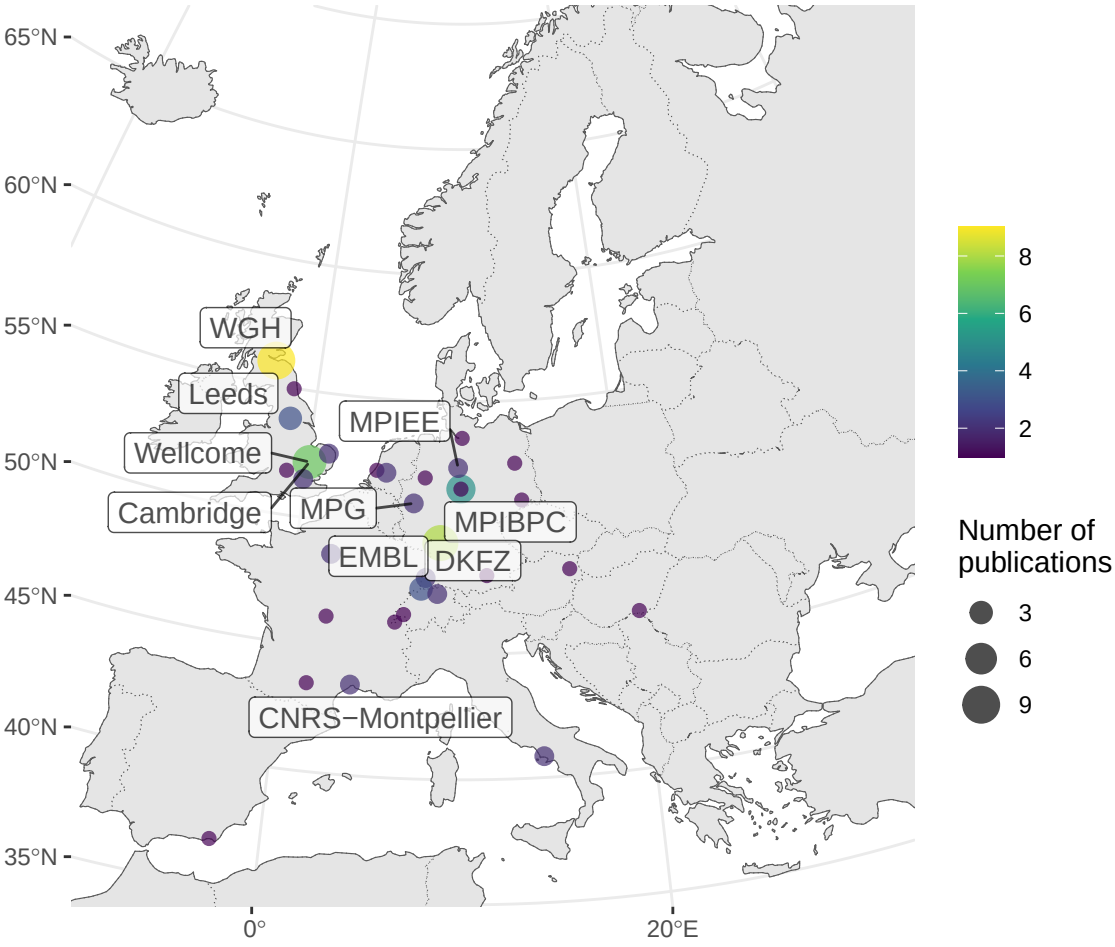

Figure 4.12: Number of prequel publications in western Europe, with top contributing institutions labeled.

shown. Here it's clear that prequel techniques are used by many different institutions (Figure 4.15). In contrast, as will be shown in Chapter 6, most current era techniques never spread beyond their institutions of origin. The LCM study comes from Allen Brain Institute's atlases for Allen's mouse sleep deprivation atlas [126] and human glioblastoma atlas [127]; although LCM is a current era technique, those two studies are in the prequel sheet because they also have ISH atlases.

References

1. Lein E, Borm LE, and Linnarsson S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. 2017. DOI: 10.1126/science.aan6827

2. Liao J, Lu X, Shao X, Zhu L, and Fan X. Uncovering an Organ's Molecular Architecture at Single-Cell Resolution by Spatially Resolved Transcriptomics.

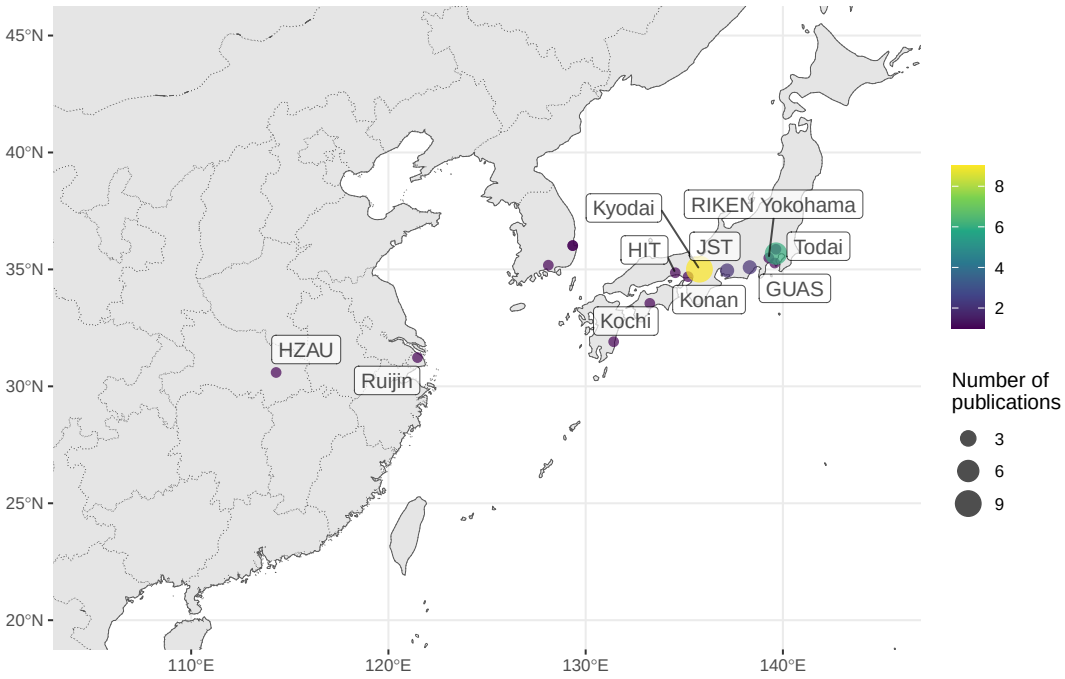

Figure 4.13: Number of prequel publications in northeast Asia, with top contributing institutions labeled.

Trends in Biotechnology 2020 Jun. DOI: 10.1016/j.tibtech.2020.05.006. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167779920301402

3. Crosetto N, Bienko M, and Van Oudenaarden A. Spatially resolved transcriptomics and beyond. 2015. DOI: 10.1038/nrg3832

4. Luo L, Salunga RC, Guo H, Bittner A, Joy K, Galindo JE, Xiao H, Rogers KE, Wan JS, Jackson MR, and Erlander MG. Gene expression profiles of laser-captured adjacent neuronal subtypes. Nature Medicine 1999 Jan; 5:117–22. DOI: 10.1038/4806. Available from: http://www.nature.com/articles/nm0199_117

5. Sgroi DC, Teng S, Robinson G, LeVangie R, Hudson JR, and Elkahloun AG. In Vivo Gene Expression Profile Analysis of Human Breast Cancer Progression. Cancer Research 1999 Nov; 59:5656 LP –5661. Available from: http://cancerres.aacrjournals.org/content/59/22/5656.abstract

6. Ohyama H, Zhang X, Kohno Y, Alevizos I, Posner M, Wong D, and Todd R. Laser Capture Microdissection-Generated Target Sample for High-Density Oligonucleotide Array Hybridization. BioTechniques 2000 Sep; 29:530–6. DOI: 10.2144/00293st05. Available from: https://www.future-science.com/doi/10.2144/00293st05

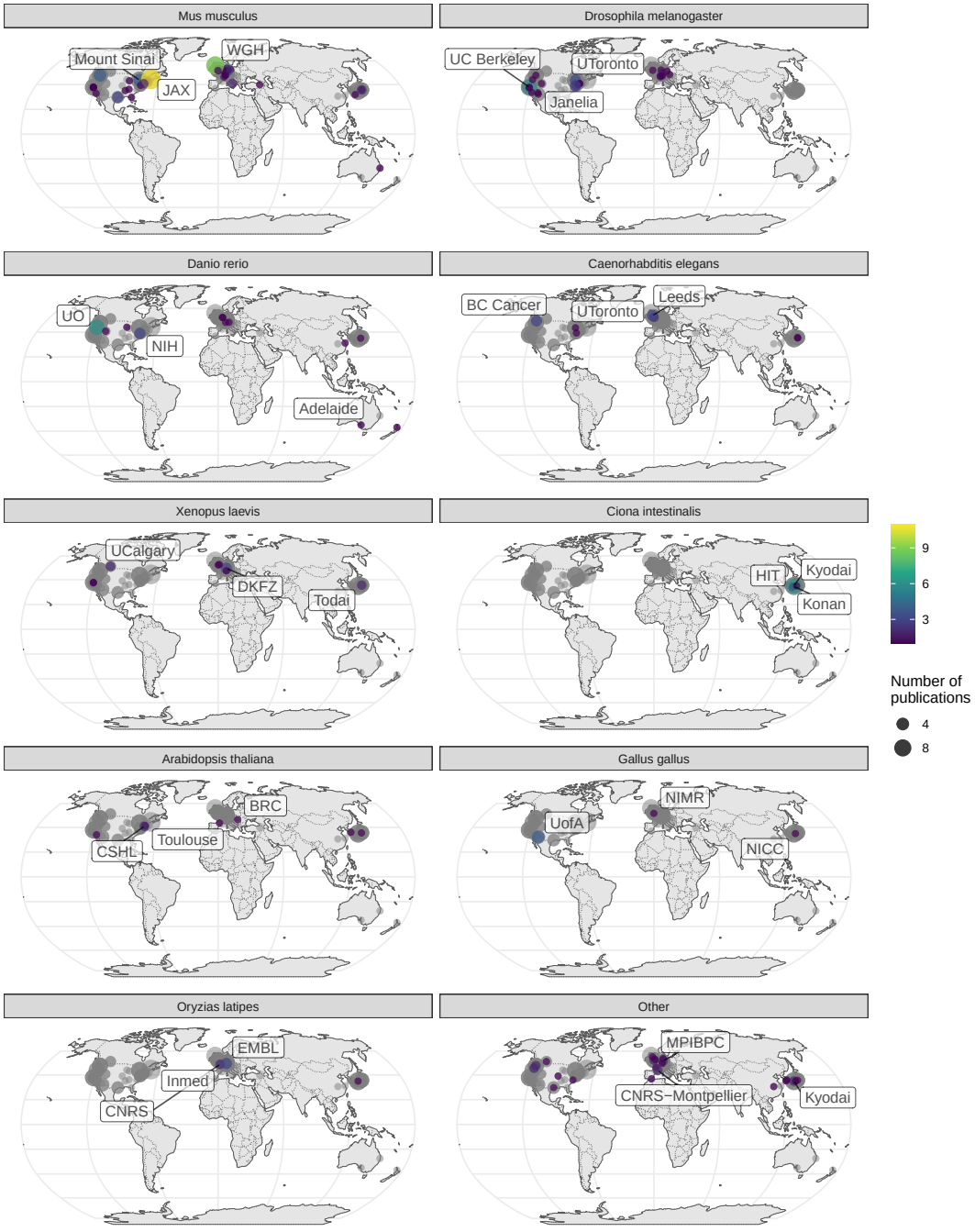

Figure 4.14: Number of prequel publications per city broken down by species. Gray points are the overall number as a reference of contributions from each city and region.

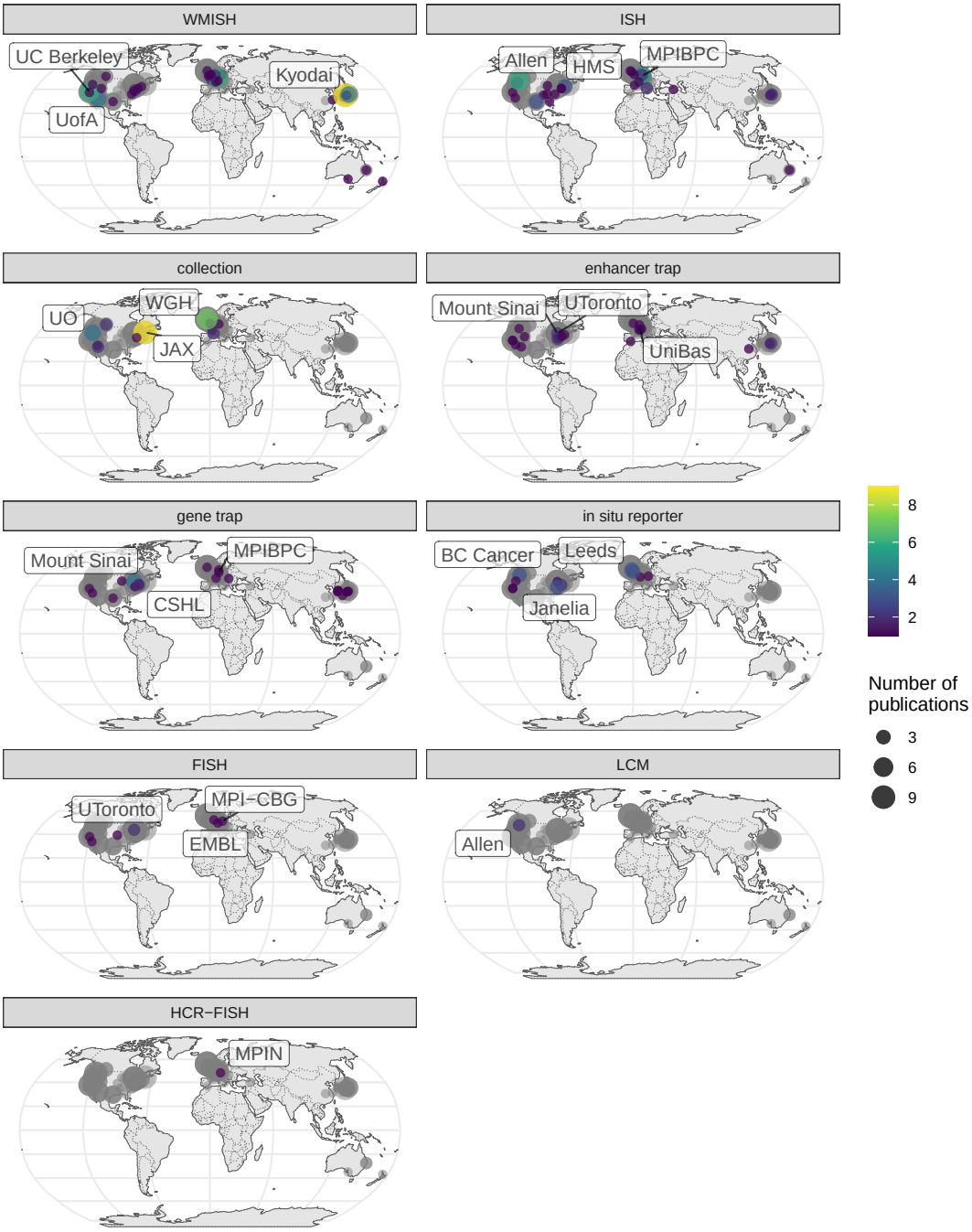

Figure 4.15: Number of prequel publications per city broken down by technique. Gray points are the overall number as a reference of contributions from each city and region.

7. Kitahara O, Furukawa Y, Tanaka T, Kihara C, Ono K, Yanagawa R, Nita ME, Takagi T, Nakamura Y, and Tsunoda T. Alterations of Gene Expression during Colorectal Carcinogenesis Revealed by cDNA Microarrays after Laser-Capture Microdissection of Tumor Tissues and Normal Epithelia. Cancer Research 2001 May; 61:3544 LP –3549. Available from: http://cancerres.aacrjournals.org/content/61/9/3544.abstract

8. Satija R, Farrell JA, Gennert D, Schier AF, and Regev A. Spatial reconstruction of single-cell gene expression data. Nature Biotechnology 2015 May; 33:495–502. DOI: 10.1038/nbt.3192. Available from: https://doi.org/10.1038/nbt.3192

9. Achim K, Pettit JB, Saraiva LR, Gavriouchkina D, Larsson T, Arendt D, and Marioni JC. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nature Biotechnology 2015 May; 33:503–9. DOI: 10.1038/nbt.3209. Available from: https://doi.org/10.1038/nbt.3209

10. Karaiskos N, Wahle P, Alles J, Boltengagen A, Ayoub S, Kipar C, Kocks C, Rajewsky N, and Zinzen RP. The Drosophila embryo at single-cell transcriptome resolution. Science 2017 Oct; 358:194–9. DOI: 10.1126/science.aan3235. Available from: https://science.sciencemag.org/content/358/6360/194

11. Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, Zwan J van der, Häring M, Braun E, Borm LE, La Manno G, Codeluppi S, Furlan A, Lee K, Skene N, Harris KD, Hjerling-Leffler J, Arenas E, Ernfors P, Marklund U, and Linnarsson S. Molecular Architecture of the Mouse Nervous System. Cell 2018 Aug; 174:999–1014. DOI: 10.1016/j.cell.2018.06.021. Available from: https://doi.org/10.1016/j.cell.2018.06.021

12. Ortiz C, Navarro JF, Jurek A, Märtin A, Lundeberg J, and Meletis K. Molecular atlas of the adult mouse brain. Science Advances 2020 Jun; 6:eabb3446. DOI: 10.1126/sciadv.abb3446. Available from: www.brain-map.org%20https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.abb3446

13. Chen WT, Lu A, Craessaerts K, Pavie B, Sala Frigerio C, Corthout N, Qian X, Laláková J, Kühnemund M, Voytyuk I, Wolfs L, Mancuso R, Salta E, Balusu S, Snellinx A, Munck S, Jurek A, Fernandez Navarro J, Saido TC, Huitinga I, Lundeberg J, Fiers M, and De Strooper B. Spatial Transcriptomics and In Situ Sequencing to Study Alzheimer's Disease. Cell 2020 Jul; 0. DOI: 10.1016/j.cell.2020.06.038. Available from: http://www.cell.com/article/S0092867420308151/fulltext

14. Jenett A, Rubin GM, Ngo TT, Shepherd D, Murphy C, Dionne H, Pfeiffer BD, Cavallaro A, Hall D, Jeter J, Iyer N, Fetter D, Hausenfluck JH, Peng H, Trautman ET, Svirskas RR, Myers EW, Iwinski ZR, Aso Y, DePasquale GM, Enos A, Hulamm P, Lam SCB, Li HH, Laverty TR, Long F, Qu L, Murphy SD, Rokicki K, Safford T, Shaw K, Simpson JH, Sowell A, Tae S, Yu Y, and

Zugates CT. A GAL4-Driver Line Resource for Drosophila Neurobiology. Cell Reports 2012 Oct; 2:991–1001. DOI: 10.1016/j.celrep.2012.09.011. Available from: http://dx.doi.org/10.1016/j.celrep.2012.09.011

15. Bravo González-Blas C, Quan XJ, Duran-Romaña R, Taskiran II, Koldere D, Davie K, Christiaens V, Makhzami S, Hulselmans G, Waegeneer M de, Mauduit D, Poovathingal S, Aibar S, and Aerts S. Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. Molecular Systems Biology 2020 May; 16:e9438. DOI: 10.15252/msb.20209438. Available from: https://doi.org/10.15252/msb.20209438

16. Bowes JB, Snyder KA, Segerdell E, Jarabek CJ, Azam K, Zorn AM, and Vize PD. Xenbase: Gene expression and improved integration. Nucleic Acids Research 2009 Oct; 38:D607–D612. DOI: 10.1093/nar/gkp953. Available from: https://academic.oup.com/nar/article/38/suppl_1/D607/3112302

17. XDB3. 2004. Available from: http://xenopus.nibb.ac.jp/

18. Satou Y, Takatori N, Yamada L, Mochizuki Y, Hamaguchi M, Ishikawa H, Chiba S, Imai K, Kano S, Murakami SD, Nakayama A, Nishino A, Sasakura Y, Satoh G, Shimotori T, Shin-i T, Shoguchi E, Suzuki MM, Takada N, Utsumi N, Yoshida N, Saiga H, Kohara Y, and Satoh N. Gene expression profiles in Ciona intestinalis tailbud embryos. Development 2001 Aug; 128:2893 LP –2904. Available from: http://dev.biologists.org/content/128/15/2893.abstract

19. Sprague J. The Zebrafish Information Network (ZFIN): the zebrafish model organism database. Nucleic Acids Research 2003 Jan; 31:241–3. DOI: 10.1093/nar/gkg027. Available from: http://zfin.org/zf_info/%20https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg027

20. Belmamoune M and Verbeek FJ. Data Integration for Spatio-Temporal Patterns of Gene Expression of Zebrafish development: the GEMS database. Journal of Integrative Bioinformatics 2008 Jun; 5:35–48. DOI: 10.1515/jib-2008-92. Available from: http://www.degruyter.com/view/j/jib.2008.5.issue-2/biecoll-jib-2008-92/biecoll-jib-2008-92.xml

21. Henrich T. MEPD: a Medaka gene expression pattern database. Nucleic Acids Research 2003 Jan; 31:72–4. DOI: 10.1093/nar/gkg017. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg017

22. Bell GW, Yatskievych TA, and Antin PB. GEISHA, a whole-mount in situ hybridization gene expression screen in chicken embryos. Developmental Dynamics 2004 Mar; 229:677–87. DOI: 10.1002/dvdy.10503. Available from: http://doi.wiley.com/10.1002/dvdy.10503

23. Lovell PV, Wirthlin M, Kaser T, Buckner AA, Carleton JB, Snider BR, McHugh AK, Tolpygo A, Mitra PP, and Mello CV. ZEBrA: Zebra finch Expression Brain Atlas—A resource for comparative molecular neuroanatomy and brain evolution studies. Journal of Comparative Neurology 2020 Aug; 528:2099–131. DOI: 10.1002/cne.24879. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.24879

24. Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu SQ, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, and Rubin GM. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome biology 2002 Dec; 3:research0088.1. DOI: 10.1186/gb-2002-3-12-research0088. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-12-research0088

25. Luengo Hendriks CL, Keränen SV, Fowlkes CC, Simirenko L, Weber GH, DePace AH, Henriquez C, Kaszuba DW, Hamann B, Eisen MB, Malik J, Sudar D, Biggin MD, and Knowles DW. Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution I: Data acquisition pipeline. Genome Biology 2006 Dec; 7:R123. DOI: 10.1186/gb-2006-7-12-r123. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2006-7-12-r123

26. Ardini-Poleske ME, Clark RF, Ansong C, Carson JP, Corley RA, Deutsch GH, Hagood JS, Kaminski N, Mariani TJ, Potter SS, Pryhuber GS, Warburton D, Whitsett JA, Palmer SM, and Ambalavanan N. LungMAP: The Molecular Atlas of Lung Development Program. American Journal of Physiology-Lung Cellular and Molecular Physiology 2017 Nov; 313:L733–L740. DOI: 10.1152/ajplung.00139.2017. Available from: https://www.physiology.org/doi/10.1152/ajplung.00139.2017

27. Blackshaw S, Harpavat S, Trimarchi J, Cai L, Huang H, Kuo WP, Weber G, Lee K, Fraioli RE, Cho SH, Yung R, Asch E, Ohno-Machado L, Wong WH, and Cepko CL. Genomic Analysis of Mouse Retinal Development. PLoS Biology 2004 Jun; 2. Ed. by William Harris:e247. DOI: 10.1371/journal.pbio.0020247. Available from: https://dx.plos.org/10.1371/journal.pbio.0020247

28. Harding SD, Armit C, Armstrong J, Brennan J, Cheng Y, Haggarty B, Houghton D, Lloyd-MacGilp S, Pi X, Roochun Y, Sharghi M, Tindal C, McMahon AP, Gottesman B, Little MH, Georgas K, Aronow BJ, Potter SS, Brunskill EW, Southard-Smith EM, Mendelsohn C, Baldock RA, Davies JA, and Davidson D. The GUDMAP database - an online resource for genitourinary research. Development 2011 Jul; 138:2845–53. DOI: 10.1242/

dev.063594. Available from: http://golgi.ana.%20http://dev.biologists.org/cgi/doi/10.1242/dev.063594

29. Ahmed A, Ward NJ, Moxon S, Lopez-Gomollon S, Viaut C, Tomlinson ML, Patrushev I, Gilchrist MJ, Dalmay T, Dotlic D, Münsterberg AE, and Wheeler GN. A Database of microRNA Expression Patterns in Xenopus laevis. PLOS ONE 2015 Oct; 10:e0138313. Available from: https://doi.org/10.1371/journal.pone.0138313

30. Karali M, Peluso I, Gennarino VA, Bilio M, Verde R, Lago G, Dollé P, and Banfi S. miRNeye: a microRNA expression atlas of the mouse eye. BMC Genomics 2010 Dec; 11:715. DOI: 10.1186/1471-2164-11-715. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-11-715

31. Diez-Roux G, Banfi S, Sultan M, Geffers L, Anand S, Rozado D, Magen A, Canidio E, Pagani M, Peluso I, Lin-Marq N, Koch M, Bilio M, Cantiello I, Verde R, De Masi C, Bianchi SA, Cicchini J, Perroud E, Mehmeti S, Dagand E, Schrinner S, Nürnberger A, Schmidt K, Metz K, Zwingmann C, Brieske N, Springer C, Hernandez AM, Herzog S, Grabbe F, Sieverding C, Fischer B, Schrader K, Brockmeyer M, Dettmer S, Helbig C, Alunni V, Battaini MA, Mura C, Henrichsen CN, Garcia-Lopez R, Echevarria D, Puelles E, Garcia-Calero E, Kruse S, Uhr M, Kauck C, Feng G, Milyaev N, Ong CK, Kumar L, Lam M, Semple CA, Gyenesei A, Mundlos S, Radelof U, Lehrach H, Sarmientos P, Reymond A, Davidson DR, Dollé P, Antonarakis SE, Yaspo ML, Martinez S, Baldock RA, Eichele G, and Ballabio A. A High-Resolution Anatomical Atlas of the Transcriptome in the Mouse Embryo. PLoS Biology 2011 Jan; 9. Ed. by Barsh GS:e1000582. DOI: 10.1371/journal.pbio.1000582. Available from: https://dx.plos.org/10.1371/journal.pbio.1000582

32. Aboobaker AA, Tomancak P, Patel N, Rubin GM, and Lai EC. Drosophila microRNAs exhibit diverse spatial expression patterns during embryonic development. Proceedings of the National Academy of Sciences 2005 Dec; 102:18017–22. DOI: 10.1073/pnas.0508823102. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0508823102

33. Wienholds E. MicroRNA Expression in Zebrafish Embryonic Development. Science 2005 Jul; 309:310–1. DOI: 10.1126/science.1114519. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1114519

34. Darnell DK, Kaur S, Stanislaw S, Konieczka JK, Yatskievych TA, and Antin PB. MicroRNA expression during chick embryo development. Developmental Dynamics 2006 Nov; 235:3156–65. DOI: 10.1002/dvdy.20956. Available from: http://doi.wiley.com/10.1002/dvdy.20956

35. Johnson AA, Hibberd JM, Gay C, Essah PA, Haseloff J, Tester M, and Guiderdoni E. Spatial control of transgene expression in rice (Oryza sativa L.) using the GAL4 enhancer trapping system. The Plant Journal 2005 Feb; 41:779–89. DOI: 10.1111/j.1365-313X.2005.02339.x. Available from: http://doi.wiley.com/10.1111/j.1365-313X.2005.02339.x

36. Nakayama N, Arroyo JM, Simorowski J, May B, Martienssen R, and Irish VF. Gene Trap Lines Define Domains of Gene Regulation in Arabidopsis Petals and Stamens. The Plant Cell 2005 Sep; 17:2486–506. DOI: 10.1105/tpc.105.033985. Available from: http://www.plantcell.org/lookup/doi/10.1105/tpc.105.033985

37. Lis JT, Simon JA, and Sutton CA. New heat shock puffs and β=galactosidase activity resulting from transformation of Drosophila with an hsp70-lacZ hybrid gene. Cell 1983 Dec; 35:403–10. DOI: 10.1016/0092-8674(83)90173-3. Available from: https://linkinghub.elsevier.com/retrieve/pii/0092867483901733

38. O'Kane CJ and Gehring WJ. Detection in situ of genomic regulatory elements in Drosophila. Proceedings of the National Academy of Sciences of the United States of America 1987 Dec; 84:9123–7. DOI: 10.1073/pnas.84.24.9123. Available from: https://www.pnas.org/content/84/24/9123

39. Spradling A and Rubin G. Transposition of cloned P elements into Drosophila germ line chromosomes. Science 1982 Oct; 218:341–7. DOI: 10.1126/science.6289435. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.6289435

40. Bellen HJ, O'Kane CJ, Wilson C, Grossniklaus U, Pearson RK, and Gehring WJ. P-element-mediated enhancer detection: a versatile method to study development in Drosophila. Genes & Development 1989 Sep; 3:1288–300. DOI: 10.1101/gad.3.9.1288. Available from: http://www.genesdev.org/cgi/doi/10.1101/gad.3.9.1288

41. Wilson C, Pearson RK, Bellen HJ, O'Kane CJ, Grossniklaus U, and Gehring WJ. P-element-mediated enhancer detection: an efficient method for isolating and characterizing developmentally regulated genes in Drosophila. Genes & Development 1989 Sep; 3:1301–13. DOI: 10.1101/gad.3.9.1301. Available from: http://www.genesdev.org/cgi/doi/10.1101/gad.3.9.1301

42. Gossler A, Joyner A, Rossant J, and Skarnes W. Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. Science 1989 Apr; 244:463–5. DOI: 10.1126/science.2497519. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.2497519

43. Stanford WL, Cohn JB, and Cordes SP. Gene-trap mutagenesis: past, present and beyond. Nature Reviews Genetics 2001 Oct; 2:756–68. DOI: 10.1038/35093548. Available from: www.nature.com/reviews/genetics%20http://www.nature.com/articles/35093548

44. Bier E, Vaessin H, Shepherd S, Lee K, McCall K, Barbel S, Ackerman L, Carretto R, Uemura T, and Grell E. Searching for pattern and mutation in the Drosophila genome with a P-lacZ vector. Genes & Development 1989 Sep; 3:1273–87. DOI: 10.1101/gad.3.9.1273. Available from: http://www.genesdev.org/cgi/doi/10.1101/gad.3.9.1273

45. Allen ND, Cran DG, Barton SC, Hettle S, Reik W, and Surani MA. Transgenes as probes for active chromosomal domains in mouse development. Nature 1988; 333:852–5. DOI: 10.1038/333852a0. Available from: https://www.nature.com/articles/333852a0

46. Sundaresan V, Springer P, Volpe T, Haward S, Jones JD, Dean C, Ma H, and Martienssen R. Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. Genes & Development 1995 Jul; 9:1797–810. DOI: 10.1101/gad.9.14.1797. Available from: http://www.genesdev.org/cgi/doi/10.1101/gad.9.14.1797

47. Friedrich G and Soriano P. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. Genes & Development 1991 Sep; 5:1513–23. DOI: 10.1101/gad.5.9.1513. Available from: http://www.genesdev.org/cgi/doi/10.1101/gad.5.9.1513

48. Wurst W, Rossant J, Prideaux V, Kownacka M, Joyner A, Hill DP, Guillemot F, Gasca S, Cado D, and Auerbach A. A large-scale gene-trap screen for insertional mutations in developmentally regulated genes in mice. Genetics 1995; 139. Available from: https://www.genetics.org/content/139/2/889

49. Skarnes WC, Moss JE, Hurtley SM, and Beddington RS. Capturing genes encoding membrane and secreted proteins important for mouse development. Proceedings of the National Academy of Sciences 1995 Jul; 92:6592–6. DOI: 10.1073/pnas.92.14.6592. Available from: https://www.pnas.org/content/92/14/6592

50. Forrester LM, Nagy A, Sam M, Watt A, Stevenson L, Bernstein A, Joyner AL, and Wurst W. An induction gene trap screen in embryonic stem cells: Identification of genes that respond to retinoic acid in vitro. Proceedings of the National Academy of Sciences of the United States of America 1996 Feb; 93:1677–82. DOI: 10.1073/pnas.93.4.1677. Available from: https://www.pnas.org/content/93/4/1677

51. Stanford WL, Caruana G, Vallis KA, Inamdar M, Hidaka M, Bautch VL, and Bernstein A. Expression Trapping: Identification of Novel Genes Expressed in Hematopoietic and Endothelial Lineages by Gene Trapping in ES Cells.

Blood 1998 Dec; 92:4622–31. DOI: 10.1182/blood.V92.12.4622. Available from: https://ashpublications.org/blood/article/92/12/4622/245124/Expression-Trapping-Identification-of-Novel-Genes

52. Leighton PA, Mitchell KJ, Goodrich LV, Lu X, Pinson K, Scherz P, Skarnes WC, and Tessier-Lavigne M. Defining brain wiring patterns and mechanisms through gene trapping in mice. Nature 2001 Mar; 410:174–9. DOI: 10.1038/35065539. Available from: http://www.nature.com/articles/35065539

53. Giani AM, Gallo GR, Gianfranceschi L, and Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. Computational and Structural Biotechnology Journal 2020 Jan; 18:9–19. DOI: 10.1016/j.csbj.2019.11.002. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2001037019303277

54. Frohman MA, Dush MK, and Martin GR. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proceedings of the National Academy of Sciences 1988 Dec; 85:8998–9002. DOI: 10.1073/pnas.85.23.8998. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.85.23.8998

55. Myers EW. A Whole-Genome Assembly of Drosophila. Science 2000 Mar; 287:2196–204. DOI: 10.1126/science.287.5461.2196. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.287.5461.2196

56. Waterston RH et al. Initial sequencing and comparative analysis of the mouse genome. Nature 2002 Dec; 420:520–62. DOI: 10.1038/nature01262. Available from: http://www.nature.com/articles/nature01262

57. Lander ES et al. Initial sequencing and analysis of the human genome. Nature 2001 Feb; 409:860–921. DOI: 10.1038/35057062. Available from: http://www.nature.com/articles/35057062

58. Venter JC et al. The Sequence of the Human Genome. Science 2001 Feb; 291:1304–51. DOI: 10.1126/science.1058040. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1058040

59. Pérez-Martín F, Yuste-Lisbona FJ, Pineda B, Angarita-Díaz MP, García-Sogo B, Antón T, Sánchez S, Giménez E, Atarés A, Fernández-Lozano A, Ortíz-Atienza A, García-Alcázar M, Castañeda L, Fonseca R, Capel C, Goergen G, Sánchez J, Quispe JL, Capel J, Angosto T, Moreno V, and Lozano R. A collection of enhancer trap insertional mutants for functional genomics in tomato. Plant Biotechnology Journal 2017 Nov; 15:1439–52. DOI: 10.1111/pbi.12728. Available from: http://doi.wiley.com/10.1111/pbi.12728

60. Hiwatashi Y, Nishiyama T, Fujita T, and Hasebe M. Establishment of gene-trap and enhancer-trap systems in the moss Physcomitrella patens. The Plant Journal 2001 Dec; 28:105–16. DOI: 10.1046/j.1365-313X.2001.01121.x. Available from: http://doi.wiley.com/10.1046/j.1365-313X.2001.01121.x

61. Kawakami K, Abe G, Asada T, Asakawa K, Fukuda R, Ito A, Lal P, Mouri N, Muto A, Suster ML, Takakubo H, Urasaki A, Wada H, and Yoshida M. ZTrap: Zebrafish gene trap and enhancer trap database. BMC Developmental Biology 2010 Oct; 10:105. DOI: 10.1186/1471-213X-10-105. Available from: http://bmcdevbiol.biomedcentral.com/articles/10.1186/1471-213X-10-105

62. Marquart GD, Tabor KM, Brown M, Strykowski JL, Varshney GK, LaFave MC, Mueller T, Burgess SM, Higashijima Si, and Burgess HA. A 3D Searchable Database of Transgenic Zebrafish Gal4 and Cre Lines for Functional Neuroanatomy Studies. Frontiers in Neural Circuits 2015 Nov; 9:1–17. DOI: 10.3389/fncir.2015.00078. Available from: http://journal.frontiersin.org/Article/10.3389/fncir.2015.00078/abstract

63. Hope IA. 'Promoter trapping' in Caenorhabditis elegans. Tech. rep. 1991 :399–408. Available from: https://dev.biologists.org/content/develop/113/2/399.full.pdf

64. Lynch AS, Briggs D, and Hope IA. Developmental expression pattern screen for genes predicted in the C. elegans genome sequencing project. Nature Genetics 1995 Nov; 11:309–13. DOI: 10.1038/ng1195-309. Available from: http://www.nature.com/articles/ng1195-309

65. Sulston J, Du Z, Thomas K, Wilson R, Hillier L, Staden R, Halloran N, Green P, Thierry-Mieg J, Qiu L, Dear S, Coulson A, Craxton M, Durbin R, Berks M, Metzstein M, Hawkins T, Ainscough R, and Waterston R. The C. elegans genome sequencing project: A beginning. Nature 1992; 356:37–41. DOI: 10.1038/356037a0. Available from: https://www.nature.com/articles/356037a0

66. The C. elegans Sequencing Consortium. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science 1998 Dec; 282:2012–8. DOI: 10.1126/science.282.5396.2012. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.282.5396.2012

67. Yang XW, Model P, and Heintz N. Homologous recombination based modification in Escherichia coli and germline transmission in transgenic mice of a bacterial artificial chromsome. Nature Biotechnology 1997 Sep; 15:859–65. DOI: 10.1038/nbt0997-859. Available from: http://www.nature.com/articles/nbt0997-859

68. Gong S, Zheng C, Doughty ML, Losos K, Didkovsky N, Schambra UB, Nowak NJ, Joyner A, Leblanc G, Hatten ME, and Heintz N. A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. Nature 2003 Oct; 425:917–25. DOI: 10.1038/nature02033. Available from: http://www.nature.com/articles/nature02033

69. Siegert S, Scherf BG, Del Punta K, Didkovsky N, Heintz N, and Roska B. Genetic address book for retinal cell types. Nature Neuroscience 2009 Sep; 12:1197–204. DOI: 10.1038/nn.2370. Available from: http://www.nature.com/articles/nn.2370

70. Skarnes WC, Rosen B, West AP, Koutsourakis M, Bushell W, Iyer V, Mujica AO, Thomas M, Harrow J, Cox T, Jackson D, Severin J, Biggs P, Fu J, Nefedov M, Jong PJ de, Stewart AF, and Bradley A. A conditional knockout resource for the genome-wide study of mouse gene function. Nature 2011 Jun; 474:337–42. DOI: 10.1038/nature10163. Available from: http://www.nature.com/articles/nature10163

71. A Mouse for All Reasons. Cell 2007 Jan; 128:9–13. DOI: 10.1016/j.cell.2006.12.018. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867406016126

72. White JK et al. Genome-wide Generation and Systematic Phenotyping of Knockout Mice Reveals New Roles for Many Genes. Cell 2013 Jul; 154:452–64. DOI: 10.1016/j.cell.2013.06.022. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867413007617

73. West DB, Pasumarthi RK, Baridon B, Djan E, Trainor A, Griffey SM, Engelhard EK, Rapp J, Li B, Jong PJd, and Lloyd KK. A lacZ reporter gene expression atlas for 313 adult KOMP mutant mouse lines. Genome Research 2015 Apr; 25:598–607. DOI: 10.1101/gr.184184.114. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.184184.114

74. Tuck E, Estabel J, Oellrich A, Maguire AK, Adissu HA, Souter L, Siragher E, Lillistone C, Green AL, Wardle-Jones H, Carragher DM, Karp NA, Smedley D, Adams NC, Bussell JN, Adams DJ, Ramirez-Solis R, Steel KP, Galli A, and White JK. A gene expression resource generated by genome-wide lacZ profiling in the mouse. Disease Models & Mechanisms 2015 Nov; 8:1467–78. DOI: 10.1242/dmm.021238. Available from: http://dmm.biologists.org/cgi/doi/10.1242/dmm.021238

75. Visel A, Taher L, Girgis H, May D, Golonzhka O, Hoch RV, McKinsey GL, Pattabiraman K, Silberberg SN, Blow MJ, Hansen DV, Nord AS, Akiyama JA, Holt A, Hosseini R, Phouanenavong S, Plajzer-Frick I, Shoukry M, Afzal V, Kaplan T, Kriegstein AR, Rubin EM, Ovcharenko I, Pennacchio LA, and Rubenstein JL. A high-resolution enhancer atlas of the developing telencephalon. Cell 2013 Feb; 152:895–908. DOI: 10.1016/j.cell.2012.12.041. Available from: http://dx.doi.org/10.1016/j.cell.2012.12.041

76. Kvon EZ, Kazmar T, Stampfel G, Yáñez-Cuna JO, Pagani M, Schernhuber K, Dickson BJ, and Stark A. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature 2014 Aug; 512:91–5. DOI: 10.1038/nature13395. Available from: http://www.nature.com/articles/nature13395

77. Hunt-Newbury R, Viveiros R, Johnsen R, Mah A, Anastas D, Fang L, Halfnight E, Lee D, Lin J, Lorch A, McKay S, Okada HM, Pan J, Schulz AK, Tu D, Wong K, Zhao Z, Alexeyenko A, Burglin T, Sonnhammer E, Schnabel R, Jones SJ, Marra MA, Baillie DL, and Moerman DG. High-Throughput In Vivo Analysis of Gene Expression in Caenorhabditis elegans. PLoS Biology 2007 Sep; 5. Ed. by Sulston J:e237. DOI: 10.1371/journal.pbio.0050237. Available from: https://dx.plos.org/10.1371/journal.pbio.0050237

78. Gall JG, Lou M, Kline P, and Giles NH. Formation and Detecction of RNA-DNA Hybrid Molecules in Cytological Preparations. PNAS 1969; 63:378–83. DOI: 10.1073/pnas.63.2.378. Available from: https://www.pnas.org/content/63/2/378

79. John HA, Birnstiel ML, and Jones KW. RNA-DNA hybrids at the cytological level. Nature 1969; 223:582–7. DOI: 10.1038/223582a0. Available from: https://www.nature.com/articles/223582a0

80. Harrison P, Conkie D, Paul J, and Jones K. Localisation of cellular globin messenger RNA by in situ hybridisation to complementary DNA. FEBS Letters 1973 May; 32:109–12. DOI: 10.1016/0014-5793(73)80749-5. Available from: https://doi.org/10.1016/0014-5793(73)80749-5

81. Huber D, Voith von Voithenberg L, and Kaigala G. Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? Micro and Nano Engineering 2018 Nov; 1:15–24. DOI: 10.1016/j.mne.2018.10.006. Available from: https://linkinghub.elsevier.com/retrieve/pii/S259000721830008X

82. Langer-Safer PR, Levine M, and Ward DC. Immunological method for mapping genes on Drosophila polytene chromosomes. Proceedings of the National Academy of Sciences 1982 Jul; 79:4381–5. DOI: 10.1073/pnas.79.14.4381. Available from: https://www.pnas.org/content/79/14/4381

83. Singer RH and Ward DC. Actin gene expression visualized in chicken muscle tissue culture by using in situ hybridization with a biotinated nucleotide analog. Proceedings of the National Academy of Sciences of the United States of America 1982 Dec; 79:7331–5. DOI: 10.1073/pnas.79.23.7331. Available from: http://www.pnas.org/content/79/23/7331.abstract

84. Tautz D and Pfeifle C. A non-radioactive in situ hybridization method for the localization of specific RNAs in Drosophila embryos reveals translational control of the segmentation gene hunchback. Chromosoma 1989 Aug; 98:81–5. DOI: 10.1007/BF00291041. Available from: https://link.springer.com/article/10.1007/BF00291041

85. Rosen B and Beddington RS. Whole-mount in situ hybridization in the mouse embryo: gene expression in three dimensions. Trends in Genetics 1993 May; 9:162–7. DOI: 10.1016/0168-9525(93)90162-B. Available from: https://linkinghub.elsevier.com/retrieve/pii/016895259390162B

86. Seydoux G and Fire A. Soma-germline asymmetry in the distributions of embryonic RNAs in Caenorhabditis elegans. Development 1994 Oct; 120:2823 LP –2834. Available from: http://dev.biologists.org/content/120/10/2823.abstract

87. Stapleton M. The Drosophila Gene Collection: Identification of Putative Full-Length cDNAs for 70% of D. melanogaster Genes. Genome Research 2002 Aug; 12:1294–300. DOI: 10.1101/gr.269102. Available from: http://www.genome.org/cgi/doi/10.1101/gr.269102

88. Gawantka V, Pollet N, Delius H, Vingron M, Pfister R, Nitsch R, Blumenstock C, and Niehrs C. Gene expression screening in Xenopus identifies molecular pathways, predicts gene function and provides a global view of embryonic patterning. Mechanisms of Development 1998 Sep; 77:95–141. DOI: 10.1016/S0925-4773(98)00115-4. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0925477398001154

89. Bettenhausen B and Gossler A. Efficient Isolation of Novel Mouse Genes Differentially Expressed in Early Postimplantation Embryos. Genomics 1995 Aug; 28:436–41. DOI: 10.1006/geno.1995.1172. Available from: https://linkinghub.elsevier.com/retrieve/pii/S088875438571172X

90. Carson JP, Thaller C, and Eichele G. A transcriptome atlas of the mouse brain at cellular resolution. Current Opinion in Neurobiology 2002 Oct; 12:562–5. DOI: 10.1016/S0959-4388(02)00356-2. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0959438802003562

91. Kopczynski CC, Noordermeer JN, Serano TL, Chen WY, Pendleton JD, Lewis S, Goodman CS, and Rubin GM. A high throughput screen to identify secreted and transmembrane proteins involved in Drosophila embryogenesis. Proceedings of the National Academy of Sciences 1998 Aug; 95:9973–8. DOI: 10.1073/pnas.95.17.9973. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.95.17.9973

92. Birchall PS, Fishpool RM, and Albertson DG. Expression patterns of predicted genes from the C. elegans genome sequence visualized by FISH in whole organisms. 1995 :314–20. DOI: 10.1038/ng1195-314. Available from: https://doi.org/10.1038/ng1195-314

93. Carter MG. In Situ-Synthesized Novel Microarray Optimized for Mouse Stem Cell and Early Developmental Expression Profiling. Genome Research 2003 May; 13:1011–21. DOI: 10.1101/gr.878903. Available from: http://www.genome.org/cgi/doi/10.1101/gr.878903

94. Yoshikawa T, Piao Y, Zhong J, Matoba R, Carter MG, Wang Y, Goldberg I, and Ko MS. High-throughput screen for genes predominantly expressed in the ICM of mouse blastocysts by whole mount in situ hybridization. Gene Expression Patterns 2006 Jan; 6:213–24. DOI: 10.1016/j.modgep.2005.06.003. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1567133X05000682

95. Lein ES. Defining a Molecular Atlas of the Hippocampus Using DNA Microarrays and High-Throughput In Situ Hybridization. Journal of Neuroscience 2004 Apr; 24:3879–89. DOI: 10.1523/JNEUROSCI.4710-03.2004. Available from: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4710-03.2004

96. Lein ES et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature 2007 Jan; 445:168–76. DOI: 10.1038/nature05453. Available from: http://www.nature.com/articles/nature05453

97. Valoczi A. Sensitive and specific detection of microRNAs by northern blot analysis using LNA-modified oligonucleotide probes. Nucleic Acids Research 2004 Dec; 32:e175–e175. DOI: 10.1093/nar/gnh171. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gnh171

98. Kloosterman WP, Wienholds E, Bruijn E de, Kauppinen S, and Plasterk RHA. In situ detection of miRNAs in animal embryos using LNA-modified oligonucleotide probes. Nature Methods 2006 Jan; 3:27–9. DOI: 10.1038/nmeth843. Available from: http://www.nature.com/articles/nmeth843

99. McKee AE, Minet E, Stern C, Riahi S, Stiles CD, and Silver PA. A genome-wide in situ hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain. BMC Developmental Biology 2005 Jul; 5:14. DOI: 10.1186/1471-213X-5-14. Available from: http://bmcdevbiol.biomedcentral.com/articles/10.1186/1471-213X-5-14

100. Yaylaoglu MB, Titmus A, Visel A, Alvarez-Bolado G, Thaller C, and Eichele G. Comprehensive expression atlas of fibroblast growth factors and their receptors generated by a novel robotic in situ hybridization platform. Developmental Dynamics 2005 Oct; 234:371–86. DOI: 10.1002/dvdy.20441. Available from: http://doi.wiley.com/10.1002/dvdy.20441

101. Cankaya M, Hernandez A, Ciftci M, Beydemir S, Ozdemir H, Budak H, Gulcin I, Comakli V, Emircupani T, Ekinci D, Kuzu M, Jiang Q, Eichele G, and Kufrevioglu O. An analysis of expression patterns of genes encoding proteins with catalytic activities. BMC Genomics 2007 Jul; 8:232. DOI: 10.1186/1471-2164-8-232. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-8-232

102. Yokoyama S, Ito Y, Ueno-Kudoh H, Shimizu H, Uchibe K, Albini S, Mitsuoka K, Miyaki S, Kiso M, Nagai A, Hikata T, Osada T, Fukuda N, Yamashita S, Harada D, Mezzano V, Kasai M, Puri PL, Hayashizaki Y, Okado H, Hashimoto M, and Asahara H. A Systems Approach Reveals that the Myogenesis Genome Network Is Regulated by the Transcriptional Repressor RP58. Developmental Cell 2009 Dec; 17:836–48. DOI: 10.1016/j.devcel.2009.10.011. Available from: https://linkinghub.elsevier.com/retrieve/pii/S153458070900433X

103. Geffers L, Tetzlaff B, Cui X, Yan J, and Eichele G. METscout: a pathfinder exploring the landscape of metabolites, enzymes and transporters. Nucleic Acids Research 2012 Sep; 41:D1047–D1054. DOI: 10.1093/nar/gks886. Available from: http://academic.oup.com/nar/article/41/D1/D1047/1073820/METscout-a-pathfinder-exploring-the-landscape-of

104. Şişecioğlu M, Budak H, Geffers L, Çankaya M, Çiftci M, Thaller C, Eichele G, Küfrevioğlu Öİ, and Özdemir H. A compendium of expression patterns of cholesterol biosynthetic enzymes in the mouse embryo. Journal of Lipid Research 2015 Aug; 56:1551–9. DOI: 10.1194/jlr.M059634. Available from: http://www.jlr.org/lookup/doi/10.1194/jlr.M059634

105. Shcherbatyy V, Carson J, Yaylaoglu M, Jäckle K, Grabbe F, Brockmeyer M, Yavuz H, and Eichele G. A Digital Atlas of Ion Channel Expression Patterns in the Two-Week-Old Rat Brain. Neuroinformatics 2015 Jan; 13:111–25. DOI: 10.1007/s12021-014-9247-0. Available from: http://www.genepaint.org/%20http://link.springer.com/10.1007/s12021-014-9247-0

106. Thut CJ, Rountree RB, Hwa M, and Kingsley DM. A Large-Scale in Situ Screen Provides Molecular Evidence for the Induction of Eye Anterior Segment Structures by the Developing Lens. Developmental Biology 2001 Mar; 231:63–76. DOI: 10.1006/dbio.2000.0140. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0012160600901404

107. Fowlkes CC, Hendriks CLL, Keränen SV, Weber GH, Rübel O, Huang MY, Chatoor S, DePace AH, Simirenko L, Henriquez C, Beaton A, Weiszmann R, Celniker S, Hamann B, Knowles DW, Biggin MD, Eisen MB, and Malik J. A Quantitative Spatiotemporal Atlas of Gene Expression in the Drosophila Blastoderm. Cell 2008 Apr; 133:364–74. DOI: 10.1016/j.cell.2008.01.

053. Available from: https://linkinghub.elsevier.com/retrieve/pii/S009286740800281X

108. Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, Hughes TR, Tomancak P, and Krause HM. Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function. Cell 2007 Oct; 131:174–87. DOI: 10.1016/j.cell.2007.08.003. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867407010227

109. Boer BA de, Ruijter JM, Voorbraak FPJM, and Moorman AFM. More than a decade of developmental gene expression atlases: where are we now? Nucleic Acids Research 2009 Dec; 37:7349–59. DOI: 10.1093/nar/gkp819. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp819

110. Clarkson MD. Representation of anatomy in online atlases and databases: a survey and collection of patterns for interface design. BMC Developmental Biology 2016 Dec; 16:18. DOI: 10.1186/s12861-016-0116-y. Available from: http://bmcdevbiol.biomedcentral.com/articles/10.1186/s12861-016-0116-y

111. Baldock R, Bard J, Kaufman M, and Davidson D. What's New? A real mouse for your computer. BioEssays 1992 Jul; 14:501–2. DOI: 10.1002/bies.950140713. Available from: http://doi.wiley.com/10.1002/bies.950140713

112. Ringwald M, Baldock R, Bard J, Kaufman M, Eppig JT, Richardson JE, Nadeau JH, and Davidson D. A database for mouse development. Science 1994 Sep; 265:2033–4. DOI: 10.1126/science.8091224. Available from: https://science.sciencemag.org/content/265/5181/2033.abstract

113. Ringwald M, Davis GL, Smith AG, Trepanier LE, Begley DA, Richardson JE, and Eppig JT. The mouse gene expression database GXD. Seminars in Cell & Developmental Biology 1997; 8:489–97. DOI: https://doi.org/10.1006/scdb.1997.0177. Available from: http://www.sciencedirect.com/science/article/pii/S1084952197901774

114. Janning W. FlyView, aDrosophilaimage database, and otherDrosophiladatabases. Seminars in Cell & Developmental Biology 1997; 8:469–75. DOI: https://doi.org/10.1006/scdb.1997.0172. Available from: http://www.sciencedirect.com/science/article/pii/S1084952197901725

115. Martinelli SD, Brown CG, and Durbin R. Gene expression and development databases forC. elegans. Seminars in Cell & Developmental Biology 1997 Oct; 8:459–67. DOI: 10.1006/scdb.1997.0171. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1084952197901713

116. Westerfield M, Doerry E, Kirkpatrick AE, Driever W, and Douglas SA. An on-line database for zebrafish development and genetics research. Seminars in Cell & Developmental Biology 1997 Oct; 8:477–88. DOI: 10.1006/scdb.1997.0173. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1084952197901737

117. Howe DG, Bradford YM, Eagle A, Fashena D, Frazer K, Kalita P, Mani P, Martin R, Moxon ST, Paddock H, Pich C, Ramachandran S, Ruzicka L, Schaper K, Shao X, Singer A, Toro S, Van Slyke C, and Westerfield M. The Zebrafish Model Organism Database: New support for human disease models, mutation details, gene expression phenotypes and searching. Nucleic Acids Research 2017 Jan; 45:D758–D768. DOI: 10.1093/nar/gkw1116. Available from: https://academic.oup.com/nar/article/45/D1/D758/2605740

118. Kumar S, Konikoff C, Sanderford M, Liu L, Newfeld S, Ye J, and Kulathinal RJ. FlyExpress 7: An Integrated Discovery Platform To Study Co-expressed Genes Using in situ Hybridization Images in Drosophila. G3; Genes|Genomes|Genetics 2017 Aug; 7:2791–7. DOI: 10.1534/g3.117.040345. Available from: http://g3journal.org/lookup/doi/10.1534/g3.117.040345

119. Gilchrist MJ, Christensen MB, Bronchain O, Brunet F, Chesneau A, Fenger U, Geach TJ, Ironfield HV, Kaya F, Kricha S, Lea R, Massé K, Néant I, Paillard E, Parain K, Perron M, Sinzelle L, Souopgui J, Thuret R, Ymlahi-Ouazzani Q, and Pollet N. Database of queryable gene expression patterns for Xenopus. Developmental Dynamics 2009 Jun; 238:1379–88. DOI: 10.1002/dvdy.21940. Available from: http://doi.wiley.com/10.1002/dvdy.21940

120. Tassy O, Dauga D, Daian F, Sobral D, Robin F, Khoueiry P, Salgado D, Fox V, Caillol D, Schiappa R, Laporte B, Rios A, Luxardi G, Kusakabe T, Joly JS, Darras S, Christiaen L, Contensin M, Auger H, Lamy C, Hudson C, Rothbacher U, Gilchrist MJ, Makabe KW, Hotta K, Fujiwara S, Satoh N, Satou Y, and Lemaire P. The ANISEED database: Digital representation, formalization, and elucidation of a chordate developmental program. Genome Research 2010 Oct; 20:1459–68. DOI: 10.1101/gr.108175.110. Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.108175.110

121. Ringwald M, Mangan ME, Eppig JT, Kadin JA, and Richardson JE. GXD: a Gene Expression Database for the laboratory mouse. Nucleic Acids Research 1999 Jan; 27:106–12. DOI: 10.1093/nar/27.1.106. Available from: https://doi.org/10.1093/nar/27.1.106

122. Kawashima T. MAGEST: MAboya Gene Expression patterns and Sequence Tags. Nucleic Acids Research 2000 Jan; 28:133–5. DOI: 10.1093/nar/28.1.133. Available from: http://www.genome.ad.jp/magest/%20https:

//academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.133

123. Salgado D, Gimenez G, Coulier F, and Marcelle C. COMPARE, a multi-organism system for cross-species data comparison and transfer of information. Bioinformatics 2008 Feb; 24:447–9. DOI: 10.1093/bioinformatics/btm599. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btm599

124. Haudry Y, Berube H, Letunic I, Weeber PD, Gagneur J, Girardot C, Kapushesky M, Arendt D, Bork P, Brazma A, Furlong EEM, Wittbrodt J, and Henrich T. 4DXpress: a database for cross-species expression pattern comparisons. Nucleic Acids Research 2007 Dec; 36:D847–D853. DOI: 10.1093/nar/gkm797. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkm797

125. Agapite J, Albou LP, Aleksander S, Argasinska J, Arnaboldi V, Attrill H, Bello SM, Blake JA, Blodgett O, Bradford YM, Bult CJ, Cain S, Calvi BR, Carbon S, Chan J, Chen WJ, Cherry JM, Cho J, Christie KR, Crosby MA, Pons JD, Dolan ME, Santos GD, Dunn B, Dunn N, Eagle A, Ebert D, Engel SR, Fashena D, Frazer K, Gao S, Gondwe F, Goodman J, Gramates LS, Grove CA, Harris T, Harrison MC, Howe DG, Howe KL, Jha S, Kadin JA, Kaufman TC, Kalita P, Karra K, Kishore R, Laulederkind S, Lee R, MacPherson KA, Marygold SJ, Matthews B, Millburn G, Miyasato S, Moxon S, Mueller HM, Mungall C, Muruganujan A, Mushayahama T, Nash RS, Ng P, Paulini M, Perrimon N, Pich C, Raciti D, Richardson JE, Russell M, Gelbart SR, Ruzicka L, Schaper K, Shimoyama M, Simison M, Smith C, Shaw DR, Shrivatsav A, Skrzypek M, Smith JR, Sternberg PW, Tabone CJ, Thomas PD, Thota J, Toro S, Tomczuk M, Tutaj M, Tutaj M, Urbano JM, Auken KV, Slyke CEV, Wang SJ, Weng S, Westerfield M, Williams G, Wong ED, Wright A, and Yook K. Alliance of Genome Resources Portal: unified model organism research platform. Nucleic Acids Research 2020 Jan; 48:D650–D658. DOI: 10.1093/nar/gkz813. Available from: https://academic.oup.com/nar/article/48/D1/D650/5573549

126. Thompson C, Wisor J, Lee CK, Pathak S, Gerashchenko D, Smith K, Fischer S, Kuan C, Sunkin S, Ng L, Lau C, Hawrylycz M, Jones A, Kilduff T, and Lein E. Molecular and Anatomical Signatures of Sleep Deprivation in the Mouse Brain. 2010. Available from: https://www.frontiersin.org/article/10.3389/fnins.2010.00165

127. Puchalski RB, Shah N, Miller J, Dalley R, Nomura SR, Yoon JG, Smith KA, Lankerovich M, Bertagnolli D, Bickley K, Boe AF, Brouner K, Butler S, Caldejon S, Chapin M, Datta S, Dee N, Desta T, Dolbeare T, Dotson N, Ebbert A, Feng D, Feng X, Fisher M, Gee G, Goldy J, Gourley L, Gregor BW, Gu G, Hejazinia N, Hohmann J, Hothi P, Howard R, Joines K, Kriedberg A, Kuan L, Lau C, Lee F, Lee H, Lemon T, Long F, Mastan N, Mott E, Murthy C, Ngo K, Olson E, Reding M, Riley Z, Rosen D, Sandman D,

Shapovalova N, Slaughterbeck CR, Sodt A, Stockdale G, Szafer A, Wakeman W, Wohnoutka PE, White SJ, Marsh D, Rostomily RC, Ng L, Dang C, Jones A, Keogh B, Gittleman HR, Barnholtz-Sloan JS, Cimino PJ, Uppin MS, Keene CD, Farrokhi FR, Lathia JD, Berens ME, Iavarone A, Bernard A, Lein E, Phillips JW, Rostad SW, Cobbs C, Hawrylycz MJ, and Foltz GD. An anatomic transcriptional atlas of human glioblastoma. Science 2018 May; 360:660 LP –663. DOI: 10.1126/science.aaf2666. Available from: http://science.sciencemag.org/content/360/6389/660.abstract

Chapter 5

DATA ANALYSIS IN THE PREQUEL ERA

From the earliest days of enhancer and gene traps to the (WM)ISH atlases, identifying genes with spatially and temporally variable expression patterns, comparing and classifying the patterns, identifying new marker genes of cell types and developmental stages, and using gene expression to redefine cell types have been among the goals of the studies [1, 2, 3, 4, 5, 6, 7]. In the prequel era, these were typically done manually, which, with the growing size of atlases in the 2000s, was time consuming and potentially inconsistent between curators. Thus, computational methods were developed to analyze images from the (WM)ISH atlases. This chapter reviews data analysis methods designed for (WM)ISH atlases and does not involve scRNA-seq data; methods involving both (WM)ISH and scRNA-seq are reviewed in Chapter 9 for the current era because scRNA-seq is at present a popular and rapidly growing field, too in vogue to be considered "prequel". If our collection is representative, then the rise of prequel data analysis methods arrived much later than that of data collection (Figure 5.1).

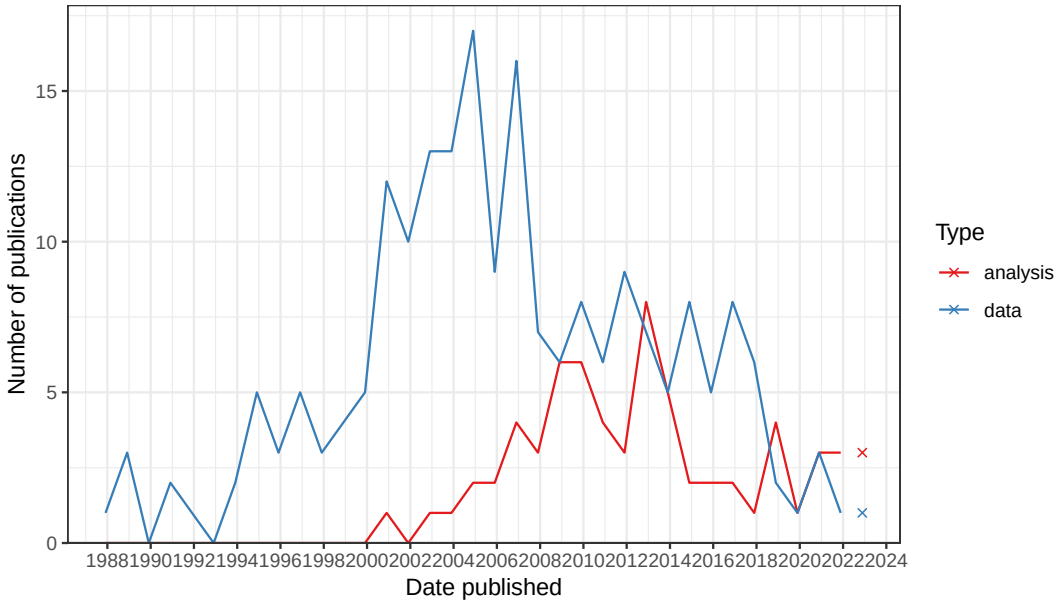

Figure 5.1: Comparing trends in data collection and data analysis in the prequel era. Bin width is 365 days. The x-shaped points show the number of publications from the last bin, which is not yet full.

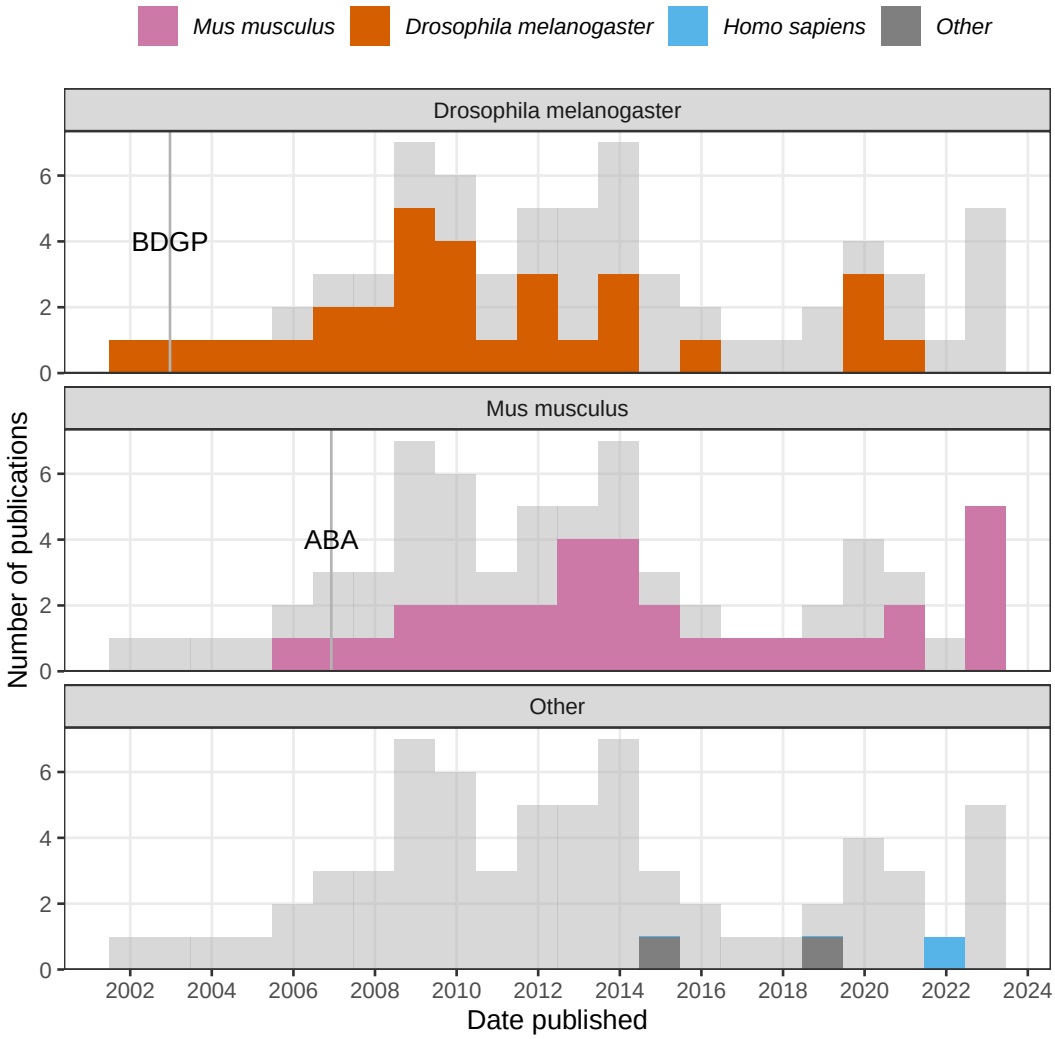

Figure 5.2: Gray histogram in the background is overall histogram of prequel data analysis literature. Number of publications in each time bin for each species is highlighted in the facets.

Except for one study on *Platynereis dumereilii* in 2014 [8], on *Xenopus tropicalis* in 2018 [9], one on post mortem human brain in 2021 [10], all data analysis methods in our collection were designed for either *Drosophila melanogaster* or *Mus musculus* (Figure 5.2). There seem to have been two waves; the first for *Drosophila*, peaking in the late 2000s, mostly concerning the BDGP in situ atlas, and the second for mice, peaking in early 2010s, mostly concerning ABA (Figure 5.2). The apparent rise since 2019 is in part driven by deep learning methods to annotate gene expression patterns or infer gene interactions. Given the small number of publications in this category and potential incompleteness of the curation, the trends should be taken with a grain of salt.

5.1 Gene patterns

The most common goal of these data analysis methods was to annotate and compare gene expression patterns, especially to automate annotation of the BDGP atlas (Figure 5.3). It seems reasonable to focus on 4 phases in this category: first, in early to mid 2000s, after image registration, the images were binarized into "expressed" and "not expressed" regions, and the shapes of the expressed regions were summarized and compared. Metrics to summarize the shapes included moment invariant [11, 12], Hamming distance [13], and a weighted score involving L1 distance between column or row histograms of two images [14]. These unsupervised methods enabled clustering of patterns and querying genes with similar patterns to a given gene.

Second, from the mid 2000s to mid 2010s, many supervised and unsupervised methods for gene expression pattern annotation or comparison were developed. In supervised methods, extensive feature engineering more sophisticated than binarization was performed on registered images for image annotation with machine learning classification. These methods were trained with existing BDGP annotations and developed to automatically annotate the BDGP expression patterns with controlled vocabulary (CV) of anatomical regions where genes were expressed. In BDGP, a gene gets annotated with a CV if the gene was deemed expressed in the anatomical region and developmental stage denoted by the CV, so the annotation typically contained a list of CVs.

The feature engineering can be based on the wavelet transform [15] and Fourier coefficients [16], but a particularly popular feature engineering method was scale-invariant feature transform (SIFT) [17, 18, 19, 20]. A method published in 2009 that used SIFT followed by bag of words where "word" is a k means cluster (code book) was quite influential [20]; several later methods were inspired by this method, with improved code books [21, 22, 23, 24]. The most common classifier that take in the features to predict annotations is support vector machine (SVM) [21, 23] or multi-label variants of it [18, 20].

Unsupervised methods rely on clustering algorithms after images are registered on a common mesh, such as affinity propagation clustering [25] and co-clustering (rows and columns of matrix are clustered simultaneously) [26, 27].

Third, another notable type of the feature engineering is dimension reduction. In 2006, some methods applied dimension reduction methods such as principal component analysis (PCA) and independent component analysis (ICA) to the registered

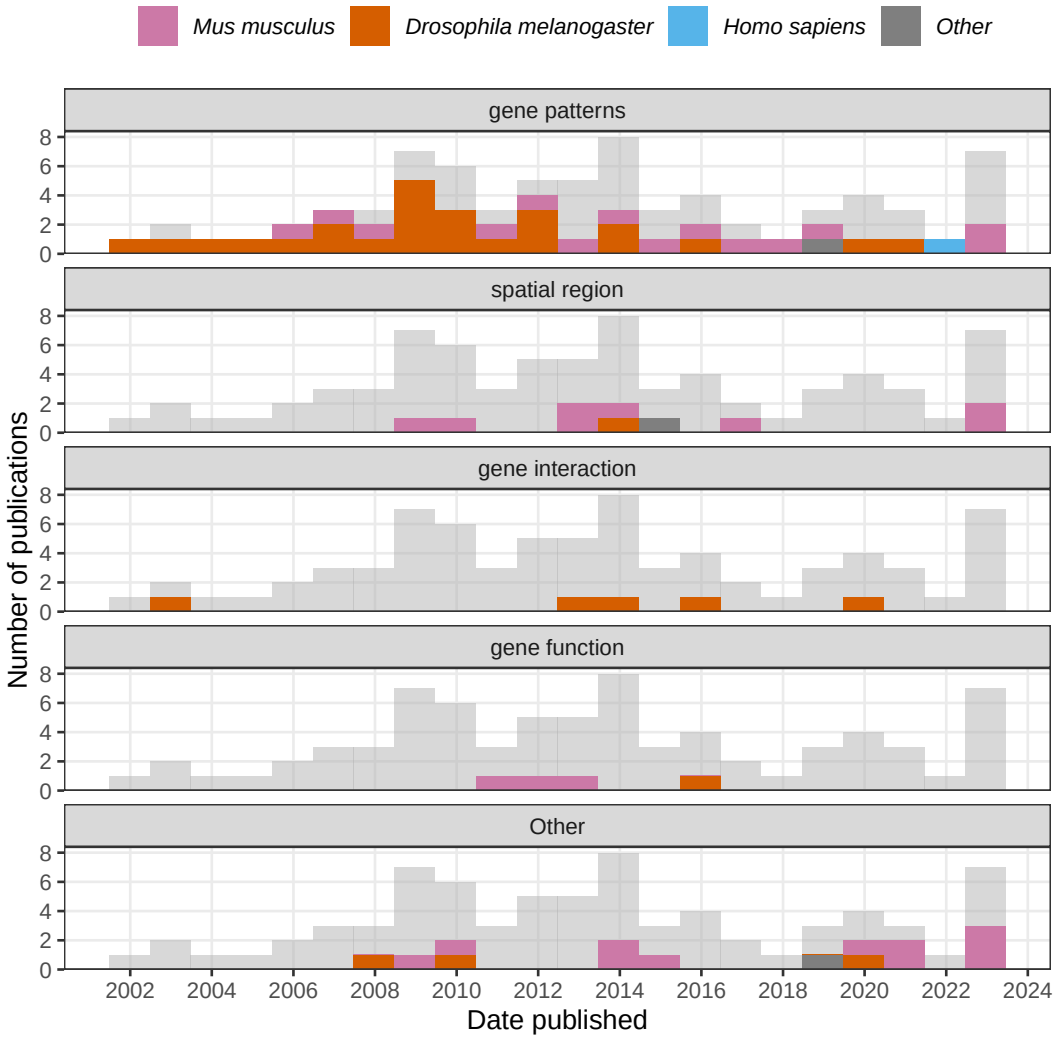

Figure 5.3: Number of publications in each time bin for each category of data analysis is highlighted in the facets.

images to find "eigen" patterns [28, 29]. Instead of PCA or ICA, the dimension reduction can also be sparse Bayesian factor analysis [30], sparse dictionary learning [31], and non-negative matrix factorization (NMF) [32, 33]. The dimension reduction can be used for unsupervised clustering of genes [28, 29, 30], as well as supervised classification methods such as SVM and logistic regression to annotate gene expression patterns with controlled vocabulary [30, 33]. Notably, in NMF, both the matrix for basis patterns and the coefficient matrix for the genes tend to exhibit block structures; the blocks in the gene coefficient matrix have been used to cluster genes [32].

Fourth, since 2015, convolutional neural networks (CNNs) have been adopted to

analyze gene expression patterns. Typically, a pre-trained model, such as ResNet50, OverFeat, or Alexnet is used. With some modifications or retraining of the original model, the model can be used to extract features for gene pattern annotation with logistic regression [34], classifying new patterns [35], and predicting interactions between genes [36].

5.2 Spatial regions

Closely related to classifying gene expression patterns are these questions: What are the implications of gene expression patterns to traditional anatomical regions as in the CV? Can we discover novel anatomical regions from gene expression? How well do expression-based regions correspond to the traditional regions? A few studies, which we call "spatial region", tried to answer these questions in the ABA (Figure 5.3). Clusters of expression patterns of cell type specific genes [37], or the most localized genes [38], principal components of the patterns [39], or patterns of coexpression modules were compared to traditional anatomy [38]. At least in the mouse brain, with the principal components, these clusters may correspond to traditional anatomy quite well [39]. However, when cell types are taken into account in clustering, gene expression seems to be able to refine traditional anatomy [37, 38].

A clustering strategy for identifying spatial regions that takes the spatial neighborhood into account is Markov random field (MRF). In MRFs, nearby voxels can be made to be more likely to share a label, which can be cell type or histological region, and the probability of a voxel taking each of the labels only depends on labels of neighboring voxels. MRFs were used to delineate spatial regions in a 3D FISH atlas of the developing *Platynereis dumereilii* brain [8], with 86 high quality genes. The images in the atlas were aligned into a 3D model and broken into voxels 3 μm per side, which is smaller than a typical single-cell; the spatial neighborhood graph is the 3D square grid of the voxels. As FISH is not very quantitative, the gene expression was manually binarized. Expression of each gene at each voxel is modeled with a Bernoulli distribution, and the 86 genes are assumed to be independent. Cluster label assignment is modeled with Potts model, a type of MRF in which only neighboring voxels with the same label contribute to the probability distribution of the labels, thus favoring neighbors with the same label. The parameters, such as interaction strength between neighboring voxels for the Potts model and the probability parameter of the Bernoulli distributions are estimated with expectation maximization (EM).

5.3 Gene interactions

While not single-cell resolution, (WM)ISH atlases provide transcriptomes within the tissue at a resolution far higher than that of typical bulk RNA-seq and bulk microarray, thus opening the way to studying coexperssion and interaction between genes within the tissue. There are a few methods that aim to decide whether two genes interact according to (WM)ISH images, some dating published long before the popularization of scRNA-seq. Already in 2002, an early method that compares binarized gene expression patterns was used to identify interactions among genes by comparing patterns from wild type and mutant backgrounds [13].

However, as mutant lines are harder to obtain than wild type images, the simplest method is to set a threshold in Pearson correlation coefficient between two genes to decide an edge should be drawn on the gene coexpression graph [33, 40].

Alternatively, a sparse Markov network whose nodes are genes and edges are presence of interaction can be learnt from expression profiles in each voxel [41], or a CNN can be trained on known interactions and predict new interactions based on gene expression patterns [36]. There are other types of analyses, such as inferring gene function from expression pattern, identifying spatially variable genes, and gene expression imputation at locations. The latter two are still important topics in current era data analysis.

5.4 Decline

What contributed to the decline of the golden age of prequel data analysis? Partly a lack of usage of the methods developed, which was exacerbated by the decline of the golden age of (WM)ISH atlases in the 2010s so there were fewer new atlases where the methods can be applied (Figure 4.3). While many methods to automate gene expression pattern annotation for BDGP were developed before 2013, for the 2013 BDGP update that added images of 708 transcription factors, the BDGP annotated the new images with human curators instead of the automated methods [42]. Nor did BDGP use the new methods to compare and classify the new gene expression patterns; instead, the curator assigned CV annotations were used for analysis [42, 43]. BDGP did not have a major update after 2013; as existing images have already been annotated, there is no need to automate annotations.

There are additional possible reasons why these methods were not used: First, it is unclear from the publications of the methods where the software implementation can be obtained. Second, a reason why most prequel analysis methods were developed

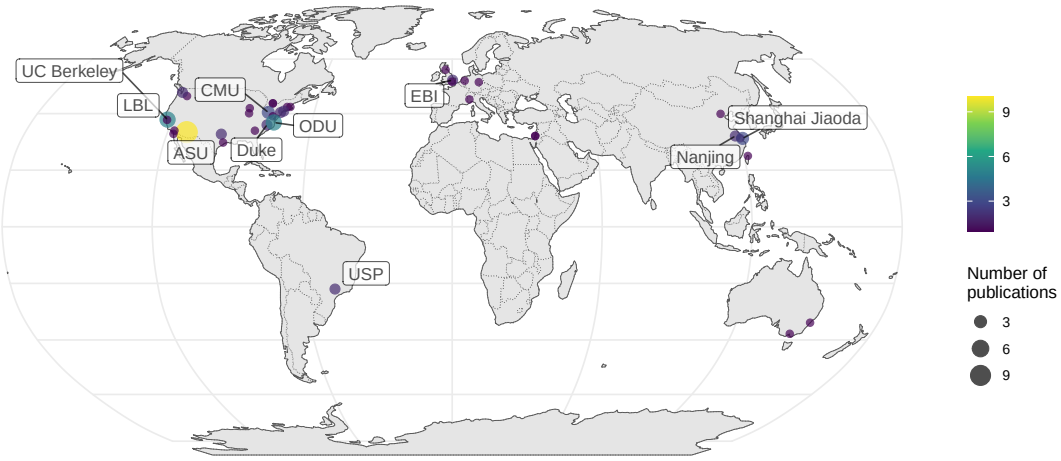

Figure 5.4: Number of publications per city for prequel data analysis.

for either BDGP or ABA is that since one gene is stained for in one embryo/section at a time, the images must be registered and standardized for different genes to be comparable; BDGP, through FlyExpress [44], and ABA, provide images that have already been registered and standardized, while many other atlases, such as GEISHA, do not. Due to challenges in image registration in other organisms, the automated gene expression pattern analysis methods can't be applied. Third, lack of usage of these methods can also be due to insufficient accuracy; from 2009 to 2013, the area under the curve (AUC) of the automated annotations is typically around 0.8 and rarely exceeded 0.9 [20, 30, 23, 21], which means when using such tools to annotate new images, extensive human review would still be required.

5.5 Geography of prequel data analysis

If our collection is representative, then contribution to prequel data analysis concentrates in a few institutions (Figure 5.4), not all of which are elite.

When broken down by species, it seems that distinct institutions contributed to data analysis of *Drosophila* and mouse data. UC Berkeley and Lawrence Berkeley National Laboratory (LBL) are responsible for BDGP, and Allen is responsible for ABA. However, among the top contributors are other institutions such as Arizona State University (ASU) and Old Dominion University (ODU) (Figure 5.6).

References

1. O'Kane CJ and Gehring WJ. Detection in situ of genomic regulatory elements in Drosophila. Proceedings of the National Academy of Sciences of the United

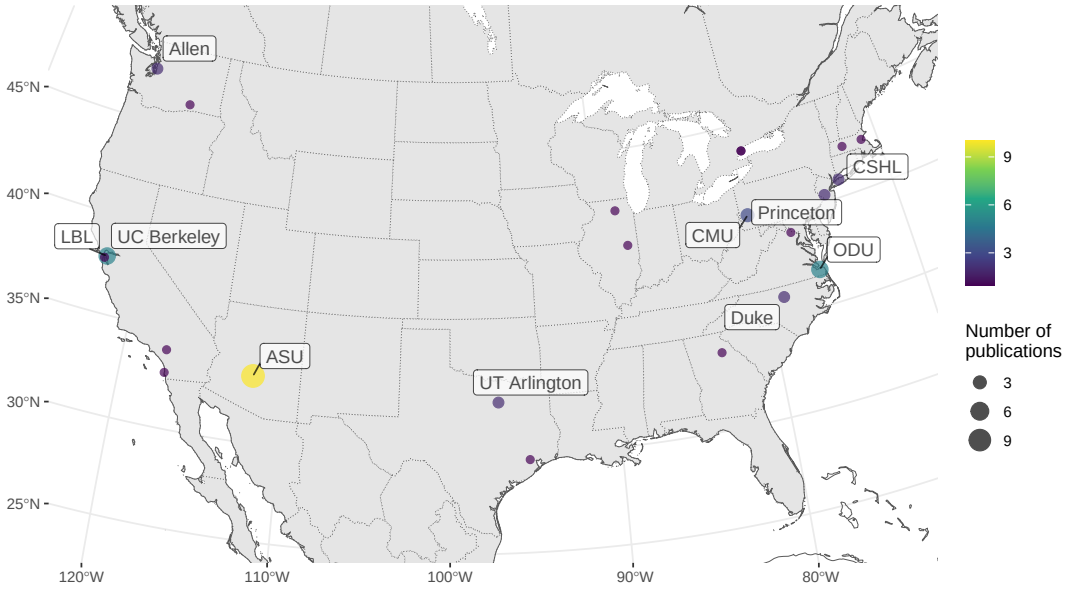

Figure 5.5: Number of publications per city for prequel data analysis in the US.

States of America 1987 Dec; 84:9123–7. DOI: 10.1073/pnas.84.24.9123. Available from: https://www.pnas.org/content/84/24/9123

2. Gossler A, Joyner A, Rossant J, and Skarnes W. Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. Science 1989 Apr; 244:463–5. DOI: 10.1126/science.2497519. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.2497519

3. Wurst W, Rossant J, Prideaux V, Kownacka M, Joyner A, Hill DP, Guillemot F, Gasca S, Cado D, and Auerbach A. A large-scale gene-trap screen for insertional mutations in developmentally regulated genes in mice. Genetics 1995; 139. Available from: https://www.genetics.org/content/139/2/889

4. Sundaresan V, Springer P, Volpe T, Haward S, Jones JD, Dean C, Ma H, and Martienssen R. Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. Genes & Development 1995 Jul; 9:1797–810. DOI: 10.1101/gad.9.14.1797. Available from: http://www.genesdev.org/cgi/doi/10.1101/gad.9.14.1797

5. Gawantka V, Pollet N, Delius H, Vingron M, Pfister R, Nitsch R, Blumenstock C, and Niehrs C. Gene expression screening in Xenopus identifies molecular pathways, predicts gene function and provides a global view of embryonic patterning. Mechanisms of Development 1998 Sep; 77:95–141. DOI: 10.1016/S0925-4773(98)00115-4. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0925477398001154

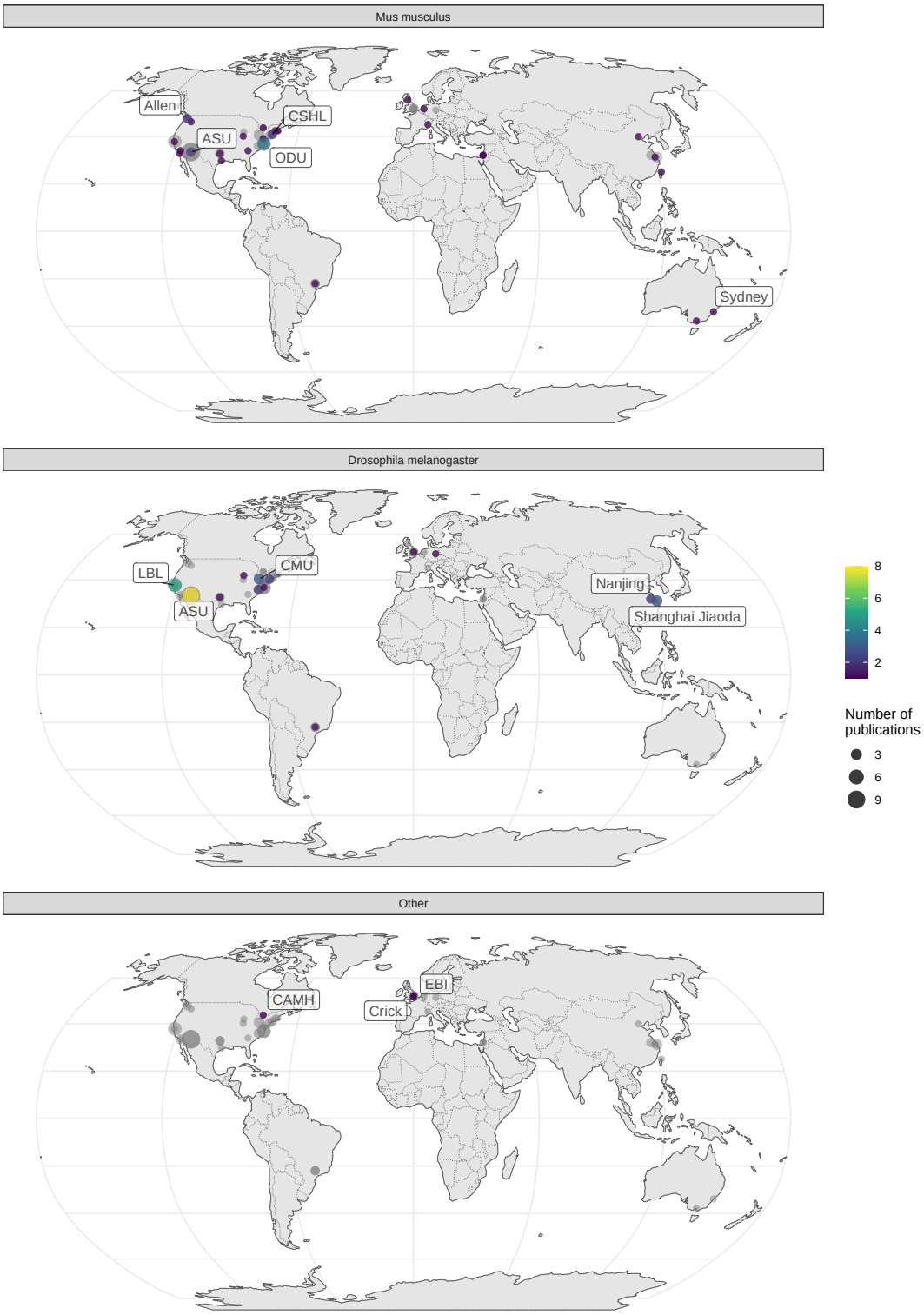

Figure 5.6: Number of publications per city for prequel data analysis broken down by species of interest.

6. Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu SQ, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, and Rubin GM. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome biology 2002 Dec; 3:research0088.1. DOI: 10.1186/gb-2002-3-12-research0088. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-12-research0088

7. Lein ES et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature 2007 Jan; 445:168–76. DOI: 10.1038/nature05453. Available from: http://www.nature.com/articles/nature05453

8. Pettit JB, Tomer R, Achim K, Richardson S, Azizi L, and Marioni J. Identifying Cell Types from Spatially Referenced Single-Cell Expression Datasets. PLoS Computational Biology 2014 Sep; 10. Ed. by Morris Q:e1003824. DOI: 10.1371/journal.pcbi.1003824. Available from: https://dx.plos.org/10.1371/journal.pcbi.1003824

9. Patrushev I, James-Zorn C, Ciau-Uitz A, Patient R, and Gilchrist MJ. New methods for computational decomposition of whole-mount in situ images enable effective curation of a large, highly redundant collection of Xenopus images. PLoS Computational Biology 2018 Aug; 14:e1006077. DOI: 10.1371/journal.pcbi.1006077. Available from: https://doi.org/10.1371/journal.pcbi.1006077

10. Abed-Esfahani P, Darwin BC, Howard D, Wang N, Kim E, Lerch J, and French L. Evaluation of deep convolutional neural networks for in situ hybridization gene expression image representation. bioRxiv 2021 Jan :2021.01.22.427860. DOI: 10.1101/2021.01.22.427860. Available from: http://biorxiv.org/content/early/2021/01/25/2021.01.22.427860.abstract

11. Jayaraman K, Panchanathan S, and Kumar S. Classification and Indexing of Gene Expression Images. Tech. rep. 2001

12. Gurunathan R, Van Emden B, Panchanathan S, and Kumar S. Identifying spatially similar gene expression patterns in early stage fruit fly embryo images: Binary feature versus invariant moment digital representations. BMC Bioinformatics 2004 Dec; 5:202. DOI: 10.1186/1471-2105-5-202. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-202

13. Kumar S, Jayaraman K, Panchanathan S, Gurunathan R, Marti-Subirana A, and Newfeld SJ. BEST: A Novel Computational Approach for Comparing Gene Expression Patterns From Early Stages of Drosophila melanogaster Development. Genetics 2002 Dec; 162:2037 LP –2047. Available from: http://www.genetics.org/content/162/4/2037.abstract

14. Liu Z, Yan SF, Walker JR, Zwingman TA, Jiang T, Li J, and Zhou Y. Study of gene function based on spatial co-expression in a high-resolution mouse brain atlas. BMC Systems Biology 2007 Apr; 1:19. DOI: 10.1186/1752-0509-1-19. Available from: http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-1-19

15. Zhou J and Peng H. Automatic recognition and annotation of gene expression patterns of fly embryos. Bioinformatics 2007 Mar; 23:589–96. DOI: 10.1093/bioinformatics/btl680. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl680

16. Heffel A, Stadler PF, Prohaska SJ, Kauer G, and Kuska JP. Process flow for classification and clustering of fruit fly gene expression patterns. *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008 :721–4. DOI: 10.1109/ICIP.2008.4711856. Available from: http://ieeexplore.ieee.org/document/4711856/

17. Lowe DG. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 2004 Nov; 60:91–110. DOI: 10.1023/B:VISI.0000029664.99615.94. Available from: http://link.springer.com/10.1023/B:VISI.0000029664.99615.94

18. Ji S, Sun L, Jin R, Kumar S, and Ye J. Automated annotation of Drosophila gene expression patterns using a controlled vocabulary. Bioinformatics 2008 Sep; 24:1881–8. DOI: 10.1093/bioinformatics/btn347. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btn347

19. Li YX, Ji S, Kumar S, Ye J, and Zhou ZH. Drosophila Gene Expression Pattern Annotation through Multi-Instance Multi-Label Learning. IJCAI : proceedings of the conference 2009 Jan; 2009:1445–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20824158%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2932460

20. Ji S, Li YX, Zhou ZH, Kumar S, and Ye J. A bag-of-words approach for Drosophila gene expression pattern annotation. BMC Bioinformatics 2009 Apr; 10:119. DOI: 10.1186/1471-2105-10-119. Available from: http://www.biomedcentral.com/1471-2105/10/119

21. Sun Q, Muckatira S, Yuan L, Ji S, Newfeld S, Kumar S, and Ye J. Image-level and group-level models for Drosophila gene expression pattern annotation. BMC Bioinformatics 2013 Dec; 14:350. DOI: 10.1186/1471-2105-14-350. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-350

22. Ji S, Yuan L, Li YX, Zhou ZH, Kumar S, and Ye J. Drosophila gene expression pattern annotation using sparse features and term-term interactions. *Proceedings of the 15th ACM SIGKDD international conference on Knowl-*

edge discovery and data mining - KDD '09. New York, New York, USA: ACM Press, 2009 :407. DOI: 10.1145/1557019.1557068. Available from: http://portal.acm.org/citation.cfm?doid=1557019.1557068

23. Yuan L, Woodard A, Ji S, Jiang Y, Zhou ZH, Kumar S, and Ye J. Learning Sparse Representations for Fruit-Fly Gene Expression Pattern Image Annotation and Retrieval. BMC Bioinformatics 2012 May; 13:107. DOI: 10.1186/1471-2105-13-107. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-107

24. Liscovitch N, Shalit U, and Chechik G. FuncISH: learning a functional representation of neural ISH images. Bioinformatics 2013 Jul; 29:i36–i43. DOI: 10.1093/bioinformatics/btt207. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt207

25. Frise E, Hammonds AS, and Celniker SE. Systematic image-driven analysis of the spatial Drosophila embryonic expression landscape. Molecular Systems Biology 2010 Jan; 6:345. DOI: 10.1038/msb.2009.102. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1038/msb.2009.102

26. Jagalur M, Pal C, Learned-Miller E, Zoeller RT, and Kulp D. Analyzing in situ gene expression in the mouse brain with image registration, feature extraction and block clustering. BMC Bioinformatics 2007 Dec; 8:S5. DOI: 10.1186/1471-2105-8-S10-S5. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-S10-S5

27. Zhang W, Feng D, Li R, Chernikov A, Chrisochoides N, Osgood C, Konikoff C, Newfeld S, Kumar S, and Ji S. A mesh generation and machine learning framework for Drosophila gene expression pattern image analysis. BMC Bioinformatics 2013 Dec; 14:372. DOI: 10.1186/1471-2105-14-372. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-372

28. Pan JY, Guilherme A, Balan R, Xing EP, Traina AJM, and Faloutsos C. Automatic mining of fruit fly embryo images. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*. Vol. 2006. New York, New York, USA: ACM Press, 2006 :693. DOI: 10.1145/1150402.1150489. Available from: http://portal.acm.org/citation.cfm?doid=1150402.1150489

29. Hanchuan Peng, Fuhui Long, Eisen M, and Myers E. Clustering Gene Expression Patterns of Fly Embryos. *3rd IEEE International Symposium on Biomedical Imaging: Macro to Nano, 2006.* Vol. 2006. IEEE, 2006 :1144–7. DOI: 10.1109/ISBI.2006.1625125. Available from: http://ieeexplore.ieee.org/document/1625125/

30. Pruteanu-Malinici I, Mace DL, and Ohler U. Automatic Annotation of Spatial Expression Patterns via Sparse Bayesian Factor Models. PLoS Computational Biology 2011 Jul; 7. Ed. by Bader JS:e1002098. DOI: 10.1371/journal.pcbi.1002098. Available from: https://dx.plos.org/10.1371/journal.pcbi.1002098

31. Li Y, Chen H, Jiang X, Li X, Lv J, Peng H, Tsien JZ, and Liu T. Discover mouse gene coexpression landscapes using dictionary learning and sparse coding. Brain Structure and Function 2017 Dec; 222:4253–70. DOI: 10.1007/s00429-017-1460-9. Available from: http://link.springer.com/10.1007/s00429-017-1460-9

32. Noto T, Barnagian D, and Castro JB. Genome-scale investigation of olfactory system spatial heterogeneity. PLOS ONE 2017 May; 12. Ed. by Matsunami H:e0178087. DOI: 10.1371/journal.pone.0178087. Available from: https://dx.plos.org/10.1371/journal.pone.0178087

33. Wu S, Joseph A, Hammonds AS, Celniker SE, Yu B, and Frise E. Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. Proceedings of the National Academy of Sciences 2016 Apr; 113:4290–5. DOI: 10.1073/pnas.1521171113. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1521171113

34. Zeng T, Li R, Mukkamala R, Ye J, and Ji S. Deep convolutional neural networks for annotating gene expression patterns in the mouse brain. BMC Bioinformatics 2015 Dec; 16:147. DOI: 10.1186/s12859-015-0553-9. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0553-9

35. Long W, Li T, Yang Y, and Shen H. FlyIT: Drosophila Embryogenesis Image Annotation based on Image Tiling and Convolutional Neural Networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2021; 18:194–204. DOI: 10.1109/TCBB.2019.2935723

36. Yang Y, Fang Q, and Shen HB. Predicting gene regulatory interactions based on spatial gene expression data and deep learning. PLOS Computational Biology 2019 Sep; 15. Ed. by Krishnaswamy S:e1007324. DOI: 10.1371/journal.pcbi.1007324. Available from: https://dx.plos.org/10.1371/journal.pcbi.1007324

37. Ko Y, Ament SA, Eddy JA, Caballero J, Earls JC, Hood L, and Price ND. Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain. Proceedings of the National Academy of Sciences 2013 Feb; 110:3095–100. DOI: 10.1073/pnas.1222897110. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1222897110

38. Grange P, Bohland JW, Okaty BW, Sugino K, Bokil H, Nelson SB, Ng L, Hawrylycz M, and Mitra PP. Cell-type-based model explaining coexpression patterns of genes in the brain. Proceedings of the National Academy of Sciences 2014 Apr; 111:5397–402. DOI: 10.1073/pnas.1312098111. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1312098111

39. Bohland JW, Bokil H, Pathak SD, Lee CK, Ng L, Lau C, Kuan C, Hawrylycz M, and Mitra PP. Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. Methods 2010 Feb; 50:105–12. DOI: 10.1016/j.ymeth.2009.09.001. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1046202309002035

40. Campiteli MG, Comin CH, Costa LDF, Babu MM, and Cesar RM. A methodology to infer gene networks from spatial patterns of expression – an application to fluorescence in situ hybridization images. Molecular BioSystems 2013 Jun; 9:1926. DOI: 10.1039/c3mb25475e. Available from: http://xlink.rsc.org/?DOI=c3mb25475e

41. Puniyani K and Xing EP. GINI: From ISH Images to Gene Interaction Networks. PLoS Computational Biology 2013 Oct; 9. Ed. by Chechik G:e1003227. DOI: 10.1371/journal.pcbi.1003227. Available from: https://dx.plos.org/10.1371/journal.pcbi.1003227

42. Hammonds AS, Bristow CA, Fisher WW, Weiszmann R, Wu S, Hartenstein V, Kellis M, Yu B, Frise E, and Celniker SE. Spatial expression of transcription factors in Drosophila embryonic organ development. Genome Biology 2013 Dec; 14:R140. DOI: 10.1186/gb-2013-14-12-r140. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-12-r140

43. Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, Hartenstein V, Celniker SE, and Rubin GM. Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biology 2007 Jul; 8:R145. DOI: 10.1186/gb-2007-8-7-r145. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-7-r145

44. Kumar S, Konikoff C, Sanderford M, Liu L, Newfeld S, Ye J, and Kulathinal RJ. FlyExpress 7: An Integrated Discovery Platform To Study Coexpressed Genes Using in situ Hybridization Images in Drosophila. G3; Genes|Genomes|Genetics 2017 Aug; 7:2791–7. DOI: 10.1534/g3.117.040345. Available from: http://g3journal.org/lookup/doi/10.1534/g3.117.040345

Chapter 6

FROM THE PAST TO THE PRESENT

6.1 Legacy of the prequel era

The current era continues many of the quests of the prequel era, such as to profile the transcriptome in space, to identify genes with restricted expression, to classify gene expression patterns, to build reference gene expression atlases for model systems, and to infer anatomical regions based on gene expression. While the prequel era also sought to identify cell type markers, this has been taken over by non-spatial transcriptomics, which has been used to identify marker genes to stain for with spatial transcriptomics methods not easily scalable to the whole transcriptome. As already mentioned, (WM)ISH atlases can be understood as an improved alternative to microarray and *in situ* reporter screens, and the latter can be in turn understood as an improved alternative to enhancer and gene traps. To some extent, current era spatial transcriptomics started as an improved alternative to (WM)ISH atlases, to profile the whole transcriptome in the same cells [1, 2]. On the other hand, part of current era of spatial transcriptomics can be seen as an improvement to bulk microarray or RNA-seq [3, 1, 4, 5], and lower throughput single-cell biology [6, 7].

How does the current era relate to the prequel era in general? The current era has undergone massive growth unseen in the prequel era (Figure 6.2). Unlike in the prequel era, current era technologies are typically highly multiplexed to quantify hundreds to thousands of genes if not the transcriptome within the same piece of tissue. While cell segmentation of bright field (WM)ISH images is challenging, in some current era technologies, transcripts can be traced back to the individual cells of origin. Moreover, cost of NGS has greatly decreased, and the most popular current era techniques—LCM followed by RNA-seq, and 10X Visium—rely on NGS to quantify transcripts, thus making it much more efficient to profile transcriptomes in space than with (WM)ISH, let alone enhancer and gene traps.

Again, we may take inspiration from histories of other technologies that have no doubt undergone revolutions to illustrate where we are in what appears to be a revolution in progress in part propelled by NGS and greater computing power. The growing popularity and greatly improved efficiency of current era techniques compared to those of prequel techniques in applications to thousands of genes may

be akin to how the safety bicycle relates to the penny-farthing. The advances from the former has rendered the latter virtually obsolete, and nearly all bikes we see on the streets today are much more like the safety bicycle in both appearance and mechanism than the penny-farthing. Since the 1890s, when the safety bicycle became popular, bicycle technologies have drastically improved. However, most histories of cycling do mention the penny-farthing and its ancestors such as the bone shaker, velocipedes (where the "velo" in "velodrome" comes from), and the hobby horse. Some histories mention bicycles propelled by treadles rather than the familiar cranks and a 17th century four wheeled human powered vehicle propelled by pulling ropes from within. Despite these vehicles' drastically different mechanisms from the modern bicycle, as these are earlier and less successful attempts to achieve the goal of devising a human powered land vehicle that travels faster than walking, which is still one of the primary goals of modern bicycle-related technologies. These histories are really histories of the quest to achieve that goal.

Moreover, when you see a lightening fast high-tech aerodynamic carbon fiber time trial bike tested in an aerospace wind tunnel, the penny-farthing is not to be forgotten, because the former still benefits from legacies from the latter. Roads used to be unpaved and very rough, and in the US, the paved roads, road signs guiding travelers, and interstates originated from advocacy by the League of American Wheelmen (LAW) since 1880, which was the penny-farthing era [8]. The same may be said for the UK [9]. As the automobile replaced the safety bicycle as the favored mode of transportation in the 20th century (another revolution in transportation), drivers are not only benefiting from the better roads advocated by early cyclists but also the pneumatic tire originally popularized by the safety bicycle. Finally, without the legacy of LAW's advocacy, the modern form of fast road racing for which the high-tech carbon fiber time trial bike is built would not be possible. Today, LAW, which has been renamed League of American Bicyclists (LAB), is still operational as a cycling advocacy group.

On the one hand, just like the penny-farthing, which is now obsolete except in some hobbyist niches, prequel techniques can be seen as earlier and less successful attempts to achieve the goal to profile expression of as many genes as possible while preserving spatial context in tissue, less successful attempts whose disadvantages are addressed in newer and more successful attempts. While enhancer and gene traps and ISH atlases never completely died off (Figure 4.3), there is no doubt that to profile expression of larger number of genes in new studies, prequel techniques have

by and large been replaced by current era techniques. On the other hand, just like LAW/B, the legacies of prequel spatial transcriptomics directly benefit the current era.

How has the prequel era influenced the current era? The direct influence does not seem profound overall when considering all biological systems studied, but is nevertheless sizable. For mouse brain studies, the Allen Brain Institute and its ABA do seem to have a bigger influence. The most obvious institutional continuation between the prequel and current eras is the Allen Brain Institute, which used ISH for the mouse ISH atlases, LCM and microarray for the human and macaque atlases, and generated bulk and scRNA-seq datasets as part of the atlases. Allen scRNA-seq data is often used to benchmark computational methods to map dissociated scRNA-seq cells to spatial locations in tissue and/or to impute gene expression in space (the two related tasks are collectively called spatial reconstruction of scRNA-seq here), with STARmap [10], osmFISH [11], MERFISH [12], and/or Visium [13] mouse cortex data as the spatial reference [14, 15, 16, 17, 13]; this is an institutional legacy from the prequel era. Another prequel era institutional legacy is the Jackson Lab (JAX), home of the prequel GXD, and where many lab mice come from. JAX has also contributed to the current era with the recent Visium mouse urinary bladder atlas [18]. The data might soon be available for online exploration with `cellxgene` on the JAX single-cell Portal but is not yet available as of writing. However, for the most part, as shown later in this chapter, prequel and current era data collection techniques were developed and used in distinct institutions, suggesting that the two eras are largely sociologically distinct (Figure 6.12). This is not surprising given that different techniques in the current era are also often developed and used in largely distinct institutions (e.g. 7.27).

The influence is mainly usage of prequel resources in current era data analysis, mostly in spatial reconstruction of scRNA-seq data and cross referencing to validate or interpret computational results. As already mentioned in Chapter 4, early scRNA-seq spatial reconstruction methods used binarized prequel style WMISH atlases for zebrafish embryos (Seurat v1) and *Platynereis* and whole mount FISH atlas BDTNP for *Drosophila* as spatial references. Thereafter the Seurat v1 zebrafish WMISH atlas and BDTNP have been used to benchmark several new spatial reconstruction methods [15, 19, 17], including methods developed for the DREAM challenge to map cells to locations with smaller number of informative genes ([20, 21, 22]. However, such benchmarks do not seem to indicate interest in studying the biology

of zebrafish and *Drosophila* development in 3D, as the purpose of such benchmark is more to validate computational methods than to perform biological inferences. Furthermore, zebrafish and *Drosophila* only take up very small proportions of all current era studies compared to mouse and human (Figure 6.4). For the mouse brain, the ABA mouse ISH atlas has been used as the spatial reference to quantitatively map scRNA-seq cell types to spatial locations [23]. The Allen developing mouse brain ISH atlas was also used as spatial reference to map human brain organoid scRNA-seq cells to space and mouse developmental stages for interpretation [24]. Here, unlike the WMISH atlases and BDTNP, the ABA is not binarized before spatial mapping.

With staining for around 20,000 genes, the ABA is more frequently used to qualitatively confirm that the computationally imputed gene expression patterns recapitulate the ISH staining of the same genes [14, 25, 16, 26, 27]. EMAGE eMouseAtlas [28] has been used to qualitative validate Geo-seq [29] and DBiT-seq [30] results, but usage of EMAGE is rare in the current era. In current era ST and Visium, an H&E image of the tissue accompanies the spatial transcriptome. The H&E image of mouse brains has been used to manually or computationally align the dataset to the Allen Mouse Brain Common Coordinate Framework (CCF) [31] to integrate ABA's brain anatomical ontologies to new datasets to facilitate interpretation of the data [32, 33, 27]. Even without H&E, Allen ontologies have been used to manually annotate HybISS data from the developing mouse brain based on marker gene expression [34]. In the mouse primary motor cortex (MOp) MERFISH atlas [12], Allen CCF was used to select the MOp region.

There are over 15 extant mouse databases from the prequel era (Figure 4.9), yet ABA is exceptional in its impact on the current era. We have never seen any mention of other prequel mouse databases, such as Eurexpress [35], and GenePaint [36] in current era literature. This may be due to the following reasons: First, the ABA is the most comprehensive prequel atlas for the adult mouse brain, with around 20,000 genes for adult mice (P56). As of August 2021, EMAGE has ISH images for 17,554 genes, Eurexpress has 19,440 (that's the number of assays, but it seems that each gene typically has one assay, so it should be close to the number of genes), and GenePaint contains Eurexpress data and can query ISH data from several other databases including ABA. EMAGE covers a wide range of developmental stages, from E0.5 to E18, but not later stages and adults. Eurexpress only covers E14.5. The Allen developing mouse brain atlas covers from E11.5 to P28, though only with

about 2000 genes. When the ABA is used in the current era, most of the time the adult mouse atlas is used.

Second, the ABA has much better infrastructure to facilitate quantitative analyses of the atlas than the other prequel mouse atlases. Both EMAGE and Eurexpress have detailed annotation of ISH results for many genes and allow searching for genes with similar expression patterns. In addition, Eurexpress shows ISH for many consecutive sections in 3D, and EMAGE has 3D histology models (for morphology rather than gene expression) at different developmental stages. In addition to these functionalities, ABA quantified ISH staining and registered the quantified ISH to the CCF, so just like in scRNA-seq, each voxel would have a vector of gene expression values. Usage of ABA in the current era mentioned above would not be possible without the CCF. ABA also has an application programming interface (API) to automate retrieval of such quantitative data for analyses suitable for the quantitative nature of the current era. In contrast, we are unaware of such quantification, registration, and API in other prequel mouse atlases, thus restricting their uses to be more qualitative. A similar pattern can be seen in *Drosophila* prequel databases. BDTNP registered staining from thousands of embryos stained for different genes onto a common coordinate system. BDTNP data could also be easily downloaded as csv-like files that can be easily parsed, though as of August 2021, the BDTNP website is not responsive. As seen in 5, a reason why FlyExpress was commonly used was that images for different genes were registered in FlyExpress.

6.2 Metadata of the current era

The current era started with LCM followed by microarray in 1999 [5]. Due to the immense popularity of LCM followed by microarray or RNA-seq, the body of LCM literature is too vast for unbiased and comprehensive manual curation, so the curated database does not include most LCM literature, which was instead collected from a PubMed search and text mined (Figure 8.3, Chapter 8). Because the search results—without manual inspection and curation—may contain irrelevant entries and miss relevant ones, they are separated from the curated database in our analyses. Current era literature in the curated database is classified into Microdissection, smFISH, ISS, Array, and No Imaging, to be defined in detail in their corresponding sections below.

Table 6.1: Summary of spatial transcriptomics techniques in the current era

Method	Date published	Category	Max # genes	Min spot diameter (μm)
voxelation	2002-02-01	ROI selection	Tx wide	NA
PA-GFP	2010-11-12	ROI selection	Tx wide	NA
SRM seqFISH	2012-06-03	smFISH	32	single-cell
Tomo-array	2012-09-19	ROI selection	Tx wide	NA
iceFISH	2013-02-17	smFISH	20	single-cell
ISS	2013-07-14	ISS	222	single-cell
Tomo-seq	2013-08-12	ROI selection	Tx wide	NA
bDNA-smFISH	2013-10-06	smFISH	928	single-cell
TIVA	2014-01-12	ROI selection	Tx wide	NA
FISSEQ	2014-03-21	ISS	8102	single-cell
seqFISH	2014-03-28	smFISH	10421	single-cell
MERFISH	2015-04-24	smFISH	4209	single-cell
Puzzle Imaging	2015-07-20	De novo	NA	NA
Geo-seq	2016-03-21	ROI selection	Tx wide	NA
corrFISH	2016-06-06	smFISH	10	single-cell
ST	2016-07-01	NGS barcoding	Tx wide	100
HCR-seqFISH	2016-10-19	smFISH	249	single-cell
punch	2017-06-28	ROI selection	Tx wide	NA
SGA	2017-11-28	smFISH	35	single-cell
APEX-RIP	2017-12-14	De novo	NA	NA
Niche-seq	2017-12-22	ROI selection	Tx wide	NA
ExM-MERFISH	2018-03-19	smFISH	10050	single-cell
STARmap	2018-07-27	ISS	1020	single-cell
Paired-cell sequencing	2018-09-17	De novo	NA	NA
osmFISH	2018-10-30	smFISH	33	single-cell

Table 6.1: Summary of spatial transcriptomics techniques in the current era

Method	Date published	Category	Max # genes	Min spot diameter (μm)
seqFISH+	2019-03-25	smFISH	10000	single-cell
slide-seq	2019-03-29	NGS barcoding	Tx wide	10
bDNA-MERFISH	2019-05-25	smFISH	130	single-cell
GeoMX DSP	2019-06-21	ROI selection	2093	NA
DNA microscopy	2019-06-27	De novo	NA	NA
APEX-seq	2019-07-11	De novo	NA	NA
INSTA-seq	2019-08-06	ISS	NA	single-cell
PARSIFT	2019-09-04	De novo	NA	NA
HDST	2019-09-09	NGS barcoding	Tx wide	2
GaST-seq	2019-10-10	ROI selection	Tx wide	NA
BARseq	2019-10-17	ISS	107	single-cell
PIC-seq	2020-03-09	De novo	NA	NA
miRNA nanowell	2020-05-09	NGS barcoding	9	300
split-FISH	2020-06-15	smFISH	317	single-cell
Visium	2020-06-22	NGS barcoding	Tx wide	55
ZipSeq	2020-07-06	ROI selection	Tx wide	NA
SMD-seq	2020-08-11	ROI selection	Tx wide	NA
HybISS	2020-09-29	smFISH	199	single-cell
DBiT-seq	2020-10-19	NGS barcoding	Tx wide	10
C-FISH	2020-10-23	smFISH	2	single-cell
SCRINSHOT	2020-11-20	smFISH	177	single-cell
slide-seq2	2020-12-07	NGS barcoding	Tx wide	10
Stereo-seq	2021-01-19	NGS barcoding	Tx wide	0.22
GeoMX WTA	2021-01-25	ROI selection	20175	NA
Seq-Scope	2021-01-27	NGS barcoding	Tx wide	0.5

Table 6.1: Summary of spatial transcriptomics techniques in the current era

Method	Date published	Category	Max # genes	Min spot diameter (μm)
ExSeq	2021-01-29	ISS	297	single-cell
BOLORAMIS	2021-03-08	ISS	96	single-cell
Pick-Seq	2021-03-09	ROI selection	Tx wide	NA
nanoneedles	2021-03-10	ROI selection	9	NA
CISI	2021-04-15	smFISH	37	single-cell
STRP-seq	2021-04-19	ROI selection	Tx wide	NA
XYZeq	2021-04-21	NGS barcoding	Tx wide	500
electro-seq	2021-04-23	ISS	201	single-cell
BARseq2	2021-05-10	ISS	65	single-cell
ClumpSeq	2021-05-24	De novo	NA	NA
sci-Space	2021-07-02	NGS barcoding	Tx wide	73.2
CIM-seq	2021-07-12	De novo	NA	NA
PIC	2021-07-20	ROI selection	Tx wide	NA
par-seqFISH	2021-08-13	smFISH	105	single-cell
SPACECAT	2021-08-17	ROI selection	Tx wide	NA
RNAscope	2021-09-29	smFISH	95	single-cell
Molecular Cartography	2021-10-12	smFISH	100	single-cell
Visium protein	2021-10-16	NGS barcoding	Tx wide	55
coppaFISH	2021-10-24	smFISH	72	single-cell
Raman2RNA	2021-12-01	smFISH	9	single-cell
EASI-FISH	2021-12-06	smFISH	29	single-cell
Halo-seq	2021-12-07	De novo	NA	NA
OpTAG-seq	2021-12-30	ROI selection	Tx wide	NA
MOSAICA	2022-01-10	smFISH	10	single-cell
LoRNA	2022-01-25	De novo	NA	NA

Table 6.1: Summary of spatial transcriptomics techniques in the current era

Method	Date published	Category	Max # genes	Min spot diameter (μm)
manual dissection with velocimetry and cell tracking	2022-01-31	ROI selection	Tx wide	NA
SM-Omics	2022-02-10	NGS barcoding	Tx wide	100
FUNseq	2022-02-22	ROI selection	Tx wide	NA
centrifugation on 384 well plate	2022-02-23	ROI selection	Tx wide	NA
Space-TREX	2022-02-24	NGS barcoding	Tx wide	55
MERR APEX-seq	2022-03-03	De novo	NA	NA
vCatFISH	2022-03-16	smFISH	21	single-cell
SPARC-seq	2022-03-23	NGS barcoding	Tx wide	50
GPS-seq	2022-04-05	NGS barcoding	Tx wide	NA
scStereo-seq	2022-05-04	NGS barcoding	Tx wide	0.22
Select-seq	2022-05-09	ROI selection	Tx wide	NA
HybRISS	2022-05-13	smFISH	175	single-cell
punch2	2022-06-17	ROI selection	Tx wide	NA
STARmap PLUS	2022-06-22	ISS	2766	single-cell
STcEM	2022-06-27	smFISH	287	single-cell
PIXEL-seq	2022-07-04	NGS barcoding	Tx wide	1.22
SmT	2022-07-18	NGS barcoding	Tx wide	55
CosMX	2022-07-19	smFISH	1020	single-cell
SHM-seq	2022-07-19	NGS barcoding	Tx wide	100
scNaST	2022-07-22	NGS barcoding	Tx wide	NA
PHYTOMap	2022-07-30	smFISH	28	single-cell

Table 6.1: Summary of spatial transcriptomics techniques in the current era

Method	Date published	Category	Max # genes	Min spot diameter (μm)
Matrix-seq	2022-08-05	NGS barcoding	Tx wide	50
xDbit	2022-09-01	NGS barcoding	Tx wide	50
ARTseq-FISH	2022-09-14	smFISH	67	single-cell
EEL FISH	2022-09-22	smFISH	445	single-cell
TEMPOmap	2022-09-27	ISS	991	single-cell
CBSST-Seq	2022-10-05	NGS barcoding	Tx wide	50
GeoMX SPG	2022-10-06	ROI selection	21000	NA
Light-Seq	2022-10-10	ROI selection	Tx wide	NA
Slide-TCR-seq	2022-10-11	NGS barcoding	Tx wide	10
clampFISH 2.0	2022-10-24	smFISH	10	single-cell
Spatial-seq	2022-10-30	ROI selection	Tx wide	NA
sphere-seq	2022-11-01	De novo	NA	NA
STRS	2022-11-03	NGS barcoding	Tx wide	55
Xenium	2022-11-03	smFISH	313	single-cell
SPRINTseq	2022-11-17	ISS	108	single-cell
LR-Spatial VDJ	2022-11-24	NGS barcoding	Tx wide	55
SR-Spatial VDJ	2022-11-24	NGS barcoding	Tx wide	55
mFISH3D	2022-11-24	smFISH	6	single-cell
STOmics-GenX	2022-12-08	NGS barcoding	Tx wide	0.22
Spectrum-FISH	2022-12-13	smFISH	33	single-cell
TATTOO-seq	2022-12-14	ROI selection	Tx wide	NA
DRaqL	2022-12-16	ROI selection	Tx wide	NA
SPOTS	2023-01-02	NGS barcoding	Tx wide	55
MiP-Seq	2023-01-07	ISS	217	single-cell

Table 6.1: Summary of spatial transcriptomics techniques in the current era

Method	Date published	Category	Max # genes	Min spot diameter (μm)
USeqFISH	2023-01-26	smFISH	30	single-cell
SMA	2023-01-27	NGS barcoding	Tx wide	NA
RRST	2023-01-31	NGS barcoding	Tx wide	55
IISS	2023-02-15	ISS	40	single-cell
Spatial-CITE-seq	2023-02-23	NGS barcoding	Tx wide	25
BMKMANU S1000	2023-02-26	NGS barcoding	Tx wide	NA
Spatial ATAC–RNA-seq	2023-03-15	NGS barcoding	Tx wide	20
Spatial CUT&Tag–RNA-seq	2023-03-15	NGS barcoding	Tx wide	NA
fs-LM	2023-03-16	ROI selection	Tx wide	NA
SiT	2023-03-17	NGS barcoding	Tx wide	55

Chronologically, in the curated database, microdissection came first, with voxelation in 2002 ([3]V. M. Brown et al. 2002), followed by smFISH, ISS, no imaging, and NGS barcoding (Figure 6.1). Despite an early start in the midst of the (WM)ISH golden age, if not including non-curated LCM literature, the current era did not really take off until around 2014 (Figure 6.2). Ever since, its has seen drastic growth, far exceeding that of the prequel era in the 1990s and 2000s (Figure 6.2). Growth in microdissection and NGS barcoding seemed to have contributed the most to this overall drastic growth (Figure 6.1). All techniques in the curated database, along with their classification, maximum number of genes, spatial resolution, and references are listed in Table 6.1.

A timeline of foundational or influential techniques in the current era is shown in Figure 6.3. This is not meant to be a timeline of all current era techniques, but

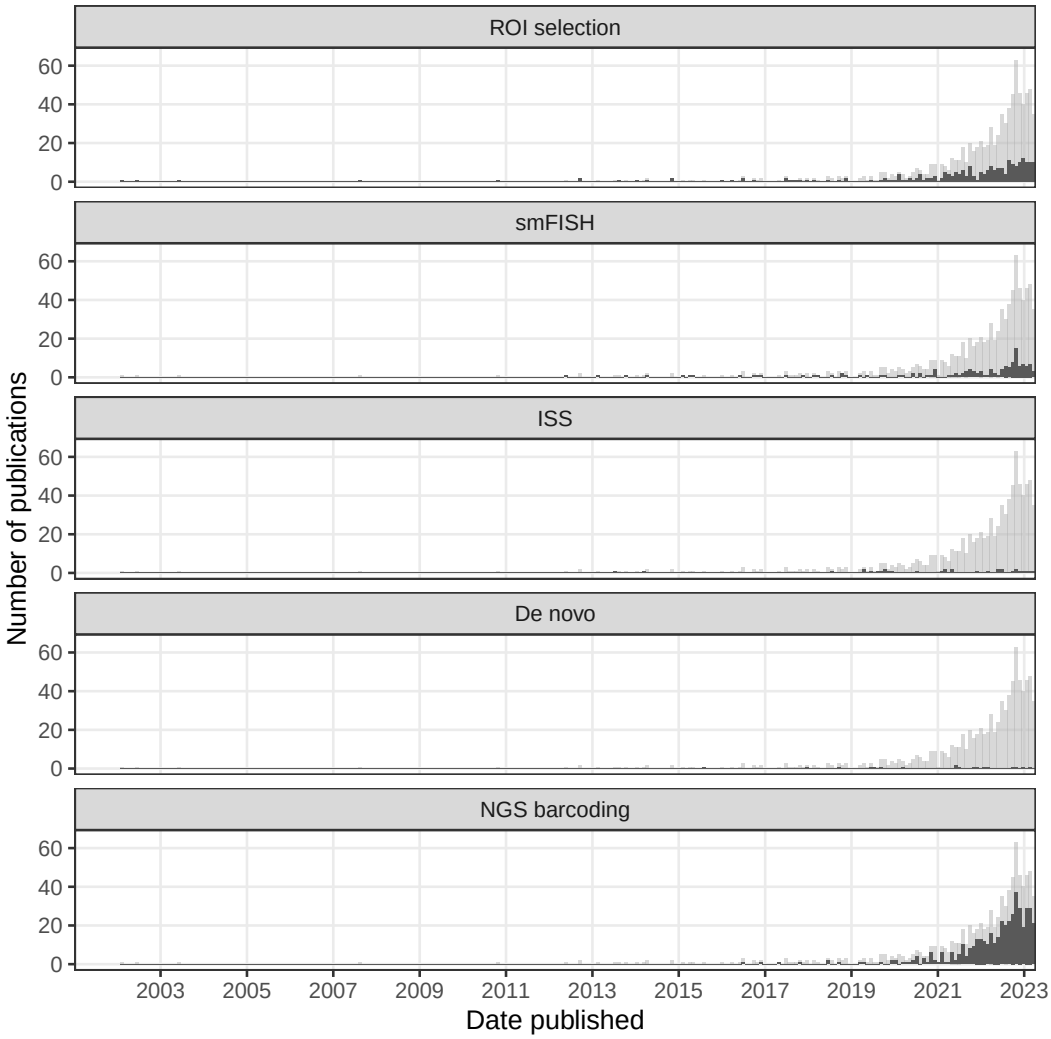

Figure 6.1: Number of publications over time in the current era. The gray histogram in the background is the overall trend of all current era literature. Each facet highlights a category, ordered chronologically in terms of first report. Bin width is 30 days. Plots in this figure include curated LCM literature, but not the non-curated literature.

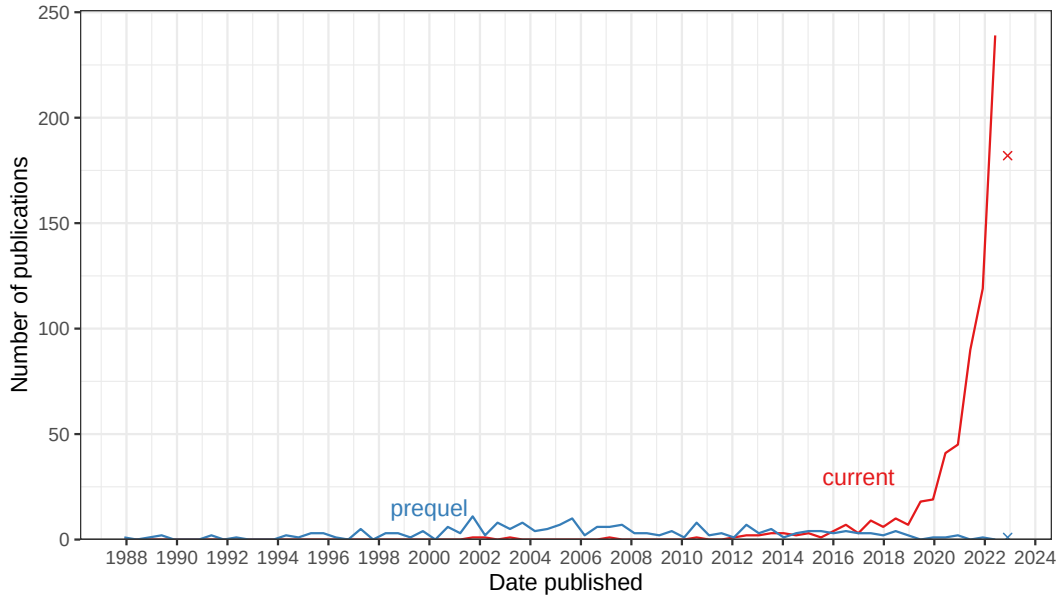

Figure 6.2: Comparing number of publications over time in the prequel and the current eras. Bin width is 180 days. The x-shaped points show the number of publications from the last bin, which is not yet full.

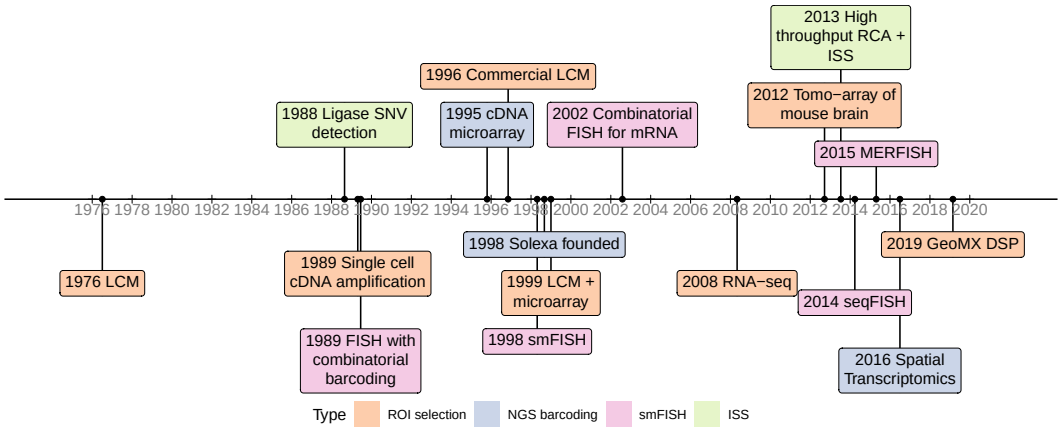

Figure 6.3: Timeline of major techniques related to the current era.

only of techniques that either laid the foundation of popular current era techniques (e.g. Solexa, later Illumina, sequencing) or very influential within a category of techniques (e.g. MERFISH for smFISH-based techniques, and ST for NGS barcoding). Just like the "revolution" of current era spatial transcriptomics, each item in the timeline must not be understood as works of the "solitary genius". Rather, each of the landmark innovations in the timeline occurred in its own historical context, with influences from predecessors, which are not plotted in the timeline for the sake of brevity.

The prequel era started with untargeted screens and grew into atlases and databases striving to be comprehensive. Screens are still a theme in the current era and spatial transcriptomics is still used in untargeted searches for genes involved in development of model organisms, but with highly multiplexed technology, this can also be done for pathological and human tissues (Figure 6.4, Figure 6.5). Thanks to multiplexing, while mouse was the most popular species in the prequel era, in the current era, there are more studies on human tissues than those on mice and the vast majority of studies are on either humans or mice (Figure 6.4). Furthermore, there are datasets for a wider range of organs in mice in the current era (e.g. colon, liver, uterus, and etc.) than in the prequel era though there still is more interest in the brain (Figure 6.6, Figure 4.7).

Drosophila is no longer as commonly used in the current era (Figure 6.4). Whole mount smFISH has been applied *Drosophila* brains but without multiplexing [37], zebrafish embryos [38], and embryonic mouse organs [39]. For *Drosophila* tissue sections, while microdissection, smFISH, and ISS may be applied, the resolution of ST and Visium may be too low to discern sufficiently fine patterns in such a small model organism. Besides low resolution Tomo-seq along one body axis [40, 41], current era *Drosophila* datasets come from subcellular resolution technologies, such as smFISH on YFP trap lines (whole mount nervous system, not multiplexed) [37], *in situ* sequencing (retina) [42], ExSeq (embryo, might be whole mount with tissue clearing and expansion) [43], and Stereo-seq (whole embryo sectioned along the anterior-posterior axis) [44]. The reason why *Drosophila* is less popular may be that the most popular commercial technologies are unsuitable, as Visium (Figure 6.7) has too low a resolution and MERFISH is not whole mount.

Atlases have been made with current era technology, such as MERFISH [12], HybISS [34], ST [32], Visium [47], GeoMX DSP [48], and Slide-seq2 [49] described and analyzed with similar language to that of (WM)ISH atlases. Also as in the prequel era, the brain is still the most favored healthy organ (Figure 6.5, Figure 6.6). Among pathological tissues, breast tumors are the most used (Figure 6.5). Note that these anatograms only includes organs available in the R package `gganatogram`. Datasets from organs unavailable in the package are not shown. For metastases, the organ used for plotting here is the destination of metastases, so a liver metastasis of breast cancer would be plotted in the liver. More recently, in the wake of the SARS-CoV-2 pandemic, a number of studies using GeoMX Digital Spatial Profiler (DSP) to profile spatial transcriptomes of lungs of COVID victims have been published

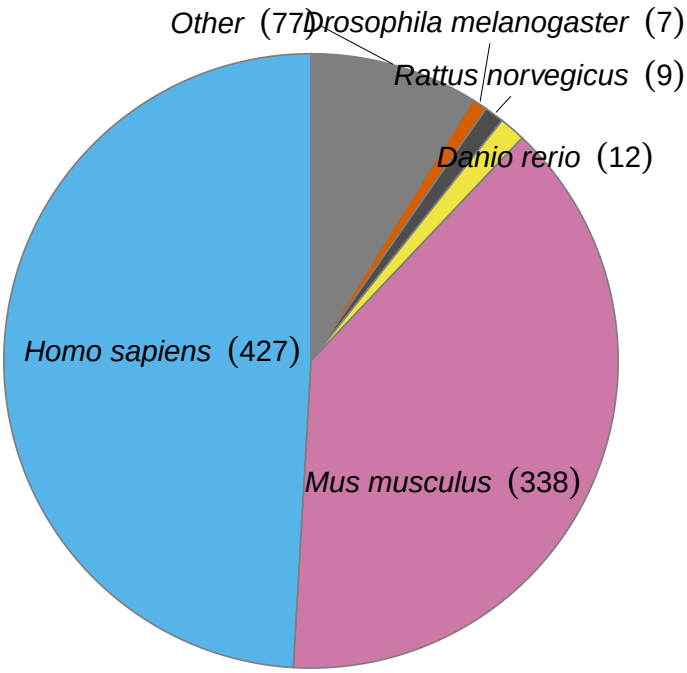

Figure 6.4: Number of publications per species.

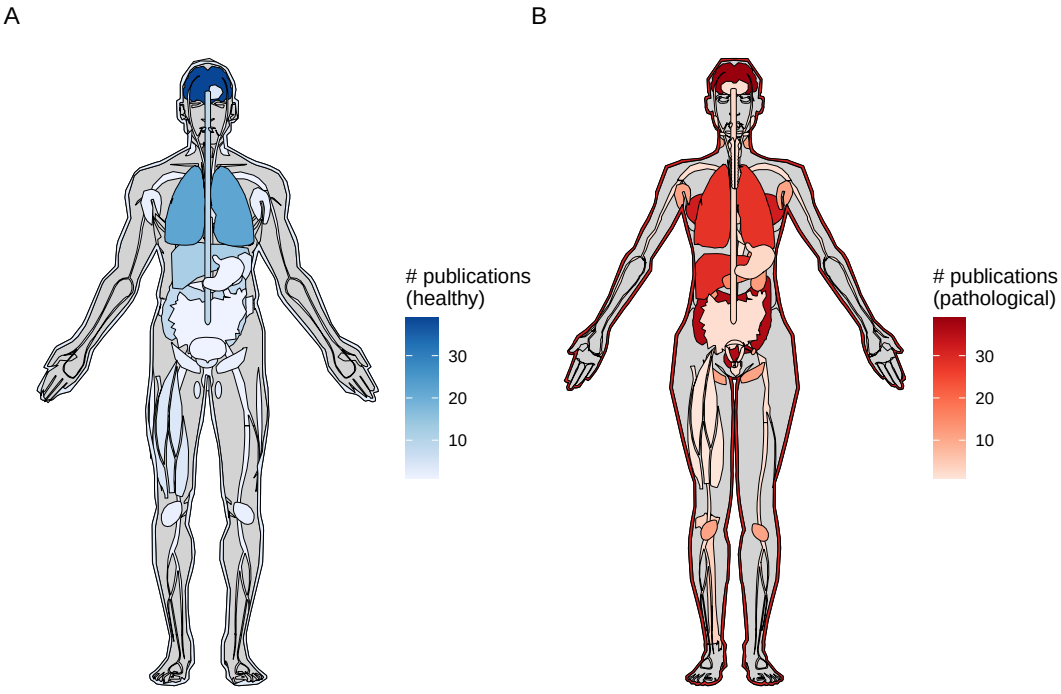

Figure 6.5: A) Number of publications for each healthy organ in human (male shown here, as there is no study on healthy female specific organs in humans at present). B) Number of publications for pathological organs in human (female shown here, but there are at least two studies on prostate cancer [45, 46]).

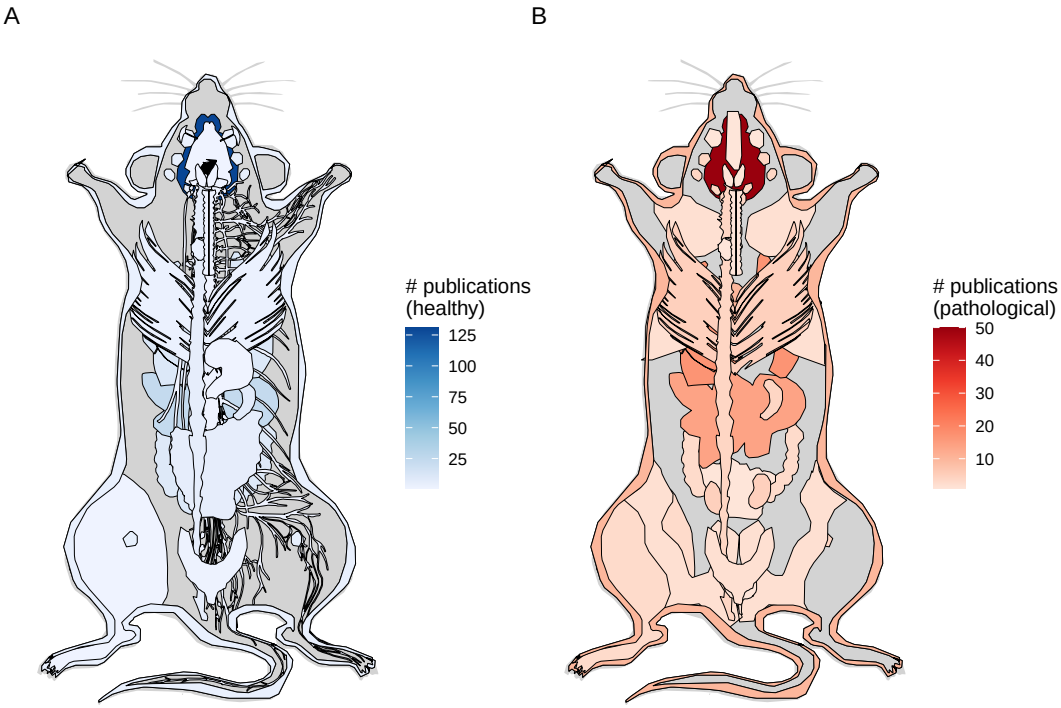

Figure 6.6: A) Number of publications per healthy organ in the mouse. B) Number of publications for pathological organs in mouse.

[50, 51, 48, 52].

However, unlike in the prequel era, in which older technologies were adapted to larger scale to produce the screens and atlases, the current era has another major theme—new techniques, due to the challenges to be discussed in the following sections; the number of new techniques published each year has grown steadily in the past few years (Figure 6.9). However, this difference might be due to bias in curation and change in culture. In the prequel era, very different enhancer and gene trap vectors were lumped together into enhancer or gene trap in our database, and there might have been many different early non-radioactive ISH protocols not included in our database because they were not used to profile a sufficiently large number of genes. Furthermore, in the current era, authors like to give techniques new names, making related techniques seem distinct rather than lumped together in a wider category like enhancer or gene trap.

While a few techniques other than LCM have become popular, such as ISS (2013), Tomo-seq (2013), MERFISH (2015), ST (late 2016), GeoMX DSP (2019, only showing transcriptomics studies here), and Visium (first preprint in 2020), most techniques never or rarely spread beyond their institutions of origin (Figures 6.7,

6.9). Furthermore, except for Visium and LCM, prequel (WM)ISH, enhancer trap, and gene trap have been used by more institutions than current era techniques (6.8). This might be because there has not been enough time for recently published new techniques to be implemented elsewhere, or if they have been implemented, there has not been enough time for the relevant studies to be published, or that there has been much less time for relatively new commercial techniques like MERFISH to spread to more institutions compared to (WM)ISH. Furthermore, usage of Visium and GeoMX DSP might have been spread by commercialization and core facilities and usage of Tomo-seq might have been spread by relative ease of implementation with standard lab equipment; implementing complex current era techniques that require custom built equipment such as custom fluidics systems independently may be more challenging, thus hampering their widespread adoption. This is analogous to a well-tested and fool proof commercial cake mix widely available at grocery stores that only calls for standard kitchen equipment such as the oven and the hand mixer as opposed to a cake recipe that is not only very complicated but also requires the home cook to build custom kitchen equipment. Even if instructions to assemble the custom equipment is available, most people would probably prefer to buy the pre-assembled product when feasible. The average home cook would most likely prefer the former to the latter. Having a core facility perform a procedure is like ordering a cake from a bakery, which is much easier than DIY trials and errors and building custom equipment.

Protocols of WMISH (as used in GEISHA) [53], ISH (as used in GenePaint and ABA) [36], Visium, and MERFISH [54] all have numerous steps. What (WM)ISH and Visium seem to have in common besides that they are widely adopted is that a significant part of the protocol is taken care of, by commercial automated systems (for (WM)ISH) or core facilities (for Visium), so there is less DIY hassle. Commercial automated ISH systems are commonly used by large scale (WM)ISH atlases. For example, GEISHA used the Abimed *In Situ* Pro [53], and GenePaint [55], ABA [56], and LungMap [57] used the Tecan EVO liquid handling platform (or its pre-commercial version), to automate ISH staining of numerous sections or embryos and genes. Several major institutions have core facilities that perform Visium [58, 59], and even if the core facility does not perform Visium as a whole, NGS core facilities are common. Furthermore, the Visium protocol does not require custom made equipment that cannot be purchased from 10X itself and major lab equipment companies such as Bio-Rad and VWR. Visium involves scanning the H&E image of the tissue section, which can be done by a histology core. As library preparation

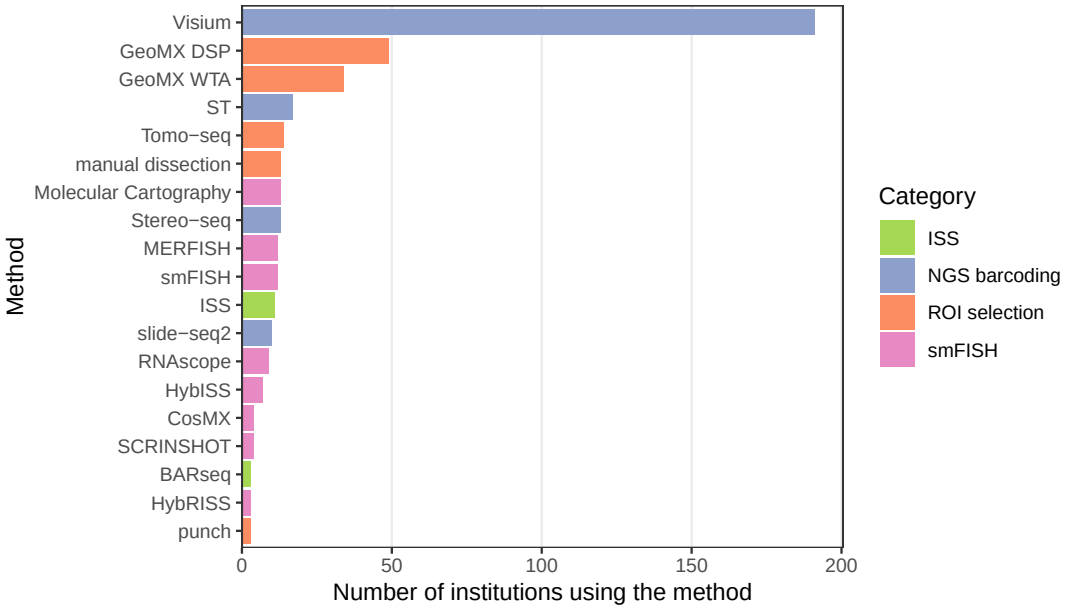

Figure 6.7: Techniques used by at least 3 institutions and the number of institutions that have used them.

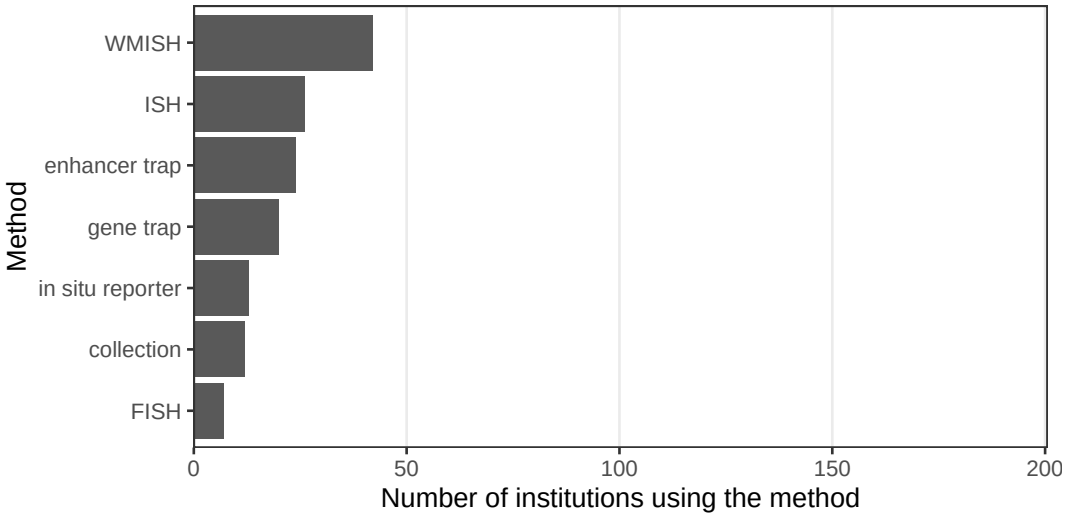

Figure 6.8: Prequel techniques used by at least 3 institutions and the number of institutions that have used them.

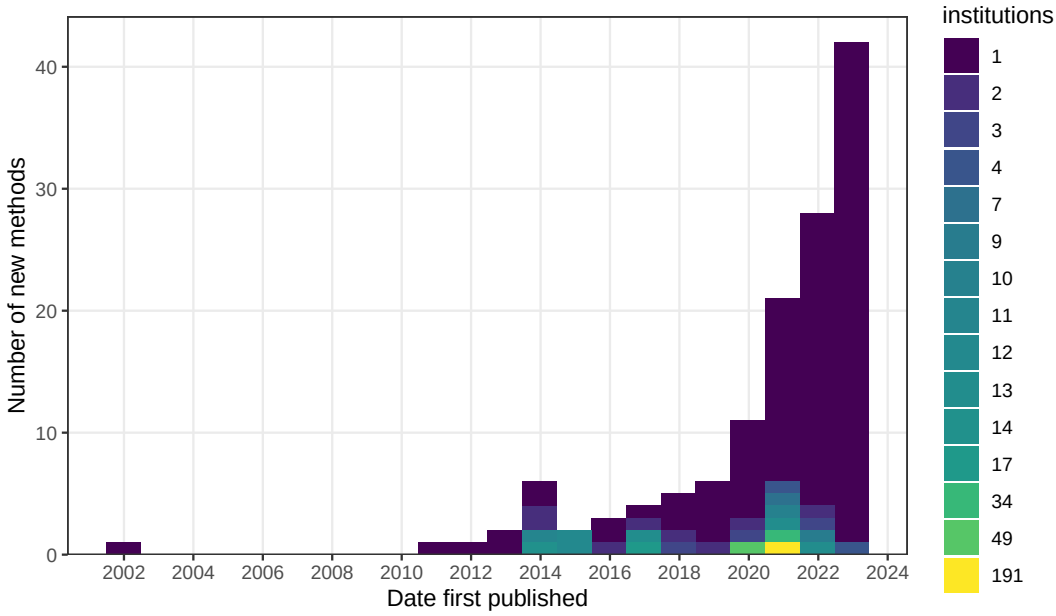

Figure 6.9: Number of new methods per year, colored by the number of institutions that have used the method.

of the forerunner of Visium, ST, can be automated [60], it would be reasonable to say that Visium library preparation can be automated. In contrast, the MERFISH protocol involves a custom built fluidics system to automate the imaging and liquid handling and long imaging time that might not be appropriate for a microscopy core facility. However, as MERFISH is getting commercialized by Vizgen and automated with the MERSCOPE product, it might become more widely adopted in the near future as the commercial package removes a lot of DIY hassle to independently implement MERFISH.

While in prequel (WM)ISH atlases, the images are themselves *the* data, current era data goes beyond visualization of gene expression in space. NGS based current era data has the sequencing reads in fastq files, which can be re-processed for RNA velocity and isoform analyses. The fastq files are often deposited in data repositories such as GEO and ENA, where they can be downloaded for re-processing. However, for some human data, to protect patients' privacy, the fastq files are not available or have controlled access. While the fastq files from around half of published papers for NGS based current era datasets are available in a data repository (Figure 6.10), the fastq files from most NGS based current era preprints are not available, especially the older preprints (Figure 6.11). Sometimes preprints state that the data will be deposited on GEO upon acceptance of the manuscript (e.g. [61]).

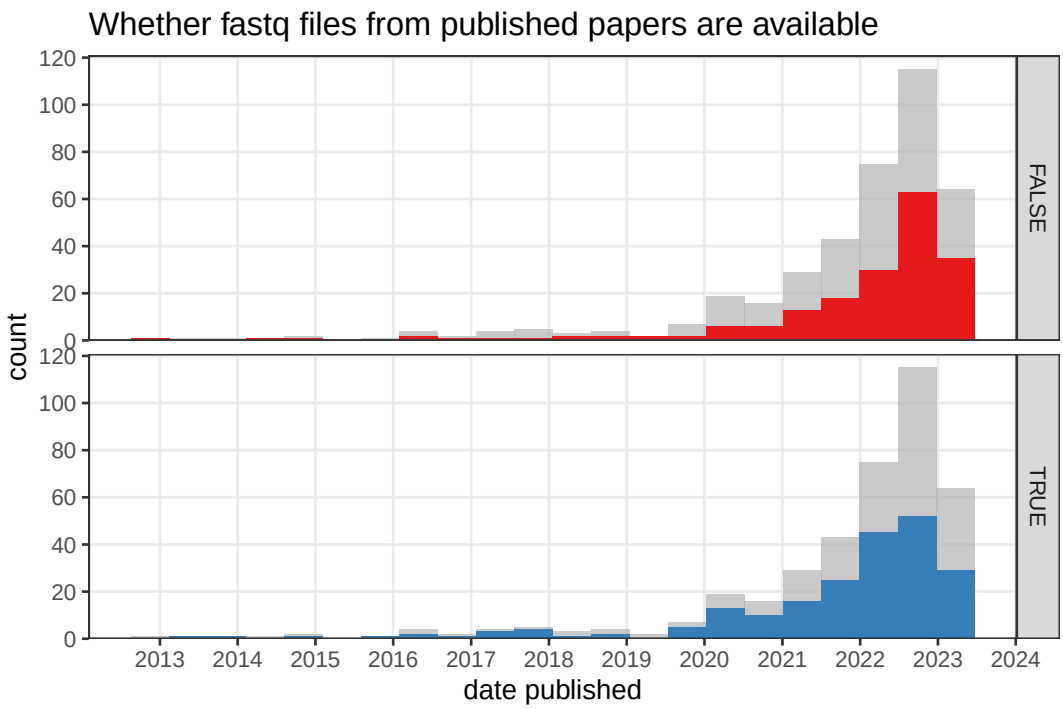

Figure 6.10: Whether fastq files from published NGS based papers (no preprints) are available on a public data repository such as GEO over time. Bin width is 180 days.

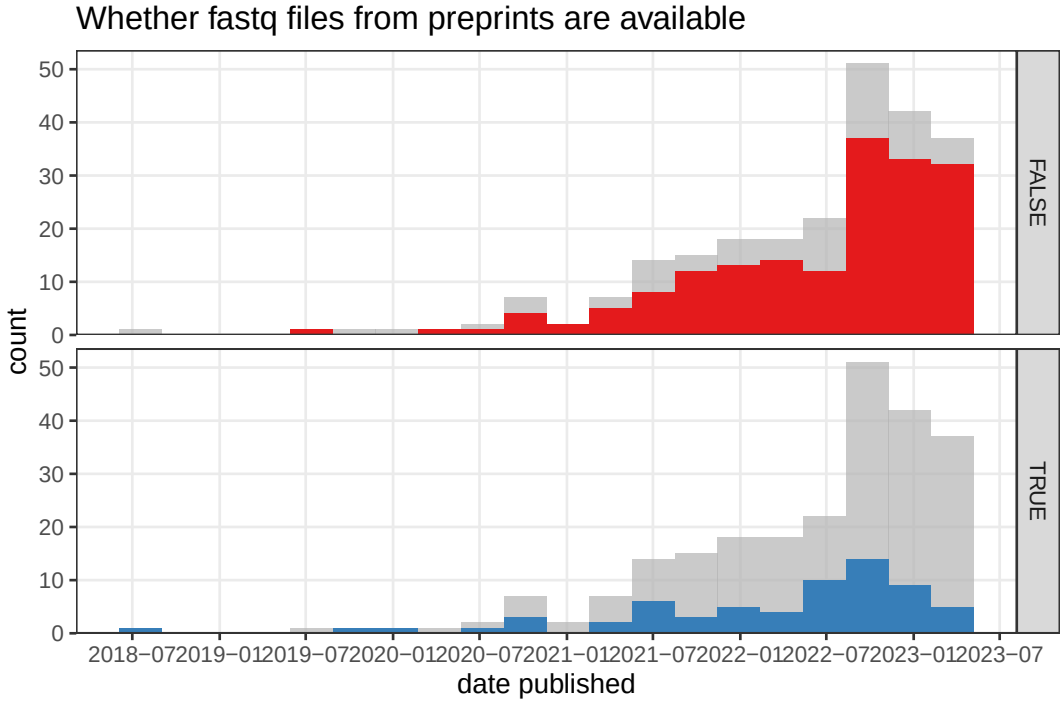

Figure 6.11: Whether fastq files from published NGS based preprints are available on a public data repository such as GEO over time. Bin width is 90 days.

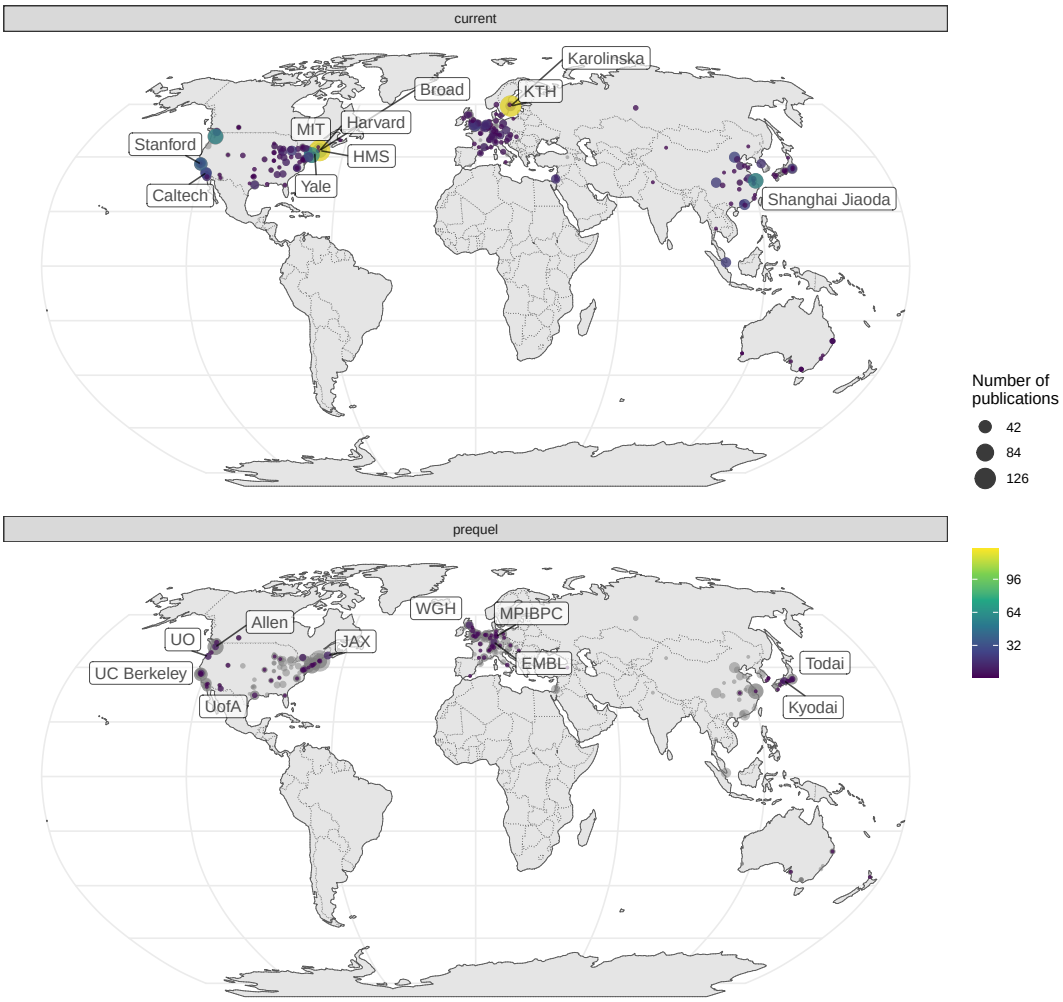

Figure 6.12: World map of institutions. Area of the point is proportional to the number of publications from that city. Gray points are sum of both prequel and current eras for each city. Top 10 institutions in each era are labeled.

Especially in the US, research in the current era tends to be more concentrated in a few elite institutions despite the mainstreaming of spatial transcriptomics to many less well-known institutions, while some top contributors in the prequel era were some less well-known institutions (Figure 6.12, 6.13). Among the top contributing institutions in the prequel era are those hosting databases, such as Allen Institute for ABA, University of Oregon (UO) for ZFIN, UC Berkeley and Lawrence Berkeley National Laboratory (LBL) for BDGP, University of Arizona (UofA) for GEISHA, Jackson Laboratory (JAX) for GXD, Western General Hospital (WGH) for EMAGE, and Kyoto University (Kyodai) for GHOST (Figure 6.13). By and large, in western Europe and northeast Asia, prequel and current era research was conducted in different institutions as well (Figure 6.14, Figure 6.15).

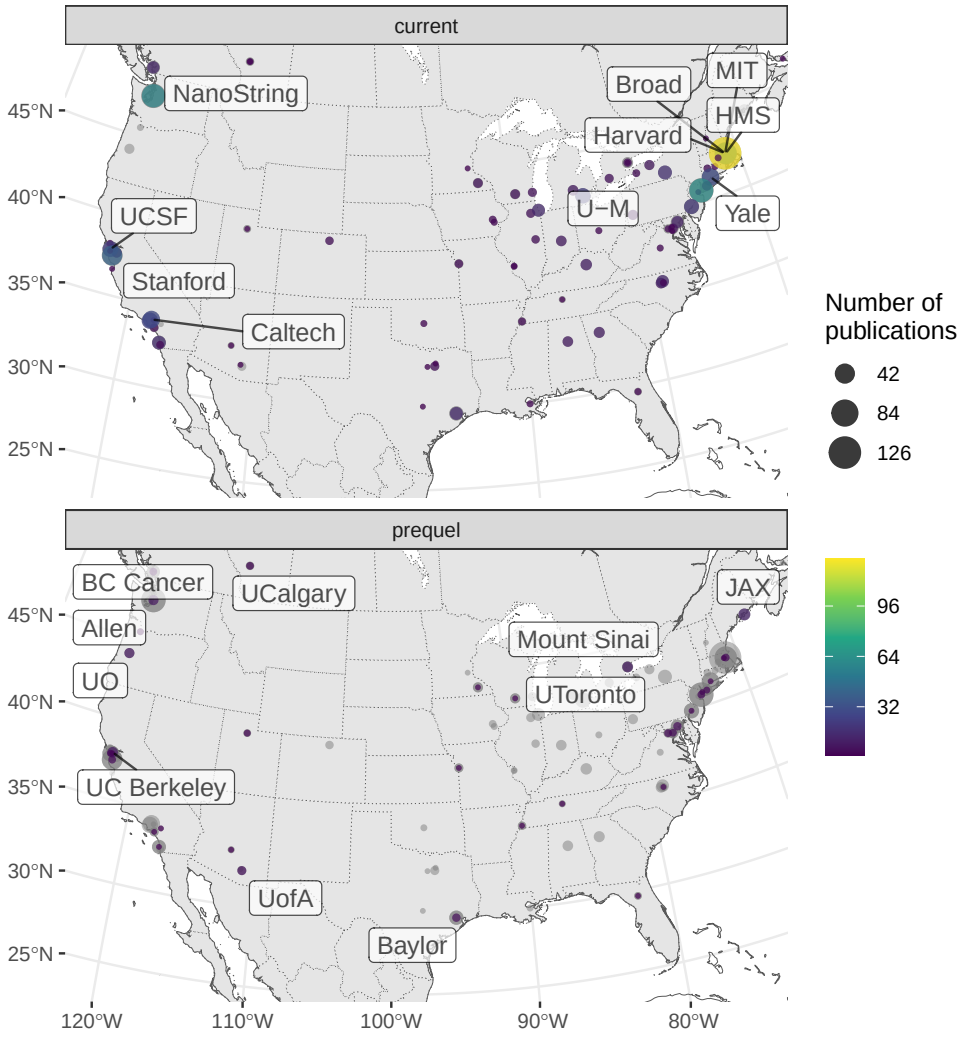

Figure 6.13: Map of institutions around continental US. Area of the point is proportional to the number of publications from that city. Gray points are sum of both prequel and current eras for each city. Top 10 institutions in each era are labeled.

6.3 Learning from the past

What can we learn from the history of the prequel era? We might be able to learn something from the past, as people in the past have come up with good ideas that have been mostly forgotten in the present era. An example of such an idea in the history of cycling is the 1930s network of at least 280 miles of cycleways separated from motor traffic in the UK, forgotten even by the Ministry of Transport itself; with the new wave of bike advocacy since the 1970s, there have been recent efforts to resurrect these old cycleways [62]. Furthermore, the past can illustrate what might happen next and what to do to get better outcome during similar developments at present, such as how the 1918 Spanish flu pandemic has been compared to the

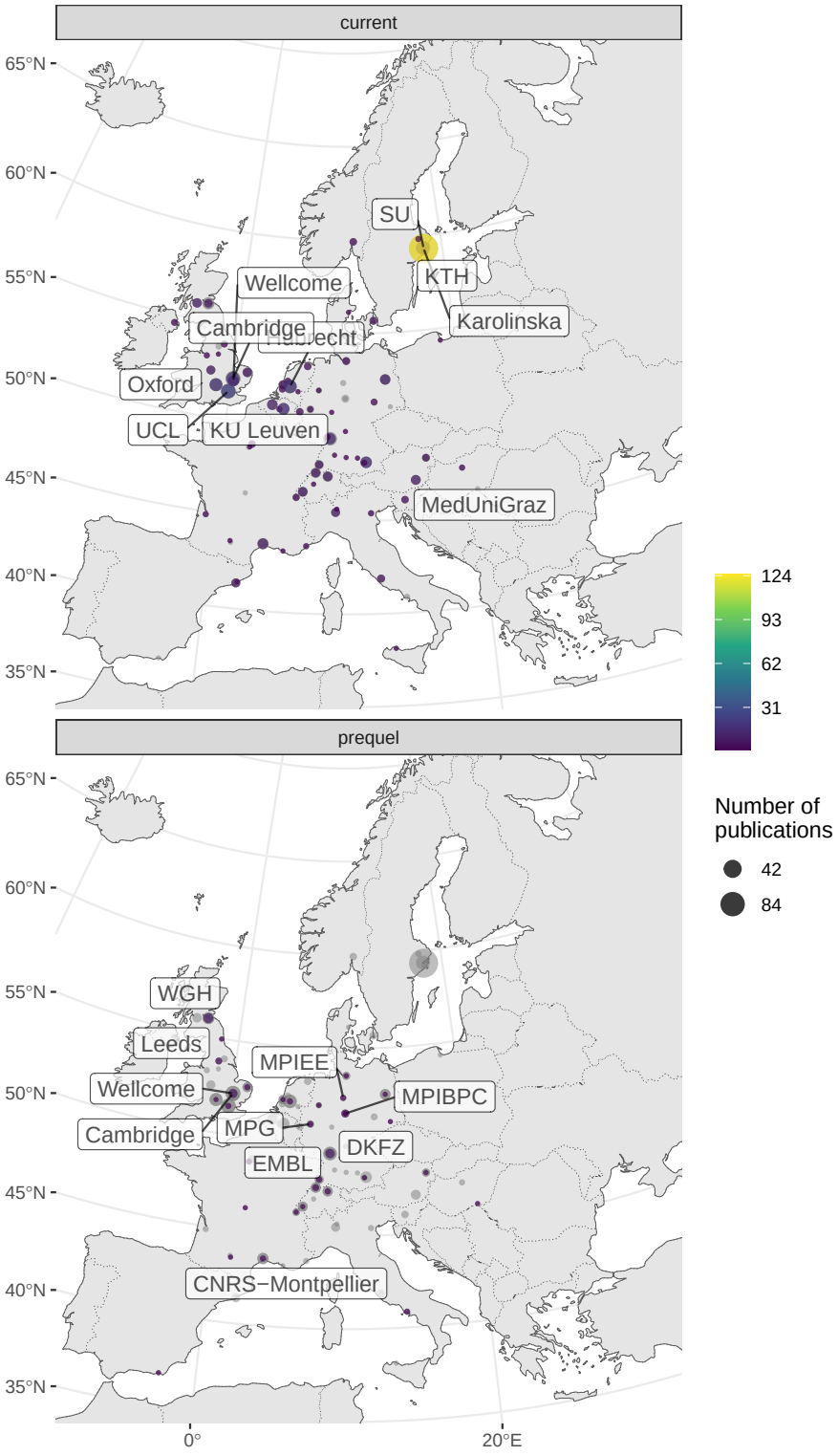

Figure 6.14: Map of institutions around western Europe. Area of the point is proportional to the number of publications from that city. Gray points are sum of both prequel and current eras for each city. Top 10 institutions in each era are labeled.

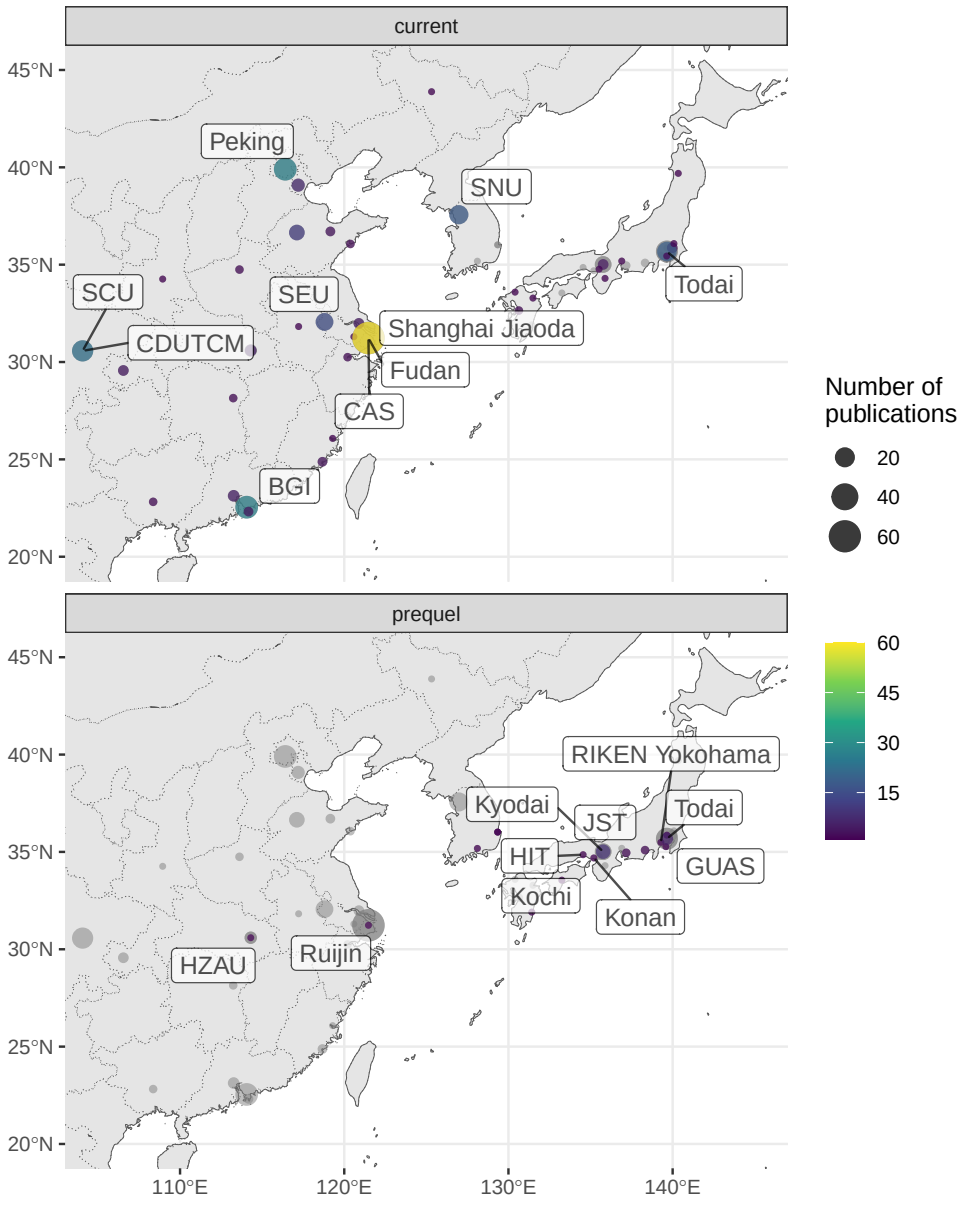

Figure 6.15: Map of institutions in northeast Asia. Area of the point is proportional to the number of publications from that city. Gray points are sum of both prequel and current eras for each city. Top 10 institutions in each era are labeled.

current COVID pandemic to point to strategies (e.g. [63, 64].

First, prequel (WM)ISH atlases by nature require thousands of animals to stain for thousands of genes, and often show photos from multiple animals stained for the same gene, sometimes showing variability in staining and morphology (especially in BDGP and GEISHA as the embryos are small and can be stained *en masse*), giving some qualitative sense of reproducibility of the staining and pattern and how generalizable a pattern seen in the atlas is to the wider population of the model organism. In contrast, current era datasets and atlases from model organisms tend to use much smaller numbers of animals thanks to multiplexing and cost and do not tend to discuss biological differences and reproducibility of results between the animals. For instance, in the Molecular Atlas of the Adult Mouse Brain [32], 3 male C57BL/6 mice were used. The online viewer of the Molecular Atlas shows gene expression in coronal sections from different mice all registered to the Allen CCF; adjacent spots from different mice sometimes have quite different expression of the same gene. However, such variability is not discussed in the paper. The MERFISH MOp atlas has 32 sections from each of the 2 mice used and reproducibility of results in the 2 mice is not discussed. The HybISS developing mouse atlas [34] only used one E10.5 mouse embryo.

Second, while there are databases for current era data, as discussed in Section 7.9, they do not provide the querying functionalities and systematic annotations of the ABA, EMAGE, and Eurexpress. While SpatialDB [65] provides easy access to and visualization of processed current era data from several different technologies, as of August 2021, SpatialDB does not seem to have updated since 2020 and does not contain new datasets. As of writing, the Human Cell Atlas (HCA) [66] has data from 5 Visium studies and at least one HybISS study [67]. The Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative - Cell Census Network (BICCN) [68] has data from MERFISH, osmFISH, seqFISH, and etc. While the studies can be queried, as of writing, in HCA and BICCN, unlike in the prequel atlases, genes can't be queried to open a webpage to show expression patterns in different data sources, nor can one search for other genes with similar expression patterns. Gene expression patterns are also not annotated. Given the massive volume of scRNA-seq and current era spatial data in HCA and BICCN, it would be more challenging to enable gene annotation, search, and comparison as this would involve analyzing and comparing hundreds of scRNA-seq datasets. However, for current era spatial data, for each organ of each species, the number

of datasets in the order of dozens per organ seems to be more manageable for such analyses and comparisons that would enable prequel database style gene queries (Figures 6.5, 6.6). Specifically, mouse brain data can be registered to the CCF to facilitate comparison between datasets, studies, and subjects within one study. Furthermore, the massive volume of quantitative current era may be used to refine prequel gene annotations such as ontologies of developmental stages and anatomical regions.

Third, while most extant prequel (WM)ISH atlases, such as ABA, LungMAP, and GUDMAP, are hosted in online databases for query and view, most current era datasets—including those that claim to be atlases—cannot be viewed online, which can be useful in cases such as to easily look up more information about genome wide association study (GWAS) candidate genes associated with phenotype or expression quantitative trait loci (eQTL) and about differentially expressed (DE) genes from non-spatial transcriptomic or proteomic studies. Even if comparison and analysis of current era data for gene annotation and query is challenging, a web portal that searches multiple datasets for gene expression patterns, merely linking to the gene expression plots in the original data visualization websites of the datasets, would still be helpful. Besides datasets in SpatialDB, some current era datasets can be queried and visualized online, plotting gene expression values in space (dataset description is linked to the online data visualization portal), such as zebrafish Tomo-seq [1], mid-gastrula mouse embryo Geo-seq [29], mouse cortex osmFISH [11], ST molecular atlas of the adult mouse brain [32], and ST and Cartana ISS for Alzheimer's disease [33]. However, many other current era atlases do not provide online visualization, such as the MERFISH MOp atlas [12] and the Visium breast cancer atlas [47].

Finally, what if another revolution in spatial genomics comes? What in the current era will be remembered like the ABA, and what will be forgotten? We may take clues from the impact of the ABA in the current era and how other prequel atlases seem to be forgotten. To recap, the ABA is still relevant in the current era because of its comprehensiveness, quantification of ISH staining, registration to the CCF, and API to programatically query the database. Some of the hallmarks of the current era are quantitative data and multiplexing. With quantification and the CCF, ABA data resembles such hallmarks, although ABA's quantification and CCF began in 2006, long before the current era really took off around the mid 2010s. The API makes the data easier to download for analyses than the images from (WM)ISH databases that don't have APIs. The comprehensiveness makes the ABA relevant

to qualitative comparisons to current era results. Furthermore, the Allen Institute itself is participating in the current era by not only producing bulk and single-cell RNA-seq data for the atlas but also hosting the data catalog for the BICCN. In contrast, we are unaware of other prequel atlas consortia, such as EMAGE and BDGP, participating in the current era. We cannot foresee what the next revolution would be like. However, from ABA, perhaps we may say that for data, resources, and institutions from the current era to not to be forgotten when the next era comes, they should resemble or adapt to the hallmarks of the next era.

References

1. Junker JP, Noël ES, Guryev V, Peterson KA, Shah G, Huisken J, McMahon AP, Berezikov E, Bakkers J, and Van Oudenaarden A. Genome-wide RNA Tomography in the Zebrafish Embryo. Cell 2014. DOI: 10.1016/j.cell.2014.09.038

2. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, and Church GM. Highly Multiplexed Subcellular RNA Sequencing in Situ. Science 2014 Mar; 343:1360 LP –1363. DOI: 10.1126/science.1250212. Available from: http://science.sciencemag.org/content/343/6177/1360.abstract

3. Brown VM, Ossadtchi A, Khan AH, Yee S, Lacan G, Melega WP, Cherry SR, Leahy RM, and Smith DJ. Multiplex Three-Dimensional Brain Gene Expression Mapping in a Mouse Model of Parkinson's Disease. Genome Research 2002 May; 12:868–84. DOI: 10.1101/gr.229002. Available from: http://www.genome.org/cgi/doi/10.1101/gr.229002

4. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, Mollbrink A, Linnarsson S, Codeluppi S, Borg Å, Pontén F, Costea PI, Sahlén P, Mulder J, Bergmann O, Lundeberg J, and Frisén J. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 2016 Jul; 353:78–82. DOI: 10.1126/science.aaf2403. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aaf2403

5. Luo L, Salunga RC, Guo H, Bittner A, Joy K, Galindo JE, Xiao H, Rogers KE, Wan JS, Jackson MR, and Erlander MG. Gene expression profiles of laser-captured adjacent neuronal subtypes. Nature Medicine 1999 Jan; 5:117–22. DOI: 10.1038/4806. Available from: http://www.nature.com/articles/nm0199_117

6. Lubeck E and Cai L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. Nature Methods 2012 9:7 2012 Jun; 9:743–8.

DOI: 10.1038/nmeth.2069. Available from: https://www.nature.com/articles/nmeth.2069

7. Chen KH, Boettiger AN, Moffitt JR, Wang S, and Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 2015. DOI: 10.1126/science.aaa6090

8. Guroff M. American Drivers Have Bicyclists to Thank for a Smooth Ride to Work. 2016. Available from: https://www.smithsonianmag.com/travel/american-drivers-thank-bicyclists-180960399/

9. Reid C. 19th century cyclists paved the way for modern motorists' roads. 2011

10. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, Nolan GP, Bava FA, and Deisseroth K. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science 2018 Jul; 361:eaat5691. DOI: 10.1126/science.aat5691. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aat5691

11. Codeluppi S, Borm LE, Zeisel A, La Manno G, Lunteren JA van, Svensson CI, and Linnarsson S. Spatial organization of the somatosensory cortex revealed by osmFISH. Nature Methods 2018. DOI: 10.1038/s41592-018-0175-z

12. Zhang M, Eichhorn SW, Zingg B, Yao Z, Zeng H, Dong H, and Zhuang X. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics. bioRxiv 2020 Jan :2020.06.04.105700. DOI: 10.1101/2020.06.04.105700. Available from: http://biorxiv.org/content/early/2020/06/05/2020.06.04.105700.abstract

13. Dou J, Liang S, Mohanty V, Cheng X, Kim S, Choi J, Li Y, Rezvani K, Chen R, and Chen K. Unbiased integration of single cell multi-omics data. bioRxiv 2020 Jan :2020.12.11.422014. DOI: 10.1101/2020.12.11.422014. Available from: http://biorxiv.org/content/early/2020/12/11/2020.12.11.422014.abstract

14. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, and Satija R. Comprehensive Integration of Single-Cell Data. Cell 2019 Jun; 177:1888–902. DOI: 10.1016/j.cell.2019.05.031. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867419305598

15. Cang Z and Nie Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. Nature Communications 2020; 11:2084. DOI: 10.1038/s41467-020-15968-5. Available from: https://doi.org/10.1038/s41467-020-15968-5

16. Abdelaal T, Mourragui S, Mahfouz A, and Reinders MJT. SpaGE: Spatial Gene Enhancement using scRNA-seq. Nucleic Acids Research 2020 Oct; 48:e107–e107. DOI: 10.1093/nar/gkaa740. Available from: https://doi.org/10.1093/nar/gkaa740

17. Okochi Y, Sakaguchi S, Nakae K, Kondo T, and Naoki H. Model-based prediction of spatial gene expression via generative linear mapping. Nature Communications 2021; 12:3731. DOI: 10.1038/s41467-021-24014-x. Available from: https://doi.org/10.1038/s41467-021-24014-x

18. Baker D, Al-Naggar IM, Sivajothi S, Flynn WF, Amiri A, Luo D, Hardy CC, Kuchel GA, Smith PP, and Robson P. A Cellular Reference Resource for the Mouse Urinary Bladder. bioRxiv 2021 Jan :2021.09.20.461121. DOI: 10.1101/2021.09.20.461121. Available from: http://biorxiv.org/content/early/2021/09/23/2021.09.20.461121.abstract

19. Sharma G, Colantuoni C, Goff LA, Fertig EJ, and Stein-O'Brien G. projectR: an R/Bioconductor package for transfer learning via PCA, NMF, correlation and clustering. Bioinformatics 2020 Jun; 36:3592–3. DOI: 10.1093/bioinformatics/btaa183. Available from: https://doi.org/10.1093/bioinformatics/btaa183

20. Tanevski J, Nguyen T, Truong B, Karaiskos N, Ahsen ME, Zhang X, Shu C, Xu K, Liang X, Hu Y, Pham HVV, Xiaomei L, Le TD, Tarca AL, Bhatti G, Romero R, Karathanasis N, Loher P, Chen Y, Ouyang Z, Mao D, Zhang Y, Zand M, Ruan J, Hafemeister C, Qiu P, Tran D, Nguyen T, Gabor A, Yu T, Guinney J, Glaab E, Krause R, Banda P, Stolovitzky G, Rajewsky N, Saez-Rodriguez J, and Meyer P. Gene selection for optimal prediction of cell position in tissues from single-cell transcriptomics data. Life Science Alliance 2020 Nov; 3:e202000867. DOI: 10.26508/lsa.202000867. Available from: http://www.life-science-alliance.org/content/3/11/e202000867.abstract

21. Alonso AM, Carrea A, and Diambra L. Prediction of cell position using single-cell transcriptomic data: an iterative procedure [version 2; peer review: 2 approved]. F1000Research 2020; 8. DOI: 10.12688/f1000research.20715.2. Available from: http://openr.es/jqr

22. Pham D, Tan X, Xu J, Grice LF, Lam PY, Raghubar A, Vukovic J, Ruitenberg MJ, and Nguyen Q. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. bioRxiv 2020 Jan :2020.05.31.125658. DOI: 10.1101/2020.05.31.125658. Available from: http://biorxiv.org/content/early/2020/05/31/2020.05.31.125658.abstract

23. Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, Zwan J van der, Häring M, Braun E, Borm LE, La Manno G, Codeluppi S, Furlan A, Lee K, Skene N, Harris KD, Hjerling-Leffler J, Arenas E, Ernfors P, Marklund U, and Linnarsson S. Molecular Architecture of the Mouse Nervous System.

Cell 2018 Aug; 174:999–1014. DOI: 10.1016/j.cell.2018.06.021. Available from: https://doi.org/10.1016/j.cell.2018.06.021

24. Fleck JS, Sanchís-Calleja F, He Z, Santel M, Boyle MJ, Camp JG, and Treutlein B. Resolving organoid brain region identities by mapping single-cell genomic data to reference atlases. Cell Stem Cell 2021; 28:1148–59. DOI: https://doi.org/10.1016/j.stem.2021.02.015. Available from: https://www.sciencedirect.com/science/article/pii/S1934590921000655

25. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, and Macosko EZ. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell 2019 Jun; 177:1873–87. DOI: 10.1016/j.cell.2019.05.006. Available from: https://doi.org/10.1016/j.cell.2019.05.006

26. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh Pr, and Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. Nature Methods 2019; 16:1289–96. DOI: 10.1038/s41592-019-0619-0. Available from: https://doi.org/10.1038/s41592-019-0619-0

27. Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z, Sanger A, Tokcan N, Vanderburg CR, Segerstolpe A, Zhang M, Avraham-Davidi I, Vickovic S, Nitzan M, Ma S, Buenrostro J, Brown NB, Fanelli D, Zhuang X, Macosko EZ, and Regev A. Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with Tangram. bioRxiv 2020 Jan :2020.08.29.272831. DOI: 10.1101/2020.08.29.272831. Available from: http://biorxiv.org/content/early/2020/09/24/2020.08.29.272831.abstract

28. Armit C, Richardson L, Venkataraman S, Graham L, Burton N, Hill B, Yang Y, and Baldock RA. eMouseAtlas: An atlas-based resource for understanding mammalian embryogenesis. Developmental Biology 2017 Mar; 423:1–11. DOI: 10.1016/j.ydbio.2017.01.023. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0012160616308582

29. Peng G, Suo S, Chen J, Chen W, Liu C, Yu F, Wang R, Chen S, Sun N, Cui G, Song L, Tam PP, Han JDJ, and Jing N. Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. Developmental Cell 2016 Mar; 36:681–97. DOI: 10.1016/j.devcel.2016.02.020. Available from: http://dx.doi.org/10.1016/j.devcel.2016.02.020

30. Liu Y, Yang M, Deng Y, Su G, Enninful A, Guo CC, Tebaldi T, Zhang D, Kim D, Bai Z, Norris E, Pan A, Li J, Xiao Y, Halene S, and Fan R. High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. Cell 2020; 183:1665–81. DOI: https://doi.org/10.1016/j.

cell.2020.10.026. Available from: https://www.sciencedirect.com/science/article/pii/S0092867420313908

31. Wang Q, Ding SL, Li Y, Royall J, Feng D, Lesnar P, Graddis N, Naeemi M, Facer B, Ho A, Dolbeare T, Blanchard B, Dee N, Wakeman W, Hirokawa KE, Szafer A, Sunkin SM, Oh SW, Bernard A, Phillips JW, Hawrylycz M, Koch C, Zeng H, Harris JA, and Ng L. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. Cell 2020 May; 181:936–53. DOI: 10.1016/j.cell.2020.04.007. Available from: https://doi.org/10.1016/j.cell.2020.04.007

32. Ortiz C, Navarro JF, Jurek A, Märtin A, Lundeberg J, and Meletis K. Molecular atlas of the adult mouse brain. Science Advances 2020 Jun; 6:eabb3446. DOI: 10.1126/sciadv.abb3446. Available from: www.brain-map.org%20https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.abb3446

33. Chen WT, Lu A, Craessaerts K, Pavie B, Sala Frigerio C, Corthout N, Qian X, Laláková J, Kühnemund M, Voytyuk I, Wolfs L, Mancuso R, Salta E, Balusu S, Snellinx A, Munck S, Jurek A, Fernandez Navarro J, Saido TC, Huitinga I, Lundeberg J, Fiers M, and De Strooper B. Spatial Transcriptomics and In Situ Sequencing to Study Alzheimer's Disease. Cell 2020 Jul; 0. DOI: 10.1016/j.cell.2020.06.038. Available from: http://www.cell.com/article/S0092867420308151/fulltext

34. La Manno G, Siletti K, Furlan A, Gyllborg D, Vinsland E, Mossi Albiach A, Mattsson Langseth C, Khven I, Lederer AR, Dratva LM, Johnsson A, Nilsson M, Lönnerberg P, and Linnarsson S. Molecular architecture of the developing mouse brain. Nature 2021; 596:92–6. DOI: 10.1038/s41586-021-03775-x. Available from: https://doi.org/10.1038/s41586-021-03775-x

35. Diez-Roux G, Banfi S, Sultan M, Geffers L, Anand S, Rozado D, Magen A, Canidio E, Pagani M, Peluso I, Lin-Marq N, Koch M, Bilio M, Cantiello I, Verde R, De Masi C, Bianchi SA, Cicchini J, Perroud E, Mehmeti S, Dagand E, Schrinner S, Nürnberger A, Schmidt K, Metz K, Zwingmann C, Brieske N, Springer C, Hernandez AM, Herzog S, Grabbe F, Sieverding C, Fischer B, Schrader K, Brockmeyer M, Dettmer S, Helbig C, Alunni V, Battaini MA, Mura C, Henrichsen CN, Garcia-Lopez R, Echevarria D, Puelles E, Garcia-Calero E, Kruse S, Uhr M, Kauck C, Feng G, Milyaev N, Ong CK, Kumar L, Lam M, Semple CA, Gyenesei A, Mundlos S, Radelof U, Lehrach H, Sarmientos P, Reymond A, Davidson DR, Dollé P, Antonarakis SE, Yaspo ML, Martinez S, Baldock RA, Eichele G, and Ballabio A. A High-Resolution Anatomical Atlas of the Transcriptome in the Mouse Embryo. PLoS Biology 2011 Jan; 9. Ed. by Barsh GS:e1000582. DOI: 10.1371/journal.pbio.1000582. Available from: https://dx.plos.org/10.1371/journal.pbio.1000582

36. Yaylaoglu MB, Titmus A, Visel A, Alvarez-Bolado G, Thaller C, and Eichele G. Comprehensive expression atlas of fibroblast growth factors and their receptors generated by a novel robotic in situ hybridization platform. Developmental Dynamics 2005 Oct; 234:371–86. DOI: 10.1002/dvdy.20441. Available from: http://doi.wiley.com/10.1002/dvdy.20441

37. Titlow JS, Kiourlappou M, Palanca A, Lee JY, Gala DS, Ennis D, Yu JJS, Young FL, Miguel D, Pinto S, Garforth S, Francis HS, Strivens F, Mulvey H, Dallman-Porter A, Thornton S, Arman D, Järvelin AI, Thompson MK, Kounatidis I, Parton RM, Taylor S, and Davis I. Systematic analysis of YFP gene traps reveals common discordance between mRNA and protein across the nervous system. bioRxiv 2022 Mar :2022.03.21.485142. DOI: 10.1101/2022.03.21.485142. Available from: https://www.biorxiv.org/content/10.1101/2022.03.21.485142v2%20https://www.biorxiv.org/content/10.1101/2022.03.21.485142v2.abstract

38. Oka Y and Sato TN. Whole-mount single molecule FISH method for zebrafish embryo. eng. Scientific reports 2015 Feb; 5:8571. DOI: 10.1038/srep08571. Available from: https://pubmed.ncbi.nlm.nih.gov/25711926%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4339797/

39. Wang C, Lu T, Emanuel G, Babcock HP, and Zhuang X. Imaging-based pooled CRISPR screening reveals regulators of lncRNA localization. Proceedings of the National Academy of Sciences of the United States of America 2019. DOI: 10.1073/pnas.1903808116

40. Combs PA and Eisen MB. Sequencing mRNA from Cryo-Sliced Drosophila Embryos to Determine Genome-Wide Spatial Patterns of Gene Expression. PLoS ONE 2013 Aug; 8. Ed. by Jennings B:e71820. DOI: 10.1371/journal.pone.0071820. Available from: https://dx.plos.org/10.1371/journal.pone.0071820

41. Combs PA and Fraser HB. Spatially varying cis-regulatory divergence in Drosophila embryos elucidates cis-regulatory logic. PLOS Genetics 2018 Nov; 14. Ed. by Desplan C:e1007631. DOI: 10.1371/journal.pgen.1007631. Available from: https://dx.plos.org/10.1371/journal.pgen.1007631

42. Fürth D, Hatini V, and Lee JH. In Situ Transcriptome Accessibility Sequencing (INSTA-seq). bioRxiv 2019 Jan :722819. DOI: 10.1101/722819. Available from: http://biorxiv.org/content/early/2019/08/06/722819.abstract

43. Alon S, Goodwin DR, Sinha A, Wassie AT, Chen F, Daugharthy ER, Bando Y, Kajita A, Xue AG, Marrett K, Prior R, Cui Y, Payne AC, Yao CC, Suk HJ, Wang R, Yu CC, Tillberg P, Reginato P, Pak N, Liu S, Punthambaker S, Iyer EP, Kohman RE, Miller JA, Lein ES, Lako A, Cullen N, Rodig S, Helvie K, Abravanel DL, Wagle N, Johnson BE, Klughammer

J, Slyper M, Waldman J, Jané-Valbuena J, Rozenblatt-Rosen O, Regev A, Church GM, Marblestone AH, and Boyden ES. Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. Science 2021 Jan; 371. DOI: 10.1126/SCIENCE.AAX2656/SUPPL{_}FILE/AAX2656{_}TABLESS1-S6ANDS9-S14.XLSX. Available from: https://www.science.org/doi/10.1126/science.aax2656

44. Wang M, Hu Q, Lv T, Wang Y, Lan Q, Xiang R, Tu Z, Wei Y, Han K, Shi C, Guo J, Liu C, Yang T, Du W, An Y, Cheng M, Xu J, Lu H, Li W, Zhang S, Chen A, Chen W, Li Y, Wang X, Xu X, Hu Y, and Liu L. High-resolution 3D spatiotemporal transcriptomic maps of developing Drosophila embryos and larvae. Developmental Cell 2022 May; 57:1271–83. DOI: 10.1016/J.DEVCEL.2022.04.006

45. Burgess DJ. Spatial transcriptomics coming of age. Nature Reviews Genetics 2019 Jun; 20:317–7. DOI: 10.1038/s41576-019-0129-z. Available from: https://www.nature.com/articles/s41576-019-0129-z

46. Brady L, Kriner M, Coleman I, Morrissey C, Roudier M, True LD, Gulati R, Plymate SR, Zhou Z, Birditt B, Meredith R, Geiss G, Hoang M, Beechem J, and Nelson PS. Inter- and intra-tumor heterogeneity of metastatic prostate cancer determined by digital spatial gene expression profiling. Nature Communications 2021; 12:1426. DOI: 10.1038/s41467-021-21615-4. Available from: https://doi.org/10.1038/s41467-021-21615-4

47. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, Thennavan A, Wang C, Torpy JR, Bartonicek N, Wang T, Larsson L, Kaczorowski D, Weisenfeld NI, Uytingco CR, Chew JG, Bent ZW, Chan CL, Gnanasambandapillai V, Dutertre CA, Gluch L, Hui MN, Beith J, Parker A, Robbins E, Segara D, Cooper C, Mak C, Chan B, Warrier S, Ginhoux F, Millar E, Powell JE, Williams SR, Liu XS, O'Toole S, Lim E, Lundeberg J, Perou CM, and Swarbrick A. A single-cell and spatially resolved atlas of human breast cancers. Nature Genetics 2021; 53:1334–47. DOI: 10.1038/s41588-021-00911-1. Available from: https://doi.org/10.1038/s41588-021-00911-1

48. Delorey TM et al. COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. Nature 2021; 595:107–13. DOI: 10.1038/s41586-021-03570-8. Available from: https://doi.org/10.1038/s41586-021-03570-8

49. Lake BB, Menon R, Winfree S, Hu Q, Ferreira RM, Kalhor K, Barwinska D, Otto EA, Ferkowicz M, Diep D, Plongthongkum N, Knoten A, Urata S, Naik AS, Eddy S, Zhang B, Wu Y, Salamon D, Williams JC, Wang X, Balderrama KS, Hoover P, Murray E, Vijayan A, Chen F, Waikar SS, Rosas S, Wilson FP, Palevsky PM, Kiryluk K, Sedor JR, Toto RD, Parikh C, Kim EH, Macosko EZ, Kharchenko PV, Gaut JP, Hodgin JB, Eadon MT, Dagher PC, El-Achkar TM, Zhang K, Kretzler M, Jain S, and ftK consortium for the. An atlas of

healthy and injured cell states and niches in the human kidney. bioRxiv 2021 Jan :2021.07.28.454201. DOI: 10.1101/2021.07.28.454201. Available from: http://biorxiv.org/content/early/2021/07/29/2021.07.28.454201.abstract

50. Park J, Foox J, Hether T, Danko D, Warren S, Kim Y, Reeves J, Butler DJ, Mozsary C, Rosiene J, Shaiber A, Afshinnekoo E, MacKay M, Bram Y, Chandar V, Geiger H, Craney A, Velu P, Melnick AM, Hajirasouliha I, Beheshti A, Taylor D, Saravia-Butler A, Singh U, Wurtele ES, Schisler J, Fennessey S, Corvelo A, Zody MC, Germer S, Salvatore S, Levy S, Wu S, Tatonetti N, Shapira S, Salvatore M, Loda M, Westblade LF, Cushing M, Rennert H, Kriegel AJ, Elemento O, Imielinski M, Borczuk AC, Meydan C, Schwartz RE, and Mason CE. Systemic Tissue and Cellular Disruption from SARS-CoV-2 Infection revealed in COVID-19 Autopsies and Spatial Omics Tissue Maps. bioRxiv 2021 Jan :2021.03.08.434433. DOI: 10.1101/2021.03.08.434433. Available from: http://biorxiv.org/content/early/2021/03/09/2021.03.08.434433.abstract

51. Butler D, Mozsary C, Meydan C, Foox J, Rosiene J, Shaiber A, Danko D, Afshinnekoo E, MacKay M, Sedlazeck FJ, Ivanov NA, Sierra M, Pohle D, Zietz M, Gisladottir U, Ramlall V, Sholle ET, Schenck EJ, Westover CD, Hassan C, Ryon K, Young B, Bhattacharya C, Ng DL, Granados AC, Santos YA, Servellita V, Federman S, Ruggiero P, Fungtammasan A, Chin CS, Pearson NM, Langhorst BW, Tanner NA, Kim Y, Reeves JW, Hether TD, Warren SE, Bailey M, Gawrys J, Meleshko D, Xu D, Couto-Rodriguez M, Nagy-Szakal D, Barrows J, Wells H, O'Hara NB, Rosenfeld JA, Chen Y, Steel PAD, Shemesh AJ, Xiang J, Thierry-Mieg J, Thierry-Mieg D, Iftner A, Bezdan D, Sanchez E, Campion TR, Sipley J, Cong L, Craney A, Velu P, Melnick AM, Shapira S, Hajirasouliha I, Borczuk A, Iftner T, Salvatore M, Loda M, Westblade LF, Cushing M, Wu S, Levy S, Chiu C, Schwartz RE, Tatonetti N, Rennert H, Imielinski M, and Mason CE. Shotgun transcriptome, spatial omics, and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification, and drug interactions. Nature Communications 2021; 12:1660. DOI: 10.1038/s41467-021-21361-7. Available from: https://doi.org/10.1038/s41467-021-21361-7

52. Margaroli C, Benson P, Sharma NS, Madison MC, Robison SW, Arora N, Ton K, Liang Y, Zhang L, Patel RP, and Gaggar A. Spatial mapping of SARS-CoV-2 and H1N1 Lung Injury Identifies Differential Transcriptional Signatures. eng. Cell reports. Medicine 2021 Mar :100242. DOI: 10.1016/j.xcrm.2021.100242. Available from: https://pubmed.ncbi.nlm.nih.gov/33778787%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7985929/

53. Bell GW, Yatskievych TA, and Antin PB. GEISHA, a whole-mount in situ hybridization gene expression screen in chicken embryos. Developmental

Dynamics 2004 Mar; 229:677–87. DOI: 10.1002/dvdy.10503. Available from: http://doi.wiley.com/10.1002/dvdy.10503

54. Moffitt JR and Zhuang X. RNA Imaging with Multiplexed Error-Robust Fluorescence in Situ Hybridization (MERFISH). Methods in Enzymology 2016. DOI: 10.1016/bs.mie.2016.03.020

55. Geffers L and Eichele G. High-Throughput In Situ Hybridization: Systematical Production of Gene Expression Data and Beyond BT - In Situ Hybridization Methods. In Situ Hybridization Methods 2015. Ed. by Hauptmann G:221–45. DOI: 10.1007/978-1-4939-2303-8{_}11. Available from: https://doi.org/10.1007/978-1-4939-2303-8_11

56. Lein ES et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature 2007 Jan; 445:168–76. DOI: 10.1038/nature05453. Available from: http://www.nature.com/articles/nature05453

57. Ljungberg MC, Sadi M, Wang Y, Aronow BJ, Xu Y, Kao RJ, Liu Y, Gaddis N, Ardini-Poleske ME, Umrod T, Ambalavanan N, Nicola T, Kaminski N, Ahangari F, Sontag R, Corley RA, Ansong C, and Carson JP. Spatial distribution of marker gene activity in the mouse lung during alveolarization. Data in Brief 2019; 22:365–72. DOI: https://doi.org/10.1016/j.dib.2018.10.150. Available from: https://www.sciencedirect.com/science/article/pii/S2352340918313672

58. 10X VISIUM SPATIAL TRANSCRIPTOMICS. Available from: https://biotech.illinois.edu/htdna/applications/10x-visium-spatial-transcriptomics

59. ADVANCED GENOMICS CORE PRICING. Available from: https://brcf.medicine.umich.edu/cores/advanced-genomics/price-list/

60. Jemt A, Salmén F, Lundmark A, Mollbrink A, Fernández Navarro J, Ståhl PL, Yucel-Lindberg T, and Lundeberg J. An automated approach to prepare tissue-derived spatially barcoded RNA-sequencing libraries. Scientific Reports 2016. DOI: 10.1038/srep37137

61. Zuo Q, Mogol AN, Liu YJ, Casiano AS, Chien C, Drnevich J, Imir OB, Kulkoyluoglu-Cotul E, Park NH, Shapiro DJ, Park BH, Ziegler Y, Katzenellenbogen BS, Aranda E, O'Neill JD, Raghavendra AS, Tripathy D, and Erdogan ZM. Targeting metabolic adaptations in the breast cancer–liver metastatic niche using dietary approaches to improve endocrine therapy efficacy. bioRxiv 2021 Jan :2021.09.07.458711. DOI: 10.1101/2021.09.07.458711. Available from: http://biorxiv.org/content/early/2021/09/07/2021.09.07.458711.abstract

62. Laskow S. Resurrecting the Forgotten Bike Highways of 1930s Britain. 2017

63. Sharma A, Ghosh D, Divekar N, Gore M, Gochhait S, and Shireshi SS. Comparing the socio-economic implications of the 1918 Spanish flu and the COVID-19 pandemic in India: A systematic review of literature. eng. International social science journal 2021 Mar :10.1111/issj.12266. DOI: 10.1111/issj.12266. Available from: https://pubmed.ncbi.nlm.nih.gov/34230684%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8251181/

64. Robinson KR. Comparing the Spanish flu and COVID-19 pandemics: Lessons to carry forward. Nursing Forum 2021 Apr; 56:350–7. DOI: https://doi.org/10.1111/nuf.12534. Available from: https://doi.org/10.1111/nuf.12534

65. Fan Z, Chen R, and Chen X. SpatialDB: a database for spatially resolved transcriptomes. Nucleic Acids Research 2020 Jan; 48:D233–D237. DOI: 10.1093/nar/gkz934. Available from: https://doi.org/10.1093/nar/gkz934

66. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, Clevers H, Deplancke B, Dunham I, Eberwine J, Eils R, Enard W, Farmer A, Fugger L, Göttgens B, Hacohen N, Haniffa M, Hemberg M, Kim S, Klenerman P, Kriegstein A, Lein E, Linnarsson S, Lundberg E, Lundeberg J, Majumder P, Marioni JC, Merad M, Mhlanga M, Nawijn M, Netea M, Nolan G, Pe'er D, Phillipakis A, Ponting CP, Quake S, Reik W, Rozenblatt-Rosen O, Sanes J, Satija R, Schumacher TN, Shalek A, Shapiro E, Sharma P, Shin JW, Stegle O, Stratton M, Stubbington MJT, Theis FJ, Uhlen M, Oudenaarden A van, Wagner A, Watt F, Weissman J, Wold B, Xavier R, Yosef N, and Participants HCAM. The Human Cell Atlas. eLife 2017; 6. Ed. by Gingeras TR:e27041. DOI: 10.7554/eLife.27041. Available from: https://doi.org/10.7554/eLife.27041

67. Langseth CM, Gyllborg D, Miller JA, Close JL, Long B, Lein ES, Hilscher MM, and Nilsson M. Comprehensive in situ mapping of human cortical transcriptomic cell types. Communications Biology 2021; 4:998. DOI: 10.1038/s42003-021-02517-z. Available from: https://doi.org/10.1038/s42003-021-02517-z

68. Callaway EM et al. A multimodal cell census and atlas of the mammalian primary motor cortex. Nature 2021 598:7879 2021 Oct; 598:86–102. DOI: 10.1038/s41586-021-03950-0. Available from: https://www.nature.com/articles/s41586-021-03950-0

Chapter 7

CURRENT ERA TECHNOLOGIES

7.1 ROI selection

A simple way to preserve spatial information is to isolate the samples from known locations in the tissue, and the act of selection and isolation is the only means to preserve the locations. The samples can be isolated physically or by molecular techniques. The known locations can be targeted, for cells with certain histological characteristics, or untargeted, on a grid over the tissue.

History of LCM

Microdissection

LCM, also known as laser microdissection (LMD), is by far the most commonly used method of microdissection. Before LCM, manual microdissection could isolate small pieces of tissue, but the process was laborious [1]. Laser microdissection predates ISH, though it was not used for spatial transcriptomics until it was possible to profile the transcriptome from small quantity of tissue.

A precursor to laser microdissection is the 1912 "Strahlenstich", which focused a conventional light source to a spot a few micrometers in size to cut tissues [2]. Soon after the invention of the laser in 1960, ruby laser was used to manipulate mitochondria, and a ruby laser microdissection system was commercialized by Zeiss in 1965 [2]. UV laser was used to create chromosomal lesions in 1969 [3]. The first use of UV laser to cut tissue was in 1976 [4] (Figure 6.3).

At present, there are two main types of LCM: IR and UV. IR LCM was introduced in 1996 [5]. It utilizes a cap with thermoplastic film which is placed over an area of interest, and an IR laser to briefly heat select areas of tissues to 90 °C so the film melts in the area and fuses to the area of tissue of interest [5] (Figure 7.1 A). This was commercialized as the Arcturus PixCell II LCM System in 1997, which was used in several early LCM studies including the first one in 1999 [6, 7, 8, 9] (Figure 6.3).

UV LCM is also known as laser microbeam microdissection (LMM) due to the microbeam of UV laser used. A popular commercial UV LCM system is the Robot-Microbeam (P.A.L.M. Wolfratshausen, Germany), now Zeiss PALM Microbeam. In

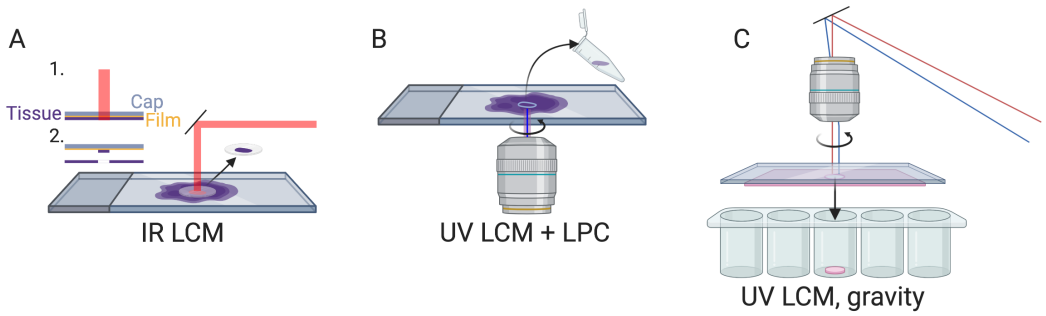

Figure 7.1: A) IR LCM schematic. B) UV LCM and LPC schematic, like in Zeiss PALM Microbeam. C) UV LCM, letting microdissected region fall by gravity, like in Leica LMD. All schematics in this book, i.e. anything not made with `ggplot2`, were created with BioRender.com

this method, a narrow UV laser beam ablates a narrow strip of tissue surrounding the area of interest, isolating the area of interest from the rest of the section, so the area of interest is minimally heated. Then, the area of interest is removed from the slide into the collection vial with laser pressure catapult (LPC), avoiding physical contact so as to prevent cross contamination (Figure 7.1 B). An early version of this system was first used in 1996 to isolate single-cells from gastric tumors, followed by PCR to analyze E-cadherin mutations, but the cells were removed with a needle rather than LPC [10]. Another popular commercial UV LCM system is the Leica LMD; unlike the PALM system, the Leica system lets the isolated tissue fall into collection vials by gravity, still avoiding physical contact (Figure 7.1 C). UV LCM was used in some early LCM spatial transcriptomics studies as well [11], and remains popular in recent years while IR LCM seems to have fallen out of favor [12, 13, 14].

Recent versions of the Arcturus LCM system have both IR and UV, which can be used in conjunction. UV can be used to cut the region of interest (ROI) and IR can then be used to fuse the region to the film at a few points for removal [15].

Amplification

The minuscule amount of transcripts from microdissected tissues, which can be single-cells, needs to be amplified to be detected by microarray or RNA-seq. Indeed, RNA amplification is a part of one of the most prevalent topics in LCM related search results (Figures 8.3, 8.4). To this day, there are two main strategies of amplification of minuscule amount of mRNA or cDNA—*in vitro* transcription (IVT) of cDNA (linear amplification) and PCR (exponential amplification), or a combination of both [16]. These two strategies have coexisted since their beginnings in 1989 and 1990

(Figure 6.3).

Heterogeneous cDNAs can be amplified with PCR by appending known sequences to one or both ends of the cDNA so primers with known sequences can be used to amplify the heterogeneous cDNAs. Early approaches meant for single-cells or small number of cells include tailing the cDNA with poly-dA [17] or poly-dG [18] after reverse transcription, and use as PCR primers sequences containing poly-dT (both poly-dA tail and reverse transcription (RT) primer of poly-A mRNAs) or poly-dC (poly-dG tail) and poly-dT (RT primer). Alternatively, lone linkers ("lone" because they are designed to prevent linker polymerization) could be ligated to both ends of the DNA fragments of interest to anneal to PCR primers [19]. Some of the early single-cell (or small number of cells) transcriptomic studies used PCR amplification, prior to quantification or differential expression analyses with Southern blot with radiolabeled cDNA probes hybridizing to cDNA clones of interest screened from plaque lift hybridization of a phage cDNA library [20], or with cDNA microarray [21, 22]. LCM was used to isolate the single-cells in [22]. Before the advent of CEL-seq, early scRNA-seq methods also used PCR amplification [23, 24]. An influential method is switching mechanism at the 5' end of the RNA transcript (SMART) [25], for construction of cDNA (clone) libraries covering the full length of mRNAs, though not originally for single-cells. The full length scRNA-seq method Smart-seq(2) [26] is based on SMART but adapted to the minuscule amount of transcripts from single-cells, with PCR amplification of the cDNA. Smart-seq(2) is one of the most commonly used library preparation methods for LCM since the 2010s, and was used for RNA-seq of LCM isolated single-cells [27].

Alternatively, transcripts can be amplified by IVT, with a T7 RNA polymerase promoter attached to the 5' end of the poly-dT primer, so the RNA polymerase transcribes the cDNAs into many copies of antisense RNAs (aRNA) [28, 29]. Some of the early single-cell (or small number of cells) transcriptomic studies used IVT amplification. Quantification and differential expression analyses of the aRNA can be performed with differential display [30, 31], cDNA microarray [32, 33], or with "expression profiling" [29, 31], i.e. reverse northern blot with radiolabeled aRNAs hybridizing to cDNA clones of interest, where the cDNA clones can be blotted onto a Southern blot membrane in a macroarray, which may have inspired the development of the cDNA microarray printed on glass [31]. LCM was used to isolate the single-cells in [33]. Since the 2010s, Cell Expression by Linear amplification and Sequencing (CEL-seq) [34] and derivatives (e.g. CEL-seq2, MARS-seq, and

SORT-seq), which use IVT amplification, have been commonly used for library preparation for microdissected or *de facto* microdissected samples such as from LCM [35], Tomo-seq [36], and Niche-seq [37].

Usage of LCM

Usage trends of LCM as reflected in PubMed and bioRxiv search results are analyzed in Chapter 8. LCM can be used to isolate targeted ROIs based on histology, or to create a grid for untargeted search of gene expression patterns in space, and examples of both are highlighted here. Moreover, the three themes of screening, atlas curation, and new technique development, are all represented in LCM literature. In the "screening" theme, LCM is used to isolate cell populations of interest based on histology (targeted) to discover genes associated with pathological conditions such as cancer metastasis [11] and cell types [38], or to discover cell type localization in healthy tissue difficult to other spatial transcriptomics techniques such as the bone marrow [14].

LCM can also be used to dissect the tissue in a grid, not targeting very specific histological regions (untargeted), to identify genes associated with locations on the grid [39, 40] or transcriptomically defined regions [13, 40], or to map cells from scRNA-seq to spatial locations [13, 40]. The untargeted studies can also touch upon the "atlas" theme, providing an online interface to query and explore the spatial transcriptomes [40].

However, targeted approaches can also be used for the "atlas" theme, such as in the human [41, 42]; [42] and macaque [43] atlases of the ABA, isolating histologically annotated regions for microarray profiling to build systematic resources for exploration. This addresses the limitation of bright field ISH that only one gene can be stained per section thus requiring large number of brains, which is too costly for primates; in LCM, while often not single-cell resolution, the same brain can be used to profile the whole transcriptome. The "technique development" theme is evident in the text mining results (Figure 8.3), and contributes to some of the advantages of LCM as discussed below.

As shown in Chapter 8, LCM transcriptomics has spread far and wide, and has been used on many research topics rarely featured in (WM)ISH atlases. These include cancer and botany (Figure 8.2, Figure 8.3). The following advantages of LCM might have contributed to its popularization: first, as already mentioned, both IR and UV LCM systems have been commercialized prior to their use for

transcriptomics, making setup convenient. Second, while LCM equipment can be expensive and require specialized training to use, many institutions have core facilities that can perform LCM [44, 45, 46, 47], reducing cost and personnel training time in individual laboratories.

Third, in some cases, especially in the clinical setting, only archival formalin fixed, paraffin embedded (FFPE) tissues are available. While in 2020, newer current era technologies such as Visium [48] and GeoMX DSP [49] have been demonstrated on FFPE tissues, LCM followed by microarray was already demonstrated on FFPE tissues in 2007 [50] and with RNA-seq by 2014 [51]. As a result, for several years, LCM may have been the only option to perform spatial transcriptomics on FFPE samples. In addition, LCM might still be the only way to profile transcriptomes of single-cells in FFPE samples. With scRNA-seq library preparation methods such as Smart-seq2 [27], and CEL-seq [35] it is possible to profile the transcriptome in minuscule amount of LCM isolated tissue, and even single-cells [27]. With Smart-3SEQ, LCM single-cell transcriptomics has been made possible for FFPE tissues as well, even for samples that are several years old [52].

Finally, despite its long history, LCM cannot yet be replaced by newer spatial transcriptomics technologies. Unlike smFISH or ISS based techniques, LCM followed by RNA-seq is not restricted to known genes and allows for transcriptome wide profiling and other omics. Unlike ST and Visium, LCM can have single-cell resolution, and unlike array based techniques with resolution of the size of a cell or higher, such as Slide-seq(2) and HDST, LCM can more unequivocally isolate individual cells or nuclei based on histology.

LCM has a number of disadvantages, some of which are addressed by other current era spatial transcriptomics technologies. First, compared to droplet based scRNA-seq and highly multiplexed barcoding, using LCM to isolate single-cells is still too laborious, limiting its throughput. Second, LCM requires tissue sections, while preparation of many slides to cover a 3D volume can be laborious and it can be challenging to reconstruct 3D structures from tissue sections. To reiterate, sections of blastoderm stage embryos are hard to interpret, which motivated WMISH. Third, because it can be challenging to segment cells based on hematoxylin and eosin (H&E) or immunohistochemistry (IHC) staining and parts of different cells can be stacked within the thickness of the section even in thin sections, single-cells isolated by LCM can have contents of other cells.

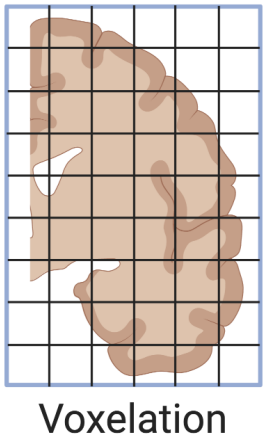

Figure 7.2: Voxelation of human brain, as in [53].

Physical microdissection

Voxelation

LCM did not completely replace microdissection with a physical blade. Voxelation was one of the alternatives to LCM developed to profile spatial transcriptomes in 3D and address the limitation of throughput of ISH. In voxelation, a grid of steel blades is used to cut tissue into cubes for microarray profiling, but the resolution is low. Human brains were first cut into 8 mm thick slabs and then a grid of 1 cm per side [53, 54], and mouse brains were first cut into 1 mm thick slabs and then a grid of 1 mm per side [55, 54, 56] (Figure 7.2). With low resolution, it's easier to use voxelation to profile large 3D tissues of multiple slabs that would be much more laborious with LCM's thinner sections and higher resolution [55]. As the human voxels were quite large (almost 1 ml) and corresponding voxels of 20 to 30 mice were pooled [55, 56] to get enough transcripts, the voxelation studies did not mention T7-based PCR amplification of transcripts, unlike for LCM samples [11]. To the best of our knowledge, voxelation never spread beyond its institution of origin, UCLA School of Medicine, and has not been used in a publication to generate new data since 2007 [56] and for data analysis since 2009 [57].

Tomo-seq

Another alternative to LCM is Tomo-seq/array, which has continued to be utilized in recent years. In this approach, the tissue is sectioned with a cryotome like tomography (hence the "Tomo"), and the transcripts in each section are extracted for microarray (Tomo-array) or RNA-seq (Tomo-seq) profiling; the resolution is limited

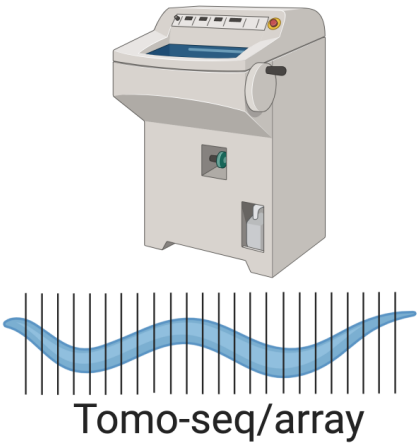

Figure 7.3: Tomo-seq, here showing C. elegans.

by section thickness, which has gone down to 8 μm [58]. Three-D expression maps can be reconstructed from sections along the anterior-posterior (AP), dorsal-ventral (DV), and left-right (LR) axes. All three themes, namely screening, atlas curation, and new technique development, are present in Tomo-seq/array literature.

Tomo-array was first used in 2012 to build a 3D mouse brain transcriptome atlas, attempting to address difficulties in image registration in ISH atlases, low resolution of voxelation, and limitation of LCM to specific regions [59] (Figure 6.3). Mouse brains were sectioned along all three axes and 200 adjacent 5 μm sections were pooled as "fractions" for microarray; again, PCR amplification was not mentioned. Fractions from the three axes were then used to reconstruct a 3D atlas.

Tomo-seq was first demonstrated in 2013, on *Drosophila melanogaster* embryos, with 60 and 25 μm sections, again in response to the difficulty to scale ISH atlases to the whole transcriptome [60]. Genes patterned along the AP axis were identified, and the data is stored in an online database. However, Tomo-seq is more commonly credited to a 2014 method first demonstrated on zebrafish embryos, with 18 μm sections [36]. Gene expression patterns along the AP axis of straightened embryos were identified, and sections along all three axes were used for 3D reconstruction of embryos that were not straightened. The data and the 3D reconstruction are also stored in an online database, though the 3D reconstruction algorithm produced many artefacts.

Since then, Tomo-seq has been used in several different biological systems, typically when one axis is of primary interest. Tomo-seq has been used in *C. elegans* [61], developing zebrafish hearts [62], *Drosophila* embryos [63], ischemic mouse hearts

[64], and *Pristionchus pacificus* [65] to identify genes associated with that axis of interest. Tomo-seq was also used on mouse [58] and human [66] gastruloids to demonstrate the viability of this in vitro and potentially high-throughput model for developmental biology. Again, due to the minuscule amount of tissue in each section, library preparation methods designed for scRNA-seq, such as CEL-seq(2) [36, 65, 61] have been adapted to Tomo-seq.

Other methods of physical microdissection

Algorithms inspired by reconstruction of ray-based computerized tomography have been used to reconstruct spatial patterns of gene expression from Tomo-seq-like slices cut from different angles of the same tissue with a stereotypical structure. This was first attempted with Gene Expression Tomography (GET) [67], though only on qPCR quantification of one gene in those slices. More recently, this kind of idea was used in Spatial Transcriptomics by Reoriented Projections and sequencing (STRP-seq), in response to the limited number of genes of smFISH and ISS based techniques, degradation of RNA and technical complexity of LCM, and number of specimens required by and inadequacy of the 2014 Tomo-seq 3D reconstruction [68]. This has been shown to perform better than the 2014 Tomo-seq 3D reconstruction method, and was demonstrated on the brain of a non-model organism, the lizard *Pogona vitticeps*.

Because of the specialized equipment and technical complexity of LCM and degradation of RNA, other methods of physical microdissection have been developed. Examples of such techniques are Cell and Tissue Acquisition System (CTAS), which uses a disposable capillary unit connect to the vacuum to aspirate tissue [69], and an automated micropuncch system that collects samples of tissue with diameter of 110 μm at 300 μm intervals [70]. In addition, for similar reasons, manual microdissection is still used (Figure 6.7), such as to dissect leaves on a grid of distances from a lesion to characterize response to infection [71, 72]. Manual microdissection of pre-defined anatomical regions was also used to create low resolution gene expression atlases of *Xenopus laevis* [73] and *Xenopus tropicalis* [74] embryos, to avoid sectioning as required for LCM and artefacts in Tomo-seq 3D reconstruction.

***De facto* microdissection**

Some methods have been developed that do not directly cut tissues. Instead, cells, or ROIs judged from histology, are optically and molecularly marked so that only

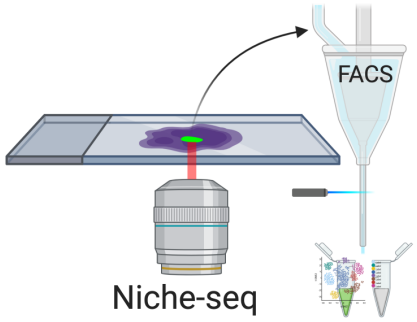

Figure 7.4: Niche-seq schematics. Green: cells with photoactivated PA-GFP.

transcripts or cells from the marked regions are captured. Because these methods involve selection of pre-defined ROIs within the section, we call them *de facto* microdissection.

Transcriptome *in vivo* analysis (TIVA) from 2014 can be viewed as the first of these methods [75]. Live cell culture is incubated with the photoactivable cage with a poly-U sequence that captures poly-A transcripts. Select cells are photoactivated by 405 nm laser and the captured transcripts are sequenced. TIVA is widely cited, perhaps because it is one of the earliest single-cell resolution and transcriptome wide methods, predating RNA-seq from LCM isolated single-cells. However, because TIVA has only been demonstrated on fewer than a dozen cells per sample, to the best of our knowledge it has not been used in any other publication to collect new data.

A *de facto* microdissection method that has spread beyond its institution of origin is Niche-seq, which was developed as LCM is still usually used to isolate groups of cells rather than single-cells and involves tissue fixation [37]. Select regions of *ex vivo* tissues from transgenic mice expressing photoactivable GFP (PA-GFP), here lymph node and spleen B cell and T cell niches, are photoactivated at 820 nm with two photon irradiation. Then the tissue is dissociated and cells with photoactivated PA-GFP are collected from flow cytometry-based fluorescence-activated cell sorting (FACS) for scRNA-seq with MARS-seq (Figure 7.4). This approach was originally used in 2010 to isolate B cells from light and dark zones of the lymph node followed by transcriptome profiling with microarray in bulk [76]; the difference in Niche-seq is scRNA-seq of the sorted cells. After its inception, Niche-seq has been used once more in lymph node niches [77]. However, as Niche-seq requires transgenic mice expressing PA-GFP and living tissue, it cannot be applied to human tissues, to fixed

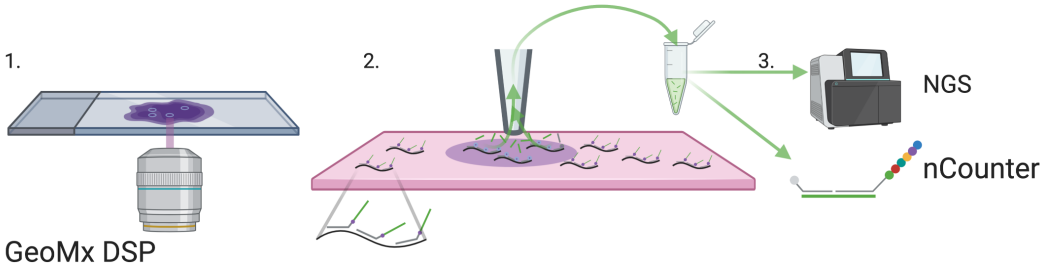

Figure 7.5: GeoMX DSP schematics, inspired by figures in [78]. Black: transcripts in tissue. Gray: probes. Green: indexing oligo.

tissues, or when a PA-GFP line is unavailable. This might limit further growth of Niche-seq. Moreover, the spatial context of cells within the photoactivated region is lost, limiting spatial resolution.

Another method that spread beyond its institution of origin is the commercial GeoMX DSP from NanoString [78], which can be used for both high throughput immunofluorescence and transcript quantification in FFPE tissue sections. While GeoMX DSP does not physically isolate relevant parts of the tissue, it is discussed in this section because like other microdissection based techniques, GeoMX DSP is primarily ROI based, and spatial location is known from selection of the ROI. For transcript quantification, probes are attached to indexing oligos with a UV cleavable linker (Figure 7.5). The selected ROI is illuminated by UV to remove the index oligos from the probes. Then the released index oligos are aspirated and quantified with either NGS or NanoString nCounter. This can be repeated for multiple ROIs, which can be a grid for unbiased profiling [78]. The probes tile the transcripts, and each probe has a distinct index oligo, so in NGS, each tile is counted separately, enabling isoform quantification [78]. The number of genes profiled by GeoMx DSP depends on the gene panel used; the Cancer Transcriptome Atlas panel with over 1800 genes have been used in several studies (e.g. [79, 80], and with the human or mouse Whole Transcriptome Atlas (WTA) panel, transcripts of 18190 genes can be quantified, nearly covering the whole transcriptome [81]. In GeoMX WTA, the UV cleaved index oligo must be sequenced with NGS to identify the gene each transcript is from. As pre-defined probes are required, unlike in RNA-seq, novel transcripts cannot be quantified. Ready made probe sets for oncology, immunology, and neuroscience are sold by NanoString [82]. Although GeoMx DSP was published in 2019, it has spread to several different institutions, and has been used on pancreatic ductal adenocarcinoma (PDAC) [49], hepatocellular carcinoma (HCC)

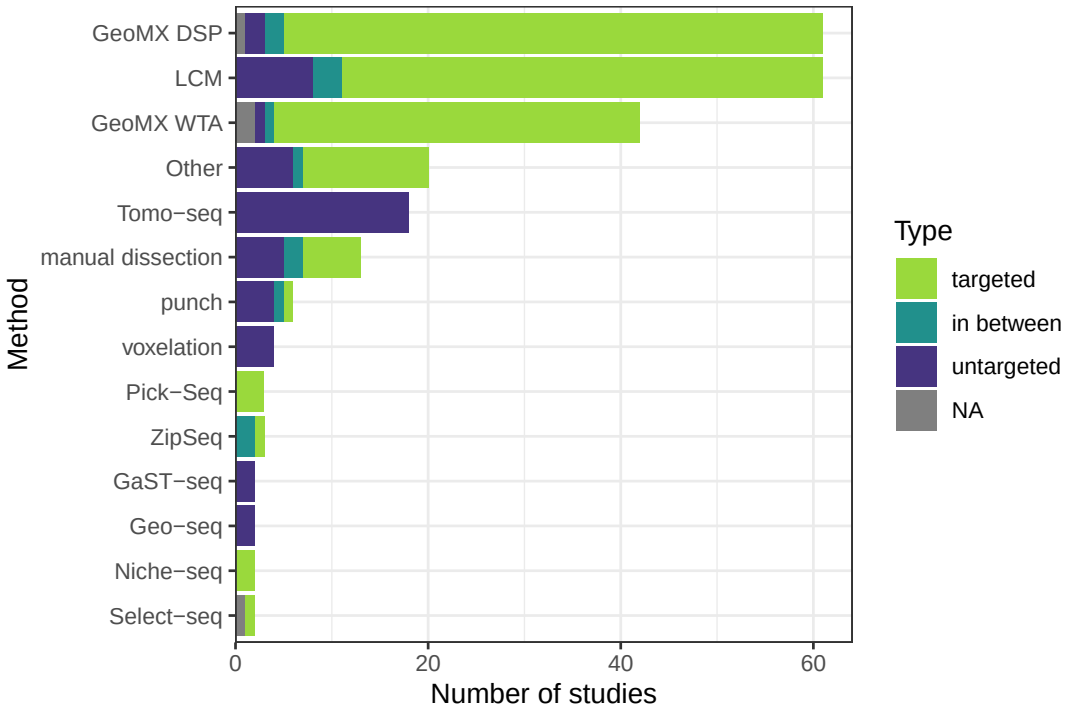

Figure 7.6: Number of studies of each of the three types: targeted, in between, and untargeted, using each microdissection based technique plus GeoMX DSP. Techniques used in less than two studies or two types are lumped into Other.

[83], reactive lymph nodes [84], and COVID infected lungs from autopsy [80, 85, 86, 79].

Targeted vs. untargeted

Some methods can only be used in a regular grid, such as Tomo-seq, while some can be used either in a regular grid or in targeted ROIs, such as LCM and GeoMX DSP (Figure 7.6. Some are primarily used for targeted ROIs, such as Niche-seq. Sometimes a targeted ROI in the section may be chosen, which is then divided into smaller regular parts, in between targeted and untargeted.

After LCM, GeoMX DSP/WTA is the most popular targeted ROI based technique, and as already mentioned, GeoMX DSP has been used in several COVID autopsy studies. GeoMX DSP is often used to profile proteins, which is beyond the scope of this book; our database only contains metadata for GeoMX DSP transcriptomic datasets. As of writing, all GeoMX DSP datasets in our database are from human, and are from predominantly pathological FFPE tissues (Figures 7.7, 7.8). Because of COVID, GeoMX DSP is more used on the lungs for transcriptomics than other tissues (Figure 7.9).

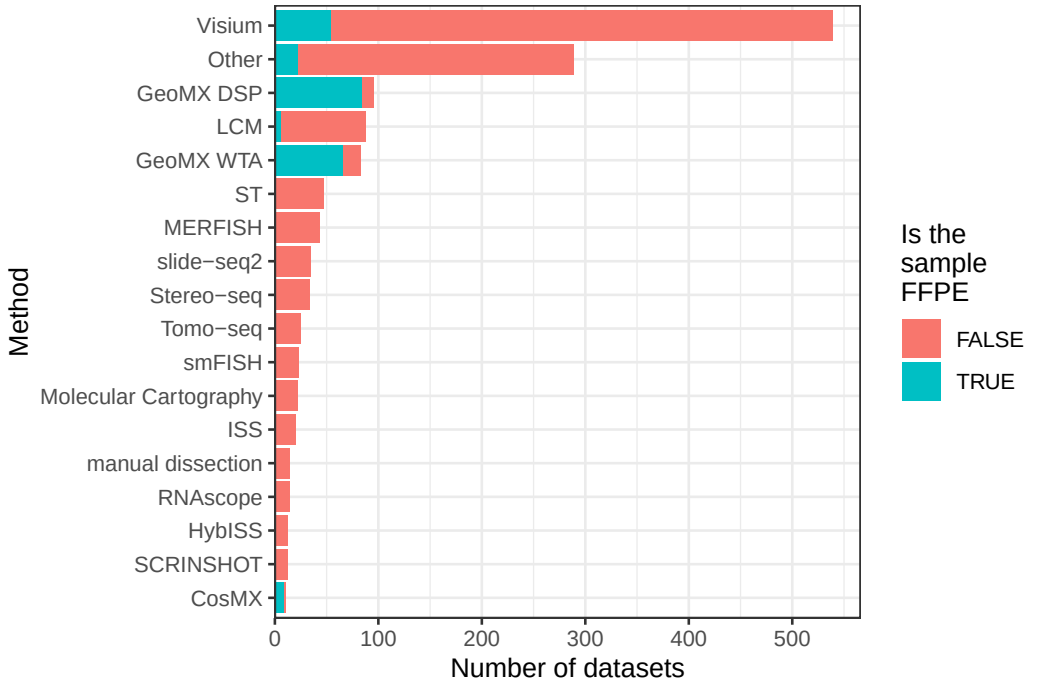

Figure 7.7: Number of FFPE and frozen section datasets from each current era technique; techniques used in fewer than 10 datasets are lumped into Other. LCM is only for curated LCM literature and does not include all search results in Chapter 6.

In an earlier version of this book, in the current era, ROI selection (formerly Microdissection) was the most widely used type of techniques. However, NGS barcoding has surpassed ROI selection more recently due to the rapid growth of popularity of Visium (Figure 7.10). Excluding LCM, GeoMX DSP and Tomo-seq are the most popular techniques after ST and Visium (Figure 6.7). ROI selection has not been replaced by other seemingly more sophisticated techniques such as ST and MERFISH, and is still popular in 2020 and 2021 (Figure 6.1, Figure 8.1). ROI selection techniques generally do not have single-cell resolution, but combined with scRNA-seq or snRNA-seq data, cell type compositions of ROIs can be computationally deconvoluted [14, 49]. The popularity may be due to availability of commercial platforms (LCM and GeoMX DSP), core facilities (LCM, NGS, and Nanostring nCounter for GeoMX DSP), Nanostring's Technology Access Platform (TAP), a commercial data collection and analysis service for GeoMX DSP [87], not requiring specialized equipment (Tomo-seq, manual microdissection), or disadvantages of other techniques discussed later in this chapter.

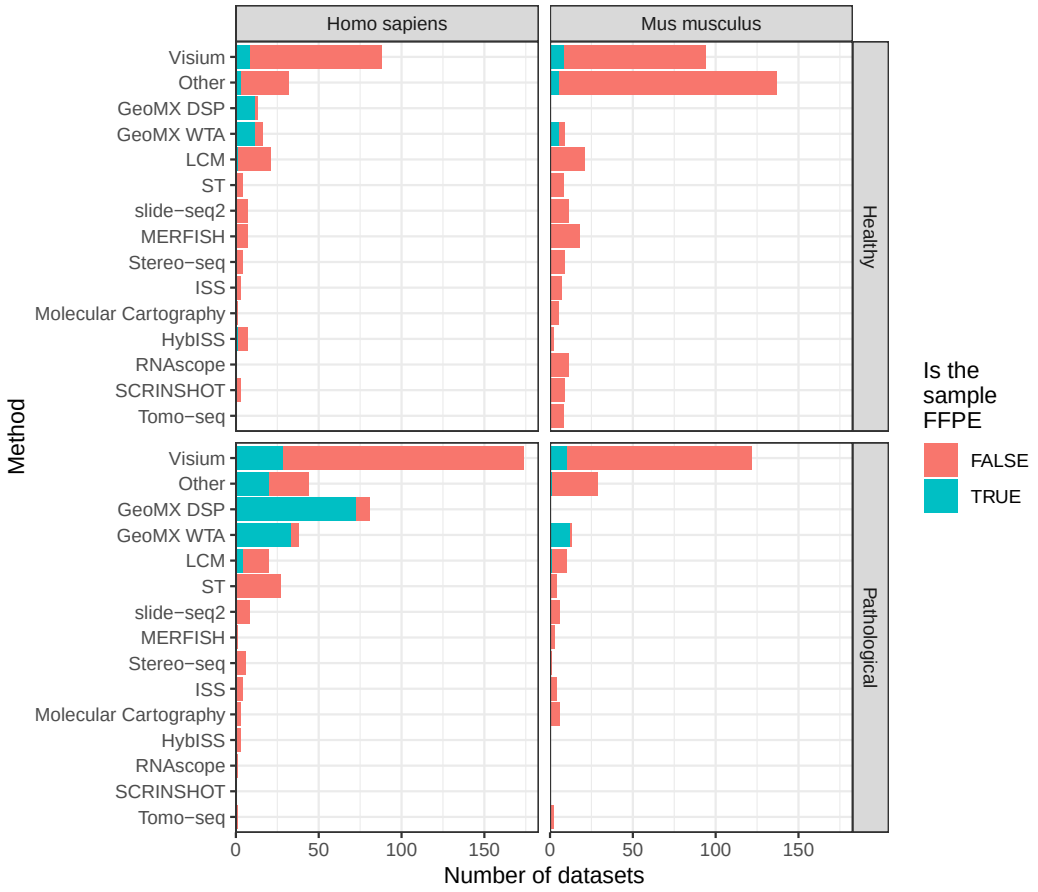

Figure 7.8: Number of FFPE and frozen section datasets from each current era technique in humans and mice healthy and pathological tissues; techniques used in fewer than 10 datasets are lumped into Other. LCM is only for curated LCM literature and does not include all search results in Chapter 6.

7.2 Single molecular FISH

One quantitative approach to transcript abundance estimation is to display individual transcripts as distinct puncta with FISH and count them. Prior to smFISH, transmission electron microscopy was used to visualize individual mRNA molecules in fibroblasts by labeling the poly-A tail with a single large colloidal gold particle and the *in situ* reverse transcribed cDNA with small gold particles [88]. That FISH can be used to visualize single mRNA molecules was first demonstrated in 1998 [89] (Figure 6.3). Five or more probes targeting adjacent parts of the transcript, each about 50 nt long and labeled with 5 fluorophores were hybridized to the transcripts. The puncta seen were shown to be likely individual mRNA molecules, as the fluorescence intensity of each punctum was consistent with the number of fluorophores, and the number of puncta for β-actin was consistent with the number of β-actin

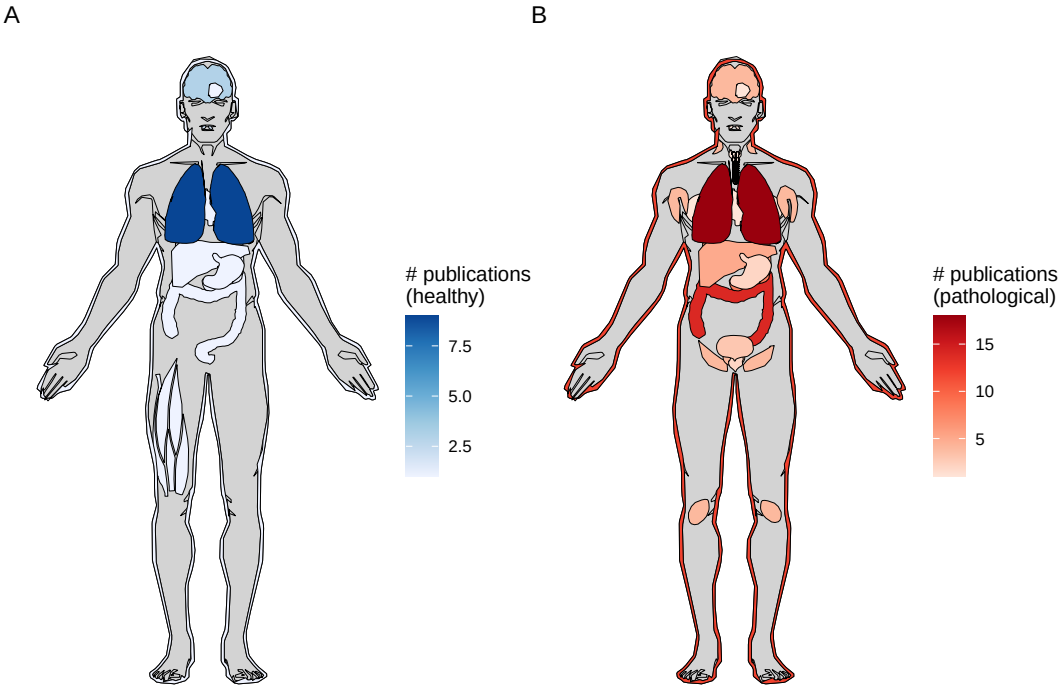

Figure 7.9: Number of GeoMX DSP or WTA studies for healthy and pathological human organs. Male is shown here because there are studies for the prostate but not for female specific organs.

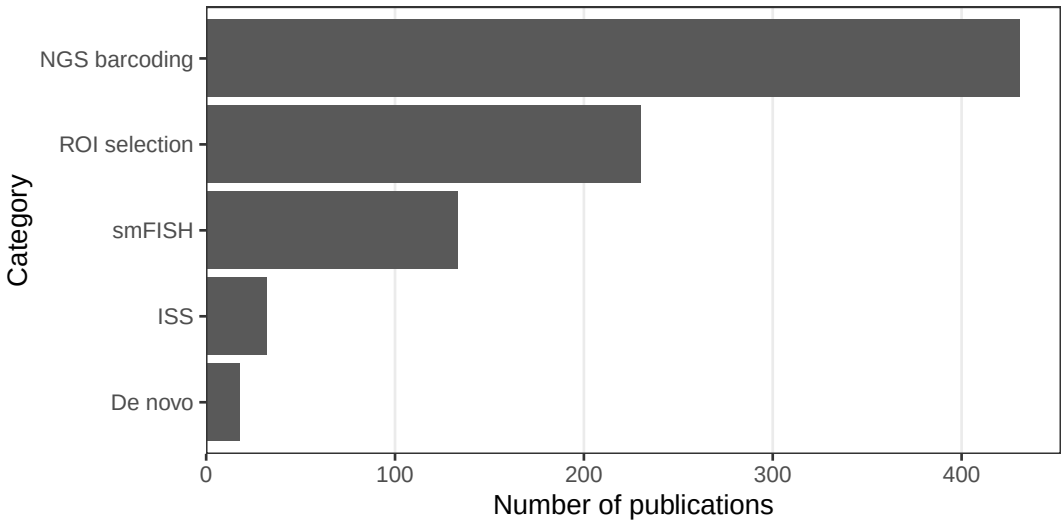

Figure 7.10: Number of publications per category of techniques in the current era. Non-curated LCM literature is excluded.

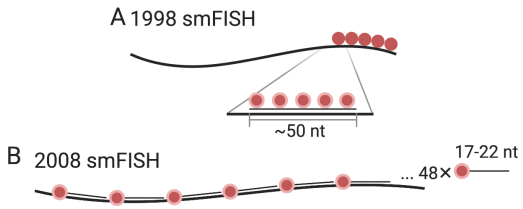

Figure 7.11: A) Schematic of smFISH from [89]. The long thick line stands for the mRNA, and short think line stands for DNA oligo probe. B) smFISH with singly labeled probes from [90].

transcripts measured by other means, and the colors of puncta seen from probes with different colored fluorophores targeting different parts of the transcript were consistent with organization of the fluorophores on the transcript (Figure 7.11).

The 1998 approach had a number of disadvantages, leading to development of an alternative approach in 2008 [90]. First, probes labeled with multiple fluorophore moieties are difficult to synthesize and purify. Second, the multiple fluorophores on the same probe can interact with each other and self-quench. Third, out of the 5 probes per transcript, only 1 or 2 may have actually hybridized to the transcript in most cases, making it difficult to distinguish between true signal and non-specific binding. In the 2008 method, each 17-22 nt probe is labeled with one fluorophore at the 3' end, and a larger number of probes (48 or more) targeting tandem sequences of the transcript were used to improve signal to noise ratio (Figure 7.11). The probes were computationally designed and ordered from Biosearch Technologies. This method influenced later highly multiplexed smFISH techniques; computational probe design and commercial synthesis would remain crucial.

Barcoding strategies

To use smFISH to quantify transcripts transcriptome wide, there is an obvious challenge—how to distinguish among over 20,000 genes with only about 5 easily distinguishable colors? Various strategies using multiple colors and/or rounds of hybridization or imaging have been devised to drastically expand the palette. The first attempt to do so was in 1989, using 3 colors to visualize 4 chromosomes in immunological DNA FISH [91] (Figure 6.3). Each probe can be labeled with one or two of the 3 haptens: biotin, 2-acetyl aminofluorene (AAF), and Chemiprobe. Red fluorophore was attached to avidin to target biotin label, and blue and green to different secondary antibodies targeting, respectively, mouse anti-Chemiprobe and rabbit anti-AAF primary antibodies (Figure 7.12). Then with one doubly labeled and

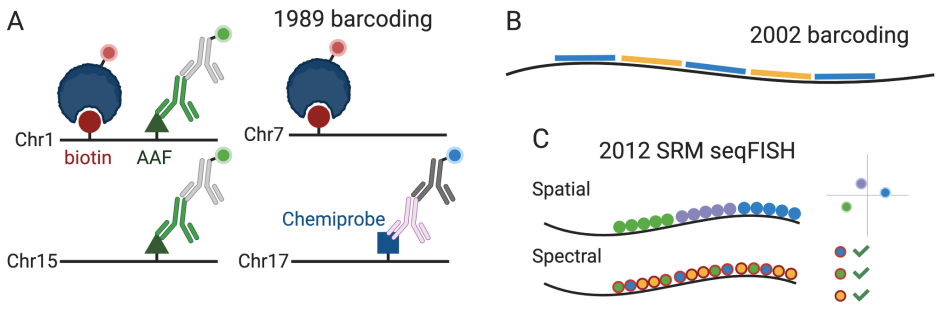

Figure 7.12: A) Combinatorial barcoding in immunological DNA FISH, as described in [91]. The line stands for the probe and the circle, triangle, and square stand for haptens. Not to scale, and only one hapten of each kind is shown on one probe. B) Combinatorial barcoding in [92]. Short colored lines stand for probes with fluorophores of the color. C) Schematic of SRM seqFISH as described in [93].

3 singly labeled probes, imaged with different excitation wavelengths or channels, 3 colors can distinguish 4 chromosomes. However, with this method, the palette size is limited by the number of haptens available and the number of their combinations.

For transcript detection, to our best knowledge, the first attempt was in 2002 [92]; fluorophore labeled probes were synthesized as in the 1998 smFISH method, and either probes of one color or a mixture of probes of 2 colors were hybridized to the transcript, and imaged with different channels, to visualize transcription foci in the nucleus (Figure 7.12). This way, combinations of 2 of the 4 available colors plus blank were used to encode 10 different transcripts.

The above mentioned historical works in smFISH and combinatorial barcoding laid foundation to smFISH-based spatial transcriptomics. The first attempt to quantify transcripts with combinatorial barcoding at single molecular resolution was in 2012 by Long Cai's group, which later developed seqFISH and its variants [93]. Like in the 2008 smFISH study, singly labeled probes purchased from Biosearch were used, but forming blocks of different colors as in the 1998 smFISH β-actin experiment. Then the transcripts were imaged with super-resolution microscopy (SRM), in particular stochastic optical reconstruction microscopy (STORM). In the spatial barcoding strategy, the ordering of the colors in space would distinguish between transcripts, but would require linearization of the transcripts and high resolution (20 nm) (Figure 7.12). To improve signal to noise ratio, cyanine dye–based photoswitchable dye pairs [94] was used so both the activator and the emitter fluorophores must be present and adjacent for the fluorophores to be reactivated. In the spectral barcoding approach, the pairs of fluorophores are spread across the transcript, so the transcripts

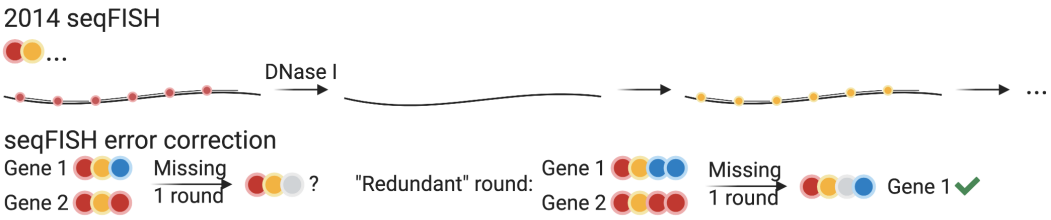

Figure 7.13: Probe structures of 2014 seqFISH ([95]Lubeck et al. 2014) and seqFISH error correction.

are recognized by the pairs of fluorophores detected (Figure 7.12). The spectral approach requires lower resolution (100 nm) and does not require linearization, but because the ordering of the colors is not used, the number of possible barcodes from the same number of colors is smaller than in the spatial approach. With spectral barcoding, transcripts of 32 genes were quantified in yeast, with 3 color barcodes chosen from 7 available colors. To the best of our knowledge, after its inception, this SRM method has not been used to generate new data, perhaps because it requires specialized equipment for SRM. None of the later methods in our curated database used SRM.

Thus far, probes with fluorophores of different colors were hybridized to mRNAs at the same time, without multiple rounds of hybridization. To obtain single molecular resolution but without SRM, there is a challenge of needing to use multiple probes of the same color to strengthen signal, which requires transcripts that are long enough to accommodate probes of different colors. The more colors that are used to encode more genes, the longer the transcripts must be.

This changed in 2014, with the advent of seqFISH [95]. Twenty four singly labeled probes were designed for each gene, and 12 genes were encoded with 4 colors and 2 rounds of hybridization (Figure 7.13). After imaging the first round of hybridization and DAPI staining for DNA, the probes are removed with DNase I, and then probes for the second round are hybridized. Let F denote the number of fluorophores or colors, and N denote the number of rounds of hybridization, then the number of genes that can be barcoded is F^N. However, with longer barcodes to encode more genes, error can build up.

The most common error in multi-round smFISH is missing signal, most likely in one round [96, 97]. If all F^N barcodes are used and one round is missing for a mRNA molecule, then the existing signal of this molecule is consistent to F genes, so it cannot be uniquely identified. If a small proportion of barcodes are intentionally

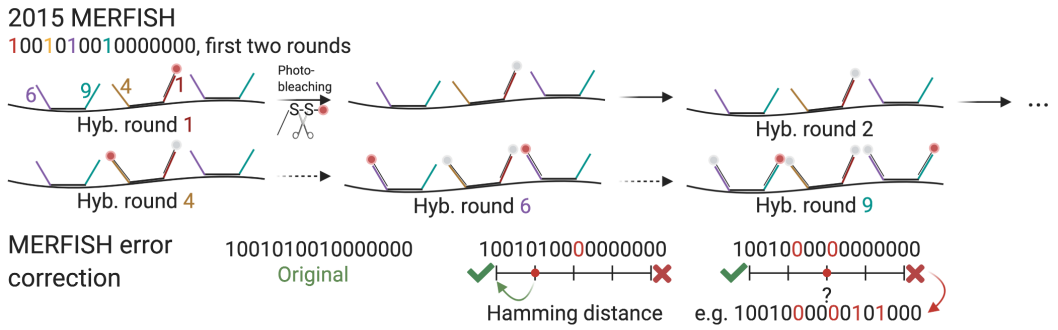

Figure 7.14: Schematic of MERFISH ([97]K. H. Chen et al. 2015; [99]Jeffrey R. Moffitt et al. 2016) and MERFISH error correction.

left out to control for false positives, as was done in this first version of seqFISH (4 out of 16), then error correction is still not guaranteed. A further defense against errors in 2014 seqFISH was to repeat the 2 rounds of hybridization 3 times, so 6 rounds were performed. This filtered out false positives where repeated rounds didn't match, and barring false positives, this can recover the original 2 barcoding rounds if up to 2 of the 6 total rounds have missing signal.

Another error correction scheme was introduced in 2016, with hybridization chain reaction (HCR) seqFISH [96], and was used in seqFISH+ [98] as well. One more round of hybridization than necessary to encode the number of genes of interest was used, and the barcodes are designed so that if one of the rounds is missing, the remaining rounds still uniquely identify the gene (Figure 7.13). For example, with 5 colors, 3 rounds are enough to encode 100 genes, as 125 barcodes are possible. However, a fourth round is used, so missing one round can still result in 3 remaining rounds that uniquely identify the gene.

An alternative to seqFISH was developed with error correction in mind – multiplexed error-robust FISH (MERFISH) [97]. In MERFISH each encoding probe has a 30 nt long region that targets the transcript, and 2 or 3 20 nt [99] readout sequences to bind to readout probes (Figure 7.14). First, the encoding probes are hybridized to the transcripts. For each round of hybridization, readout probes, singly labeled, are hybridized to the readout sequences on the encoding probes and imaged. Then the fluorescence of the previous round is either photobleached (version 1) [97] or when the fluorophore is bound to the readout probe with a disulfide bond, cleaved off with a reducing agent such as Tris(2-carboxyethyl)phosphine (TCEP) (version 2) [99]. The readout probes are not stripped, and in the next round, new readout probes are hybridized to new readout sequences and imaged.

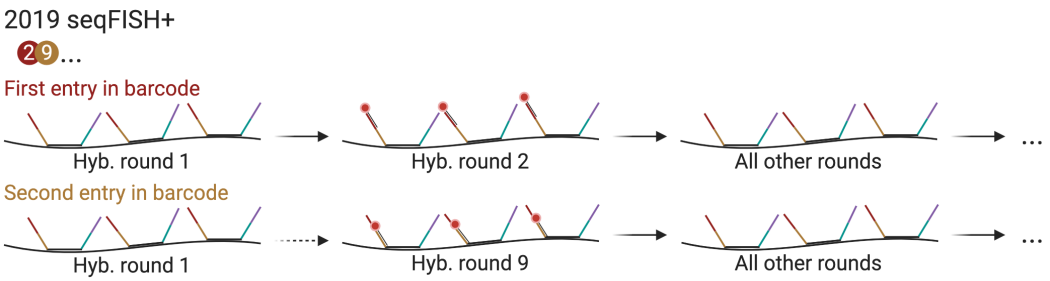

Figure 7.15: Schematic of seqFISH with pseudocolors.

The MERFISH barcodes are binary, with "1" for a round with fluorescence, and "0" without, and must differ from other barcodes at least 4 places, i.e. with Hamming distance of at least 4 (HD4). As missing signal is the most common error, each barcode has 4 1's, or Hamming weight 4. This way, when one round is missing, the gene can still be uniquely identified, but when 2 rounds are missing, the remaining barcode is equally distant to 2 genes, so the error cannot be corrected (Figure 7.14). Sixteen rounds of imaging, or 16 bits, would result in 140 barcodes. In this case, there are 16 different readout sequences, and each gene is assigned 4 of them, for the 4 1's in the barcode. If the code is expanded to 69 bits, then about 10,000 genes can be encoded, and by using 3 colors to image 3 bits per round, only 23 rounds of imaging are needed to cover the 69 bits, cutting imaging time to a third [100]. An HD2 code, i.e. barcodes are at least hamming distance 2 away from each other, can also be used, but errors can only be recognized but not corrected. All variants of MERFISH use this type of binary barcoding.

More recently, a new variant of seqFISH was devised to scale up to 10,000 genes [101]. The barcoding and hybridization scheme enabling such scale was first introduced in vitro in 2017 as RNA SPOTs [102], and was then adapted to cultured cells in 2018, targeting introns of nascent transcripts of over 10,000 genes [101]. In 2019, this scheme was used to profile mature transcripts of 10,000 genes in both cell culture and the mouse brain, and with super-resolution [98]. Super-resolution beyond the diffraction limit can be achieved by computationally super-resolving the transcript spots with a radial center algorithm [103] when spot density is very high to help with decoding barcodes; the super-resolution version is known as seqFISH+. While this new version of seqFISH can reduce optical crowding and greatly expand the palette, the super-resolution algorithm that can further reduce crowding does not have to be used to locate the transcript spots when density is low. This version of seqFISH was again used to visualize genomic loci (super-resolution) [104] and

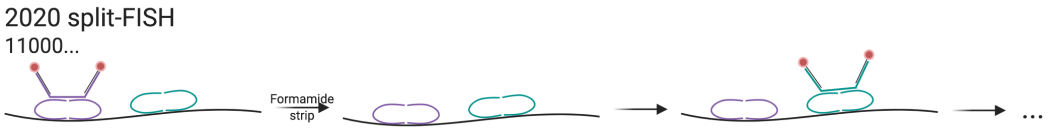

Figure 7.16: Schematic of split-FISH.

mature transcripts of a smaller number of genes (not super-resolution) [105].

This method is quite different from previous seqFISH variants, and is in some ways reminiscent of MERFISH. Like previous versions of seqFISH, each barcode is a series of colors, but a large number of "pseudocolors", specifically 20 per channel in the seqFISH+ study, are used rather than the 5 fluorophores, so 3 rounds of hybridization can encode 20^3 or 8000 genes per channel. Any number of pseudocolors and rounds can be used depending on the number of genes profiled. Each primary probe has a 28 nt region targeting the transcript and 4 readout sites of 15 nt. Each readout site has as many different sequences as there are pseudocolors, and the 4 sites correspond to the series of 4 pseudocolors in the barcode. First, 24 primary probes are hybridized to the transcripts. Then for each place of the barcode, 20 (or whatever number of pseudocolors) rounds of hybridization with readout probes are performed, stripping with formamide between rounds. In these 20 rounds, each gene should light up only once, and its place in the 20 rounds is its pseudocolor (Figure 7.15). This way, in each image, only 1 out of 20 molecules of interest imaged in the channel fluoresce, reducing optical crowding. For the entire barcode of length 4, there would be 80 rounds of hybridization. In contrast, in MERFISH, with the 16 bit barcode, this would be 1 out of 4. Like in MERFISH, a larger number of real colors, or channels, can be used to increase throughput, to image multiple pseudocolors simultaneously. So with 3 channels, 24,000 genes can be encoded. The same error correction method as in HCR seqFISH was used, so while a barcode of length 3 is sufficient, length 4 was used.

Another new method, called split-FISH [106] was devised to reduce off target hybridization, and thus background noise and some barcoding errors. For each encoding probe or bridge probe like in MERFISH, a pair of split probes hybridize to the transcript itself, inspired by the Z probes of RNAscope (Figure 7.16). Half of the split probes would bind to the transcript, and the other half bind to the bridge probe. Then as in MERFISH, the bridge probe has 2 readout sequences and singly labeled readout probes bind to the bridge probe for imaging. This method reduces

off target hybridization because the bridge probe can only indirectly bind to the transcript if both of the split probes hybridize to the transcript. To encode 317 genes, 2 places out of 26 in binary barcodes are chosen to be "1", resulting in 325 possible barcodes; 8 of them are left blank to control for false positives. Error correction is not mentioned.

Despite the availability of the above barcoding schemes, when the number of genes stained for is not too large, each gene can still be encoded by only one round of hybridization and one color. When the number of genes is larger than the number of colors, each round of hybridization stains for as many genes as there as colors, and the probes are stripped so the next round stains for a different set of genes. This has been done in osmFISH [107] staining for 33 genes, in a non-barcoded adaptation of HCR-seqFISH called Spatial Genomic Analysis (SGA) [108] staining for 35 genes, and in Expansion-Assisted Iterative Fluorescence *In Situ* Hybridization (EASI-FISH) 26 genes [109].

Signal amplification

As already mentioned, in smFISH, a large number of singly labeled probes can be used to boost signal, but not all transcripts are long enough to accommodate this number of probes. Furthermore, isoform specific exons are often not long enough to accommodate these probes for isoform specific staining. Without increasing the number of probes, background reduction such as by tissue clearing, split probes (e.g. in split-FISH), and using fluorophores with colors very different from the color of autofluorescence [99] can increase signal to noise ratio. There are also ways to boost signals without increasing the number of probes, the most common of which are branched DNA (bDNA), rolling circle amplification (RCA), and HCR. All of these methods non-covalently attach numerous fluorophores to the probe to amplify signal. Background reduction and signal amplification can be used in conjunction.

Branched DNA

Dating back at least as far back as to 1993 [110], early use of bDNA in ISH was to detect low copy number of viral genomes, eventually down to single copies [111]. bDNA signal amplification involves several steps of hybridization (Figure 7.17). First, usually some sort of bridge probe binds to the transcript itself. Then the primary amplifier binds to the bridge probe, leaving a long overhang. Then multiple secondary amplifiers bind to the primary amplifier on the overhang of the primary

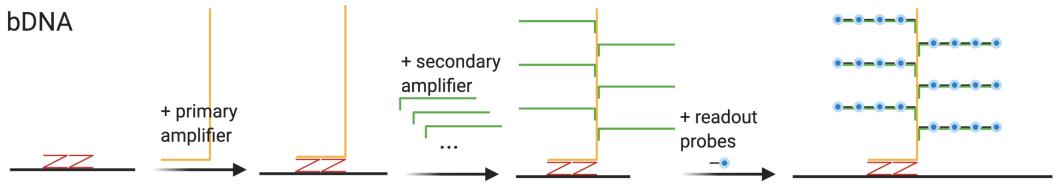

Figure 7.17: Schematic of bDNA. The Z probes are specific to RNAscope, but the other parts are generic to bDNA.

amplifier, and each secondary amplifier also leaves an overhang. Finally, multiple labeled readout probes bind to each secondary amplifier. This way, space available for hybridization of the readout probes is drastically expanded, allowing for more fluorophores per unit transcript length.

For FISH, a particularly influential bDNA method is RNAscope, introduced in 2012 for FFPE tissues, and is now commercially available from ACD [112]. In addition to bDNA amplification, RNAscope reduces background noise from non-specific hybridization by using 2 bridge Z probes in between the transcript and the primary amplifier, so the primary amplifier will only bind when both Z probes are present. An smFISH RNAscope method has been used to profile around 1000 genes in cell culture [113] and 49 genes in the mouse somatosensory cortex [114], although these experiments were not highly multiplexed and only one or a handful of genes distinguishable by fluorophore color were stained for in the same cells or sections; numerous cells and sections were stained to cover all genes in the gene panels. ACD RNAscope HiPlex v2 can profile 12 targets, but without barcoding. Up to 4 targets are imaged with 4 different fluorescent channels per round of imaging, then the fluorophores are cleaved for the next round of imaging. With fresh frozen tissue, this can be applied to up to 48 targets. bDNA has also made its way into more highly multiplexed smFISH, as a variant of MERFISH [115]. Here, the primary amplifier binds to the readout regions of the MERFISH encoding probe. Like in regular MERFISH (v2), the fluorophores are attached to the readout probes by a disulfide bond and removed by TCEP after each round of hybridization; the bDNA moiety is not removed. With bDNA amplification, only 16 probes per gene can detect about as many transcripts as with 92 unamplified probes [115].

Rolling circle amplification

Chronologically, the next of the popular signal amplification method is padlock probe RCA. Padlock probe was introduced in 1994 by Mats Nilsson as a way to

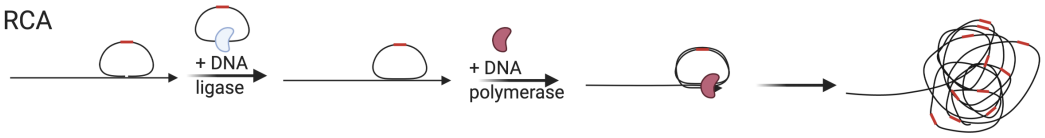

Figure 7.18: Schematic of RCA, here shown with target priming though a separate primer can also be used. Red segment is the gene barcode.

reduce background in ISH and to detect single nucleotide variants (SNVs) [116]. Both ends of of the padlock probe must hybridize to the target without terminal mismatches for the ligase to connect the ends of the probe to form a circle (Figure 7.18); thus padlock probe and RCA can detect SNPs and point mutations [117, 118]. The circle encloses the target like a padlock on a string, hence the name "padlock probe". Then probes that are not circularized are digested by an exonuclease. RCA was introduced in 1995 as a way to create tandem repeats and potentially point to the origins of tandem repeats in genomes, not seeming to have signal amplification in mind [119]. A primer anneals to circularized DNA and is then elongated by Φ29 DNA polymerase, and as the polymerase goes around the circle many times, many copies of the complimentary sequences of the circle are made (Figure 7.18). In 1998, padlock probes and RCA were united to create a method of signal amplification [120, 118].

In spatial transcriptomics, padlock probe and RCA were initially used for *in situ* sequencing (ISS) [121], but more recently adapted to smFISH. The padlock probe with the gene barcode is hybridized to *in situ* reverse transcribed cDNA as in ISS and hybridization-based ISS (HybISS) [122], or the mRNA itself as in SCRINSHOT [123], hybridization-based RNA ISS (HybRISS) [124], and barcoded oligonucleotides ligated on RNA amplified for multiplexed and parallel *in situ* analyses (BOLORAMIS) [125]. RCA can be initiated with the target cDNA itself as a primer or with a separate primer when the target is mRNA. Then readout probes are hybridized to the RCA amplified gene barcode, with [122] or without [123] a bridge probe. In Hyb(R)ISS and SCRINSHOT, multiple rounds of readout hybridization encode each gene with a sequence of colors as in seqFISH; although error correction is not discussed, the seqFISH error correction scheme can be easily adapted. Perhaps because of larger number of copies of the gene barcode sequence produced by RCA, Hyb(R)ISS and SCRINSHOT use 5 probes per gene, each with a 30 nt (HybISS, target sequences are proprietary information of CARTANA for HybRISS) or 40 nt (SCRINSHOT) region to target the transcript. While we are unaware of isoform

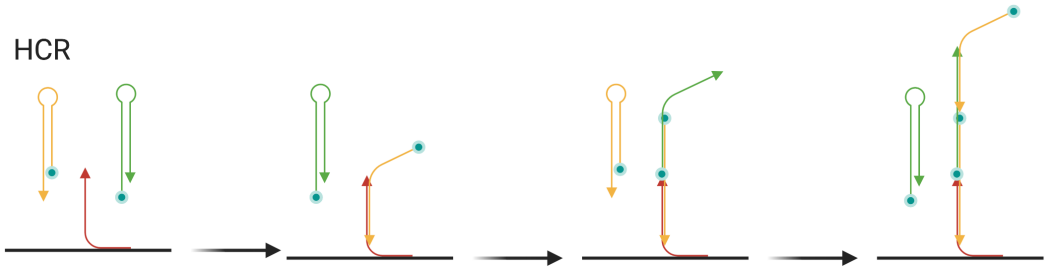

Figure 7.19: Schematic of HCR, showing 3 cycles, but this can continue indefinitely until H1 and H2 are exhausted. Arrow shows 5' to 3' direction.

specific studies conducted with Hyb(R)ISS or SCRINSHOT, isoform specific exons may more realistically accommodate the 5 probes.

Hybridization chain reaction

A third signal amplification method is HCR, introduced in 2004 [126], which has been adapted to seqFISH, giving rise to HCR-seqFISH. EASI-FISH also uses HCR for signal amplification. In singly labeled hairpins, the long stem is protected by the short stem, but can also hybridize with short stems of other hairpins (Figure 7.19). The long stem of H1 can hybridize to the short stem of H2, and vice versa (Figure 7.19). First, an initiator probe is hybridized to the transcript (24 per gene in the 2016 HCR-seqFISH study). Then the long stem of H1 hybridizes to the part of initiator not hybridized to the transcript, now leaving the short stem vacant. Then the long stem of H2 hybridizes to the vacant short stem of H1, and now the short stem of H2 is vacant for another H1. This cycle can continue indefinitely until H1 and H2 are depleted. This way, many fluorophores are tethered to the target transcript without increasing the number of probes bound to the transcript, thus amplifying signal.

Similarly, RCA can continue indefinitely until DNA polymerase is inhibited or removed or when deoxynucleotides are depleted. In contrast, the bDNA moiety has a controlled size and does not grow indefinitely until stopped. In both bDNA and HCR, the amplified moiety is still anchored on the target transcript. In contrast, since when the padlock probe encloses the target, the DNA polymerase is inhibited [120], the padlock must be dissociated from the target before RCA, or in the case of target priming, the target cDNA itself grows into the RCA hairball. As the hairball is not anchored to the original target, it can drift away and obscure the original location of the target. BOLORAMIS crosslinks the RCA amplicon to the cellular matrix to prevent the amplicon from drifting away.

Primer exchange reaction

Chronologically, a fourth signal amplification method is the primer-exchange reaction (PER), introduced in 2017 [127]. In PER, a hairpin with an overhang of domain A' and double strand enclosed domain B is used. Primer A complementary to domain A' of the hairpin anneals to the overhang, and a strand displacing polymerase copies domain B, extending domain A, thus creating a concatenation of A and B. Then the copied domain B competes with domain B in the hairpin until the concatemer AB is displaced by the hairpin's domain B. Then another hairpin with domain B' as the overhang can continue to extend the concatemer in the next cycle of the PER reaction. PER is used in smFISH method signal amplification by exchange reaction (SABER) [128] for signal amplification, where the primer is the target sequence binding to the transcript has a domain A at the 3' end, and the hairpin has a domain A' overhang and another A and A' in double strand instead of B and B', so multiple copies of domain A is concatenated to the primer. Then fluorescent readout probes anneal to the multiple copies of domain A from PER, thus greatly increasing the number of fluorophores that can bind to the same transcript target. Branched probes as in bDNA can be applied to the PER concatemers for additional signal amplification. The short readout probes can be stripped without stripping the longer primary probes binding to the transcripts for multiple rounds of hybridization to image more genes than fluorophores.

Optical crowding

As we have seen, smFISH-based spatial transcriptomics has been scaled to around 10,000 genes and can potentially be scaled to the whole transcriptome. With increasing number of mRNA molecules visualized, it's also increasingly likely for different target molecules to be so close to each other that their fluorescent spots overlap or are even within the diffraction limit of the optical microscope and appear as one point. This is the problem of optical crowding, and some existing ways to mitigate this problem are summarized below.

As already mentioned, SRM is not susceptible to this problem [93], though access to SRM is not as common as access to regular confocal or epifluorescent microscopes. Another simple strategy is to select the most highly expressed genes from RNA-seq. These genes are imaged separately with smFISH, with one color and one round of hybridization per gene instead of combinatorial barcoding, as was done in the first MERFISH study [97]. However, with increasing number of highly expressed

genes, this method becomes increasingly laborious. Also as already mentioned, in seqFISH+, only 1 in 60 mRNA molecules of interest light up in each channel and round of hybridization (20 pseudocolors per channel and 3 channels), and the transcript spots can be computationally super-resolved, thus reducing optical crowding [98].

Another strategy is to allow transcript spots to overlap but computationally resolve them, as in corrFISH [129], BarDensr [130], ISTDECO [131], and Composite In Situ Imaging (CISI) [132]. In corrFISH, Transcripts of highly expressed genes encoding ribosomal proteins were visualized with sequential hybridization and 2 colors but not every gene lights up in each round of hybridization; each gene is encoded by one color and a sequence of 0's (absence of fluorescence) and 1's (presence) of that color. Then images from different rounds of hybridization in the same FOV are correlated to identify transcripts that are 1's in both rounds amidst transcripts that are not 1's in both rounds. To the best of our knowledge, after its conception, corrFISH has not been applied to generate any new high throughput dataset.

A more recent method, BarDensr, models the observed brightness of potentially mixed spots in terms of the point spread function (PSF), codebook, unknown spot density, probe washing, background, and per round per channel gain. Then the unknown spot density and deconvolution of barcodes at mixed spots are inferred by maximizing sparsity of the spots in space (most voxels don't have spots) while keeping reconstruction loss of the observed brightness sufficiently low. BarDensr is very recently published, and, as of writing, we are unaware of studies that used the method. ISTDECO is similar but only uses a Gaussian PSF, codebook, and background.

CISI uses seqFISH-like barcoding, but does not even require spot detection. Gene abundance is computationally inferred with compressed sensing. First, an autoencoder is trained on composite images with different channels. Then in the latent space inferred by the autoencoder, the channels are decompressed with compressed sensing principles and decoded into genes with the decoder branch of the trained autoencoder. The barcodes and genes must be carefully chosen from an existing dataset. The genes must be described by a small number of coexpression modules so module activity is sparse. Inferring the sparse module activity before inferring individual gene levels at the decompression step is more tractable than directly inferring individual gene abundances.

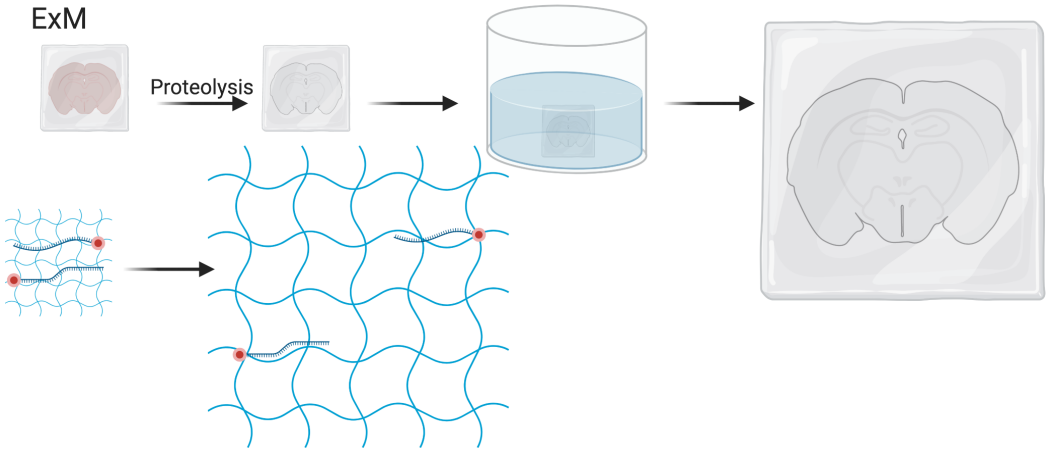

Figure 7.20: Schematic of expansion microscopy.

A strategy that has been reused is expansion microscopy (ExM). When a polyelectrolyte gel is dialyzed in water, it expands as its polymer network changes into extended conformations [133]. First, the tissue is infused with monomers of the gel. Then with small molecule linkers, molecules of interest such as fluorophores and RNAs can be covalently incorporated to the polymer network over the course of free radical polymerization. After the gel forms, proteins in the tissue are digested to homogenize mechanical properties of the gel and to clear the tissue to reduce autofluorescent background. Then the gel is soaked in water to expand, linearly expanding 3 to 4.5 times on each side [133, 134] (Figure 7.20). This way, transcripts attached to the gel are physically separated, avoiding optical crowding. ExM has thus been adapted to MERFISH for this purpose [135], as well as EASI-FISH. In addition, EASI-FISH was used to quantify transcripts in 300 μm thick brain slices and imaging was accelerated with light sheet microscopy. However, a disadvantage of ExM is that each FOV now covers less of the original tissue, thus increasing imaging time. Furthermore, the expanded gel would continue to expand during the rounds of hybridization. As the expansion is non-linear and non-isotropic, barcode decoding is challenging as it's difficult to match transcript spots across rounds of hybridizations.

Usage of smFISH-based techniques

As already noted, the number of genes whose transcripts can be possibly quantified simultaneously in the same piece of tissue with highly multiplexed smFISH-based technology has increased over time (Figure 7.21). The number of cells that can be imaged in one study has also increased (Figure 7.22). However, in practice, the

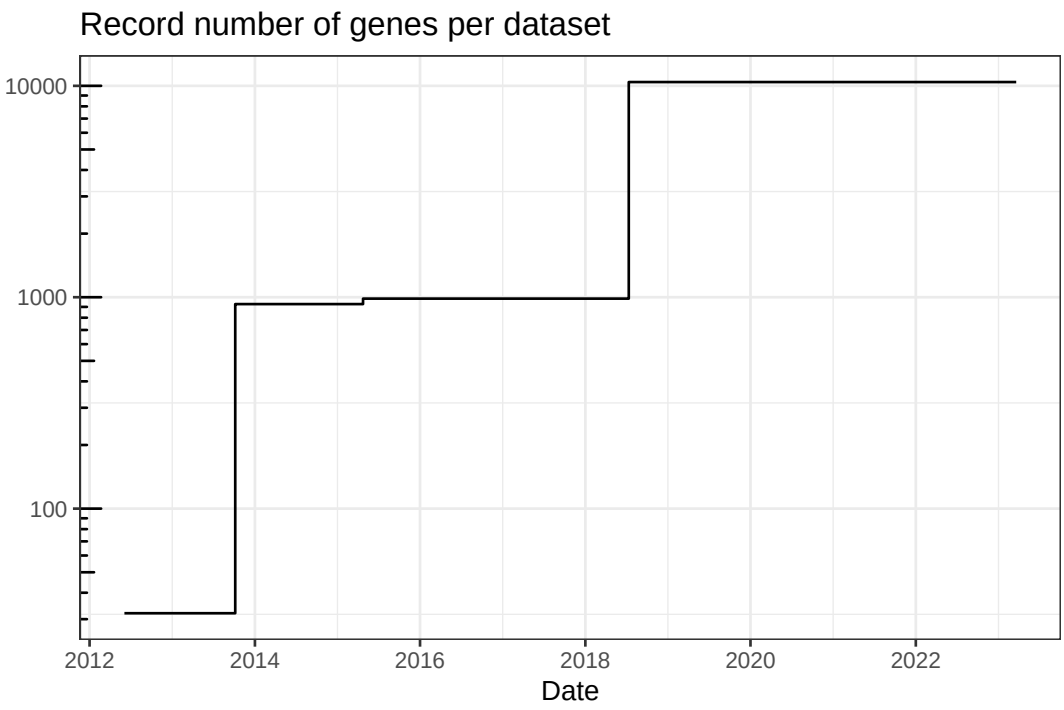

Figure 7.21: Record number of genes per dataset quantified by smFISH-based techniques over time.

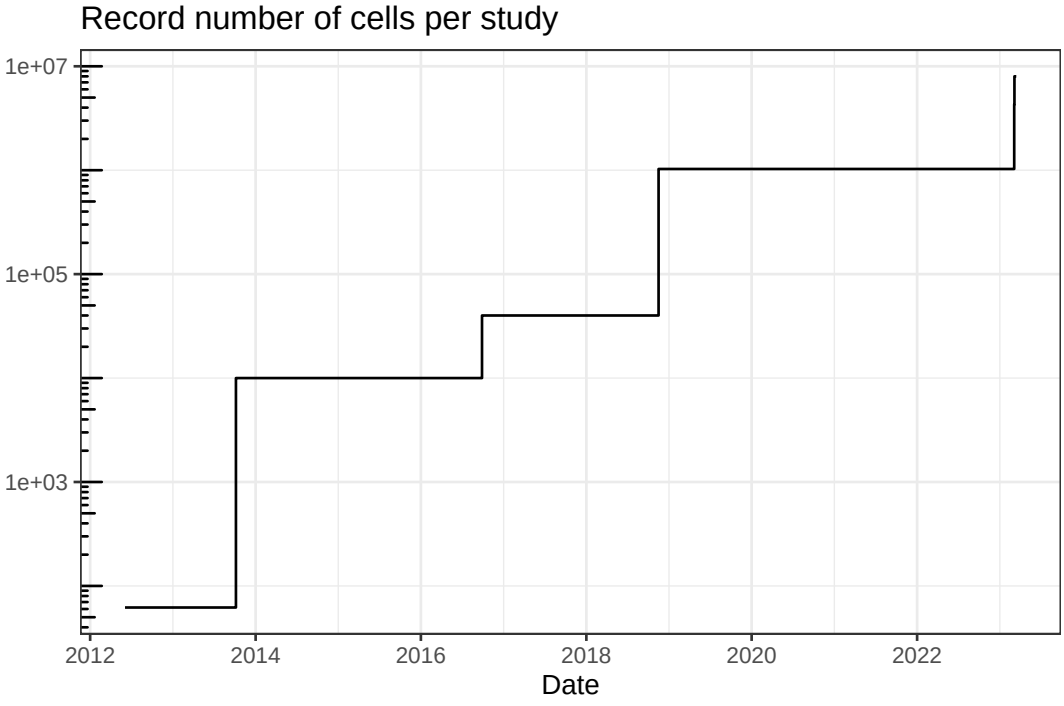

Figure 7.22: Record total number of cells per study profiled by smFISH-based techniques over time.

actual number of genes and cells profiled has not significantly increased (Figure 7.23, Figure 7.24). These plots only show papers that reported the number of cells and genes in the main text; if we download and process all publicly available datasets associated with such papers, the trends might change, although figures of papers that do not report the number of cells (number of genes is usually reported in smFISH and ISS studies) don't seem to indicate that the trend would change significantly. Moreover, as discussed in Section 7.8, some of the studies used smFISH-based methods to visualize DNA loci and 3D chromatin structure alongside transcripts. The number of genes here is for the transcripts, including when only introns are targeted.

An earlier version of the plot of number of genes over time plotted the mean number of genes for each study, due to difficulty in defining what constitutes a dataset. However, since that version caused confusion as sometimes one study profiled very different number of genes in different experiments, we decided to give some definition of "dataset" and not to plot the mean. Here a "dataset" means either a different tissue, cell type, experimental or clinical condition, or a separate experiment profiling a different number or set of genes in the same study. One dataset can involve multiple sections and individuals.

The trend line looks pretty flat. Although studies quantifying a very large number of genes tend to be recent, many other studies profiling fewer genes pulled the line down. The slope (with all data, outliers and all) is not significantly different from 0 (t-test), after log transforming the number of genes per dataset.

```
##
## Call:
## lm(formula = log(n_genes) ~ date_published, data = smfish)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.868 -1.196  0.033  1.042  4.735
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.566e+00  2.338e+00   1.525    0.129
## date_published 5.364e-05  1.249e-04   0.429    0.668
##
```

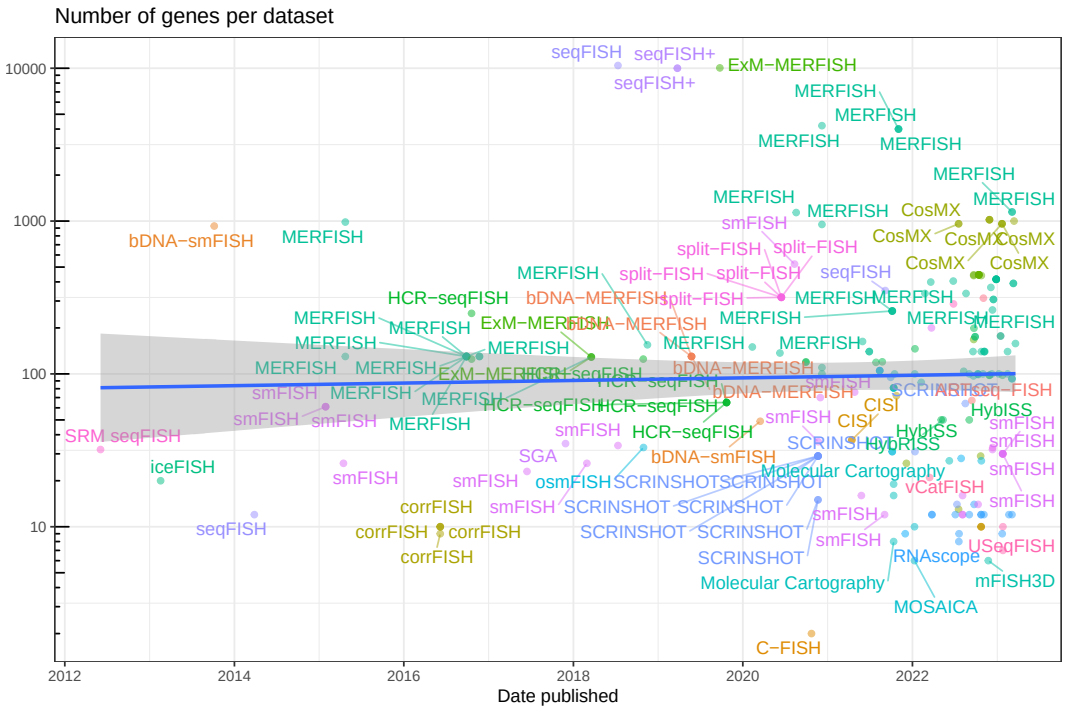

Figure 7.23: Number of genes per datasets in each study, over time. Gray ribbon is 95% confidence interval (CI). The points are translucent; more opaque points are multiple datasets from the same study.

```
## Residual standard error: 1.583 on 222 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.0008302,  Adjusted R-squared:  -0.003671
## F-statistic: 0.1845 on 1 and 222 DF,  p-value: 0.668
```

How total number of cells profiled in each study that reported the number of cells in the main text is shown here. The total number across datasets is used because sometimes number of cells per dataset is not reported.

After log transforming the total number of cells per study (when reported), whose distribution is very right skewed, it does seem that the total number of cells increased with time (Figure 7.24). New smFISH-based techniques in our database since 2021 are all optimized for features other than larger number of genes and are applied to relative small numbers of genes in demonstration. For instance, EASI-FISH is optimized for thick brain sections ([109]Y. Wang et al. 2021). par-seqFISH is optimized for bacteria ([136]Dar et al. 2021). CISI is optimized for reducing the number of imaging cycles and avoiding direct spot calling and has not been demonstrated on large number of genes [132]. The distinctive feature of MOSAICA

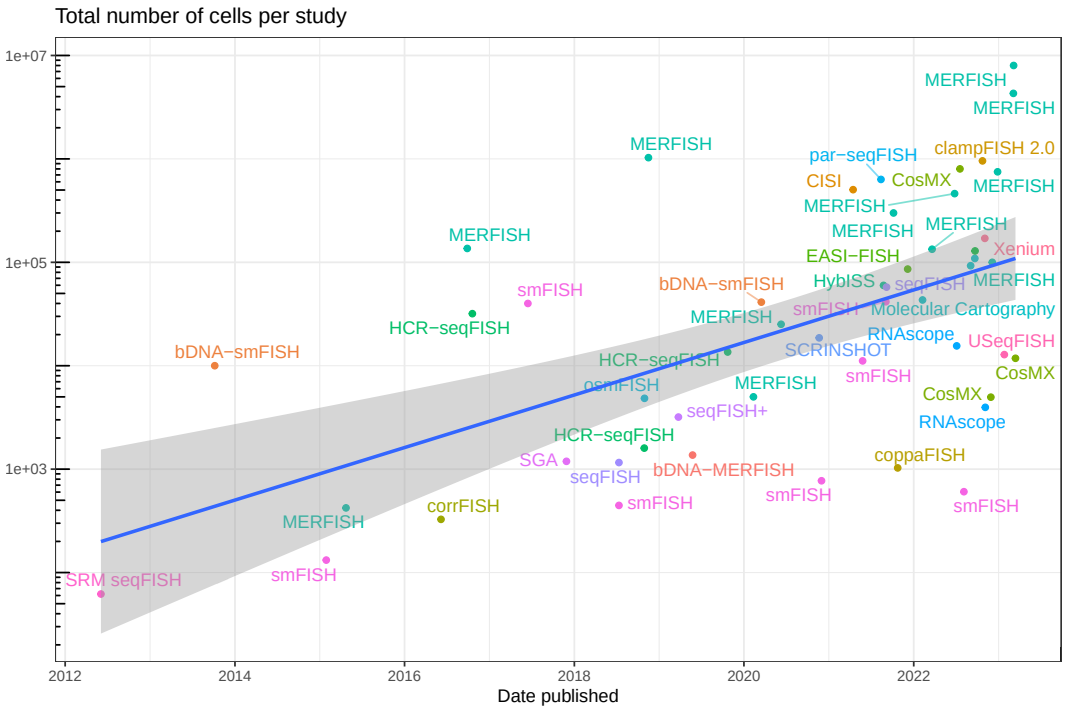

Figure 7.24: Total number of cells per study profiled by smFISH-based techniques over time.

is to use both the color and the lifetime of the fluorophores and is only demonstrated to be 10-plex [137]. Recent applications of existing techniques also tend to feature larger number of cells but only hundreds of genes (e.g. 368 genes in [138]), where the MERFISH dataset is complementary to scRNA-seq datasets of the same tissue, using marker genes from scRNA-seq clusters.

```
##
## Call:
## lm(formula = log(n_cells) ~ date_published, data = sum_cells)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8383 -1.6338  0.0016  1.8604  4.7738
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.951e+01  6.067e+00  -3.216  0.00233 **
## date_published  1.601e-03  3.286e-04   4.872 1.24e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.273 on 48 degrees of freedom
## Multiple R-squared:  0.3309, Adjusted R-squared:  0.317
## F-statistic: 23.74 on 1 and 48 DF,  p-value: 1.244e-05
```

MERFISH is the smFISH-based technique used in the most institutions (Figure 6.7), although most of the smFISH-based techniques barely spread beyond their institutions of origin, if at all (Figure 7.26). The following advantages and disadvantages of smFISH-based techniques may explain these trends in usage. Advantages and disadvantages of individual smFISH-based techniques reviewed so far are summarized in Table 7.1.

MERFISH has been commercialized by Vizgen and has spread much more far and wide than seqFISH and HybISS; another commercial technology, Molecular Cartography also also spread far and wide (Figure 7.27). While MERFISH is mostly used in the US, Molecular Cartography is mostly used in Europe, in accordance with the location of their companies.

smFISH-based techniques have the following advantages. First, smFISH, especially with larger number of probes, have nearly 100% detection efficiency of transcripts [93], i.e. detecting almost all transcripts that are present. Different ways to evaluate efficiency of spatial transcriptomics techniques have been reported. The reported "efficiency" of MERFISH was estimated by the average ratio between the number of transcripts per segmented cell detected by MERFISH and those detected by smFISH in the same cell type for 10 genes. With combinatorial barcoding, however, the efficiency is decreased. Studies for other techniques may use different ways to estimate efficiency. Compared to smFISH, MERFISH version 2 with HD4 code has about 95% detection efficiency on 130 genes and 92 probes per gene, although the efficiency dropped to ~25% with the HD2 code that can encode nearly 1000 genes but can only identify but not correct errors [99, 140]. When scaled to 10,050 genes, MERFISH has around 79% detection efficiency [115]. As for HCR-seqFISH, the efficiency is around 84% (smFISH and HCR-seqFISH were performed in the same cell for 5 genes) [96], and for seqFISH+, around 49% (slope of line fitted to average transcript count per cell in seqFISH+ vs. smFISH for 60 genes) [98]. Nevertheless, this is much better than the efficiency of ST, which is around 6.9% compared to smFISH in the 2016 ST study (transcript counts for 3 genes in ST spots

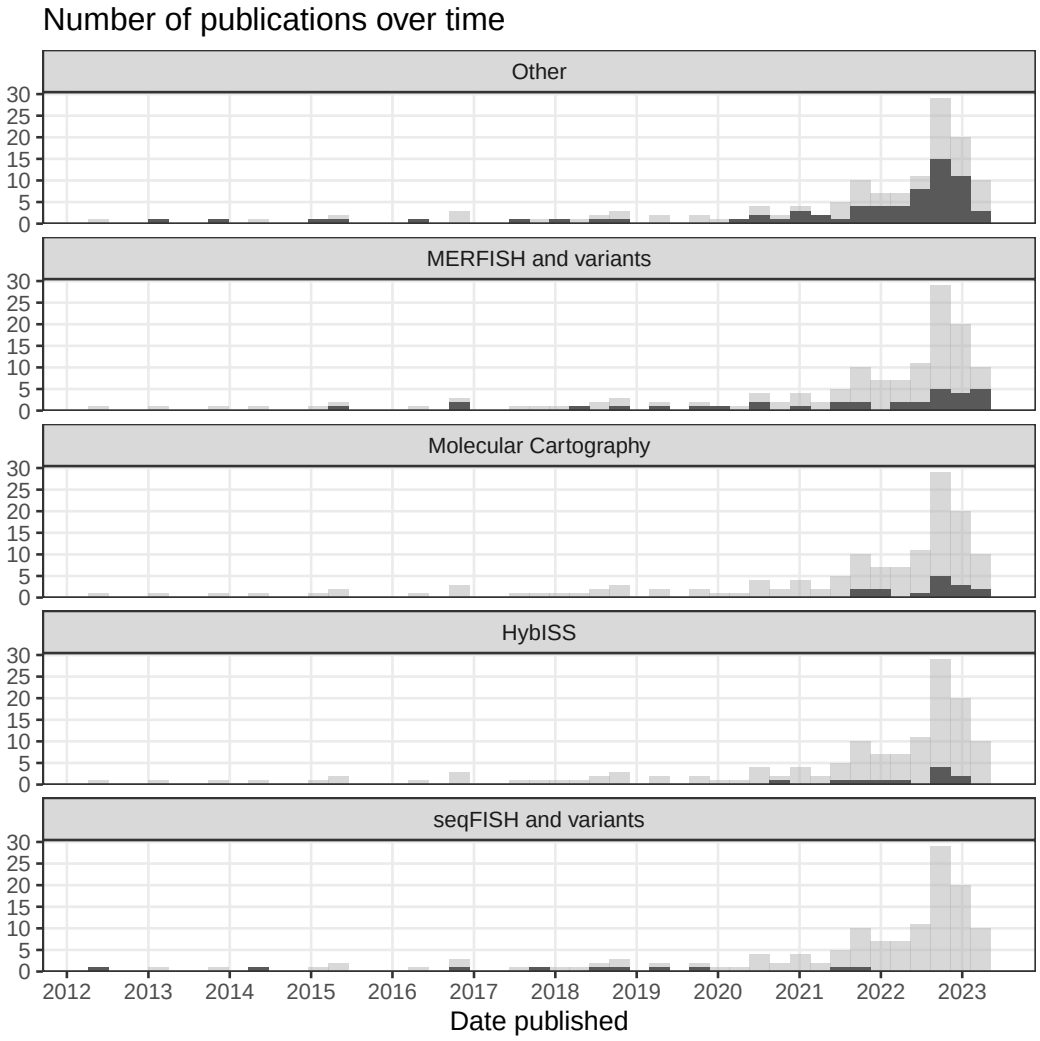

Figure 7.25: Number of publications over time, broken down by technique type. Preprints are included, and the gray histogram in the background is the overall trend of all smFISH-based techniques. Bin width is 90 days.

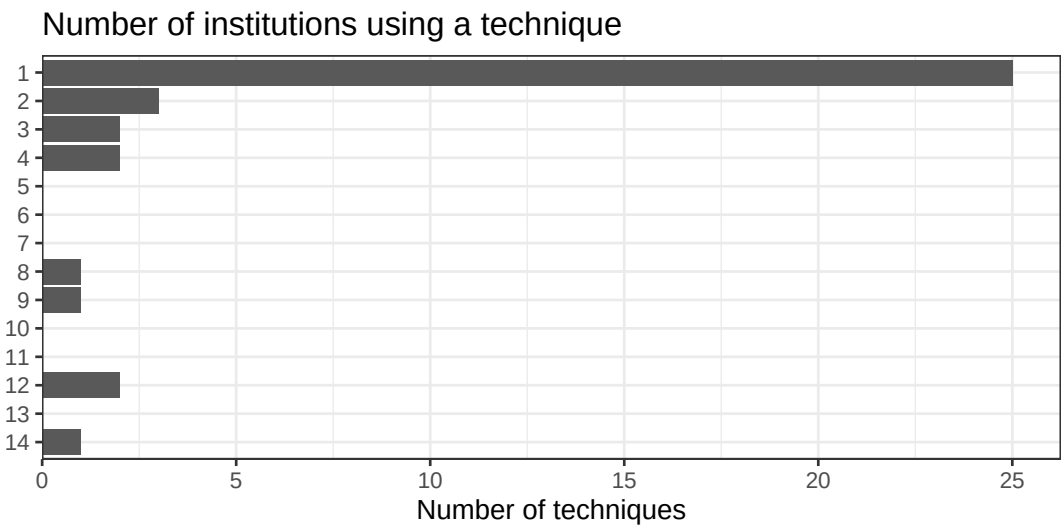

Figure 7.26: Number of techniques that have been used by each number of institutions; most techniques have only been used by 1 institution, i.e. the institution of origin.

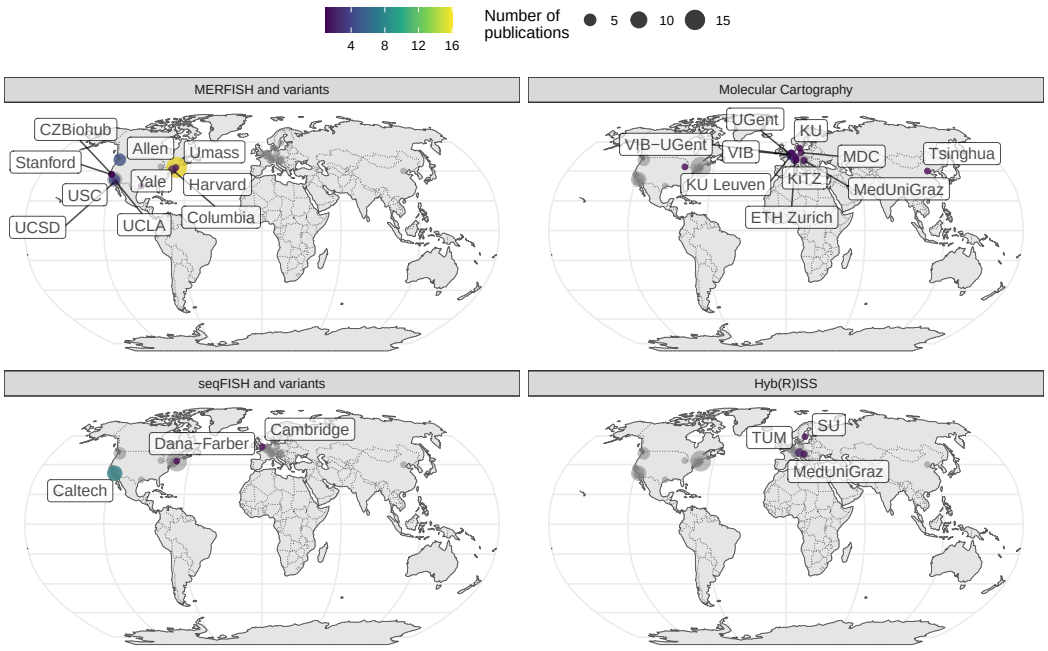

Figure 7.27: Geographical locations of institutions that used certain techniques. Point area is proportional to number of publication from the city of interest. Gray points in the background is all publications using smFISH-based techniques. The cities and institutions labeled are those of the first author. Note that for seqFISH, the hidden Markov random field (HMRF) study at Dana Faber [139] and the mouse embryo study [105] had collaboration with Long Cai's group at Caltech, so the dataset was most likely still collected at Caltech.

were compared to those from smFISH of 100 μm diameter discs at comparable brain regions in an adjacent section) [141]. To put the 6.9% in context, from ERCC spike ins and in some cases comparison to smFISH, scRNA-seq methods such as Drop-seq, 10X, inDrop, CEL-seq, and CEL-seq2 have capture efficiency of between 3% and 25% [142, 143, 144, 145, 146]. Thus smFISH-based spatial transcriptomics methods can be much more efficient than scRNA-seq, though efficiency of RCA based smFISH compared to regular smFISH has not been reported.

Second, since individual transcripts are imaged and counted, smFISH-based methods are highly quantitative and records subcellular localization of transcripts. While most smFISH-based spatial transcriptomics studies analyze data at the cellular gene count level, not using subcellular transcript localization, cells have been shown to show great variation in subcellular localization of transcripts of the same set of genes and a number of "archetypal" patterns have been described [147, 148, 149].

The following disadvantages may explain why smFISH-based spatial transcriptomics has not been widely used on large number of genes (Figure 7.23), and why MERFISH is the most used technique (Figure 7.25). First, multiple rounds of hybridization and high magnification mean that data collection is time consuming. MERFISH version 2 greatly sped up imaging, as version 1 requires higher magnification and needs to photobleach fields of view (FOV) one at a time; one FOV in version 1 is 40 μm $\times$ 40 μm, while one FOV in version 2 is 223 μm $\times$ 223 μm. Version 2 also cut imaging time in half by using 2 colors, targeting 2 bits per round. This way, for 130 genes and 40,000 cells, MERFISH took about 18 hours [99], while HCR-seqFISH would take days because of overnight hybridization after probes are stripped for each round of hybridization although the seqFISH barcode is much shorter. When scaled to 10,000 genes, MERFISH takes 23 rounds of hybridization [100], while seqFISH+ takes 80 rounds [98], although because ExM was used for MERFISH in this case to reduce optical crowding, expanding the area to be images ~4 fold, the actual imaging time of ExM-MERFISH and seqFISH+ here may have been comparable. Perhaps MERFISH has been scaled to larger number of cells and used in more studies beyond the institution of origin (Figure 7.27) because of the higher detection efficiency and shorter imaging time.

Second, with increasing area of tissue and number of genes covered, smFISH-based spatial transcriptomics generates terabytes of images—for each FOV, there is an image for each channel, z-plane, and round of hybridization. Images from the MERFISH dataset of 40,000 cells and 130 genes took 2 to 3 days to process on a

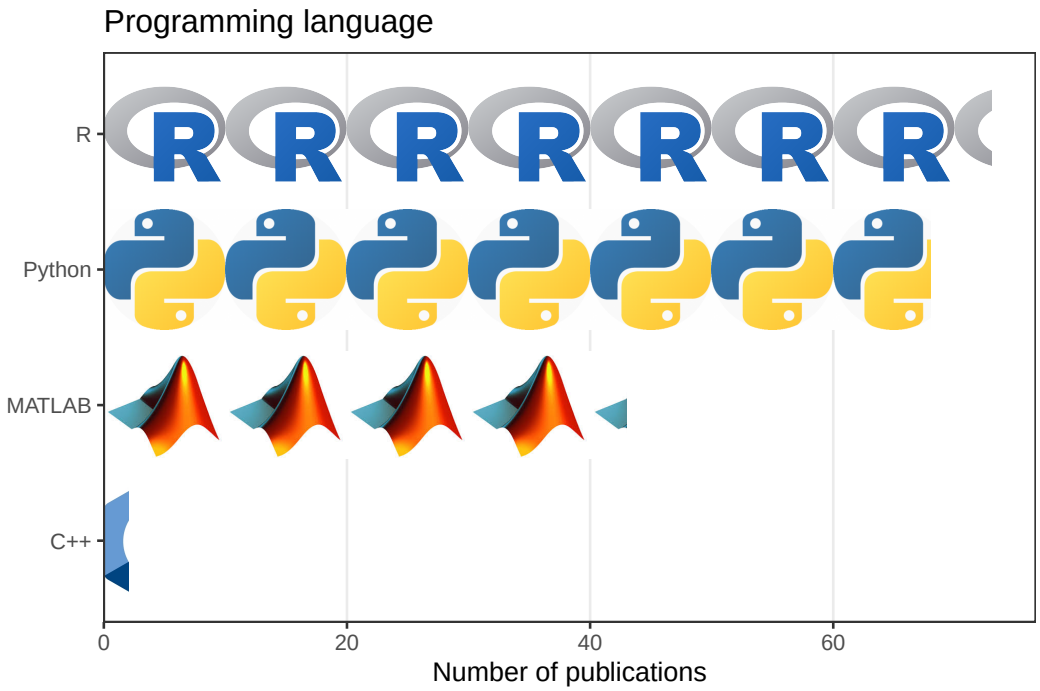

Figure 7.28: Number of publications using smFISH-based techniques that used each of the 50 most common programming languages. Each icon stands for 5 publications.

multi-core server, although the number of cores was not stated [99]. In contrast, it takes hours, or even just minutes, to process the fastq files of a scRNA-seq dataset to get the gene count matrix [150], nor do the fastq files take up so much disk space. So for the user, processing the most upstream form of data is much more challenging for highly multiplexed smFISH than scRNA-seq. Until 2019, software to process such images and to decode the combinatorial barcodes was typically written in the proprietary programming language MATLAB (Figure 7.28), and poorly documented, so it was difficult for people outside the lab of origin to use.

More recently, Python is replacing MATLAB as the programming language of choice to write such image processing software. The Chan Zuckerberg Initiative developed starfish in Python as a unified framework to process smFISH-based spatial transcriptomics data [151]. However, image processing pipelines specific to each technology have been developed instead, such as MERlin for MERFISH [100] and IRIS for ISS [152], and image stitching is performed separately such as with MIST [153] or BigStitcher [154] if needed as `starfish` does not directly support multiple FOVs. `starfish` has been used by the HybISS group [155, 122] for spot calling,

decoding, and cell segmentation, and by the CISI group for spot calling [132] and CellProfiler for cell segmentation. In contrast, for scRNA-seq, there are popular data processing tools that apply across technologies, such as STAR (wrapped by Cell Ranger) [156], alevin [157], and `kallisto` [150]. Furthermore, even with an open source and interoperable image processing pipeline, cell segmentation, which is essential to obtaining the gene count matrix commonly used in data analysis, is challenging.

Third, custom fluidics systems have been used for the numerous rounds of hybridization [98, 99, 107]. These custom fluidics and pump systems are not commercially available and need to be built by any lab that wishes to adopt the smFISH-based technologies. To the best of our knowledge, there are no core facilities that perform smFISH-based spatial transcriptomics. Thus for the user, adopting an smFISH-based spatial transcriptomics technique means not only learning a new syntax to process images, made difficult in some cases by the cost of MATLAB and lack of documentation, but also setting up a complex custom fluidics system integrated to a microscope, which may not be feasible at microscopy cores. However, this is changing with commercial Vizgen MERFISH, the Rebus Esper spatial omics platform, Molecular Cartography of Resolve Biosciences, 10X Xenium, and Nanostring CosMX, with convenient automated imaging machines and reagent kits. Rebus Esper was used to automate osmFISH in [158], and claims to have less than one hour of hands on time and be able to return a gene count matrix for 100,000 cells with spatial coordinates of the cells within 2 days. While Molecular Cartography is smFISH-based, it's not clear from its website how it works and it only profiles 100 genes. Aria from Fluigent can also be potentially used to automate highly multiplexed FISH. MERFISH and Molecular Cartography have spread far and wide after commercialization, and we expect other commercial smFISH platforms to spread as well.

Fourth, to profile large numbers of genes, numerous probes need to be designed, especially when dozens of probes are used for each gene to enhance signal. Probes with fluorophores are expensive as well and larger quantity of them are needed with signal amplification. These probes are an expensive one time purchase, and might not be worthwhile if a lab does not perform highly multiplexed smFISH very often. A core facility with a good collection of probes can reduce cost to individual labs, but to reiterate, as of writing, we are unaware of any core facility performing highly multiplexed smFISH techniques such as MERFISH (except NeuroTechnology Studio

Table 7.1: Pros and cons of smFISH-based techniques.

Technique	Pro	Con
HCR-seqFISH	Relatively high efficiency (84%), fewer rounds of hybridization, error correction	Lower efficiency than MERFISH, time consuming to re-hybridize probes to target after stripping
seqFISH+	Avoids optical crowding, scalable	Lower efficiency (49%), numerous rounds of hybridization
MERFISH	High efficiency (95%) with HD4 code, error correction, version 2 relatively fast, scalable, commercialized	Numerous rounds of hybridization, numerous probes requiring long transcripts though this is resolved by bDNA signal amplification
ExM-MERFISH	Avoids optical crowding, clears tissue	Each FOV contains less of the original tissue
HybISS	Only 5 probes per gene, applicable to isoform specific exons, padlock probe reduces background, lower magnification when imaging (20x and 40x, while MERFISH uses 60x), can discern SNPs	Error correction not reported, amplicon takes up space and might drift away if not cross linked
HybRISS	Avoids inefficiency of reverse transcription, better signal to noise ratio and more transcripts detected then HybISS.	Padlock probe sequences are proprietary to CARTANA
bDNA-smFISH	Commercial RNAscope kit, reduces background and amplifies signal, amplified moiety does not grow indefinitely	Except for bDNA-MERFISH, it has not been used in a highly multiplexed setting

at Brigham Health for MERFISH) and seqFISH. Cost of probes could be a reason why recent applications of highly multiplexed smFISH techniques did not profile larger number of genes. Finally, smFISH-based techniques require a pre-defined list of genes and probes, so unlike in RNA-seq, novel transcripts would be missed.

So far we have reviewed studies that showcase new techniques and technical improvements such as signal amplification and resolving optical crowding. Some smFISH-based techniques have been used in studies that focus on biological problems rather than new techniques. HCR-seqFISH has been used twice in biological studies, in chicken neural tube (35 genes) [108] and mouse T cell precursors (65 genes) [159] though both were conducted within Caltech, the institution of origin. Moreover, spatial location of cells is not necessarily a reason to use HCR-seqFISH; Zhou et al. used HCR-seqFISH because of the high detection efficiency compared to scRNA-seq in dissociated FACS sorted T cell progenitors, so when spatial information is already lost. More recently, pseudocolor seqFISH was used in a mouse embryo atlas at University of Cambridge (though Long Cai is still a coauthor), finally moving beyond the stage of testing into new biological research [105]. Combinatorial barcoding has also been used to profile bacterial species in the microbiome by targeting rRNAs, though this does not profile the transcriptome, nor is it single molecular [160, 161]. For spatial transcriptome in bacteria, a new version of seqFISH, par-seqFISH, was developed to profile 105 genes in the biofilm bacterium *Pseudomonas aeruginosa* [136]. This may open the way to spatial transcriptomics in not only biofilms, but in the microbiome in general.

MERFISH has been used more broadly in biological studies. Within Harvard, the institution of origin, MERFISH has been used to create atlases of the hypothalamic preoptic region (155 genes) [162] and the primary motor cortex (MOp) (258 genes) [163] in mice, and adapted to stain for chromatin conformation and transcription foci (introns) [164]. Outside Harvard (Figure 7.27), MERFISH has been used to study how gene expression variability relates to cell state in cell culture [140] and used in conjunction with smFISH-based chromatin tracing to study the relationships between chromatin compartmentalization and gene expression [165].

After its inception, HybISS became part of a single-cell atlas of the developing mouse nervous system [166]. This atlas is mostly scRNA-seq data, but 119 genes were stained with HybISS to validate secondary organizers discovered via scRNA-seq. As part of the HCA, HybISS has also been used for the human adult temporal lobe [167] and fetal forebrain [155].

7.3 *In situ* sequencing

In contrast to smFISH-based techniques, techniques reviewed in this section determine the sequences of the target transcript or the gene specific barcode by *in*

situ sequencing by ligation (SBL) or sequencing by synthesis (SBS) to distinguish between transcripts of different genes. This section reviews 3 *in situ* SBL strategies, SOLiD, cPAL, and SEDAL, and the spatial transcriptomics techniques using them.

SBL relies on the specificity of the DNA ligase, so ligation only occurs when both sequences to ligate match the template in the vicinity of the site of ligation. Prior to SBL, this specificity was used to detect SNVs that would otherwise be missed as ISH probes can tolerate some mismatches. A technique using ligation of two oligonucleotides to detect SNVs was introduced in 1988 by Ulf Landegren [168], laying the foundation of SBL (Figure 6.3). The padlock probe came in the same tradition of SNV detection, and Mats Nilsson worked with Landergren when creating the padlock probe [116].

Almost all spatial transcriptomics techniques based on SBL require *in situ* reverse transcription of the mRNAs as ligation with RNA as template is inefficient. As already mentioned in Section 7.1, IVT amplification of transcripts from single-cells for expression profiling originated in the Eberwine group [28, 29], where rather than LCM, the cDNAs from the single-cells were reverse transcribed *in situ* during electrophysiological recording before the cellular content was aspirated for IVT amplification. This was built upon the *in situ* reverse transcription technique from the Eberwine group in 1988 [169], where the cDNA of proopiomelanocortin (POMC) was radiolabeled so the spatial distribution of the mRNA was visualized on an autoradiograph. This made most *in situ* SBL techniques possible, which instead of radioactivity, use gene barcodes to locate the transcripts in a multiplexed and safer way.

SOLiD and FISSEQ

The earliest proposal of SBL we are able to locate is a patent filed in 1995 describing a method similar to sequencing by oligo ligation detection (SOLiD). An initiator oligonucleotide hybridizes to the template to be sequenced, and is extended by ligation to a 9-mer probe with a label such as a fluorophore that indicates one or two nucleotides of the probe [170]. The probe has a blocking moiety so only one probe is ligated in each cycle. Then the blocking moiety is removed so the initiator can be further extended by ligation in the next cycle. As mismatches in the probe inhibit ligation, the nucleotide of interest in the probe can be read off from the label after probes that are not ligated are removed. This can determine every 9th nucleotide in the template, and with 9 different initiators, each out of phase by one nucleotide, the

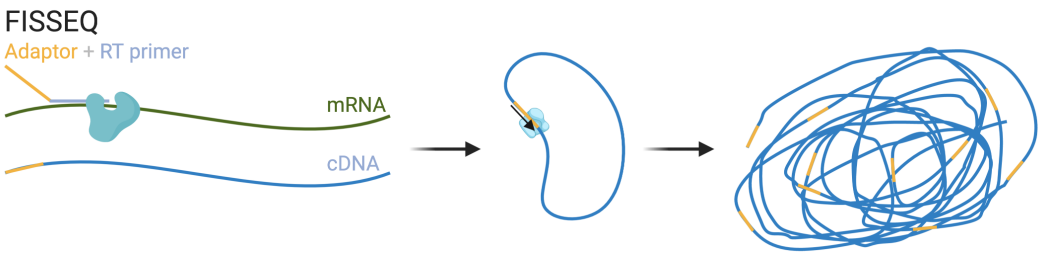

Figure 7.29: Schematic of RCA in FISSEQ.

sequence of the entire template can be determined. However, this method existed only on paper, while since 2006, Applied Biosystems (Applera) seemed to have developed SOLiD independently from that patent after acquiring Agencourt, which developed a sequencing by ligation method that would be the foundation of SOLiD [171].

In 2014, single-cell resolution and transcriptome wide spatial transcriptomics was far out of reach (Figure 7.21). An attempt to reach this goal was fluorescent in situ sequencing (FISSEQ) [172]. A universal adapter and random hexamer reverse transcription (RT) primer was hybridized to the mRNAs to reverse transcribe them into cDNA (Figure 7.29). Then the cDNA, now with the adaptor on the 5' end, is circularized, and amplified with RCA with a primer complementary to the adaptor. Then again, with sequencing primers receding into the adaptor, SOLiD is used to sequence the cDNA amplicons in situ.

In SOLiD, color of the fluorophore encodes the two 3'-most bases of 8-mer probes with other bases degenerate (Figure 7.30). Once a probe perfectly matching the target right after the primer, the probe is ligated to the primer and the fluorescent signal is recorded. Then the fluorophore and the nearest 3 bases of the probe are cleaved off. In the next cycle, a new matching probe is ligated to the now extended primer. This is continued until the end of the target, for 7 cycles per primer in the case of FISSEQ [173]. For the first 7 cycles, the primer matches the adaptor (N). Then the primer N, extended for 7 cycles is stripped, and a new primer receding one nucleotide to the 5' end of the adaptor (N-1) is added in cycle 8. Again 7 cycles of ligation are performed and the extended primer N-1 is stripped after cycle 14 to make room for N-2. For N-2, N-3, and N-4, a bridge oligo is used so the target with unknown sequence, rather than the adaptor with known sequence, is interrogated by the probes. With N through N-4, the entire target is covered. With the fluorescent signals recorded from the rounds of ligation, and the knowledge of the last nucleotide

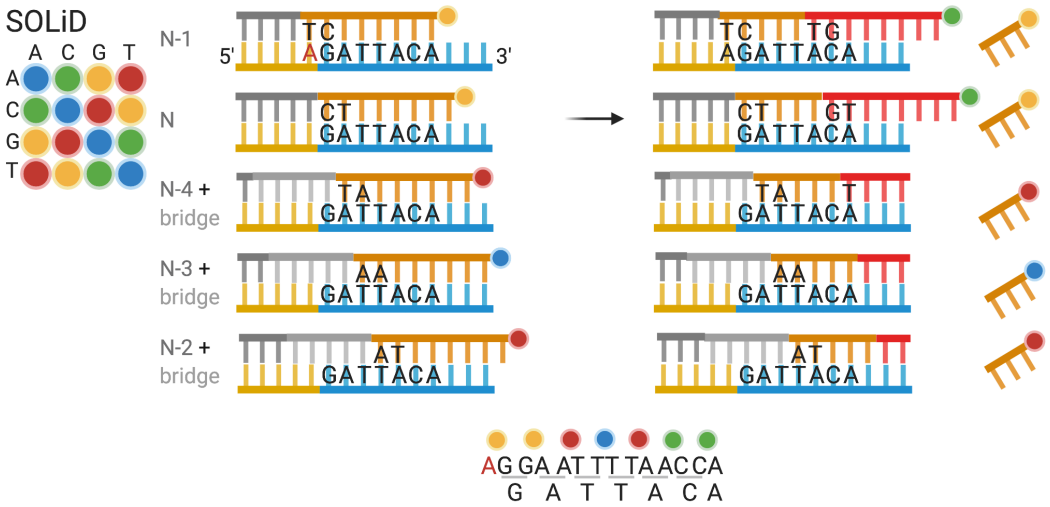

Figure 7.30: Schematic of SOLiD sequencing, determining the sequence GATTACA. The rows are arranged in the order of 5' to 3' positions of the first fluorescent probe, but the actual hybridization and ligation can take a different order. As part of the constant region, the 'A' highlighted in red is known.

of the adaptor interrogated by the first ligation to primer N-1, the sequence of the target can be determined. Figure 7.30 shows how SOLiD determines the sequence "GATTACA". As already mentioned in the smFISH section, with increasing number of genes profiled, optical crowding is increasingly a problem. To mitigate optical crowding, the primer N can have one or more degenerate bases at the 5' end reaching into the target; with one degenerate base, only 1/4 of the amplicons are sequenced. With two bases, this would be 1/16. This is repeated to cover all transcripts, but increases imaging time.

While FISSEQ may seem a promising approach to reach the goal of single-cell resolution and transcriptome wide spatial transcriptomics that unlike smFISH-based techniques, is not limited by pre-defined gene panels, it has been largely dormant since its inception due to the following disadvantages. First, SOLiD has fallen out of favor because of limited read length when used in situ (5-30 nt), propagation of errors from previous cycles [174], and difficulty in sequencing panlindromic sequences [175]. SOLiD was chosen for FISSEQ because it works well at room temperature; though SBS supports longer read lengths, it requires a heated stage [173]. Second, FISSEQ is extremely inefficient, over 20 times less sensitive than scRNA-seq and two orders of magnitude less sensitive than 2013 Nilsson ISS (discussed later in this section) [173], in part because of inefficiency of random RT priming [172] and tight packing of amplicons [174]. Furthermore, as ribosomal RNA (rRNA)

is not depleted, ~40-80% of FISSEQ reads are rRNA [172, 173]. As about 200 mRNA reads are detected per cell in FISSEQ without rRNA depletion, compared to about 40,000 in scRNA-seq, and suppose detection efficiency of scRNA-seq is 10%, then detection efficiency of FISSEQ might be around 0.005% [173]. Third, highly abundant genes involved in translation and splicing is depleted in FISSEQ compared to bulk RNA-seq [172]. Finally, FISSEQ imaging is time consuming, taking 2 to 3 weeks if performed manually [173].

With expansion microscopy, the idea of FISSEQ was revived in ExSeq [174]. Just like in ExM-MERFISH, transcripts are incorporated into a polyelectrolyte gel, which is expanded, so the amplicons are no longer so tightly packed. This eliminated the depletion of highly abundant genes compared to bulk RNA-seq, and the detection efficiency and proportion of rRNA reads of ExSeq seem on par with randomly primed bulk RNA-seq of adjacent sections. In addition to SOLiD sequencing as in FISSEQ, the amplicons are also sequenced ex situ with Illumina SBS. The in situ sequences are matched to ex situ sequences and only unique matches are kept, to more effectively align amplicons to the genome and to localize mRNA sequence variations such as alternative splicing that are more difficult to detect with SOLiD's short read length. There is also a targeted version of ExSeq, in which padlock probes with gene specific barcodes are RCA amplified and the barcodes are sequenced in situ by either SOLiD or Illumina SBS, profiling up to 297 genes; the detection efficiency is 62% compared to smFISH (for 4 genes in the same 60 cells to which both targeted ExSeq and HCR-smFISH were performed, the number of transcripts detected by ExSeq is about 62% compared to HCR-smFISH), which is high compared to ~5% for 2013 Nilsson ISS but lower than that of MERFISH (HD4) and HCR-seqFISH [174, 176]. Eight probes were designed for each gene, and the transcripts must be at least 960 nt long, shorter than required by MERFISH (without bDNA) and seqFISH variants. To our best knowledge, ExSeq has yet been used to collect new datasets after its inception. Just like ExM-MERFISH, ExSeq has disadvantages from expansion microscopy, such as increased imaging time as there is less tissue per unit area and that the expansion is non-isotropic and continues through the rounds of imaging.

In INSTA-seq [177], recessed sequencing primers and multiple rounds hybridization and sequencing like in SOLiD were used. However, unlike in SOLiD, each fluorescent probe only queries one base, and the ligation extends the sequencing primer on both the 5' and 3' ends. To select for poly-adenylated mRNAs, oligo-dT

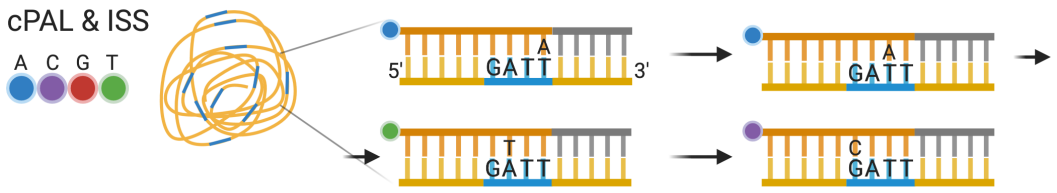

Figure 7.31: Schematic of cPAL as used in ISS.

is used as the primer for reverse transcription (RT), and the cDNA is circularized and RCA amplified. Oligo-dT is then used as the sequencing primer, to sequence RT start and stop site of the particular cDNA *in situ*, giving rise to a barcode of the each amplicon. Then NGS can be used to determine the full sequence of each amplicon and matched to the *in situ* sequenced barcodes and thus spatial locations. As RT is terminated where the transcript is crosslinked to RNA binding proteins (RBP), INSTA-seq can profile RBP motifs near the 3' UTR in space.

cPAL and ISS

An alternative SBL scheme is combinatorial probe anchor ligation (cPAL), which to our best knowledge, was first demonstrated in 2005 [178]. In cPAL, an anchor primer is hybridized to a constant region immediately adjacent to the target. T4 DNA ligase requires matching base pairing up to 6 bases from the ligation junction when ligating from 5' to 3' and 7 bases when ligating from 3' to 5'. The first base of the target 5' to the constant region is interrogated by a 9-mer probe whose 5' most base is represented by the color of a fluorophore and ligated to the primer if a perfect match is present (Figure 7.31). Then the ligated construct is stripped and a new primer is hybridized to the constant region. The second base is interrogated by a 9-mer probe whose second 5' most base is represented by the fluorophore. This can carry on until the 6th base on the 5' direction. When the constant region is 5' to the target, bases 3' to the constant region are interrogated in a similar fashion. With constant regions flanking a target so primers bind in both direction, a 13 nt target can be sequenced this way, and the read length can be somewhat increased by adding degenerate bases to the anchor primer extending into the target [179].

The only *in situ* sequencing method that was reused after its inception was originally demonstrated in 2013 by Mats Nilsson's group [121], which we call ISS here (Figure 6.7). First, padlock probes are hybridized to *in situ* reverse transcribed cDNAs and RCA amplified (Figure 7.18). The padlock probe can carry a gene specific 4 nt barcode (barcode version), or leave a 4 nt gap between the ends of the probe after

it's hybridized to the cDNA to be filled when the probe is circularized (gap filling version). Then the barcode or the filled gap is sequenced in situ, with an anchor primer binding 3' to the target, with cPAL. Because of limited read length of cPAL, short sequences uniquely identify each gene and isoform for the gap filling approach becomes difficult to find with increasing number of genes and isoforms. In contrast, a barcode with length n can encode 4^n genes and isoforms. As a result, the barcode approach was repeatedly used after the inception of ISS and was commercialized by CARTANA, which was recently acquired by 10X Genomics.

The barcode approach was initially used to profile 39 genes [121], but has been used to profile up to 222 genes in human brains affected by Alzheimer's disease [180]. Although, as already mentioned, ISS has much lower detection efficiency than smFISH-based methods, because of RCA and this low detection efficiency, the density of imaged amplicons is lower, allowing for imaging at lower resolution (20x; MERFISH uses 60x) and thus facilitating profiling large areas of tissues such as whole mouse brain coronal sections [181, 182]. ISS has also been used in conjunction with spatial transcriptomics techniques that are transcriptome wide but lack single-cell resolution, such as ST. Panels of usually fewer than 100 genes of interest are selected from ST and scRNA-seq data, to be profiled with ISS for more in depth characterization of these genes [180, 183]. In addition, because of the specificity conferred by the padlock probe and the small number of probes required per gene (usually 5 per gene but can be fewer), ISS has been used to quantify isoforms from isoform specific exons and exon-exon junctions [184].

cPAL sequencing has also been used in BOLORAMIS [125] to profile transcripts of 96 genes and 77 miRNAs. Efficiency of padlock probe ligation when the template is RNA is improved with the SplintR ligase and careful placement of the ligation junction in the target region; the inefficiency of reverse transcription in ISS is avoided. With target sequence of 25 nt, shorter than that of STARmap (next section), BOLORAMIS has been adapted to target miRNAs that are 18-23 nt long, but barcode error rate for miRNAs is higher than that for mRNAs. While cPAL was used for demonstration, in principle, hybridization and SBS may be used to detect the barcodes. In terms of average number of spots per cell for 3 genes in BOLORAMIS vs. smFISH, efficiency of BOLORAMIS is 11% for GAPDH, 35% for POLR2A, and 12% for TFRC.

The number of genes that can be profiled by ISS is limited by the barcode length. Just like in seqFISH, only a small subset of all possible barcodes given a barcode

length is used for error correction. As a result, to profile the entire transcriptome of over 20,000 genes, the barcode should be at least 8 nt long (65,536 barcodes), while in one direction, cPAL can only sequence 6 or 7 nt and degenerate bases. It is possible in theory to lengthen the barcode to up to 13 nt by sequencing from both ends of the barcode as in the original 2005 method [178]. However, with increasing number of transcripts comes the problem of optical crowding, which is exacerbated by the physical size of the RCA amplicon. Perhaps ExM can be used here to mitigate optical crowding just like in ExSeq. To address the limitation in barcode length, HybISS, i.e. hybridization-based in situ sequencing, was devised [122] so the now seqFISH-like barcode can be arbitrarily lengthened by increasing the number of rounds of hybridization. HybISS has already been reviewed in Section7.2; despite the "ISS" in its name, HybISS is classified as smFISH-based because it does not involve SBL or SBS. HybISS also has up to 5 fold higher signal to noise ratio than ISS, and has somewhat higher detection efficiency than ISS though the improvement is less than 2 fold (average number of amplicons detected per cell for each channel in HybISS compared to ISS) [122]. Comparison between HybISS and smFISH has not been reported. Nevertheless, HybISS has not yet been scaled to more than 120 genes and ExM may still be needed for transcriptome wide profiling.

SEDAL and STARmap

Both SOLiD and cPAL have some drawbacks. As the gene specific barcode does not have to be long to encode all genes in the genome, when the barcode is used, limits in read length is not a major limitation. Because one color encodes two bases, SOLiD is very accurate [185], but error in one cycle propagates to later cycles. At least in the mouse brain, SOLiD also has high background [135]. In contrast, cPAL does not have an inbuilt error rejection mechanism; the barcode must be elongated to allow for error correction, much like in the error correction scheme of seqFISH. Furthermore, in ISS, the mRNA is first reverse transcribed into cDNA because ligation of the padlock probe is inefficient when the template is RNA [117]. However, the efficiency of RT depends on the gene of interest and the variability of RT efficiency depends on RNA concentration [186, 187].

A new method of in situ sequencing, namely sequencing with error-reduction by dynamic annealing and ligation (SEDAL) in spatially-resolved transcript amplicon readout mapping (STARmap), was devised to address these shortcomings [188]. In STARmap, the specific amplification of nucleic acids via intramolecular ligation (SNAIL) probe is a derivative of the original padlock probe that avoids RT altogether.

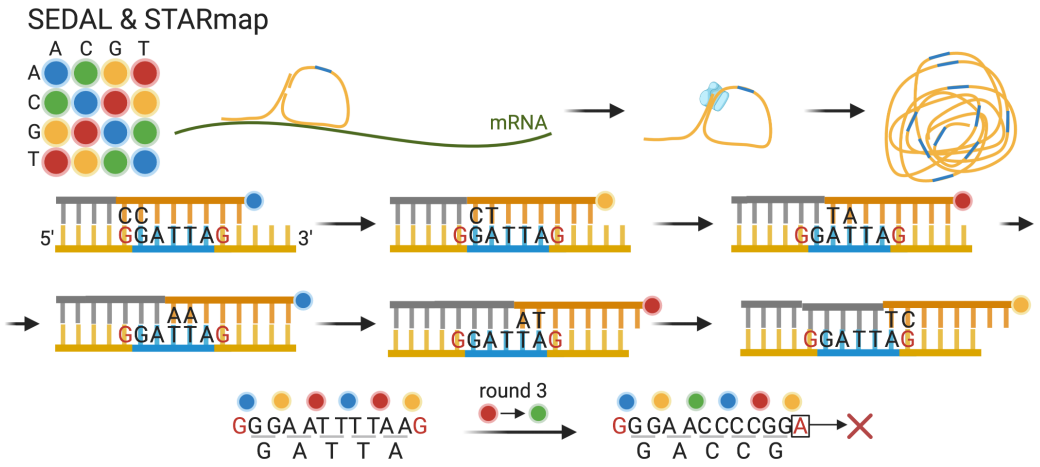

Figure 7.32: Schematic of RCA of SNAIL probe and SEDAL. Also showing error propagation and identification of 2 base encoding. As part of the constant region, the 'G' highlighted in red is known.

A primer partially hybridizes to the mRNA, and partially to the padlock probe (Figure 7.32). The padlock probe carrying a 5 nt gene specific barcode hybridizes to the mRNA adjacent to the primer, but both ends of the padlock probe hybridize to the primer instead, so when the ends are ligated together, the template is DNA rather than RNA, thus avoiding both RT and inefficiency of ligation with RNA template, and then the primer is used to initiate RCA. As both the primer and the padlock probe must match the mRNA template for RCA to occur, SNAIL probes are specific and background of non-specific binding is eliminated. To reduce background autofluorescence and prevent the RCA amplicons from moving, the amplicons are crosslinked into a hydrogel and the tissue is cleared of proteins and lipids.

Then SEDAL is used to sequence the gene specific barcodes. The sequences flanking the gene barcode are known. In the first round an anchor or reading probe binds to the constant region 5' to the barcode, one base away from the barcode (Figure 7.32). The decoding probes are 8-mers labeled with a fluorophore at the 5' end whose color represents the 2 nucleotides at the 3' end that interrogates the barcode; the other bases are degenerate. If the decoding probe matches the barcode, then it is ligated to the reading probe and the fluorescent signal is recorded. In the first round, the decoding probe interrogates the last base of the constant region and the first base of the barcode, as the last base of the constant region is necessary to decode the sequence of colors. Then the reading and decoding probes are stripped. In the

second round, the reading probe stops right where the barcode starts. In the third round, the reading probe has a degenerate base extending into the barcode. Reading probes of the following rounds extend further into the barcode with degenerate bases. In the last round, the decoding probe interrogates the last base of the barcode and the first base of the following constant region. Like in SOLiD, with 2 base encoding, an error in a previous round propagates into later rounds; with propagation, when there is an error when decoding, then the first base of the constant region after the barcode would be incorrectly decoded, so the error is identified and rejected. Comparison of detection efficiency of STARmap with that of smFISH has not been reported; the efficiency is reported (average number of transcripts per cell for 151 cell type marker genes) to be somewhat better, at least not worse, than that of scRNA-seq, suggesting that STARmap is perhaps more efficient than ISS, but most likely much less efficient than MERFISH (HD4) and seqFISH.

Sequencing by synthesis

While most *in situ* sequencing techniques use SBL, some use SBS, indeed with a heated stage to perform SBS in situ. Because Illumina SBS is much more well-known and widely used than SBL for NGS, we will not recap it here. SBS has been tried to sequence DNA barcodes of antibodies in highly multiplexed immunofluorescence [189]. BARseq [190], a method to trace neuron projections is also based on SBS. In BARseq, the gap filling version of ISS (Section 7.3) is used and the filled gap that is the projection tracing barcode is sequenced with Illumina SBS chemistry. BARseq has also been adapted to profile endogenous transcripts (up to 79 genes as of writing) and image projection barcodes in the same neurons (BARseq2) [191, 130]; gene expression and projection can be correlated in some though not all cells. For endogenous transcripts, the mRNA is first reverse transcribed, and the barcode version of ISS (Section 7.3) is used to amplify the barcodes (in the padlock probe but not the cDNA) with RCA, which are then sequenced *in situ* with SBS. For transcripts, BARseq2 detects slightly more copies of mRNAs than 10X v3 scRNA-seq for the same gene in the same tissue.

7.4 NGS with spatial barcoding

This section reviews techniques that capture transcripts from a permeabilized tissue section on a spatially organized array for RNA-seq. These techniques are similar to 3' based scRNA-seq, with amplification and sequencing handle, barcode, UMI, and poly-T to capture polyadenylated transcripts, except that each spot in the array

has its own barcode, rather than each droplet. As the spots are not organized in a regular array in Slide-seq, PIXEL-seq, and Seq-scope, this section more generally concerns techniques that pre-determine spatial locations of each spatial barcode before capturing polyadenylated transcripts, so the spatial location is encoded by the barcode rather than selection and isolation of ROIs. These techniques are generally transcriptome wide, but do not have single-cell resolution; the resolution is the size and shape of the spots and spacing between the spots. In ST and Visium, the array is constructed by printing the capture sequences onto commercial microarray slides, so the 5' end of the sequences are attached to the slide; where each spatial barcode is placed is known. In DBiT-seq, the array is constructed by depositing barcodes specific to each microfluidic channel with orthogonal channels. Alternatively, the capture sequences can be attached to beads like in droplet scRNA-seq, as in Slide-seq and HDST. The beads are randomly placed on a slide in a single layer, and the location of barcodes are determined before library preparation when the capture sequences and transcripts are released from the slide.

ST [141] and Visium are the most popular NGS barcoding based techniques worldwide (Figure 6.7, Figure 7.33). In ST, the printed spots have diameter of 100 μm and are 200 μm apart from center to center (Figure 7.35). Multiple sections can be mounted to the same slide, separated by a rubber mask. For each section, there are 1007 spots covering an area of 6200×6600 μm. The 5' end of the capture sequence is a linker to be cleaved to release the transcripts, followed by amplification and sequencing handle, an 18 nt spatial barcode, a 9 nt UMI, and poly-T (Figure 7.36). For the genes quantified with smFISH, ST's detection efficiency is around 6.9% compared to smFISH (transcript count per area for 3 genes in corresponding regions in adjacent sections), within the range of the efficiency of scRNA-seq techniques. Despite the low resolution, ST is popular probably due to transcriptome wide profiling, ease to apply to larger area of tissue, not requiring specialized equipment such as SRM and custom fluidics systems, commercial kits, possible automation of library preparation [192], availability of a documented and open source data preprocessing pipeline called ST Pipeline [193], and the extra information from H&E staining before library preparation.

After its inception, ST has been used in a wide range of clinical pathological tissues, such as heart after heart failure [194], peritonitis-affected gingival tissue [195], prostate cancer [196], breast cancer [197], arthritic joint biopsies [198], lymph nodes affected by melanoma metastasis [199], spinal cords [200] and cerebellums

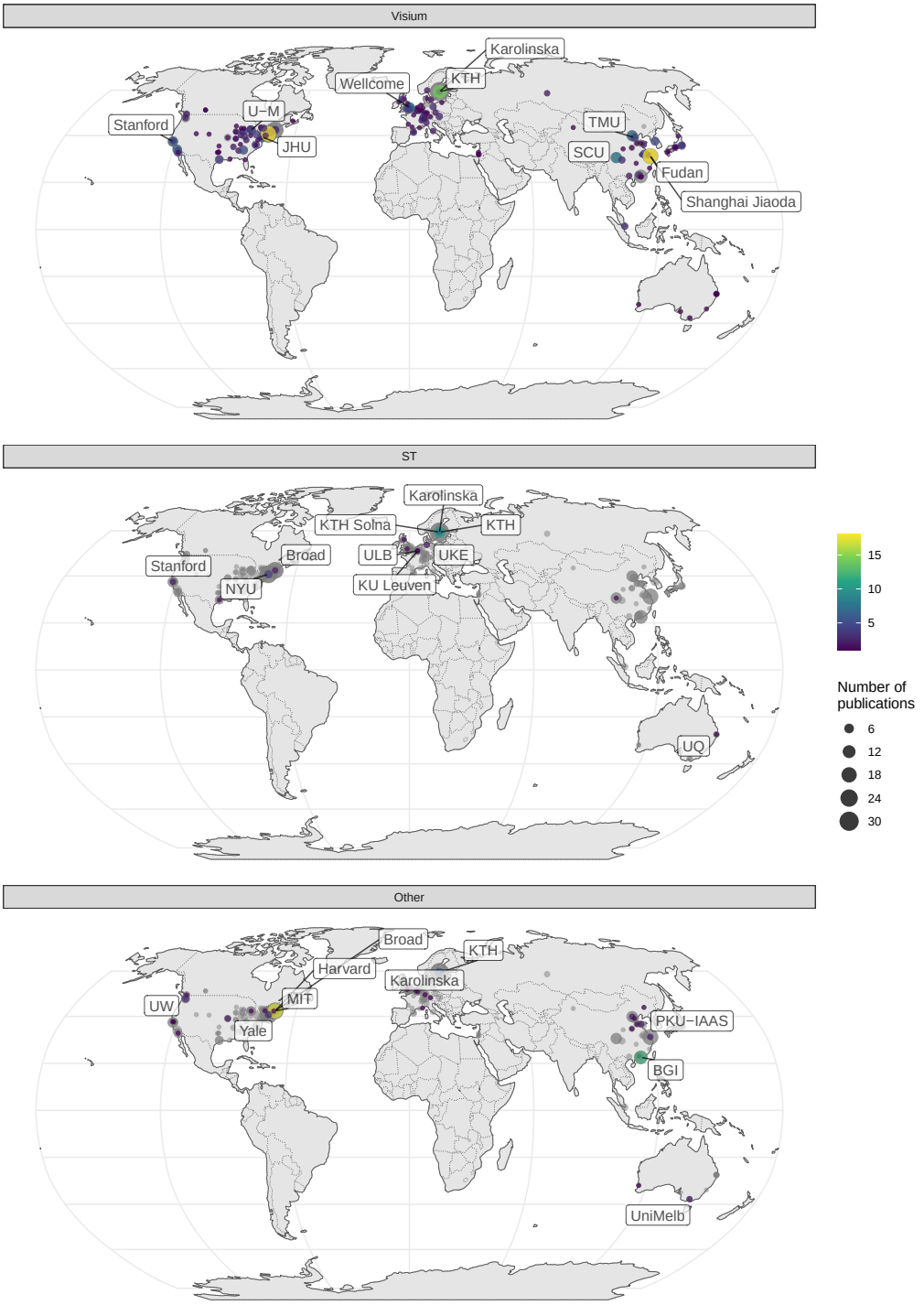

Figure 7.33: Cities and institutions using the two most popular technologies worldwide, the two methods used by the most institutions. Preprints are included.

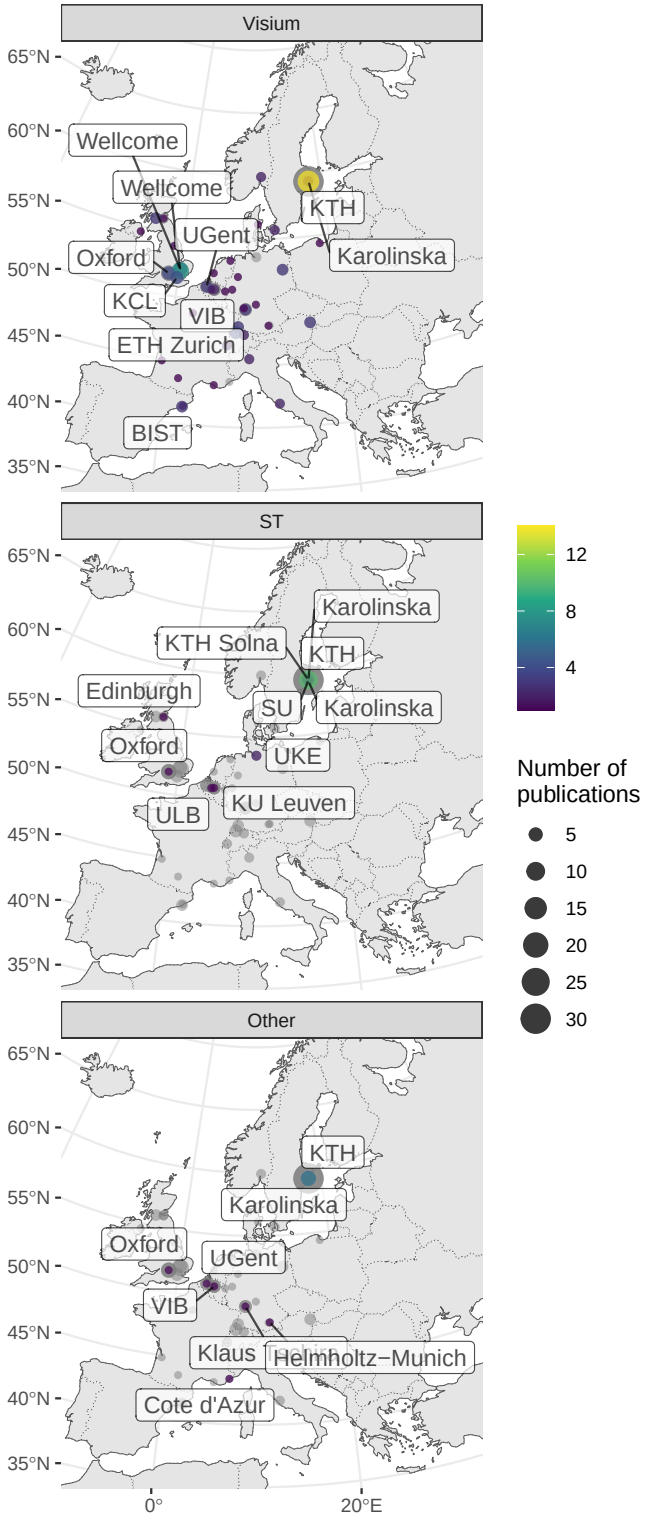

Figure 7.34: Cities and institutions using the two most popular tehcnologies in western Europe. Preprints are included.

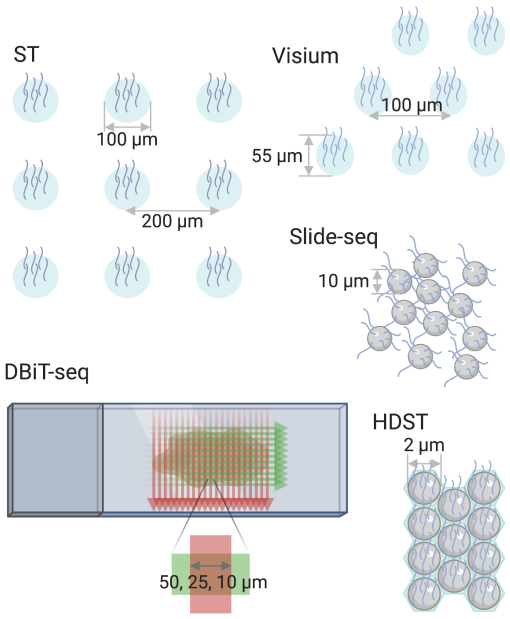

Figure 7.35: Schematic of spot construction and size of array based techniques.

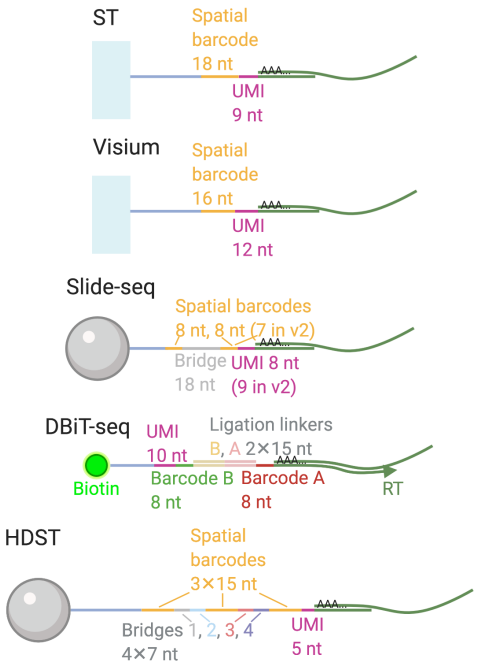

Figure 7.36: Barcode and UMI structure and lengths of array based techniques.

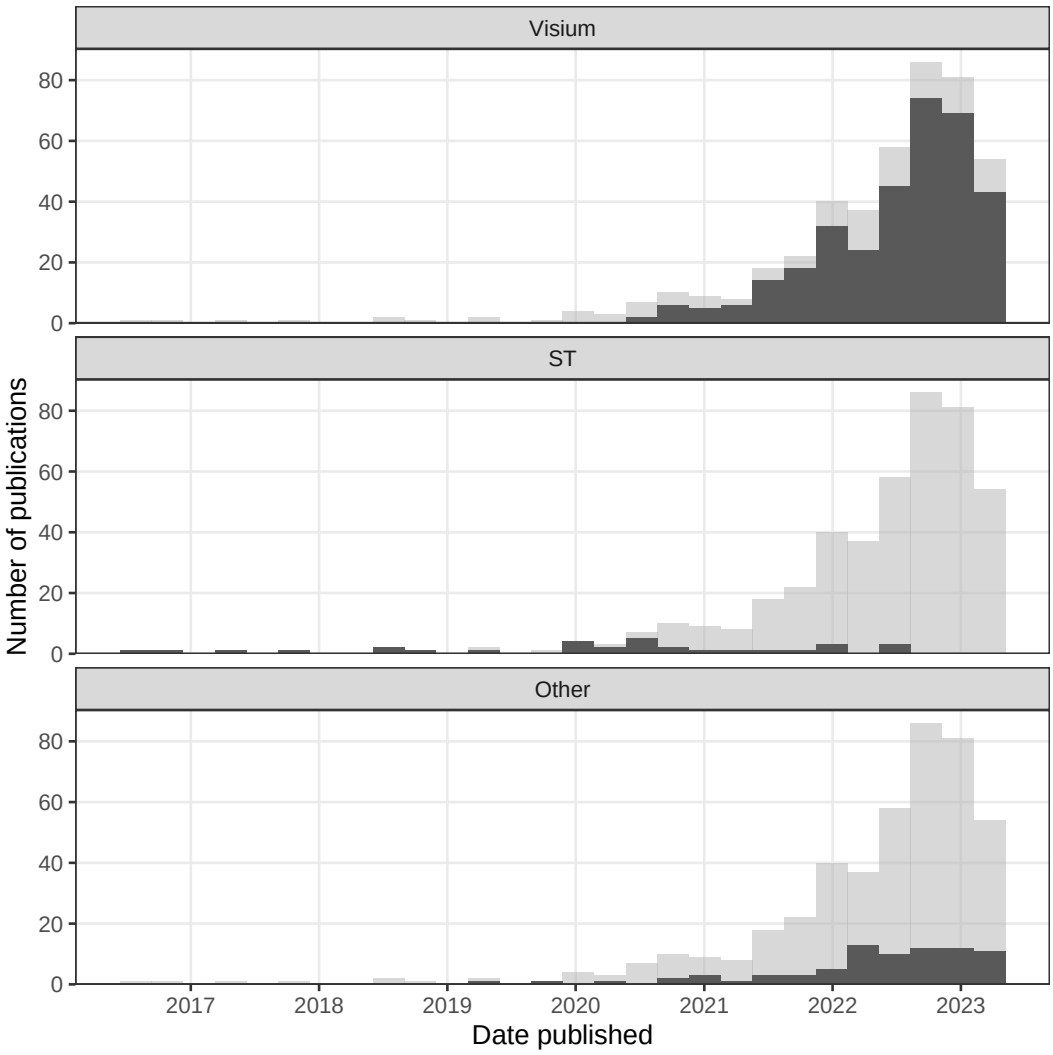

Figure 7.37: Number of publications over time, broken down by technique. The facets are ordered by total usage of the technique. Bin width is 90 days.

[201] affected by amyotrophic lateral sclerosis (ALS), and squamous cell carcinoma [202]. ST has also been used to construct gene expression atlases of healthy tissues such as the developing human heart [183] and the mouse brain [203]. Common downstream data analyses include identifying differentially expressed (DE) genes between diseased and healthy regions, gene set enrichment analysis (GSEA) among DE genes, and cell type deconvolution of the spots by integrating ST and scRNA-seq data. Data analysis methods designed specifically for ST or Visium will be reviewed in more detail in Chapter 9.

After 10X Genomics acquired ST in December 2018, the 10X Visium has quickly gained popularity and spread to multiple institutions (Figure 6.7). Visium has

superseded ST and has become the most popular current era technology (Figure 7.33, Figure 7.37). While ST is still the second most popular NGS barcoding technology in Europe (Figure ref(fig:array-europe)), usage of ST seems concentrated in Sweden, where ST comes from. In contrast, usage of Visium is more decentralized (Figure 7.33). Visium is similar to ST and shares the advantages of ST, but with higher spatial resolution. The spots are 100 μm apart center to center, each with a diameter of 55 μm, arranged in a hexagonal configuration (Figure 7.35). After adjusting for spot area, Visium seems to capture somewhat more transcripts and genes compared to ST [204], but more datasets in the same tissues and accounting for sequencing depths are needed to make a fairer comparison. In addition, Visium's growth in popularity may be due to core facilities at multiple institutions providing Visium services [205, 206]. As a new version of ST, Visium was originally designed for fresh frozen OCT embedded tissue and 3' Illumina sequencing. However, Visium has more recently been adapted to FFPE tissue [48] (the now commercial Visium FFPE has a very different chemistry from [48]), as well as to Nanopore long read sequencing to quantify isoforms [184, 207], although Visium is still predominantly used on fresh frozen tissues for 3' end sequencing (Add figure about FFPE for all current era methods).

Visium studies in our database are almost exclusively on humans and mice, and mostly on humans (Figure 7.38). For a long time, in both humans and mice, and both the healthy and pathological cases, the brain is again the most studied organ (Figures 7.39, 7.40), but more studies have been performed on other organs in humans more recently. This is in stark contrast with usage of GeoMX DSP, which was used in several COVID lung studies but not much in brain (Figure 7.9).

In response to the low resolution of ST, Slide-seq was developed to increase the resolution of array based spatial transcriptomics [209]). Beads like those used in Drop-seq [142] with diameter 10 μm are spread on a slide in a single layer, not necessarily in a regular grid, and bead barcodes are generated with 16 rounds of split pool, each round adding one nucleotide, broken into 2 blocks of 8 nt (2 blocks of 8 and 7 nt in version 2) (Figure 7.35, Figure 7.36). As the location of each barcode is not pre-determined, the slide is imaged and the barcodes are sequenced in situ with SOLiD. Then the OCT frozen tissue section is mounted on the layer of beads on the slide and the beads are removed for library preparation. The first version of Slide-seq is very inefficient; for the genes compared, the Slide-seq only detects 2 to 3 orders of magnitude fewer transcripts per cell for 3 genes than smFISH in an

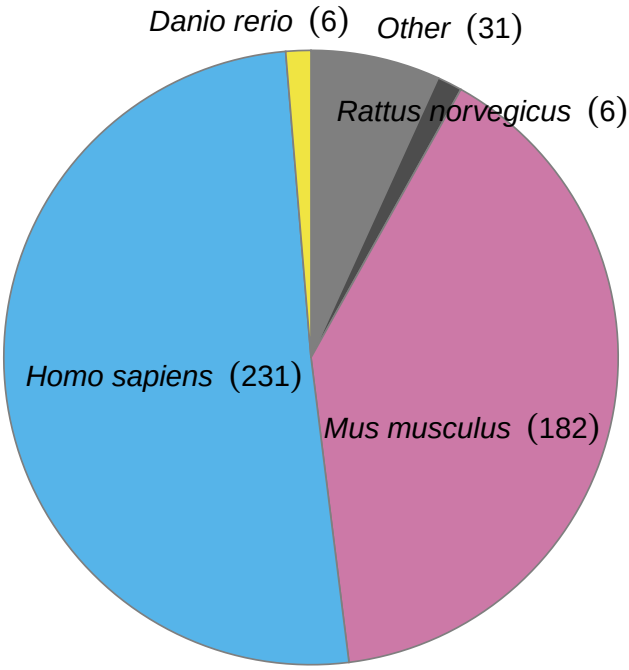

Figure 7.38: Number of publication per species.

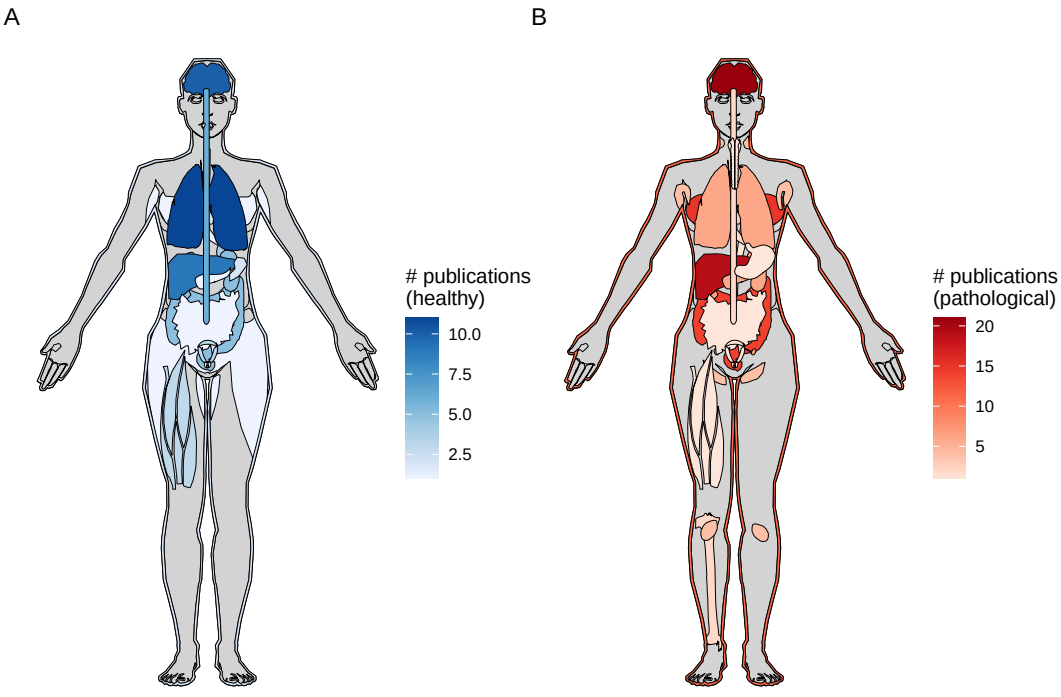

Figure 7.39: Number of Visium studies for healthy (A) and pathological (B) human organs. Female is shown here due to several breast cancer and ovary studies. There is one human prostate Visium study in our database [208].

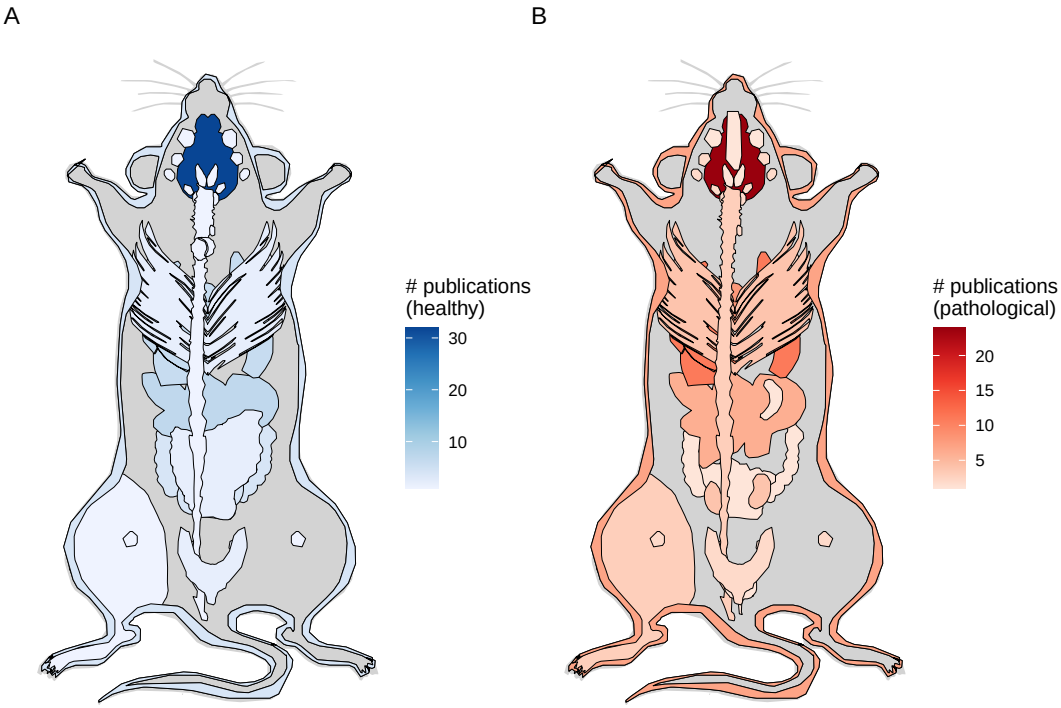

Figure 7.40: Number of Visium studies for healthy (A) and pathological (B) mouse organs. Female is shown here as there is a uterus study while there is no study on male specific organs in our database.

adjacent section and about 2.7% compared to Drop-seq for the same cell type from CA1 [209].

In the second version of Slide-seq (Slide-seq2) [210], the barcodes are sequenced by SEDAL (like in Figure 7.32, but with one color per base) rather than SOLiD, which increased the efficiency of spatial mapping of Illumina reads, probably because of error propagation in the 2 base encoding of SOLiD. Moreover, bead synthesis is further optimized and a second strand synthesis step is added to the library preparation to increase the number of cDNAs for PCR amplification. Efficiency is improved in Slide-seq2, which is ~9.3x higher than version 1, about on par with Drop-seq, 1 order of magnitude lower than that of smFISH, and somewhat better than Visium in the dataset chosen. Here "efficiency" means number of UMIs or transcripts for 3 genes from a fixed area in CA1. The official software to process the in situ sequencing images is written in MATLAB, which is proprietary. Although the size of the bead is close to the size of a single-cell, Slide-seq does not have single-cell resolution as one bead can capture transcripts from more than one cells nearby, so cell type deconvolution of beads is still needed. After its inception, Slide-seq2 has been used on mouse and human testes, at the institution of origin

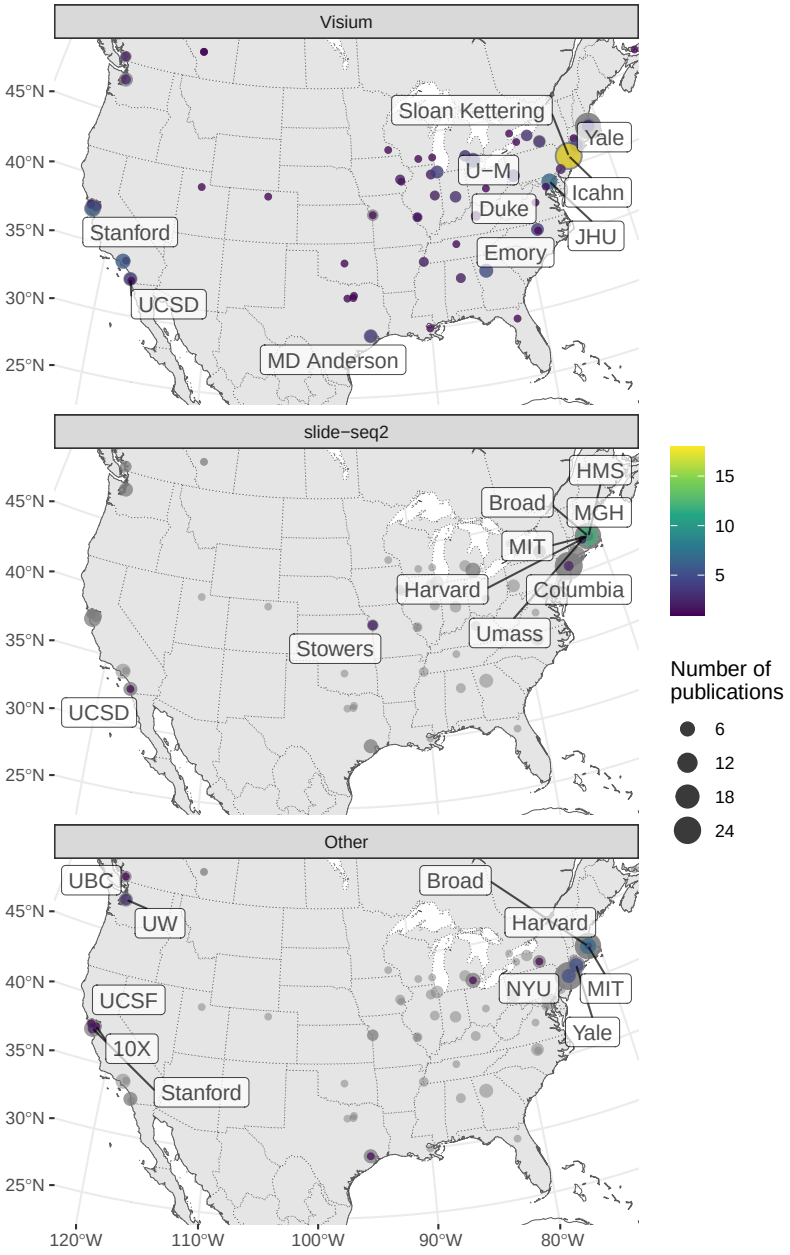

Figure 7.41: Cities and institutions using the two most popular technologies around continental US. Preprints are included.

[180].

Slide-seq is the second most popular NGS barcoding technique around the US, but unlike Visium (the most popular), Slide-seq is still more concentrated around Broad, Harvard, and MIT, where it was first developed (Figure 7.41).

Spatial resolution of array based techniques has been further increased with HDST,

with a resolution of 2 μm [211], which is smaller than a single-cell. Like in Slide-seq, beads like those used in droplet scRNA-seq are used. The diameter of each bead is 2 μm, and hexagonal wells with diameter 2.05 μm are carved into a slides so each well contains one bead (Figure 7.35). The spatial barcodes are generated by 3 rounds of split-pool, each round adding 15 nt from the barcode pool (Figure 7.36). The UMI is only 5 nt but such a small area does not contain that many transcripts. As the beads are randomly placed in the wells, the locations of barcodes need to be determined. Four rounds of FISH, with combinations of red, green, and no color, encode each of the 3 barcodes on each bead. Again, HDST was originally designed for fresh frozen OCT embedded tissue rather than FFPE. HDST is very inefficient; for the genes compared, the detection efficiency is only ~1.3% compared to smFISH per bead area. To our best knowledge, HDST has not been used for new datasets after its inception.

In response to the low efficiency and complicated procedure to localize barcodes of Slide-seq and HDST, Deterministic Barcoding in Tissue for spatial omics sequencing (DBiT-seq) was developed, with resolution up to 10 μm [204]. Let i, j denote the index of channel in each direction. Barcode A_i, attached to poly-T, is flown across the slide in microfluidic channels and RT is performed (Figure 7.35). Then barcode B_j, attached to the UMI, PCR handle, and biotin, is flown across the slide in microfluidic channels perpendicular to those that delivered barcode A_i, and barcode B_j is ligated to barcode A_i and the cDNA (Figure 7.35, Figure 7.36). Then the ligated barcodes and cDNA can be purified by streptavidin-coated magnetic beads. Each microfluidic channel carries a different barcode, so where the channels for barcodes A_i and B_j intersect, an array is created and the location of each spot is encoded by i, j. The resolution is limited by the width of the channels and the spacing between them; widths of 50, 25, and 10 μm have been tested. Per unit spot area, DBiT-seq seems to detect at least 3 times more genes and UMIs than ST and Visium and the improvement is even starker at the 10 μm resolution. For the genes compared, DBiT-seq's detection efficiency is ~15.5% of that of smFISH per unit area, making it relative more sensitive among the array based methods reviewed here. DBiT-seq has also been adapted to FFPE, although just like in Visium, RNAs in FFPE tissues are more degraded than in fresh frozen tissues and fewer genes and UMIs are detected per unit area in comparable tissue types [212].

The record resolution of array based techniques is ever increasing (Figure 7.42); sub-micron techniques are appearing in 2021. The record is broken by Stereo-seq

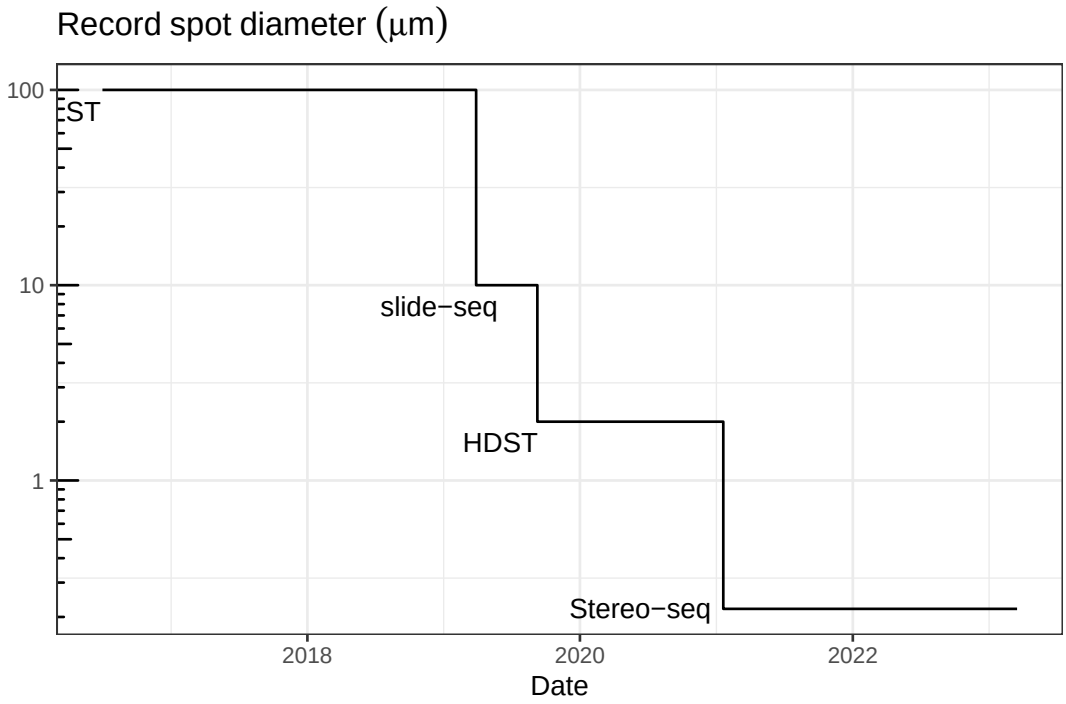

Figure 7.42: Record spot diameter of array based methods over time.

in January 2021, reporting a spot diameter of 220 nm although the distance between spots is 500 or 715 nm [213]. In Stereo-seq, circularized DNA containing a random 25 nt barcode is RCA amplified and deposited into an etched grid. The barcode is sequenced and then oligos with polyT and molecular ID are hybridized to the barcode to capture polyA transcripts from the mounted tissue. The reported capture efficiency is around 170 UMIs per 100 μm^2 in mouse brain, on par with that of the Visium mouse brain dataset from the 10X website reanalyzed in the same study.

In Northeast Asia, Visium is the most popular method and is used in many less well-known institutions across different countries. The second most popular technology is Stereo-seq, which has been commercialized by BGI, although its use is concentrated around BGI Shenzhen where it was developed (Figure 7.43).

Another sub-micron array capture method is Seq-Scope [214], which creates clusters of polyT capture sequences each with its own spatial barcode (20-32 nt) from Illumina bridge amplification on a repurposed Illumina flow cell. The spatial barcode is sequenced with SBS. Then the flow cell is dismantled so the tissue can be mounted for transcript capture. The captured transcripts are then sequenced with NGS. The clusters can have a diameter down to 0.5 μm, and the clusters are randomly seeded, not distributed in a grid. The reported capture efficiency is around

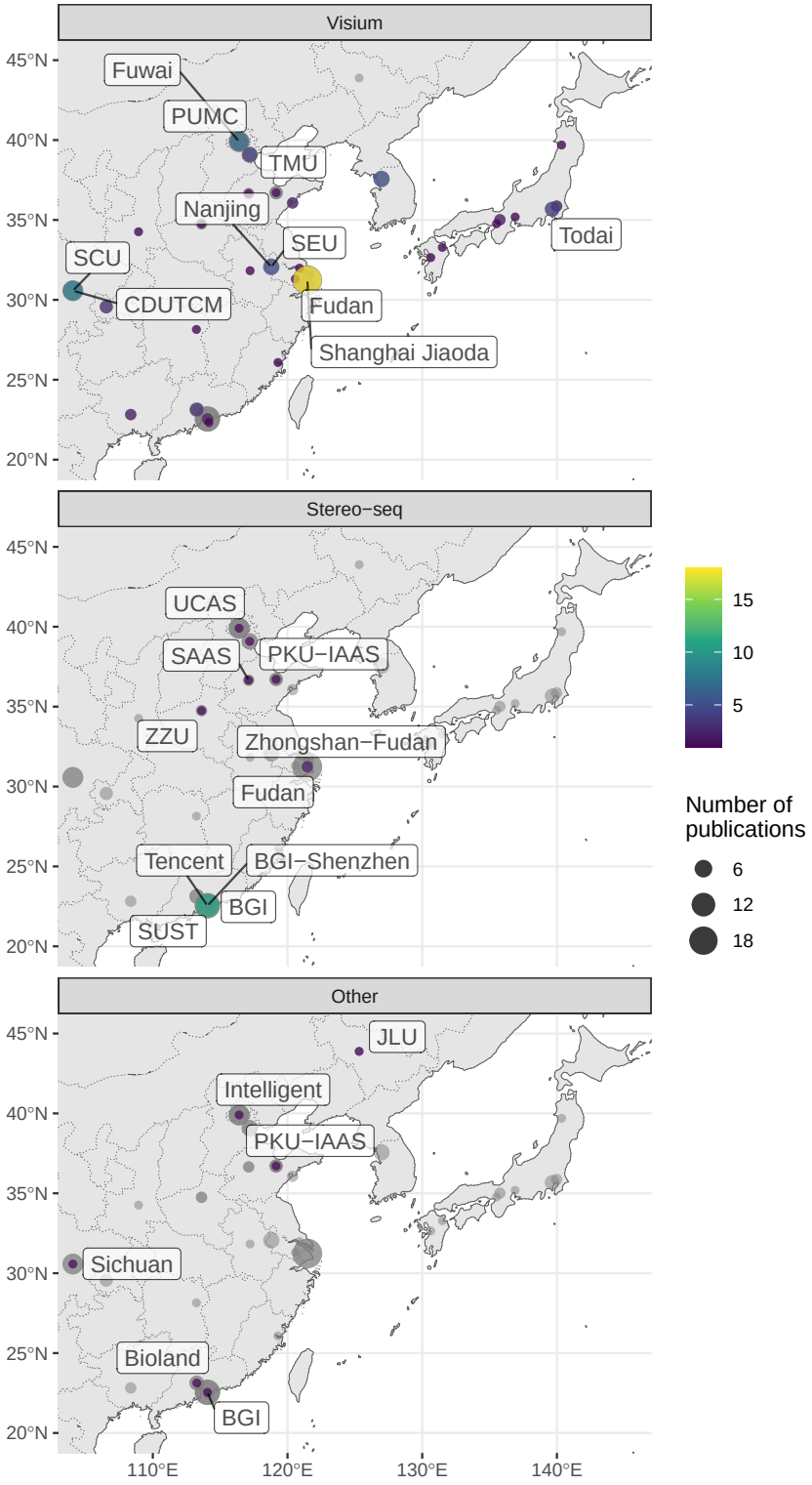

Figure 7.43: Cities and institutions using the two most popular tehcnologies in Northeast Asia. Preprints are included.

1000 and up to 2000 UMIs per 100 μm^2 in mouse colon, much higher than that of Stereo-seq, although we are not sure whether colon data is comparable to brain data here.

A more recent nearly sub-micron technique is PIXEL-seq [215]. Again, as in the Illumina flow cell, PIXEL-seq amplifies each randomly seeded spatial barcode (24 nt) and polyT capture sequence into polonies. However, here a crosslinked polyacrylamide gel (rather than a linear one in Illumina) is used, to form continuous polonies without much space between their "territories" rather than discrete clusters. The spatial barcodes are also first sequenced with SBS before the tissue is mounted for transcript capture. On average, the polony is around 1.17 μm^2 in area, so assuming it is circular, then the diameter is 1.22 μm. The reported capture efficiency is around 1000 UMIs per μm^2 in mouse brain, which might be comparable to that of Seq-Scope.

While such sub-micron techniques have subcellular resolution, in practice, the data is binned into much larger grids for standard scRNA-seq analysis, such as $36\mu m \times 36\mu m$ in Stereo-seq and $10\mu m \times 10\mu m$ or $7\mu m \times 7\mu m$ or $5\mu m \times 5\mu m$ in Seq-Scope. The subcelluar information was not directly used in the analyses, although even with binning, the resolution is still higher than that of ST and Visium.

All these array and NGS based techniques reviewed so far capture polyadenylated transcripts. While miRNAs form a major topic in LCM literature (Figure 8.3) and are profiled in some prequel era ISH atlases, current era techniques mostly preclude miRNA quantification. BOLORAMIS has been demonstrated on 77 miRNAs, but the barcode error rate is higher than in mRNAs. Without a poly-A tail, miRNAs are precluded by NGS based techniques that rely on poly-A capture. To quantify miRNAs in space, an array based technique was developed as an alternative to LCM and designed for FFPE tissues [216]. The tissue is pixelated, and each pixel is 300 μm $\times$ 300 μm. Within each pixel is a smaller 3 $\times$ 3 array, each spot of which has probes for one miRNA; the locations of the spots within each pixel can be easily discerned with a fluorescent microscope. This way, up to 9 miRNAs can be profiled in the same tissue section at the same time, although the 9 miRNAs are from somewhat nearby cells but not the same cells.

Gray areas and single-cell resolution

In Section 7.1, the definition of "microdissection" is relaxed, so that cell sorting can be some kind of "microdissection", and GeoMX DSP is described there because

like most microdissection methods, it's primarily used for targeted ROIs rather than regular grids and spatial location is known from selection of the ROI. Due to the fuzziness of "microdissection", some techniques that assign spatial barcodes to a regular grid but dissociate each spot in the array into single-cells may or may not be considered *de facto* "microdissection". Because of the regular grid as in Visium and DBiT-seq and that spatial location is known by pre-determined spot barcode rather than selection of ROI, these techniques are summarized in this section.

In Visium, all cells within the same spot get the same spot barcode, so the transcriptome of each spot is from mixture of different cells, often different cell types. Cell types can be computationally deconvolved with software such as Stereoscope [217] and CIBERSORT [218] with a reference of transcriptomes of known cell types. To address this problem, some new array capture spatial techniques impart each spot a spatial barcode before the cells or nuclei in the spots are dissociated and assigned another cell specific barcode for scRNA-seq, so the transcriptomes have single-cell resolution, though not single-cell spatial resolution as the location of each cell within the spot is not recorded and the spatial resolution is lower than that of Visium. In XYZeq [219], spot barcodes with UMIs and poly-T capture sequences are deposited in microwells 500 μm center to center arranged in an array. The tissue sections are fixed with dithio-bis(succinimidyl propionate), which preserves RNA integrity for scRNA-seq. The tissue is permeablized and incubated in an microarray hybridization chamber for the spatial barcodes to hybridize to polyadenylated transcripts and for reverse transcription, so the cDNA acquires the spatial barcode. Then the cells are sorted into PCR wells and the cell barcode is added from a PCR primer. A related method is sci-Space [220], where spatial "hashing" oligos are spotted in a an array on a slide covered with dried agarose. The spots are on average about 73.2 μm in radius and about 222 μm apart center to center. The hashing oligos diffuse into the nuclei in the tissue mounted on the slide, and the spatially hashed nuclei are dissociated for sci-RNA-seq.

7.5 Detection efficiencies

In the previous few sections we have mentioned detection efficiency many times. To recap, these are common methods to estimate detection efficiency of spatial transcriptomics techniques:

1. Gold standard smFISH and the single molecule resolution technique of interest are performed in the same cells for a small number of genes and the numbers

of transcripts spots detected in each segmented cell from smFISH and the technique of interest are compared. This was performed for HCR-seqFISH and ExSeq.

2. Gold standard smFISH and the single molecule resolution technique of interest are performed on different cells of the same type or on adjacent sections for a relatively small number of genes. Then average transcript counts of each gene per cell among these cells are compared between smFISH and the technique of interest. This was performed for MERFISH and seqFISH+, and was used to compare efficiencies of HybISS and HybRISS [124].
3. Gold standard smFISH is performed on an adjacent section for a small number of genes. Transcript spot counts from smFISH and UMI counts from the NGS based technique per unit area in the same tissue type are compared. The unit area can have the same shape and size of the transcript capture spot, or can contain multiple spots and averaged over the spots. This was performed for ST, HDST, and Slide-seq(2).
4. UMI or transcript spot counts of select marker genes per cell in the spatial techniques of interest are compared to those in scRNA-seq of the same cell type. This was performed for STARmap and Slide-seq(2). In Slide-seq(2), as the tissue section is imaged, nuclei can be segmented and counted so the number of cells in the ROI compared is known and an equivalent number of cells from scRNA-seq is sampled for comparison.
5. Number of all UMI and genes detected per unit area in one NGS based spatial technique is compared to those of other NGS based spatial techniques. This was performed for DBiT-seq, Visium (FFPE), Stereo-seq, PIXEL-seq, and Seq-Scope. A caveat is that sequencing depth is not always considered.
6. The number of reads per cell is compared between scRNA-seq and FISSEQ. This is only known to be performed for FISSEQ.

In summary, a putative ranking, from high to low, of capture efficiencies of current era techniques, noting which methods above are used to estimate the efficiencies, is:

smFISH (~100%) > MERFISH (2, HD4, ~95%) > HCR-seqFISH (1, ~86%) > ExSeq (1, targeted, 62%) > seqFISH+ (2, ~49%) > (maybe) Seq-Scope (5) ~ PIXEL-seq (5) > (maybe) DBiT-seq (3, 5, ~15%) ~ Visium ~ Stereo-seq (5) > (maybe) HybRISS (2 with HybISS) > HybISS (2 with ISS and HybRISS) ~ (maybe) STARmap (4) ~ (maybe) scRNA-seq ~ Slide-seq2 (3, 4) ~ ST (3, ~6.9%) ~ ISS (~5%) > HDST (3, ~1.3%) > Slide-seq1 (3, 4, ~1%) > FISSEQ (6, 0.005%)

A percentage is not shown where it is not reported. This is putative because this is based on reports in the main text. There are conflicting reports of capture efficiency of Visium and DBiT-seq. Furthermore, comparison of different tissues and different genes from those studies may be problematic. For some of the technologies, the capture efficiency is compared to that of smFISH with only a few genes. Multiple datasets from each technology for as similar a tissue as possible for the same set of genes should be compared to get a better idea about the capture efficiency of each technique. Moreover, other factors such as tissue handling, sequencing depth, and data processing software may influence the results.

While fresh frozen tissues are predominantly used in the current era, DBiT-seq, Visium, and LCM have been adapted to FFPE tissues. RNAscope can be used on FFPE tissue for up to 12 targets. GeoMX DSP has been predominantly used on FFPE tissues. FFPE is a common way to archive tissue specimen, and sometimes the only tissues available is FFPE, sometimes years if not decades old. From techniques that have both fresh frozen and FFPE protocols, FFPE and storage of FFPE samples seem to significantly degrade the transcripts and reduce detection efficiency, but there can still be enough information preserved to identify cell types in the spots and correlation between gene expression measured in FFPE and fresh frozen tissues is usually high. In the pre-commercial FFPE Visium mouse brain dataset from the protocol of [48], at a sequencing depth of ~50,000 reads per tissue covered spot, the spots have on average ~1200 genes and ~2200 UMIs detected. In contrast, in a similar fresh frozen mouse brain section, with sequencing depth of ~115,000 reads per tissue covered spot, the spots have on average ~6000 genes and ~27200 UMIs detected [48]. While the fresh frozen sample has higher sequencing depth, FFPE seems to reduce the number of genes and UMIs detected beyond the impact of sequencing depth. As Visium captures the transcripts on spots printed to a glass slide, the transcripts need to be dissociated from the tissue, and in the case of FFPE, it means de-crosslinking. However, commercial FFPE Visium today has a very difference chemistry from [48] and is said to be much more efficient. In FFPE DBiT-seq, the transcripts don't have to be de-crosslinked as the barcodes are deposited into the tissue. FFPE reduced cDNA length from an average of ~1400 nt in PFA fixed fresh frozen tissue to about ~600 nt on average. In mouse embryos, while DBiT-seq on PFA fixed fresh frozen tissue gives on average 2100 genes and 4688 UMIs per 25 μm spot, FFPE tissue gives only 355 genes and 520 UMIs in the same sized spots on average [212]. However, sequencing depth is not discussed. From these studies, in both Visium and DBiT-seq, FFPE might decrease

detection efficiency in terms of number of UMIs detected per unit area by about 5 (if considering sequencing depth in the Visium study) to 10 folds.

7.6 *De novo* reconstruction of spatial locations

The techniques reviewed above, involve either imaging (e.g. LCM, smFISH, ISS, Slide-seq, and HDST) or prior knowledge of locations (e.g. Tomo-seq, ST, Visium, and DBiT-seq). Some spatial transcriptomics techniques have been developed that require neither imaging nor prior knowledge of locations, and we review these in this section. While techniques that deposit spatial barcodes in an array at known locations such as Visium and DBiT-seq do not require imaging to know the location of gene expression, the spatial barcode locations are known *a priori*. In contrast, techniques reviewed in this section do not involve *a priori* knowledge of locations.

It is possible to reconstruct relative locations of cells or transcripts from colocalization without imaging, albeit imperfectly. These techniques are reviewed in more details in [221]; we will only briefly summarize techniques that do not require DNA bound to a surface so they can be applied in cells and tissues. An early method to do so is Puzzle Imaging, published in 2015 [222]. Here "colocalization" can mean whether two neurons have axons in the same voxel or whether two neurons are synaptically connected. The spatial reconstruction is framed as a dimension reduction problem; each voxel is represented as a vector with n dimensions, where n stands for the number of neurons, and these vectors are to be projected into 2 or 3 dimensions, representing spatial dimensions, for reconstruction. Puzzle Imaging was only demonstrated in synthetic datasets, but not real biological datasets. Such reconstruction was made possible for transcripts with DNA microscopy [223, 224]. Transcripts are reverse transcribed in situ, and the cDNA, with an UMI added, is PCR amplified in situ. The amplified products diffuse and encounter amplified products from other transcripts. The nearby cDNAs are concatenated with overlap extension PCR, with additional random sequences in the overlapping primers to encode each concatenation event, called unique event identifier (UEI). When the concatenated cDNAs are sequenced, the two UMIs and the UEI are recorded. Because amplified products from two nearby transcripts are more likely to be concatenated than those from two transcripts that are far apart, the number of UEIs between two UMIs can be used to reconstruct relative distance between transcripts.

Techniques have also been developed to quantify transcripts from subcellular compartments, such as APEX-RIP [225] and APEX-seq [226]. Although these tech-

niques do not record or reconstruct spatial coordinates, they are included in this review because the publications describing them described them with terms such as “spatial”, “localization”, and “spatial transcriptome”. APEX is an engineered ascorbate peroxidase, which can be targeted to specific cellular compartments by expressing a fusion of APEX and a protein targeted to the compartment of interest. With substrates H_2O_2 and biotin-phenol (BP), APEX catalyzes formation of biotin-phenoxyl radicals that can biotinylate nearby proteins, which can be isolated with streptavidin. In APEX-RIP, mRNAs are cross linked to nearby proteins and thus isolated after isolating biotinylated proteins. In contrast, in APEX-seq, the mRNAs are directly biotinylated. Compared to APEX-RIP, APEX-seq better discerns transcript localization in compartments not bound by membrane. However, both APEX-RIP and APEX-seq were originally designed for bulk rather than single-cell samples and was tested only on cell culture. Also, because a fusion protein is required, they cannot be performed in human tissue sections.

Rare cell types are difficult to characterize with most spatial transcriptomics techniques. ST and Visium lack single-cell resolution and signal from rare cell types may be diluted by signal from common cell types in the same spot. LCM is still typically not used on single-cells and rare cell types may or may not be easily discernible with H&E. smFISH-based techniques and targeted ISS require a pre-defined panel of genes, often selected from scRNA-seq and well-known markers, but such selection is more challenging for rare cell types, which may not be well-studied enough to begin with due to challenges in other transcriptomics techniques. However, spatial pattern of genes expressed in rare cell types can be characterized by deliberately creating doublets or multiplets involving both common and rare cell types, as in paired cell sequencing [227] and ClumpSeq [228]. Earlier, RNA-seq has been performed to cell multiplets to identify physical interactions between cell types in the mouse bone marrow in ProximID, but reconstruction of spatial locations was not attempted [229]. Spatial patterns of genes expressed in common cell types such as hepatocytes and small intestine enterocytes are already known from smFISH or LCM and spatial reconstruction of scRNA-seq data [230, 39]. Genes expressed in the rare cell types are identified from genes much more highly expressed in the multiplet than in individual cells from common cell types in scRNA-seq, or markers of rare cell types from scRNA-seq if such data exists. Then the multiplets are mapped to spatial locations with patterned genes expressed by common cell types and existing smFISH or LCM data as reference. Then rare cell types and their characteristic gene programs are mapped to spatial locations as well and their patterns can be

characterized without directly imaging these cells.

7.7 Overall comparisons

In the previous sections, we have discussed pros and cons of types of technologies, but have not discussed relative pros and cons when comparing across types. With so many technologies being developed, which one should an interested user choose? Disclaimer: As we have never performed the protocols of any current era spatial transcriptomics technology in the wet lab, we don't know whether some steps in some protocols are more prone to failure or require more hands on experiences to perform well. Below are comparisons across categories or subcategories when relevant:

ROI based

This includes when microdissection techniques are applied to targeted and histologically informed ROIs and GeoMX DSP/WTA:

Compared to techniques that neither target specific ROIs nor have single-cell spatial resolution such as Tomo-seq, Visium, and Slide-seq(2):

Pros:

1. Cell type deconvolution of voxels at the border of different histological regions is unnecessary, as the ROIs are selected based on histological regions.
2. LCM and GeoMX DSP/WTA are most commonly used for ROI based studies. Both have commercial platforms. LCM followed by RNA-seq may be performed by core facilities and GeoMX can be performed by Nanostring TAP. Visium also has this advantage.
3. Both LCM + RNA-seq and GeoMX are compatible with and widely used on FFPE tissues, which may be the only specimen available in some cases. In contrast, Visium is predominantly used on frozen sections.
4. Though the ROIs are often larger than Visium spots, in principle the ROIs can be chosen to be smaller if the transcriptomics method is sensitive enough. However, the resolution might not exceed that of the new sub-micron methods such as PIXEL-seq and Seq-Scope.

Cons:

1. In ROI based studies, typically only a small number of ROIs are used. It can be labor intensive to study the histological staining to manually select too many ROIs.
2. Unlike in techniques that rely on spatial barcodes that can be pooled and later demultiplexed, when spatial locations are only known from selection of ROIs, scaling to larger number of ROIs becomes more challenging as samples need to be collected ROI to ROI. This also applies to Tomo-seq.

Compared to techniques with single-cell and single molecule resolution, including those based on smFISH and ISS:

Pros:

1. When RNA-seq or cDNA microarray is performed after ROI selection, then it's transcriptome wide. Even when it's not transcriptome wide, such as with some gene panels in GeoMX DSP, well over 1000 genes can be profiled at a time, while smFISH and ISS based methods are typically used on fewer than 300 genes. Furthermore, for LCM, with RNA-seq, new transcripts and isoforms can be discovered as there are no probes confined to known transcripts.
2. Again, existing well-established commercial platforms and core facilities for LCM and GeoMX DSP, though this is changing with MERFISH, Rebus Esper, Molecular Cartography, Xenium, CosMX, and commercial probe panels.
3. Because at present, the rawest data the user sees from smFISH and ISS is the images, while the rawest data the user sees from NGS (for LCM and higher-plexed GeoMX DSP/WTA) is fastq files, processing the raw data to get a gene count matrix is more difficult and time consuming for smFISH and ISS.
4. Compatible with FFPE tissues, but this is also changing as CosMX is FFPE compatible.

Cons:

1. Usually not single-cell resolution. Locations of individual cells within an ROI is lost.
2. Subcellular localization of transcripts is lost.
3. Low z resolution for 3D profiling. EASI-FISH has been applied to cleared 300 μm (pre-expansion) sections, and STARmap has been applied to 150

μm sections. For ROI based methods, the thick section would need to be first sectioned into thinner sections for ROI selection. Also, within the thin section, the ROI is effectively 2D while boundaries between cells vary through the z-plane. In frozen sections, the section thickness is usually at least 10 μm, and in FFPE, the section thickness is usually at least 4 μm, so the z resolution is lower than that of confocal and light sheet microscopy. However, 3D thick sections are also challenging for smFISH and ISS. Only relatively small numbers of genes have been profiled in the thick sections (26 for EASI-FISH and 28 for STARmap).

4. Lower detection efficiency than some of the smFISH-based methods, though ISS based methods also tend to be inefficient.

NGS barcoding

In the previous subsection, cons of ROI based methods compared to NGS barcoding would be the pros of NGS barcoding compared to ROI based methods. Compared to smFISH and ISS:

Pros:

1. Transcriptome wide, and can discover new transcripts and isoforms. Visium has been adapted for full length sequencing.
2. When MERFISH and seqFISH were scaled to around 10,000 genes, gene expression was only profiled in relatively small numbers of cells (1000 something or fewer per dataset). In contrast, a coronal section of one hemisphere of a mouse brain can fit into one tissue capture area of a Visium slide. One Visium tissue capture area has 4992 spots, and if each spot contains 3 cells, then when all spots are covered by tissue, transcriptomes from nearly 15,000 cells are captured. Even when the area is not fully covered, many more cells are captured than in the 10,000 gene MERFISH and seqFISH datasets.
3. Visium is commercially available and is performed by some core facilities. Library preparation of Visium can also be automated. While (Hyb)ISS and Xenium are also commercialized by 10X, Visium is much more popular, perhaps for its other advantages.
4. Thanks to the popularity of Visium and ST and their similarity with scRNA-seq, there's more software designed for Space Ranger output and plotting gene expression at the spots with the H&E staining in the background, such as Seurat, STUtility, and SpatialExperiment.

5. The rawest form of data the users see is usually the fastq files, which are easier to process to get a gene count matrix than the smFISH and ISS images.

Cons:

1. New techniques not yet commercialized such as PIXEL-seq and XYZeq might also be just hard to independently implement.
2. For the most part, there is no single-cell spatial resolution. Even for the sub-micron techniques, the spatial resolution is lower than the diffraction limit of visible light.
3. As the tissue section is mounted onto a slide with spatial barcodes in 2D, z resolution is limited to section thickness. Even with the sub-micron techniques, z positions of transcripts are lost.
4. Lower detection efficiency compared to some smFISH-based techniques. Sequencing cost would increase with greater sequencing depth for greater sensitivity, and to cover larger areas of tissue.

smFISH

Compared to techniques without single-cell resolution:

Additional pros not already mentioned:

1. From imaging, potentially interesting information such as cell and nuclei morphology and subcellular transcript localization is available.
2. single-cell resolution, in 3D, especially when the cell membrane is visualized by staining for a membrane protein.

Compared to ISS:

Pros:

1. Higher detection efficiency
2. More genes profiled with high detection efficiency. FISSEQ is extremely inefficient.

Cons:

1. With RCA signal amplification and lower detection efficiency, ISS and HybISS can be applied to larger areas of tissues more easily as lower magnification is used when imaging.
2. Untargeted ExSeq and INSTA-seq coupled with NGS can be used to explore unknown transcripts and isoforms not confined to probes for known transcripts.

Choosing the right technique

For the prospective user, which technique is right? Perhaps consider the following questions:

1. How important is single-cell resolution to your study? Is it so important that it justifies more labor intensive non-commercial techniques or buying newly released commercial equipment that have not yet entered core facilities? Would computational cell type deconvolution be satisfactory?
2. How important is high detection efficiency to your study? Lower efficiency as in scRNA-seq can still discern more common cell types.
3. How important is profiling the whole transcriptome in space to your study? If the spatial dataset is not transcriptome wide, then would data integration with scRNA-seq to impute gene expression in space be satisfactory for now given constraints of the current state of spatial transcriptomics? Moreover, focusing on a panel of genes rather than profiling the whole transcriptome is not necessarily a bad thing when only the panel is of interest to reduce sequencing cost. Visium has an option to profile panels of genes rather than the whole transcriptome.
4. How important are subcellular transcript localization and cell and nuclei morphology to your study?

If these distinctive features of highly multiplexed smFISH are not as important, then perhaps stick to a common commercial options such as Visium, LCM, and GeoMX DSP.

5. What are the spatial axes of interest? If there is only one axis of interest and the other axes are much less important for the questions asked, then perhaps Tomo-seq is a good choice.
6. Are your samples FFPE? If so, then consider one of Visium, GeoMX DSP, CosMX, LCM, and DBiT-seq.

7. Are you aiming to characterize the tissue section in an untargeted way, or are you more focused on specific histological regions? If the latter, then perhaps ROI based techniques.
8. Are you trying to profile the transcriptome of a relatively large area of tissue, such as an entire coronal section of a mouse brain hemisphere?
9. How important are novel transcripts and isoforms to your study?

Note that "important" here really means what kind of trade-offs are better for the purpose of a study, as the extra biological information in the above questions can all be important to more fully understanding the biological system of interest. However, at present, as discussed in the pros and cons throughout this chapter, trade-offs are necessary among cost, convenience, detection efficiency, spatial resolution, area of tissue covered, number of genes profiled, and FFPE compatibility. We are unaware of any technique that excels in all these.

Furthermore, all current era techniques have limits in some of the above aspects. For instance, in terms of tissue area, "large" means something like the size of a mouse brain, in the scale of several mm's per side. However, human tissues are often much larger, in the scale of several cm's per side. While large microscope slides for human brain sections up to 178 mm $\times$ 127 mm are available (e.g. ClariTex Super Mega Slides), we are unaware of current or prequel era datasets of human brain sections not divided into smaller parts. Even the ABA human ISH atlas is from smaller parts of the human brain and not registered to a CCF. Getting from 6.5 mm $\times$ 6.5 mm in one Visium tissue capture area to 6.5 cm $\times$ 6.5 cm in 2D means 100 times more imaging, which is even more time consuming with the high magnification used in smFISH so each FOV is only about 200 μm per side. As highly multiplexed smFISH studies already report days of image processing even with multiple CPU cores (e.g. 4 days for osmFISH of mouse cortex, which is a relatively large area compared to most other highly multiplexed smFISH datasets [107], so 100 times more imaging would mean months to years of image processing. While deep learning based cell segmentation with GPU might speed up image processing, some steps in image processing do not typically use deep learning, such as image stitching, which becomes more onerous with 100 times more FOVs. For NGS based techniques, this means 100 times more sequencing, with the increased cost and challenges in data processing.

7.8 Spatial multi-omics

Some spatial transcriptomics techniques have been adapted to collect data of other modalities, such as proteomics, neuron projection (connectomics), and 3D chromatin conformation. Here multi-omics means that data from different modalities are collected from the same piece of tissue, rather than from adjacent sections. These modalities can give a fuller picture of cell state than transcriptomics alone.

In both MERFISH [135] and GeoMx DSP [78], a panel of proteins can be quantified with oligonucleotide tagged antibodies, and the oligo tag is detected and counted as spots just like mRNA. In GeoMX DSP, if using all the 10-plex antibody panels from Nanostring plus the 4 antibodies from the core panel, then 144 proteins can be quantified at once and the barcodes need to be quantified with NGS. Antibodies tagged with oligonucleotides with poly-A tails can also be incorporated into ST as SM-Omics [231], and into DBiT-seq [204]. Now Visium can be performed on immunofluorescence tissue sections, as well as with an antibody panel for proteins [232]. We have already mentioned adaptation of MERFISH targeting introns and genomic DNA to determine 3D chromatin conformation [165, 164], and pseudocolor seqFISH and seqFISH+ have been used for this purpose in cell culture as well [101, 104]. In MERFISH, traditional Nissl or poly-A based staining miss cellular processes, but neuron projection tracing can be performed prior to MERFISH. In the mouse motor cortex MERFISH atlas [163], axons are first visualized by injecting cholera toxin subunit b (CTb) conjugated to 3 different dyes into 3 cortical areas as a retrograde tracer, tracing from terminals of the axons to the cell bodies. After imaging the axons, transcripts are imaged and quantified with MERFISH so neuronal projection can be related to the transcriptome. Viruses can be used for anterograde tracing, i.e. from cell bodies to axon terminals [233], and can in theory be performed prior to MERFISH imaging. Axonal projections are traced in BARseq(2), as already mentioned in 7.3.

Other types of measurements more sophisticated than H&E staining but are not an -omics per se have been performed on the same tissue along with spatial transcriptomics as well. The Allen Institute profiled 4 modalities in the same cells in explanted human brain slices [234], albeit at a small scale: HCR-smFISH for up to 9 genes, action potential recording of up to 5 neurons with multi patch-clamp, intrinsic membrane properties by step depolarization, and morphology of axons and dendrites by biocytin/streptavidin staining. smFISH and biocytin staining were performed after the cells were fixed after the electrophysiological profiling. Electro-

physiological recordings from cultured cardiomyocytes in space has been coupled to STARmap (201 genes) with electro-seq; again, the recording is performed before the cells are fixed and cleared for STARmap [235]. With soft electronics that integrates into the cultured tissue in hydrogel, electro-seq is less invasive than patch-clamp recording, which breaks the membrane and causes transcripts to leak.

7.9 Databases of the current era

The database holding various spatial gene expression data was proposed early in the prequel era (1990s), when enhancer and gene trap data was proliferating and major WMISH atlas projects were in progress. In contrast, in the current era, databases only emerged after datasets from various techniques have already proliferated. One of such databases is SpatialDB [236], published in late 2019, which holds gene count matrices from ST, LCM, Tomo-seq, and etc. and spatially variable genes identified with SpatialDE [237] and trendsceek [238]. In addition, the SpatialDB website provides interactive visualization of gene expression in space. Data can be queried by gene symbols, species, and data collection techniques. Unfortunately, it seems that SpatialDB has not been updated since 2020.

Another database is the Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative - Cell Census Network (BICCN) [239]. This is an international collaboration providing and generating multi-modal data for the mouse, human, and non-human primate brain, collected with scRNA-seq, ATAC-seq, neuron projection tracing, MRI, IHC, MERFISH [163], osmFISH, seqFISH, etc. The database website is hosted by the Allen Institute, and thus may be considered a continuation of the ABA. Data can be queried by species, technique, modality, and the lab that generated the data, but not by gene symbols. Also, to recap, the HCA has human Visium and HybISS data, though spatial transcriptomics does not seem to be its focus.

While current era mouse brain atlases still reference the prequel ABA ontologies [203, 180, 231, 105], data cannot be queried by ontology in the current era databases, nor by a reference gene expression pattern as in the prequel database FlyExpress [240]. With more quantitative and comprehensive data, the traditional ontology may need to be revised. Unlike prequel databases such as ABA, EMAGE, and FlyExpress, to the best of our knowledge, current era spatial data has not been systematically registered to a 3D model for integrative analysis across datasets and for visualization.

References

1. Bidarimath M, Edwards AK, and Tayade C. Laser Capture Microdissection for Gene Expression Analysis. Methods in Molecular Biology 2015; 1219:115–37. DOI: 10.1007/978-1-4939-1661-0{_}10. Available from: http://link.springer.com/10.1007/978-1-4939-1661-0_10

2. Greulich KO. Introduction: The history of using light as a working tool. Micromanipulation by Light in Biology and Medicine 1999 :1–6. DOI: 10.1007/978-1-4612-4110-2{_}1. Available from: http://link.springer.com/10.1007/978-1-4612-4110-2_1

3. Berns MW, Olson RS, and Rounds DE. In vitro Production of Chromosomal Lesions with an Argon Laser Microbeam. Nature 1969 Jan; 221:74–5. DOI: 10.1038/221074a0. Available from: http://www.nature.com/articles/221074a0

4. Meier-Ruge W, Bielser W, Remy E, Hillenkamp F, Nitsche R, and Unsold R. The laser in the Lowry technique for microdissection of freeze-dried tissue slices. The Histochemical Journal 1976 Jul; 8:387–401. DOI: 10.1007/BF01003828. Available from: http://link.springer.com/10.1007/BF01003828

5. Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, and Liotta LA. Laser Capture Microdissection. Science 1996 Nov; 274:998–1001. DOI: 10.1126/science.274.5289.998. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.274.5289.998

6. Luo L, Salunga RC, Guo H, Bittner A, Joy K, Galindo JE, Xiao H, Rogers KE, Wan JS, Jackson MR, and Erlander MG. Gene expression profiles of laser-captured adjacent neuronal subtypes. Nature Medicine 1999 Jan; 5:117–22. DOI: 10.1038/4806. Available from: http://www.nature.com/articles/nm0199_117

7. Ohyama H, Zhang X, Kohno Y, Alevizos I, Posner M, Wong D, and Todd R. Laser Capture Microdissection-Generated Target Sample for High-Density Oligonucleotide Array Hybridization. BioTechniques 2000 Sep; 29:530–6. DOI: 10.2144/00293st05. Available from: https://www.future-science.com/doi/10.2144/00293st05

8. Sgroi DC, Teng S, Robinson G, LeVangie R, Hudson JR, and Elkahloun AG. In Vivo Gene Expression Profile Analysis of Human Breast Cancer Progression. Cancer Research 1999 Nov; 59:5656 LP –5661. Available from: http://cancerres.aacrjournals.org/content/59/22/5656.abstract

9. Kitahara O, Furukawa Y, Tanaka T, Kihara C, Ono K, Yanagawa R, Nita ME, Takagi T, Nakamura Y, and Tsunoda T. Alterations of Gene Expression during Colorectal Carcinogenesis Revealed by cDNA Microarrays after Laser-Capture Microdissection of Tumor Tissues and Normal Epithelia. Cancer Research 2001 May; 61:3544 LP –3549. Available from: http://cancerres.aacrjournals.org/content/61/9/3544.abstract

10. Becker I, Becker KF, Röhrl MH, Minkus G, Schütze K, and Höfler H. Single-cell mutation analysis of tumors from stained histologic slides. Laboratory Investigation 1996 Dec; 75:801–7. Available from: https://pubmed.ncbi.nlm.nih.gov/8973475/

11. Nakamura T, Furukawa Y, Nakagawa H, Tsunoda T, Ohigashi H, Murata K, Ishikawa O, Ohgaki K, Kashimura N, Miyamoto M, Hirano S, Kondo S, Katoh H, Nakamura Y, and Katagiri T. Genome-wide cDNA microarray analysis of gene expression profiles in pancreatic cancers using populations of tumor cells and normal ductal epithelial cells selected for purity by laser microdissection. Oncogene 2004 Mar; 23:2385–400. DOI: 10.1038/sj.onc.1207392. Available from: http://www.nature.com/articles/1207392

12. Moor AE, Golan M, Massasa EE, Lemze D, Weizman T, Shenhav R, Baydatch S, Mizrahi O, Winkler R, Golani O, Stern-Ginossar N, and Itzkovitz S. Global mRNA polarization regulates translation efficiency in the intestinal epithelium. Science 2017 Sep; 357:1299–303. DOI: 10.1126/science.aan2399. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aan2399

13. Zechel S, Zajac P, Lönnerberg P, Ibáñez CF, and Linnarsson S. Topographical transcriptome mapping of the mouse medial ganglionic eminence by spatially resolved RNA-seq. Genome biology 2014 Oct; 15:486. DOI: 10.1186/s13059-014-0486-z. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0486-z

14. Baccin C, Al-Sabah J, Velten L, Helbling PM, Grünschläger F, Hernández-Malmierca P, Nombela-Arrieta C, Steinmetz LM, Trumpp A, and Haas S. Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. Nature Cell Biology 2020 Jan; 22:38–48. DOI: 10.1038/s41556-019-0439-6. Available from: https://doi.org/10.1038/s41556-019-0439-6

15. Arcturus XT Laser Capture Microdissection (LCM) Instrument - US. Available from: //www.thermofisher.com/us/en/home/life-science/gene-expression-analysis-genotyping/laser-capture-microdissection/arcturus-laser-capture-microdissection-lcm-instrument.html

16. Tang F, Lao K, and Surani MA. Development and applications of single-cell transcriptome analysis. eng. Nature methods 2011 Apr; 8:S6–S11. DOI: 10.1038/nmeth.1557. Available from: https://pubmed.ncbi.nlm.

nih.gov/21451510%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3408593/

17. Belyavsky A, Vinogradova T, and Rajewsky K. PCR-based cDNA library construction: general cDNA libraries at the level of a few cells. eng. Nucleic acids research 1989 Apr; 17:2919–32. DOI: 10.1093/nar/17.8.2919. Available from: https://pubmed.ncbi.nlm.nih.gov/2471144%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC317702/

18. Brady G, Barbara M, and Iscove NN. Representative in Vitro cDNA Amplification From Individual Hemopoietic Cells and Colonies. Methods in Molecular and Cellular Biology 1990; 2:17–25

19. Ko MSH, Ko SBH, Takahashi N, Nishiguchi K, and Abe K. Unbiased amplification of a highly complex mixture of DNA fragments by 'lone linker'-tagged PCR. Nucleic Acids Research 1990 Jul; 18:4293. DOI: 10.1093/nar/18.14.4293. Available from: https://doi.org/10.1093/nar/18.14.4293

20. Dulac C and Axel R. A novel family of genes encoding putative pheromone receptors in mammals. Cell 1995; 83:195–206. DOI: https://doi.org/10.1016/0092-8674(95)90161-2. Available from: https://www.sciencedirect.com/science/article/pii/0092867495901612

21. Klein CA, Seidl S, Petat-Dutter K, Offner S, Geigl JB, Schmidt-Kittler O, Wendler N, Passlick B, Huber RM, Schlimok G, Baeuerle PA, and Riethmüller G. Combined transcriptome and genome analysis of single micrometastatic cells. Nature Biotechnology 2002; 20:387–92. DOI: 10.1038/nbt0402-387. Available from: https://doi.org/10.1038/nbt0402-387

22. Tietjen I, Rihel JM, Cao Y, Koentges G, Zakhary L, and Dulac C. Single-Cell Transcriptional Analysis of Neuronal Progenitors. Neuron 2003; 38:161–75. DOI: https://doi.org/10.1016/S0896-6273(03)00229-0. Available from: https://www.sciencedirect.com/science/article/pii/S0896627303002290

23. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, and Surani MA. mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods 2009; 6:377–82. DOI: 10.1038/nmeth.1315. Available from: https://doi.org/10.1038/nmeth.1315

24. Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, and Linnarsson S. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. eng. Genome research 2011 Jul; 21:1160–7. DOI: 10.1101/gr.110882.110. Available from: https://pubmed.ncbi.nlm.nih.gov/21543516%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3129258/

25. Zhu YY, Machleder EM, Chenchik A, Li R, and Siebert PD. Reverse Transcriptase Template Switching: A SMART™ Approach for Full-Length cDNA Library Construction. BioTechniques 2001 Apr; 30:892–7. DOI: 10.2144/01304pf02. Available from: https://doi.org/10.2144/01304pf02

26. Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, Schroth GP, and Sandberg R. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. eng. Nature biotechnology 2012 Aug; 30:777–82. DOI: 10.1038/nbt.2282. Available from: https://pubmed.ncbi.nlm.nih.gov/22820318%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3467340/

27. Nichterwitz S, Chen G, Aguila Benitez J, Yilmaz M, Storvall H, Cao M, Sandberg R, Deng Q, and Hedlund E. Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. Nature Communications 2016 Nov; 7:12139. DOI: 10.1038/ncomms12139. Available from: http://www.nature.com/articles/ncomms12139

28. Van Gelder RN, Zastrow ME von, Yool A, Dement WC, Barchas JD, and Eberwine JH. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proceedings of the National Academy of Sciences 1990 Mar; 87:1663 LP –1667. Available from: http://www.pnas.org/content/87/5/1663.abstract

29. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, and Coleman P. Analysis of gene expression in single live neurons. Proceedings of the National Academy of Sciences 1992 Apr; 89:3010 LP –3014. DOI: 10.1073/pnas.89.7.3010. Available from: http://www.pnas.org/content/89/7/3010.abstract

30. Liang P and Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 1992 Aug; 257:967 LP –971. DOI: 10.1126/science.1354393. Available from: http://science.sciencemag.org/content/257/5072/967.abstract

31. Kacharmina JE, Crino PB, and Eberwine JBTMiE. Preparation of cDNA from single cells and subcellular regions. cDNA Preparation and Characterization 1999; 303:3–18. DOI: https://doi.org/10.1016/S0076-6879(99)03003-7. Available from: https://www.sciencedirect.com/science/article/pii/S0076687999030037

32. Hemby SE, Ginsberg SD, Brunk B, Arnold SE, Trojanowski JQ, and Eberwine JH. Gene Expression Profile for Schizophrenia: Discrete Neuron Transcription Patterns in the Entorhinal Cortex. Archives of General Psychiatry 2002 Jul; 59:631–40. DOI: 10.1001/archpsyc.59.7.631. Available from: https://doi.org/10.1001/archpsyc.59.7.631

33. Kamme F, Salunga R, Yu J, Tran DT, Zhu J, Luo L, Bittner A, Guo HQ, Miller N, Wan J, and Erlander M. Single-Cell Microarray Analysis in Hippocampus CA1: Demonstration and Validation of Cellular Heterogeneity. The Journal of Neuroscience 2003 May; 23:3607 LP –3615. DOI: 10.1523/JNEUROSCI.23-09-03607.2003. Available from: http://www.jneurosci.org/content/23/9/3607.abstract

34. Hashimshony T, Wagner F, Sher N, and Yanai I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. Cell Reports 2012; 2:666–73. DOI: https://doi.org/10.1016/j.celrep.2012.08.003. Available from: https://www.sciencedirect.com/science/article/pii/S2211124712002288

35. Tzur YB, Winter E, Gao J, Hashimshony T, Yanai I, and Colaiácovo MP. Spatiotemporal Gene Expression Analysis of the Caenorhabditis elegans Germline Uncovers a Syncytial Expression Switch. Genetics 2018 Oct; 210:587–605. DOI: 10.1534/genetics.118.301315. Available from: http://www.genetics.org/lookup/doi/10.1534/genetics.118.301315

36. Junker JP, Noël ES, Guryev V, Peterson KA, Shah G, Huisken J, McMahon AP, Berezikov E, Bakkers J, and Van Oudenaarden A. Genome-wide RNA Tomography in the Zebrafish Embryo. Cell 2014. DOI: 10.1016/j.cell.2014.09.038

37. Medaglia C, Giladi A, Stoler-Barak L, De Giovanni M, Salame TM, Biram A, David E, Li H, Iannacone M, Shulman Z, and Amit I. Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. Science 2017 Dec; 358:1622–6. DOI: 10.1126/science.aao4277. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aao4277

38. Aguila J, Cheng S, Kee N, Cao M, Deng Q, and Hedlund E. Spatial transcriptomics and in silico random pooling identify novel markers of vulnerable and resistant midbrain dopamine neurons. bioRxiv 2018 Oct :334417. DOI: 10.1101/334417. Available from: https://doi.org/10.1101/334417

39. Moor AE, Harnik Y, Ben-Moshe S, Massasa EE, Rozenberg M, Eilam R, Bahar Halpern K, and Itzkovitz S. Spatial Reconstruction of Single Enterocytes Uncovers Broad Zonation along the Intestinal Villus Axis. Cell 2018 Nov; 175:1156–67. DOI: 10.1016/j.cell.2018.08.063. Available from: https://doi.org/10.1016/j.cell.2018.08.063

40. Peng G, Suo S, Chen J, Chen W, Liu C, Yu F, Wang R, Chen S, Sun N, Cui G, Song L, Tam PP, Han JDJ, and Jing N. Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. Developmental Cell 2016 Mar; 36:681–97. DOI: 10.1016/j.devcel.2016.02.020. Available from: http://dx.doi.org/10.1016/j.devcel.2016.02.020

41. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, Lagemaat LN van de, Smith KA, Ebbert A, Riley ZL, Abajian C, Beckmann CF, Bernard A, Bertagnolli D, Boe AF, Cartagena PM, Chakravarty MM, Chapin M, Chong J, Dalley RA, Daly BD, Dang C, Datta S, Dee N, Dolbeare TA, Faber V, Feng D, Fowler DR, Goldy J, Gregor BW, Haradon Z, Haynor DR, Hohmann JG, Horvath S, Howard RE, Jeromin A, Jochim JM, Kinnunen M, Lau C, Lazarz ET, Lee C, Lemon TA, Li L, Li Y, Morris JA, Overly CC, Parker PD, Parry SE, Reding M, Royall JJ, Schulkin J, Sequeira PA, Slaughterbeck CR, Smith SC, Sodt AJ, Sunkin SM, Swanson BE, Vawter MP, Williams D, Wohnoutka P, Zielke HR, Geschwind DH, Hof PR, Smith SM, Koch C, Grant SGN, and Jones AR. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature 2012 Sep; 489:391–9. DOI: 10.1038/nature11405. Available from: http://www.nature.com/articles/nature11405

42. Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall JJ, Aiona K, Arnold JM, Bennet C, Bertagnolli D, Brouner K, Butler S, Caldejon S, Carey A, Cuhaciyan C, Dalley RA, Dee N, Dolbeare TA, Facer BAC, Feng D, Fliss TP, Gee G, Goldy J, Gourley L, Gregor BW, Gu G, Howard RE, Jochim JM, Kuan CL, Lau C, Lee CK, Lee F, Lemon TA, Lesnar P, McMurray B, Mastan N, Mosqueda N, Naluai-Cecchini T, Ngo NK, Nyhus J, Oldre A, Olson E, Parente J, Parker PD, Parry SE, Stevens A, Pletikos M, Reding M, Roll K, Sandman D, Sarreal M, Shapouri S, Shapovalova NV, Shen EH, Sjoquist N, Slaughterbeck CR, Smith M, Sodt AJ, Williams D, Zöllei L, Fischl B, Gerstein MB, Geschwind DH, Glass IA, Hawrylycz MJ, Hevner RF, Huang H, Jones AR, Knowles JA, Levitt P, Phillips JW, Šestan N, Wohnoutka P, Dang C, Bernard A, Hohmann JG, and Lein ES. Transcriptional landscape of the prenatal human brain. Nature 2014 Apr; 508:199–206. DOI: 10.1038/nature13185. Available from: http://brainspan.org%20http://www.nature.com/articles/nature13185

43. Bakken TE, Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A, Dalley RA, Royall JJ, Lemon T, Shapouri S, Aiona K, Arnold J, Bennett JL, Bertagnolli D, Bickley K, Boe A, Brouner K, Butler S, Byrnes E, Caldejon S, Carey A, Cate S, Chapin M, Chen J, Dee N, Desta T, Dolbeare TA, Dotson N, Ebbert A, Fulfs E, Gee G, Gilbert TL, Goldy J, Gourley L, Gregor B, Gu G, Hall J, Haradon Z, Haynor DR, Hejazinia N, Hoerder-Suabedissen A, Howard R, Jochim J, Kinnunen M, Kriedberg A, Kuan CL, Lau C, Lee CK, Lee F, Luong L, Mastan N, May R, Melchor J, Mosqueda N, Mott E, Ngo K, Nyhus J, Oldre A, Olson E, Parente J, Parker PD, Parry S, Pendergraft J, Potekhina L, Reding M, Riley ZL, Roberts T, Rogers B, Roll K, Rosen D, Sandman D, Sarreal M, Shapovalova N, Shi S, Sjoquist N, Sodt AJ, Townsend R, Velasquez L, Wagley U, Wakeman WB, White C, Bennett C, Wu J, Young R, Youngstrom BL, Wohnoutka P, Gibbs RA, Rogers J, Hohmann JG, Hawrylycz MJ, Hevner RF, Molnár Z, Phillips JW, Dang C, Jones AR,

Amaral DG, Bernard A, and Lein ES. A comprehensive transcriptional map of primate brain development. Nature 2016 Jul; 535:367–75. DOI: 10.1038/nature18637. Available from: http://www.blueprintnhpatlas.org

44. Translational Pathology Core Laboratory (TPCL). Available from: https://www.uclahealth.org/pathology/tpcl-services

45. Veritas Laser Capture Microdissection (LCM) and Laser Cutting System from Applied Biosystems. Available from: https://www.unthsc.edu/research/flow-cytometry-and-laser-capture-microdissection-core-facility/beckman-coulter-cytomics-fc500-flow-cytometry-analyzer/veritas-laser-capture

46. Dana-Farber Core Facilities. Available from: https://www.dana-farber.org/research/core-facilities/

47. Johns Hopkins Cell Imaging Core Facility. Available from: https://www.hopkinsmedicine.org/kimmel_cancer_center/research/shared_resources/cell_imaging.html

48. Villacampa EG, Larsson L, Kvastad L, Andersson A, Carlson J, and Lundeberg J. Genome-wide Spatial Expression Profiling in FFPE Tissues. bioRxiv 2020 Jul :2020.07.24.219758. DOI: 10.1101/2020.07.24.219758. Available from: https://doi.org/10.1101/2020.07.24.219758

49. Hwang WL, Jagadeesh KA, Guo JA, Hoffman HI, Yadollahpour P, Mohan R, Drokhlyansky E, Van Wittenberghe N, Ashenberg O, Farhi S, Schapiro D, Reeves J, Zollinger DR, Eng G, Schenkel JM, Freed-Pastor WA, Rodrigues C, Gould J, Lambden C, Porter C, Tsankov A, Dionne D, Abbondanza D, Waldman J, Cuoco M, Nguyen L, Delorey T, Phillips D, Ciprani D, Kern M, Mehta A, Fuhrman K, Fropf R, Beechem J, Loeffler JS, Ryan DP, Weekes CD, Ting DT, Ferrone CR, Wo JY, Hong TS, Aguirre AJ, Rozenblatt-Rosen O, Mino-Kenudson M, Fernandez-Del Castillo C, Liss AS, Jacks T, and Regev A. Single-nucleus and spatial transcriptomics of archival pancreatic cancer reveals multi-compartment reprogramming after neoadjuvant treatment. bioRxiv 2020 Aug :2020.08.25.267336. DOI: 10.1101/2020.08.25.267336. Available from: https://doi.org/10.1101/2020.08.25.267336

50. Coudry RA, Meireles SI, Stoyanova R, Cooper HS, Carpino A, Wang X, Engstrom PF, and Clapper ML. Successful Application of Microarray Technology to Microdissected Formalin-Fixed, Paraffin-Embedded Tissue. The Journal of Molecular Diagnostics 2007 Feb; 9:70–9. DOI: 10.2353/jmoldx.2007.060004. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1525157810603637

51. Morton ML, Bai X, Merry CR, Linden PA, Khalil AM, Leidner RS, and Thompson CL. Identification of mRNAs and lincRNAs associated with lung cancer progression using next-generation RNA sequencing from laser micro-

dissected archival FFPE tissue specimens. Lung Cancer 2014 Jul; 85:31–9. DOI: 10.1016/j.lungcan.2014.03.020. Available from: https://linkinghub.elsevier.com/retrieve/pii/S016950021400141X

52. Foley JW, Zhu C, Jolivet P, Zhu SX, Lu P, Meaney MJ, and West RB. Gene expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ. Genome Research 2019 Nov; 29:1816–25. DOI: 10.1101/gr.234807.118. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.234807.118

53. Brown VM. High-Throughput Imaging of Brain Gene Expression. Genome Research 2002 Feb; 12:244–54. DOI: 10.1101/gr.204102. Available from: http://www.genome.org/cgi/doi/10.1101/gr.204102

54. Singh RP, Brown VM, Chaudhari A, Khan AH, Ossadtchi A, Sforza DM, Meadors A, Cherry SR, Leahy RM, and Smith DJ. High-resolution voxelation mapping of human and rodent brain gene expression. Journal of Neuroscience Methods 2003 May; 125:93–101. DOI: 10.1016/S0165-0270(03)00045-1. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0165027003000451

55. Brown VM, Ossadtchi A, Khan AH, Yee S, Lacan G, Melega WP, Cherry SR, Leahy RM, and Smith DJ. Multiplex Three-Dimensional Brain Gene Expression Mapping in a Mouse Model of Parkinson's Disease. Genome Research 2002 May; 12:868–84. DOI: 10.1101/gr.229002. Available from: http://www.genome.org/cgi/doi/10.1101/gr.229002

56. Chin MH, Geng AB, Khan AH, Qian WJ, Petyuk VA, Boline J, Levy S, Toga AW, Smith RD, Leahy RM, and Smith DJ. A genome-scale map of expression for a mouse brain section obtained using voxelation. Physiological Genomics 2007 Aug; 30:313–21. DOI: 10.1152/physiolgenomics.00287.2006. Available from: https://www.physiology.org/doi/10.1152/physiolgenomics.00287.2006

57. An L, Xie H, Chin MH, Obradovic Z, Smith DJ, and Megalooikonomou V. Analysis of multiplex gene expression maps obtained by voxelation. BMC Bioinformatics 2009 Apr; 10:S10. DOI: 10.1186/1471-2105-10-S4-S10. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-S4-S10

58. Brink SC van den, Alemany A, Batenburg V van, Moris N, Blotenburg M, Vivié J, Baillie-Johnson P, Nichols J, Sonnen KF, Martinez Arias A, and Oudenaarden A van. Single-cell and spatial transcriptomics reveal somitogenesis in gastruloids. Nature 2020 Feb; 582:405. DOI: 10.1038/s41586-020-2024-3. Available from: https://doi.org/10.1038/s41586-020-2024-3

59. Okamura-Oho Y, Shimokawa K, Takemoto S, Hirakiyama A, Nakamura S, Tsujimura Y, Nishimura M, Kasukawa T, Masumoto Kh, Nikaido I, Shigeyoshi Y, Ueda HR, Song G, Gee J, Himeno R, and Yokota H. Transcriptome Tomography for Brain Analysis in the Web-Accessible Anatomical Space. PLoS ONE 2012 Sep; 7. Ed. by Hayasaka S:e45373. DOI: 10.1371/journal.pone.0045373. Available from: https://dx.plos.org/10.1371/journal.pone.0045373

60. Combs PA and Eisen MB. Sequencing mRNA from Cryo-Sliced Drosophila Embryos to Determine Genome-Wide Spatial Patterns of Gene Expression. PLoS ONE 2013 Aug; 8. Ed. by Jennings B:e71820. DOI: 10.1371/journal.pone.0071820. Available from: https://dx.plos.org/10.1371/journal.pone.0071820

61. Ebbing A, Vertesy A, Betist M, Spanjaard B, Junker JP, Berezikov E, Oudenaarden A van, and Korswagen H. Spatial transcriptomics of C. elegans males and hermaphrodites identifies novel fertility genes. bioRxiv 2018 Jun :348201. DOI: 10.1101/348201. Available from: https://www.biorxiv.org/content/10.1101/348201v1

62. Burkhard SB and Bakkers J. Spatially resolved RNA-sequencing of the embryonic heart identifies a role for Wnt/β-catenin signaling in autonomic control of heart rate. eLife 2018 Feb; 7. DOI: 10.7554/eLife.31515

63. Combs PA and Fraser HB. Spatially varying cis-regulatory divergence in Drosophila embryos elucidates cis-regulatory logic. PLOS Genetics 2018 Nov; 14. Ed. by Desplan C:e1007631. DOI: 10.1371/journal.pgen.1007631. Available from: https://dx.plos.org/10.1371/journal.pgen.1007631

64. Lacraz GP, Junker JP, Gladka MM, Molenaar B, Scholman KT, Vigil-Garcia M, Versteeg D, Ruiter H de, Vermunt MW, Creyghton MP, Huibers MM, Jonge N de, Oudenaarden A van, and Rooij E van. Tomo-Seq Identifies SOX9 as a Key Regulator of Cardiac Fibrosis During Ischemic Injury. Circulation 2017 Oct; 136:1396–409. DOI: 10.1161/CIRCULATIONAHA.117.027832. Available from: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.117.027832

65. Rödelsperger C, Ebbing A, Sharma DR, Okumura M, Sommer RJ, and Korswagen HC. Spatial transcriptomics of nematodes identifies sperm cells as a source of genomic novelty and rapid evolution. Molecular Biology and Evolution 2020 Aug. DOI: 10.1093/molbev/msaa207. Available from: https://doi.org/10.1093/molbev/msaa207

66. Moris N, Anlas K, Brink SC van den, Alemany A, Schröder J, Ghimire S, Balayo T, Oudenaarden A van, and Martinez Arias A. An in vitro model of early anteroposterior organization during human development. Nature 2020 Jun; 582:410–5. DOI: 10.1038/s41586-020-2383-9. Available from: https://doi.org/10.1038/s41586-020-2383-9

67. Brown VM, Ossadtchi A, Khan AH, Gambhir SS, Cherry SR, Leahy RM, and Smith DJ. Gene expression tomography. Physiological Genomics 2002 Feb; 8:159–67. DOI: 10.1152/physiolgenomics.00090.2001. Available from: https://doi.org/10.1152/physiolgenomics.00090.2001

68. Schede HH, Schneider CG, Stergiadou J, Borm LE, Ranjak A, Yamawaki TM, David FPA, Lonnerberg P, Laurent G, Tosches MA, Codeluppi S, and La Manno G. Spatial tissue profiling by imaging-free molecular tomography. bioRxiv 2020 Aug :2020.08.04.235655. DOI: 10.1101/2020.08.04.235655. Available from: http://biorxiv.org/content/early/2020/08/04/2020.08.04.235655.abstract

69. Kudo LC, Vi N, Ma Z, Fields T, Avliyakulov NK, Haykinson MJ, Bragin A, and Karsten SL. Novel Cell and Tissue Acquisition System (CTAS): Microdissection of Live and Frozen Brain Tissues. PLoS ONE 2012 Jul; 7. Ed. by Aldabe R:e41564. DOI: 10.1371/journal.pone.0041564. Available from: https://dx.plos.org/10.1371/journal.pone.0041564

70. Yoda T, Hosokawa M, Takahashi K, Sakanashi C, Takeyama H, and Kambara H. Site-specific gene expression analysis using an automated tissue microdissection punching system. Scientific Reports 2017 Dec; 7:4325. DOI: 10.1038/s41598-017-04616-6. Available from: http://www.nature.com/articles/s41598-017-04616-6

71. Giolai M, Verweij W, Lister A, Heavens D, Macaulay I, and Clark MD. Spatially resolved transcriptomics reveals plant host responses to pathogens. Plant Methods 2019 Dec; 15:114. DOI: 10.1186/s13007-019-0498-5. Available from: https://plantmethods.biomedcentral.com/articles/10.1186/s13007-019-0498-5

72. Lukan T, Pompe-Novak M, Baebler Š, Tušek-Žnidarič M, Kladnik A, Križnik M, Blejec A, Zagorščak M, Stare K, Dušak B, Coll A, Pollmann S, Morgiewicz K, Hennig J, and Gruden K. Precision transcriptomics of viral foci reveals the spatial regulation of immune-signaling genes and identifies RBOHD as an important player in the incompatible interaction between potato virus Y and potato. The Plant Journal 2020 Nov; 104:645–61. DOI: 10.1111/tpj.14953. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tpj.14953

73. Plouhinec JL, Medina-Ruiz S, Borday C, Bernard E, Vert JP, Eisen MB, Harland RM, and Monsoro-Burq AH. A molecular atlas of the developing ectoderm defines neural, neural crest, placode, and nonneural progenitor identity in vertebrates. PLOS Biology 2017 Oct; 15. Ed. by Briscoe J:e2004045. DOI: 10.1371/journal.pbio.2004045. Available from: https://dx.plos.org/10.1371/journal.pbio.2004045

74. Blitz IL, Paraiso KD, Patrushev I, Chiu WT, Cho KW, and Gilchrist MJ. A catalog of Xenopus tropicalis transcription factors and their regional expression in the early gastrula stage embryo. Developmental Biology 2017 Jun; 426:409–17. DOI: 10.1016/j.ydbio.2016.07.002. Available from: https://linkinghub.elsevier.com/retrieve/pii/S001216061630118X

75. Lovatt D, Ruble BK, Lee J, Dueck H, Kim TK, Fisher S, Francis C, Spaethling JM, Wolf JA, Grady MS, Ulyanova AV, Yeldell SB, Griepenburg JC, Buckley PT, Kim J, Sul JY, Dmochowski IJ, and Eberwine J. Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. Nature Methods 2014 Feb; 11:190–6. DOI: 10.1038/nmeth.2804. Available from: https://www.nature.com/articles/nmeth.2804

76. Victora GD, Schwickert TA, Fooksman DR, Kamphorst AO, Meyer-Hermann M, Dustin ML, and Nussenzweig MC. Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter. eng. Cell 2010 Nov; 143:592–605. DOI: 10.1016/j.cell.2010.10.032. Available from: https://pubmed.ncbi.nlm.nih.gov/21074050%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3035939/

77. De Giovanni M, Cutillo V, Giladi A, Sala E, Maganuco CG, Medaglia C, Di Lucia P, Bono E, Cristofani C, Consolo E, Giustini L, Fiore A, Eickhoff S, Kastenmüller W, Amit I, Kuka M, and Iannacone M. Spatiotemporal regulation of type I interferon expression determines the antiviral polarization of CD4+ T cells. Nature Immunology 2020; 21:321–30. DOI: 10.1038/s41590-020-0596-6. Available from: https://doi.org/10.1038/s41590-020-0596-6

78. Merritt CR, Ong GT, Church SE, Barker K, Danaher P, Geiss G, Hoang M, Jung J, Liang Y, McKay-Fleisch J, Nguyen K, Norgaard Z, Sorg K, Sprague I, Warren C, Warren S, Webster PJ, Zhou Z, Zollinger DR, Dunaway DL, Mills GB, and Beechem JM. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. Nature Biotechnology 2020 38:5 2020 May; 38:586–99. DOI: 10.1038/s41587-020-0472-9. Available from: https://www.nature.com/articles/s41587-020-0472-9

79. Margaroli C, Benson P, Sharma NS, Madison MC, Robison SW, Arora N, Ton K, Liang Y, Zhang L, Patel RP, and Gaggar A. Spatial mapping of SARS-CoV-2 and H1N1 Lung Injury Identifies Differential Transcriptional Signatures. eng. Cell reports. Medicine 2021 Mar :100242. DOI: 10.1016/j.xcrm.2021.100242. Available from: https://pubmed.ncbi.nlm.nih.gov/33778787%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7985929/

80. Park J, Foox J, Hether T, Danko D, Warren S, Kim Y, Reeves J, Butler DJ, Mozsary C, Rosiene J, Shaiber A, Afshinnekoo E, MacKay M, Bram Y, Chandar V, Geiger H, Craney A, Velu P, Melnick AM, Hajirasouliha I, Beheshti A, Taylor D, Saravia-Butler A, Singh U, Wurtele ES, Schisler J,

Fennessey S, Corvelo A, Zody MC, Germer S, Salvatore S, Levy S, Wu S, Tatonetti N, Shapira S, Salvatore M, Loda M, Westblade LF, Cushing M, Rennert H, Kriegel AJ, Elemento O, Imielinski M, Borczuk AC, Meydan C, Schwartz RE, and Mason CE. Systemic Tissue and Cellular Disruption from SARS-CoV-2 Infection revealed in COVID-19 Autopsies and Spatial Omics Tissue Maps. bioRxiv 2021 Jan :2021.03.08.434433. DOI: 10.1101/2021.03.08.434433. Available from: http://biorxiv.org/content/early/2021/03/09/2021.03.08.434433.abstract

81. Roberts K, Aivazidis A, Kleshchevnikov V, Li T, Fropf R, Rhodes M, Beechem JM, Hemberg M, and Bayraktar OA. Transcriptome-wide spatial RNA profiling maps the cellular architecture of the developing human neocortex. bioRxiv 2021 Jan :2021.03.20.436265. DOI: 10.1101/2021.03.20.436265. Available from: http://biorxiv.org/content/early/2021/03/20/2021.03.20.436265.abstract

82. Gene Expression Panels | NanoString Technologies. Available from: https://www.nanostring.com/products/gene-expression-panels/gene-expression-panels-overview

83. Sharma A, Seow JJW, Dutertre CA, Pai R, Blériot C, Mishra A, Wong RMM, Singh GSN, Sudhagar S, Khalilnezhad S, Erdal S, Teo HM, Khalilnezhad A, Chakarov S, Lim TKH, Fui ACY, Chieh AKW, Chung CP, Bonney GK, Goh BKP, Chan JKY, Chow PKH, Ginhoux F, and DasGupta R. Onco-fetal Reprogramming of Endothelial Cells Drives Immunosuppressive Macrophages in Hepatocellular Carcinoma. Cell 2020 Oct; 183:377–94. DOI: 10.1016/j.cell.2020.08.040. Available from: https://doi.org/10.1016/j.cell.2020.08.040

84. Tripodo C, Zanardi F, Iannelli F, Mazzara S, Vegliante M, Morello G, Di Napoli A, Mangogna A, Facchetti F, Sangaletti S, Chiodoni C, VanShoiack A, Jeyasekharan AD, Casola S, Colombo MP, Ponzoni M, and Pileri SA. A Spatially Resolved Dark- versus Light-Zone Microenvironment Signature Subdivides Germinal Center-Related Aggressive B Cell Lymphomas. iScience 2020 Oct; 23. DOI: 10.1016/j.isci.2020.101562. Available from: https://doi.org/10.1016/j.isci.2020.101562

85. Butler D, Mozsary C, Meydan C, Foox J, Rosiene J, Shaiber A, Danko D, Afshinnekoo E, MacKay M, Sedlazeck FJ, Ivanov NA, Sierra M, Pohle D, Zietz M, Gisladottir U, Ramlall V, Sholle ET, Schenck EJ, Westover CD, Hassan C, Ryon K, Young B, Bhattacharya C, Ng DL, Granados AC, Santos YA, Servellita V, Federman S, Ruggiero P, Fungtammasan A, Chin CS, Pearson NM, Langhorst BW, Tanner NA, Kim Y, Reeves JW, Hether TD, Warren SE, Bailey M, Gawrys J, Meleshko D, Xu D, Couto-Rodriguez M, Nagy-Szakal D, Barrows J, Wells H, O'Hara NB, Rosenfeld JA, Chen Y, Steel PAD, Shemesh AJ, Xiang J, Thierry-Mieg J, Thierry-Mieg D, Iftner A, Bezdan D, Sanchez E, Campion TR, Sipley J, Cong L, Craney A, Velu P, Melnick AM, Shapira S, Hajirasouliha I, Borczuk A, Iftner T, Salvatore M,

Loda M, Westblade LF, Cushing M, Wu S, Levy S, Chiu C, Schwartz RE, Tatonetti N, Rennert H, Imielinski M, and Mason CE. Shotgun transcriptome, spatial omics, and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification, and drug interactions. Nature Communications 2021; 12:1660. DOI: 10.1038/s41467-021-21361-7. Available from: https://doi.org/10.1038/s41467-021-21361-7

86. Delorey TM et al. COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. Nature 2021; 595:107–13. DOI: 10.1038/s41586-021-03570-8. Available from: https://doi.org/10.1038/s41586-021-03570-8

87. DSP Technology Access Program (TAP)

88. Bassell GJ, Powers CM, Taneja KL, and Singer RH. Single mRNAs visualized by ultrastructural in situ hybridization are principally localized at actin filament intersections in fibroblasts. eng. The Journal of cell biology 1994 Aug; 126:863–76. DOI: 10.1083/jcb.126.4.863. Available from: https://pubmed.ncbi.nlm.nih.gov/7914201%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2120111/

89. Femino AM, Fay FS, Fogarty K, and Singer RH. Visualization of Single RNA Transcripts in Situ. Science 1998 Apr; 280:585 LP –590. DOI: 10.1126/science.280.5363.585. Available from: http://science.sciencemag.org/content/280/5363/585.abstract

90. Raj A, Bogaard P van den, Rifkin SA, Oudenaarden A van, and Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. Nature Methods 2008 Oct; 5:877–9. DOI: 10.1038/nmeth.1253. Available from: http://www.nature.com/articles/nmeth.1253

91. Nederlof PM, Flier S van der, Wiegant J, Raap AK, Tanke HJ, Ploem JS, and Ploeg M van der. Multiple fluorescence in situ hybridization. Cytometry 1990; 11:126–31. DOI: 10.1002/cyto.990110115. Available from: http://doi.wiley.com/10.1002/cyto.990110115

92. Levsky JM. Single-Cell Gene Expression Profiling. Science 2002 Aug; 297:836–40. DOI: 10.1126/science.1072241. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1072241

93. Lubeck E and Cai L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. Nature Methods 2012 9:7 2012 Jun; 9:743–8. DOI: 10.1038/nmeth.2069. Available from: https://www.nature.com/articles/nmeth.2069

94. Bates M, Dempsey GT, Chen KH, and Zhuang X. Multicolor Super-Resolution Fluorescence Imaging via Multi-Parameter Fluorophore Detection. ChemPhysChem 2012 Jan; 13:99–107. DOI: 10.1002/cphc.201100735. Available from: http://doi.wiley.com/10.1002/cphc.201100735

95. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, and Cai L. Single-cell in situ RNA profiling by sequential hybridization. 2014 Mar. DOI: 10.1038/nmeth.2892. Available from: https://www.nature.com/articles/nmeth.2892

96. Shah S, Lubeck E, Zhou W, and Cai L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. Neuron 2016. DOI: 10.1016/j.neuron.2016.10.001

97. Chen KH, Boettiger AN, Moffitt JR, Wang S, and Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 2015. DOI: 10.1126/science.aaa6090

98. Eng CHL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, Yun J, Cronin C, Karp C, Yuan GC, and Cai L. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature 2019 Apr; 568:235–9. DOI: 10.1038/s41586-019-1049-y. Available from: https://www.nature.com/articles/s41586-019-1049-y

99. Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, and Zhuang X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. Proceedings of the National Academy of Sciences of the United States of America 2016. DOI: 10.1073/pnas.1612826113

100. Xia C, Fan J, Emanuel G, Hao J, and Zhuang X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proceedings of the National Academy of Sciences of the United States of America 2019. DOI: 10.1073/pnas.1912459116

101. Shah S, Takei Y, Zhou W, Lubeck E, Yun J, Eng CHL, Koulena N, Cronin C, Karp C, Liaw EJ, Amin M, and Cai L. Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. Cell 2018. DOI: 10.1016/j.cell.2018.05.035

102. Eng CHL, Shah S, Thomassie J, and Cai L. Profiling the transcriptome with RNA SPOTs. Nature Methods 2017; 14:1153–5. DOI: 10.1038/nmeth.4500. Available from: https://doi.org/10.1038/nmeth.4500

103. Parthasarathy R. Rapid, accurate particle tracking by calculation of radial symmetry centers. Nature Methods 2012; 9:724–6. DOI: 10.1038/nmeth.2071. Available from: https://doi.org/10.1038/nmeth.2071

104. Takei Y, Yun J, Ollikainen N, Zheng S, Pierson N, White J, Shah S, Thomassie J, Eng CHL, Guttman M, Yuan GC, and Cai L. Global architecture of the nucleus in single cells by DNA seqFISH+ and multiplexed immunofluorescence. bioRxiv 2020 Jan :2020.11.29.403055. DOI: 10.1101/2020.11.29.403055. Available from: http://biorxiv.org/content/early/2020/11/30/2020.11.29.403055.abstract

105. Lohoff T, Ghazanfar S, Missarova A, Koulena N, Pierson N, Griffiths JA, Bardot ES, Eng CH, Tyser RC, Argelaguet R, Guibentif C, Srinivas S, Briscoe J, Simons BD, Hadjantonakis AK, Göttgens B, Reik W, Nichols J, Cai L, and Marioni JC. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. Nature Biotechnology 2021 40:1 2021 Sep; 40:74–85. DOI: 10.1038/s41587-021-01006-2. Available from: https://www.nature.com/articles/s41587-021-01006-2

106. Goh JJL, Chou N, Seow WY, Ha N, Cheng CPP, Chang YC, Zhao ZW, and Chen KH. Highly specific multiplexed RNA imaging in tissues with split-FISH. Nature Methods 2020 Jul; 17:689–93. DOI: 10.1038/s41592-020-0858-0. Available from: https://doi.org/10.1038/s41592-020-0858-0

107. Codeluppi S, Borm LE, Zeisel A, La Manno G, Lunteren JA van, Svensson CI, and Linnarsson S. Spatial organization of the somatosensory cortex revealed by osmFISH. Nature Methods 2018. DOI: 10.1038/s41592-018-0175-z

108. Lignell A, Kerosuo L, Streichan SJ, Cai L, and Bronner ME. Identification of a neural crest stem cell niche by Spatial Genomic Analysis. Nature Communications 2017 Dec; 8:1830. DOI: 10.1038/s41467-017-01561-w. Available from: http://www.nature.com/articles/s41467-017-01561-w

109. Wang Y, Eddison M, Fleishman G, Weigert M, Xu S, Henry FE, Wang T, Lemire AL, Schmidt U, Yang H, Rokicki K, Goina C, Svoboda K, Myers EW, Saalfeld S, Korff W, Sternson SM, and Tillberg PW. Expansion-Assisted Iterative-FISH defines lateral hypothalamus spatio-molecular organization. bioRxiv 2021 Jan :2021.03.08.434304. DOI: 10.1101/2021.03.08.434304. Available from: http://biorxiv.org/content/early/2021/03/08/2021.03.08.434304.abstract

110. Urdea MS. Synthesis and Characterization of Branched DNA (bDNA) for the Direct and Quantitative Detection of CMV, HBV, HCV, and HIV. Clinical Chemistry 1993 Apr; 39:725–6. DOI: 10.1093/clinchem/39.4.725. Available from: https://academic.oup.com/clinchem/article/39/4/725/5646913

111. Player AN, Shen LP, Kenny D, Antao VP, and Kolberg JA. Single-copy gene detection using branched DNA (bDNA) in situ hybridization. Journal of Histochemistry and Cytochemistry 2001 May; 49:603–11. DOI: 10.1177/002215540104900507. Available from: http://journals.sagepub.com/doi/10.1177/002215540104900507

112. Wang F, Flanagan J, Su N, Wang LC, Bui S, Nielson A, Wu X, Vo HT, Ma XJ, and Luo Y. RNAscope. The Journal of Molecular Diagnostics 2012 Jan; 14:22–9. DOI: 10.1016/j.jmoldx.2011.08.002. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1525157811002571

113. Battich N, Stoeger T, and Pelkmans L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. Nature Methods 2013 Nov; 10:1127–36. DOI: 10.1038/nmeth.2657. Available from: https://www.nature.com/articles/nmeth.2657

114. Bayraktar OA, Bartels T, Holmqvist S, Kleshchevnikov V, Martirosyan A, Polioudakis D, Ben Haim L, Young AM, Batiuk MY, Prakash K, Brown A, Roberts K, Paredes MF, Kawaguchi R, Stockley JH, Sabeur K, Chang SM, Huang E, Hutchinson P, Ullian EM, Hemberg M, Coppola G, Holt MG, Geschwind DH, and Rowitch DH. Astrocyte layers in the mammalian cerebral cortex revealed by a single-cell in situ transcriptomic map. Nature Neuroscience 2020 Apr; 23:500–9. DOI: 10.1038/s41593-020-0602-1. Available from: https://doi.org/10.1038/s41593-020-0602-1

115. Xia C, Babcock HP, Moffitt JR, and Zhuang X. Multiplexed detection of RNA using MERFISH and branched DNA amplification. Scientific Reports 2019 Dec; 9:1–13. DOI: 10.1038/s41598-019-43943-8. Available from: https://www.nature.com/articles/s41598-019-43943-8

116. Nilsson M, Malmgren H, Samiotaki M, Kwiatkowski M, Chowdhary B, and Landegren U. Padlock probes: circularizing oligonucleotides for localized DNA detection. Science 1994 Sep; 265:2085–8. DOI: 10.1126/science.7522346. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.7522346

117. Larsson C, Grundberg I, Söderberg O, and Nilsson M. In situ detection and genotyping of individual mRNA molecules. Nature Methods 2010 May; 7:395–7. DOI: 10.1038/nmeth.1448. Available from: http://www.nature.com/articles/nmeth.1448

118. Lizardi PM, Huang X, Zhu Z, Bray-Ward P, Thomas DC, and Ward DC. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. Nature Genetics 1998 Jul; 19:225–32. DOI: 10.1038/898. Available from: http://www.nature.com/articles/ng0798_225

119. Fire A and Xu SQ. Rolling replication of short DNA circles. Proceedings of the National Academy of Sciences 1995 May; 92:4641–5. DOI: 10.1073/pnas.92.10.4641. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.92.10.4641

120. Baner J, Nilsson M, Mendel-Hartvig M, and Landegren U. Signal amplification of padlock probes by rolling circle replication. Nucleic Acids Research 1998 Nov; 26:5073–8. DOI: 10.1093/nar/26.22.5073. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/26.22.5073

121. Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wählby C, and Nilsson M. In situ sequencing for RNA analysis in preserved tissue and

cells. Nature Methods 2013 Sep; 10:857–60. DOI: 10.1038/nmeth.2563. Available from: https://www.nature.com/articles/nmeth.2563

122. Gyllborg D, Langseth CM, Qian X, Salas SM, Hilscher M, Lein E, and Nilsson M. Hybridization-based In Situ Sequencing (HybISS): spatial transcriptomic detection in human and mouse brain tissue. bioRxiv 2020 Feb :2020.02.03.931618. DOI: 10.1101/2020.02.03.931618. Available from: https://doi.org/10.1101/2020.02.03.931618

123. Sountoulidis A, Liontos A, Nguyen HP, Firsova AB, Fysikopoulos A, Qian X, Seeger W, Sundström E, Nilsson M, and Samakovlis C. SCRINSHOT, a spatial method for single-cell resolution mapping of cell states in tissue sections. bioRxiv 2020 Feb :2020.02.07.938571. DOI: 10.1101/2020.02.07.938571. Available from: https://www.biorxiv.org/content/biorxiv/early/2020/02/07/2020.02.07.938571.full.pdf

124. Lee H, Salas SM, Gyllborg D, and Nilsson M. Direct RNA targeted transcriptomic profiling in tissue using Hybridization-based RNA In Situ Sequencing (HybRISS). bioRxiv 2020 Jan :2020.12.02.408781. DOI: 10.1101/2020.12.02.408781. Available from: http://biorxiv.org/content/early/2020/12/02/2020.12.02.408781.abstract

125. Liu S, Punthambaker S, Iyer EPR, Ferrante T, Goodwin D, Fürth D, Pawlowski AC, Jindal K, Tam JM, Mifflin L, Alon S, Sinha A, Wassie AT, Chen F, Cheng A, Willocq V, Meyer K, Ling KH, Camplisson CK, Kohman RE, Aach J, Lee JH, Yankner BA, Boyden ES, and Church GM. Barcoded oligonucleotides ligated on RNA amplified for multiplexed and parallel in situ analyses. Nucleic Acids Research 2021 Jun; 49:e58–e58. DOI: 10.1093/nar/gkab120. Available from: https://doi.org/10.1093/nar/gkab120

126. Dirks RM and Pierce NA. From The Cover: Triggered amplification by hybridization chain reaction. Proceedings of the National Academy of Sciences 2004 Oct; 101:15275–8. DOI: 10.1073/pnas.0407024101. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0407024101

127. Kishi JY, Schaus TE, Gopalkrishnan N, Xuan F, and Yin P. Programmable autonomous synthesis of single-stranded DNA. Nature Chemistry 2018; 10:155–64. DOI: 10.1038/nchem.2872. Available from: https://doi.org/10.1038/nchem.2872

128. Kishi JY, Lapan SW, Beliveau BJ, West ER, Zhu A, Sasaki HM, Saka SK, Wang Y, Cepko CL, and Yin P. SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. Nature Methods 2019; 16:533–44. DOI: 10.1038/s41592-019-0404-0. Available from: https://doi.org/10.1038/s41592-019-0404-0

129. Coskun AF and Cai L. Dense transcript profiling in single cells by image correlation decoding. Nature Methods 2016 Jul; 13:657–60. DOI: 10.1038/

nmeth.3895. Available from: https://www.nature.com/articles/nmeth.3895

130. Chen S, Loper J, Chen X, Vaughan A, Zador AM, and Paninski L. BARcode DEmixing through Non-negative Spatial Regression (BarDensr). PLOS Computational Biology 2021 Mar; 17:e1008256. Available from: https://doi.org/10.1371/journal.pcbi.1008256

131. Andersson A, Diego F, Hamprecht FA, and Wählby C. ISTDECO: In Situ Transcriptomics Decoding by Deconvolution. bioRxiv 2021 Jan :2021.03.01.433040. DOI: 10.1101/2021.03.01.433040. Available from: http://biorxiv.org/content/early/2021/03/02/2021.03.01.433040.abstract

132. Cleary B, Simonton B, Bezney J, Murray E, Alam S, Sinha A, Habibi E, Marshall J, Lander ES, Chen F, and Regev A. Compressed sensing for highly efficient imaging transcriptomics. Nature Biotechnology 2021. DOI: 10.1038/s41587-021-00883-x. Available from: https://doi.org/10.1038/s41587-021-00883-x

133. Chen F, Tillberg PW, and Boyden ES. Expansion microscopy. Science 2015 Jan; 347:543–8. DOI: 10.1126/science.1260088. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1260088

134. Chen F, Wassie AT, Cote AJ, Sinha A, Alon S, Asano S, Daugharthy ER, Chang JB, Marblestone A, Church GM, Raj A, and Boyden ES. Nanoscale imaging of RNA with expansion microscopy. Nature Methods 2016 Jul; 13:679–84. DOI: 10.1038/nmeth.3899. Available from: https://www.nature.com/articles/nmeth.3899

135. Wang G, Moffitt JR, and Zhuang X. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. Scientific Reports 2018. DOI: 10.1038/s41598-018-22297-7

136. Dar D, Dar N, Cai L, and Newman DK. In situ single-cell activities of microbial populations revealed by spatial transcriptomics. bioRxiv 2021 Jan :2021.02.24.432792. DOI: 10.1101/2021.02.24.432792. Available from: http://biorxiv.org/content/early/2021/02/25/2021.02.24.432792.abstract

137. Vu T, Vallmitjana A, Gu J, La K, Xu Q, Flores J, Zimak J, Shiu J, Hosohama L, Wu J, Douglas C, Waterman M, Ganesan A, Hedde PN, Gratton E, and Zhao W. Spatial transcriptomics using combinatorial fluorescence spectral and lifetime encoding, imaging and analysis. bioRxiv 2021 Jan :2021.06.22.449468. DOI: 10.1101/2021.06.22.449468. Available from: http://biorxiv.org/content/early/2021/06/23/2021.06.22.449468.1.abstract

138. Choi J, Li J, Ferdous S, Liang Q, Moffitt JR, and Chen R. Spatial organization of the mouse retina at single cell resolution. bioRxiv 2022 Dec :2022.12.04.518972. DOI: 10.1101/2022.12.04.518972. Available

from: https://www.biorxiv.org/content/10.1101/2022.12.04.518972v1%20https://www.biorxiv.org/content/10.1101/2022.12.04.518972v1.abstract

139. Zhu Q, Shah S, Dries R, Cai L, and Yuan GC. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nature Biotechnology 2018 Dec; 36:1183–90. DOI: 10.1038/nbt.4260. Available from: https://www.nature.com/articles/nbt.4260

140. Foreman R and Wollman R. Mammalian gene expression variability is explained by underlying cell state. bioRxiv 2019. DOI: 10.1101/626424

141. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, Mollbrink A, Linnarsson S, Codeluppi S, Borg Å, Pontén F, Costea PI, Sahlén P, Mulder J, Bergmann O, Lundeberg J, and Frisén J. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 2016 Jul; 353:78–82. DOI: 10.1126/science.aaf2403. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aaf2403

142. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, and McCarroll SA. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015 May; 161:1202–14. DOI: 10.1016/j.cell.2015.05.002. Available from: http://dx.doi.org/10.1016/j.cell.2015.05.002

143. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, and Bielas JH. Massively parallel digital transcriptional profiling of single cells. Nature Communications 2017 Apr; 8:14049. DOI: 10.1038/ncomms14049. Available from: http://www.nature.com/articles/ncomms14049

144. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, and Kirschner MW. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell 2015 May; 161:1187–201. DOI: 10.1016/j.cell.2015.04.044. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867415005000

145. Hashimshony T, Senderovich N, Avital G, Klochendler A, Leeuw Y de, Anavy L, Gennert D, Li S, Livak KJ, Rozenblatt-Rosen O, Dor Y, Regev A, and Yanai I. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biology 2016 Dec; 17:77. DOI: 10.1186/s13059-016-

0938-8. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0938-8

146. Grün D, Kester L, and Oudenaarden A van. Validation of noise models for single-cell transcriptomics. Nature Methods 2014 Jun; 11:637–40. DOI: 10.1038/nmeth.2930. Available from: https://www.nature.com/articles/nmeth.2930

147. Samacoits A, Chouaib R, Safieddine A, Traboulsi AM, Ouyang W, Zimmer C, Peter M, Bertrand E, Walter T, and Mueller F. A computational framework to study sub-cellular RNA localization. Nature Communications 2018 Dec; 9:4584. DOI: 10.1038/s41467-018-06868-w. Available from: http://www.nature.com/articles/s41467-018-06868-w

148. Stoeger T, Battich N, Herrmann MD, Yakimovich Y, and Pelkmans L. Computer vision for image-based transcriptomics. Methods 2015 Sep; 85:44–53. DOI: 10.1016/j.ymeth.2015.05.016. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1046202315002091

149. Cabili MN, Dunagin MC, McClanahan PD, Biaesch A, Padovan-Merhar O, Regev A, Rinn JL, and Raj A. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. Genome Biology 2015 Dec; 16:20. DOI: 10.1186/s13059-015-0586-4. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0586-4

150. Melsted P, Booeshaghi AS, Liu L, Gao F, Lu L, Min KH(, Veiga Beltrame E da, Hjörleifsson KE, Gehring J, and Pachter L. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nature Biotechnology 2021. DOI: 10.1038/s41587-021-00870-2. Available from: https://doi.org/10.1038/s41587-021-00870-2

151. Perkel JM. Starfish enterprise: finding RNA patterns in single cells. Nature 2019 Aug; 572:549–51. DOI: 10.1038/D41586-019-02477-9

152. Zhou C, Huang R, Zhou X, and Xing D. Sensitive and specific microRNA detection by RNA dependent DNA ligation and rolling circle optical signal amplification. Talanta 2020; 216:120954. DOI: https://doi.org/10.1016/j.talanta.2020.120954. Available from: http://www.sciencedirect.com/science/article/pii/S0039914020302459

153. Chalfoun J, Majurski M, Blattner T, Bhadriraju K, Keyrouz W, Bajcsy P, and Brady M. MIST: Accurate and Scalable Microscopy Image Stitching Tool with Stage Modeling and Error Minimization. Scientific Reports 2017; 7:4988. DOI: 10.1038/s41598-017-04567-y. Available from: https://doi.org/10.1038/s41598-017-04567-y

154. Hörl D, Rojas Rusak F, Preusser F, Tillberg P, Randel N, Chhetri RK, Cardona A, Keller PJ, Harz H, Leonhardt H, Treier M, and Preibisch S. BigStitcher: reconstructing high-resolution image datasets of cleared and expanded samples.

Nature Methods 2019; 16:870–4. DOI: 10.1038/s41592-019-0501-0. Available from: https://doi.org/10.1038/s41592-019-0501-0

155. Bruggen D van, Pohl F, Langseth CM, Kukanja P, Lee H, Kabbe M, Meijer M, Hilscher MM, Nilsson M, Sundström E, and Castelo-Branco G. Developmental landscape of human forebrain at a single-cell level unveils early waves of oligodendrogenesis. bioRxiv 2021 Jan :2021.07.22.453317. DOI: 10.1101/2021.07.22.453317. Available from: http://biorxiv.org/content/early/2021/07/22/2021.07.22.453317.abstract

156. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2012 Oct; 29:15–21. DOI: 10.1093/bioinformatics/bts635. Available from: https://doi.org/10.1093/bioinformatics/bts635

157. Srivastava A, Malik L, Smith T, Sudbery I, and Patro R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. Genome Biology 2019 Dec; 20:65. DOI: 10.1186/s13059-019-1670-y. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1670-y

158. Bhaduri A, Sandoval-Espinosa C, Otero-Garcia M, Oh I, Yin R, Eze UC, Nowakowski TJ, and Kriegstein AR. An Atlas of Cortical Arealization Identifies Dynamic Molecular Signatures. bioRxiv 2021 Jan :2021.05.17.444528. DOI: 10.1101/2021.05.17.444528. Available from: http://biorxiv.org/content/early/2021/05/18/2021.05.17.444528.abstract

159. Zhou W, Yui MA, Williams BA, Yun J, Wold BJ, Cai L, and Rothenberg EV. Single-Cell Analysis Reveals Regulatory Gene Expression Dynamics Leading to Lineage Commitment in Early T Cell Development. Cell Systems 2019 Oct; 9:321–37. DOI: 10.1016/j.cels.2019.09.008. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2405471219303163

160. Shi H, Shi Q, Grodner B, Lenz JS, Zipfel WR, Brito IL, and De Vlaminck I. Highly multiplexed spatial mapping of microbial communities. Nature 2020; 588:676–81. DOI: 10.1038/s41586-020-2983-4. Available from: https://doi.org/10.1038/s41586-020-2983-4

161. Cao Z, Zuo W, Wang L, Chen J, Qu Z, Jin F, and Dai L. Spatial profiling of microbial communities by sequential FISH with error-robust encoding. bioRxiv 2021 Jan :2021.05.27.445923. DOI: 10.1101/2021.05.27.445923. Available from: http://biorxiv.org/content/early/2021/05/27/2021.05.27.445923.abstract

162. Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, Rubinstein ND, Hao J, Regev A, Dulac C, and Zhuang X. Molecular, spa-

tial, and functional single-cell profiling of the hypothalamic preoptic region. Science 2018. DOI: 10.1126/science.aau5324

163. Zhang M, Eichhorn SW, Zingg B, Yao Z, Zeng H, Dong H, and Zhuang X. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics. bioRxiv 2020 Jan :2020.06.04.105700. DOI: 10.1101/2020.06.04.105700. Available from: http://biorxiv.org/content/early/2020/06/05/2020.06.04.105700.abstract

164. Su J and Song Q. DSTG: Deconvoluting Spatial Transcriptomics Data through Graph-based Artificial Intelligence. bioRxiv 2020 Jan :2020.10.20.347195. DOI: 10.1101/2020.10.20.347195. Available from: http://biorxiv.org/content/early/2020/10/21/2020.10.20.347195.abstract

165. Liu M, Lu Y, Yang B, Chen Y, Radda JS, Hu M, Katz SG, and Wang S. Multiplexed imaging of nucleome architectures in single cells of mammalian tissue. Nature Communications 2020 Dec; 11:1–14. DOI: 10.1038/s41467-020-16732-5. Available from: https://doi.org/10.1038/s41467-020-16732-5

166. La Manno G, Siletti K, Furlan A, Gyllborg D, Vinsland E, Mossi Albiach A, Mattsson Langseth C, Khven I, Lederer AR, Dratva LM, Johnsson A, Nilsson M, Lönnerberg P, and Linnarsson S. Molecular architecture of the developing mouse brain. Nature 2021; 596:92–6. DOI: 10.1038/s41586-021-03775-x. Available from: https://doi.org/10.1038/s41586-021-03775-x

167. Langseth CM, Gyllborg D, Miller JA, Close JL, Long B, Lein ES, Hilscher MM, and Nilsson M. Comprehensive in situ mapping of human cortical transcriptomic cell types. Communications Biology 2021; 4:998. DOI: 10.1038/s42003-021-02517-z. Available from: https://doi.org/10.1038/s42003-021-02517-z

168. Landegren U, Kaiser R, Sanders J, and Hood L. A ligase-mediated gene detection technique. Science 1988 Aug; 241:1077 LP –1080. DOI: 10.1126/science.3413476. Available from: http://science.sciencemag.org/content/241/4869/1077.abstract

169. Tecott LH, Barchas JD, and Eberwine JH. In situ transcription: specific synthesis of complementary DNA in fixed tissue sections. Science 1988 Jun; 240:1661 LP –1664. DOI: 10.1126/science.2454508. Available from: http://science.sciencemag.org/content/240/4859/1661.abstract

170. Macevicz S. DNA sequencing by parallel olgonucleotide extensions. 1995. Available from: https://patents.google.com/patent/US5969119A/en

171. Alsup WH. Applera v. Illumina. Tech. rep. 2009. Available from: https://www.leagle.com/decision/infco20100325236

172. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, and Church GM. Highly Multiplexed Subcellular RNA Sequencing in Situ. Science 2014 Mar; 343:1360 LP –1363. DOI: 10.1126/science.1250212. Available from: http://science.sciencemag.org/content/343/6177/1360.abstract

173. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, Turczyk BM, Yang JL, Lee HS, Aach J, Zhang K, and Church GM. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. Nature Protocols 2015 Mar; 10:442–58. DOI: 10.1038/nprot.2014.191. Available from: http://www.nature.com/articles/nprot.2014.191

174. Alon S, Goodwin DR, Sinha A, Wassie AT, Chen F, Daugharthy ER, Bando Y, Kajita A, Xue AG, Marrett K, Prior R, Cui Y, Payne AC, Yao CC, Suk HJ, Wang R, Yu CC, Tillberg P, Reginato P, Pak N, Liu S, Punthambaker S, Iyer EP, Kohman RE, Miller JA, Lein ES, Lako A, Cullen N, Rodig S, Helvie K, Abravanel DL, Wagle N, Johnson BE, Klughammer J, Slyper M, Waldman J, Jané-Valbuena J, Rozenblatt-Rosen O, Regev A, Church GM, Marblestone AH, and Boyden ES. Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. Science 2021 Jan; 371. DOI: 10.1126/SCIENCE.AAX2656/SUPPL{_}FILE/AAX2656{_}TABLESS1-S6ANDS9-S14.XLSX. Available from: https://www.science.org/doi/10.1126/science.aax2656

175. Giani AM, Gallo GR, Gianfranceschi L, and Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. Computational and Structural Biotechnology Journal 2020 Jan; 18:9–19. DOI: 10.1016/j.csbj.2019.11.002. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2001037019303277

176. Lein E, Borm LE, and Linnarsson S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. 2017. DOI: 10.1126/science.aan6827

177. Fürth D, Hatini V, and Lee JH. In Situ Transcriptome Accessibility Sequencing (INSTA-seq). bioRxiv 2019 Jan :722819. DOI: 10.1101/722819. Available from: http://biorxiv.org/content/early/2019/08/06/722819.abstract

178. Shendure J. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. Science 2005 Sep; 309:1728–32. DOI: 10.1126/science.1117389. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1117389

179. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, Dahl F, Fernandez A, Staker B, Pant KP, Baccash J, Borcherding AP, Brownley A, Cedeno R, Chen L, Chernikoff D, Cheung A, Chirita R, Curson B, Ebert JC, Hacker CR, Hartlage R, Hauser B, Huang S, Jiang Y, Karpinchyk V, Koenig M, Kong C, Landers T, Le C, Liu J, McBride CE, Morenzoni M, Morey RE, Mutch K, Perazich H, Perry K, Peters BA, Peterson J, Pethiyagoda CL, Pothuraju K, Richter C, Rosenbaum AM, Roy S, Shafto J, Sharanhovich U, Shannon KW, Sheppy CG, Sun M, Thakuria JV, Tran A, Vu D, Zaranek AW, Wu X, Drmanac S, Oliphant AR, Banyai WC, Martin B, Ballinger DG, Church GM, and Reid CA. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science 2010 Jan; 327:78–81. DOI: 10.1126/science.1181498. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1181498

180. Chen WT, Lu A, Craessaerts K, Pavie B, Sala Frigerio C, Corthout N, Qian X, Laláková J, Kühnemund M, Voytyuk I, Wolfs L, Mancuso R, Salta E, Balusu S, Snellinx A, Munck S, Jurek A, Fernandez Navarro J, Saido TC, Huitinga I, Lundeberg J, Fiers M, and De Strooper B. Spatial Transcriptomics and In Situ Sequencing to Study Alzheimer's Disease. Cell 2020 Jul; 0. DOI: 10.1016/j.cell.2020.06.038. Available from: http://www.cell.com/article/S0092867420308151/fulltext

181. Qian X, Harris KD, Hauling T, Nicoloutsopoulos D, Muñoz-Manchado AB, Skene N, Hjerling-Leffler J, and Nilsson M. Probabilistic cell typing enables fine mapping of closely related cell types in situ. Nature Methods 2020 Jan; 17:101–6. DOI: 10.1038/s41592-019-0631-4. Available from: https://doi.org/10.1038/s41592-019-0631-4

182. Partel G, Hilscher M, Milli G, Solorzano L, Klemm A, Nilsson M, and Wählby C. Identification of spatial compartments in tissue from in situ sequencing data. bioRxiv 2019 Sep :765842. DOI: 10.1101/765842. Available from: https://doi.org/10.1101/765842

183. Asp M, Giacomello S, Larsson L, Wu C, Fürth D, Qian X, Wärdell E, Custodio J, Reimegård J, Salmén F, Österholm C, Ståhl PL, Sundström E, Åkesson E, Bergmann O, Bienko M, Månsson-Broberg A, Nilsson M, Sylvén C, and Lundeberg J. A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. Cell 2019 Dec; 179:1647–60. DOI: 10.1016/j.cell.2019.11.025. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867419312826

184. Lebrigand K, Bergenstråhle J, Thrane K, Mollbrink A, Barbry P, Waldmann R, and Lundeberg J. The spatial landscape of gene expression isoforms in tissue sections. bioRxiv 2020 Aug :2020.08.24.252296. DOI: 10.1101/2020.08.24.252296. Available from: https://doi.org/10.1101/2020.08.24.252296

185. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, and Law M. Comparison of Next-Generation Sequencing Systems. Journal of Biomedicine and Biotechnology 2012; 2012. Ed. by Oefner PJ:251364. DOI: 10.1155/2012/251364. Available from: https://doi.org/10.1155/2012/251364

186. Schwaber J, Andersen S, and Nielsen L. Shedding light: The importance of reverse transcription efficiency standards in data interpretation. Biomolecular Detection and Quantification 2019 Mar; 17:100077. DOI: 10.1016/j.bdq.2018.12.002. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2214753517302188

187. Bustin S, Dhillon HS, Kirvell S, Greenwood C, Parker M, Shipley GL, and Nolan T. Variability of the Reverse Transcription Step: Practical Implications. Clinical Chemistry 2015 Jan; 61:202–12. DOI: 10.1373/clinchem.2014.230615. Available from: https://academic.oup.com/clinchem/article/61/1/202/5611431

188. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, Nolan GP, Bava FA, and Deisseroth K. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science 2018 Jul; 361:eaat5691. DOI: 10.1126/science.aat5691. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aat5691

189. Kohman RE and Church GM. Fluorescent in situ sequencing of DNA barcoded antibodies. bioRxiv 2020 Apr; 20:2020.04.27.060624. DOI: 10.1101/2020.04.27.060624. Available from: https://doi.org/10.1101/2020.04.27.060624

190. Chen X, Sun YC, Zhan H, Kebschull JM, Fischer S, Matho K, Huang ZJ, Gillis J, and Zador AM. High-Throughput Mapping of Long-Range Neuronal Projection Using In Situ Sequencing. Cell 2019 Oct; 179:772–86. DOI: 10.1016/j.cell.2019.09.023. Available from: https://doi.org/10.1016/j.cell.2019.09.023

191. Sun YC, Chen X, Fischer S, Lu S, Gillis J, and Zador AM. Integrating barcoded neuroanatomy with spatial transcriptional profiling reveals cadherin correlates of projections shared across the cortex. bioRxiv 2020 Jan :2020.08.25.266460. DOI: 10.1101/2020.08.25.266460. Available from: http://biorxiv.org/content/early/2020/08/26/2020.08.25.266460.abstract

192. Jemt A, Salmén F, Lundmark A, Mollbrink A, Fernández Navarro J, Ståhl PL, Yucel-Lindberg T, and Lundeberg J. An automated approach to prepare tissue-derived spatially barcoded RNA-sequencing libraries. Scientific Reports 2016. DOI: 10.1038/srep37137

193. Navarro JF, Sjöstrand J, Salmén F, Lundeberg J, and Ståhl PL. ST Pipeline: an automated pipeline for spatial mapping of unique transcripts. Bioinformatics (Oxford, England) 2017. DOI: 10.1093/bioinformatics/btx211

194. Asp M, Salmén F, Ståhl PL, Vickovic S, Felldin U, Löfling M, Fernandez Navarro J, Maaskola J, Eriksson MJ, Persson B, Corbascio M, Persson H, Linde C, and Lundeberg J. Spatial detection of fetal marker genes expressed at low level in adult human heart tissue. Scientific Reports 2017; 7:12941. DOI: 10.1038/s41598-017-13462-5. Available from: https://doi.org/10.1038/s41598-017-13462-5

195. Lundmark A, Gerasimcik N, Båge T, Jemt A, Mollbrink A, Salmén F, Lundeberg J, and Yucel-Lindberg T. Gene expression profiling of periodontitis-affected gingival tissue by spatial transcriptomics. Scientific Reports 2018; 8:9370. DOI: 10.1038/s41598-018-27627-3. Available from: https://doi.org/10.1038/s41598-018-27627-3

196. Berglund E, Maaskola J, Schultz N, Friedrich S, Marklund M, Bergenstråhle J, Tarish F, Tanoglidi A, Vickovic S, Larsson L, Salmén F, Ogris C, Wallenborg K, Lagergren J, Ståhl P, Sonnhammer E, Helleday T, and Lundeberg J. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nature Communications 2018. DOI: 10.1038/s41467-018-04724-5

197. He B, Bergenstråhle L, Stenbeck L, Abid A, Andersson A, Borg Å, Maaskola J, Lundeberg J, and Zou J. Integrating spatial gene expression and breast tumour morphology via deep learning. Nature Biomedical Engineering 2020 Jun :1–8. DOI: 10.1038/s41551-020-0578-x. Available from: https://doi.org/10.1038/s41551-020-0578-x

198. Carlberg K, Korotkova M, Larsson L, Catrina AI, Ståhl PL, and Malmström V. Exploring inflammatory signatures in arthritic joint biopsies with Spatial Transcriptomics. Scientific Reports 2019; 9:18975. DOI: 10.1038/s41598-019-55441-y. Available from: https://doi.org/10.1038/s41598-019-55441-y

199. Thrane K, Eriksson H, Maaskola J, Hansson J, and Lundeberg J. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. Cancer Research 2018. DOI: 10.1158/0008-5472.CAN-18-0747

200. Maniatis S, Äijö T, Vickovic S, Braine C, Kang K, Mollbrink A, Fagegaltier D, Andrusivová Ž, Saarenpää S, Saiz-Castro G, Cuevas M, Watters A, Lundeberg J, Bonneau R, and Phatnani H. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. Science 2019 Apr; 364:89 LP –93. DOI: 10.1126/science.aav9776. Available from: http://science.sciencemag.org/content/364/6435/89.abstract

201. Gregory JM, McDade K, Livesey MR, Croy I, Marion de Proce S, Aitman T, Chandran S, and Smith C. Spatial transcriptomics identifies spatially dysregulated expression of GRM3 and USP47 in amyotrophic lateral sclerosis. Neuropathology and Applied Neurobiology 2020 Aug; 46:441–57. DOI: 10.1111/nan.12597. Available from: https://doi.org/10.1111/nan.12597

202. Ji AL, Rubin AJ, Thrane K, Jiang S, Reynolds DL, Meyers RM, Guo MG, George BM, Mollbrink A, Bergenstråhle J, Larsson L, Bai Y, Zhu B, Bhaduri A, Meyers JM, Rovira-Clavé X, Hollmig ST, Aasi SZ, Nolan GP, Lundeberg J, and Khavari PA. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell 2020; 182:497–514. DOI: https://doi.org/10.1016/j.cell.2020.05.039. Available from: http://www.sciencedirect.com/science/article/pii/S0092867420306723

203. Ortiz C, Navarro JF, Jurek A, Märtin A, Lundeberg J, and Meletis K. Molecular atlas of the adult mouse brain. Science Advances 2020 Jun; 6:eabb3446. DOI: 10.1126/sciadv.abb3446. Available from: www.brain-map.org%20https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.abb3446

204. Liu Y, Yang M, Deng Y, Su G, Enninful A, Guo CC, Tebaldi T, Zhang D, Kim D, Bai Z, Norris E, Pan A, Li J, Xiao Y, Halene S, and Fan R. High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. Cell 2020; 183:1665–81. DOI: https://doi.org/10.1016/j.cell.2020.10.026. Available from: https://www.sciencedirect.com/science/article/pii/S0092867420313908

205. 10X VISIUM SPATIAL TRANSCRIPTOMICS. Available from: https://biotech.illinois.edu/htdna/applications/10x-visium-spatial-transcriptomics

206. ADVANCED GENOMICS CORE PRICING. Available from: https://brcf.medicine.umich.edu/cores/advanced-genomics/price-list/

207. Joglekar A, Prjibelski A, Mahfouz A, Collier P, Lin S, Schlusche AK, Marrocco J, Williams SR, Haase B, Hayes A, Chew JG, Weisenfeld NI, Wong MY, Stein AN, Hardwick S, Hunt T, Bent Z, Fedrigo O, Sloan SA, Risso D, Jarvis ED, Flicek P, Luo W, Pitt GS, Frankish A, Smit AB, Ross ME, and Tilgner HU. Cell-type, single-cell, and spatial signatures of brain-region specific splicing in postnatal development. bioRxiv 2020 Jan :2020.08.27.268730. DOI: 10.1101/2020.08.27.268730. Available from: http://biorxiv.org/content/early/2020/08/27/2020.08.27.268730.abstract

208. McCray T, Pacheco JV, Loitz CC, Garcia J, Baumann B, Schlicht MJ, Valyi-Nagy K, Abern MR, and Nonn L. Vitamin D sufficiency enhances differentiation of patient-derived prostate epithelial organoids. iScience 2021;

24:101974. DOI: https://doi.org/10.1016/j.isci.2020.101974. Available from: https://www.sciencedirect.com/science/article/pii/S2589004220311718

209. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, and Macosko EZ. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science 2019. DOI: 10.1126/science.aaw1219

210. Stickels R, Murray E, Kumar P, Li J, Marshall J, Di Bella D, Arlotta P, Macosko E, and Chen F. Sensitive spatial genome wide expression profiling at cellular resolution. bioRxiv 2020 Mar :2020.03.12.989806. DOI: 10.1101/2020.03.12.989806. Available from: https://doi.org/10.1101/2020.03.12.989806

211. Vickovic S, Eraslan G, Salmén F, Klughammer J, Stenbeck L, Schapiro D, Äijö T, Bonneau R, Bergenstråhle L, Navarro JF, Gould J, Griffin GK, Borg Å, Ronaghi M, Frisén J, Lundeberg J, Regev A, and Ståhl PL. High-definition spatial transcriptomics for in situ tissue profiling. Nature Methods 2019. DOI: 10.1038/s41592-019-0548-y

212. Liu Y, Enninful A, Deng Y, and Fan R. Spatial transcriptome sequencing of FFPE tissues at cellular level. bioRxiv 2020 Jan :2020.10.13.338475. DOI: 10.1101/2020.10.13.338475. Available from: http://biorxiv.org/content/early/2020/10/19/2020.10.13.338475.abstract

213. Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, Yang J, Li W, Xu J, Hao S, Lu H, Chen X, Liu X, Huang X, Lin F, Tang X, Li Z, Hong Y, Fu D, Jiang Y, Peng J, Liu S, Shen M, Liu C, Li Q, Wang Z, Wang Z, Huang X, Yuan Y, Volpe G, Ward C, Muñoz-Cánoves P, Thiery JP, Zhao F, Li M, Kuang H, Wang O, Lu H, Wang B, Ni M, Zhang W, Mu F, Yin Y, Yang H, Lisby M, Cornall RJ, Uhlen M, Esteban MA, Li Y, Liu L, Xu X, and Wang J. Large field of view-spatially resolved transcriptomics at nanoscale resolution. bioRxiv 2021 Jan :2021.01.17.427004. DOI: 10.1101/2021.01.17.427004. Available from: http://biorxiv.org/content/early/2021/01/24/2021.01.17.427004.abstract

214. Cho CS, Xi J, Park SR, Hsu JE, Kim M, Jun G, Kang HM, and Lee JH. Seq-Scope: Submicrometer-resolution spatial transcriptomics for single cell and subcellular studies. bioRxiv 2021 Jan :2021.01.25.427807. DOI: 10.1101/2021.01.25.427807. Available from: http://biorxiv.org/content/early/2021/01/27/2021.01.25.427807.abstract

215. Fu X, Sun L, Chen JY, Dong R, Lin Y, Palmiter RD, Lin S, and Gu L. Continuous Polony Gels for Tissue Mapping with High Resolution and RNA Capture Efficiency. bioRxiv 2021 Jan :2021.03.17.435795. DOI: 10.1101/2021.03.17.435795. Available from: http://biorxiv.org/content/early/2021/03/17/2021.03.17.435795.abstract

216. Nagarajan MB, Tentori AM, Zhang WC, Slack FJ, and Doyle PS. Spatially resolved and multiplexed MicroRNA quantification from tissue using nanoliter well arrays. Microsystems & Nanoengineering 2020; 6:51. DOI: 10.1038/s41378-020-0169-8. Available from: https://doi.org/10.1038/s41378-020-0169-8

217. Andersson A, Bergenstråhle J, Asp M, Bergenstråhle L, Jurek A, Fernández Navarro J, and Lundeberg J. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. Communications Biology 2020; 3:565. DOI: 10.1038/s42003-020-01247-y. Available from: https://doi.org/10.1038/s42003-020-01247-y

218. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, and Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nature Methods 2015; 12:453–7. DOI: 10.1038/nmeth.3337. Available from: https://doi.org/10.1038/nmeth.3337

219. Lee Y, Bogdanoff D, Wang Y, Hartoularos GC, Woo JM, T. MC, M. NH, S. LD, Yang S, James L, Sadaf M, Joshua C, Eric S, N. ND, L. RT, S. SY, Alexander M, D. CE, and Jimmie YC. XYZeq: Spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment. Science Advances 2021 Sep; 7:eabg4755. DOI: 10.1126/sciadv.abg4755. Available from: https://doi.org/10.1126/sciadv.abg4755

220. Srivatsan SR, Regier MC, Barkan E, Franks JM, Packer JS, Grosjean P, Duran M, Saxton S, Ladd JJ, Spielmann M, Lois C, Lampe PD, Shendure J, Stevens KR, and Trapnell C. Embryo-scale, single-cell spatial transcriptomics. Science 2021; 373:111–7. DOI: 10.1126/science.abb9536

221. Boulgakov AA, Ellington AD, and Marcotte EM. Bringing Microscopy-By-Sequencing into View. Trends in Biotechnology 2020; 38:154–62. DOI: https://doi.org/10.1016/j.tibtech.2019.06.001. Available from: http://www.sciencedirect.com/science/article/pii/S0167779919301349

222. Glaser JI, Zamft BM, Church GM, and Kording KP. Puzzle Imaging: Using Large-Scale Dimensionality Reduction Algorithms for Localization. PLOS ONE 2015 Jul; 10. Ed. by Najbauer J:e0131593. DOI: 10.1371/journal.pone.0131593. Available from: https://dx.plos.org/10.1371/journal.pone.0131593

223. Weinstein JA, Regev A, and Zhang F. DNA Microscopy: Optics-free Spatiogenetic Imaging by a Stand-Alone Chemical Reaction. Cell 2019 Jun; 178:229–41. DOI: 10.1016/j.cell.2019.05.019. Available from: https://doi.org/10.1016/j.cell.2019.05.019

224. Hoffecker IT, Yang Y, Bernardinelli G, Orponen P, and Högberg B. A computational framework for DNA sequencing microscopy. Proceedings of the National Academy of Sciences of the United States of America 2019 Sep;

116:19282–7. DOI: 10.1073/pnas.1821178116. Available from: https://www.pnas.org/content/116/39/19282.abstract

225. Kaewsapsak P, Shechner DM, Mallard W, Rinn JL, and Ting AY. Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. eLife 2017 Dec; 6. DOI: 10.7554/eLife.29224. Available from: https://elifesciences.org/articles/29224

226. Fazal FM, Han S, Parker KR, Kaewsapsak P, Xu J, Boettiger AN, Chang HY, and Ting AY. Atlas of Subcellular RNA Localization Revealed by APEX-Seq. Cell 2019 Jul; 178:473–90. DOI: 10.1016/j.cell.2019.05.027. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867419305550

227. Halpern KB, Shenhav R, Massalha H, Toth B, Egozi A, Massasa EE, Medgalia C, David E, Giladi A, Moor AE, Porat Z, Amit I, and Itzkovitz S. Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. Nature Biotechnology 2018 Nov; 36:962. DOI: 10.1038/nbt.4231. Available from: https://www.nature.com/articles/nbt.4231

228. Manco R, Averbukh I, Porat Z, Halpern KB, Amit I, and Itzkovitz S. Clump sequencing exposes the spatial expression programs of intestinal secretory cells. bioRxiv 2020 Aug :2020.08.05.237917. DOI: 10.1101/2020.08.05.237917. Available from: https://www.biorxiv.org/content/10.1101/2020.08.05.237917v1

229. Boisset JC, Vivié J, Grün D, Muraro MJ, Lyubimova A, and Oudenaarden A van. Mapping the physical network of cellular interactions. Nature Methods 2018; 15:547–53. DOI: 10.1038/s41592-018-0009-z. Available from: https://doi.org/10.1038/s41592-018-0009-z

230. Halpern KB, Shenhav R, Matcovitch-Natan O, Tóth B, Lemze D, Golan M, Massasa EE, Baydatch S, Landen S, Moor AE, Brandis A, Giladi A, Stokar-Avihail A, David E, Amit I, and Itzkovitz S. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. Nature 2017 Feb; 542:1–5. DOI: 10.1038/nature21065. Available from: https://www.nature.com/articles/nature21065

231. Vickovic S, Lötstedt B, Klughammer J, Segerstolpe Å, Rozenblatt-Rosen O, and Regev A. SM-Omics: An automated platform for high-throughput spatial multi-omics. bioRxiv 2020 Jan :2020.10.14.338418. DOI: 10.1101/2020.10.14.338418. Available from: http://biorxiv.org/content/early/2020/10/15/2020.10.14.338418.abstract

232. Guilliams M, Bonnardel J, Haest B, Vanderborght B, Wagner C, Remmerie A, Bujko A, Martens L, Thoné T, Browaeys R, De Ponti FF, Vanneste B, Zwicker C, Svedberg FR, Vanhalewyn T, Gonçalves A, Lippens S, Devriendt B, Cox E, Ferrero G, Wittamer V, Willaert A, Kaptein SJ, Neyts J, Dallmeier K, Geldhof P, Casaert S, Deplancke B, Dijke P ten, Hoorens A, Vanlander A, Berrevoet

F, Van Nieuwenhove Y, Saeys Y, Saelens W, Van Vlierberghe H, Devisscher L, and Scott CL. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. Cell 2022 Jan; 185:379–96. DOI: 10.1016/J.CELL.2021.12.018

233. Xu X, Holmes TC, Luo MH, Beier KT, Horwitz GD, Zhao F, Zeng W, Hui M, Semler BL, and Sandri-Goldin RM. Viral Vectors for Neural Circuit Mapping and Recent Advances in Trans-synaptic Anterograde Tracers. Neuron 2020; 107:1029–47. DOI: https://doi.org/10.1016/j.neuron.2020.07.010. Available from: http://www.sciencedirect.com/science/article/pii/S0896627320305274

234. Kim MH, Radaelli C, Thomsen ER, Mahoney JT, Long B, Taormina MJ, Kebede S, Gamlin C, Sorensen SA, Campagnola L, Dee N, D'Orazi F, Ko AL, Ojemann JG, Silbergeld DL, Gwinn RP, Cobbs C, Keene CD, Jarsky T, Murphy G, Zeng H, Nicovich PR, Ting JT, Levi BP, and Lein ES. Molecular and genetic approaches for assaying human cell type synaptic connectivity. bioRxiv 2020 Jan :2020.10.16.343343. DOI: 10.1101/2020.10.16.343343. Available from: http://biorxiv.org/content/early/2020/10/17/2020.10.16.343343.abstract

235. Li Q, Lin Z, Liu R, Tang X, Huang J, He Y, Zhou H, Sheng H, Shi H, Wang X, and Liu J. In situ electro-sequencing in three-dimensional tissues. bioRxiv 2021 Jan :2021.04.22.440941. DOI: 10.1101/2021.04.22.440941. Available from: http://biorxiv.org/content/early/2021/04/23/2021.04.22.440941.abstract

236. Fan Z, Chen R, and Chen X. SpatialDB: a database for spatially resolved transcriptomes. Nucleic Acids Research 2020 Jan; 48:D233–D237. DOI: 10.1093/nar/gkz934. Available from: https://doi.org/10.1093/nar/gkz934

237. Svensson V, Teichmann SA, and Stegle O. SpatialDE: Identification of spatially variable genes. Nature Methods 2018 Apr; 15:343–6. DOI: 10.1038/nmeth.4636. Available from: https://www.nature.com/articles/nmeth.4636

238. Edsgärd D, Johnsson P, and Sandberg R. Identification of spatial expression trends in single-cell gene expression data. Nature Methods 2018 May; 15:339–42. DOI: 10.1038/nmeth.4634. Available from: http://www.nature.com/articles/nmeth.4634

239. Callaway EM et al. A multimodal cell census and atlas of the mammalian primary motor cortex. Nature 2021 598:7879 2021 Oct; 598:86–102. DOI: 10.1038/s41586-021-03950-0. Available from: https://www.nature.com/articles/s41586-021-03950-0

240. Kumar S, Konikoff C, Sanderford M, Liu L, Newfeld S, Ye J, and Kulathinal RJ. FlyExpress 7: An Integrated Discovery Platform To Study Co-expressed Genes Using in situ Hybridization Images in Drosophila. G3; Genes|Genomes|Genetics 2017 Aug; 7:2791–7. DOI: 10.1534/g3.117.040345. Available from: http://g3journal.org/lookup/doi/10.1534/g3.117.040345

Chapter 8

TEXT MINING LCM TRANSCRIPTOMICS ABSTRACTS

This analysis was performed in 2021 and has not been updated. If it's run again in 2023, the results might be different as spatial transcriptomics has evolved in the past two years.

To analyze trends in LCM followed by microarray or RNA-seq, abstracts were downloaded from the PubMed API, with search term `"((laser capture microdissection) OR (laser microdissection)) AND ((microarray) OR (transcriptome) OR (RNA-seq))"`. For preprints, abstracts from the search term "laser microdissection" were downloaded from bioRxiv. Because bioRiv's advanced search does not acknowledge parentheses, a more complicated search term was not used. Upon random inspection, the retrieved abstracts mostly seem relevant. The number of LCM transcriptomics search results dwarfs the number of publications for other methods of spatial transcriptomics and seems to show two peaks, one around 2012, and the other in 2020 and 2021 (Figure 8.1); the LCM corpus contains 2252 abstracts as of March 26, 2021, while there are between 500 and 600 papers in the curated database.

LCM transcriptomics is also more geographically diffuse and spread out into many less well-known institutions and some developing countries, though some elite institutions are among the top contributors, such as Harvard Medical School and Massachusetts General Hospital (Boston), Columbia University, NYU, Rockefeller, and Sloan-Kettering (New York), NIH (Bethesda), and Cambridge University (Cambridge, UK) (Figure 8.2).

After identifying common and relevant phrases in the abstracts, the abstracts were tokenized into unigrams. We used the `stm` R package [1] to identify topics. The cities in which the research was conducted, date published or posted on bioRxiv (linear, not transformed), and journal (including bioRxiv) were used as covariates for topic prevalence, because labs and journals may have preferred topics and city is a proxy to institution, and it's reasonable to assume that prevalence of at least some topic changes through time, such as due to evolution of technology. Cities and journals with fewer than 5 papers were lumped into "Other". From a trade-off between held out likelihood and residual, and between topic exclusivity and semantic

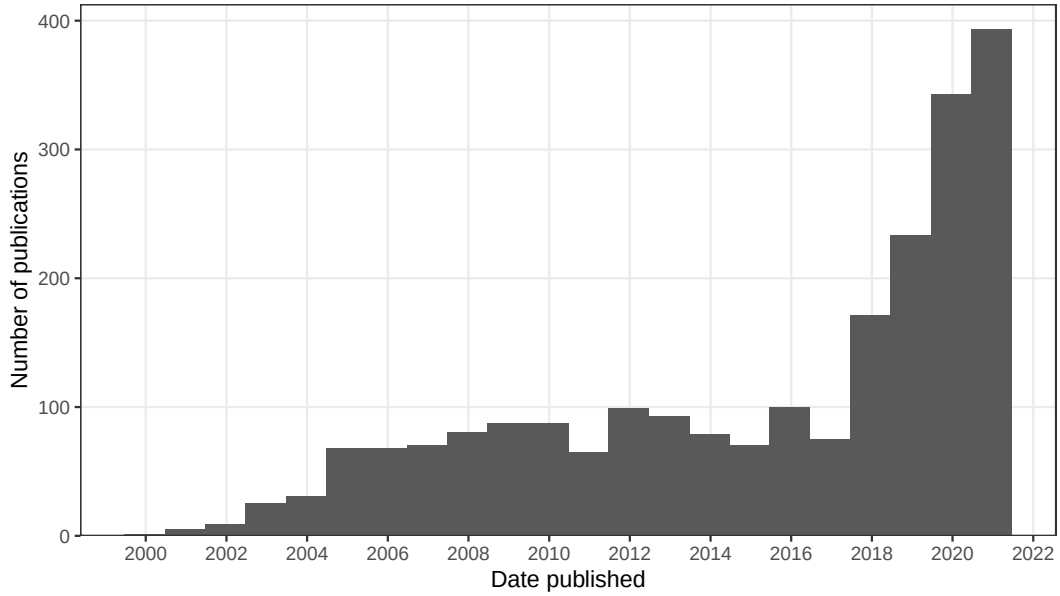

Figure 8.1: Number of publications in LCM transcriptomics PubMed search results over time. Bin width is 365 days.

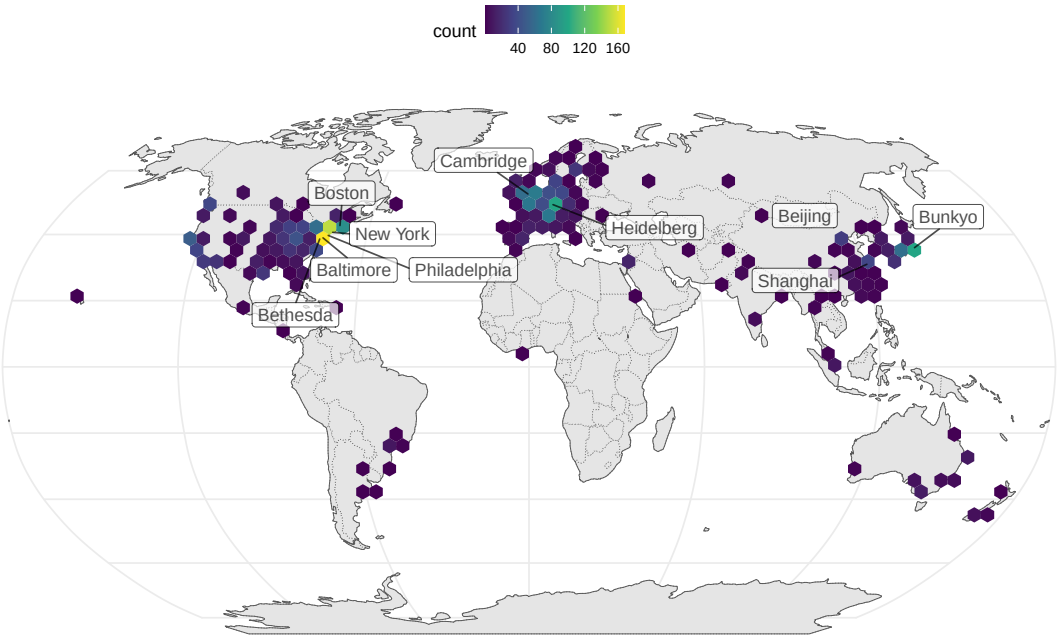

Figure 8.2: Geographic distribution of LCM transcriptomics research, with top 10 cities labeled. Number of publications is binned over longitude and latitude.

coherence, we chose 50 topics. Code used to find this can be found here.

Here `stm` stands for structural text mining. A generative model of word counts is fitted with the word counts in each abstract as well as abstract level covariates, here date, city, and journal. Among parameters of the model estimated are the proportion of each topic in each abstract after accounting for covariates (θ), topic proportions in the corpus (γ), and probability of getting each word from each topic (β). See the `stm` vignette for more details. `stm` can not only detect topics without having a human read all the abstracts, but also find how covariates relate to topic prevalence.

8.1 Topic modeling

As already mentioned, microarray was first demonstrated on LCM samples in 1999, profiling 477 cDNAs from rat neurons [2]. Since then, LCM transcriptomics has been used on many research topics, such as various aspects of cancer (topics 5, 6, 8, 10, 11, 13, 16, 20, 24, 27, 34, 44, 50), botany (topics 9, 15, 21, 40, 43, 45), developmental biology (topics 1, 3, 17, 18, 29, 35, 39), neuroscience (topics 7, 14, 19, 23, 25, 32, 33, 36, 47), immunology (topics 12, 22, 48), miRNA (topic 5), and technical issues related to LCM (topics 4, 28, 37, 41) (Figure 8.3).

In most cases, the top 5 words in each topic give us a decent idea what the topic is about. We can also plot the probability to get top words (β) in each topic.

While in most cases, the topic is apparent from the top words, some topics are less apparent (e.g. topic 49). From the top words and quick glances of abstracts with the highest proportion of each topic, the 50 topics are summarized here in more human readable terms:

1. Stem cell and fetal development
2. GWAS, genetic screens, and genetics of complex phenotypes
3. Biomechanics, ECM, eye lens, muscles, and morphogenesis
4. Data analysis, especially of RNA-seq, but also of 3D genome structure and microarray
5. miRNAs in cancer
6. Quantitative analyses of cancer, clinical and bioinformatic
7. Hippocampus and Alzheimer's disease, sometimes related to Down syndrome
8. Prostate cancer and other stuff in molecular biology and biochemistry, probably because some prostate cancer papers have an emphasis on molecular biology

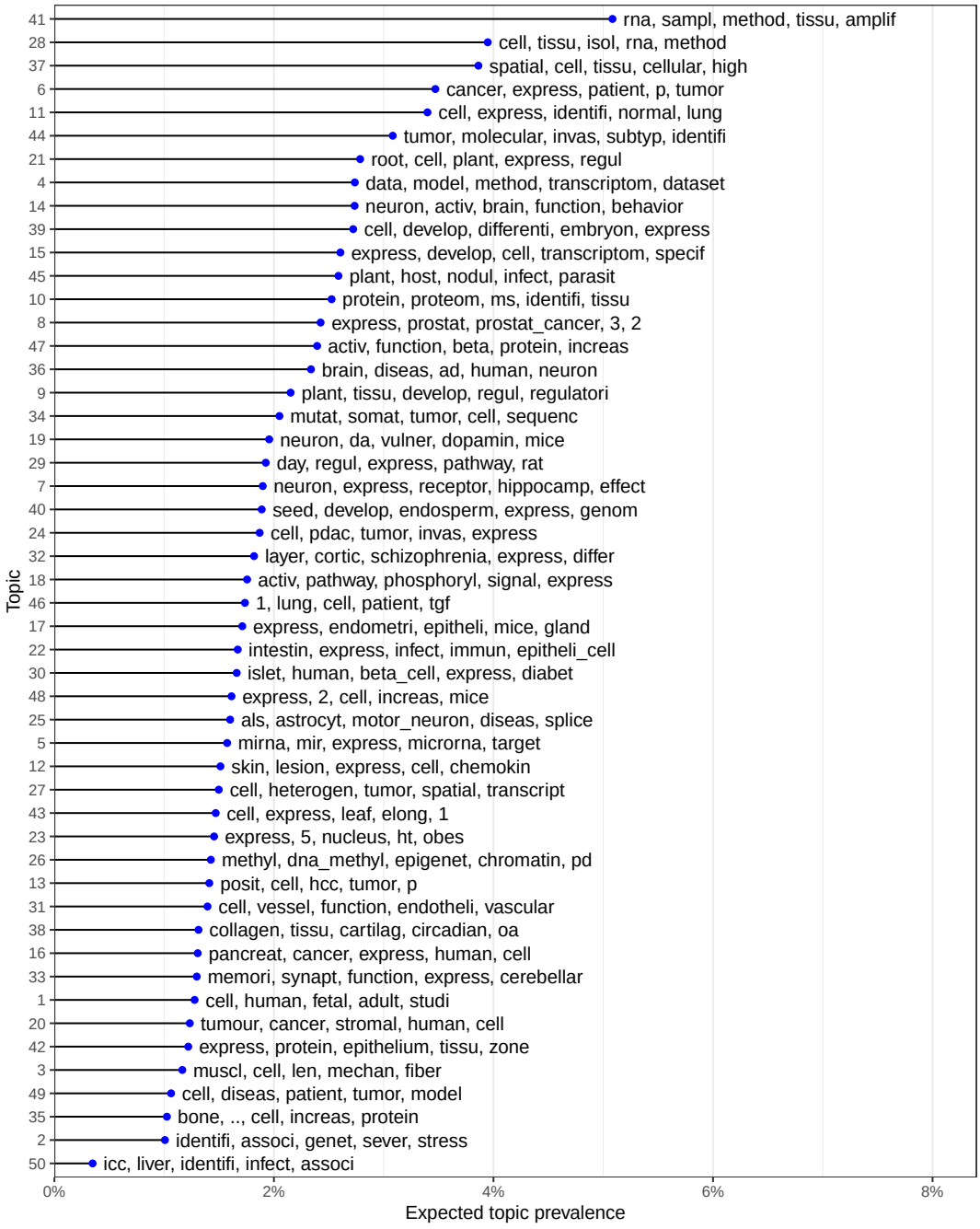

Figure 8.3: Top words for each of the 50 topics, ordered by expected topic prevalence and showing top 5 words contributing to each topic.

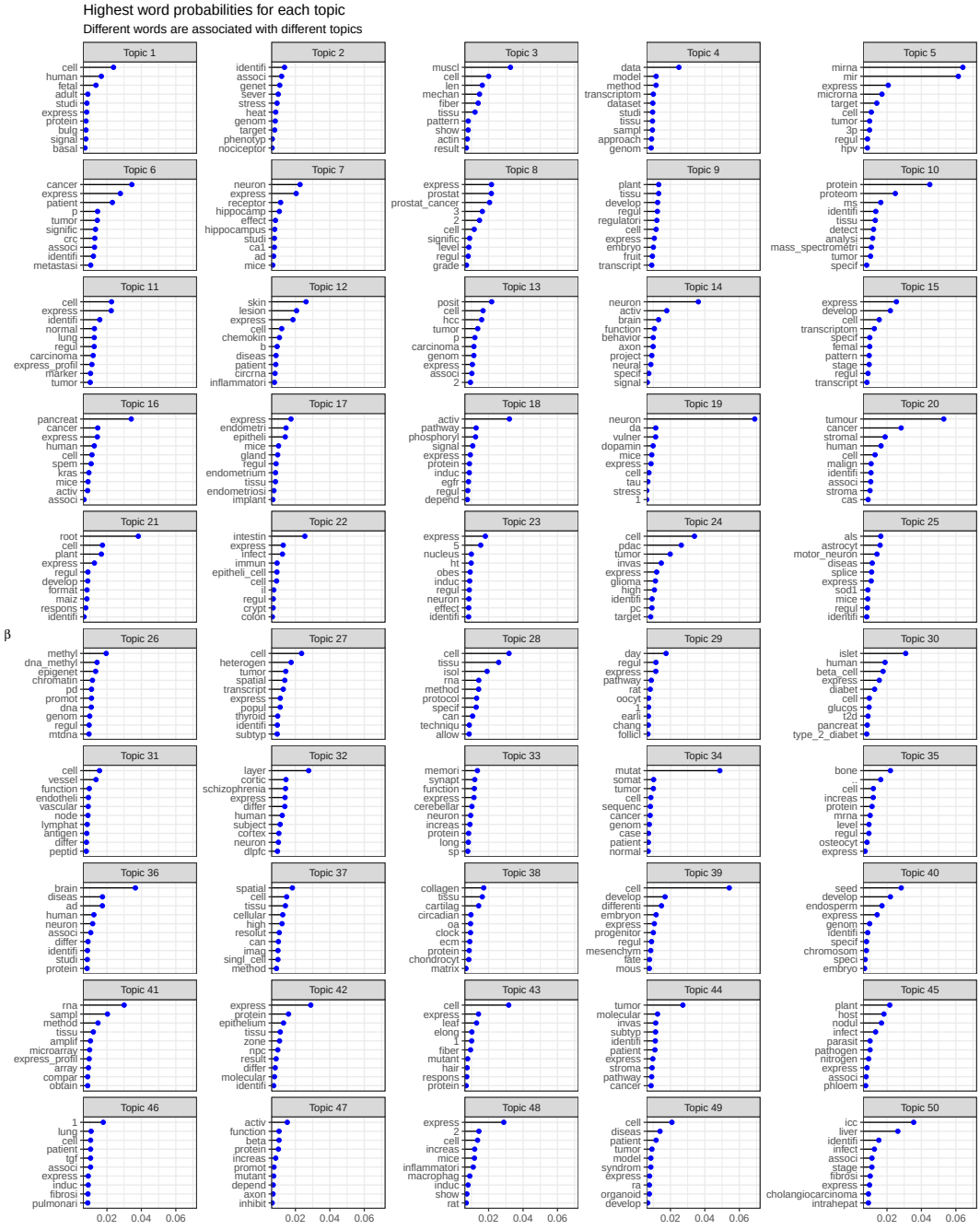

Figure 8.4: Probability of top 10 words in each topic. Zoom in or open image in new tab to see the text.

9. Plant embryos, plant development, and some stuff about evolution and ecology related to plants
10. Proteomics, especially in cancer
11. Cancer progression and diagnostics, especially lung cancer
12. Inflammation and immunology, especially in skin diseases
13. Breast cancer and liver cancer, with an emphasis in data analysis
14. Neural circuitry, neural plasticity, brain injury, and behavior
15. Plant gamitogenesis and reproduction
16. Spasmolytic polypeptide-expressing metaplasia (SPEM), oncogenes, KRAS
17. Endometrium and implantation. Somehow the top 2 entries are about hearing loss. Why? Epithelium?
18. Cell cycle, also hepatic zonation and circadian rhythm (the latter is also a cycle)
19. Neurons, especially dopaminergic
20. Tumor stroma and microenvironment
21. Plant roots
22. Intestine, especially microbiome and immune response
23. Hypothalamus, obesity, and appetite
24. PDAC, and some stuff about glioma and prostate cancer
25. ALS, and other neurodegenerative diseases affecting motor neurons
26. Epigenetics
27. Tumor single-cell profiling and cellular heterogeneity
28. Tissue isolation and preparation
29. Bone growth plate, especially recovery after radiotherapy, and some other stuff like oocytes, glaucoma, and epithelial injury
30. Pancreas and diabetes, especially T2D
31. Lymphocytes, lymphatic and blood vessels
32. Prefrontal cortex and schizophrenia
33. Synapses, dendritic spines, neuron potentiation, sometimes related to memory
34. Cancer genomics, mutations, and phylogeny
35. Bone formation, but also some other stuff about cancer and kidneys
36. Neurodegenerative diseases, Alzheimer's, Parkinson's, and multiple system atrophy
37. Spatial single-cell techniques and imaging
38. Connective tissues and ECM, and some other stuff about circadian rhythms
39. Stem cells and development

40. Plant seed development and reproduction
41. RNA extraction and amplification, especially in microarray, but also in RNA-seq
42. Lots of different stuff about epithelium
43. Plant leaves, but also other stuff about gamitogenesis
44. Cancer pathway analyses and molecular and cellular mechanisms
45. Plant nitrogen fixation and soil microbiome
46. Lots of different stuff related to fibrosis and fibroblasts, such as in lung diseases and graft rejection
47. Neuron morphogenesis, axon guidance, somehow also angiogenesis, protein signaling
48. Inflammation, immune response, especially in atherosclerosis, though there's some other content about blood vessels
49. Model organisms and in vitro model systems
50. Intrahepatic cholangiocarcinoma (ICC)

Some of them might not really be related to LCM (e.g. GWAS), and some seem to be a mixture of different topics recognized by humans but seemingly united by something else in common. There are very likely more than 50 topics present, depending on how a topic is defined. The topics can be broadly categorized into Botany, Cancer, Development, Immunology, Neuroscience, Technical, and Other, though these categories can overlap. Some of the "Other" topics seem like mixtures of multiple topics, such as topic 29, while some are very specific and relevant, such as topic 30 (pancreas and diabetes). The broad categories will be used in further analyses.

Clusters of related topics can be seen in the topic correlation plot. See documentation of `topicCorr` in the `stm` package for more details. Here we use a high-dimensional undirected graphical (HUGE) model [3] to estimate the topic correlation graph. The topic proportions (θ) are assumed to be multivariate Gaussian, and HUGE tries to identify edges connecting topics that are not independent from each other conditioned on everything else, while trying to keep the graph sparse (few edges). While θ is not Gaussian, the results from HUGE aren't unreasonable.

Indeed, cancer, botany, neuroscience, and technical topics tend to cluster together, although this is not the case for immunology and development.

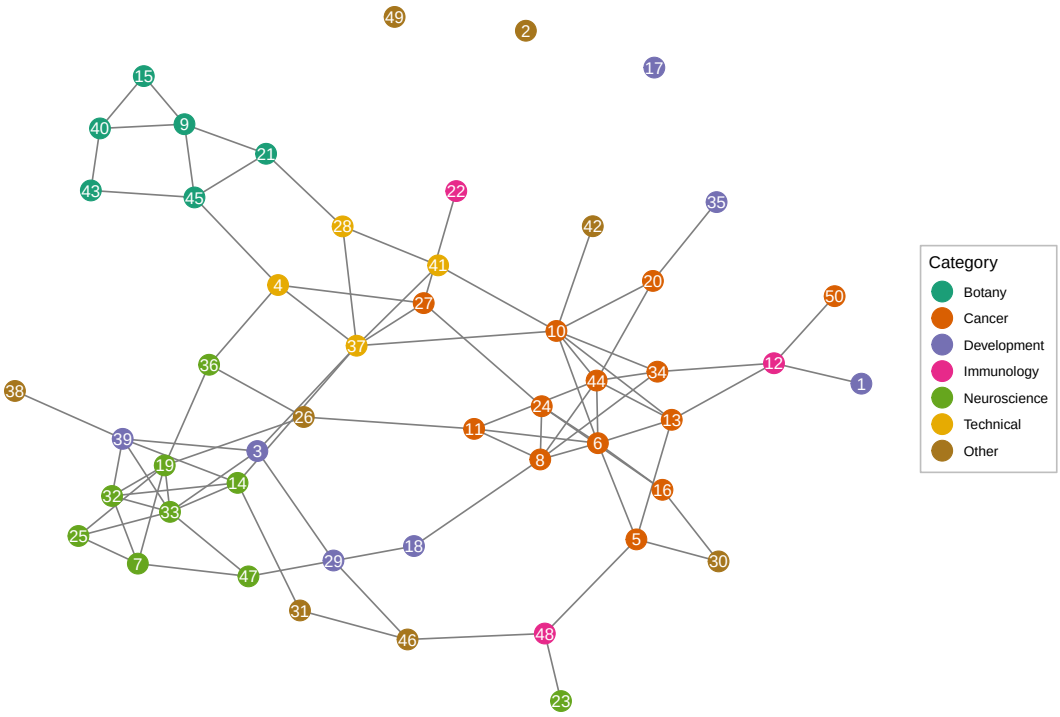

Figure 8.5: Correlation between topics.

8.2 Changes of word usage through time

We binned dates into years and tested for association of word proportion in each year with the year by fitting a logistic regression model and checking significance of the coefficient for year; word frequency per year since 2001 for the significant words (after Benjamini-Hochberg multiple testing correction) are shown in Figure 8.6. Because too many words are significant, only top 10 from words with decreasing frequency and top 10 with increasing frequency are plotted.

Here we see that words and phrases associated with microarray and RNA amplification have declined in frequency, while words associated with RNA-seq, single-cell, as well as words discussing molecular mechanisms have increased in frequency (Figure 8.6). While transcripts from LCM samples from recent studies were still amplified, the relevant terms decreased in frequency probably because more recent studies, such as ones in the curated database, tend to cite established protocols and kits of library preparation that do the amplification such as Smart-seq2 rather than discussing amplification directly. The "spatial" is associated with current era techniques. Such trends can also be clustered and shown in a heatmap.

Some words have increased in frequency, especially since 2015 (Figure 8.7). Some words sharply decreased in frequency in the early 2000s. However, some words have

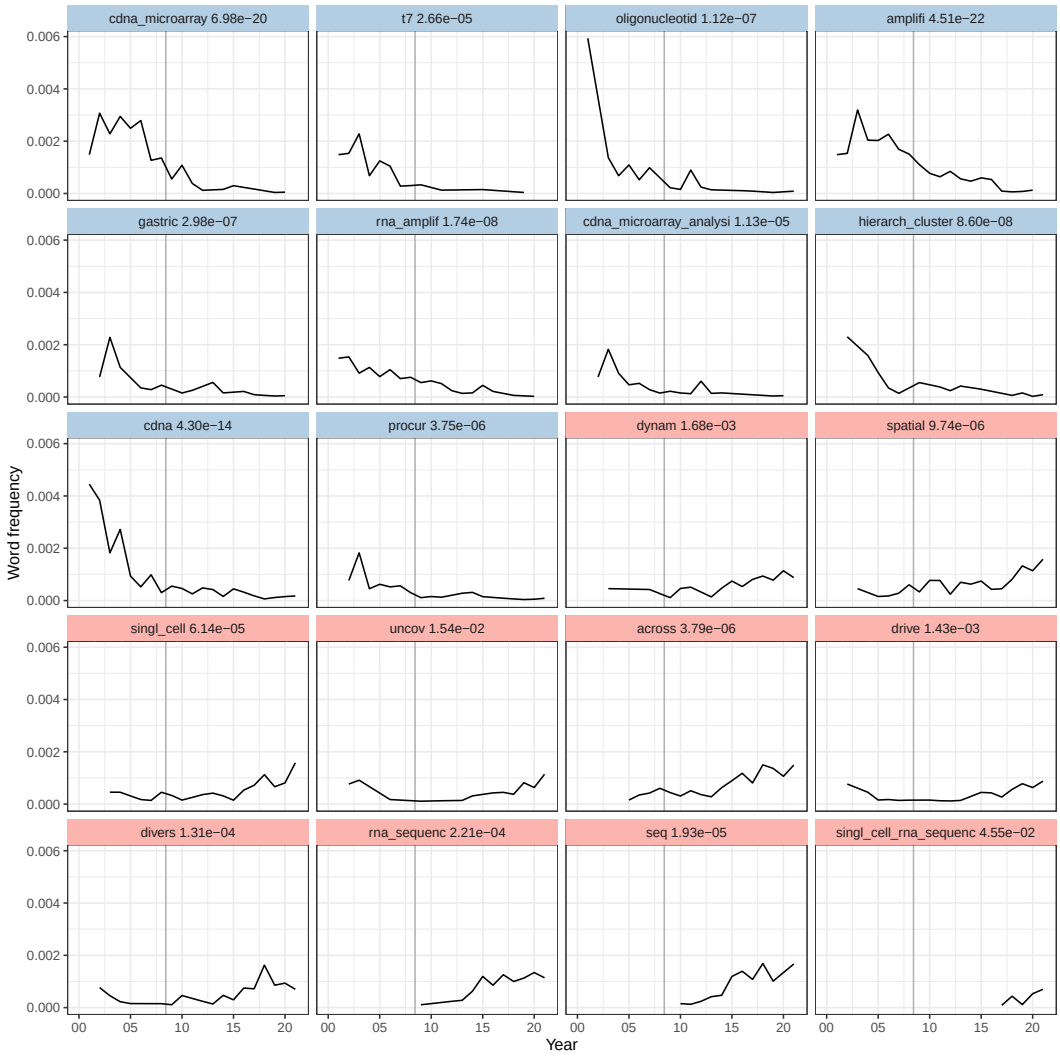

Figure 8.6: Word frequency over time since 2001 for words significantly associated with time, sorted from the most decreasing to the most increasing in frequency in time according to the slope in the model. The adjusted p-value of each word is shown. Vertical line marks June 6, 2008, when the first paper about RNA-seq was published [4].

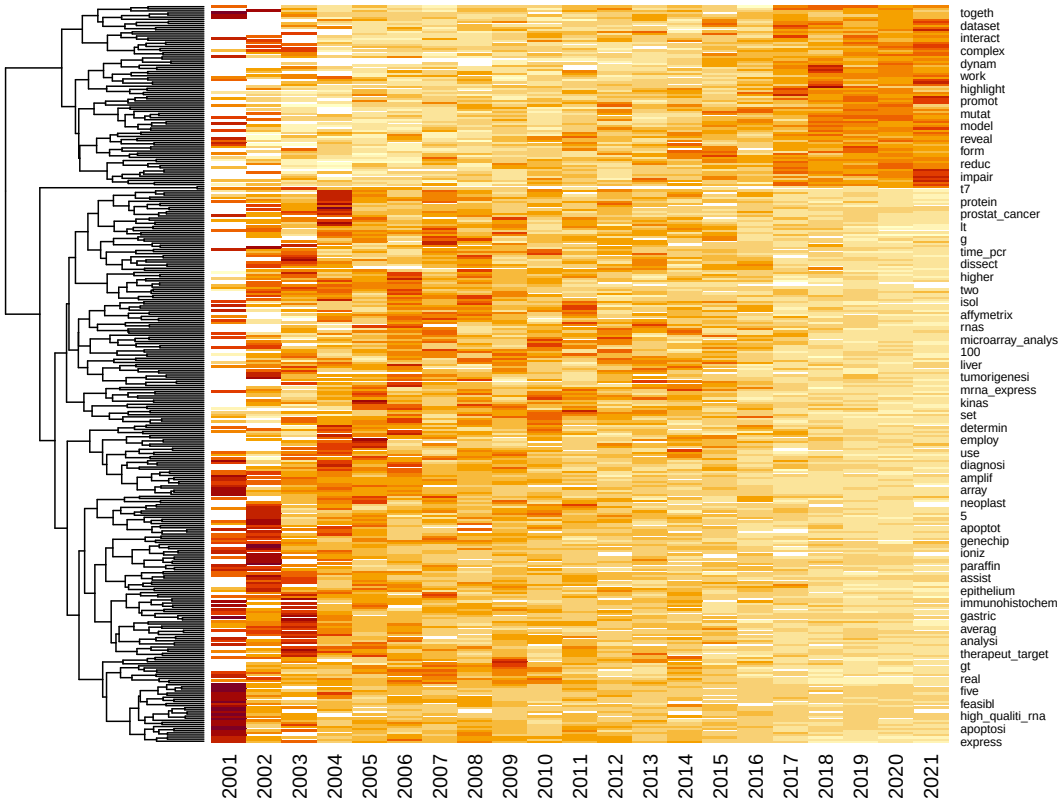

Figure 8.7: Heat map clustering changes in word frequency over time. The rows of the matrix are normalized, only showing trend rather than frequency.

increased in frequency, peaking in the late 2000s and early 2010s, before declining. Among the terms whose frequency peaked around the early 2010s are "microarray" and "microarray analysis", perhaps because while RNA-seq was introduced in 2008, microarray did not immediately become obsolete, or perhaps because microarray results are often compared to RNA-seq results, though perhaps wordings changed through the 2000s so the "cDNA" in "cDNA microarray" was omitted (Figure 8.6). Frequency of "real time PCR" also declined, probably because real time PCR was often performed along side microarray but not scRNA-seq to corroborate microarray results (e.g. [5, 6], so usage of this term declined with the decline of the cDNA microarray. Besides microarray related terms, some of the words that decreased in frequency are biological terms related to cancer. The "frequency" here is the proportion of all words from all abstracts of a year taken up by a word; the decline in proportion can either be due to decline in interest in the topics that use the word or growth in other topics that don't use the word. This will be explored further in the next section.

8.3 Changes of topic prevalence through time

We tested for association of prevalence of each of the 50 topics with time using the `estimateEffect` function in the `stm` package. Samples of the parameters were taken from the variational posterior of the `stm` model to estimate the variances of the slopes of the linear model of topic prevalence vs. date published, as well as to test whether topic prevalence is significantly associated with time. The p-values of the slopes were corrected for multiple hypothesis testing with the Benjamini-Hochberg method. While the linear model only captures monotonous changes, a more flexible model, such as b-spline transform of the date, was not used because of the modest size of this corpus; on average, each topic has only 45 abstracts, though some topics are larger and some smaller.

As many topics have statistically significant associations with time, only the top 10 most decreasing and top 10 most decreasing topics are plotted here (that's what I intended, but there were only 8 significantly decreasing topics, so top 12 increasing topics are shown). In the early 2000s, a major topic of research about LCM was reliability of T7-based PCR amplification of the small amount of transcripts from samples for microarray, but the prevalence of this topic (topic 41) has declined over time (Figure 8.6, Figure 8.8). The reason for such decline can be a combination of the following: First, other topics in neuroscience and botany emerged and grew (Figure 8.8); some of them are now among the most prevalent topics (Figure 8.3). Second, usage of terms related to microarray and RNA amplification for microarray declined while usage of terms related to RNA-seq increased after 2008 due to the advent of RNA-seq because the latter replaced microarray as the transcriptomics method of choice, so the decline is expected (Figure 8.6). Also as expected, prevalence of topics in data analysis (topic 4) and spatial single-cell and imaging technologies (topic 37) increased. Interestingly, cancer topics are among the most significantly decreasing (Figure 8.7, Figure 8.8). Because unlike cDNA microarray, these topics are still relevant today, such decline is puzzling.

Next, we checked whether whether the rise of topics not directly related to cancer may be relevant to the decline of proportions of cancer topics. In `stm`, the abstracts are not hard assigned to topics. Rather, each abstract has a proportion of each topic, and abstracts often have over 90% of one topic. Here, for simplicity, we say an abstract "has" a topic if the proportion of the topic in the abstract is at least 25%.

When the number of abstracts with each topic is plotted, the declines are less drastic or reversed while the increases became much more drastic, especially after 2015,

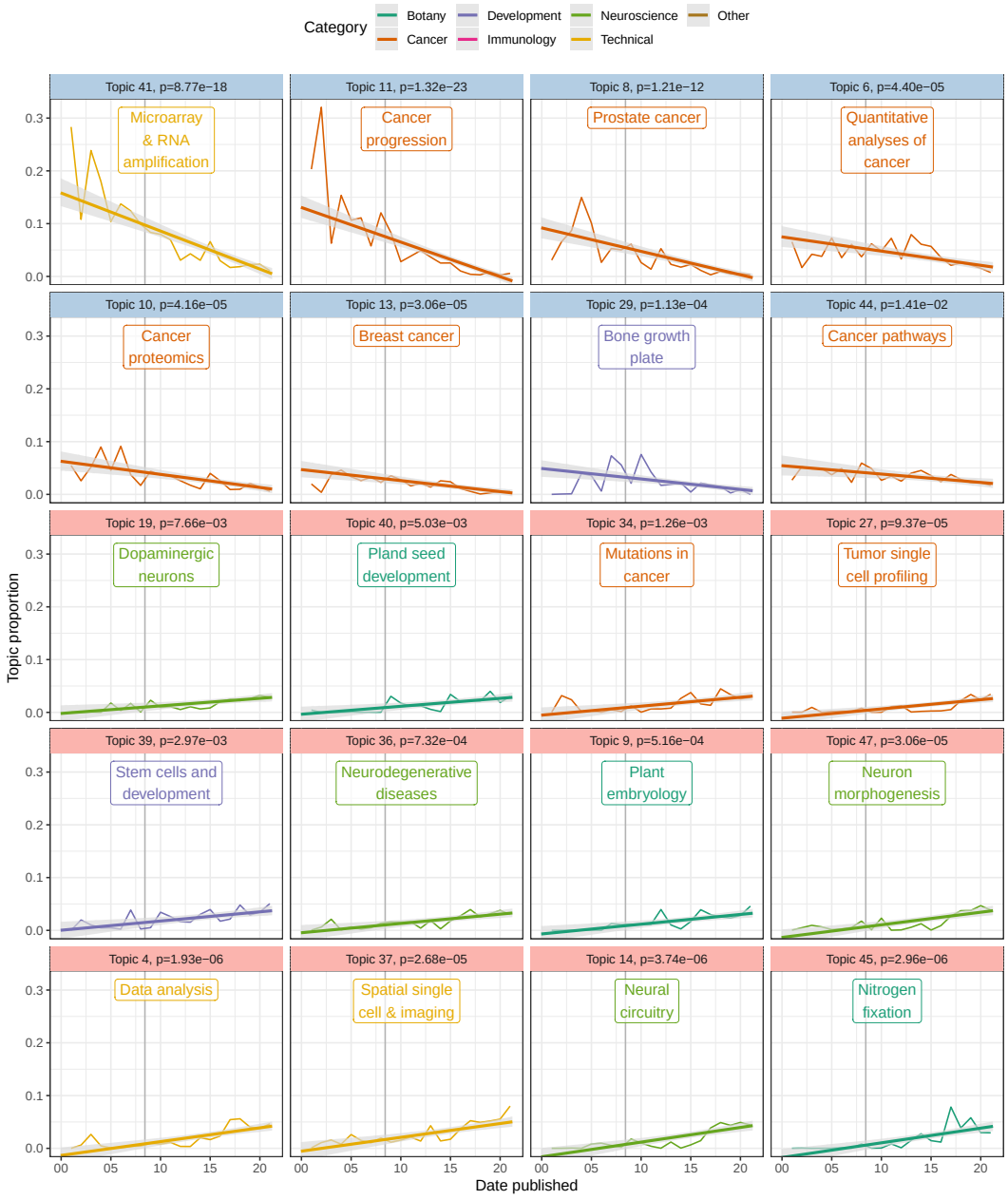

Figure 8.8: Topic prevalence over time since 2001 with fitted linear model. Gray ribbon indicates 95% confidence interval (CI) of the slope, estimated from the samples of the variational posterior of the `stm` model. Vertical line indicates advent of RNA-seq in 2008. Light blue facet strip means decreasing trend with adjusted p < 0.05, and pink strip means increasing.

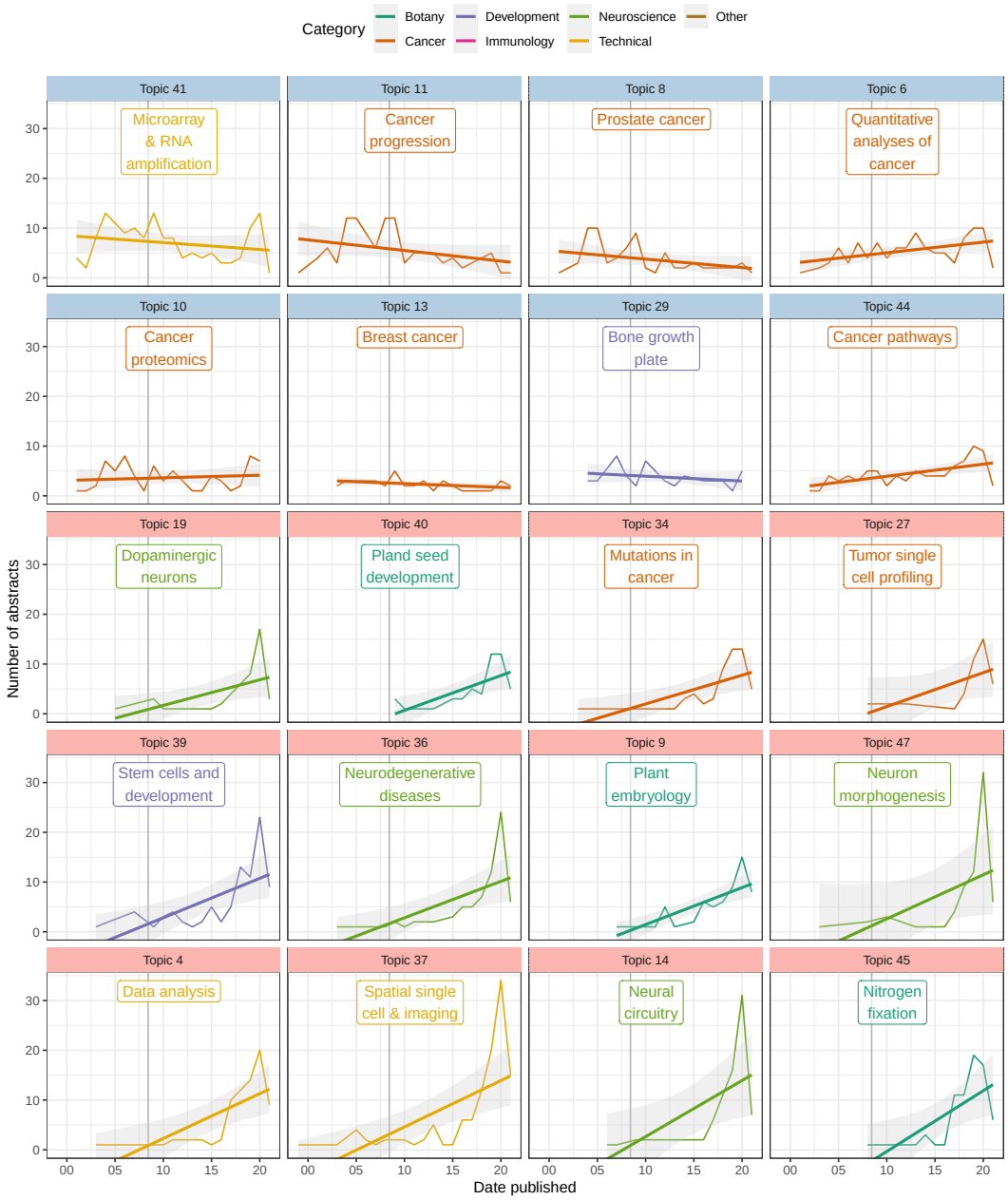

Figure 8.9: Number of abstracts with each topic whose proportions changed the most in time. Gray ribbon is the 95% CI of the line fitted to the count per year.

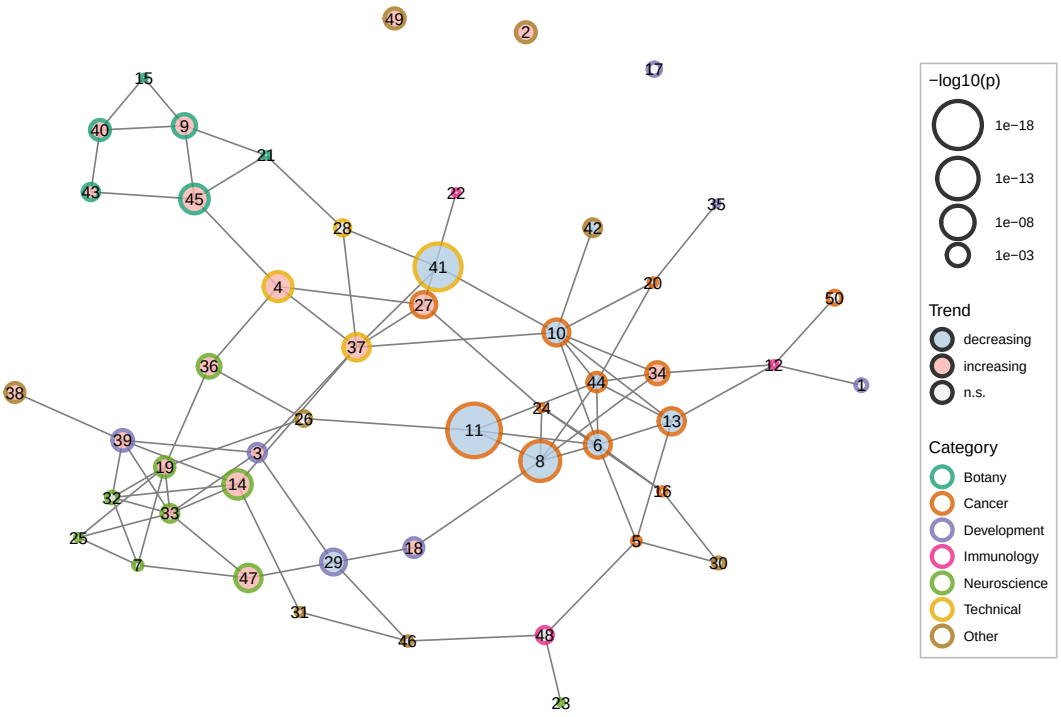

Figure 8.10: Correlation between topics colored by both broad categories of the topics and whether its proportion increased, decreased, or did not significantly change (n.s.).

perhaps due to the rise of scRNA-seq, whose library preparation methods made it possible to quantify transcripts from small amount of tissues from LCM (Figure 8.9). These trends don't necessarily correspond to the overall trend across the corpus (Figure 8.1). Then we see in recent years a diversification of topics that may be related to LCM from search results, resulting into decrease of proportion of some older topics the interest in which might not have drastically decreased if not somewhat increased, though not increasing as quickly as other topics. Nevertheless, it is clear that some cancer topics have decreased even in counts. However, remember that some of the `stm` topics seem to be mixtures of multiple topics recognizable by humans and these `stm` topics might have picked up aspects of the abstracts less readily noticed by humans. In other words, it might not be that interest in some cancers decreased per se, but thanks to scRNA-seq, the way these cancers are discussed changed, using words that contributed to other, growing topics. Furthermore, because so many different topics are drastically growing in recent years, the increase in proportion of each of them became less drastic.

Now return to the topic correlation graph, and all the 50 topics, along with their

trends, are shown (Figure 8.10). Overall, cancer topics tend to be decreasing in proportion. As already seen in Figure 8.9, this is in part due to growth in non-cancer topics but in part due to decline in some cancer topics. Botany and neuroscience topics tend to increase in proportion. This trend is also evident in the topic correlations. Microarray and RNA amplification (topic 41) is correlated with a cancer topic, while spatial single-cell and imaging (topic 37) and data analysis (topic 4) are correlated with neuroscience topics. Topic 27, which is about single-cell profiling of tumors, has grown, perhaps due to the growth of scRNA-seq. Possibly, as cancer is still relevant, the decline in some cancer topics fed into topic 27 as tumors are examined at the single-cell level.

8.4 Association of topics with city

Again, with the `estimateEffects` function, we identify cities associated with certain topics. Some topics might be more spread out, while some some may be confined to a few institutions, which are approximated by city here because it's more difficult to automatically extract institutions from the author address on PubMed than cities. Some institutions might specialize in certain topics. Also note that while for PubMed papers, the cities of the first author are used, because the first author has greater contribution to the paper, only the address of the corresponding author is available from the bioRxiv API. Furthermore, multiple institutions across continents may collaborate on one paper, so the cities here only give a rough idea where LCM related research takes place. Here only the names of the cities are used, with the state and country they are in to distinguish between cities with the same name, without the longitude and latitude, because we don't expect an association between topic and the coordinates in and of themselves, nor do we expect spatial autocorrelation of the topics.

Here we note that Center for Dementia Research, Nathan Kline Institute in Orangeburg has greatly contributed to research in hippocampal CA1 pyramidal neurons in Alzheimer's disease and Down syndrome (topic 7) (Figure 8.11). This is the first time I heard of Nathan Kline. Department of Plant Biology at Cornell, Ithaca has greatly contributed to study of plant development (topic 9). Topic 17 is a mixture of topics recognizable by humans; besides the endometrium, some of the top entries are about hearing loss, which come from University of Rochester. George Mason University in Manassas, Virginia contributed several papers about cancer pathway analysis (topic 44). University of Pittsburgh has disproportionate contribution to the study of prefrontal cortex and schizophrenia (topic 32), dating back to 2007.

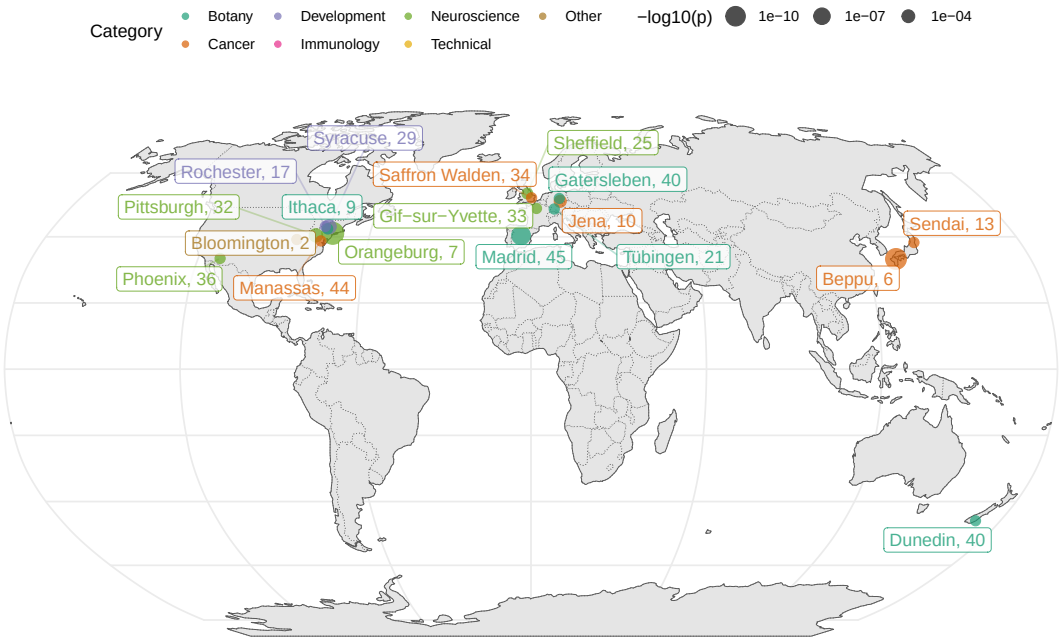

Figure 8.11: Cities associated with topics (p < 0.005) shown on a map.

Centro de Biotecnologia y Genomica de Plantas (UPM-INIA), Madrid has disproportionate contribution to the study of soil microbiome and nitrogen fixation (topic 45). University of Sheffield has a long history and disproportionate contribution to the study of neurodegenerative diseases affecting motor neurons (topic 25), dating back to 2007.

Association of a topic with an institution that used to greatly contribute to the topic but then stopped might also explain why some topics declined in prevalence over time although drastic growth in other topics might be a better explanation (Figure 8.8, 8.9). Topic 29 prominently features the bone growth plate though this `stm` topic has entries for other biological systems as well. These bone growth plate papers come from Upstate Medical University in Syracuse, New York, from 2005 to 2010. Decline in topic 29 might be related to cessation of study of the growth plate at this institution after 2010, though other institutions have not picked up this topic afterwards. Institute of Human Genetics and Anthropology, Friedrich-Schiller-University in Jena, Germany greatly contributed to cancer proteomics (topic 10) between 2004 and 2011 but then stopped, though other institutions carried on studying this topic. Kyushu University Beppu Hospital in Japan greatly contributed to quantitative analyses in cancer (topic 6) from 2005 to 2014, although other institutions continue contributing to this topic, whose paper count actually increased over time although the topic's proportion decreased due to drastic growth in other

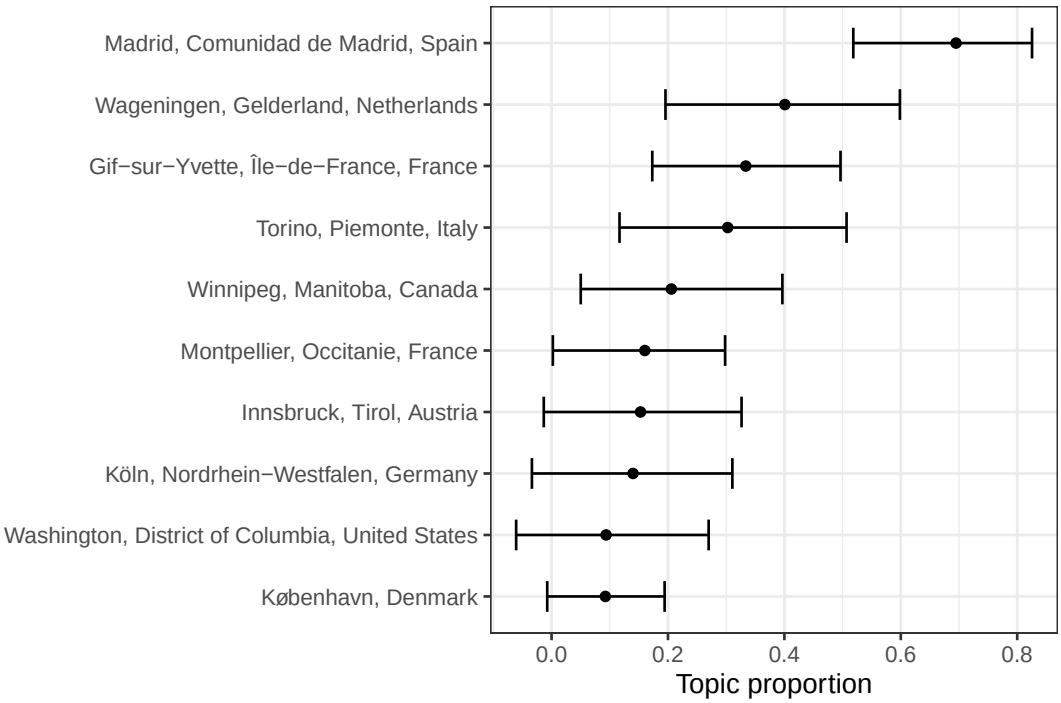

Figure 8.12: Proportion of topic 45 in each city. Error bars are 95% CI of the point estimate.

topics (8.9). The vast majority of LCM related publications from Sendai, Japan are about breast cancer (topic 13), from between 2007 to 2017, which is why Sendai is associated with this city although this topic is widespread.

Association of a city with a topic can also be visualized with topic proportion in each city from `estimateEffect` (Figure 8.12). Here topic 45 (soil microbiome and nitrogen fixation) is plotted, but readers on RStudio Cloud can try other topics.

Here "disproportionate" means disproportionate within this corpus of LCM related search results. Institutions with "disproportionate" contribution to a topic do not necessarily dominate such topic although the topic may dominate the institution, i.e. the topic takes up a very large proportion of abstracts from this institution within this corpus. Nor are these institutions necessarily elite; this analysis might be an interesting way to discover labs from not so well-known institutions that may be outstanding in some topics. The institutions are often not elite because elite institutions often greatly contribute to many topics, weakening the association of the institution to the topic. Except for growth plate in Syracuse, we have not identified topics largely confined to an institution.

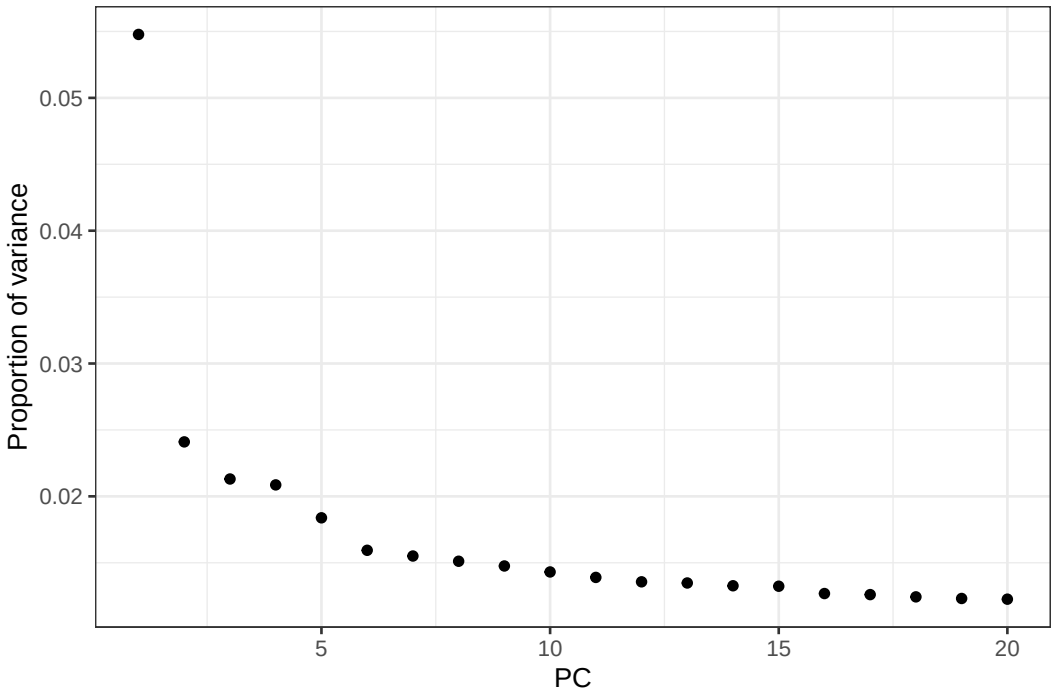

Figure 8.13: Proportion of variance explained by each of the first 20 principal components (PC).

8.5 GloVe word embedding

We used global vector (GloVe) embedding to identify linear substructures in the word vector space of the LCM transcriptomics abstract corpus and to identify contexts [7]. In GloVe embedding, words are represented by vectors. Words with similar meanings tend to be closer together in this vector space, and differences between word vectors can encode meaning as well. The "meanings" come from the context, or word co-occurrence. GloVe was devised to find a word embedding with properties like "king" - "man" + "woman" = "queen" or "ice" - "solid" + "gas" = "steam", and related words like "cancer" and "tumor" are close together but both are far from unrelated words like "flower".

This corpus was used to train a 125 dimensional embedding, and the embeddings of words occurring more than 30 times in the corpus were projected to lower dimensions with principal component analysis (PCA) to find axes explaining the most variance in the embedding, hopefully identifying dominant axes of meaning within this corpus. The words are also Louvain clustered in the embedding space to find clusters of words related in meaning.

The first principal component (PC) explains over 5% of the variance, and then the

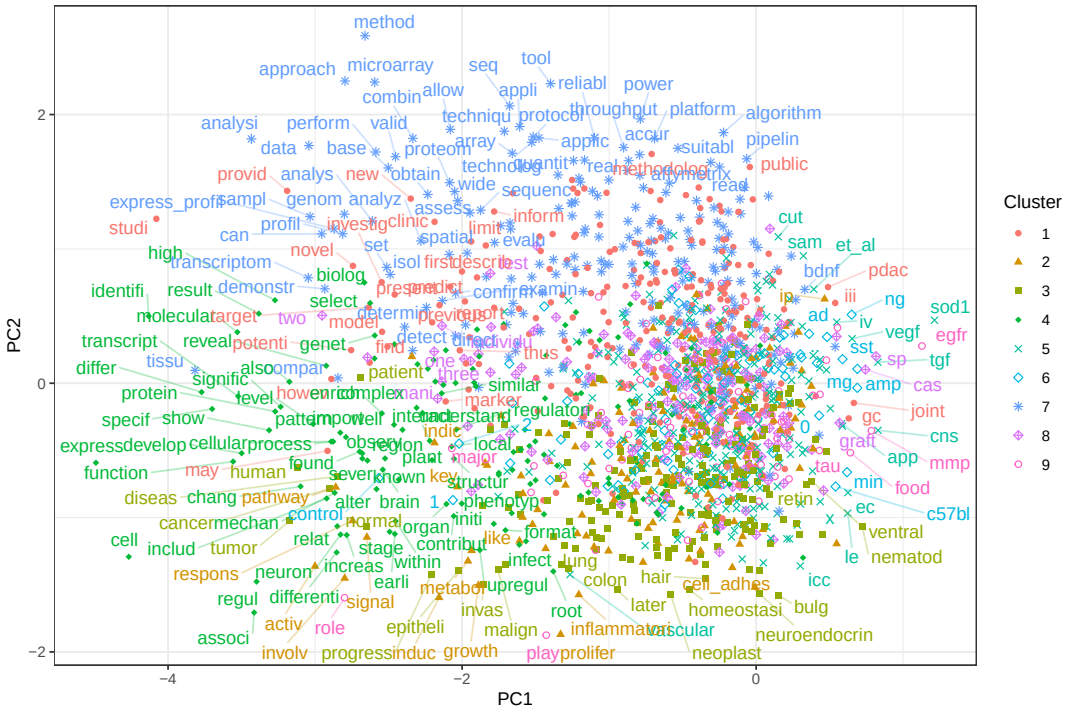

Figure 8.14: Projection of word embeddings into the first 2 PCs. Each point is a word occuring over 30 times in the corpus. Not all words are labeled to avoid overlaps in the labels. Words and points are colored by Louvain clusters.

"elbow" is at PC5.

Words more positive in PC1 are often gene names, parts of gene names, or acronyms, and names of specific biological entities or processes. In contrast, words more negative in PC1 tend to be more general and more widely used. PC2 separates the technical (cluster 2, top) from the biological (clusters 2-4) (Figure 8.14). As expected, "cancer", "tumor", and "disease" are not far from each other (bottom left), and "malignant" and "invasive" are close (bottom center). PC1 explains more variance than all other PCs; though it's only 5.5%, it picked up a very important dimension in word meanings in this corpus. PCs are arranged in decreasing order of variance explained.

PC3 separates processes and interactions (clusters 2 and 4, left) from entities of samples, tissues, organs, and diseases (clusters 3 and 7, right). PC4 separates the molecular and cellular (bottom left) and the quantitative (clusters 1 and 3, bottom right) from the qualitative (top). Some of the qualitative terms are used to discuss implications of results of the papers (clusters 1 and 4, top left) (Figure 8.15).

Now we have seen some important axes of meanings and types of words, which are

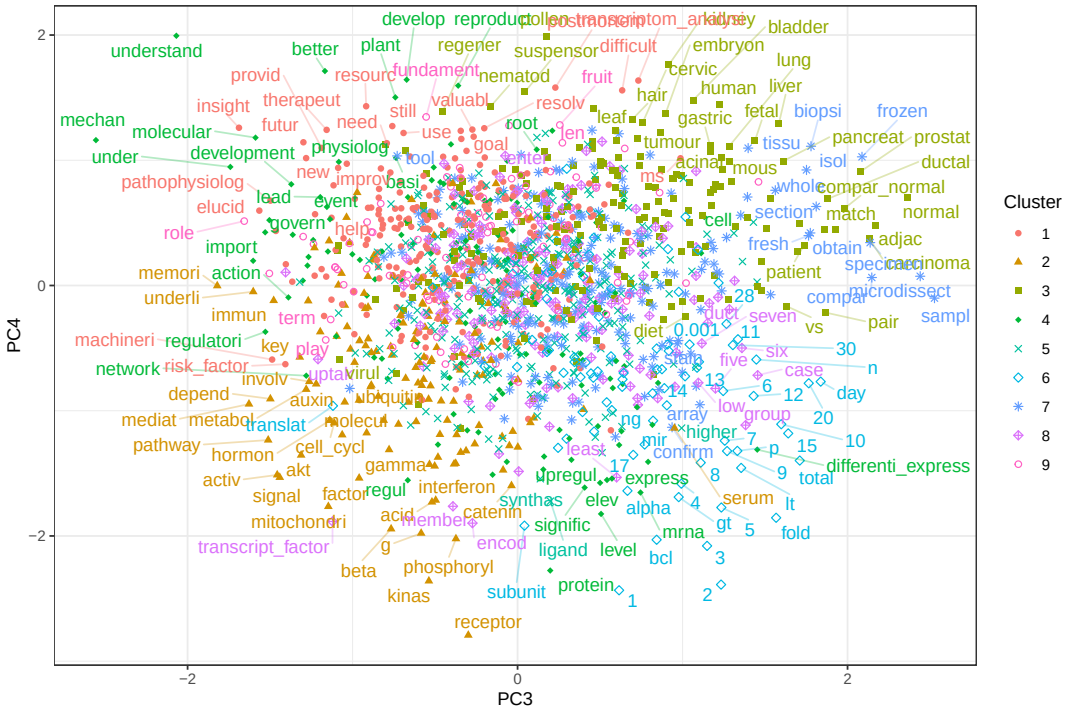

Figure 8.15: Projection of word embeddings into the 3rd and 4th PCs.

not surprising given familiarity with the general structure of abstracts and applications of LCM. There must be more axes of meaning, as the first 4 PCs only explain about 12% of the total variance of word embeddings (Figure 8.13). The clusters of words can be better visualized with UMAP, which is a non-linear dimension reduction method that tries to preserve distances between points but is most commonly used to project into 2 dimensions.

The clusters of words are easier to discern with Uniform Manifold Approximation and Projection (UMAP) (Figure 8.16). Cluster 1 is mostly terms used to discuss the results and implementations of the studies. Cluster 2 is molecular terms. Cluster 3 has many terms about cancer. Cluster 4 contains terms on quantitative molecular analyses and molecular mechanisms. Cluster 5 is biological terms. Cluster 6 is words used to describe results of studies, with many quantitative words; “p”, “0.05”, and “0.01” are found in this cluster rather than cluster 3 because p-values are results of data analyses and 0.05 and 0.01 are common thresholds of significance. Cluster 7 is words in experimental procedures. Cluster 8 has some biological terms, some quantitative terms, and numbers that are spelt out. These clusters of words give some idea about topics of the studies, but unlike `stm`, these clusters also give a glimpse into different parts of the abstract, such as summary of the results and implications

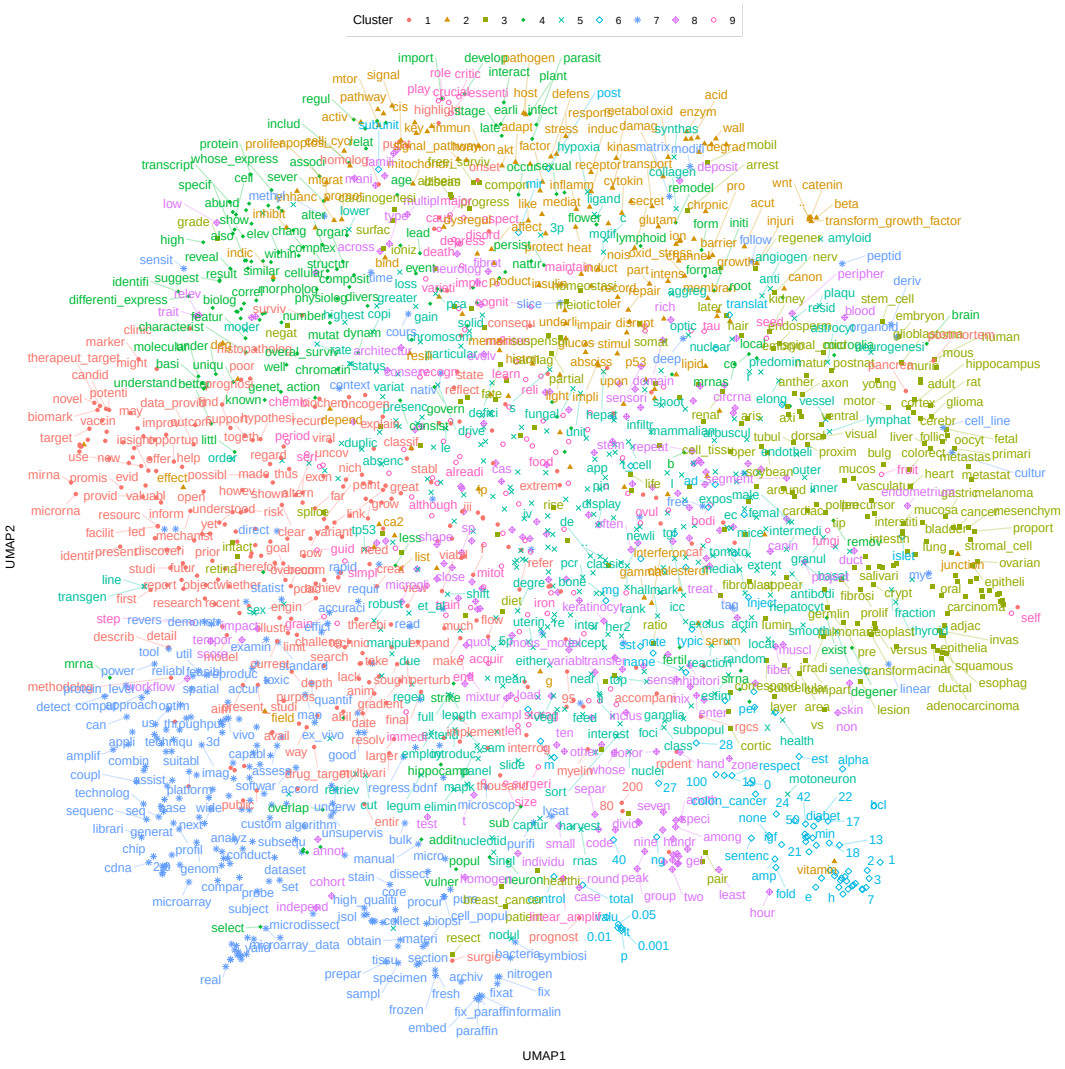

Figure 8.16: UMAP projection of word embeddings. Zoom in if reading the PDF version of this book.

of the results.

References

1. Roberts ME, Stewart BM, and Tingley D. Stm: An R package for structural topic models. Journal of Statistical Software 2019; 91. DOI: 10.18637/jss.v091.i02. Available from: https://www.jstatsoft.org/v091/i02

2. Luo L, Salunga RC, Guo H, Bittner A, Joy K, Galindo JE, Xiao H, Rogers KE, Wan JS, Jackson MR, and Erlander MG. Gene expression profiles of laser-captured adjacent neuronal subtypes. Nature Medicine 1999 Jan; 5:117–22. DOI: 10.1038/4806. Available from: http://www.nature.com/articles/nm0199_117

3. Zhao T, Liu H, Roeder K, Lafferty J, and Wasserman L. The Huge Package for High-Dimensional Undirected Graph Estimation in R. J. Mach. Learn. Res. 2012 Apr; 13:1059–62

4. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, and Snyder M. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science 2008 Jun; 320:1344–9. DOI: 10.1126/science.1158441. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1158441

5. Cunnea P, McMahon J, O'Connell E, Mashayekhi K, Fitzgerald U, and McQuaid S. Gene expression analysis of the microvascular compartment in multiple sclerosis using laser microdissected blood vessels. Acta Neuropathologica 2010; 119:601–15. DOI: 10.1007/s00401-009-0618-9. Available from: https://doi.org/10.1007/s00401-009-0618-9

6. Kitamura S, Tanahashi T, Aoyagi E, Nakagawa T, Okamoto K, Kimura T, Miyamoto H, Mitsui Y, Rokutan K, Muguruma N, and Takayama T. Response Predictors of S-1, Cisplatin, and Docetaxel Combination Chemotherapy for Metastatic Gastric Cancer: Microarray Analysis of Whole Human Genes. Oncology 2017; 93:127–35. DOI: 10.1159/000464329. Available from: https://www.karger.com/DOI/10.1159/000464329

7. Pennington J, Socher R, and Manning CD. GloVe: Global Vectors for Word Representation. Tech. rep.

Chapter 9

DATA ANALYSIS IN THE CURRENT ERA

So far we have reviewed numerous techniques to collect spatial transcriptomics data. In this chapter, we review computational methods to analyze data generated by current era techniques and methods that, while only having WMISH, FISH, or ISH as spatial data, involve scRNA-seq data as well. For a publication to be included in the "Analysis" sheet of this database, it must either focus on a data analysis method, or present alongside new data, sophisticated data analysis going beyond using existing packages. While some data analysis methods originally not designed for spatial data can be used for spatial data, this chapter is about methods designed specifically with spatial data in mind. This means that the methods should be demonstrated on a spatial transcriptomic dataset in the publication, even if not explicitly using spatial coordinates.

Since 2019, there has been a sharp increase in interest in current era data analysis (Figure 9.2). If our collection of prequel data analysis literature is somewhat

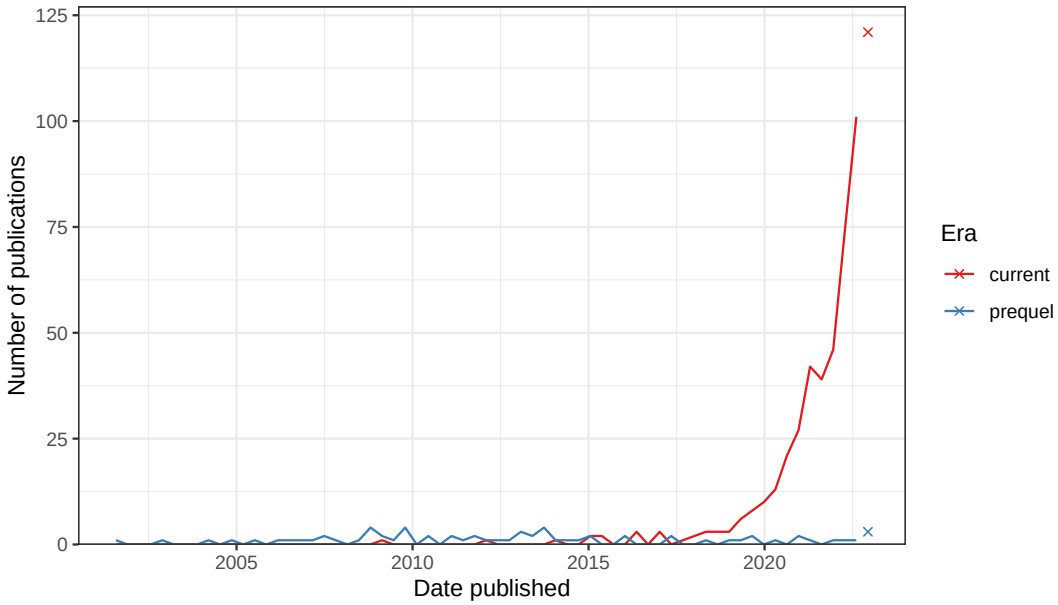

Figure 9.1: Number of publications over time for current era and prequel data analysis. Bin width is 120 days. Preprints are included for this figure. The x-shaped points show the number of publications from the last bin, which is not yet full.

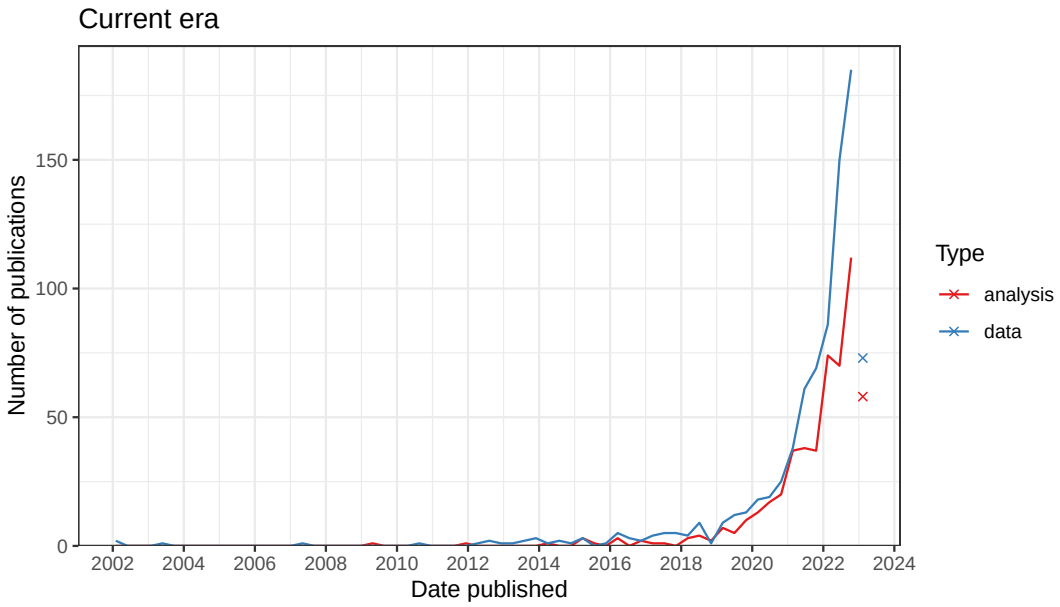

Figure 9.2: Number of publication over time for current era data collection and data analysis. Bin width is 120 days. The x-shaped points show the number of publications from the last bin, which is not yet full.

representative and complete, then interest in current era data analysis dwarfs the golden age of prequel data analysis from 2008 to 2014 (Figure 9.1). As already shown, interests in current era data collection increased sharply since 2018 (Figure 6.2, Figure 9.2), although not as sharply as in data collection (Figure 9.2).

In contrast, in the prequel era, interest in data analysis peaked after the peak for data collection, and eventually interest both eventually diminished but continues (Figure 9.3). There are many different types of data analysis, the ones with the most interest are finding spatial regions, preprocessing (including image processing and quality control), cell type inference (especially cell type deconvolution of Visium spots), and cell-cell interaction (Figure 9.4). While mapping dissociated cells to spatial locations on a spatial reference used to be at the top, there has been more interest in the other topics mentioned just now.

Several methods for cell type deconvolution in array based techniques that don't have single-cell resolution were developed (cell type inference), but the drastic growth in data analysis seems to be driven by multiple categories of analyses (Figure 9.5). Top contributors to data analysis methods in the current and prequel eras are different as well. In the current era, while many less well-known institutions have contributed to data analysis, the top contributors are an elite club. Among the top contributors in the

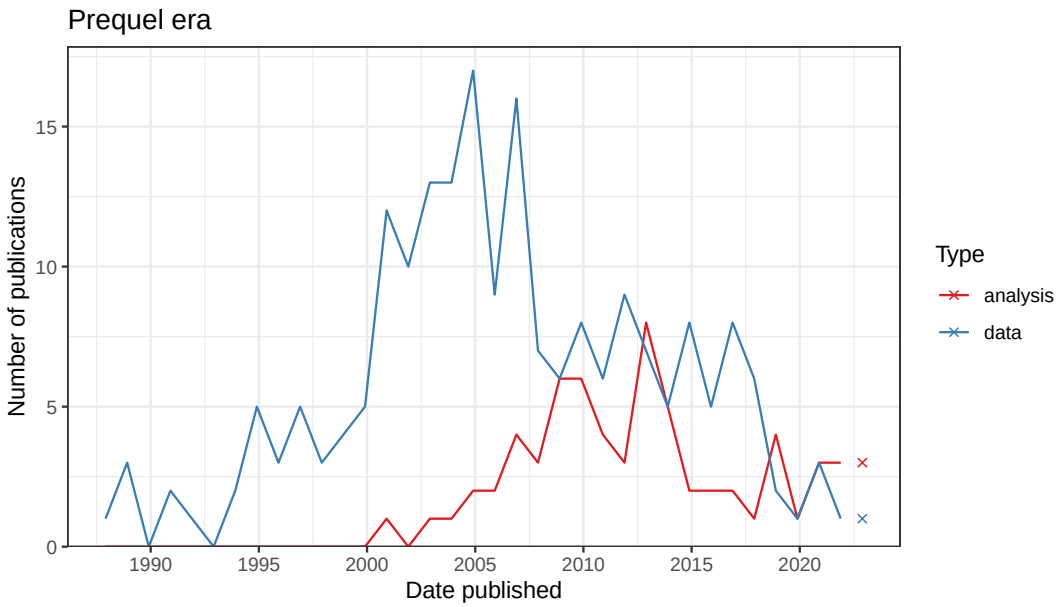

Figure 9.3: Number of publications over time for prequel data collection and data analysis. Bin width is 365 days. The x-shaped points show the number of publications from the last bin, which is not yet full.

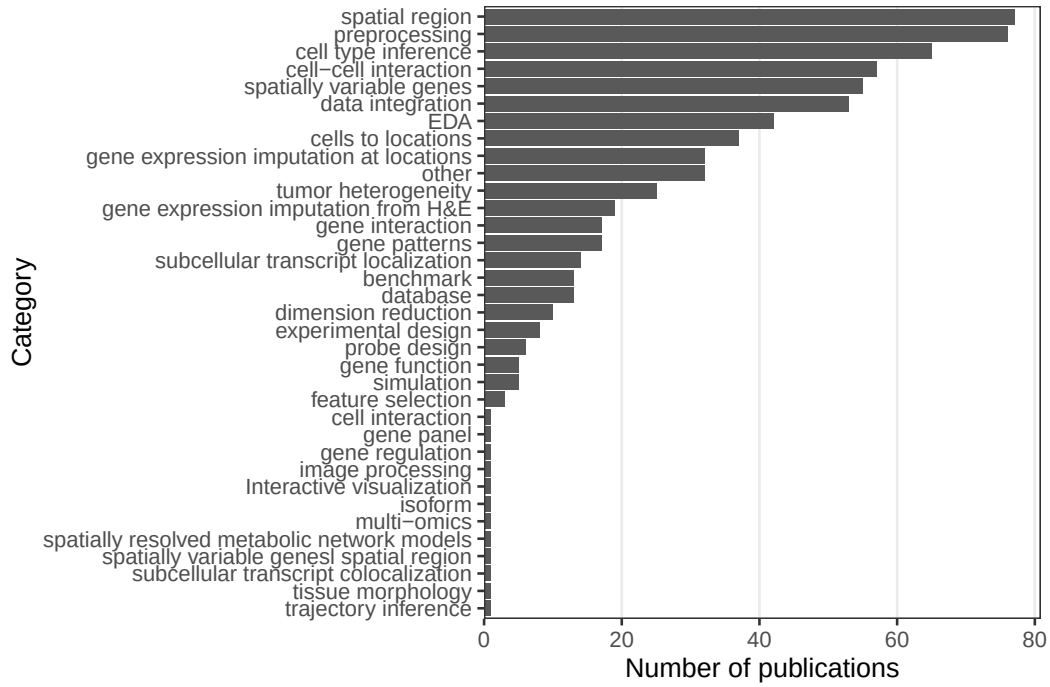

Figure 9.4: Number of publications for each category of data analysis; note that the same publication can fall into multiple categories.

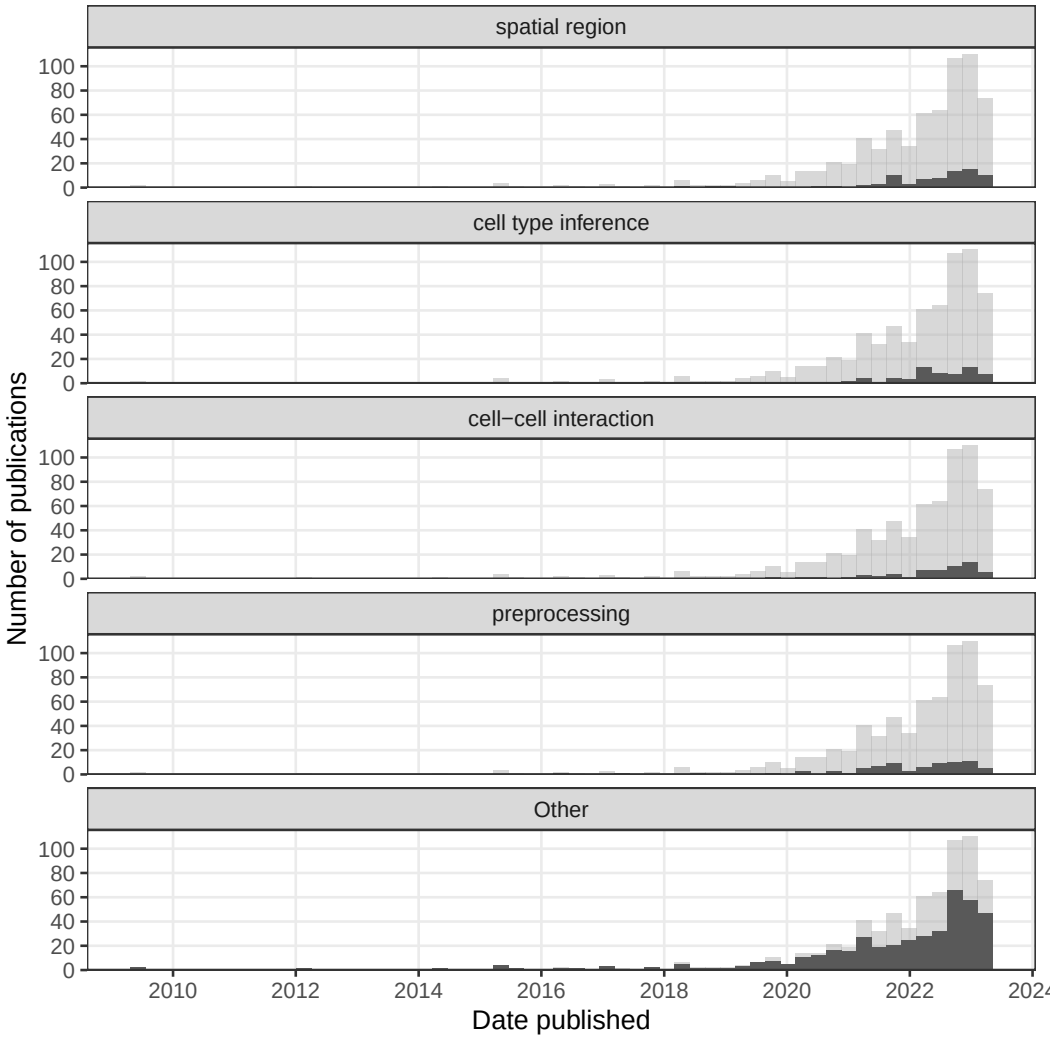

Figure 9.5: Number of publications over time broken down by type of data analysis. The 3 categories most popular in the past year are shown, and the others are lumped into 'Other'. Bin width is 90 days.

prequel era are less famous institutions such as Arizona State University (ASU), Old Dominion University (ODU), and Lawrence Berkeley National Laboratory (LBL), which developed the BDTNP and the Fly Enhancer atlases (Figure 9.6).

In our database, we have recorded programming languages used in data analysis or package development. All programming languages that played a major role in the project were recorded. For downstream analysis, this includes languages of the user interface of existing packages used and languages of new functions written for the project. For package development, this includes any language used to write the package essential to the functioning of the package. In publications that focus

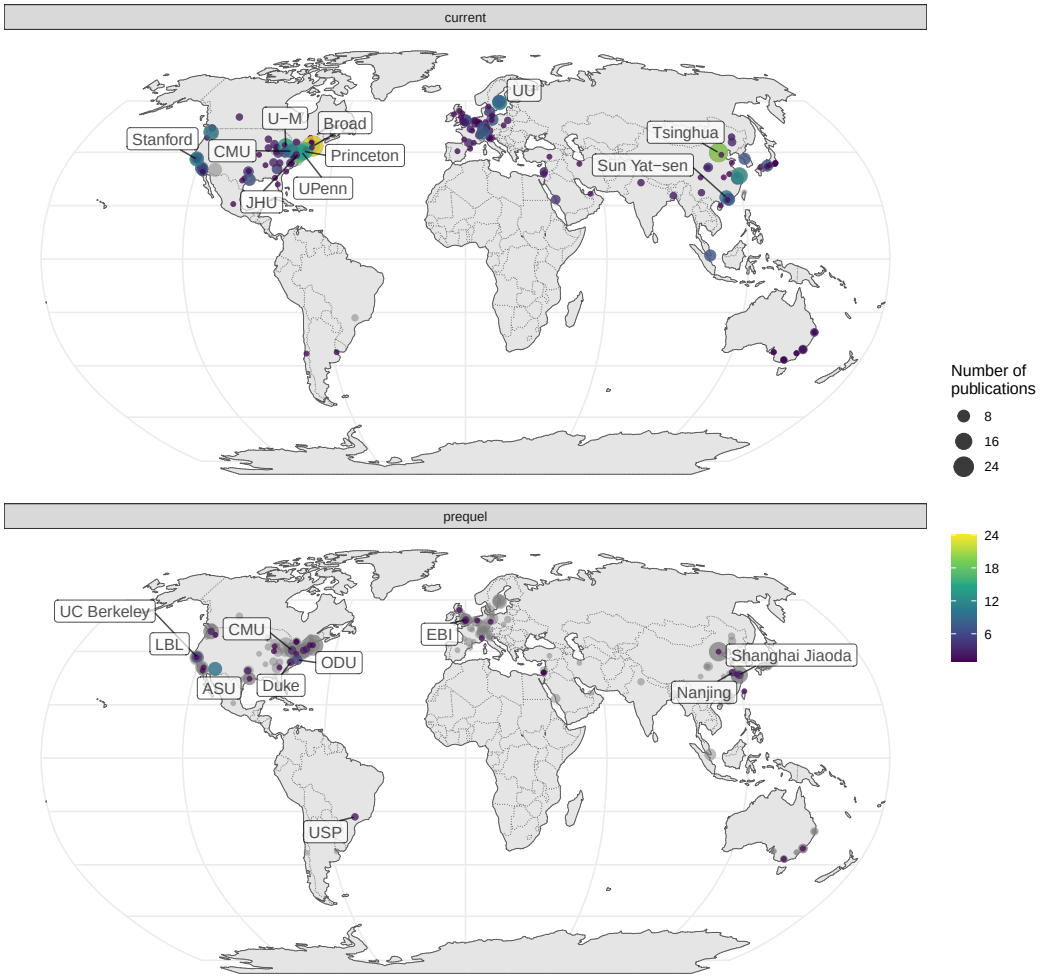

Figure 9.6: Map of where first authors of current era and prequel data analysis papers were located as of publication. Each point is a city and point size is number of publications from all institutions in the city. Top 10 institutions in each era are labeled.

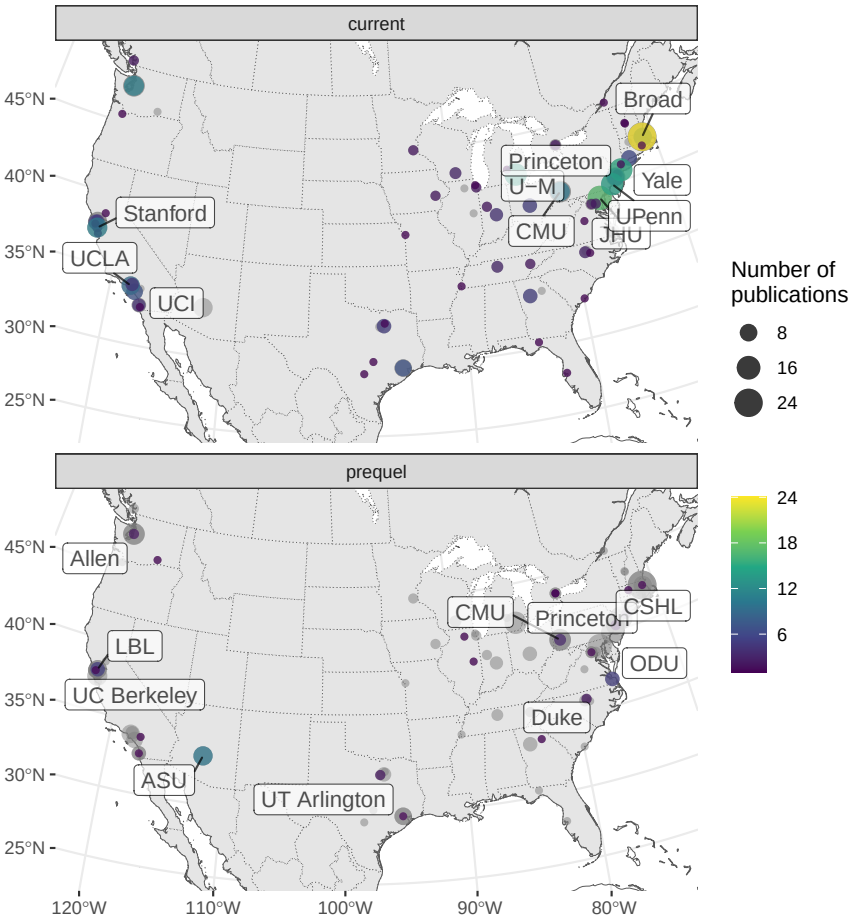

Figure 9.7: Map of where first authors of current era and prequel data analysis papers were located as of publication around continental US. Top 10 institutions in each era are labeled.

on data collection, R is by far the most popular programming language used in downstream data analysis (Figure 9.10). The second most popular is Python, and then MATLAB, which is more common in smFISH (Figure 7.28) and ISS for its image processing functionality. Python is used for both image processing and other types of analyses. C and C++ are not as common in downstream analysis.

The same top 5 programming languages are the most common for developing data analysis packages (Figure 9.11). Python is the most popular, especially for packages involving deep learning, image processing, using Torch for optimization, or are command line tools. R follows, and is more popular for exploratory data analysis (EDA) and data visualization, but sometimes both R and Python are used in the same package. Other languages aren't nearly as commonly used for packages reported on in our database. The above observations about usage of R and Python seem to

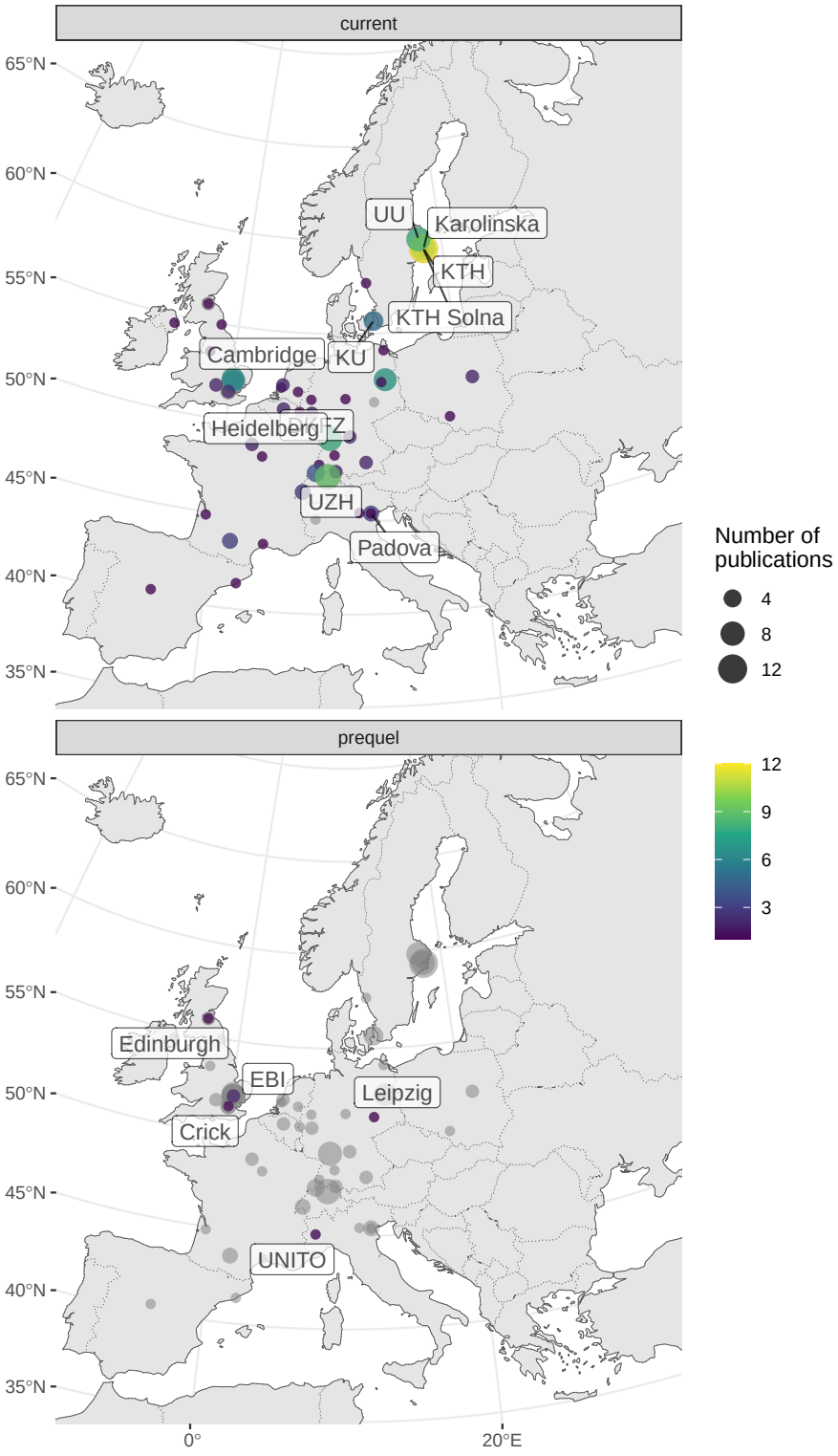

Figure 9.8: Map of where first authors of current era and prequel data analysis papers were located as of publication in western Europe. Top 10 institutions in each era are labeled.

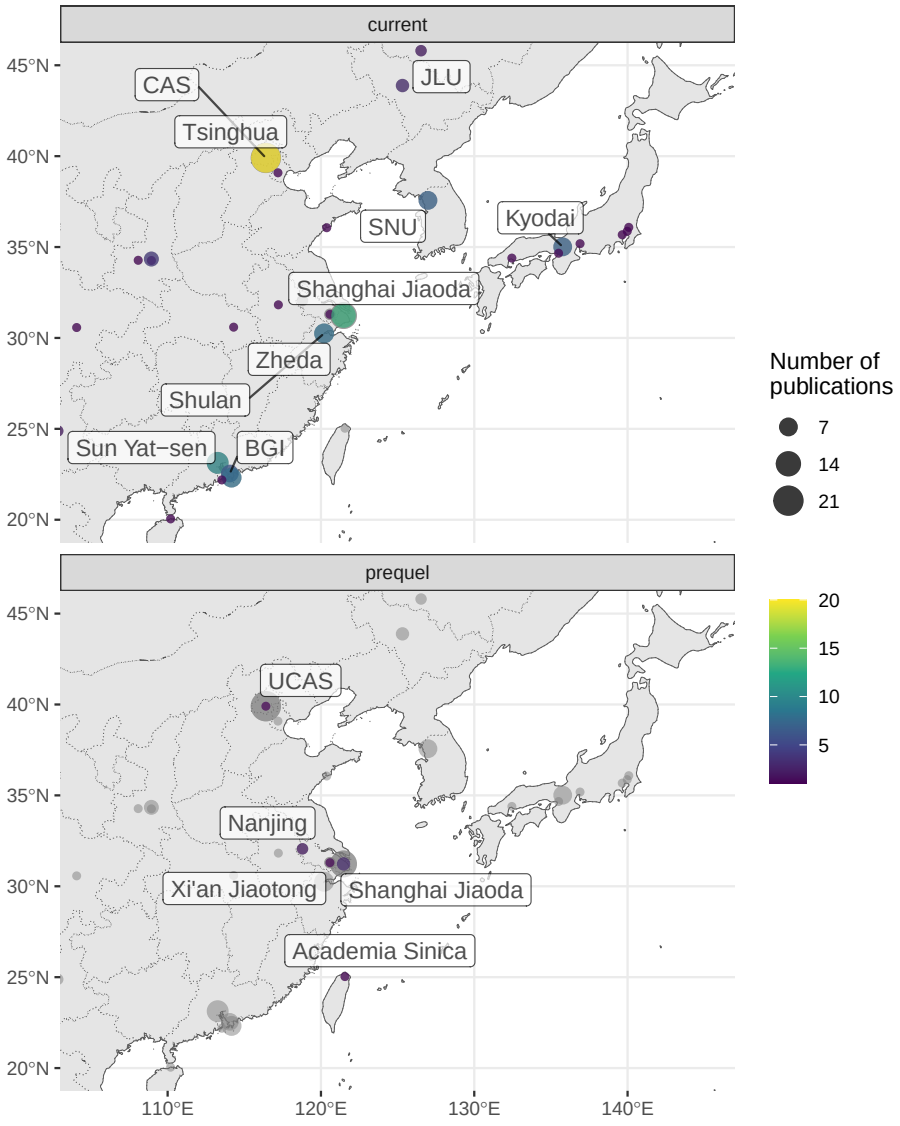

Figure 9.9: Map of where first authors of current era and prequel data analysis papers were located as of publication in northeastern Asia. Top 10 institutions in each era are labeled.

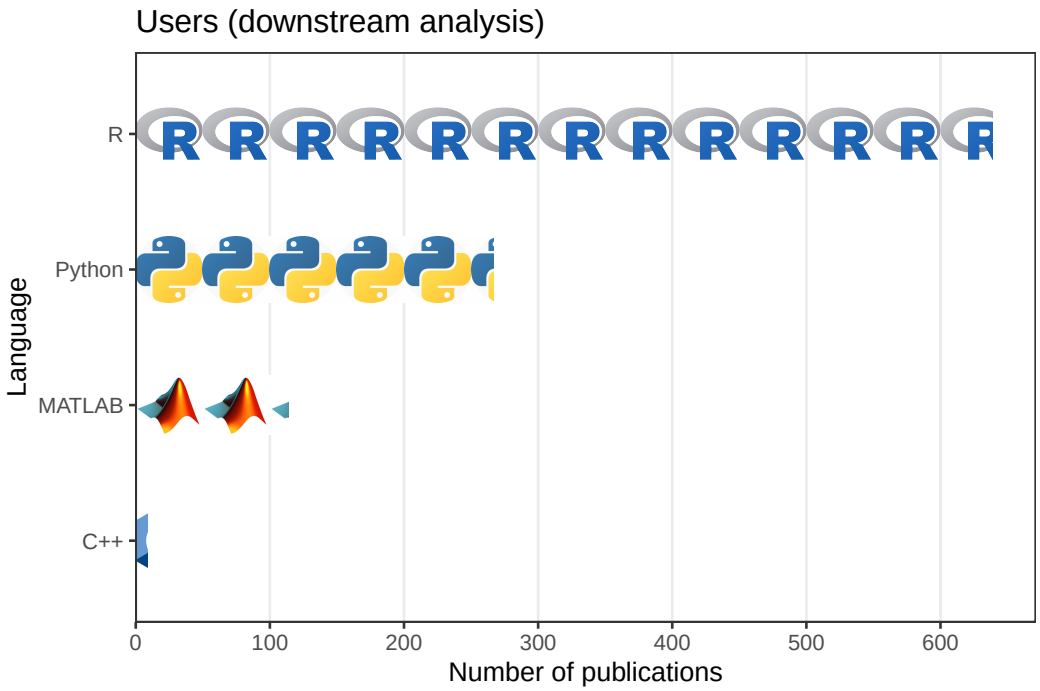

Figure 9.10: Number of publications for data collection using each of the 5 most popular programming languages for downstream data analysis.

reflect the broader cultural differences between the R and Python communities; the former caters more to the users and statisticians who do not specialize in computer science, while the latter caters more to developers and computer science specialists. MATLAB is not as commonly used for package development. While popularity of Python and R have grown (and some others such as Julia), the popularity of MATLAB seems more level (Figure 9.12). C and C++ are more common in package development than in downstream analysis, but are often used in conjunction with either R or Python or both as C and C++ are used for performance while R and Python are for user interface. With packages such as `reticulate`, `rpy2`, `basilisk`, `Rcpp`, and `Cython`, the most popular open source languages can be made interoperable to each other to some extent, making use of the best resources from each language.

We have also recorded whether the package is well documented and whether it's hosted on a public repository as a loose proxy of user friendliness and quality. Here "well documented" means at least all arguments of all functions exposed to the user are documented, though we consider it better when examples are included. Public repositories can to some extent indicate user friendliness and quality because the packages need to pass some sort of checking in order to be hosted on the repositories,

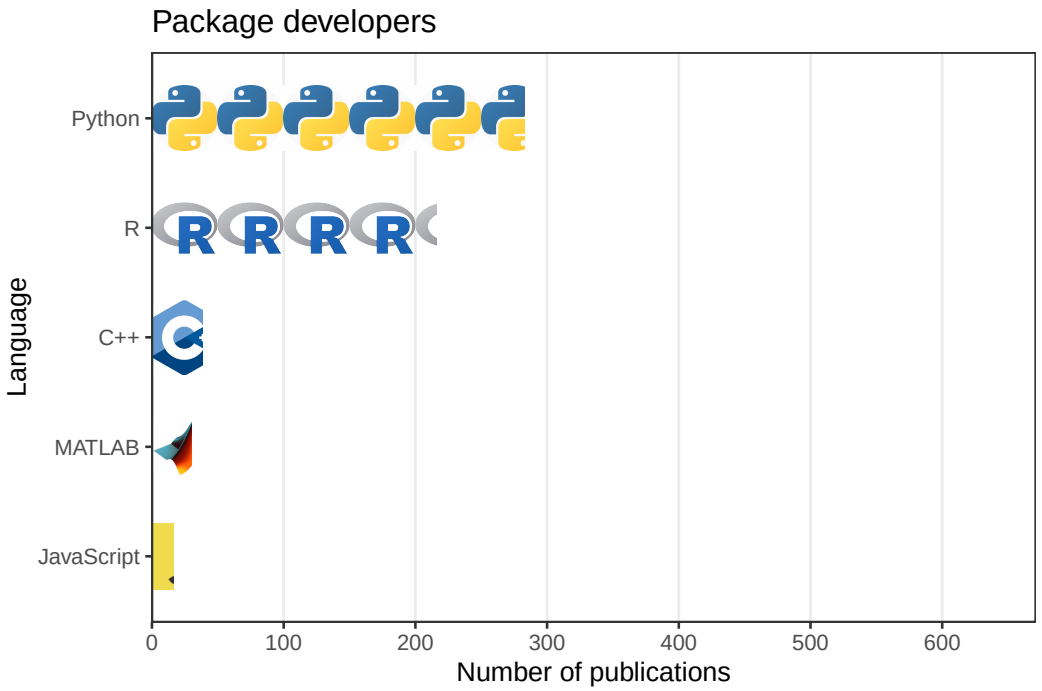

Figure 9.11: Number of publication for data analysis using each of the 5 most popular programming languages for package development. In this and the previous figure, each icon stands for 50 publications, and the x axes of both figures are aligned. Note that multiple programming languages can be used in one publication.

though some repositories, such as Bioconductor, have stricter standards than others. Moreover, installation of the package is easier when the package is on a public repository. A majority of Python packages and the vast majority of R and C++ packages are well documented, while many older MATLAB packages are not though more recent MATLAB packages are also mostly documented (Figure 9.12). Most packages are not on a public repository such as CRAN, Bioconductor, PyPI, and conda (Figure 9.13). For CRAN and especially Bioconductor, this might be due to the work required to meet standards of these repositories such as to pass automated checks, write documentation, examples, unit tests (Bioconductor), and vignettes (Bioconductor).

Some of the most popular categories of analyses (Figure 9.4) are reviewed in the rest of this section, arranged roughly in the order each task is performed in a data analysis workflow, from converting raw data to something more amenable to biological interpretations to forming biological hypotheses. The former is specific to certain types of techniques, and includes image processing (smFISH and ISS), spatial reconstruction (scRNA-seq and smFISH and ISS data that are not transcriptome

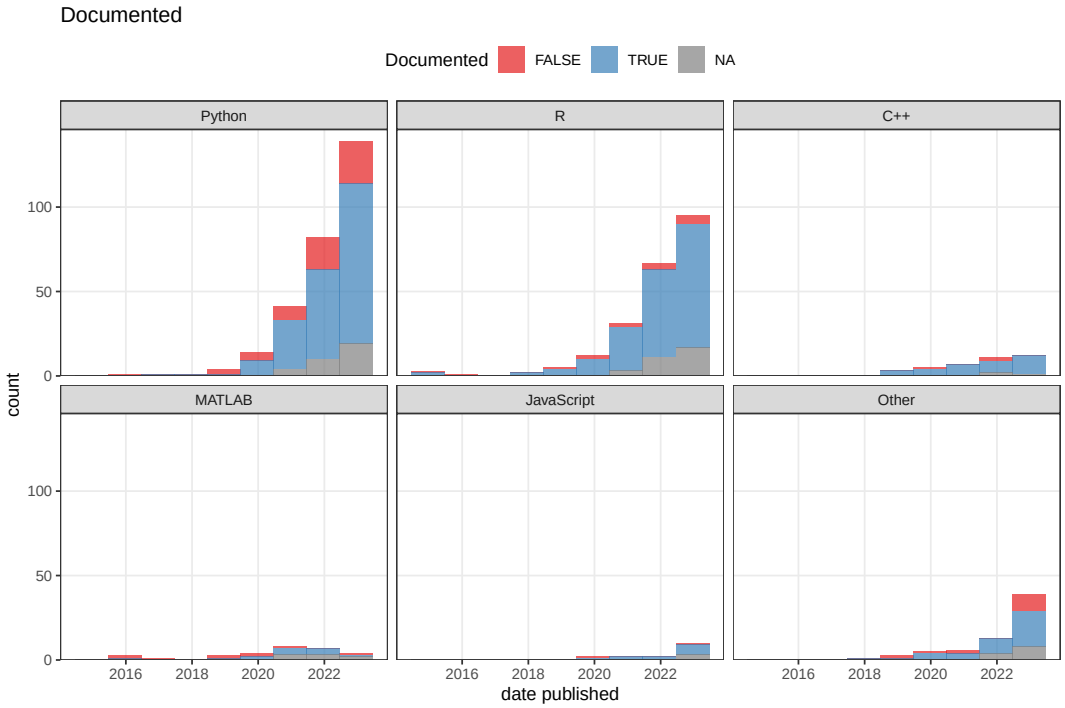

Figure 9.12: Among data analysis publications, the number of packages that are or are not well documented over time.

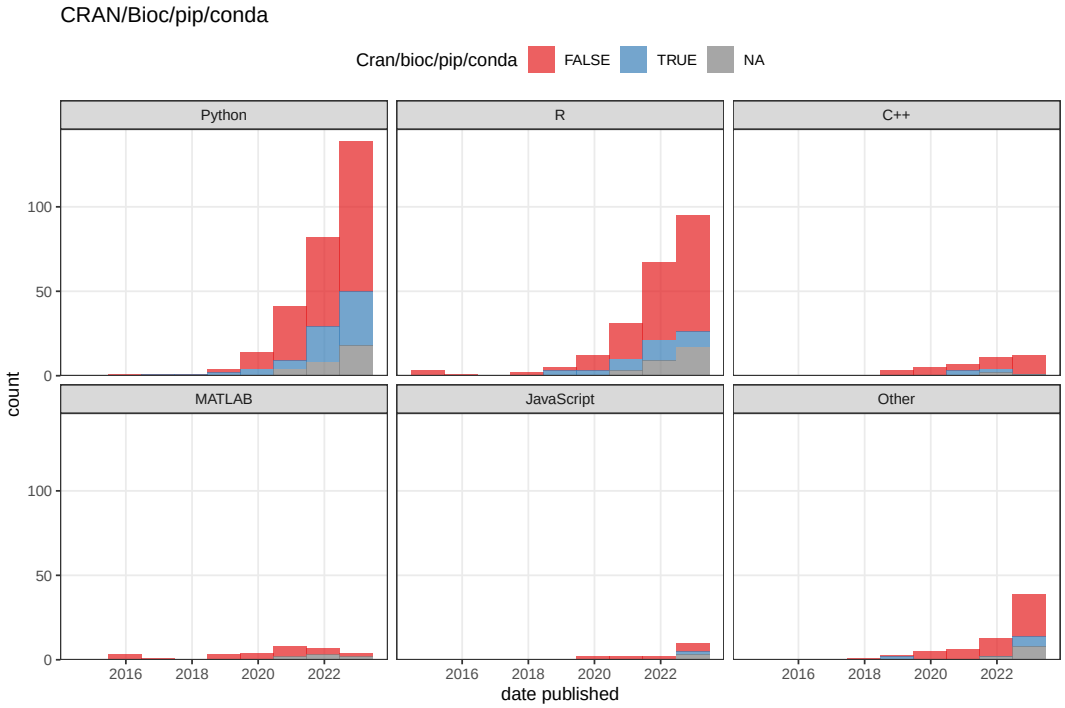

Figure 9.13: The number of packages that are or are not on a public repository such as CRAN, Bioconductor, PyPI, or conda over time. In both C and D, the bin width is 365 days. NA means the source code repository is not available.

wide), and cell type deconvolution (NGS barcoding data that are not single-cell resolution). The latter can largely be applied across types of techniques if given a gene by cell or spot matrix and cell or spot locations. Exceptions to the "largely" include analyses of subcellular transcript location which can only be applied to single molecule resolution data and spatial point process based methods which are more appropriate to model the cell or transcript locations rather than the fixed Visium spot locations. Each category will first be defined, and the common core principles will be summarized.

9.1 Preprocessing

By "preprocessing" we mean extracting information from raw data so common analysis methods can be applied. "Raw data" can mean any form of data, even if processed in some ways, that still needs to have information extracted for common analysis tasks to apply, such as PCA, clustering, and DE. Preprocessing for array based techniques that use NGS is similar to preprocessing for scRNA-seq. The same aligners can be used to align reads to the genome or pseudoalign to the transcriptome, and the spot barcodes can be demultiplexed just like in scRNA-seq; indeed, ST and Visium, the preprocessing pipelines ST Pipeline and Space Ranger wrap the STAR aligner. In addition, the transcript spots need to be aligned to the H&E image for visualization, interpretation, and using information from H&E for analyses. As microdissection based techniques also use NGS, preprocessing would not be very different from that of scRNA-seq or bulk RNA-seq data. However preprocessing of smFISH and ISS data is very different from that of NGS based data, and this would be the focus of this section.

The rawest data the user sees is images. As mentioned earlier, preprocessing of images was typically performed with poorly documented MATLAB code difficult to decipher by users. While some switched to Python recently, such as in MERlin for MERFISH, the preprocessing tool is still often specific to the technique of interest. The HybISS group has switched from MATLAB to Python and used `starfish` for spot detection and decoding, and pciSeq from this group has been reimplemented in Python (originally MATLAB) as well [1, 2]. Some groups used GUI based tools such as Fiji, ImageJ, and CellProfiler [3, 4, 5]. However, as the GUI based analyses are not recorded and shared or are manual, it is difficult to reproduce such analyses.

To provide a free, open source, and well-documented preprocessing tool applicable to data from multiple techniques, the Chan Zuckerberg Initiative developed the

Python package `starfish` implementing image registration, spot calling, barcode calling, cell segmentation, and etc. with classical image processing methods such as thresholding, image registration by translation, top hat filtering, Laplacian of Gaussian, watershed segmentation, and etc. While a good start, it's not clear how to apply `starfish` to multiple FOVs based on its tutorials. To improve `starfish`, another Python pipeline, SMART-Q was developed, with more modularity and improvements upon `starfish` such as additional parameter to mitigate over-segmentation (individual cell or nucleus broken into too many pieces) by watershed and integration with immunofluorescence images of marker genes [6]. However, SMART-Q was only demonstrated in RNAscope data without combinatorial barcoding, with one FOV at a time. Another such smFISH pipeline based on classical image processing is `dotdotdot`, which is written in MATLAB but the functions are well documented [7]. Again, `dotdotdot` was only demonstrated on RNAscope without combinatorial barcoding. There are other open source tools for one or more of the preprocessing steps, but are not meant to be a comprehensive pipeline. Below we review each step in preprocessing of smFISH and ISS raw data, how this was done in the original papers of datasets with classical image processing, and alternative and improved approaches such as ones based on deep learning or Bayesian statistics.

The packages mentioned in this section are summarized in the Table 9.1. The package names link to the code repo if available, and the titles link to the paper associated with the package. Each section in this chapter has a table like this. There are relevant packages not mentioned in this book; they can be found in the database.

Table 9.1: Packages mentioned for smFISH and ISS image processing

Name	Language	Title	Date published
corrFISH	MATLAB	Dense transcript profiling in single-cells by image correlation decoding	2016-06-06
graph-ISS	Python	Identification of spatial compartments in tissue from in situ sequencing data	2019-09-18

pciSeq	MATLAB	Probabilistic cell typing enables fine mapping of closely related cell types in situ	2019-11-18
SMART-Q	Python	SMART-Q: An Integrative Pipeline Quantifying Cell Type-Specific RNA Transcription	2020-04-29
spage2vec	Python	Spage2vec: Unsupervised detection of spatial gene expression constellations	2020-09-25
deepBlink	Python	deepBlink: Threshold-independent detection and localization of diffraction-limited spots	2020-12-15
ISTDECO	Python	ISTDECO: In Situ Transcriptomics Decoding by Deconvolution	2021-03-02
BarDensr	Python	BARcode DEmixing through Non-negative Spatial Regression (BarDensr)	2021-03-08
JSTA	Python; C	Joint cell segmentation and cell type annotation for spatial transcriptomics	2021-05-31
SSAM	Python; C++	Cell segmentation-free inference of cell types from in situ transcriptomics data	2021-06-10
Baysor	Julia	Bayesian segmentation of spatially resolved transcriptomics data	2021-10-14

Image registration

First, images of each FOV from different rounds of hybridization must be aligned; this is image registration. The images can be aligned to a reference of fiducial beads or DAPI staining, which is especially useful when "no fluorescence" is part of the barcode [8, 9]. If "no fluorescence" is not involved, then the reference can be a particular round of hybridization [10, 11]. Image registration is usually affine, i.e. images are translated, scaled, or rotated to match the reference, and often only translation is used. However, non-linear registration has been used in case the sample does not lie flat and chromatic aberration shifts spots in different channels [2].

Spot and barcode calling

Then the spots representing individual transcripts are identified (spot calling). The background of autofluorescence and non-specific hybridization is often removed by thresholding or top hat filtering, only preserving brighter pixels. Spots can be identified with multi-Gaussian fitting with fixed width, which can distinguish between partially overlapping spots [8], or tightened by Lucy-Richardson deconvolution [12], or by identifying local maxima in intensity after identifying potential spots with Laplacian of Gaussian [10, 11]. The spots can also be identified with deep learning. In Python package graph-ISS [13], a convolutional neural network (CNN) is pretrained on manually annotated candidate signal spots from another dataset, and probability that a new candidate obtained after top hat filtering and h-maxima transform is a signal is returned by the last softmax layer of the CNN. Another CNN based spot calling tool is deepBlink [14], which builds on the popular U-net architecture.

Once spots are called in each round of hybridization, spots that most likely to correspond to the same transcript are read as barcode and decoded to identify the gene encoded by the barcode (barcode calling). As image registration is imperfect, the spot coming from the same transcript may still be slightly shifted between rounds of hybridization. To identify the barcode from the rounds of hybridization, the spot in one round of hybridization is typically identified with a spot in another round if the spatial distance between the two is sufficiently small, such as less than between 1 and 3 pixels, or smaller than the distance to a barcode that contains error [10, 11, 15, 9].

In graph-ISS [13], spots identified from CNN from different rounds of hybridization are connected in a graph, with each spot in each round of hybridization a node and the edge weight decreases with increasing distance between spots across rounds up to a maximum distance. Edges connecting spots not from consecutive rounds are removed. The barcode is called by maximum flow of minimum costs between the sink and the source of the graph. Then a quality score is calculated for the barcode according to the CNN probability of spots and distance between spots from different rounds. Although graph-ISS was originally designed for ISS data, it might be adapted to seqFISH, HybISS, STARmap, and SCRINSHOT as well. However, for MERFISH and seqFISH+, in which a transcript may not have signal in some rounds of hybridization, graph-ISS would need to be altered. Alteration would also be required to decode STARmap's 2 base encoding.

For MERFISH specifically, transcript counts have been statistically modeled in the Rust package MERFISHtools, which takes errors in barcode calling into account [16]. While MERFISH's inbuilt error correction (HD4) accounts for 1 to 0 error, which is more common, 0 to 1 errors can still occur, and there are still barcodes with so many errors that they can't be matched to genes (dropout). The errors are modeled as a multinomial distribution with event probabilities as probabilities of identifying transcripts of a gene correctly with and without the inbuilt correction, misidentifying transcripts of a gene as those from each other gene with and without the inbuilt correction, and dropouts, with actual transcript counts, number of correct and incorrect identifications, and dropouts as latent variables to be estimated by Bayesian inference. The flat prior is used for now.

Computational methods to overcome optical crowding and to deconvolute spots were summarized in Section 7.2: corrFISH, BarDensr, and ISTDECO. The above mentioned spot calling methods all treat spot detection and decoding as separate tasks. In contrast, in both BarDensr and ISTDECO, the two related tasks are performed jointly.

Cell segmentation

To assign transcript spots to cells, the cells need to be segmented and spots within the segmented boundary of a cell must be assigned to that cell. For neurons, Nissl staining, which stains the cell body and dendrites but not axons, has been used for cell segmentation [10, 11]. Without Nissl staining, total poly-A staining can be used instead, and segmented with watershed transform, although poly-a staining concentrates in the cell body and misses cellular processes such as dendrites [12]. This misses some interesting biological information; dendrites can have different transcriptomes from the cell body of the same neuron, both *in vitro* and *in vivo* [17, 18, 19]. Cell segmentation can be done manually as automated methods may not be sufficiently reliable and would still require manual inspection and correction, or automated with machine learning models trained by manual segmentation of smaller number of cells such as the random forest model in Ilastik [11, 20] and CNN models such as DeepCell [21] and CellPose [22]. Watershed segmentation is more commonly used.

Without seeing the actual extent of the cell, the quality of manual segmentation is questionable, especially in regions with high cell density, thus limiting the performance of machine learning models. Sometimes problematic methods were used to

segment cells, such as 3D Voroni tessellation [10] and convex hull of Nissl staining based segmentation [11]. These are problematic because cells need not to take a convex shape so such segmentation may mis-assign transcripts from other cells, or to be conservative about mis-assigning transcripts from other cells, miss transcripts that in fact belong to the cell of interest. However, one study did specifically stain for membrane bound proteins for the actual extent of the plasma membrane and accurate cell segmentation [20].

To address the challenges of cell segmentation, segmentation methods utilizing scRNA-seq data with annotated cell types have been developed recently. One such method is Python package JSTA [23], in which a deep neural network (DNN) learns a segmentation and cell type annotation using the information from a scRNA-seq reference with cell type annotations. First, watershed is used for an initial cell segmentation, both MERFISH and scRNA-seq data are scaled and centered. Then a DNN is trained on the scRNA-seq data to predict cell type from gene expression. Then a separate DNN is trained to refine the cell boundaries iteratively with expectation maximization (EM): The cell type classifier is applied on the watershed segmented MERFISH data to classify putative cells (E). Then a random subset of the pixels are used to train the pixel classifier, maximizing a loss function comparing the new pixel cell type probabilities to the initial/previous assignment (M). The new cell type probabilities are then scaled per pixel according to distance to nuclei. Only probabilities of cell types of neighboring cells are kept and the other cell types are assigned probability 0. The new cell type probabilities of each pixel is then used as event probabilities of a multinomial distribution and randomly assign a new cell type label to the pixel. Then the new cell type assignment to pixels is used to train the pixel classifier again, until the cell type assignments converge. This may refine boundaries between neighboring cells of different types, and the initial watershed boundaries are kept for neighboring cells of the same type. A problem with this package is that inhomogeneous transcript localization is not taken into account.

Alternatives to cell segmentation

Due to the challenges in accurate cell segmentation, some analysis methods did away with cell segmentation altogether, directly using the transcript locations. In the Julia package Baysor [24], based on Markov random field (MRF), which encourages nearby transcripts to take the same label. A spatial neighborhood graph is constructed with Delaunay triangulation with each transcript as a node. The

probability of each transcript taking each label is modeled with a MRF and initial edge weights decrease with distance. This package first distinguishes between intracellular transcripts and extracellular background. Then it can also assign transcripts to cell types without cell segmentation, with a scRNA-seq reference with cell type annotations; as locations of the transcripts are known, this amounts to annotating tissue regions with cell types. It can also segment cells, with existing segmentation and staining (e.g. Nissl, DAPI, and poly-A) as priors. Cell segmentation can also be informed by cell type labels, so transcripts from different cell types are not assigned to the same cells. Each of the three functionalities, identifying intracellular transcripts, cell type annotation of transcripts, and cell segmentation, is based on a different MRF model. The parameters of the model, such as edge weights, labels of other transcripts, and etc. are estimated with EM. The drawbacks of this package are that its current implementation is limited to 2D and it does not take inhomogeneous subcellular transcript localization into account.

Besides cell type annotation of transcripts based on MRF, another segmentation-free method is also described in the Baysor paper [24], in which the k nearest neighbors of each transcript are taken to be a pseudo-cell and analyzed by standard scRNA-seq data analysis methods such as clustering, PCA, and UMAP. For ISS, transcripts can be probabilistically assigned to cells and cells to cell types, with pciSeq [2]. Briefly, spatial locations of transcripts are modeled by a Poisson point process whose intensity is scaled by a term following Gamma distribution to give the negative binomial distribution of transcript counts in cells. The intensity for each gene and each cell is also informed by distance between transcripts and nucleus centroids (from DAPI), scRNA-seq data of the cell type this cell belongs to, and the detection efficiency of ISS. The data consists of locations of transcripts and the genes they come from. The unknown parameters, such as probability of each transcript to come from each cell and each cell from each cell type, are estimated by variational Bayesian inference. Cell types and spatial domains can also be identified without scRNA-seq cell type annotations as well.

In the Python package SSAM [25], transcript density is first estimated with Gaussian kernel density, which is then projected into a square lattice. Local maxima of transcript density are taken as pseudo-cells and clustered to infer *de novo* cell types. Then tissue domains are identified by clustering sliding windows of spatial cell type maps. Tissue domains can also be identified without appealing to cell types.

In the Python package spage2vec [26], graphs are constructed by connecting each

transcript spot to its neighbors within a certain distance such that at 97% of all transcript spot are connected to at least one neighbor. Then the transcript spots with these graphs are projected by a graph neural network (GNN) into a 50 dimensional space which is informed by the graphs and thus local neighborhoods of transcripts. The transcript spots in the 50 dimensional space can then be clustered or projected to 2 or 3 dimensions with UMAP to show tissue domains.

9.2 Exploratory data analysis

Table 9.2: Packages mentioned for EDA

Name	Language	Title	Date published
Spaniel	R	Spaniel: analysis and interactive sharing of Spatial Transcriptomics data	2019-05-05
Seurat3	R	Comprehensive Integration of Single-Cell Data	2019-06-13
SpatialCPie	R	SpatialCPie: an R/Bioconductor package for spatial transcriptomics cluster evaluation	2020-04-29
STUtility	R	Seamless integration of image and molecular analysis for spatial transcriptomics workflows	2020-07-16
SPATA	R	Inferring spatially transient gene expression pattern from spatial transcriptomic studies	2020-10-21
Giotto	R	Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data	2021-03-08
Squidpy	Python	Squidpy: a scalable framework for spatial single-cell analysis	2022-01-31
SpatialExperiment	R	SpatialExperiment: infrastructure for spatially resolved transcriptomics data in R using Bioconductor	2022-04-28

After data preprocessing, as described above, for array or microdissection based

data, we get a gene count matrix with locations of voxels, and for smFISH and ISS based data, we get locations of transcripts, and if cell segmentation is performed, a gene count matrix and cell boundaries as well. For scRNA-seq, Seurat [27], scanpy, and packages surrounding SingleCellExperiment on Bioconductor such as `scran` and `scater` implement further preprocessing of the gene count matrix, such as data normalization and scaling, as well as basic EDA methods to inspect and create an overview of the data, such as quality control (QC), data visualization, finding highly variable genes, dimension reduction, and clustering, and have user friendly tutorials, consistent user interface, and decent documentation. Such integrative EDA packages, as well as more specialized data visualization packages, have emerged for spatial transcriptomics as well, and are reviewed in this section.

In practice, spatial transcriptomics data is often analyzed with standard scRNA-seq analysis at the EDA stage, with one or more of PCA, tSNE, UMAP, clustering cells or spots, and finding marker genes for clusters, and differential expression (DE) between case and control [10, 12, 28, 29, 30]. For ST and Visium, the data is also often normalized like in scRNA-seq with CPM or classical Seurat log normalization and scaling [29, 31, 30]. Seurat also implements data integration, which has been used to transfer cell type labels from scRNA-seq to Visium for cell type deconvolution [32], and can potentially be used to impute gene expression in non-transcriptome wide spatial data from scRNA-seq (discussed in Section 9.3). Then the clusters, marker genes, and genes of interest from scRNA-seq are often visualized within spatial context, and some studies proceed to other analyses that utilize the spatial information. Due to the relevance of scRNA-seq data normalization, EDA, and data integration to spatial data, the existing scRNA-seq ecosystems of Seurat, scanpy (spatial part in Squidpy [33]), and SingleCellExperiment (spatial part in SpatialExperiment [34]) are adapting to the rise of spatial transcriptomics, with new data structures, visualization of gene expression and cell metadata (e.g. total UMI counts, cluster, and cell type) on the spatial coordinates, with H&E as background for ST and Visium, and perhaps other spatial functionalities such as spatial neighborhood graphs and spatially variable genes.

There are other EDA packages not originating from an existing scRNA-seq EDA ecosystem as well. R packages Giotto [35], STUtility [36], and SPATA [37] not only support basic QC and EDA functionalities like those in Seurat, but also spatial analyses not supported by Seurat. These packages are well documented, but are not (yet?) on CRAN or Bioconductor.

Giotto has two main parts: Giotto Analyzer and Giotto Viewer. Besides basic Seurat functionalities and spatial data visualization, Giotto Analyzer implements several types of spatial analyses to be reviewed in more detail in the rest of this section: cell type enrichment in spatial data without single-cell resolution, identifying spatially variable genes, gene co-expression patterns, cellular neighborhoods, interactions between cell types and ligand-receptor pairs in such interactions, and genes whose expression is associated with cell type interactions. However, the methods implemented in Giotto tend to have simpler principles than those of more specialized packages for each of the above tasks, such as hypergeometric test for cell type enrichment and spatially coherent genes, though Giotto wraps specialized packages such as SpatialDE [38], trendsceek [39] for spatially variable genes, and smfishhmrf [40] to identify spatial cellular neighborhoods. Giotto Viewer provides interactive visualization of the data. As Giotto uses its own object class to store data, interoperability with other single-cell and spatial software becomes more challenging given the popularity of Seurat and SingleCellExperiment.

In contrast, STUtility develops upon the Seurat class, so is interoperable with other Seurat functionalities. STUtility is specific to ST and Visium, while Giotto applies to all spatial technologies with cell or spot level data. Beyond Seurat, STUtility enables masking the array to remove spots outside the tissue, alignment of multiple sections, manual annotation and alignment with `shiny`, visualization of the aligned sections in 3D, finding neighbors of spots of a given type, and using NMF to identify archetypal gene expression patterns. While Giotto and STUtility might not have the most sophisticated spatial analysis methods, their main advantage is akin to that of Seurat and SingleCellExperiment, namely that multiple analysis tasks, often with a variety of algorithms for each task, can be done with the same object class and user interface, saving the time and trouble on learning new syntax and converting objects to new classes.

SPAtial Transcriptomic Analysis (SPATA), while implementing its own class, uses Seurat for data normalization and dimension reduction. SPATA also implements functions to visualize spatial data and a `shiny` app for not only interactive data visualization but also manually setting spatial trajectories and annotation of spatial regions. It also wraps Monocle 3 [41] for pseudotime analysis and SPARK [42] for finding spatially variable genes. In addition, SPATA implements its own method of finding spatially variable genes, reviewed in Section 9.5.

Some R packages have also been written for specific visualization tasks, but not

the entire EDA process. Spaniel is a package that builds on Seurat and SingleCellExperiment for interoperability and implements QC plots that help the user to remove ST or Visium spots outside the tissue. However, Spaniel's main difference from STUtility is that Spaniel can create a `shiny` app for interactive visualization and exploration of the data. While this may make Spaniel sound unremarkable, it was written about a year before Seurat supported spatial data. Another specialized package is SpatialCPie [43], which also uses `shiny` for interactive visualization. SpatialCPie cluster ST or Visium data at multiple resolutions and plots a graph showing how clusters from one resolution relates to those from other resolutions. It also plots a pie chart at each ST or Visium spot, on top of an H&E background, showing similarity of each spot to each cluster, to give a more nuanced view than simply coloring the spots by cluster. Both packages are on Bioconductor.

9.3 Spatial reconstruction of scRNA-seq data

It may be fair to say that the holy grail of spatial transcriptomics is to profile the whole transcriptome at single-cell resolution and without dropouts. We have already seen that, with seqFISH+ and ExM-MERFISH, this goal seems to possibly be within reach. However, the goal may be further than is seemingly the case, as the smFISH-based techniques are still not generally applied to more than a few dozens to a few hundreds of genes, in the order of 10,000 cells (Figure 7.23, Figure 7.24), which only covers a small area of tissue. Meanwhile, techniques without single-cell resolution and with lower detection efficiency but which can cover large swaths of tissue have grown in popularity (Figure 7.37). Hence spatial transcriptomics has not supplanted scRNA-seq—which has also grown tremendously in popularity in recent years [44]–but remains a complement. Spatial data that is not transcriptome wide can be complemented by scRNA-seq for information of other genes; this section reviews computational methods that map cells from scRNA-seq to spatial locations with a small panel of landmark genes and/or to impute gene expression not profiled by the spatial reference in space, or in short spatial reconstruction of scRNA-seq data. These are the most common types of data analysis(Figure 9.4). The two tasks are related but distinct, as when cells from scRNA-seq are mapped to spatial locations, spatial patterns of the genes expressed in the cells are also predicted. However, gene expression can also be predicted at spatial locations without mapping cells to the locations. Spatial data that does not have single-cell resolution can be complemented by scRNA-seq for cell type deconvolution of the spots (Section 9.4). In turn, spatial data complements scRNA-seq with spatial information such as gene

expression patterns and cell neighborhoods.

Attempts at spatial reconstruction of single-cell data date back to 2014, when growth in the popularity of scRNA-seq started to pick up pace [44]. Early (2014-2017) methods tend to fall in three categories: direct dimension reduction with PCA, ad hoc scoring, and pseudotime projected into space. The first two have been by and large abandoned due to their limitations, and the third isn't commonly used. Another category is generative modeling, which we consider intermediate due to its early origin and lasting legacy as some later methods involve more sophisticated generative modeling. Later (2018-present) methods commonly involve a lower dimensional latent space shared by the scRNA-seq and the spatial data, and many different approaches have been tried to obtain the shared latent space and project it back into the higher dimensional space of gene expression. However, other principles were used as well, such as optimal transport, nonlinear direct dimension reduction, black box machine learning, mixture of experts model, and etc.

Table 9.3: Packages mentioned for spatial reconstruction of scRNA-seq data

Name	Language	Title	Date published
Seurat1	R	Spatial reconstruction of single-cell gene expression data	2015-04-13
DistMap	R	The Drosophila embryo at single-cell transcriptome resolution	2017-10-13
gimVI	Python	A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements	2019-05-06
Seurat3	R	Comprehensive Integration of Single-Cell Data	2019-06-13
LIGER	R; C++	Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity	2019-06-13

SPRESSO	R; Python	Novel computational model of gastrula morphogenesis to identify spatial discriminator genes by self-organizing map (SOM) clustering	2019-08-29
Harmony	R; C; C++	Fast, sensitive and accurate integration of single-cell data with Harmony	2019-11-18
novoSpaRc	Python	Gene expression cartography	2019-11-20
sstGPLVM	Python	A Bayesian nonparametric semi-supervised model for integration of multiple single-cell experiments	2020-01-21
SpaOTsc	Python	Inferring spatial and signaling relationships between cells from single-cell transcriptomic data	2020-04-29
st_analysis	Python	Molecular atlas of the adult mouse brain	2020-06-26
GLISS	NA	Integrative Spatial Single-cell Analysis with Graph-based Feature Learning	2020-08-13
SpaGE	Python	SpaGE: Spatial Gene Enhancement using scRNA-seq	2020-09-21
FIST	MATLAB	Imputation of Spatially-resolved Transcriptomes by Graph-regularized Tensor Completion	2021-04-07
LIGER	R; C++	Iterative single-cell multi-omic integration using online learning	2021-04-19
Tangram	Python	Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram	2021-10-28

Direct dimension reduction

As already mentioned in our summary of Puzzle Imaging, spatial reconstruction of dissociated tissue can be considered a dimension reduction problem. Here with

scRNA-seq, the high-dimensional gene expression data is directly projected to 1 to 3 dimensions that correspond to the spatial dimensions.

One of the earliest reconstruction methods (2014) maps single-cell qPCR data onto a sphere that mimics the developing mouse otocyst [45]. Ninety six genes were profiled with qPCR in single-cells, and the gene expression profiles were projected to the first 3 principal components (PCs), which are then projected onto the surface of a sphere. The sphere is oriented on the dorsal-ventral (DV), anterior-posterior (AP), and left-right (LR) axes by expression of marker genes known to be expressed in one end of those axes. At least for the otocyst, this approach seemed to recapitulate expression patterns of many genes, at least qualitatively, at the resolution of octants. This approach was later adapted to reconstruct the human [46] and mouse [47] blastocysts. A one-dimensional version of this approach was also adapted to spatially reconstruct cells from the organ of Corti along the apical and basal axis, though the PCA was performed only on DE genes between apical and basal cells and 2 PCs were projected to 1 dimension [48].

Direct dimension reduction is still used after 2018, with dimension reductions other than PCA. Another form of dimension reduction for spatial reconstruction is the self-organinzing map (SOM) as in the package SPRESSO [49]. The Geo-seq mid-gastrula mouse embryo data [50] was reconstructed in 3D with genes selected from GO terms; 18 genes selected from a few GO terms could place all microdissected samples into the correct AP/LR quadrant with SOM. However, such genes were found by checking the SOM projections from thousands of GO combinations against the Geo-seq ground truth and may not apply to other biological systems. Also, the spatial reconstruction along the DV axis was not checked, though in Geo-seq, the samples were microdissected along the DV axis with a cryotome in addition to dissection into AP/LR quadrants with LCM.

A more recent, graph based dimension reduction is GLISS [51]. After using a Laplacian score based method to identify landmark genes from spatial data (to be reviewed in Section 9.5), a graph is constructed for the scRNA-seq data based on similarity in expression profiles of the landmark genes among cells as a proxy to spatial locations. With this graph, a new set of genes whose expression depend on the structure of the graph, or spatially variable genes, are identified, and added to the landmark genes. A new similarity graph is then constructed with both the landmark genes and spatially variable genes, and the dimension reduction is the eigenvectors of the graph Laplacian of this graph, starting from the second

eigenvector. One-dimensional projection would be the second eigenvector. Two-dimensional projection would be the second and third, and so on.

Ligand-receptor (L-R) pairs have also been used for direct dimension reduction, in CSOmap [52]. Expression of L-R pairs in scRNA-seq cells is used to construct a cell-cell affinity matrix, with higher affinity meaning that two cells are more likely to be close to each other. Then an algorithm similar to tSNE is used to project the affinity matrix into 3 dimensions, corresponding to the physial dimensions. The Kullback–Leibler (KL) divergence between the affinity and probability of the two cells to be neighbors is minimized, with constraints of the minimum physical size of the cell and the amount of space available.

Ad hoc scoring

The methods above tend to only capture simple spatial patterns with simple gradients along axes, or have low resolution that is effectively restricted to octants or quadrants. More complex patterns with higher resolution can be reconstructed qualitatively with some score that measures similarity between each cell in scRNA-seq and each location in a spatial reference for the genes present in both datasets and favors genes more specific to a subset of cells. The spatial pattern of the score is the predicted gene expression pattern. As the score is qualitative and does not utilize statistical modeling of the data, this is called ad hoc scoring. The spatial reference is FISH (not smFISH) data of a panel of genes, with images for different genes registered onto a common coordinate system. As FISH is not very quantitative, both the spatial and the scRNA-seq data are binarized into "on" and "off" for each gene, and the predicted gene expression patterns based on the score is binarized as well since the score is only qualitative. Such approach is simple to implement, but the binarization misses quantitative nuances of gene expression patterns.

Ad hoc scoring has been used in *Platynereis dumerilii* brains; the FISH atlas was broken into voxels 3 μm on each side, smaller than the average single-cell, and 98 landmark genes in the atlas used to predict patterns of other genes in scRNA-seq with a score [53]. A different method, DistMap, uses a score based on Matthew correlation coefficient (MCC) was used to soft assign cells from scRNA-seq to locations in the BDTNP atlas with 84 landmark genes and to predict expression patterns of the other genes [54]. The latter method inspired the DREAM Single-cell transcriptomics challenge in 2018 [55], a competition in which participants select the most informative genes and predict cell locations with 60, 40, and 20 of the

84 BDTNP landmark genes. At least some participating teams adapted the scoring method used in the original DistMap after selecting genes with their own methods [56, 57].

Generative models

Many areas in spatial transcriptomics data analysis describe the data with a plausible statistical model and fit such a model to the data. Generative models have several advantages. First, uncertainties in parameter estimates and model predictions can be computed. Second, the model is more explainable, i.e. that humans may understand contributions of variables to the fitted model. Explainability plays an important role in models identifying spatially variable genes. As already mentioned, some of the segmentation-free smFISH or ISS analysis packages, such as pciSeq, rely on generative models. Generative models are used for spatial reconstruction of scRNA-seq data as well.

The popular scRNA-seq EDA package Seurat originated from spatial reconstruction of scRNA-seq data in 2015, to map cells from scRNA-seq to a WMISH reference with 47 landmark genes [58]. The WMISH images were mostly obtained from ZFIN, and divided into 128 bins, which was then collapsed into 64 due to LR symmetry. As WMISH is not very quantitative, the WMISH reference was binarized. Due to the sparsity of scRNA-seq data, the normalized scRNA-seq data was smoothed. Then a mixture of 2 Gaussian distributions was fitted to each gene, for the "on" and the "off" states. With such distributions, the posterior probability that each cell comes from each bin can be calculated with the probability that the cell is "on" or "off" like in the bin for the 47 genes, although cells can very well have intermediate and more nuanced gene expression. The spatial centroid of each cell is the center of mass of the spatial map of the posterior probabilities. So far, the landmark genes have been assumed to be independent, which is unrealistic. Centroids that are close to actual bins are then used to calculate a covariance matrix of a subset of the landmark genes for each bin, with which the Gaussian mixture models and posterior probabilities are updated. While this model seems reasonable, it is no longer used, likely because of the advances in highly multiplexed smFISH and ISS that produced quantitative spatial references that do no need binarization for some tissues, especially the mouse brain. Nevertheless, the scRNA-seq part of Seurat lived on. As already mentioned, WMISH or ISH atlases are the only spatial transcriptomics resources available for some biological systems and most of the atlases are not transcriptome wide, so this method can still be useful.

A different generative model was used to map scRNA-seq cells to a smFISH atlas in the mouse liver [59]. Six marker genes known to be patterned in the portal-central axis of the hepatic lobule were profiled with smFISH. Then the smFISH data was binned into 9 zones and normalized, and each gene in each zone was modeled with a gamma distribution, which was then multiplied by coefficients correcting for the fact that only part of the cell is in the tissue section for the λ of a Poisson distribution to form a negative binomial distribution. The negative binomial distribution was sampled and normalized for the whole cells in scRNA-seq and proportion of UMIs from the gene of interest, which would approximate the distribution of a cell in each zone having expression levels of the gene of interest. The prior probability of a hepatocyte originating from each zone seems to be the relative area of the concentric ring that is each zone, centered on the central vein. With the prior and the sampled distribution of expression of marker genes, the posterior probability of each cell from each zone can be calculated with Bayes rule. To impute expression of genes other than the 6 markers in each zone, the gene count matrix is multiplied to the posterior probability matrix (after weighing the probabilities). Here the 6 markers are assumed to be independent, which might not be realistic. The same approach is still used by the same lab for more recent liver datasets [60, 61], although we are unaware of its use outside that lab.

Some of the shared latent space methods are based on generative models as well, with the latent space as part of the model. In gimVI [62], which is adapted from scVI specifically to impute gene expression in space by integrating spatial and scRNA-seq data, gene expression in scRNA-seq is modeled with the negative binomial (NB) or zero inflated negative binomial (ZINB) distribution, and the spatial data is modeled with the Poisson or NB distribution (depending on the technology and dataset). The scRNA-seq and spatial data are modeled as coming from a shared latent lower dimensional space, which is decoded back to the higher dimensional gene expression space by a neural network to capture nonlinear structures as part of the mean parameters of the NB, ZINB, or Poisson distributions. The latent space is estimated when the model is fitted with variational Bayesian inference. To impute gene expression in space, the latent space is sampled and passed through the decoding neural network to get the mean parameters of the gene expression distributions for spatial data.

Another generative model with a shared latent space is semi-supervised t-distributed Gaussian process latent variable model (sstGPLVM) [63]. The scRNA-seq or spatial

data is modeled as coming from a noisy sample in high dimension from a lower dimensional shared latent space. The latent space can be concatenated to fixed covariates such as batch, technology used to collect data, spatial coordinates, and etc. and is estimated with black box variational inference. Missing data in gene expression and covariates can be estimated from the latent space, thus enabling mapping scRNA-seq cells to spatial coordinates and imputing gene expression, and the latent space can be collapsed across a covariate to remove its effect. The latent space has a Gaussian prior with identity variance. The prior of the high-dimensional noiseless space is a Gaussian process with covariance between cells defined by a kernel that is a weighted sum of Matern 1/2 and Gaussian kernels to allow for a non-smooth manifold that better represents data. The input to the kernel is a weighted sum (length scales of kernel) of l1 distance between the cells in the latent space (including the covariates). The noise added to the noiseless high-dimensional space to model actual data is a heavy tailed Student's t distribution, to account for overdispersion and non-Gaussian distribution of the data. This method is not specifically designed for spatial data, but can be used to integrate different scRNA-seq datasets as well.

Shared latent space

There are some additional methods that project scRNA-seq and spatial data into a shared latent space to impute gene expression in space but without generative modeling. Some of them are designed for data integration in general, but included here the authors demonstrated integration of scRNA-seq and spatial data, seeming to intend their packages for such usage.

In version 3 or later of Seurat [27], the scRNA-seq and spatial datasets are projected into a shared latent space by canonical correlation analysis (CCA), which finds a low-dimensional space that maximizes correlation between the two dataset, or by projecting one dataset into a low-dimensional PCA space of the other dataset. Then anchor cells are identified, as cells in the two datasets with sufficient shared neighborhood, and the weight of each anchor on each cell in the spatial dataset is calculated by ad hoc scoring favoring closeness in the latent space and more similar shared neighborhood to the anchor. Gene expression is then simply transferred from scRNA-seq to spatial data by multiplying the normalized gene count matrix of genes absent from the spatial data in scRNA-seq with the anchor weight matrix.

LIGER [64] is a different data integration method, of which a Seurat wrapper has

been implemented. The latent space is inferred by integrative NMF, which finds a set of factors unique to the scRNA-seq or the spatial dataset, and a set of factors shared by both. Gene expression is imputed in spatial data by averaging the expression of genes of interest in the k (50) nearest neighbors (kNN) from the scRNA-seq data in the space spanned by the shared factors.

In SpaGE [65], a common latent space is inferred as such: gene shared by the spatial dataset and scRNA-seq are used to do PCA independently for the two datasets. Then the cosine similarity matrix of the PCs of the two dataset is passed to singular value decomposition (SVD). Then the left and and right singular vectors are used to align the PCs to a common latent space of principal vectors. The original data is projected into the space spanned by the principal vectors of the scRNA-seq data. Then kNN is used to project gene expression from scRNA-seq to spatial data.

In Harmony [66], the data, with different batches, is first PCA projected. Then the PCA projection is clustered with an altered k-means clustering algorithm that assigns cells probabilistically to clusters and maximizes diversity in batches in each cluster. Then the batch correction is found by mixture of expert model. In each cluster, the PCA projection is modeled by a linear combination of variables in the design matrix (containing batch information), with an intercept term for batch free variation in each cluster. The batch correction term is a weighted sum of the linear model predictions excluding the intercept term, weighted by the probabilistic assignment of each cell to each cluster. Then the batch correction term is subtracted from the original PCA projection. The clustering and correction are repeated until convergence. This way, the cells from scRNA-seq and spatial data are aligned in a common latent space. Then gene expression is imputed in spatial data with kNN.

Other principles

Approaches that do not fall into the categories reviewed above are reviewed in this subsection, including projecting pseudotime into space, black box machine learning, and optimal transport.

In some biological systems, cell differentiation corresponds to physical locations of the cells, so pseudotime, which supposedly arranges cells along differentiation trajectories, have been mapped to space, thus placing dissociated cells in space. For instance, in the bone growth place, cells at different stages of differentiation are physically arranged along the length of the bone in a cylinder, so the pseudotime trajectory of the cells was simply warped into a straight line for spatial reconstruction

[67]. Similarly, in Drosophila larva, cell differentiation corresponds to the proximal-distal axis in the antenna disk and the AP axis in the eye disk, so cells from both scRNA-seq and scATAC-seq were binned according to pseudotime and assigned to the corresponding bins in the eye-antenna disk [68]. However, this would not work in tissues without such neat correspondence, such as the Drosophila embryo, in which some genes are expressed in periodic patterns to specify segments.

Deep learning libraries such as PyTorch also made it more effective to predict locations for scRNA-seq cells without a pre-conceived statistical model of the data. For instance, after data normalization and batch correction, a deep neural network can be trained on ST data with annotations of spatial regions to predict spatial regions for scRNA-seq data [69]. In addition, PyTorch's gradient-based optimization has been used to probabilistically map scRNA-seq cells to spatial locations in Tangram [70]. The spatial reference is voxelated, and a mapping matrix of probability of each cell mapping to each voxel is inferred by minimizing KL divergence between mapped and actual cell density in each voxel and favoring stronger correlation between mapped data and the spatial reference in expression of each gene across voxels and gene expression profiles of each voxel.

Thus far, the reconstruction methods do not take spatial autocorrelation—i.e. that cells physically closer to each other are more likely to have more similar gene expression profiles—in the spatial data into account. Optimal transport, i.e. finding a way to transport a pile of dirt from one place to others with minimum cost, has been used to exploit spatial autocorrelation to map scRNA-seq cells to spatial locations. In novosparc [71], neighborhood graphs are constructed for scRNA-seq in gene expression space and for spatial reference data in physical space. Then assuming spatial autocorrelation, optimal transport is used to place cells in locations to make the two graphs match. This can be done without gene expression data in the spatial grid, but can be improved with a spatial gene expression reference. In SpaOTsc [72], first an optimal transport plan from scRNA-seq cells to spatial locations is inferred with gene expression dissimilarity matrices between scRNA-seq cells and between cells and locations and a spatial distance matrix between spatial locations. Then a spatial distance matrix for scRNA-seq cells is imputed based on that optimal transport plan. The plan can also be used to impute gene expression in space. SpaOTsc also uses optimal transport to infer cell-cell interaction, to be reviewed in the Cell-cell Interaction section. A drawback of this kind of method is that because different cell types can mix in the same spatial neighborhood, such as hepatocytes

and Kupffer cells in the liver, spatial autocorrelation is not absolute.

Spatial autocorrelation can also be utilized without optimal transport, but with tensor completion in Canonical Polyadic Decomposition (CPD) form as in FIST [73]. The spatial data can be viewed as a three-dimensional tensor, with the x and y coordinates and gene expression at each location ((or 4 with z coordinate). CPD is used to improve computational efficiency. In CPD, the tensor is approximated with a sum of rank 1 tensors, i.e. cross products of 3 vectors, one for each dimension. This decomposition, with extra dimensions for unknown gene expressions, is found by minimizing the difference between the reconstructed tensor with the existing tensor for known genes and by favoring spatial autocorrelation of gene expression on a neighborhood graph and favoring similarity of expression of genes with similar functions in a protein-protein interaction graph.

9.4 Cell type deconvolution

There is another aspect to how spatial and scRNA-seq data complement each other. In array based techniques that do not have single-cell resolution, the cell type composition of each spot can be estimated with scRNA-seq data. Perhaps because of the increasing popularity of ST and Visium, several cell type deconvolution methods have been developed in the past year, falling into four categories: negative binomial models, packages built upon linear models but without negative binomial, topic modeling, and packages not explicitly using statistical modeling. While any tool designed for cell type deconvolution of bulk RNA-seq data can be used, this section specifically focuses on cell type deconvolution tools designed with spatial data in mind.

Table 9.4: Packages mentioned for cell type deconvolution

Name	Language	Title	Date published
NMFreg	Python; MATLAB	Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution	2019-03-29
Seurat3	R	Comprehensive Integration of Single-Cell Data	2019-06-13
stereoscope	Python	Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography	2020-10-09

DSTG	Python	DSTG: Deconvoluting Spatial Transcriptomics Data through Graph-based Artificial Intelligence	2021-01-22
SPOTlight	R	SPOTlight:Seeded NMF regression to Deconvolute Spatial Transcriptomics Spots with Single-Cell Transcriptomes	2021-02-05
RCTD	R	Robust decomposition of cell type mixtures in spatial transcriptomics	2021-02-18
Giotto	R	Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data	2021-03-08
AdRoit	R	AdRoit is an accurate and robust method to infer complex transcriptome composition	2021-10-22
Tangram	Python	Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram	2021-10-28
SpatialDecon	R	Advances in mixed cell deconvolution enable quantification of cell types in spatially-resolved gene expression data	2022-01-19
STRIDE	Python	STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing	2022-03-07
DestVI	Python	DestVI identifies continuums of cell types in spatial transcriptomics data	2022-04-21
STdeconvolve	R	Reference-free cell-type deconvolution of pixel-resolution spatially resolved transcriptomics data	2022-04-29

Negative binomial

Cell type deconvolution can be performed by explicitly modeling spot level gene expression in terms of individual cell types, usually the scRNA-seq cell clusters. As gene expression is over-dispersed compared to Poisson and is often well modeled with negative binomial, the negative binomial distribution is often used to model gene expression in cell type deconvolution. In stereoscope [74], a negative binomial distribution is fit to the expression of each gene in each cell type in scRNA-seq data. Then at each spot, gene expression is modeled as a weighted sum of the negative binomial distributions from each cell type, and the weights are estimated by maximum likelihood estimation (MLE).

In cell2location [75], expression of each gene at each spot is modeled as negative binomial, parameterized with rate and dispersion. The rate is a weighted sum of cell type gene expression signatures and the weights themselves are modeled with factors to group cell types for similar cell type localization. The rate is also adjusted for technology sensitivity, which can be different between scRNA-seq and the spatial technique, and additive shifts specific to the gene and spot. The model is Bayesian and the weights and the sensitivity scaling parameter have informative priors. The parameters are estimated with variational inference. The weights can be interpreted as the number of cells of each cell type at each spot.

In DestVI [76], expression of a gene in both the scRNA-seq reference and the query spatial dataset is modeled by a negative binomial model, parameterized with rate and dispersion, and the rate is informed by a low-dimensional cell type specific latent embedding from a variational autoencoder. For the spatial data, the rate involves a weighted sum of the cell type latent vectors representing average state of the cells. These weights would be cell type proportions after normalizing so they add up to 1. The scRNA-seq and the spatial data are modeled separately. To link the two models, the scRNA-seq cell type latent vectors are used as priors for those for the spatial data, and the decoder of the model trained on scRNA-seq data is used in the model for the spatial data, as transfer learning of cell state decoding.

While not negative binomial regression per se, the negative binomial model is central to AdRoit [77], so AdRoit is summarized in this subsection. First, genes informative of cell types are selected, from cell type marker genes and highly variable genes. Then for each cell type, a negative binomial distribution is fitted to expression of each gene with MLE. Then the mean and variance of this negative binomial distribution are computed. This fitting is also done to each sample (in bulk RNA-seq) or spot (ST

and Visium), and the mean and variance are computed. Then cell type proportions are roughly estimated in each sample with non-negative least square (NNLS), where the mean in the sample is a weighted sum of the means in each cell type, with a constant to make sure that the proportions add up to one. Then the log ratio between the actual sample mean and the predicted mean from the rough proportions (without the scaling constant) for each gene is computed as a gene specific scaling factor. Then a similar linear model of the mean per sample and mean per cell type for each gene is fitted with NNLS, with that gene specific scaling factor multiplied to the cell type proportion coefficients. In the loss function, weights are added to reduce contribution from genes not specific to a cell type or too variable across samples and the cell type proportion coefficients are L2 regularized due to collinearity of similar cell types. Then the cell type proportion coefficients are scaled to add up to 1.

Other (generalized) linear models

Despite the over-dispersion, the Poisson distribution is often used to model gene expression as it captures the discrete nature of transcript counts and is simpler than the negative binomial distribution. In Robust Cell Type Decomposition (RCTD) [78], gene expression at each spot is modeled as a Poisson distribution, whose mean is an expected rate scaled by total transcript count at the spot. The log rate is the sum of the log of weighted sum of mean gene expression for each cell type from a scRNA-seq reference, a fixed spot specific effect term, a gene specific platform random effect, and another gene specific random effect term for overdispersion. The parameters, including cell type weights, are then estimated with MLE.

SpatialDecon [79] is written for GeoMX DSP, and is based on log-normal regression. As gene expression is often right skewed, log transformation is commonly used to pull the tail in and make the data look a bit more normal for statistical methods that assume normal distribution of the data. After log transformation, a linear model is fitted so the observed gene expression in each ROI is a weighted sum of gene expression signatures of each cell type, with an additive baseline as intercept. The weights must be non-negative, and their p-values are calculated. To remove outliers, any gene expression value below a threshold where technical noise dominates is set to that threshold.

In both the prequel and current era, NMF is quite popular among data analysis methods as the factors (cell embeddings) and the gene loadings tend to exhibit block-like structures and the values of the basis and the loadings are enforced to

be non-negative, corresponding to the non-negative nature of gene expression data and making the results more interpretable. The blocks in the factors may reflect cell types or clusters, and the blocks in gene loadings may reflect cell type marker genes. NMF has been used for cell type deconvolution as well. To address slide-seq (version 1)'s lack of single-cell resolution and poor efficiency, NMFreg was developed to reconstruct the expression profile of each spot as a weighted sum of cell type signatures from scRNA-seq [80]. First, scRNA-seq gene count matrix of cell types of interest and cell type annotations is decomposed with NMF, and each factor is assigned to a cell type and one cell type can have multiple factors. Then non-negative least squares is used to compute the weights of the weighted sum of the factors for each spot. As such weights tend not to cleanly assign spots to cell types, perhaps due to the sparsity of scRNA-seq and slide-seq data, the weights are then thresholded. The threshold is the maximum cell loading of cells not assigned to the cell type of interest among in the factors of this cell type. The weights for this cell type are only kept if the l_2 norm of the weight vector for these factors exceed the threshold. Another NMF based method, SPOTlight [81], uses a very similar principle.

Topic modeling

In Chapter 8, we performed topic modeling of LCM related abstracts with the `stm` package, a generative model of word counts in abstracts from latent topics, and discussed proportion of each topic in each abstract, topic proportions in the entire corpus, and the probability of getting each word from each topic. The `stm` method is built upon a popular and classic topic modeling method, latent Dirichlet allocation (LDA); beyond LDA, `stm` allows for covariates in portion of topics in each abstract (e.g. discussed in Sections 8.3 and 8.4) and correlation between topics (e.g. discussed in Figure 8.5). In both `stm` and LDA, a set number of topics must be chosen before hand, and the number can be chosen based on metrics such as how well word counts are predicted in a held out portion of the dataset. LDA has been used in some cell type deconvolution methods, where cell type is analogous to topic and gene is analogous to word.

In STRIDE [82], the scRNA-seq and the spatial data are assumed to be similar enough to be projected into a shared latent space, inferred from LDA. Here topic isn't entirely the same as cell type. Contribution of each topic to each cell and the probability of getting each gene from each topic are estimated, and from the former contribution of each topic to each pre-annotated cell type can be summarized. Then

the model trained on scRNA-seq data is used to predict contribution of each topic to each spot in the spatial data, which can then be related to contribution of each cell type to each spot.

In contrast, STdeconvolve [83] does not use a scRNA-seq reference and the topics are the cell types. First, genes more likely to be informative of cell types are selected, including genes that are over-dispersed and are neither expressed in too few spots nor constitutively expressed in all spots. This is reminiscent of removing stop words (e.g. the, is, to, of, so, and) and rare words in text mining, as performed for the analyses in Chapter 8, as these words are less informative of the topics. Then with the informative genes, LDA is performed to estimate contribution of each cell type to each spot and the probability of getting each gene from each cell type.

Other principles without explicit statistical modeling

Some of the packages already mentioned in previous sections have cell type deconvolution functionalities as well. For instance, Seurat's data transfer based on anchors between datasets can also be used to transfer cell type annotations, and the *ad hoc* score for the transferred cell types has been used as a qualitative measure of cell type composition in Visium spots [32]. Giotto implements three methods for qualitative cell type deconvolution: First, a score based on fold change in expression of marker genes in a spot compared to the mean across spots. Second, another score scoring genes for specificity in both scRNA-seq cell types and ST or Visium spots and the sum of the top 100 gene scores is the cell type enrichment score for each spot. For these two methods, p-values are calculated from permutation testing. Third, given a fixed set of cell type marker genes, a hypergeometric test is used to test for enrichment of marker genes among top 5% expressed genes of the spot. In Tangram, the cell mapping matrix from scRNA-seq to ST or Visium can be inferred as the ground truth cell density per spot can be measured from H&E staining. When cells from scRNA-seq are mapped to spots in ST and Visium, the cell type annotations are also mapped.

The graph convolutional neural network (GCN) has been applied to cell type deconvolution, in DSTG [84]. First, scRNA-seq cells are randomly assigned to "spots" of 2 to 8 cells, forming a pseudo-ST dataset. Then the pseudo-ST and real ST data are projected to a CCA space, and a mutual k nearest neighbor graph is built in this space. After that both the pseudo and real ST data and the graph are fed into a GCN, trained to minimize cross entropy between imputed cell composition and actual cell

composition in the pseudo-ST spots. Finally, the trained model is used to predict cell composition in real ST data.

As already mentioned in Section 9.3, some methods exploit spatial autocorrelation in gene expression to map dissociated cells from scRNA-seq to a spatial reference or to impute gene expression in space. Also as well be discussed in Section 9.7, some methods that find spatial regions based on the transcriptome favor spatial autocorrelation in cluster labels, i.e. neighboring spots or cells tend to have the same label. Cell type colocalization is also spatially autocorrelated, but spatial autocorrelation is generally not exploited in cell type deconvolution methods.

9.5 Spatially variable genes

Some genes, such as house keeping genes, are ubiquitously expressed. Such genes, while highly variable at the single-cell level, may be interspersed in space so they may not show a spatial trend. Expression of some genes depends on spatial location, which can be due to cell type localization or variation within or independent from cell types. One of the goals of early prequel studies was to identify spatially variable genes, which was done manually, which can be inconsistent and labor intensive. With more quantitative data and data analysis methods, the current era brought identification of spatially variable genes to the next level. Simple methods to identify such genes include dividing the extent of the tissue into a grid and use Fisher's exact test to test for non-random distribution of transcript counts in the grid, or to run DE between one region—be it a grid cell or a manually annotated histological region—and another region. Some more sophisticated methods have been developed that avoid the potential arbitrariness of grids and manual annotation, taking advantages of increased resolution of spatial transcriptomics. This section reviews these computational methods that identifies genes with expression that depends on spatial locations. Two principles are the most common. One is Gaussian process regression and generalization to discrete distributions with the log mean parameter modeled as Gaussian process. Another centers on Laplacian scores of graphs. There are also some additional methods using other principles.

Table 9.5: Packages mentioned for spatially variable genes

Name	Language	Title	Date published

trendsceek	R	Identification of spatial expression trends in single-cell gene expression data	2018-03-19
SpatialDE	Python	SpatialDE: identification of spatially variable genes	2018-03-19
Seurat3	R	Comprehensive Integration of Single-Cell Data	2019-06-13
RayleighSelection	R; C++	Clustering-independent analysis of genomic data using spectral simplicial theory	2019-11-22
SPARK	R; C++	Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies	2020-01-27
GLISS	NA	Integrative Spatial Single-cell Analysis with Graph-based Feature Learning	2020-08-13
singleCellHaystack	R	A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data	2020-08-28
SPATA	R	Inferring spatially transient gene expression pattern from spatial transcriptomic studies	2020-10-21
Giotto	R	Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data	2021-03-08
MERINGUE	R; C++	Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomics data with nonuniform cellular densities	2021-05-13
BOOST-GP	R; C++	Bayesian Modeling of Spatial Molecular Profiling Data via Gaussian Process	2021-06-19

SOMDE	Python	SOMDE: A scalable method for identifying spatially variable genes with self-organizing map	2021-06-24
GPcounts	Python	Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments	2021-07-02
scGCO	Python	Identification of spatially variable genes with graph cuts	2022-09-19
singleCellHaystack	R	A universal differential expression prediction tool for single-cell and spatial genomics data	2022-11-15

Gaussian process regression

Gene expression in space can be modeled as a 2D Gaussian process. Spatial dependence of gene expression from any finite collection of locations in space can be modeled with a joint multivariate Gaussian distribution, whose covariance matrix can be defined with a kernel, which is typically defined so spatially closer cells or spots have higher covariance.

SpatialDE [38] is one of the more popular methods to identify spatially variable genes. Spatial gene expression is modeled as a Gaussian process, in which the mean is the mean expression level of the gene, and the covariance matrix has a spatial and non-spatial component. The spatial component uses the Gaussian kernel, in which the covariance decays exponentially with squared distance between cells or spots, with rate of decay controlled by a length scale parameter. In the null model, the gene expression follows a Gaussian distribution without covariance between cells or spots. Then the model likelihood of the fitted full model and the null model are compared with log likelihood ratio test. The log likelihood ratios under null model are asymptotically χ^2 distributed, and this distribution is used to calculate the p-values of the test. If a gene is found to be significantly spatially variable, then the full model can be fitted with two other kernels, linear and periodic, and compared to the Gaussian kernel with Bayesian Information Criterion (BIC) to discover linear and periodic patterns. As gene expression is discrete and not Gaussian, the data needs to be normalized before applying SpatialDE; even then, data normalization does not make the data Gaussian.

The discrete, non-Gaussian distribution of gene expression is directly modeled by

SPARK [42]. Gene expression is modeled by a Poisson distribution, with a rate parameter scaled by total transcript count at the spot or cell of interest. The log rate parameter contains a linear model for non-spatial variation in gene expression and can include cell or spot level covariates such as cell type, with non-spatial residuals. The spatial dependence is modeled by a zero mean Gaussian process with either a Gaussian or cosine kernel for the covariance matrix and 5 different length scale parameters are tried for each kernel type, so 10 kernels are tried. The model is fitted with one kernel at a time, with a penalized quasilikelihood algorithm. The p-values are estimated by Satterthwaite method, and the p-values from the 10 kernels are combined with the Cauchy p-value combination rule.

Gene expression data may better be modeled with NB or ZINB, which is done in GPcounts [85]. The log of the mean parameter of the NB or ZINB, scaled by total transcript count at the cell or spot, is modeled with a Gaussian process with Gaussian kernel for covariance. For ZINB, the dropout probability is related to the NB mean by a Michaelis-Menten equation. For one sample, the null hypothesis a constant model, a Gaussian with fixed mean and no covariance between cells or spots, i.e. gene expression does not vary in space. Spatially variable genes are identified with the log likelihood ratio test as in SpatialDE. For two samples, the null hypothesis is that two samples have the same gene expression pattern, and the alternative hypothesis is that two different Gaussian processes are required to model the two samples. Three models are fitted, one for each sample and another fit with both samples as replicates, and the SpatialDE likelihood ratio test is used to compare the separate models to the shared one. The models are fitted with a sparse approximation of variational Bayesian inference. A similar ZINB model is used in BOOST-GP [86], but instead of using the likelihood ratio test, the model is fully Bayesian. Whether a gene is spatially variable is a parameter in the model that indicates whether a kernel other than white noise (covariance among the locations is the identity matrix) is appropriate for a gene of interest, and the posterior distributions of the parameters are sampled with Markov chain monte carlo (MCMC).

The size of the covariance matrix of the cells or spots grows quadratically with the number of cells or spots. To speed up computation, SMODE aggregates cells or spots into nodes with SOM, reducing the size of the covariance matrix, before proceeding to a SpatialDE-like test [87].

Laplacian score

GLISS [51] has already been mentioned as a method to reconstruct scRNA-seq data in space by projecting scRNA-seq data into one to three dimensions that stand for spatial dimensions. The first step of GLISS is to identify spatially variable genes in the spatial reference as landmark genes. In 2005, the Laplacian score was proposed as a method of feature selection, which favors features that preserves the local structure of the data in the feature space and has large variance [88]. In GLISS, a spatial neighborhood graph is constructed on the spatial reference; two cells or spots have larger edge weight if they are physically close to each other. By default, the graph is a mutual nearest neighbor graph, in which cells or spots are nodes and an edge connects two nodes if they are mutual k nearest neighbors. Then for each gene, a Laplacian score is computed using the gene of interest and the graph Laplacian of the spatial neighborhood graph. Genes with low Laplacian scores are chosen as landmark genes, as a low score favors similarity of gene expression in nearby cells or spots and large variance among the spots, which means spatially coherent regions with high and low expression of the gene. The p-value of the gene is computed by permuting expression of the gene of interest among cells and recomputing the score.

The simplical complex is a generalization of the graph that not only includes nodes and edges but also triangles, tetrahedrons, and their higher dimensional generalizations. RayleighSelection implements generalizations of the Laplacian score for simplical complexes for clustering-free DE [89]. The one-dimensional Laplacian score, a generalization in which gene expression values are attributed to edges rather than nodes, has been used for DE in scRNA-seq data. The nodes here are clusters of cells and two nodes are connected by an edge when they intersect, as in topological data analysis (TDA) [90]. P-values of genes were computed by permutation test, permuting expression of a gene of interest among cells. For spatial data, the spatial neighborhood graph was created as the Vietoris-Rips complex. The zero-dimensional, which is the same as the original Laplacian score, was used to identify spatially variable genes. The graph was also created for cells from pairs of cell types and the Laplacian score, with feature as cell type label, was used to identify cell type colocalization.

Other principles

A spatial point pattern is the observed spatial locations of things or events, and a point process is a stochastic mechanism that generated the point pattern. As already mentioned, in pciSeq, transcript spot locations are modeled by a Poisson point

process whose intensity itself is modeled with a Gamma distribution. Cell locations can also be modeled as a point process, which is done in trendsceek [39]. Each point in a spatial point process can have additional properties other than location, such as gene expression and cell type, which are called marks. If the marks are completely randomly distributed in space, then points with one mark would not be more or less likely to be near points with the same (for categorical marks) or similar (quantitative marks) marks than to points with dissimilar marks. To identify spatial distribution of gene expression that deviates from complete randomness, trendsceek uses 4 mark-segregation summary statistics, which are functions of distance between two points, taking the expected value of a summary statistics on the marks of every pair of points sepearated by the given distance: Stoyan's mark-correlation function (squared geometric mean of marks of two points normalize by squared mean of marks), mean-mark function (average of marks in two points), variance-mark function (variance of the marks given distance between points), and mark-variogram (squared difference of marks of two points). Permutation testing is used to calculate p-values. Regions of interest in the tissue are the regions with $p < 0.05$ from the permutation testing. Perhaps due to the permutation, trendsceek seems to be less scalable and less sensitive than SpatialDE and SPARK [42, 91].

Seurat's spatial functionalities include finding spatially variable genes, which currently provides two methods, one of this is mark-variogram, inspired by trendsceek. The other is Moran's I, which is a common summary statistics of spatial autocorrelation, as spatially patterned genes also exhibit autocorrelation. MERINGUE uses a local version of Moran's I for spatially variable genes [92]. As dependence of gene expression of spatial location means spatial autocorrelation, which both Gaussian process models and Laplacian score of spatial neighborhood graphs aim to identify, Moran's I can be a simpler and hence more computationally efficient way to identify spatially variable genes. Moran's I is sometimes used to evaluate performance of more sophisticated SVG methods, with higher values being better (e.g. [86, 93]; this raises the question of whether Moran's I itself is that much worse than the more sophisticated methods in detecting SVG. SPARK is claimed to have higher power than Moran's I test (`spdep::moran.test()` in R) due to the latter's use of asymptotic normality in computing the p-values [42], but this may or may not hold for SVG methods based on other principles.

Giotto [35] implements three simple and fast methods to find spatially variable genes in addition to wrapping SpatialDE and trendsceek. First, a spatial neighborhood

graph is constructed, which can be mutual k nearest neighbors graph, a graph placing an edge when two cells are within a certain distance, or Delaunay triangulation. Then the gene expression is binarized. The first method uses the silhouette score. In clustering, a measure of whether each point should be assigned to its current cluster or it should better be assigned to a neighboring cluster. The mean silhouette score indicates how tight and segregated the cluster are. Here the clusters are cells expressing the gene of interest and those not expressing. Then a high silhouette score means that cells expressing the gene and those not expressing are well-segregated in space, which means the gene is spatially variable. The second and third method only differ in the way gene expression is binarized. The second uses k-means with $k = 2$, and the third uses a threshold. Then a contingency table M is constructed from neighboring cells in the graph expressing or not expressing the gene; each row is whether a cell expresses the gene, and each column is whether its neighbor also expresses it, so $M_{1,1}$ is the number of distinct pairs of cells both expressing the gene, $M_{1,2}$ is the number of pairs of cells in which source cell is expressing the gene and target cell is not, and so on. Fisher's exact test is used to test for dependency in gene expression on whether cells are neighbors.

The KL divergence is a measure of difference between two probability distributions. In singleCellHaystack [94], the cell density in the tissue (or a PCA, tSNE, or UMAP space) is estimated at grid points with Gaussian kernel density, and normalized to form a probability distribution of locations of cells. Then the probability distribution of whether a gene of interest is expressed at each grid point is compared to the cell density distribution with KL divergence. P-values are computed by permuting gene expression among cells. Again, this is a cluster-free DE method, not designed specifically for spatial data but can be applied to spatial data.

We have already mentioned Markov random field (MRF) models for partitioning a tissue section into cell types and cells. MRF has also been used to identify spatially variable genes, as in scGCO (single-cell graph cuts optimization) [91]. Expression values of a gene are binned into 2 to 10 categories with Gaussian mixture model clustering. Then a graph connecting cells in space is constructed over the tissue by Delaunay triangulation, and the graph, with the expression category of the gene, is modeled with a MRF. Then as the model favors neighbors in the graph with the same category, edges of the graph are cut to maximize the likelihood of the model, thus identifying not only regions of tissue with an expression category of the gene, but also genes forming such regions. As MRF enforces coherent regions

of the tissue to take the same category, while when only gene expression, without spatial information, is considered, not all cells in the region warrant the category. Then the number of cells that truly deserve the category in each region is used to calculate statistical significance of the gene's spatial variability. The null hypothesis is a homogeneous Poisson point process, in which cells (points) are completely randomly distributed in space and the location of one cell is independent from the location of any other cell. The smallest p-value of any category and any region is reported for the gene of interest.

So far the methods identifying spatially variable genes based on Gaussian process regression commonly use the Gaussian kernel for the covariance matrix, which assumes that the gene expression modeled is weakly stationary, i.e. covariance only depends on distance between cells or spots. This does not take into account anisoptropy, i.e. spatial dependence of gene expression is different in different directions, observed in tissues such as the brain cortex and the hepatic lobule in which cell functions are primarily stratified along one direction or axis. SPATA [37] implements a method to find spatially variable genes for such primary axis. With the interactive `shiny` app, the user defines this axis, which may or may not be a straight line, and cells within a certain distance from the axis are included for further analysis. Then among the included cells, gene expression and cell type annotations along the axis can be visualized in the `shiny` app. Then SPATA fits a variety of functions with known forms, e.g. linear or nonlinear descent or ascend, peaks, periodic, etc. to the gene expression along the axis. For each function, the sum of the residuals is calculated and compared to find functions that better represent the change in gene expression along the axis to identify patterns.

9.6 Gene patterns

Table 9.6: Packages mentioned for gene patterns

Name	Language	Title	Date published
SpatialDE	Python	SpatialDE: identification of spatially variable genes	2018-03-19
std-nb	C++	Charting Tissue Expression Anatomy by Spatial Transcriptome Decomposition	2018-12-28

stLearn	Python	stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues	2020-05-31
GLISS	NA	Integrative Spatial Single-cell Analysis with Graph-based Feature Learning	2020-08-13
MERINGUE	R; C++	Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomics data with nonuniform cellular densities	2021-05-13

When spatially variable genes are identified, a question naturally arises: Are there archetypal patterns among these spatially variable genes? As already reviewed in Section 5, comparing and classifying gene expression patterns was a major topic in the prequel era. Such interest persists in the current era, although we find no evidence that current era gene pattern analysis is significantly influenced by the prequel antecedents, although factor analysis and NMF have been used in both eras.

The most straightforward way to identify archetypal gene patterns is to cluster the gene expression patterns and obtain the cluster centers to represent the cluster. This has been used to analyze mouse brain voxelation data in 2009 [95]. Wavelet transform was applied to the data and the Euclidean distance between the wavelet feature vectors was used to measure gene similarity. Gene similarity between pre-defined "typical" genes and other genes was one way to find groups of similar genes and k-means clustering is another.

Some package already reviewed also have functionality to identify archetypal gene patterns. In DistMap and SPARK [54, 42], the gene patterns are clustered with hierarchical clustering, and the individual clusters are obtained by tree cut. In MERINGUE [92], spatial cross-correlation is computed for each pair of SVG. The spatial cross-correlation here is similar to Moran's I, and relates to Moran's I in a way similar to how covariance relates to variance. Then the spatial cross-correlation matrix is clustered with hierarchical clustering. In Giotto [35], a gene-gene correlation matrix (by default Pearson) is computed, which is then hierarchically clustered.

Then the mean or centroid of each cluster is taken to represent that cluster. SpatialDE [38] also clusters gene expression patterns, in automatic expression histology (AEH), which implements a Gaussian process generalization of Gaussian mixture model clustering. The number of components is set by the user, and the model is fitted to infer the mean pattern of each component. In GLISS [51], the archetypal patterns are identified in the reconstructed latent space as gene expression in the latent space is spline smoothed, and the spline coefficients are clustered.

Beyond clustering, a common way to identify archetypal gene patterns is factor analysis. This has already been done in the prequel era [96], but is further developed in the current era. Factor analysis tries to model higher dimensional data as a linear combination of a smaller number of variables called "factors", and PCA is a type of factor analysis. A prostate cancer ST dataset has been modeled with Poisson factor analysis [30]. The observed UMI counts at each spot is modeled as a sum of factors, each of which is Poisson distributed, with its own rate parameter, which in turn depends on Gamma distributed factor, gene, and spot level parameters that may account for overdispersion though this model does not entirely capture the mean-variance relationship of NB. The parameters are estimated from MCMC sampling of the posterior of this model. Once the parameters are estimated, the individual factors can be calculated from the parameters based on the model. The factors seem to indicate regions in the tumor, such as cancer, stroma, and regions with immune cell infiltration. As NB may describe gene expression better than the Poisson distribution, a NB adaptation of the above Poisson factor analysis model has been developed [97]. The observed UMI count at each spot is modeled as a sum of NB factors, whose rate parameter can incorporate gene, spot, and experiment level covariates. The package stLearn [57], which also implements methods to identify cell-cell interactions and spatial regions, uses PCA, ICA, and factor analysis to detect microenvironments in the tissue as again, the factors can correspond to specific regions in the tissue.

9.7 Spatial regions

As already mentioned in trendsceek and scGCO, the problem of identifying spatially variable genes is closely related to identifying regions in tissue defined by gene expression. When archetypal gene patterns are identified, a related question arises: Do the patterns define novel anatomical regions in the tissue? As seen in the previous section, archetypal gene patterns, such as in factors, can reflect tissue regions. There are also methods that identify such regions without first identifying spatially variable

genes and/or archetypal gene patterns. In the prequel era (Chapter 5), some studies clustered the voxels based on gene expression to identify spatial regions in the tissue, with either k-means clustering or co-clustering [98, 99, 100], or with Potts model [101]. More sophisticated clustering methods have been developed in the current era tot identify spatial regions. However, as different cell types can reside in the same spatial neighborhood, and conversely, cells from one cell type can reside in different regions of the tissue, MRF has been used to find spatially coherent regions that can contain multiple cell types.

Table 9.7: Packages mentioned for spatial regions

Name	Language	Title	Date published
smfishHmrf	R; Python; C	Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data	2018-10-29
stLearn	Python	stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues	2020-05-31
MULTILAYER	Python	Inferring biologically relevant molecular tissue substructures by agglomerative clustering of digitized spatial transcriptomes with multilayer	2021-05-07
BayesSpace	R; C++	Spatial transcriptomics at subspot resolution with BayesSpace	2021-06-03
SSAM	Python; C++	Cell segmentation-free inference of cell types from in situ transcriptomics data	2021-06-10
lisaClust	R	Spatial analysis for highly multiplexed imaging data to identify tissue microenvironments	2021-08-17

Baysor	Julia	Bayesian segmentation of spatially resolved transcriptomics data	2021-10-14
SpaGCN	Python	SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network	2021-10-28
SC-MEB	R; C++	SC-MEB: spatial clustering with hidden Markov random field using empirical Bayes	2021-11-25
RESEPT	Python	Define and visualize pathological architectures of human tissues from spatially resolved transcriptomics using deep learning	2022-08-24

Clustering

SSAM [25], already reviewed in Section 9.1, also uses clustering to identify tissue domains in smFISH or ISS data but without cell segmentation. StLearn [57] develops further on top of clustering. First, a pretrained CNN is used to extract a 2048 dimensional feature vector from the H&E image behind each ST or Visium spot. The cosine similarity between the feature vectors from neighboring spots is then calculated. To normalize data, the gene expression data is smoothed in space, and the smoothing is weighted by the cosine similarity of feature vectors between spots. Then the spots are clustered with Louvain or k-means. A spatial k nearest neighbor graph is constructed, and used to refine the clustering. If a gene expression based cluster is broken into multiple pieces in space, then those pieces would become subclusters. Singleton spots are merged with a nearby cluster if the singletons have enough spatial neighbors in that cluster.

In MULTILAYER, an agglomerative strategy is first used to find binarized regions of gene overexpression compared to average. The over- and under- (i.e. differential) expression of genes relative to the average in tissue is evaluated per spot and genes are ranked by the number of DE spots, as SVG. Then these binarized regions are compared with the Jaccard (Tanimoto) coefficient and the Dice coefficient, which become edge weights of a similarity graph. Then the graph is used in Louvain clustering to find co-expression patterns, which identifies gene expression based

spatial regions.

Markov random field

Technically, this still is a form of clustering, but the MRF is used to favor the same cluster assignment of neighboring cells or spots.

BayesSpace [102] incorporates both Gaussian mixture model clustering and MRF. The ST or Visium data is first projected to a low-dimensional space, such as by PCA. Then for each spot, the low-dimensional projection of that spot is modeled with a Gaussian mixture model, with a pre-defined number of components or clusters. The spatial neighborhood graph is simply the square grid of spots for ST and the hexagonal grid for Visium. The model has a MRF prior to encourage neighboring spots to be assigned to the same cluster. The cluster assignment is initiated with non-spatial clustering, and the parameters of the model are estimated by MCMC. In addition, BayesSpace can increase the resolution of ST and Visium. Each spot is subdivided and initiated with the dimension reduction values at the spot, and an additional parameter is added to the model that nudges the dimension reduction values at each sub-spot while preserving the sum at the spot level. The nudging parameters are estimated by MCMC along with with other parameters.

As already reviewed in Section 9.1, Baysor [24] uses MRF to delineate cell type regions in the tissue without cell segmentation. MRF is used to identify spatial regions for cell or spot level data as well. Like in the 2014 *Platynereis dumereilii* atlas [101], smfishHmrf [40] also uses Potts model for dependence of label on neighborhood. As seqFISH data is quantitative, gene expression of each cell is modeled with a Gaussian mixture model, with as many components as there are there are region labels. The data needs to be normalized, although data normalization methods don't typically turn the distribution of gene expression Gaussian. Again, the parameters, i.e. the label assignment, and mean and covariance matrices for each Gaussian component, are estimated by EM, initiated with k-means clustering of the cells.

In the SC-MEB [103] model, gene expression at each spot is Gaussian, and independent conditioned on an unknown cluster label. The cluster label then has a Potts model prior to encourage nearby spots to have the same label. The parameters, including the mean and covariance of the Gaussian distributions and the cluster labels, are estimated with an EM algorithm.

Graph convolutional network

The spatial neighborhood graph can be integrated with gene expression at the spots with a graph convolutional network (GCN), and some GCN based methods have been developed since the previous version of this chapter was written. Again, clustering is involved, but these methods are discussed in a separate subsection for the use of GCN to incorporate spatial information.

SpaGCN [93] incorporates information from the H&E staining that usually comes with ST and Visium datasets into the spatial neighborhood graph. In addition to the x and y coordinates, a weighted average of the red, green, and blue channels of the patch of H&E image behind each spot is used as a "z" coordinate. Then x, y, and "z" coordinates are used to calculate an "Euclidean" distance between spots, so physically close spots that are histologically different are considered further. Edge weights in the spatial neighborhood graph is negatively associated with the distance by a Gaussian kernel, as commonly used for the covariance matrix in Gaussian process models. Then the first 50 PCs of the gene count matrix and this spatial neighborhood graph are combined by a graph convolutional layer, whose output is Louvain clustered and iteratively refined, which would be the spatial regions.

STGATE [104] and RESEPT [105] both use graph autoencoders. In STGATE, the spatial information, in the edge weights of the spatial neighborhood graph, is learnt by a graph attention layer in the autoencoder. The latent embedddings inferred by the autoencoder can then be clustered with any clustering algorithm to give the spatial regions. In RESEPT, the autoencoder infers a 3-dimensional latent space representing the gene expression and the spatial neighborhoods, which can then be represented with the RGB channels of a color image. The image is then segmented with a convolutional neural network model, giving the spatial regions.

Spatial statistics

The rich tradition of spatial statistics, originally more used in the geographical scale, has been brought to spatial transcriptomics. These are common types of geospatial data: First, measurements at points in space, where the locations of the points are pre-determined and only the values at each location are of interest. Second, areal data, where there is one aggregated value for each areal unit such as a city or a district. Third, raster, where each cell of a regular grid (without spacing between cells) has a value. Fourth, point locations, where the locations themselves are of interest, modeled by point process models; there can be additional values associated

with each point as well. Fifth, where values can only occur on lines in a network, such as a road or river network; the data of interest can be values at pre-determined locations, or the locations themselves, or both. For each type of data, there is a well-established collection of statistical and software tools.

For spatial transcriptomics, ST and Visium data are a combination of the first two, as the locations of the spots are known and spots are often treated as points in analyses, but each spot is in fact an aggregated areal unit. smFISH and ISS data with single-cell and single molecule resolution are a combination of the second (when considering cells as areal units) and the fourth (when considering each cell as a point and studying cell locations, or when considering locations of each transcript spot). Voxelation, LCM in regular grid, and Tomo-seq data can be thought of as raster or as areal. While network spatial statistics could in principle be applied to blood vessels or axons and dendrites in the tissue, this is impractical at present because the thin sections used in almost all current era techniques only captures a tiny 2D slice of the 3D network, cutting through many of the edges, thus failing to preserve its shape and topology.

Concepts already mentioned, such as Moran's I (global and local, and the test), Gaussian process regression (used in kriging, to interpolate values between points of measurements), and MRF (including Potts model, related to autoregressive models of areal units) come from the tradition of spatial statistics, though the latter two often take on more of a machine learning sheen. More of spatial statistics has been used to identify spatial regions in tissue.

lisaClust [106] uses the spatial point process approach. Ripley's K function, as a function of distance r, is the average number of points within distance r of each point, and is a common way to assess if the points are randomly distributed or if they tend to be clustered with or repelled from each other. Where each cell is a point in the point process, and there are multiple cell types, a cross type K function (average number of cells of type j within distance r of each cell of type i) can be used to see if the two cell types attract or repel each other. The L function is a variance stabilized version of the K function. The cross type K function averages over the contribution of each cell of type i, which is the number of cells of type j within distance r, so each cell has a vector of contribution, which can then be clustered with any clustering algorithm to find spatial regions of cell type colocalizations.

Delineating spatial regions from data is not a new problem in spatial statistics. Some methods have been developed to find spatial regions in geographical space, but have

not been widely used for spatial transcriptomics to the best of our knowledge. For example, ClustGeo [107] uses dissimilarity in both feature space and physical space for hierarchical clustering, and spatialcmeans [108] performs spatially weighted c-means fuzzy clustering and only probabilistically assigns locations to clusters.

9.8 Cell-cell interaction

Related to spatial regions is cell-cell interaction: suppose a distinct neighborhood of the tissue has been identified with one of the methods in the previous section, and the neighborhood contains different cell types. Then it's natural to ask whether these cell types interact by their spatial proximity. Such information is lost in scRNA-seq. The composition of tissue neighborhoods can be characterized with existing tools. For instance, in smFISH or ISS data, we can count the number of cell types within a certain distance from each cell, as was done in the hypothalamus and the motor cortex MERFISH studies [12, 28]. We can model the data as a marked spatial point process, in which each point is a cell, with cell type annotations as marks, and use cross-type K or L function to find cell types that colocalize; the cross-type L function has been used in spicyR ([109]) for this purpose. In MERINGUE [92], spatial cross-correlation is computed for two different cell types, one with expression of a ligand, and the other with expression of a receptor, and permutation testing is used to find p-values of the ligand-receptor interactions of the cell types.

For ST and Visium, we can use one of the cell type deconvolution methods to find the number and proportion of cell types per unit area in each tissue region and cell type colocalization. When two cell types colocalize, they might interact with secreted ligands or ligands and receptors bound to the membrane. Expression of ligand-receptor (L-R) pairs in neighboring cells is often used to identify cell-cell interaction in spatial data, and the CellPhoneDB [110] database of ligands, receptors, and their interactions is often used to identify such L-R pairs. Another type of analysis going beyond colocalization tests for effects of cell-cell interaction or cell type colocalization on gene expression.

Table 9.8: Packages mentioned for cell-cell interactions

Name	Language	Title	Date published

SVCA	R; Python; C; C++; Fortran	Modeling Cell-Cell Interactions from Spatial Molecular Data with Spatial Variance Component Analysis	2019-10-01
SpaOTsc	Python	Inferring spatial and signaling relationships between cells from single-cell transcriptomic data	2020-04-29
stLearn	Python	stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues	2020-05-31
GCNG	Python	GCNG: Graph convolutional networks for inferring cell-cell interactions	2020-12-10
Giotto	R	Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data	2021-03-08
MISTy	R	Explainable multiview framework for dissecting spatial relationships from highly multiplexed data	2022-04-14
DIALOGUE	R	DIALOGUE maps multicellular programs in tissue from single-cell or spatial transcriptomics data	2022-05-05

Ligand-receptor pairs

In stLearn [57], CellPhoneDB is used to identify L-R coexpression in neighboring spots, and the p-value of the coexpression is computed by permutation testing. Then regions with diverse cell types (from Seurat label transferring or cell type deconvolution) and L-R coexpression in neighboring spots are identified as regions where cells are likely to be signaling to each other. A similar strategy is used in Giotto. Giotto identifies cell type colocalization by labeling edges of the spatial neighborhood graph as homo- or heterotypic and permutes cell type labels to find whether the cell types are more or less likely to colocalize than expected from completely random cell type localization. L-R coexpression in neighboring cells on the spatial neighborhood graph from two cell types is identified and the p-values of

the coexpression scores are computed by permutation testing, permuting locations of cells within each cell type.

While MRF, stLearn, and Giotto only use the immediate neighbors on the spatial neighborhood graph, there is a method that can capture higher order structures of the graph. In GCNG [111], the spatial neighborhood graph is constructed as an edge connects a cell to its three nearest neighbors. Then both the gene count matrix and the normalized Laplacian of the neighborhood graph are fed into a graph convolutional neural network (GCN), which is trained on known L-R pairs. The GCN can then predict novel pairs of genes involved in signaling, and if trained on the direction of interaction in the L-R pairs, it can also predict the direction of causality in the novel pairs.

SpaOTsc has already been mentioned in Section 9.3. To recapitulate, SpaOTsc uses optimal transport from scRNA-seq cells to spatial locations to impute a spatial cell-cell distance matrix for scRNA-seq cells, and the optimal transport plan can be used to impute gene expression in space. With the cell-cell distance matrix, another optimal transport plan from ligands to receptors can be inferred, interpreted as how likely one cell communicates with another. A disadvantage of spatial neighborhood graph is that common ways of construction are somewhat arbitrary. For instance, k nearest neighbor is a common way to construct the graph, but this k is somewhat arbitrary, although cell signaling can occur over a distance with secreted ligands. Here no such graph is used; the length scale of interaction is inferred by random forest. Random forest models are trained with expression of the ligand and genes correlated with a downstream target gene within a certain distance from the cells expressing the target gene are the input features. Receptor expression is the sample weights, and the target gene is to be predicted by the random forest model. Several different length scales are tried, and the one resulting into the most feature importance of the ligand is used. When L-R information is unavailable, interactions between genes can be inferred by partial information decomposition, i.e. how much unique information can a source gene provide on a target gene in a spatial neighborhood.

With a very different model, DIALOGUE [112] identifies genes that may be involved in interactions between cell types. In a niche in a tissue, different cell types can respond to the same environmental cue in a concerted manner though each cell type changes gene expression in a different way. DIALOGUE aims to identify such concerted gene programs in each cell type. First, the gene expression data

is projected into a lower dimensional space in which correlation between all pairs of cell types across niches is maximized, and the basis of this space is ordered in descending strength of correlation. This is similar to CCA, but with a penalty term to enforce sparsity in gene loading. Here the niche is a patch of cells in space with a predefined number of cells. Then each cell type has a rotation matrix that projects cells into this lower dimensional space, and different cell types from the same niche should be close to each other in this space. In this projection, for each dimension, a gene is added to the multicellular program (MCP) of each cell type if its expression among cells of this cell type correlates with the projection of this cell type in this dimension and is significantly associated with the projection of other cell types while accounting for cell type level and niche level covariates such as sample, age, and gender. Thus the MCPs could be cell type specific co-regulated gene programs. Putative signaling between cell types can be identified by finding known L-R pairs in the MCPs: each cell type is added the L-R graph as a node, and is connected to a gene if the gene is present in the MCP for this cell type. Then a path connecting one cell type to a ligand to a receptor and then to another cell type suggests signaling between the two cell types.

Genes associated with cell-cell interaction

Gene expression can be affected by several different factors, including cell type, local environment, interaction with other cells, and so on. Some packages have been developed to identify genes whose expression is associated with one or more of these factors, without using L-R databases. Within one cell type, Giotto uses classical DE (Student's t-test, Wilcoxon rank sum test, limma, and permutation of spatial locations) to find DE genes between neighbors of cells of another cell type and non-neighbors. Other packages implement more complex models that account for more of these factors associated with gene expression.

Spatial variance component analysis (SVCA) [113] models the expression of each gene of interest among the cells as a 0 mean Gaussian process. The covariance has the following components: First, the intrinsic variability, which can be cell types or continuous cell states. In the latter case, the covariance matrix of this component is the covariance between cells with genes other than the gene of interest that is modeled. Second, the spatial neighborhood. A neighborhood graph is not constructed, and the covariance matrix of this component is computed with the Gaussian kernel in which covariance decreases with distance between cells. Third, cell-cell interaction. The covariance matrix of this term is the covariance between

cells weighed by a Gaussian kernel for distance between cells, so gene expression in nearby cells contributes more to this component. Finally, the residual has an identity covariance matrix, so in the residual, cells are independent from each other. The parameter to be estimated are weights of each of these components and the length scale parameter of the Gaussian kernel, which are estimated with MLE. Significance of the cell-cell interaction component is calculated by likelihood ratio test between the full model and a reduced model without the cell-cell interaction component. Again, as gene expression is modeled as Gaussian, the data needs to be normalized before using this method.

Like SVCA, Multiview Intercellular SpaTial modeling framework (MISTy) [55] also models expression of each gene of interest among the cells, but with ensemble learning, in which any machine learning method that is explainable (i.e. feature importance can be extracted) and suitable for ensemble learning can be used. In each view, which can be intrinsic cell state, spatial neighborhood (juxtaview), or wider tissue structure (paraview), features are extracted from gene expression that represent the view and used in machine learning methods such as random forest to predict expression of a gene of interest. For intrinsic cell state, the features are expression of other genes. For juxtaview, the features are sum of expression of other genes in neighboring cells in the spatial neighborhood graph. For paraview, the feature are sum of expression of other genes in all cells in the tissue weighed by distance to each cell with a Gaussian kernel. Other views, with other feature engineering, can also be used. The full model is a linear combination of predictions of each view. In other words, the contribution of each view is determined by linear regression with prediction of each view as a covariate to predict the expression of the gene of interest. For each view, the importance of each feature is assessed as the z-score of the feature importance (e.g. from random forest) multiplied by 1 minus the p-value of the coefficient of this view in the linear regression model, so views that contribute significantly to the ensemble model and features in each of these views that are more important than other features in the same views stand out. This way, interaction among genes at different spatial scales can be identified.

9.9 Gene-gene interaction

Table 9.9: Packages mentioned for gene-gene interactions

Name	Language	Title	Date published

SpaOTsc	Python	Inferring spatial and signaling relationships between cells from single-cell transcriptomic data	2020-04-29
scHOT	R	Investigating higher-order interactions in single-cell data with scHOT	2020-07-13
MESSI	Python	Identifying signaling genes in spatial single-cell expression data	2020-09-04
GCNG	Python	GCNG: Graph convolutional networks for inferring cell-cell interactions	2020-12-10
MISTy	R	Explainable multiview framework for dissecting spatial relationships from highly multiplexed data	2022-04-14

Some of the packages already reviewed can also infer interactions between genes, such as GCNG, SpaOTsc, and MISTy. GCNG and SpaOTsc predict potential L-R pairs, and MISTy identify genes whose expression at a given spatial scale is associated with another gene of interest. The package scHOT [114] tests for association of correlation between genes with pseudotime or spatial locations by permutation testing, permuting locations of cells along pseudotime or in space. The package Mixture of Experts for Spatial Signaling genes Identification (MESSI) [73] uses a mixture of experts model to predict expression of response genes with a set of features. A spatial neighborhood graph of the cells is constructed with Delaunay triangulation The features include all genes quantified in the dataset that are also found in a L-R database, expression of genes in the L-R database in neighboring cells, cell type of neighboring cells, and etc. The response genes are all genes quantified other than the L-R genes used as features. Each cell is assigned to exactly one "expert", i.e. subtype. For each expert, expression of response genes in each cell is modeled with linear regression with the features as covariates. The parameters of the linear models and assignment of cells to experts are estimated with MLE, where the log likelihood is maximized with EM. This model can be trained in a control sample and used to predict gene expression in experimental samples. If expression of a gene is as well predicted as in the control, then signaling may not have changed in the experimental condition. If prediction becomes worse, then there may be a change in signaling involving this gene, and the experts whose

coefficients significantly differ between the control and experimental models suggest cell populations involved in the signaling change.

9.10 Subcellular transcript localization

Table 9.10: Packages mentioned for subcellular transcript localization

Name	Language	Title	Date published
FISH_quant	MATLAB	A computational framework to study sub-cellular RNA localization	2018-11-02

So far, except for segmentation free data analysis methods of smFISH and ISS images, all analysis methods are at the cellular or spot level. However, transcripts do show inhomogeneous subcellular localization that can be biologically relevant, such as whether the transcripts are translated in the endoplasmic reticulum (ER) or the cytoplasm. Thirty-four lncRNAs have been manually classified into 5 types of subcellular patterns: one or two large foci in nucleus, large foci and dispersed single molecules in nucleus, no foci in nucleus, nucleus and cytoplasm, and cytoplasmic [115]. The bDNA-smFISH study in 2013 that profiled 928 genes in cultured cells, though each gene was profiled in different cells, generated features that characterize subcellular transcript localization (mRNAs of protein coding genes) which were used to cluster cells [116]. These features include closest distance of a transcript spot to cell outline, distance to cell centroid, distance to nuclear centroid, radius to include 5%, 10%, 15%, 25%, 50%, and 75% of all spots in the cell, mean distance of a spot to other spots (related to Ripley's K or L function), and variance of distance to other spots. The package FISHquant [117] uses additional features derived from Ripley's L function of subcellular transcript localization. Then it uses these features to simulate smFISH data, cluster cells, and classify transcript localization patterns.

Whether transcripts are located in the nucleus or in the cytoplasm has also been used for RNA velocity in a MERFISH study [118]. In traditional RNA velocity based on scRNA-seq [119], when there are more transcripts with intronic reads—i.e. nascent transcripts not yet spliced—than expected from steady state in a cell, then the gene of interest is up regulated, and conversely, if there are fewer transcripts with intronic reads, then the gene may be down regulated. In other words, intronic reads not yet spliced out gives a glimpse into a near future transcriptome of the cell.

In this MERFISH study, instead of introns that require separate probes from exons, transcripts inside the nucleus are taken to be nascent, i.e. not yet exported from the nucleus, and used in lieu of intronic reads as in scRNA-seq for RNA velocity. In this study, the ER was also stained for and segmented and genes with transcripts enriched in the ER were also identified.

These studies analyzing subcellular transcript localization were all performed on cultured cells rather than in tissues, so there is no highly multiplexed smFISH data on in vivo subcellular transcript localization yet. Furthermore, as the cells cultures used in these studies grow on a plate in single layer, while cells stack on top of each other through the thickness of the section, cell segmentation in tissue can be more challenging. Some of the features used to characterize subcellular transcript localization, such as distance to cell outline and Ripley's L function (with edge correction), depend on accurate cell segmentation, which as already explained in Section 9.1, is challenging. Subcellular transcript location can be modeled as a spatial point pattern in 3D or collapsed into 2D, and analyses such as finding effects of covariates such as whether the spot is in the nucleus, distance from the nucleus, distance from the cell outline, and etc., and whether the pattern exhibits clustering (e.g. foci) or inhibition (i.e. more spaced out than expected from CSR). However, the observational window of the point process, i.e. cell segmentation, can greatly affect results of spatial point pattern analysis. For instance, when the convex hull of some spots are taken to be the observational window, then the point pattern may not appear clustered. However, if the actual observational window is much larger than the convex hull, then the point process is in fact clustered. Hence accurate cell segmentation is important to analyses of subcellular transcript localization patterns. Furthermore, some smFISH or ISS datasets only provide 2D cell segmentations, and resolution in the z axis tends to be lower than that in the x and y axes. The implications of collapsing 3D into 2D, and when 3D segmentation is available, the lower resolution in the z axis are yet to be determined.

9.11 Gene expression imputation from H&E

Table 9.11: Packages mentioned for gene expression imputation from H&E

Name	Language	Title	Date published

ST-Net	Python	Integrating Spatial Gene Expression and Breast Tumour Morphology via Deep Learning	2020-06-22
PathoMCH	Python	Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer	2020-11-02
Xfuse	Python	Super-resolved spatial transcriptomics by deep data fusion	2021-11-29

Although ST and Visium do not have single-cell resolution, the tissue sections can be H&E stained prior to library preparation, thus the transcriptomes of the spots can be mapped to H&E tissue morphology. H&E is also commonly used in clinical pathology, while ST and Visium are not used for diagnostic purposes. The package ST-Net was developed to use a pretrained CNN to extract features from H&E images behind the ST spots, and a dense neural net is trained on the extracted features and ST data to predict gene expression based on H&E images from held out patients as log normalized UMI counts [88]. Another method to predict gene expression from H&E is PathoMCH [120]. TGCA transcriptomics data is normalized and the corresponding H&E slides are labeled with the percentile of expression of each gene of interest. Then the whole slide images are broken into small tiles, all of which take the percentile label of the slide. Then the Inception v3 classification neural network is trained with the tiles and labels with very high or very low expression, and when predicting on held out images, it gives a score of gene expression from low to high in each tile. Such gene expression prediction methods can give pathologists a more nuanced view of the tissue beyond morphology.

While cell segmentation is difficult in H&E images, H&E images do have enough resolution to exhibit subcellular details. The H&E image is used in Xfuse to increase resolution of the spatial transcriptome from ST [36]. The H&E image and the corresponding transcriptomes are modeled to come from a shared latent space. Intensity of each channel at each pixel is modeled as Gaussian, and gene expression at each pixel is modeled as NB so the observed value at each spot are the sums of values at the pixels in the spot. Parameters of these distributions are mapped from the latent space through a generator CNN. The parameters are estimated with variational Bayesian inference. With the parameters, gene expression at each pixel can be predicted, thus increasing the resolution of ST.

9.12 Prospective users

With so many existing methods, and so many being developed, a user wishing to analyze spatial transcriptomics data would naturally ask, "Which method shall I use?" We suggest considering the following factors:

First, benchmarks have been performed for types of data analysis that garnered more attention (Figure reffig:analysis-cats). The benchmarking methodologies should be read carefully, as multiple benchmarks of the same type of methods can give different results. For cell type deconvolution, there seems to be a consensus among different benchmarks that cell2location, RCTD, spatialDWLS, and Tangram are top performing methods [121, 122, 123, 124]. Cell-cell interaction methods have been benchmarked here [125]. Data integration methods have been benchmarked here [126]. There are also tools to make simulated data for benchmarking, some simulating spatial distributions in addition to gene expression values [127]. However, benchmarks of spatially variable gene methods show that the different methods give disparate results, so it's not straightforward to say which method is the best [128, 129].

Second, suppose that the implementation is available (which is usually, but not always, the case), then is the implementation of the method usable? Sometimes the authors only post some scripts on GitHub that are not bundled into a package that can be easily installed. In this case, the scripts would need to be copied to the working directory of the analyses and into any future package using them. If the authors decide to update the scripts, it would be more difficult for the users to update as well or even to know about the update. In the Analysis sheet of our database, there is a metadata column "packaged", which indicates whether the scripts are bundled into a package.

Third, it's easier to learn to use a well-documented package. As shown in Figure 9.12, most packages written in R and Python are reasonably documented, which means that all arguments of functions exposed to the users are documented. However, not all of those packages have usage examples. Not all packages in our database are mentioned in this chapter; the metadata column "documented" in the database indicates whether each package is reasonably documented.

Fourth, is the package still maintained? Take caution if the package has not been updated for years, because updates in dependencies can break the code. This is a reason why CRAN and Bioconductor perform daily checks of the packages and notify the maintainer if the package fails automated checks and unit tests and if the

maintainer fails to correct the problem, the package will be removed from CRAN or Bioconductor. For this reason, if your package based on other packages is aimed for CRAN or Bioconductor, then all R dependencies must also be on CRAN or Bioconductor, and all Python dependencies (called from basilisk) must be on PyPI or conda. As most packages in our database are not on any of CRAN, Bioconductor, PyPI, or conda (Figure 9.13), this requirement constrains which methods to build upon for CRAN and Bioconductor packages. But with such requirements, the packages are more likely to be usable. That said, GitHub only package can still have high quality and be usable.

So far these factors are about usability of the implementation, but we look forward to more third-party benchmarks to compare the performances, while taking into account usability. Another factor to consider is computational speed. While there is no systematic third-party benchmark, as a rule of thumb, packages that do MCMC tend to take a long time to run due to the nature of MCMC. Gaussian process-based methods also tend to take a long time to run though often it isn't too bad. Neural network based methods tend to take a long time to run on CPU. To try out different packages, as no method performs the best in all datasets and all scenarios, perhaps try on a smaller subset of data if the packages tend to take a long time to run.

References

1. Gyllborg D, Langseth CM, Qian X, Salas SM, Hilscher M, Lein E, and Nilsson M. Hybridization-based In Situ Sequencing (HybISS): spatial transcriptomic detection in human and mouse brain tissue. bioRxiv 2020 Feb :2020.02.03.931618. DOI: 10.1101/2020.02.03.931618. Available from: https://doi.org/10.1101/2020.02.03.931618

2. Qian X, Harris KD, Hauling T, Nicoloutsopoulos D, Muñoz-Manchado AB, Skene N, Hjerling-Leffler J, and Nilsson M. Probabilistic cell typing enables fine mapping of closely related cell types in situ. Nature Methods 2020 Jan; 17:101–6. DOI: 10.1038/s41592-019-0631-4. Available from: https://doi.org/10.1038/s41592-019-0631-4

3. Shah S, Takei Y, Zhou W, Lubeck E, Yun J, Eng CHL, Koulena N, Cronin C, Karp C, Liaw EJ, Amin M, and Cai L. Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. Cell 2018. DOI: 10.1016/j.cell.2018.05.035

4. Chen WT, Lu A, Craessaerts K, Pavie B, Sala Frigerio C, Corthout N, Qian X, Laláková J, Kühnemund M, Voytyuk I, Wolfs L, Mancuso R, Salta E, Balusu S, Snellinx A, Munck S, Jurek A, Fernandez Navarro J, Saido TC, Huitinga I, Lundeberg J, Fiers M, and De Strooper B. Spatial Transcriptomics

and In Situ Sequencing to Study Alzheimer's Disease. Cell 2020 Jul; 0. DOI: 10.1016/j.cell.2020.06.038. Available from: http://www.cell.com/article/S0092867420308151/fulltext

5. Sountoulidis A, Liontos A, Nguyen HP, Firsova AB, Fysikopoulos A, Qian X, Seeger W, Sundström E, Nilsson M, and Samakovlis C. SCRINSHOT, a spatial method for single-cell resolution mapping of cell states in tissue sections. bioRxiv 2020 Feb :2020.02.07.938571. DOI: 10.1101/2020.02.07.938571. Available from: https://www.biorxiv.org/content/biorxiv/early/2020/02/07/2020.02.07.938571.full.pdf

6. Yang X, Bergenholtz S, Maliskova L, Pebworth MP, Kriegstein AR, Li Y, and Shen Y. SMART-Q: An Integrative Pipeline Quantifying Cell Type-Specific RNA Transcription. PLOS ONE 2020 Apr; 15:e0228760. Available from: https://doi.org/10.1371/journal.pone.0228760

7. Maynard KR, Tippani M, Takahashi Y, Phan BN, Hyde TM, Jaffe AE, and Martinowich K. dotdotdot: an automated approach to quantify multiplex single molecule fluorescent in situ hybridization (smFISH) images in complex tissues. Nucleic Acids Research 2020 Jun; 48:e66–e66. DOI: 10.1093/nar/gkaa312. Available from: https://doi.org/10.1093/nar/gkaa312

8. Chen KH, Boettiger AN, Moffitt JR, Wang S, and Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 2015. DOI: 10.1126/science.aaa6090

9. Eng CHL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, Yun J, Cronin C, Karp C, Yuan GC, and Cai L. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature 2019 Apr; 568:235–9. DOI: 10.1038/s41586-019-1049-y. Available from: https://www.nature.com/articles/s41586-019-1049-y

10. Shah S, Lubeck E, Zhou W, and Cai L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. Neuron 2016. DOI: 10.1016/j.neuron.2016.10.001

11. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, Nolan GP, Bava FA, and Deisseroth K. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science 2018 Jul; 361:eaat5691. DOI: 10.1126/science.aat5691. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aat5691

12. Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, Rubinstein ND, Hao J, Regev A, Dulac C, and Zhuang X. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science 2018. DOI: 10.1126/science.aau5324

13. Partel G, Hilscher M, Milli G, Solorzano L, Klemm A, Nilsson M, and Wählby C. Identification of spatial compartments in tissue from in situ sequencing data. bioRxiv 2019 Sep :765842. DOI: 10.1101/765842. Available from: https://doi.org/10.1101/765842

14. Eichenberger BT, Zhan Y, Rempfler M, Giorgetti L, and Chao JA. deepBlink: Threshold-independent detection and localization of diffraction-limited spots. bioRxiv 2020 Jan :2020.12.14.422631. DOI: 10.1101/2020.12.14.422631. Available from: http://biorxiv.org/content/early/2020/12/15/2020.12.14.422631.abstract

15. Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, and Zhuang X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. Proceedings of the National Academy of Sciences of the United States of America 2016. DOI: 10.1073/pnas.1612826113

16. Köster J, Brown M, and Liu XS. A Bayesian model for single cell transcript expression analysis on MERFISH data. Bioinformatics 2019. DOI: 10.1093/bioinformatics/bty718

17. Middleton SA, Eberwine J, and Kim J. Comprehensive catalog of dendritically localized mRNA isoforms from sub-cellular sequencing of single mouse neurons. BMC Biology 2019; 17:5. DOI: 10.1186/s12915-019-0630-z. Available from: https://doi.org/10.1186/s12915-019-0630-z

18. Ciolli Mattioli C, Rom A, Franke V, Imami K, Arrey G, Terne M, Woehler A, Akalin A, Ulitsky I, and Chekulaeva M. Alternative 3' UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. Nucleic Acids Research 2019 Mar; 47:2560–73. DOI: 10.1093/nar/gky1270. Available from: https://doi.org/10.1093/nar/gky1270

19. Farris S, Ward JM, Carstens KE, Samadi M, Wang Y, and Dudek SM. Hippocampal Subregions Express Distinct Dendritic Transcriptomes that Reveal Differences in Mitochondrial Function in CA2. Cell Reports 2019 Oct; 29:522–39. DOI: 10.1016/j.celrep.2019.08.093. Available from: https://doi.org/10.1016/j.celrep.2019.08.093

20. Lohoff T, Ghazanfar S, Missarova A, Koulena N, Pierson N, Griffiths JA, Bardot ES, Eng CH, Tyser RC, Argelaguet R, Guibentif C, Srinivas S, Briscoe J, Simons BD, Hadjantonakis AK, Göttgens B, Reik W, Nichols J, Cai L, and Marioni JC. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. Nature Biotechnology 2021 40:1 2021 Sep; 40:74–85. DOI: 10.1038/s41587-021-01006-2. Available from: https://www.nature.com/articles/s41587-021-01006-2

21. Van Valen DA, Kudo T, Lane KM, Macklin DN, Quach NT, DeFelice MM, Maayan I, Tanouchi Y, Ashley EA, and Covert MW. Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging

Experiments. PLOS Computational Biology 2016 Nov; 12:e1005177. Available from: https://doi.org/10.1371/journal.pcbi.1005177

22. Stringer C, Wang T, Michaelos M, and Pachitariu M. Cellpose: a generalist algorithm for cellular segmentation. Nature Methods 2020 18:1 2020 Dec; 18:100–6. DOI: 10.1038/s41592-020-01018-x. Available from: https://www.nature.com/articles/s41592-020-01018-x

23. Littman R, Hemminger Z, Foreman R, Arneson D, Zhang G, Gómez-Pinilla F, Yang X, and Wollman R. JSTA: joint cell segmentation and cell type annotation for spatial transcriptomics. bioRxiv 2020 Jan :2020.09.18.304147. DOI: 10.1101/2020.09.18.304147. Available from: http://biorxiv.org/content/early/2020/09/20/2020.09.18.304147.abstract

24. Petukhov V, Soldatov RA, Khodosevich K, and Kharchenko PV. Bayesian segmentation of spatially resolved transcriptomics data. bioRxiv 2020 Jan :2020.10.05.326777. DOI: 10.1101/2020.10.05.326777. Available from: http://biorxiv.org/content/early/2020/10/06/2020.10.05.326777.abstract

25. Park J, Choi W, Tiesmeyer S, Long B, Borm LE, Garren E, Nguyen TN, Codeluppi S, Schlesner M, Tasic B, Eils R, and Ishaque N. Segmentation-free inference of cell types from in situ transcriptomics data. bioRxiv 2019. DOI: 10.1101/800748

26. Partel G and Wählby C. Spage2vec: Unsupervised detection of spatial gene expression constellations. bioRxiv 2020 Feb :2020.02.12.945345. DOI: 10.1101/2020.02.12.945345. Available from: https://doi.org/10.1101/2020.02.12.945345

27. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, and Satija R. Comprehensive Integration of Single-Cell Data. Cell 2019 Jun; 177:1888–902. DOI: 10.1016/j.cell.2019.05.031. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867419305598

28. Zhang M, Eichhorn SW, Zingg B, Yao Z, Zeng H, Dong H, and Zhuang X. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics. bioRxiv 2020 Jan :2020.06.04.105700. DOI: 10.1101/2020.06.04.105700. Available from: http://biorxiv.org/content/early/2020/06/05/2020.06.04.105700.abstract

29. Moncada R, Barkley D, Wagner F, Chiodin M, Devlin JC, Baron M, Hajdu CH, Simeone DM, and Yanai I. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nature Biotechnology 2020; 38:333–42. DOI: 10.1038/s41587-019-0392-8. Available from: https://doi.org/10.1038/s41587-019-0392-8

30. Berglund E, Maaskola J, Schultz N, Friedrich S, Marklund M, Bergenstråhle J, Tarish F, Tanoglidi A, Vickovic S, Larsson L, Salmén F, Ogris C, Wallenborg K, Lagergren J, Ståhl P, Sonnhammer E, Helleday T, and Lundeberg J. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nature Communications 2018. DOI: 10.1038/s41467-018-04724-5

31. Ji AL, Rubin AJ, Thrane K, Jiang S, Reynolds DL, Meyers RM, Guo MG, George BM, Mollbrink A, Bergenstråhle J, Larsson L, Bai Y, Zhu B, Bhaduri A, Meyers JM, Rovira-Clavé X, Hollmig ST, Aasi SZ, Nolan GP, Lundeberg J, and Khavari PA. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell 2020; 182:497–514. DOI: https://doi.org/10.1016/j.cell.2020.05.039. Available from: http://www.sciencedirect.com/science/article/pii/S0092867420306723

32. Mantri M, Scuderi GJ, Nassab RA, Wang MFZ, McKellar D, Butcher JT, and De Vlaminck I. Spatiotemporal single-cell RNA sequencing of developing hearts reveals interplay between cellular differentiation and morphogenesis. bioRxiv 2020 Jan :2020.05.03.065102. DOI: 10.1101/2020.05.03.065102. Available from: http://biorxiv.org/content/early/2020/05/03/2020.05.03.065102.abstract

33. Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, Rybakov S, Ibarra IL, Holmberg O, Virshup I, Lotfollahi M, Richter S, and Theis FJ. Squidpy: a scalable framework for spatial omics analysis. en. Nat. Methods 2022 Feb; 19:171–8

34. Righelli D, Weber LM, Crowell HL, Pardo B, Collado-Torres L, Ghazanfar S, Lun AT, Hicks SC, and Risso D. SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor. Bioinformatics 2022 May; 38:3128–31. DOI: 10.1093/BIOINFORMATICS/BTAC299. Available from: https://academic.oup.com/bioinformatics/article/38/11/3128/6575443

35. Dries R, Zhu Q, Dong R, Eng Chee-Huat Linus, Li H, Liu K, Fu Y, Zhao Tianxiao, Sarkar A, Bao F, George RE, Pierson N, Cai L, and Yuan GC. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. en. Genome Biol. 2021 Mar; 22:78

36. Bergenstråhle L, He B, Bergenstråhle J, Andersson A, Lundeberg J, Zou J, and Maaskola J. Super-resolved spatial transcriptomics by deep data fusion. bioRxiv 2020 Mar :2020.02.28.963413. DOI: 10.1101/2020.02.28.963413. Available from: https://www.biorxiv.org/content/10.1101/2020.02.28.963413v1.full

37. Kueckelhaus J, Ehr J von, Ravi VM, Will P, Joseph K, Beck J, Hofmann UG, Delev D, Schnell O, and Heiland DH. Inferring spatially transient gene expression pattern from spatial transcriptomic studies. bioRxiv 2020 Jan

:2020.10.20.346544. DOI: 10.1101/2020.10.20.346544. Available from: http://biorxiv.org/content/early/2020/10/21/2020.10.20.346544.abstract

38. Svensson V, Teichmann SA, and Stegle O. SpatialDE: Identification of spatially variable genes. Nature Methods 2018 Apr; 15:343–6. DOI: 10.1038/nmeth.4636. Available from: https://www.nature.com/articles/nmeth.4636

39. Edsgärd D, Johnsson P, and Sandberg R. Identification of spatial expression trends in single-cell gene expression data. Nature Methods 2018 May; 15:339–42. DOI: 10.1038/nmeth.4634. Available from: http://www.nature.com/articles/nmeth.4634

40. Zhu Q, Shah S, Dries R, Cai L, and Yuan GC. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nature Biotechnology 2018 Dec; 36:1183–90. DOI: 10.1038/nbt.4260. Available from: https://www.nature.com/articles/nbt.4260

41. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, Trapnell C, and Shendure J. The single-cell transcriptional landscape of mammalian organogenesis. Nature 2019; 566:496–502. DOI: 10.1038/s41586-019-0969-x. Available from: https://doi.org/10.1038/s41586-019-0969-x

42. Sun S, Zhu J, and Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nature Methods 2020 Feb; 17:193–200. DOI: 10.1038/s41592-019-0701-7. Available from: https://doi.org/10.1038/s41592-019-0701-7

43. Bergenstråhle J, Bergenstråhle L, and Lundeberg J. SpatialCPie: an R/Bioconductor package for spatial transcriptomics cluster evaluation. BMC Bioinformatics 2020 Dec; 21:161. DOI: 10.1186/s12859-020-3489-7. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-3489-7

44. Svensson V, Veiga Beltrame E da, and Pachter L. A curated database reveals trends in single-cell transcriptomics. Database 2020 Jan; 2020. DOI: 10.1093/database/baaa073. Available from: https://doi.org/10.1093/database/baaa073

45. Durruthy-Durruthy R, Gottlieb A, Hartman BH, Waldhaus J, Laske RD, Altman R, and Heller S. Reconstruction of the Mouse Otocyst and Early Neuroblast Lineage at Single-Cell Resolution. Cell 2014 May; 157:964–78. DOI: 10.1016/j.cell.2014.03.036. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867414004115

46. Durruthy-Durruthy J, Wossidlo M, Pai S, Takahashi Y, Kang G, Omberg L, Chen B, Nakauchi H, Reijo Pera R, and Sebastiano V. Spatiotemporal Reconstruction of the Human Blastocyst by Single-Cell Gene-Expression Analysis Informs Induction of Naive Pluripotency. Developmental Cell 2016 Jul; 38:100–15. DOI: 10.1016/j.devcel.2016.06.014. Available from: https://doi.org/10.1016/j.devcel.2016.06.014

47. Mori T, Yamane J, Kobayashi K, Taniyama N, Tano T, and Fujibuchi W. Development of 3D Tissue Reconstruction Method from Single-cell RNA-seq Data. Genomics and Computational Biology; Vol 3 No 1 (2017) 2017. DOI: 10.18547/gcb.2017.vol3.iss1.e53. Available from: https://genomicscomputbiol.org/ojs3/GCB/article/view/23

48. Waldhaus J, Durruthy-Durruthy R, and Heller S. Quantitative High-Resolution Cellular Map of the Organ of Corti. Cell Reports 2015 Jun; 11:1385–99. DOI: 10.1016/j.celrep.2015.04.062. Available from: https://doi.org/10.1016/j.celrep.2015.04.062

49. Mori T, Takaoka H, Yamane J, Alev C, and Fujibuchi W. Novel computational model of gastrula morphogenesis to identify spatial discriminator genes by self-organizing map (SOM) clustering. Scientific Reports 2019 Dec; 9:1–10. DOI: 10.1038/s41598-019-49031-1. Available from: https://doi.org/10.1038/s41598-019-49031-1

50. Peng G, Suo S, Chen J, Chen W, Liu C, Yu F, Wang R, Chen S, Sun N, Cui G, Song L, Tam PP, Han JDJ, and Jing N. Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. Developmental Cell 2016 Mar; 36:681–97. DOI: 10.1016/j.devcel.2016.02.020. Available from: http://dx.doi.org/10.1016/j.devcel.2016.02.020

51. Zhu J and Sabatti C. Integrative Spatial Single-cell Analysis with Graph-based Feature Learning. bioRxiv 2020 Jan :2020.08.12.248971. DOI: 10.1101/2020.08.12.248971. Available from: http://biorxiv.org/content/early/2020/08/13/2020.08.12.248971.abstract

52. Ren X, Zhong G, Zhang Q, Zhang L, Sun Y, and Zhang Z. Reconstruction of cell spatial organization based on ligand-receptor mediated self-assembly. bioRxiv 2020 Jan :2020.02.13.948521. DOI: 10.1101/2020.02.13.948521. Available from: http://biorxiv.org/content/early/2020/02/14/2020.02.13.948521.abstract

53. Achim K, Pettit JB, Saraiva LR, Gavriouchkina D, Larsson T, Arendt D, and Marioni JC. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nature Biotechnology 2015 May; 33:503–9. DOI: 10.1038/nbt.3209. Available from: https://doi.org/10.1038/nbt.3209

54. Karaiskos N, Wahle P, Alles J, Boltengagen A, Ayoub S, Kipar C, Kocks C, Rajewsky N, and Zinzen RP. The Drosophila embryo at single-cell transcriptome resolution. Science 2017 Oct; 358:194–9. DOI: 10.1126/science.aan3235. Available from: https://science.sciencemag.org/content/358/6360/194

55. Tanevski J, Nguyen T, Truong B, Karaiskos N, Ahsen ME, Zhang X, Shu C, Xu K, Liang X, Hu Y, Pham HVV, Xiaomei L, Le TD, Tarca AL, Bhatti G, Romero R, Karathanasis N, Loher P, Chen Y, Ouyang Z, Mao D, Zhang Y, Zand M, Ruan J, Hafemeister C, Qiu P, Tran D, Nguyen T, Gabor A, Yu T, Guinney J, Glaab E, Krause R, Banda P, Stolovitzky G, Rajewsky N, Saez-Rodriguez J, and Meyer P. Gene selection for optimal prediction of cell position in tissues from single-cell transcriptomics data. Life Science Alliance 2020 Nov; 3:e202000867. DOI: 10.26508/lsa.202000867. Available from: http://www.life-science-alliance.org/content/3/11/e202000867.abstract

56. Alonso AM, Carrea A, and Diambra L. Prediction of cell position using single-cell transcriptomic data: an iterative procedure [version 2; peer review: 2 approved]. F1000Research 2020; 8. DOI: 10.12688/f1000research.20715.2. Available from: http://openr.es/jqr

57. Pham D, Tan X, Xu J, Grice LF, Lam PY, Raghubar A, Vukovic J, Ruitenberg MJ, and Nguyen Q. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. bioRxiv 2020 Jan :2020.05.31.125658. DOI: 10.1101/2020.05.31.125658. Available from: http://biorxiv.org/content/early/2020/05/31/2020.05.31.125658.abstract

58. Satija R, Farrell JA, Gennert D, Schier AF, and Regev A. Spatial reconstruction of single-cell gene expression data. Nature Biotechnology 2015 May; 33:495–502. DOI: 10.1038/nbt.3192. Available from: https://doi.org/10.1038/nbt.3192

59. Halpern KB, Shenhav R, Matcovitch-Natan O, Tóth B, Lemze D, Golan M, Massasa EE, Baydatch S, Landen S, Moor AE, Brandis A, Giladi A, Stokar-Avihail A, David E, Amit I, and Itzkovitz S. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. Nature 2017 Feb; 542:1–5. DOI: 10.1038/nature21065. Available from: https://www.nature.com/articles/nature21065

60. Halpern KB, Shenhav R, Massalha H, Toth B, Egozi A, Massasa EE, Medgalia C, David E, Giladi A, Moor AE, Porat Z, Amit I, and Itzkovitz S. Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. Nature Biotechnology 2018 Nov; 36:962. DOI: 10.1038/nbt.4231. Available from: https://www.nature.com/articles/nbt.4231

61. Droin C, Kholtei JE, Halpern KB, Hurni C, Rozenberg M, Muvkadi S, Itzkovitz S, and Naef F. Space-time logic of liver gene expression at sublobular scale. bioRxiv 2020 Jan :2020.03.05.976571. DOI: 10.1101/2020.03.05.976571. Available from: http://biorxiv.org/content/early/2020/10/09/2020.03.05.976571.abstract

62. Lopez R, Nazaret A, Langevin M, Samaran J, Regier J, Jordan MI, and Yosef N. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. 2019 May. Available from: http://arxiv.org/abs/1905.02269

63. Verma A and Engelhardt B. A Bayesian nonparametric semi-supervised model for integration of multiple single-cell experiments. bioRxiv 2020 Jan :2020.01.14.906313. DOI: 10.1101/2020.01.14.906313. Available from: https://doi.org/10.1101/2020.01.14.906313

64. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, and Macosko EZ. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell 2019 Jun; 177:1873–87. DOI: 10.1016/j.cell.2019.05.006. Available from: https://doi.org/10.1016/j.cell.2019.05.006

65. Abdelaal T, Mourragui S, Mahfouz A, and Reinders MJT. SpaGE: Spatial Gene Enhancement using scRNA-seq. Nucleic Acids Research 2020 Oct; 48:e107–e107. DOI: 10.1093/nar/gkaa740. Available from: https://doi.org/10.1093/nar/gkaa740

66. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh Pr, and Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. Nature Methods 2019; 16:1289–96. DOI: 10.1038/s41592-019-0619-0. Available from: https://doi.org/10.1038/s41592-019-0619-0

67. Li J, Luo H, Wang R, Lang J, Zhu S, Zhang Z, Fang J, Qu K, Lin Y, Long H, Yao Y, Tian G, and Wu Q. Systematic Reconstruction of Molecular Cascades Regulating GP Development Using Single-Cell RNA-Seq. Cell Reports 2016 May; 15:1467–80. DOI: 10.1016/j.celrep.2016.04.043. Available from: https://doi.org/10.1016/j.celrep.2016.04.043

68. Bravo González-Blas C, Quan XJ, Duran-Romaña R, Taskiran II, Koldere D, Davie K, Christiaens V, Makhzami S, Hulselmans G, Waegeneer M de, Mauduit D, Poovathingal S, Aibar S, and Aerts S. Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. Molecular Systems Biology 2020 May; 16:e9438. DOI: 10.15252/msb.20209438. Available from: https://doi.org/10.15252/msb.20209438

69. Ortiz C, Navarro JF, Jurek A, Märtin A, Lundeberg J, and Meletis K. Molecular atlas of the adult mouse brain. Science Advances 2020 Jun; 6:eabb3446. DOI: 10.1126/sciadv.abb3446. Available from: www.brain-map.org%20https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.abb3446

70. Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z, Sanger A, Tokcan N, Vanderburg CR, Segerstolpe A, Zhang M, Avraham-Davidi I, Vickovic S, Nitzan M, Ma S, Buenrostro J, Brown NB, Fanelli D, Zhuang X, Macosko EZ, and Regev A. Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with Tangram. bioRxiv 2020 Jan :2020.08.29.272831. DOI: 10.1101/2020.08.29.272831. Available from: http://biorxiv.org/content/early/2020/09/24/2020.08.29.272831.abstract

71. Nitzan M, Karaiskos N, Friedman N, and Rajewsky N. Gene expression cartography. Nature 2019 Dec; 576:132–7. DOI: 10.1038/s41586-019-1773-3. Available from: https://doi.org/10.1038/s41586-019-1773-3

72. Cang Z and Nie Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. Nature Communications 2020; 11:2084. DOI: 10.1038/s41467-020-15968-5. Available from: https://doi.org/10.1038/s41467-020-15968-5

73. Li Z, Song T, Yong J, and Kuang R. Imputation of Spatially-resolved Transcriptomes by Graph-regularized Tensor Completion. bioRxiv 2020 Jan :2020.08.05.237560. DOI: 10.1101/2020.08.05.237560. Available from: http://biorxiv.org/content/early/2020/08/05/2020.08.05.237560.abstract

74. Andersson A, Bergenstråhle J, Asp M, Bergenstråhle L, Jurek A, Fernández Navarro J, and Lundeberg J. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. Communications Biology 2020; 3:565. DOI: 10.1038/s42003-020-01247-y. Available from: https://doi.org/10.1038/s42003-020-01247-y

75. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, Lomakin A, Kedlian V, Jain MS, Park JS, Ramona L, Tuck E, Arutyunyan A, Vento-Tormo R, Gerstung M, James L, Stegle O, and Bayraktar OA. Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics. bioRxiv 2020 Jan :2020.11.15.378125. DOI: 10.1101/2020.11.15.378125. Available from: http://biorxiv.org/content/early/2020/11/17/2020.11.15.378125.abstract

76. Lopez R, Li B, Keren-Shaul H, Boyeau P, Kedmi M, Pilzer D, Jelinski A, David E, Wagner A, Addad Y, Jordan MI, Amit I, and Yosef N. Multi-resolution deconvolution of spatial transcriptomics data reveals continuous patterns of inflammation. bioRxiv 2021 Jan :2021.05.10.443517. DOI:

10.1101/2021.05.10.443517. Available from: http://biorxiv.org/content/early/2021/05/11/2021.05.10.443517.abstract

77. Yang T, Alessandri-Haber N, Fury W, Schaner M, Breese R, LaCroix-Fralish M, Kim J, Adler C, Macdonald LE, Atwal GS, and Bai Y. AdRoit: an accurate and robust method to infer complex transcriptome composition. bioRxiv 2021 Jan :2020.12.14.422697. DOI: 10.1101/2020.12.14.422697. Available from: http://biorxiv.org/content/early/2021/07/16/2020.12.14.422697.abstract

78. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, and Irizarry RA. Robust decomposition of cell type mixtures in spatial transcriptomics. bioRxiv 2020 May :2020.05.07.082750. DOI: 10.1101/2020.05.07.082750. Available from: https://doi.org/10.1101/2020.05.07.082750

79. Danaher P, Kim Y, Nelson B, Griswold M, Yang Z, Piazza E, and Beechem JM. Advances in mixed cell deconvolution enable quantification of cell types in spatially-resolved gene expression data. bioRxiv 2020 Jan :2020.08.04.235168. DOI: 10.1101/2020.08.04.235168. Available from: http://biorxiv.org/content/early/2020/09/29/2020.08.04.235168.abstract

80. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, and Macosko EZ. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science 2019. DOI: 10.1126/science.aaw1219

81. Elosua M, Nieto P, Mereu E, Gut I, and Heyn H. SPOTlight: Seeded NMF regression to Deconvolute Spatial Transcriptomics Spots with Single-Cell Transcriptomes. bioRxiv 2020 Jun :2020.06.03.131334. DOI: 10.1101/2020.06.03.131334. Available from: https://www.biorxiv.org/content/10.1101/2020.06.03.131334v1

82. Sun D, Xiao Y, Liu Z, Li T, Wu Q, and Wang C. STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing. bioRxiv 2021 Jan :2021.09.08.459458. DOI: 10.1101/2021.09.08.459458. Available from: http://biorxiv.org/content/early/2021/09/09/2021.09.08.459458.abstract

83. Miller BF, Atta L, Sahoo A, Huang F, and Fan J. Reference-free cell-type deconvolution of pixel-resolution spatially resolved transcriptomics data. bioRxiv 2021 Jan :2021.06.15.448381. DOI: 10.1101/2021.06.15.448381. Available from: http://biorxiv.org/content/early/2021/06/16/2021.06.15.448381.abstract

84. Su J and Song Q. DSTG: Deconvoluting Spatial Transcriptomics Data through Graph-based Artificial Intelligence. bioRxiv 2020 Jan :2020.10.20.347195. DOI: 10.1101/2020.10.20.347195. Available from: http://biorxiv.org/content/early/2020/10/21/2020.10.20.347195.abstract

85. BinTayyash N, Georgaka S, John ST, Ahmed S, Boukouvalas A, Hensman J, and Rattray M. Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments. bioRxiv 2020 Jan :2020.07.29.227207. DOI: 10.1101/2020.07.29.227207. Available from: http://biorxiv.org/content/early/2020/07/30/2020.07.29.227207.abstract

86. Li Q, Zhang M, Xie Y, and Xiao G. Bayesian modeling of spatial molecular profiling data via Gaussian process. Bioinformatics 2021 Jun. DOI: 10.1093/bioinformatics/btab455. Available from: https://doi.org/10.1093/bioinformatics/btab455

87. Hao M, Hua K, and Zhang X. SOMDE: A scalable method for identifying spatially variable genes with self-organizing map. bioRxiv 2021 Jan :2020.12.10.419549. DOI: 10.1101/2020.12.10.419549. Available from: http://biorxiv.org/content/early/2021/03/24/2020.12.10.419549.abstract

88. He B, Bergenstråhle L, Stenbeck L, Abid A, Andersson A, Borg Å, Maaskola J, Lundeberg J, and Zou J. Integrating spatial gene expression and breast tumour morphology via deep learning. Nature Biomedical Engineering 2020 Jun :1–8. DOI: 10.1038/s41551-020-0578-x. Available from: https://doi.org/10.1038/s41551-020-0578-x

89. Govek KW, Yamajala VS, and Camara PG. Clustering-independent analysis of genomic data using spectral simplicial theory. PLOS Computational Biology 2019 Nov; 15:e1007509. Available from: https://doi.org/10.1371/journal.pcbi.1007509

90. Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, and Rabadan R. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. Nature Biotechnology 2017; 35:551–60. DOI: 10.1038/nbt.3854. Available from: https://doi.org/10.1038/nbt.3854

91. Zhang K, Feng W, and Wang P. Identification of spatially variable genes with graph cuts. bioRxiv 2018 Jan :491472. DOI: 10.1101/491472. Available from: http://biorxiv.org/content/early/2018/12/09/491472.1.abstract

92. Miller BF, Bambah-Mukku D, Dulac C, Zhuang X, and Fan J. Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. Genome Research 2021 May. DOI: 10.1101/gr.271288.120. Available from: http://genome.cshlp.org/content/early/2021/09/20/gr.271288.120.abstract

93. Hu J, Li X, Coleman K, Schroeder A, Irwin DJ, Lee EB, Shinohara RT, and Li M. Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. bioRxiv 2020 Jan :2020.11.30.405118. DOI: 10.1101/2020.11.30.

405118. Available from: http://biorxiv.org/content/early/2020/12/02/2020.11.30.405118.abstract

94. Vandenbon A and Diez D. A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. Nature Communications 2020; 11:4318. DOI: 10.1038/s41467-020-17900-3. Available from: https://doi.org/10.1038/s41467-020-17900-3

95. An L, Xie H, Chin MH, Obradovic Z, Smith DJ, and Megalooikonomou V. Analysis of multiplex gene expression maps obtained by voxelation. BMC Bioinformatics 2009 Apr; 10:S10. DOI: 10.1186/1471-2105-10-S4-S10. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-S4-S10

96. Pruteanu-Malinici I, Mace DL, and Ohler U. Automatic Annotation of Spatial Expression Patterns via Sparse Bayesian Factor Models. PLoS Computational Biology 2011 Jul; 7. Ed. by Bader JS:e1002098. DOI: 10.1371/journal.pcbi.1002098. Available from: https://dx.plos.org/10.1371/journal.pcbi.1002098

97. Maaskola J, Bergenstråhle L, Jurek A, Navarro JF, Lagergren J, and Lundeberg J. Charting Tissue Expression Anatomy by Spatial Transcriptome Decomposition. bioRxiv 2018 Dec :362624. DOI: 10.1101/362624. Available from: https://www.biorxiv.org/content/10.1101/362624v2

98. Zhang W, Feng D, Li R, Chernikov A, Chrisochoides N, Osgood C, Konikoff C, Newfeld S, Kumar S, and Ji S. A mesh generation and machine learning framework for Drosophila gene expression pattern image analysis. BMC Bioinformatics 2013 Dec; 14:372. DOI: 10.1186/1471-2105-14-372. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-372

99. Bohland JW, Bokil H, Pathak SD, Lee CK, Ng L, Lau C, Kuan C, Hawrylycz M, and Mitra PP. Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. Methods 2010 Feb; 50:105–12. DOI: 10.1016/j.ymeth.2009.09.001. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1046202309002035

100. Ko Y, Ament SA, Eddy JA, Caballero J, Earls JC, Hood L, and Price ND. Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain. Proceedings of the National Academy of Sciences 2013 Feb; 110:3095–100. DOI: 10.1073/pnas.1222897110. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1222897110

101. Pettit JB, Tomer R, Achim K, Richardson S, Azizi L, and Marioni J. Identifying Cell Types from Spatially Referenced Single-Cell Expression Datasets. PLoS Computational Biology 2014 Sep; 10. Ed. by Morris Q:e1003824.

DOI: 10.1371/journal.pcbi.1003824. Available from: https://dx.plos.org/10.1371/journal.pcbi.1003824

102. Zhao E, Stone MR, Ren X, Guenthoer J, Smythe KS, Pulliam T, Williams SR, Uytingco CR, Taylor SEB, Nghiem P, Bielas JH, and Gottardo R. Spatial transcriptomics at subspot resolution with BayesSpace. en. Nat. Biotechnol. 2021 Nov; 39:1375–84

103. Yang Y, Shi X, Liu W, Zhou Q, Lau MC, Lim JCT, Sun L, Yeong J, and Liu J. SC-MEB: spatial clustering with hidden Markov random field using empirical Bayes. bioRxiv 2021 Jan :2021.06.05.447181. DOI: 10.1101/2021.06.05.447181. Available from: http://biorxiv.org/content/early/2021/09/07/2021.06.05.447181.abstract

104. Dong K and Zhang S. Deciphering spatial domains from spatially resolved transcriptomics with adaptive graph attention auto-encoder. bioRxiv 2021 Jan :2021.08.21.457240. DOI: 10.1101/2021.08.21.457240. Available from: http://biorxiv.org/content/early/2021/08/23/2021.08.21.457240.abstract

105. Chang Y, He F, Wang J, Chen S, Li J, Liu J, Yu Y, Su L, Ma A, Allen C, Lin Y, Sun S, Liu B, Otero J, Chung D, Fu H, Li Z, Xu D, and Ma Q. Define and visualize pathological architectures of human tissues from spatially resolved transcriptomics using deep learning. bioRxiv 2021 Jan :2021.07.08.451210. DOI: 10.1101/2021.07.08.451210. Available from: http://biorxiv.org/content/early/2021/07/16/2021.07.08.451210.abstract

106. Patrick E, Canete NP, Iyengar SS, Harman AN, Sutherland GT, and Yang P. Spatial analysis for highly multiplexed imaging data to identify tissue microenvironments. bioRxiv 2021 Jan :2021.08.16.456469. DOI: 10.1101/2021.08.16.456469. Available from: http://biorxiv.org/content/early/2021/08/17/2021.08.16.456469.abstract

107. Chavent M, Kuentz-Simonet V, Labenne A, and Saracco J. ClustGeo: an R package for hierarchical clustering with spatial constraints. Computational Statistics 2018 Jan; 33:1799–822. DOI: 10.1007/s00180-018-0791-1. Available from: http://dx.doi.org/10.1007/s00180-018-0791-1

108. Zhao F, Jiao L, and Liu H. Kernel Generalized Fuzzy C-Means Clustering with Spatial Information for Image Segmentation. Digit. Signal Process. 2013 Jan; 23:184–99. DOI: 10.1016/j.dsp.2012.09.016. Available from: https://doi.org/10.1016/j.dsp.2012.09.016

109. Canete NP, Iyengar SS, Ormerod JT, Baharlou H, Harman AN, and Patrick E. spicyR: spatial analysis of in situ cytometry data in R. en. Bioinformatics 2022 May; 38:3099–105

110. Efremova M, Vento-Tormo M, Teichmann SA, and Vento-Tormo R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nature Protocols 2020; 15:1484–

506. DOI: 10.1038/s41596-020-0292-x. Available from: https://doi.org/10.1038/s41596-020-0292-x

111. Yuan Y and Bar-Joseph Z. GCNG: Graph convolutional networks for inferring cell-cell interactions. bioRxiv 2019 Jan :2019.12.23.887133. DOI: 10.1101/2019.12.23.887133. Available from: http://biorxiv.org/content/early/2019/12/23/2019.12.23.887133.abstract

112. Jerby-Arnon L and Regev A. Mapping multicellular programs from single-cell profiles. bioRxiv 2020 Jan :2020.08.11.245472. DOI: 10.1101/2020.08.11.245472. Available from: http://biorxiv.org/content/early/2020/08/11/2020.08.11.245472.abstract

113. Arnol D, Schapiro D, Bodenmiller B, Saez-Rodriguez J, and Stegle O. Modeling Cell-Cell Interactions from Spatial Molecular Data with Spatial Variance Component Analysis. Cell Reports 2019 Oct; 29:202–11. DOI: 10.1016/j.celrep.2019.08.077. Available from: https://doi.org/10.1016/j.celrep.2019.08.077

114. Ghazanfar S, Lin Y, Su X, Lin DM, Patrick E, Han ZG, Marioni JC, and Yang JYH. Investigating higher-order interactions in single-cell data with scHOT. Nature Methods 2020; 17:799–806. DOI: 10.1038/s41592-020-0885-x. Available from: https://doi.org/10.1038/s41592-020-0885-x

115. Cabili MN, Dunagin MC, McClanahan PD, Biaesch A, Padovan-Merhar O, Regev A, Rinn JL, and Raj A. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. Genome Biology 2015 Dec; 16:20. DOI: 10.1186/s13059-015-0586-4. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0586-4

116. Battich N, Stoeger T, and Pelkmans L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. Nature Methods 2013 Nov; 10:1127–36. DOI: 10.1038/nmeth.2657. Available from: https://www.nature.com/articles/nmeth.2657

117. Samacoits A, Chouaib R, Safieddine A, Traboulsi AM, Ouyang W, Zimmer C, Peter M, Bertrand E, Walter T, and Mueller F. A computational framework to study sub-cellular RNA localization. Nature Communications 2018 Dec; 9:4584. DOI: 10.1038/s41467-018-06868-w. Available from: http://www.nature.com/articles/s41467-018-06868-w

118. Xia C, Babcock HP, Moffitt JR, and Zhuang X. Multiplexed detection of RNA using MERFISH and branched DNA amplification. Scientific Reports 2019 Dec; 9:1–13. DOI: 10.1038/s41598-019-43943-8. Available from: https://www.nature.com/articles/s41598-019-43943-8

119. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, Fan J, Borm LE, Liu Z, Bruggen D van, Guo J, He X, Barker R, Sundström E, Castelo-Branco G,

Cramer P, Adameyko I, Linnarsson S, and Kharchenko PV. RNA velocity of single cells. Nature 2018; 560:494–8. DOI: 10.1038/s41586-018-0414-6. Available from: https://doi.org/10.1038/s41586-018-0414-6

120. Levy-Jurgenson A, Tekpli X, Kristensen VN, and Yakhini Z. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. Scientific Reports 2020; 10:18802. DOI: 10.1038/s41598-020-75708-z. Available from: https://doi.org/10.1038/s41598-020-75708-z

121. Li H, Zhou J, Li Z, Chen S, Liao X, Zhang B, Zhang R, Wang Y, Sun S, and Gao X. A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. Nature Communications 2023 14:1 2023 Mar; 14:1–10. DOI: 10.1038/s41467-023-37168-7. Available from: https://www.nature.com/articles/s41467-023-37168-7

122. Yan L and Sun X. Benchmarking and integration of methods for deconvoluting spatial transcriptomic data. Bioinformatics 2023 Jan; 39. DOI: 10.1093/BIOINFORMATICS/BTAC805. Available from: https://academic.oup.com/bioinformatics/article/39/1/btac805/6900924

123. Li B, Zhang W, Guo C, Xu H, Li L, Fang M, Hu Y, Zhang X, Yao X, Tang M, Liu K, Zhao X, Lin J, Cheng L, Chen F, Xue T, and Qu K. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. Nature Methods 2022 19:6 2022 May; 19:662–70. DOI: 10.1038/s41592-022-01480-9. Available from: https://www.nature.com/articles/s41592-022-01480-9

124. Chen J, Liu W, Luo T, Yu Z, Jiang M, Wen J, Gupta GP, Giusti P, Zhu H, Yang Y, and Li Y. A comprehensive comparison on cell-type composition inference for spatial transcriptomics data. Briefings in Bioinformatics 2022 Jul; 23. DOI: 10.1093/BIB/BBAC245. Available from: https://academic.oup.com/bib/article/23/4/bbac245/6618233

125. Liu Z, Sun D, and Wang C. Evaluation of cell-cell interaction methods by integrating single-cell RNA sequencing data with spatial information. Genome Biology 2022 Dec; 23:1–38. DOI: 10.1186/S13059-022-02783-Y/FIGURES/9. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02783-y

126. Li Y, Stanojevic S, He B, Jing Z, Huang Q, Kang J, and Garmire LX. Benchmarking Computational Integration Methods for Spatial Transcriptomics Data. bioRxiv 2022 Jan :2021.08.27.457741. DOI: 10.1101/2021.08.27.457741. Available from: https://www.biorxiv.org/content/10.1101/2021.08.27.457741v2%20https://www.biorxiv.org/content/10.1101/2021.08.27.457741v2.abstract

127. Zhu J, Shang L, and Zhou X. SRTsim: spatial pattern preserving simulations for spatially resolved transcriptomics. Genome Biology 2023 Dec; 24:1–30. DOI: 10.1186/S13059-023-02879-Z/FIGURES/7. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-02879-z%20http://creativecommons.org/publicdomain/zero/1.0/

128. Chen C, Kim HJ, and Yang P. Evaluating spatially variable gene detection methods for spatial transcriptomics data. bioRxiv 2022 Nov :2022.11.23.517747. DOI: 10.1101/2022.11.23.517747. Available from: https://www.biorxiv.org/content/10.1101/2022.11.23.517747v1%20https://www.biorxiv.org/content/10.1101/2022.11.23.517747v1.abstract

129. Charitakis N, Salim A, Piers AT, Watt KI, Porrello ER, Elliott DA, Ramialison M, and Elliott D. Disparities in spatially variable gene calling highlight the need for benchmarking spatial transcriptomics methods. bioRxiv 2022 Nov :2022.10.31.514623. DOI: 10.1101/2022.10.31.514623. Available from: https://www.biorxiv.org/content/10.1101/2022.10.31.514623v1%20https://www.biorxiv.org/content/10.1101/2022.10.31.514623v1.abstract

Chapter 10

FROM THE PAST TO THE PRESENT TO THE FUTURE

The quest to profile the transcriptome in space with high resolution is not new. It started with the enhancer and gene trap screens in the late 1980s and the 1990s, before the genomes of metazoans were sequenced. However, in the prequel era, challenges with the existing technology made the dream of profiling the transcriptome in space hard to reach, as the technologies were not highly-multiplexed and not very quantitative. Over 30 years later, this dream seems to be more within reach, though with some caveats. We have come so far, because of so many strands of ideas and technologies coming together since the late 2010s. Highly multiplexed smFISH that can profile 10000 genes at a time would not have been possible without the reference genome sequence to screen for off target binding, the reference transcriptome and genome annotation with which to design the probes, the technology to synthesize DNA oligos, smFISH, confocal microscopy, digital photography, combinatorial barcoding, and the computing resources to store and process terabytes of images. ST and Visium would not have been possible without microarray technology, scRNA-seq techniques designed for small amount of RNA from each spot, NGS, and the computing power to process the data. Some of these strands are older than others, and each of them would not have been possible without more preceding strands coming together. For instance, smFISH would not have been possible without the development of non-radioactive FISH in the late 1970s and the 1980s and techniques to synthesize fluorophore labeled probes. The field of spatial transcriptomics has grown tremendously since the late 2010s, as this is the time when a wide array of technologies truly started to add up to more than the sum of their parts.

Where are we right now in terms of the development of this rapidly unfolding field? Again, we may take inspiration from and draw parallels with development of other technologies that have much longer histories. From such comparisons, we find that the field of spatial transcriptomics is coming of age. First, in several fields, there have been less successful early attempts to achieve the goal of the field that had never become very popular, and the field did not become vastly popular until the right strands of technologies came together. In the history of cycling, the hobby horse and the penny farthing are among such early attempts which were more dangerous and less efficient, and the breakthrough of the safety bicycle, with

the convergence of technologies such as the pneumatic tire, the tangent spoke, and the chain and sprocket, as well as disadvantages of horses and immaturity of the automobile, led to the bike boom in the 1890s, though there are many other important advances such as derailleurs, disc brakes, clipless pedals, and carbon fiber technology, important but not as revolutionary. In the history of elevators, there have been the Archimedes screw and the paternoster, which are no longer commonly seen as passenger elevators due to their disadvantages. The Archimedes screw elevator was very slow and costly, and the paternoster was dangerous. There have also been hand pulled elevators since the era of the Roman Empire. Convergence of several strands of technologies and social changes led to mainstreaming of the elevator, including urbanization, the steam engine, hydraulic propulsion, and electric motors. Here in spatial transcriptomics, considering the drastic growth in the late 2010s, perhaps we may say that for the purpose of profiling expression of large number of genes in tissue, prequel techniques such as enhancer and gene traps, *in situ* reporters, and (WM)ISH are among the less successful early attempts which have never seen the popularity of some current era techniques and which have gone out of favor due to their disadvantages. We have come to a time where the right technologies converge to make achieving the goal of profiling the transcriptome in space efficient enough for a much wider audience, though there are still challenges.

Second, in several fields that are no doubt mature, while many different technologies to solve the same problem are available, a small number of such technologies, often sold by a small number of companies, tend to dominate. This could be a sign of maturity of the field as companies have enough time to become well-established in the field and factors that lead to dominance such as cultural inertia and network effect have enough time and popularity to form. That these companies get to dominate at all means that this field is already popular enough to be profitable. The dominating technologies are not necessarily the best in all rounds and many factors beyond how well the technology or company currently solves the problem (e.g. historical contributions, cost, marketing, cultural inertia, and monopolistic business practices) led to dominance. The obvious example in our field is NGS; while there were many sequencing start-ups and many different ways proposed to make sequencing more efficient in the 1990s (e.g. cPAL and SOLiD as already mentioned, and sequencing by hybridization [1]), today Illumina dominates. For scRNA-seq, while there are Drop-seq, inDrops, CEL-seq, MARS-seq, SMART-seq, and etc., 10X Chromium dominates and is used in most scRNA-seq studies we have come across. When we began curating the database in January 2020, we were surprised by the common

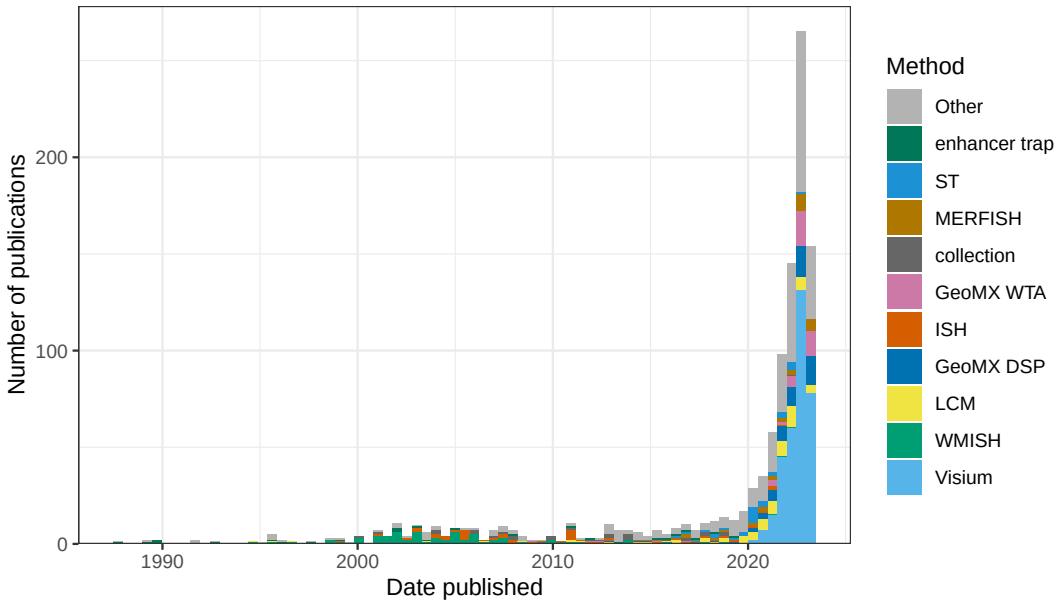

Figure 10.1: Number of publications (including preprints) using each technique to collect new data in both prequel and current era. Only the top 10 in terms of number of publications of all time are colored, and the rest are lumped into Other. Bin width is 180 days, or about half a year. The LCM is for curated LCM literature, which might not be representative of all LCM literature given LCM's long term popularity.

usage of Tomo-seq and LCM, but we have witnessed the rapid rise and spread of Visium and GeoMX DSP over the course of the past year. In the current era, from about 2014 to 2019, a variety of techniques were used to collect new data and it was hard to say which ones dominated. In contrast, since about 2020, a substantial portion of publications for new data used Visium, a portion not previously seen after the golden age of WMISH in the 2000s and since the current era began to take off in the mid 2010s (Figures 10.1, 10.2). With many institutions using their products, 10X and Nanostring have become relatively newly well-established in the field of spatial transcriptomics. Especially for Visium, open source developers (e.g. for Seurat, SpatialExperiment, and BayesSpace) are catering to the output format of Space Ranger (the official preprocessing software for Visium). This is a nod to Visium's establishment akin to the earlier establishment of 10X Chromium and Cell Ranger.

Spatial transcriptomics still faces many challenges. First, there still is the trade-off between quantity and quality. ST and Visium, which have limited resolution and low detection efficiency, can be more easily applied to larger areas of tissue and the whole transcriptome. ISS has been applied to whole mouse brain sections, because

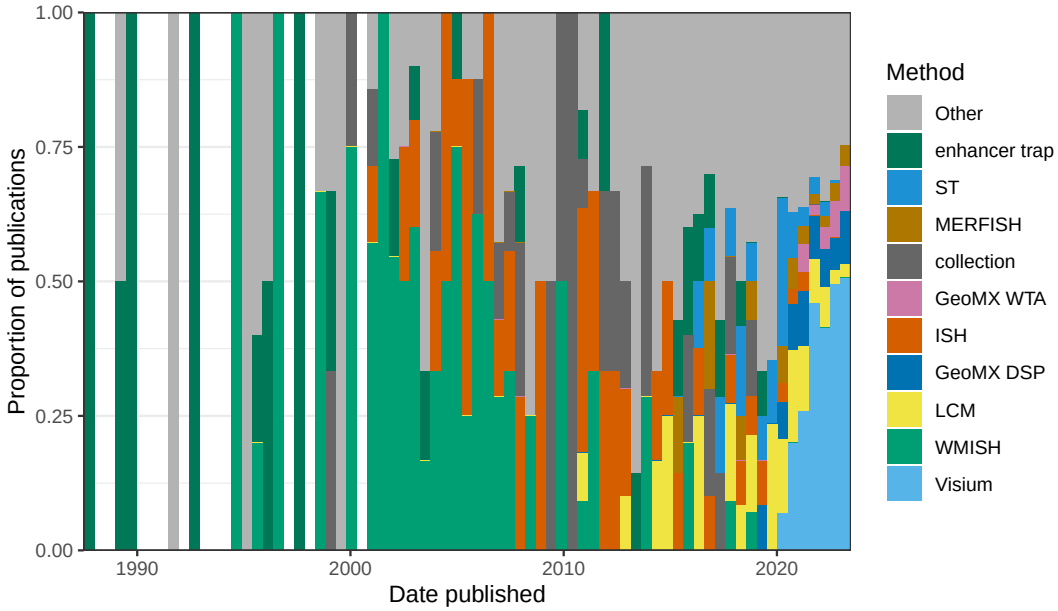

Figure 10.2: Proportion of publications per bin using each of the top 10 techniques for data collection.

while it has lower detection efficiency than smFISH, the amplified and less crowded signals can be detected at lower magnification. In contrast, while smFISH-based techniques have subcellular resolution and often over 80% detection efficiency, the efficiency is compromised when applied to 10000 genes and these techniques are more difficult to apply to larger areas of tissue. As there are still challenges, new techniques to collect data are constantly being developed. Second, compared to the prequel era, the current era is more elitist. While commercial LCM, ST, and Visium have spread far and wide, the various high quality smFISH-based techniques mostly failed to spread beyond their usually elite institutions of origin. This might be due to difficulty in building custom equipment, challenges in customizing the protocols to different tissues, limits in number of genes and cells profiled, lack of core facilities for these techniques, and lack of unified, efficient, open source, and well documented software platform to process the data. However, with the rise of commercial platforms for highly multiplexed smFISH such as MERFISH, Rebus Esper, and Molecular Cartography, this might soon change.

Data analysis has also come a long way, from PCA and ICA in the early 2000s to much more sophisticated techniques today. Many ideas that originated in other fields such as computer vision, machine learning, and statistics, including geospatial statistics, have been adapted to spatial transcriptomics in recent years. Ideas

from computer vision include SIFT, NMF, CNN, and to some extent also PCA and ICA. Ideas from machine learning include SVM, neural networks, bag of words, variational autoencoders (for some cases of latent space), mixture of experts model, k nearest neighbor, and clustering. Ideas from statistics include CCA, permutation testing, MCMC, factor analysis, generalized linear models, and hierarchical modeling. Ideas from geospatial statistics include Gaussian process model (usually used for kriging), spatial point process, and MRF. Other ideas include Laplacian score and optimal transport. Conceivably, more ideas can be adapted to spatial transcriptomics. For instance, spatiotemporal statistics can be adapted to analyze multiple aligned sections of the same tissue to address the difference in covariance between the z axis and the x and y axes. Well established methods in geospatial statistics, such as the semivariogram, J function, G function, and other point process models are also promising for spatial transcriptomics.

We have reviewed many different types of data analysis, using a diverse arsenal of principles. However, integrated analysis pipelines like Seurat are still immature for spatial transcriptomics; Seurat only supports the most rudimentary analyses and the user still needs to learn different syntax and convert data to different formats to use many of the other more specialized and advanced tools, many of which are not well documented. However, the open source culture is flourishing and growing. Most prequel data analysis publications did not link to a repository of the implementation of the software, while most current era data analysis publications do. While the proprietary MATLAB language is still in use, most, especially more recent, current era publication use R, Python, C++, and in some cases Julia and Rust, which are open source and free. Open source software and freely available data may enable less privileged individuals and institutions to perform data analysis and develop new data analysis tools.

What would an ideal future of spatial transcriptomics look like? Data collection would have subcellular resolution, be transcriptome wide, have nearly 100% detection efficiency, and is scalable to large areas of tissues in 3D. Even better, it's multi-omic, profiling not only transcriptome, but also epigenome, proteome, metabolome, etc., with equally high quality and throughput for the other omics. Moreover, the data collection technique is easy to use, such as coming in easy to use kits, and affordable, so it can spread far and wide into non-elite institutions. It should also be open source and transparent, so it would be easier for others to improve it. While we have reviewed many data analysis methods, a comprehensive

benchmark of the methods for each analysis task and evaluation of user experience would be helpful for users to choose a method to use and for developers to compare their new methods to existing methods.

Data analysis would have the same user-friendly user interface for different data types and different methods for the same task. Also, the package should be modular, so dependencies are only installed if needed. It should also be extensible, so users can add additional modules or additional tools for existing tasks to the integrative framework. This would be like SeuratWrappers, which provides Seurat interfaces to data integration and RNA velocity methods not implemented by Seurat. Or like caret and tidymodels, which provide a uniform user interface to numerous machine learning methods. This can be achieved with guidelines such as those used by Bioconductor, encouraging developers to reuse existing data structures and methods in Bioconductor rather than reinventing the wheel. It should also be effective at its task, scalable, well documented, open source, unit tested, easy to install, and portable, again, as enforced to some extent by the Bioconductor guideline. It should be implemented in easy to read code, so developers can more easily fix bugs and improve the package. In addition, it should be interoperable, so tools written in different programming languages can be integrated, combining their strengths and bridging cultural differences between the programming language communities. It should have elegant data visualization, both static for publications and interactive for data exploration and sharing. The data visualization should also be accessible, such as using redundant encoding and colorblind friendly palettes and providing alternatives to those who are visual impaired. Finally, it should be integrated with a graphical user interface (GUI) like iSee so the data can be shared with colleagues who do not code.

We don't live in the ideal world. Then what might the actual future of spatial transcriptomics look like given current trends? Visium might soon become to spatial transcriptomics what Chromium is to scRNA-seq, while LCM and GeoMX DSP live on by the side for ROI based studies. Perhaps largely with Visium, spatial transcriptomics might soon become as mainstream as scRNA-seq is today. However, just like the cDNA microarray, which was the transcriptomics method of choice in the 2000s and early 2010s and was replaced by RNA-seq which is more quantitative and sensitive, Visium might be replaced by some other technique in a few years after more technological advances that address Visium's drawbacks such as lack of single-cell resolution and low detection efficiency, though we don't know what

that new technique would be. At present, 10X has plans for what might be based on smFISH or ISS and have single molecule resolution and Visium HD which has single-cell resolution. Then we anticipate 10X to hold a substantial market share of spatial transcriptomics in the near future.

However, if 10X fails to ride the new trends, or if another company develops something much better, then it might replace 10X as the dominant company in spatial transcriptomics. Then what might replace Visium? If the commercial highly multiplexed smFISH platforms take off and become adopted by core facilities so the individual lab no longer has to invest in new equipment and the pricey probe collection, then the possibility that they may compete with Visium can't be ruled out. Moreover, as Illumina sequencing also involves image processing and matching fluorescent spots from different rounds, image processing for smFISH might no longer be a bottleneck in the near future. Commercial probe sets for highly multiplexed smFISH much as the probe sets on commercial cDNA microarrays might emerge for use with the automated platforms and core facilities. Back in the golden age of the cDNA microarray, probes of known sequences on the array were used to profile the transcriptome. Also, at present, most scRNA-seq and spatial transcriptomics studies only care about known genes and existing genome annotations, so not being able to find novel isoforms might not be a significant drawback to most users. In contrast, lack of single-cell resolution in Visium is indeed a serious drawback, because cell type deconvolution of the spots is commonly performed and many computational tools have been developed for this purpose. As we don't know how this rapidly developing field will unfold in the next few years, these are just possibilities and we cannot make specific predictions.

In addition, realistically speaking, where are we on the way to pursue the holy grail of low cost, convenient, high spatial resolution, high detection efficiency, larger area of tissue, transcriptome wide profiling, 3D tissue, and multi-omics? As already discussed in Section 7.7, trade-offs can't be avoided at present. Considering the more recent novel techniques, such as CISI, MOSAICA, sci-Space, BOLORAMIS, PIXEL-seq, and etc., we don't find the new techniques in the entire field of spatial transcriptomics going in a single direction in what to prefer in the trade-offs.

Some areas do not seem to pursue some of the objectives of the holy grail. For instance, we do not see smFISH-based techniques applied to an increasing number of genes over time (Figure 7.23), while there may be more interest in profiling larger number of cells (Figure 7.24) and novel proofs of principle (e.g. in CISI, MOSAICA,

and SABER). Instead, highly multiplexed smFISH datasets with a smaller number genes are complementary to scRNA-seq data from the same studies (e.g. [2, 3, 4, 5].

However, there are developments that reduce the competition between some areas of the trade-offs without eliminating the trade-offs. So far there seems to be less interest in *in situ* sequencing due to its inefficiency. ISS and HybISS were developed by the same group and are both in Cartana, but recent atlases that could have used either favored HybISS, which has somewhat higher detection efficiency than ISS, and with RCA amplification and a relatively low detection efficiency, can be applied to larger areas of tissue and imaged at lower magnification. For *in situ* sequencing, there also seems to be a trend to avoid the inefficiency of reverse transcription, as in HybRISS and BOLORAMIS.

New NGS barcoding techniques seem to have more emphasis on high resolution (e.g. single-cell but not high spatial resolution in XYZeq), if not high spatial resolution, but different studies seem to have different emphases on the other objectives in the holy grail. For instance, while all aiming for higher spatial resolution, the Slide-seq and Stereo-seq papers emphasize scalability to more tissue (indeed these techniques have lower detection efficiency), while the PIXEL-seq paper emphasizes not compromising detection efficiency, and the Seq-Scope paper emphasizes "easy-to-implement". Slide-seq2 then emphasizes better detection efficiency than the first version of Slide-seq, though the improved efficiency is still low. No NGS based spatial technique has attempted to rival smFISH detection efficiency. Again, the different emphases highlight the trade-offs, which will most likely stay with us for a long time. If that is the case, then spatial transcriptomics might evolve into different branches, with different types of techniques each with its own trade-offs better suited to different types of studies.

References

1. Mirzabekov AD. DNA sequencing by hybridization — a megasequencing method and a diagnostic tool? Trends in Biotechnology 1994 Jan; 12:27–32. DOI: 10.1016/0167-7799(94)90008-6. Available from: https://doi.org/10.1016/0167-7799(94)90008-6

2. La Manno G, Siletti K, Furlan A, Gyllborg D, Vinsland E, Mossi Albiach A, Mattsson Langseth C, Khven I, Lederer AR, Dratva LM, Johnsson A, Nilsson M, Lönnerberg P, and Linnarsson S. Molecular architecture of the developing

mouse brain. Nature 2021; 596:92–6. DOI: 10.1038/s41586-021-03775-x. Available from: https://doi.org/10.1038/s41586-021-03775-x

3. Bhaduri A, Sandoval-Espinosa C, Otero-Garcia M, Oh I, Yin R, Eze UC, Nowakowski TJ, and Kriegstein AR. An Atlas of Cortical Arealization Identifies Dynamic Molecular Signatures. bioRxiv 2021 Jan :2021.05.17.444528. DOI: 10.1101/2021.05.17.444528. Available from: http://biorxiv.org/content/early/2021/05/18/2021.05.17.444528.abstract

4. Lu Y, Liu M, Yang J, Weissman SM, Pan X, Katz SG, and Wang S. Spatial transcriptome profiling by MERFISH reveals fetal liver hematopoietic stem cell niche architecture. Cell Discovery 2021; 7:47. DOI: 10.1038/s41421-021-00266-1. Available from: https://doi.org/10.1038/s41421-021-00266-1

5. Bruggen D van, Pohl F, Langseth CM, Kukanja P, Lee H, Kabbe M, Meijer M, Hilscher MM, Nilsson M, Sundström E, and Castelo-Branco G. Developmental landscape of human forebrain at a single-cell level unveils early waves of oligodendrogenesis. bioRxiv 2021 Jan :2021.07.22.453317. DOI: 10.1101/2021.07.22.453317. Available from: http://biorxiv.org/content/early/2021/07/22/2021.07.22.453317.abstract

Part II

Methods and tools for spatial transcriptomics

Chapter 11

FROM SINGLE-CELL TO SPATIAL TRANSCRIPTOMICS

I have lived through the very history of spatial transcriptomics that I have documented and analyzed in the previous part. This chapter summarizes my journey through this history, from improving an early method to map dissociated cells from scRNA-seq to a prequel spatial atlas to contributing to scalable software tools with a consistent user interface to data from multiple technologies, to bring us one step closer to building an "ideal" package outlined in Chapter 10.

11.1 Spatial reconstruction of *Drosophila* embryo

As already mentioned, 2014-2017 is a period transitioning from the prequel era to the rise of the current era, when the then new current era methods mapping dissociated cells from scRNA-seq data to a prequel (WM)ISH reference 9.3. One of these early methods is DistMap [1], which maps dissociated cells *Drosphila* to the BDTNP spatial reference [2], and which I tried to improve with our lab's flagship RNA-seq read pseudoalignment method `kallisto` [3]. BDTNP was created with FISH, and has single-cell resolution.

In the most popular scRNA-seq technology 10X Chromium, as well as Visium, the transcripts are captured by their polyA tails, reverse transcribed into cDNAs, which are then polymerase chain reaction (PCR) amplified for sequencing. During the course of PCR, the primer matching the polyT (from the original polyA) introduces a cell or spatial barcode and a unique molecular identifier (UMI), so each read from a copy of the cDNA can be traced back to the cell or Visium spot and mRNA molecule it comes from. Read alignment methods such as Cell Ranger produce gene count matrices, where unique molecular identifiers (UMIs) are assigned to each gene in each cell. However, assigning UMIs to a unique gene is not trivial, as some UMIs can map to multiple genes. Furthermore, while short read scRNA-seq capturing transcripts with 3' end polyA sequence loses much of the isoform information, such information is not entirely lost. In kallisto, the reads are pseudoaligned to the transcriptome, so the UMIs can be assigned to equivalence classes (ECs), which are sets of known transcripts the UMI is compatible to. From the ECs, it's possible to infer which isoforms of a gene a transcript may have come from. The matrix with ECs instead of genes as rows and cells or samples as columns is known as the

transcript compatibility counts (TCC) matrix.

In [1], the gene count matrix from a Drop-seq *Drosophila* dataset was used, disregarding isoform information. In 2018, when I was rotating in the lab, I reimplemented DistMap in order to understand it and get introduced to spatial transcriptomics and single-cell data analysis, while using my implementation to predict expression of ECs rather than genes in space as an improvement. `kallisto` was originally designed for bulk RNA-seq. In order to run it for scRNA-seq, I used sircel to error correct the Drop-seq cell barcodes and indirectly run `kallisto` to obtain the TCC matrices. Then I used Seurat v2 to log normalize data and used multi canonical correlation analysis (CCA) and dynamic warping [4] to correct for batch effect in the Drop-seq data. The Drop-seq and BDTNP data have two spacies, *Drosophila melanogaster* and *Drosophila virilis*, but I focused on the former. Cells from the two species can be clearly separated and doublets were removed. The Drop-seq and BDTNP data somewhat overlap in developmental stage, in stage 6 and late stage 5 soon after cell membranes form around the nuclei in the early syncytical stage and soon before gastrulation. The datasets also overlap by 84 genes, which was nearly all the genes in BDTNP. To verify if my reimplementation can recapitulate spatial patterns of genes in BDTNP, I initially collapsed the ECs into genes when all transcripts in the EC map to the same gene. Since none of the ECs had transcripts that all mapped to one of the genes, 83 genes were used for the mapping.

In DistMap, first the log normalized data for each gene was manually thresholded individually so the binarized pattern matches that found in the WMISH images on BDGP. Then the Drop-seq data was also binarized, with a threshold chosen to maximize root mean squared error (RMSE) between the correlation matrix of the binarized BDTNP data and that of the binarized Drop-seq data. Then the Matthews correlation coefficient (MCC) for binary data was used to map Drop-seq cells to the BDTNP reference. Expression values of the 83 genes were permuted in each cell 100 times and the permuted data was thresholded again to compute a null distribution of the MCC scores. This way I obtained a pseuto-p-value at each location; nearly all of them had $p < 0.05$ after Benjamini-Hochberg correction.

Next the MCC scores were exponentiated so they are all positive. Let M denote this exponentiated MCC matrix with cells in columns and spatial locations in rows. Let G denote the TCC matrix, and B the binarized TCC matrix where all non-zero entries are set to 1 to normalize for the number of cells expressing a gene. Then compute matrices $D_1 = MG^T$ and $D_2 = MB^T$, and the normalized score matrix

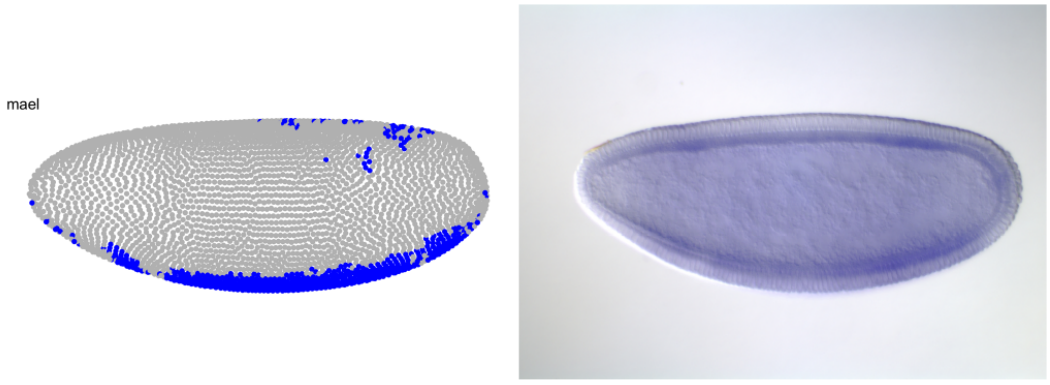

Figure 11.1: Predicted (left) expression pattern of *mael* and observed pattern in BDGP (right).

$Q = D_1/D_2$ (element-wise division). D_1 is a matrix with locations in rows and genes in columns; its entries take higher values when the MCC is high between cells and locations for the 83 landmark genes, or when many cells express a gene of interest, hence the D_2. The final score is $S = Q/(1 + Q)$ (element-wise division), which is then also binarized, with default threshold of the 75% quantile. The binarized patterns, the columns of this matrix, are the predictions of DistMap. There are many binarization steps because of the less quantitative nature of FISH (not smFISH) data. For gene *mael* which is not among the 83 landmark genes, the prediction matched the WMISH from BDGP (Fig. 11.1).

More interestingly, when predicting for ECs outside the 83 landmark genes, sometimes different ECs that map to different isoforms of the same gene have different predicted patterns, such as Inx2 (Fig. 11.2). I designed probes specific to some of the isoforms and a member of the Angela Stathopoulos lab performed WMISH with the probes to verify the predictions, although the experiment failed.

However, DistMap is not completely satisfactory for the following reasons. First, from plotting the score prior to binarization, I found that some predicted patterns are lost in binarization, although such patterns cannot be discerned in BDGP images, which may be due to the less quantitative nature of colorimetric WMISH. Second, you may feel somewhat disoriented by my discription of the DistMap procedure above, because it is an *ad hoc* score, which may not apply well to more quantitative spatial references from the current era. Hence I attempted to create a novel method to quantitatively predict expression of genes in a non-spatial scRNA-seq dataset but not in a spatial reference that profiles a limited number of genes. I first denoised and

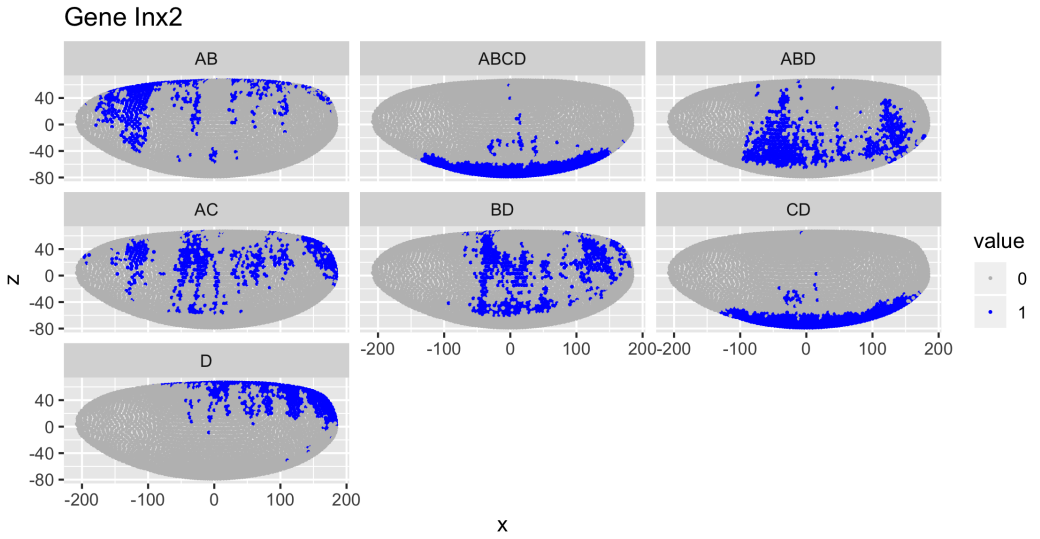

Figure 11.2: Predicted expression patterns of different ECs of Inx2. The letters correspond to isoforms of this gene and each combination of the letters is a set of isoforms the EC is compatible to.

normalized the Drop-seq data with DCA [5], as the FISH data is also normalized. Then I fitted a lasso regularized linear model with the normalized gene expression for the 83 landmark genes as the predictor and the normalized expression for any gene of interest as the response. Then I used this model on the BDTNP data to predict expression of the new gene in space. In some cases, such as Alh, the prediction seems to give the correct qualitative patterns, but in some cases, such as pyd3, the pattern is wrong (Fig. 11.3).

A further problem is that even if the pattern prediction is correct, the numbers in the predicted values may not make sense, because FISH and Drop-seq are different data types, and FISH is not very quantitative. I also reasoned that the reason why the prediction failed for some genes is that the 83 landmark gene might not be coregulated with all the other genes, and linear regression is too simplistic a model. Moreover, spatial information was not used in the model.

Around this time, the preprint of Seurat v3 [6] came out, with a different *ad hoc* score in the shared CCA or PCA space, which gave decent qualitative pattern prediction. The novosparc preprint appeared around the same time, using optimal

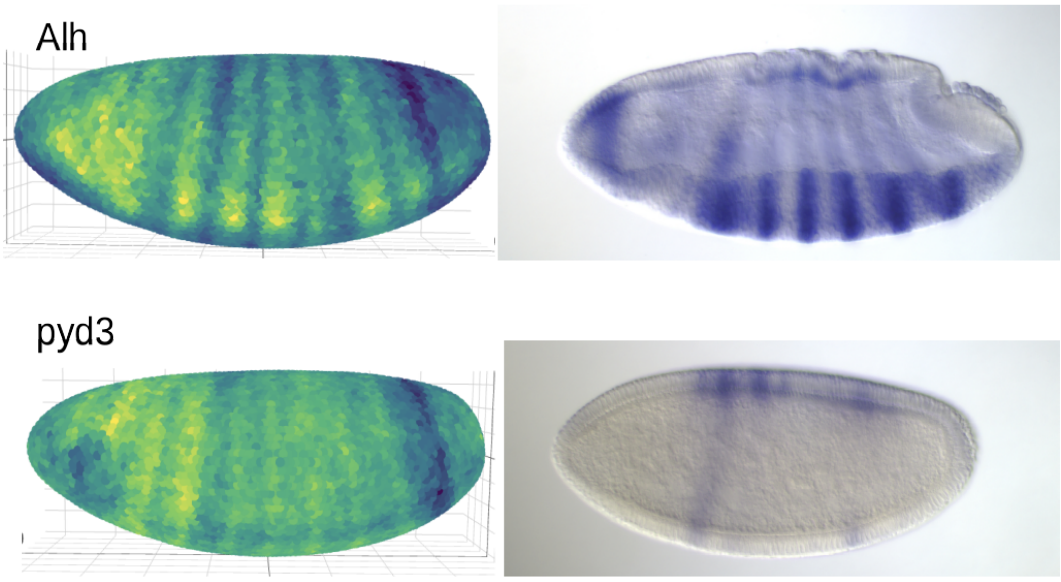

Figure 11.3: Linear regression prediction of gene expression in space (left) and BDGP WMISH images (right) for genes Alh and pyd3. In the predictions, yellow is higher value and dark blue is lower value.

transport to exploit positive spatial autocorrelation to predict gene expression and reconstruct spatial coordiantes even without a spatial reference [7]. In DistMap, I found that mapping cells to locations (MCC) and predicting gene expression (the final binarized) are different but related; it is a problem of data integration across different modalities, because once we correct for the effects of the different modalities, gene expression in the spatial dataset can be imputed from similar cells in scRNA-seq. This is the perspective taken by packages such as Seurat v3 and gimVI, although it does not use spatial information. While I moved on to the `kallisobustools` project (Section 11.2), new methods that either explicitly perform spatial reconstruction or do so under the auspice of data integration have been written, and interest in this area remains strong (Fig. 9.4), as transcriptome-wide scRNA-seq data, smFISH-based spatial data with single-cell resolution with a few hundred genes, and transcriptome-wide spatial data without single-cell resolution are complementary in some studies [8]. Some of these methods are reviewed in Section 9.3.

11.2 `kallisto` goes single-cell

While this project is on non-spatial scRNA-seq, it built skills and mindsets for my later Voyager project (Chapter 13) in the emphasis on computational efficiency, uniform user interface for data from multiple technologies, and comprehensive and reproducible documentation.

The `BUSpaRse` package

When I ran `kallisto` via `sircel` to pseudoalign the sequencing reads from Drop-seq in the previous section, it was not designed to scale to the increasing size of scRNA-seq data. It took hours to run `sircel` on each of the batches which had about 1000 cells. Our lab devised the Barcode, UMI, Set (BUS) format, where the "set" stands for ECs, to represent pseudoalignment for scRNA-seq data from different technologies [9]. `kallisto` writes a compressed binary output for the scRNA-seq pseudoalignment, and the first version of the `bustools` package sorts the binary and converts it into a text file. To convert the text file into a gene count matrix, a Python script was written to iterate through the file line by line and aggregate UMIs into ECs and genes. To aggregate ECs into genes, all transcripts must match the same gene; other transcripts were discarded. This is very efficient and only took a few minutes to run on a laptop for a typical scRNA-seq dataset with thousands of cells.

Because as already mentioned, this field is split between R and Python, in order to cater to R users, I wrote an R implementation of this process converting the BUS text file to the gene count or TCC matrix. Implementing the iteration in C++ and calling it from R via Rcpp, this took less than a minute to run. This later became the BUSpaRse package, now on Bioconductor [10]. Because `bustools` also requires a file mapping transcripts to genes, BUSpaRse implements functions to query genome annotations from Ensembl transcriptome FASTA file sequence names, TxDB, and EnsemblDB; the latter two are genome annotation resources on Bioconductor. As `kallisto` and `bustools` were later adapted to generate the spliced and unspliced gene count matrices for RNA velocity much more efficiently [11] than the original implementation [12], I also implemented functions to get the transcriptome FASTA file with flanked introns by extracting the genomic ranges from the genome sequence and the corresponding transcript to gene file in BUSpaRse.

Later, the Python prototype was translated into efficient C++ code in a newer version of `bustools`, so this functionality of BUSpaRse was superseded. However, perhaps

because the BUS format can be used for purposes outside scRNA-seq or because of the genome annotation querying functions, or because some people used BUSpaRse for the gene or TCC count matrix anyway, BUSpaRse has been downloaded by around 200 distinct IP addresses per month for the past 2 years.

Comparisons with Cell Ranger

In the Museum of Spatial Transcriptomics, 10X Visium is by far the most popular current era spatial transcriptomis technology. The 10X company started with non-spatial scRNA-seq, and its Chromium product is very popular. The official read alignment software for Chromium is Cell Ranger. To demonstrate that `kallisto bustools` can obtain the gene count matrix much faster and using less memory than Cell Ranger while getting similar results. We downloaded over 20 scRNA-seq datasets from different technologies and with a wide range of number of cells, some from the 10X website, to benchmark `kallisto bustools` , Cell Ranger, STARsolo, and Alevin, and showed that `kallisto bustools` is the fastest and most memory efficient for these datasets [11].

However, if the results are inaccurate, then the speed is no good. Then we compared the `kallisto bustools` results to Cell Ranger results and showed that the gene counts and dimension reductions are similar (Fig. 11.4) and that the downstream biological analysis results were similar as well. I performed the latter, focusing on a mouse 10k neurons Chromium dataset from the 10X website, but one analyses was run was all datasets to show the similarity in the biological results: I concatenated the gene count matrices from `kallisto bustools` and Cell Ranger for this dataset, and performed differential expression (DE) between the two matrices with the Wilcoxon rank sum test as is typical in scRNA-seq. Then I performed gene set enrichment analysis (GSEA) with gene ontology (GO) terms as the gene sets on the DE genes, to see what kind of genes tend differ in the two methods.

This is shown in Figure 11.4g: a quantile–quantile plot comparing the observed distribution of p values of GSEA, after Bonferroni correction for multiple testing across ontologies and datasets, with the expected distribution of a uniform distribution between 0 and 1. If the observed distribution does not deviate from the expected distribution, then the points should lie close to the diagonal line, y = x. The gray ribbon around the line is the 95% confidence interval. Here most Gene Ontology (GO) terms have adjusted p = 1, meaning that most GO terms are very depleted of genes differentially expressed between the `kallisto bustools` and

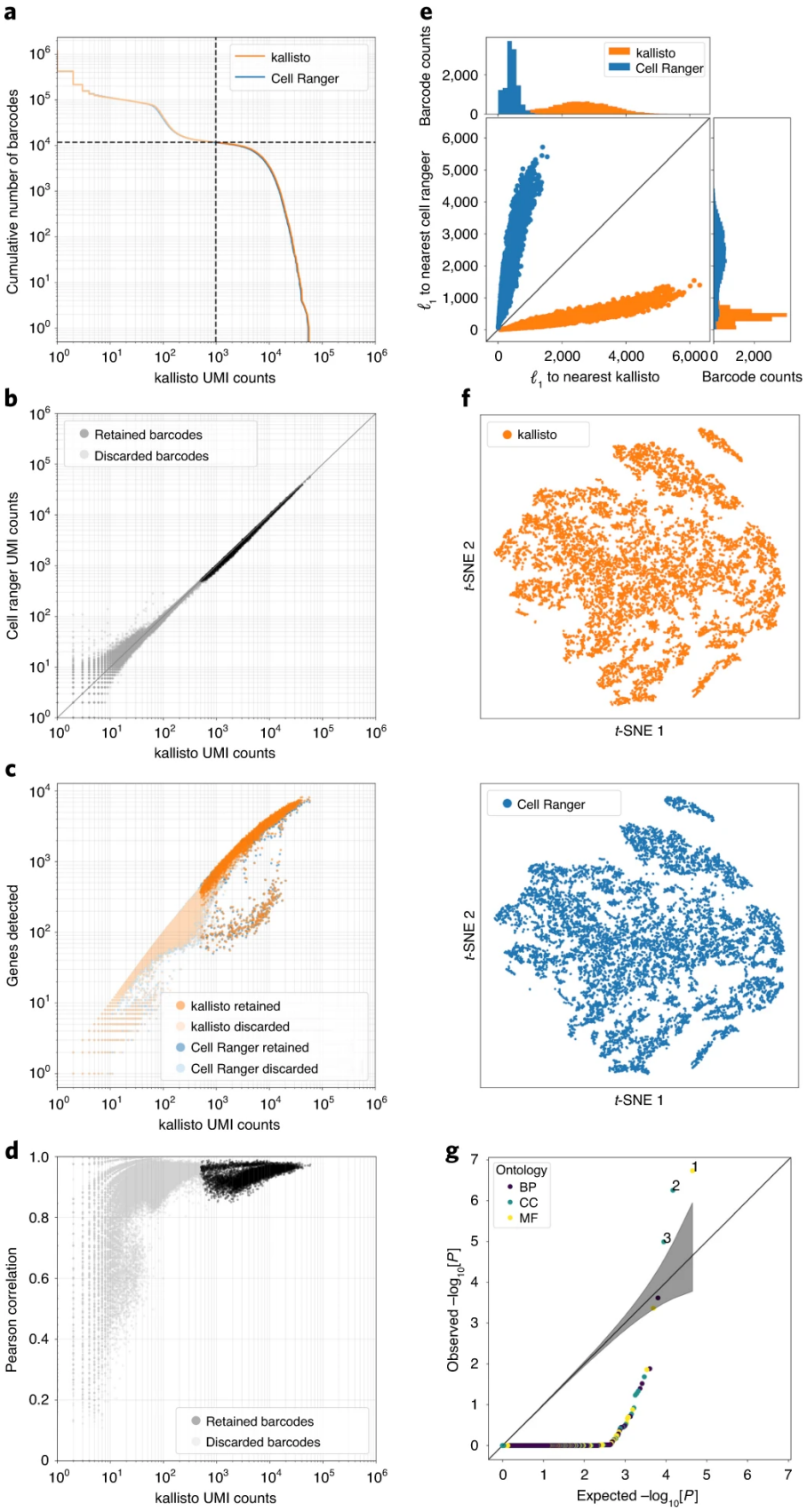

Figure 11.4: Reproduced from Figure 2 of [11].

Cell Ranger matrices. GO terms above y = x are labeled. Generally, GO terms significantly enriched among differentially expressed genes are related to ribosomal proteins; specifically, the GO terms 1, 2, and 3 correspond to structural constituent of ribosome, cytosolic large ribosomal subunit, and cytosolic small ribosomal subunit. The points are colored by ontology: biological processes (BP), cellular components (CC), and molecular functions (MF). Across datasets, the DE genes in enriched GO terms tend to be depleted in the `kallisto bustools` results compared to Cell Ranger, perhaps because `kallisto bustools` discards UMIs that don't map to a unique gene and these genes tend to have many paralogs, increasing the chance of confusion between genes. In many other datasets, no GO term is significantly enriched in the DE genes. Hence unless ribosomal genes are of interest, `kallisto bustools` should lead to biological results equivalent to those from Cell Ranger. Results from other datasets are shown in Supplementary Figure 3 of [11].

For the 10k neuron dataset specifically, I compared the `kallisto` and Cell Ranger matrices for dimension reduction, Leiden clustering, cluster marker genes, and slingshot pseudotime [13, 14], and showed that these biological analyses yielded similar results between the two matrices (Supplementary Figures 6-7). For each gene present in both matrices, I ran Pearson and Spearman correlation and found the correlation to be well above 0.95 for most genes. I also performed pseudotime analysis with slingshot , and found similar results (Supplementary Figure 7.4). Finally, I compared the gene count matrices for a mixed human and mouse dataset, and found that `kallisto bustools` can distinguish between cells from either species just as well as Cell Ranger and that UMI assignment to either species is similar (Supplementary Figure 8).

`kallisto bustools` tutorials

Next we wrote tutorials demonstrating usage of `kallisto bustools` on many datasets and demonstrate downstream analysis tasks. In the beginning, we wrote some tutorials and hosted one hour long in person and virtual workshops running and explaining the tutorials live, all the way from downloading the FASTQ files to performing downstream analyses, in both R and Python. This would not be possible with Cell Ranger, which takes hours and many CPU cores to run for the typical scRNA-seq dataset. I wrote the R tutorials, some of which can be found on this website [15]. However, since these tutorials were built locally, they were not checked to be fully reproducible.

Later we built a more comprehensive `kallisto bustools` documentation website [16]. There's a basic tutorial running `kallisto bustools` and then perform basic data filtering and normalization on a small dataset, the first thing to get started. Some further tutorials are about advanced upstream steps in running `kallisto bustools` , such as building the index for the transcritome or for RNA velocity, processing multiple FASTQ files, and processing multi-species data. Some further tutorials demonstrate downstream tasks such as pseudotime and RNA velocity analysis. Former student Joseph Min wrote a Python wrapper for `kallisto bustools` that greatly simplifies the upstream processing especially for human and mouse data. The tutorials are run on Google Colab all the way from downloading the FASTQ files to performing downstream analysis; since the free plan of Google Colab has limited computational resources and packages need to be installed from scratch, this show that the `kallisto bustools` workflow is scalable and that the tutorials are reproducible. Most tutorials are in both R and Python. The R notebooks were adapted from the first website I mentioned. The tutorials are organized by analysis tasks.

Writing the `kallisto bustools` tutorials informed our design of the Voyager documentation website (Section 13.2), where the tutorials are very comprehensive and are organized by data collection technology and spastial analysis method. The Voyager tutorials are built from scratch on the cloud with limited computational resources to ensure reproducibility and scalability, and can be run as Colab notebooks.

11.3 Cosmodrome: unified image processing pipeline for smFISH-based spatial transcriptomics

Spatial point process analyses

smFISH-based spatial transcriptomics data is a treasure trove of information. While most people base their analyses on the gene count matrix and cell centroid coordinates, because each mRNA molecule is visualized as a puncta, where the molecules are located in the cell can be interesting, as already discussed in Section 9.10. The Voyager project (Chapter 13) began with spatial point process analysis of subcellular transcript localization in a MERFISH mouse primary motor cortex dataset (MOp) [17], which provides transcript spot coordinates in 3D and cell segmentation polygons in one of the 7 z-planes. Spatial point processes model the locations where certain events occur, such as where a species is observed in ecology and where galaxies are observed in astronomy. It is the locations themselves that are studied. We can ask whether the observations tend to cluster or if they repel each other and

thus becoming more like a regular grid. The null hypothesis is "complete spatial randomness" (CSR), in which the locations of observations are independent from each other, and is formally a homogeneous Poisson point process.

In spatial point process analyses, the observational window is crucial, and cannot be imputed as the convex hull of the observations. For example, if the observations are only in a corner of the actual observational window, then they must be very clustered. However, if the observational window is taken to be the convex hull, then the clustering may not be observed. The cell segmentation can be used as the observational window. However, since not all smFISH spatial transcriptomics publications provide the cell segmentation masks or polygons, and those that do only provide in one out of several z-planes, datasets suitable for such an analysis are rare despite the proliferating publications.

Ripley's L is a summary function to find if a point process is clustered or regular and at which lengthscales. In a nutshell, it's a variance normalized form of the K function, which is the average number of points within a distance from each other, as a function of distance. With CSR, the theoretical K function is $K = \pi r^2$, and the theoretical L function is $L = r$. If the observed L function is above the theoretical line, then it means that there are more points within a distance than expected, indicating clusteirng. If the observed L function is below the theoretical line, then it means that fewer points are within a distance than expected, indicating regularity.

I have attempted to compute Ripley's L for transcript spots of select genes in each cell. Without a 3D cell segmentation, I pretended that the segmentation is the same in all z-planes, as too few transcript are present in each z-plane. Transcripts of many genes are not uniformly distributed inside the cell, and some tend to the cell periphery, as if outside the nucleus, although this cannot be verified without nuclei segmentation (Fig. 11.5). Overall, for one gene, across thousands of cells, the observed L functions tend to be above the theoretical line under CSR (Fig. 11.5), indicating clustering. Some cells have stronger clustering for this gene than others, and it would be interesting to see if they are biologically different as well.

Unfortunately, this did not go far because transcript spots assigned to one cell based on segmentation from one z-plane means that many of the transcripts in fact belong to a neighboring cell. Furthermore, existing spatial point process packages, such as `spdep`, may not scale to the many gigabytes of transcript spot data across hundreds of thousands of cells. The z-plane problem motivated me to perform image process-

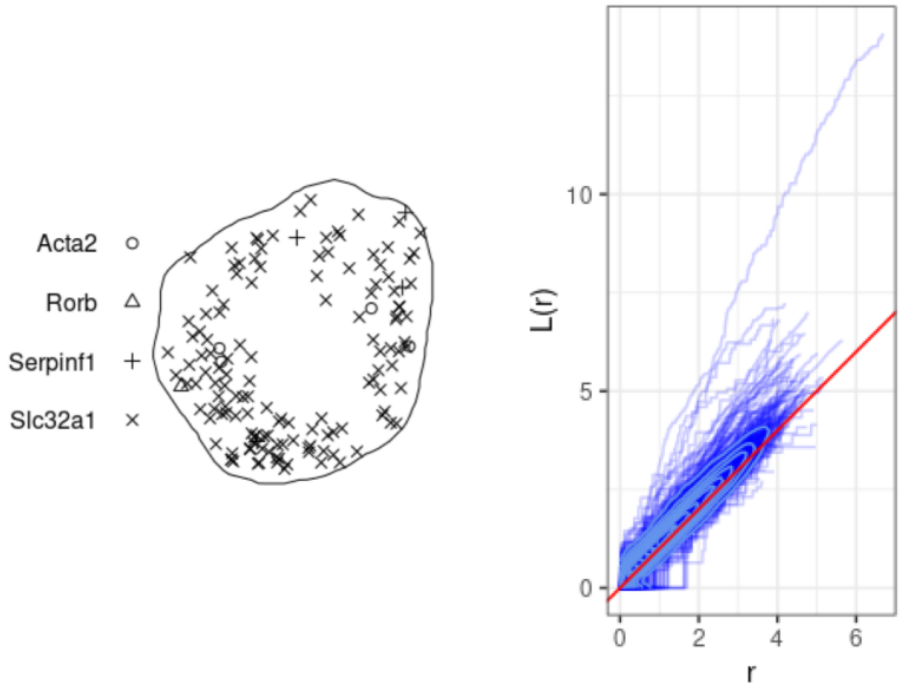

Figure 11.5: Left: Locations of transcript spots of 4 genes in a cell. Right: L functions of one gene in thousands of cells in the dataset. The red line is the theoretical L function under CSR, $L = r$.

ing from scratch to obtain 3D cell segmentations. However, as discussed in Section 9.1, image processing software tends to be very specific to one smFISH or ISS based technology. Inspired by `kallisto` `bustools` and the Tidymodels machine learning framework that brings many machine learning methods under a common user interface, I decided to write Cosmodrome, an image processing pipeline for highly multiplexed smFISH or ISS that can apply across different technologies. The name "Cosmodrome", because "Voyager" is meant to explore space, and "Cosmodrome" is the beginning of space exploration.

PanoTx format, extending `starfish`

An existing attempt to unify highly multiplexed smFISH image processing is `starfish` [18], which requires the images to be organized in the SpaceTx format. The images are organized by assay, field of view (FOV), channel, round of hybridization, and z-planes; each assay, FOV, channel, round, and z-plane has its own TIFF file, which may or may not be the most efficient when processing the images. Reformatting data into SpaceTx is not trivial; images in this MERFISH

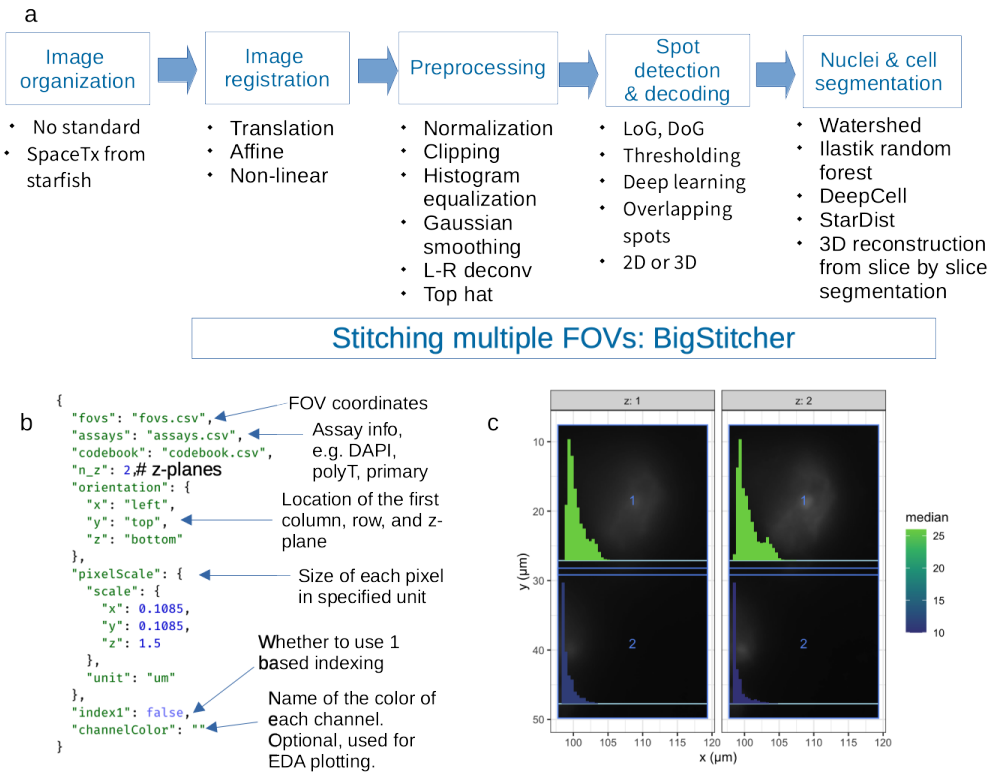

Figure 11.6: a, Schematic of highly multiplexed smFISH image processing. b, Example top level JSON file. c, Example QC plot showing multiple FOVs, showing pixel intensity histograms on top of the images, colored by median pixel intensity.

dataset have one TIFF stack for each FOV over all rounds of hybridization, while images in the seqFISH study [19] are in multi-file OME-TIFF files with one stack per FOV per round for the z-planes. Another drawback is that as of writing, `starfish` has very limited support for multiple FOVs, nor was there a way to simultaneously plot multiple FOVs in quality control (QC), whereas with the growing number of cells of smFISH-based data (Figure 7.24), numerous FOVs is the norm.

Therefore I tried to extend the SpaceTx format into PanoTx, "Pano" for panorama, because working with multiple FOVs should be front and center. Highly multiplexed smFISH image processing consists of the following steps (Fig. 11.6c):

1. Image organization, such as the SpaceTx format.

2. Image registration between channels and rounds, which canbe affine or non-linear.

3. Preprocessing, which includes clipping, Gaussian smoothing, and top hat filtering to make the image more amenable for spot calling.

4. Spot detection, which can use Laplacian of Gaussian, as described in Section 9.1, then the spots need to be matched across rounds of hybridization to decode the gene barcode.

5. Cell and nuclei segmentation, so the transcript spots can be assigned to cells.

For multiple FOVs, underlying all these steps is image stitching to obtain global coordinates in the entire tissue section, nor only in the FOV. The transformations of each FOVs can be computed without loading all images into memory, as is done in [20], and with the transformations, the global coordinates can be computed. Moreover, since the JSON format is less human readable than tables, such as in CSV files, while the top level JSON file for PanoTx is similar to that of SpaceTx, the FOV coordinates, assays, and codebook are specified as CSV instead of JSON in PanoTx (Figure 11.6b). I have written an R package that interacts with the PanoTx files and plot QC metrics for multiple FOVs, such as plotting a thumbnail of the image with their histograms arranged in space with their FOV coordinates; different z-planes appear in different facets of the plot ((Figure 11.6c), so the users can check if the FOVs are in the correct place.

There are different ways to do each step in Fig. 11.6, such as Laplacian of Gaussian and deep learning methods to detect transcript spots, and multiple deep learning packages for cell and nuclei segmentation, as discussed in Section 9.1. There are also many different image stitching methods, including ImageJ stitching [21], BigStitch [22], MIST [23], and ASHLAR [24]. Because any one method may not perform best for all datasets, I wished to provide a uniform user interface to multiple methods at each step. To do so, inspired by the proteomics image processing pipeline MCMICRO [25], which uses Nextflow to coordinate steps of image processing inside Docker containers, I built a Docker container for BigStitcher to use in a similar Nextflow workflow. While MCMICRO does not decode multiple rounds of hybridization, I would need to write a somewhat different workflow to accommodate this.

Yet these image stitching tools themselves require input file formats different from SpaceTx. While MIST and ASHLAR can only process one z-plane, BigStitcher can perform image registration through time in spatiotemporal images and can process z-stacks, making it more suitable for the kind of MERFISH image I tried to re-process.

However, it's not clear how BigStitcher can be hacked to register across assays and rounds of hybridization while stitching the FOVs at the same time. Moreover, BigStitcher resaves the images into its own type of HDF5 format, which can be time consuming. While it may be nice to have a consistent user interface to different image stitching methods, as one method might not perform best for all datasets, the disparate input file format requirements of these methods and their differences with SpaceTx make it very difficult to implement the consistent user interface, because file format conversion can be time and disk space consuming for large datasets.

All of the image stitching tools mentioned above are ImageJ plugins except for ASHLAR. From using these tools, I learnt about the OME-TIFF format, an existing TIFF format with standardized metadata to represent fluorescent microscopy images that organizes the frames of a TIFF stack by channel and z-planes. FOV coordinates can be stored in the metadata as well, and the metadata can govern multiple OME-TIFF files from different FOVs. I also learnt about the well-established and performant tool BioFormats that reads and converts between microscopy imaging file formats. While there's no standard in OME-TIFF on rounds of hybridization, the existing features of OME-TIFF and BioFormats are relevant. I was trying to reinvent the wheel.

My original interest, as expressed in the spatial point process analysis, is downstream spatial analysis rather than image processing, so after finding the file format woe, I set spatial point process analysis of transcript spots aside for now and moved on to other types of spatial data analysis, which led to the last chapters of this thesis. While my downstream spatial data analysis packages are maturing, a potential solution to the imaging file format woe is emerging, in the OME-Zarr format [26], continuing the OME standards but with the chunked and cloud optimized Zarr format already used for large imaging data. One of the OME-Zarr example datasets comes from *in situ* genome sequencing, with multiple rounds of *in situ* sequencing imaging [27]. Even better, BigStitcher will gain support for OME-Zarr[26]. Nor is Zarr the only format for efficient operations on large image data. The geospatial field has long been working with large raster data from remote sensing; borrowing from this field, the Samui browser uses cloud-optimized GeoTIFF and OpenLayers to visualize large spatial transcriptomics data [28].

The file format woe is not unique to image processing. In downstream analyses, some packages implement their own data structures instead of reusing existing, standard ones, forcing users to convert between data structures and learn new syntaxes.

For example, in the R ecosystem, something similar to the `SpatialExperiment` class [29] has been reinvented multiple times, such as in Giotto [30] which did not reuse the existing `SingleCellExperiment` (SCE) or `Seurat` classes, and in `semla`, the successor to `STUtility` [31]. However, I find these existing classes inadequate for fully utilizing opportunities from the spatial information. So I wrote the `SpatialFeatureExperiment` (SFE), which extends `SpatialExperiment` (SPE) with the Simple Features represenation of geometries (Chapter 12). By extending SCE and conforming to its styles and conventions, SFE does not create additional file format woes.

References

1. Karaiskos N, Wahle P, Alles J, Boltengagen A, Ayoub S, Kipar C, Kocks C, Rajewsky N, and Zinzen RP. The Drosophila embryo at single-cell transcriptome resolution. Science 2017 Oct; 358:194–9. DOI: 10.1126/science.aan3235. Available from: https://science.sciencemag.org/content/358/6360/194

2. Fowlkes CC, Hendriks CLL, Keränen SV, Weber GH, Rübel O, Huang MY, Chatoor S, DePace AH, Simirenko L, Henriquez C, Beaton A, Weiszmann R, Celniker S, Hamann B, Knowles DW, Biggin MD, Eisen MB, and Malik J. A Quantitative Spatiotemporal Atlas of Gene Expression in the Drosophila Blastoderm. Cell 2008 Apr; 133:364–74. DOI: 10.1016/j.cell.2008.01.053. Available from: https://linkinghub.elsevier.com/retrieve/pii/S009286740800281X

3. Bray NL, Pimentel H, Melsted P, and Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 2016 34:5 2016 Apr; 34:525–7. DOI: 10.1038/nbt.3519. Available from: https://www.nature.com/articles/nbt.3519

4. Butler A, Hoffman P, Smibert P, Papalexi E, and Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology 2018; 36:411–20. DOI: 10.1038/nbt.4096. Available from: https://doi.org/10.1038/nbt.4096

5. Eraslan G, Simon LM, Mircea M, Mueller NS, and Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nature Communications 2019; 10:390. DOI: 10.1038/s41467-018-07931-2. Available from: https://doi.org/10.1038/s41467-018-07931-2

6. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, and Satija R. Comprehensive Integration of Single-Cell Data. Cell 2019 Jun; 177:1888–902. DOI: 10.1016/j.cell.2019.05.031. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867419305598

7. Nitzan M, Karaiskos N, Friedman N, and Rajewsky N. Gene expression cartography. Nature 2019 Dec; 576:132–7. DOI: 10.1038/s41586-019-1773-3. Available from: https://doi.org/10.1038/s41586-019-1773-3

8. Guilliams M, Bonnardel J, Haest B, Vanderborght B, Wagner C, Remmerie A, Bujko A, Martens L, Thoné T, Browaeys R, De Ponti FF, Vanneste B, Zwicker C, Svedberg FR, Vanhalewyn T, Gonçalves A, Lippens S, Devriendt B, Cox E, Ferrero G, Wittamer V, Willaert A, Kaptein SJ, Neyts J, Dallmeier K, Geldhof P, Casaert S, Deplancke B, Dijke P ten, Hoorens A, Vanlander A, Berrevoet F, Van Nieuwenhove Y, Saeys Y, Saelens W, Van Vlierberghe H, Devisscher L, and Scott CL. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. Cell 2022 Jan; 185:379–96. DOI: 10.1016/J.CELL.2021.12.018

9. Melsted P, Ntranos V, and Pachter L. The barcode, UMI, set format and BUStools. Bioinformatics 2019 Nov; 35:4472–3. DOI: 10.1093/BIOINFORMATICS/BTZ279. Available from: https://academic.oup.com/bioinformatics/article/35/21/4472/5487510

10. Moses L and Pachter L. BUSpaRse: kallisto | bustools R utilities. 2023. DOI: 10.18129/B9.bioc.BUSpaRse. Available from: https://bioconductor.org/packages/BUSpaRse

11. Melsted P, Booeshaghi AS, Liu L, Gao F, Lu L, Min KH(, Veiga Beltrame E da, Hjörleifsson KE, Gehring J, and Pachter L. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nature Biotechnology 2021. DOI: 10.1038/s41587-021-00870-2. Available from: https://doi.org/10.1038/s41587-021-00870-2

12. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, Fan J, Borm LE, Liu Z, Bruggen D van, Guo J, He X, Barker R, Sundström E, Castelo-Branco G, Cramer P, Adameyko I, Linnarsson S, and Kharchenko PV. RNA velocity of single cells. Nature 2018; 560:494–8. DOI: 10.1038/s41586-018-0414-6. Available from: https://doi.org/10.1038/s41586-018-0414-6

13. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, and Dudoit S. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics 2018; 19:477. DOI: 10.1186/s12864-018-4772-0. Available from: https://doi.org/10.1186/s12864-018-4772-0

14. Saelens W, Cannoodt R, Todorov H, and Saeys Y. A comparison of single-cell trajectory inference methods. Nature Biotechnology 2019; 37:547–54. DOI: 10.1038/s41587-019-0071-9. Available from: https://doi.org/10.1038/s41587-019-0071-9

15. Lambda Moses. kallisto | bustools R notebooks. Available from: https://bustools.github.io/BUS_notebooks_R/

16. Pachter L. kallisto | bustools. Available from: https://www.kallistobus.tools/

17. Zhang M, Eichhorn SW, Zingg B, Yao Z, Zeng H, Dong H, and Zhuang X. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics. bioRxiv 2020 Jan :2020.06.04.105700. DOI: 10.1101/2020.06.04.105700. Available from: http://biorxiv.org/content/early/2020/06/05/2020.06.04.105700.abstract

18. Perkel JM. Starfish enterprise: finding RNA patterns in single cells. Nature 2019 Aug; 572:549–51. DOI: 10.1038/D41586-019-02477-9

19. Lohoff T, Ghazanfar S, Missarova A, Koulena N, Pierson N, Griffiths JA, Bardot ES, Eng CH, Tyser RC, Argelaguet R, Guibentif C, Srinivas S, Briscoe J, Simons BD, Hadjantonakis AK, Göttgens B, Reik W, Nichols J, Cai L, and Marioni JC. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. Nature Biotechnology 2021 40:1 2021 Sep; 40:74–85. DOI: 10.1038/s41587-021-01006-2. Available from: https://www.nature.com/articles/s41587-021-01006-2

20. Qian X, Harris KD, Hauling T, Nicoloutsopoulos D, Muñoz-Manchado AB, Skene N, Hjerling-Leffler J, and Nilsson M. Probabilistic cell typing enables fine mapping of closely related cell types in situ. Nature Methods 2020 Jan; 17:101–6. DOI: 10.1038/s41592-019-0631-4. Available from: https://doi.org/10.1038/s41592-019-0631-4

21. Preibisch S, Saalfeld S, and Tomancak P. Globally optimal stitching of tiled 3D microscopic image acquisitions. Bioinformatics 2009 Jun; 25:1463–5. DOI: 10.1093/BIOINFORMATICS/BTP184. Available from: https://academic.oup.com/bioinformatics/article/25/11/1463/332497

22. Hörl D, Rojas Rusak F, Preusser F, Tillberg P, Randel N, Chhetri RK, Cardona A, Keller PJ, Harz H, Leonhardt H, Treier M, and Preibisch S. BigStitcher: reconstructing high-resolution image datasets of cleared and expanded samples. Nature Methods 2019; 16:870–4. DOI: 10.1038/s41592-019-0501-0. Available from: https://doi.org/10.1038/s41592-019-0501-0

23. Chalfoun J, Majurski M, Blattner T, Bhadriraju K, Keyrouz W, Bajcsy P, and Brady M. MIST: Accurate and Scalable Microscopy Image Stitching Tool with Stage Modeling and Error Minimization. Scientific Reports 2017; 7:4988. DOI: 10.1038/s41598-017-04567-y. Available from: https://doi.org/10.1038/s41598-017-04567-y

24. Muhlich JL, Chen YA, Yapp C, Russell D, Santagata S, and Sorger PK. Stitching and registering highly multiplexed whole-slide images of tissues and tumors using ASHLAR. Bioinformatics 2022 Sep; 38:4613–21. DOI: 10.

1093/BIOINFORMATICS/BTAC544. Available from: https://academic.oup.com/bioinformatics/article/38/19/4613/6668278

25. Schapiro D, Sokolov A, Yapp C, Chen YA, Muhlich JL, Hess J, Creason AL, Nirmal AJ, Baker GJ, Nariya MK, Lin JR, Maliga Z, Jacobson CA, Hodgman MW, Ruokonen J, Farhi SL, Abbondanza D, McKinley ET, Persson D, Betts C, Sivagnanam S, Regev A, Goecks J, Coffey RJ, Coussens LM, Santagata S, and Sorger PK. MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging. Nature Methods 2021 19:3 2021 Nov; 19:311–5. DOI: 10.1038/s41592-021-01308-y. Available from: https://www.nature.com/articles/s41592-021-01308-y

26. Moore J, Basurto-Lozada D, Besson S, Bogovic J, Bragantini J, Brown EM, Burel JM, Moreno XC, Medeiros Gd, Diel EE, Gault D, Ghosh SS, Gold I, Halchenko YO, Hartley M, Horsfall D, Keller MS, Kittisopikul M, Kovacs G, Yoldaş AK, Kyoda K, Villegeorges AlTdl, Li T, Liberali P, Lindner D, Linkert M, Lüthi J, Maitin-Shepard J, Manz T, Marconato L, McCormick M, Lange M, Mohamed K, Moore W, Norlin N, Ouyang W, Özdemir B, Palla G, Pape C, Pelkmans L, Pietzsch T, Preibisch S, Prete M, Rzepka N, Samee S, Schaub N, Sidky H, Solak AC, Stirling DR, Striebel J, Tischer C, Toloudis D, Virshup I, Walczysko P, Watson AM, Weisbart E, Wong F, Yamauchi KA, Bayraktar O, Cimini BA, Gehlenborg N, Haniffa M, Hotaling N, Onami S, Royer LA, Saalfeld S, Stegle O, Theis FJ, and Swedlow JR. OME-Zarr: a cloud-optimized bioimaging file format with international community support. bioRxiv 2023 May :2023.02.17.528834. DOI: 10.1101/2023.02.17.528834. Available from: https://www.biorxiv.org/content/10.1101/2023.02.17.528834v4%20https://www.biorxiv.org/content/10.1101/2023.02.17.528834v4.abstract

27. Payne AC, Chiang ZD, Reginato PL, Mangiameli SM, Murray EM, Yao CC, Markoulaki S, Earl AS, Labade AS, Jaenisch R, Church GM, Boyden ES, Buenrostro JD, and Chen F. In situ genome sequencing resolves DNA sequence and structure in intact biological samples. Science 2021 Feb; 371. DOI: 10.1126/SCIENCE.AAY3446/SUPPL{_}FILE/AAY3446-PAYNE-SM.PDF. Available from: https://www.science.org/doi/10.1126/science.aay3446

28. Sriworarat C, Nguyen A, Eagles NJ, Collado-Torres L, Martinowich K, Maynard KR, and Hicks SC. Performant web-based interactive visualization tool for spatially-resolved transcriptomics experiments. bioRxiv 2023 Feb :2023.01.28.525943. DOI: 10.1101/2023.01.28.525943. Available from: https://www.biorxiv.org/content/10.1101/2023.01.28.525943v2%20https://www.biorxiv.org/content/10.1101/2023.01.28.525943v2.abstract

29. Righelli D, Weber LM, Crowell HL, Pardo B, Collado-Torres L, Ghazanfar S, Lun AT, Hicks SC, and Risso D. SpatialExperiment: infrastructure for

spatially-resolved transcriptomics data in R using Bioconductor. Bioinformatics 2022 May; 38:3128–31. DOI: 10.1093/BIOINFORMATICS/BTAC299. Available from: https://academic.oup.com/bioinformatics/article/38/11/3128/6575443

30. Dries R, Zhu Q, Dong R, Eng Chee-Huat Linus, Li H, Liu K, Fu Y, Zhao Tianxiao, Sarkar A, Bao F, George RE, Pierson N, Cai L, and Yuan GC. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. en. Genome Biol. 2021 Mar; 22:78

31. Bergenstråhle J, Larsson L, and Lundeberg J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. BMC Genomics 2020 Dec; 21:482. DOI: 10.1186/s12864-020-06832-3. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-06832-3

Chapter 12

SPATIALFEATUREEXPERIMENT: BRINGING SIMPLE FEATURES TO SPATIAL TRANSCRIPTOMICS

The `SpatialFeatureExperiment` (SFE) data structure is the foundation of exploratory spatial data analyses (ESDA) in the R package Voyager (Chapter 13). `SingleCellExperiment` (SCE) is a data structure designed to represent scRNA-seq data in R [1]. The raw and normalized gene count matrices are in the assays slot, with genes in rows and cells in columns (Fig. 12.1). Any matrix-like class can be used in the assays, including dense matrices, sparse matrices (e.g. `dgCMatrix` from the Matrix R package), data frames, and `DelayedArray` which allows for on-disk operations. Cell metadata is in the `colData` field, and row metadata is in the `rowData` field. In addition, analogous to `colData`, cell embedding matrices from dimension reductions are stored in the `reducedDims` field. PCA loadings are stored as attributes of the PCA cell embedding matrix. SCE implements getters and setters for these fields; getters and setters for the new fields in SFE for geometries conform to the conventions and styles of SCE.

`SpatialExperiment` (SPE) is an existing class that extends SCE for spatial -omics data, adding the spatialCoords field for spatial coordinates of cell or spot centroids, and imgData to organize images associated with the spatial dataset [2]. The images can be on disk or remote and are thus not loaded into memory unless necessary. When the image is read into memory, it is a matrix of color hex codes. In addition, there's a special column in colData, `sample_id`, to distinguish between coordinates from different tissue sections, as different sections can have overlapping numeric values of coordinates. The SPE package also implements functions to mirror and rotate the images. Accompanying SPE is the `ggspavis` package to visualize gene expression and cell attributes in space.

The geospatial tradition is central to how SFE extends SPE. The `sf` package is the R interface to Simple Features, a standard way to represent vector geometries in the geospatial field. In `sf`, geometries and their attributes can be stored in a data frame with a special column for the geometries. Furthermore, `sf` supports optimized geometric IO and operations, with the GDAL and GEOS C++ libraries, respectively. With geometric operations, the SFE object can be subsetted or cropped

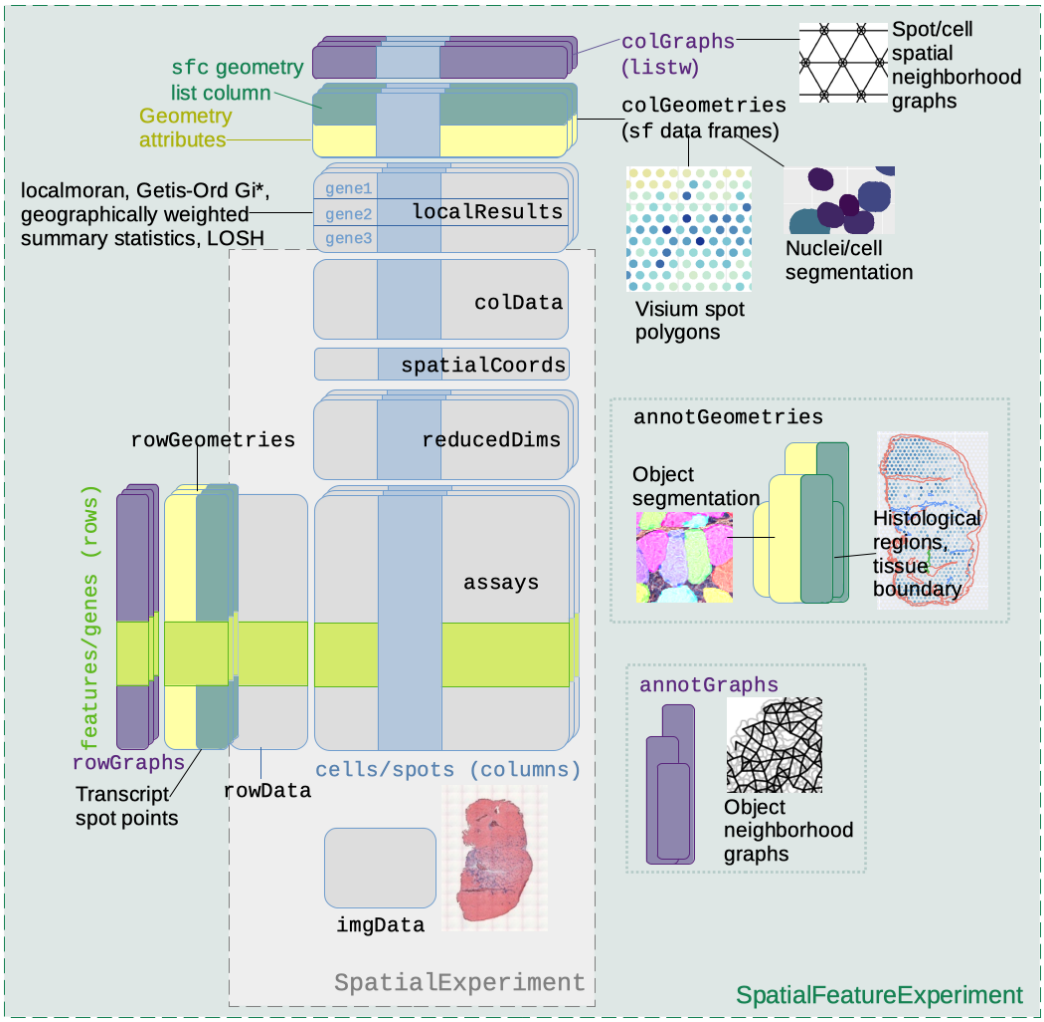

Figure 12.1: Schematic showing the structure of the SpatialFeatureExperiment object and how it extends SpatialExperiment. Details are written in the main text.

with a geometry.

SFE adds `colGeometries`, analogous to colData, essentially a collection of `sf` data frames, for geometries associated with columns of the gene count matrix such as cells and Visium spots (Fig. 12.1). A cell can be associated with multiple geometries, such as cell and nucleus segmentation polygons. This way, with the efficient geometric operations, it's possible to find problematic and likely low quality cells and nuclei with unusual sizes or multiple pieces in quality control, and to find characteristics of cells that may be associated to gene expression, such as morphological metrics and proportion of cell area occupied by the nucleus. Furthermore, in data visualization, the cell segmentation polygons can be plotted instead of centroids, to visualize cell morphology and avoid overlaps between points in areas with high

cell density. The `rowGeometries` field is for geometries associated with genes or features. It is implemented although not currently used; it can potentially be used for transcript spots from smFISH-based datasets, including those not assigned to cells. However, the transcript spot data is often very large and would benefit from on disk representations of geometries.

The `annotGeometries` field is for geometries associated with the dataset but do not directly correspond to columns of the gene count matrix. These can be cell segmentation polygons in a Visium dataset, the tissue boundary, or pathologist annotated histological regions from other software such as QuPath and ImageJ. With geometric operations, we can find the number of cells or nuclei in each Visium spot, the histological region cells or spots belong to, and the proportion of each Visium spot that is in the tissue or histological region. These can then be related to gene expression in the EDA process. Voyager can also compute univariate spatial statistics on numeric columns of colData, colGeometries, annotGeometries, and dimension reduction cell embeddings.

The neighborhood view of spatial analyses requires a spatial neighborhood graph, whereas SPE does not have a field to organize the graphs. In SFE, the `colGraphs` field stores the spatial neighborhood graph for entities associated with columns of the gene count matrix. The SFE package wraps all methods to find spatial neighborhood graphs in `spdep`, one of the core spatial analysis packages in R, including polygon contiguity, triangulation followed by edge pruning, k nearest neighbors, and distance based neighbors, as well as different types of edge weights, such as binary, row normalization, row and column normalization, and distance based edge weights. In addition, SFE has a much faster implementation of finding distance based edge weights after finding the k nearest neighbor or distance based graph. The `annotGraphs` field is for spatial neighborhood graphs of annotation geometries, so spatial analyses can be performed on attributes of these geometries such as cell area. The `rowGraphs` field is for graphs associated with genes or features, although it is not currently used. The `sample_id` is important here because the spatial neighborhood graph only makes sense within one tissue section. The graphs are represented as `listw` objects as implemented in `spdep`, so no conversion is required when using the numerous `spdep` methods in Voyager.

Results can be organized within the SFE object to link results to features from which they were computed and to facilitate visualization. Local spatial statistics returns one set of results for each cell or spot. To organize the results, the `localResults` field

was introduced, analogous to `reducedDims`. Like `reducedDims`, `localResults` are organized by the spatial analysis method, but unlike `reducedDims`, the results for each method are organized by genes or features. Univariate and bivariate local results are stored in the `localResults` field, while multivariate results are stored in `reducedDims` or optionally colData if they are vectors or data frames. Univariate global results are stored in rowData for gene expression, and in the metadata or attributes of the corresponding fields for colData, geometries, and dimension reductions. These metadata and attributes can be accessed with getter functions such as `colFeatureData()`. However, bivariate global results are not currently stored in the SFE object due to the great variability in output format.

Space Ranger is the official alignment and preprocessing software for Visium. While the SPE package can read Space Ranger output, SFE does so somewhat differently. Because the Space Ranger output includes spot diameter in pixels in the full resolution image, SFE constructs Visium spot polygons from the centroids and the diameter. In addition, the pixels can be converted to microns, based on spacing between spots which is known to be 100 microns. Furthermore, unlike SPE, SFE uses the `terra` package to manage images. The `terra` package is designed for raster geospatial data which often can't fit into memory. The images are read as SpatRaster objects, which are pointers to the images on disk, so the images are not loaded into memory unless necessary. When the image is plotted, it's not entirely loaded into memory if a lower resolution suffices. When an SPE object is converted into SFE, the images are converted to SpatRaster. SFE can also directly read Vizgen MERFISH output, which has fluorescent images that may not fit into memory, benefiting from `terra`. Like SCE and SPE before it, SFE will continue to evolve to meet the demands of ESDA.

References

1. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, Marini F, Rue-Albrecht K, Risso D, Soneson C, Waldron L, Pagès H, Smith ML, Huber W, Morgan M, Gottardo R, and Hicks SC. Orchestrating single-cell analysis with Bioconductor. en. Nat. Methods 2020 Feb; 17:137–45

2. Righelli D, Weber LM, Crowell HL, Pardo B, Collado-Torres L, Ghazanfar S, Lun AT, Hicks SC, and Risso D. SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor. Bioinformatics 2022 May; 38:3128–31. DOI: 10.1093/BIOINFORMATICS/BTAC299. Available from: https://academic.oup.com/bioinformatics/article/38/11/3128/6575443

Chapter 13

VOYAGER: EXPLORATORY SPATIAL DATA ANALYSIS FROM GEOSPATIAL TO SPATIAL -OMICS

13.1 Introduction

From the developing embryo to the hepatic lobule, spatial organization of cells is essential to the functions of many tissues. Recent years have seen a drastic rise in interest in technology development, data collection, and development of data analysis tools in spatial transcriptomics [1]. Overarching data analysis frameworks for data organization and exploratory data analysis (EDA) have been developed for spatial -omics data, such as Seurat [2], squidpy [3], Giotto [4], and semla (formerly STUtility [5]), inheriting from the tradition of single cell RNA-seq (scRNA-seq), but also adding some spatial visualization, basic spatial data analysis functionalities, and implementation or wrappers for further downstream spatial analyses such as finding spatially variable genes and cell type interactions. In addition, for the purpose of EDA, many visualization tools have been developed for spatial -omics data. Many of these visualization tools are designed to be scalable and interactive, for large imaging based data such as MERFISH and imaging mass cytometry, plotting gene expression and cell metadata in space [6, 7, 8, 9]. Some of the interactive tools can also perform cell clustering, plot 2D scatterplots of gene expressions and cell attributes, and analyze cell type colocalization [6]. Some utilize virtual reality for an immersive data visualization experience [10]. Some are focused on a particular tissue [11], or implement a very specific type of visualization [12, 13].

Existing EDA tools for spatial -omics don't fully take advantage of the opportunities presented by the spatial information. EDA is an approach to the data without many preconceived ideas, theories, or hypotheses [14], and is a mindset where the analyst asks questions, tries to answer the questions by visualization, transforming, and modeling the data, which can lead to more refined and further questions, without a formal process or a strict set of rules [15]. The importance of EDA is that "it is important to understand what you CAN DO before you learn to measure how WELL you have seem to DONE it" [16]. Exploratory spatial data analysis (ESDA) is EDA explicitly focusing on spatial aspects of the data, especially spatial autocorrelation, where nearby observations are not independent from each other [14]. Before the

rise of spatial -omics, ESDA has long been used in geography, where a rich tradition has been developed. The widely used spatial autocorrelation metrics Moran's I [17] and Geary's C [18] are among the global univariate spatial statistics, which give one set of results for the entire dataset, whose characteristics have been further elaborated upon over the years [19, 20, 21]. In addition, there are tools to explore the length scale of spatial autocorrelation, such as the correlogram [22] and variogram [23]. Local versions of spatial statistics, such as local Moran's I [24] and Getis-Ord Gi* [25] can be used to explore local spatial heterogeneity and find spatial clusters of high or low values, giving a set of results at each location. There are also spatially informed global and local bivariate and multivariate statistics that account for spatial autocorrelation and correlation between features simultaneously, such as Lee's L [26] and MULTISPATI PCA [27].

Much of this ESDA tradition that may benefit spatial -omics has not been utilized in Seurat, squidpy, and Giotto. For example, while Seurat and squidpy can perform global Moran's I, squidpy implements Ripley's L to analyze cell type clustering in space, Giotto independently implemented something similar to Lee's L for spatially informed gene co-expression, and semla implemented Moran's I and Getis-Ord Gi, they have not systematically utilized much of the ESDA tradition such as many methods to explore length scales and local spatial heterogeneity of spatial autocorrelation. Furthermore, packages that specialize in visualization typically don't go in depth if at all into ESDA. Many other spatial -omics data analysis packages focus on image processing or a specialized task such as cell type deconvolution of Visium spots, analyzing cell-cell interactions, integrating data from different modalities or tissue sections, and predicting gene expression from H&E image. Therefore, there is a gap in ESDA in the field of spatial -omics data analysis to be filled.

Spatial -omics data analysis is largely split between R and Python [1]. The most commonly used EDA framework in R is Seurat, and in Python it's scanpy (and squidpy for spatial). However, because of hidden defaults and implementation details most users may be unaware of, Seurat and scanpy can give different results that can lead to different biological conclusions, such as in log fold change in gene expression across clusters [28]. In addition, many packages are not well-documented and most are not on a standard repository such as CRAN, Bioconductor, PyPI, or conda, lacking quality checks and making installation more cumbersome [1]. Some packages have divergent syntaxes, requiring users to use a different new data structure instead of reusing existing ones, increasing the learning curve for users while forcing

developers to reinvent the wheel.

Here we present Voyager, an R package that systematically brings the geospatial ESDA tradition to spatial -omics, with a consistent user interface for different spatial data analysis methods. Univariate, bivariate, and multivariate global and local methods are included. These methods can be applied to the output of any spatial -omics technology as long as a gene count matrix and spatial coordinates of cells or spots are available. Voyager reuses and extends the existing SingleCellExperiment (SCE) [29] and SpatialExperiment (SPE) [30] data structures, with the new SpatialFeatureExperiment (SFE) data structure that brings Simple Features [31] to SPE to allow for geometric operations on the cells and annotation geometries and to organize results from spatial analyses. Inheriting from SCE which some other spatial methods such as BayesSpace [32] are based on, Voyager ESDA is meant to complement other types of spatial and non-spatial analyses. Voyager implements plotting functions for gene expression, cell attributes, annotation geometries, the histology image, and dimension reduction colored in histological space with colorblind friendly default palettes [33, 34, 35], as well as plotting functions for spatial analysis results. We have also implemented a Python version of Voyager with the core functionalities of the R version and made sure that results from core functionalities match across languages. Voyager has comprehensive documentation and tutorials for common spatial -omics technologies and spatial analysis methods. The R packages Voyager, SpatialFeatureExperiment, and `SFEData` which provides example datasets for the tutorials are available on Bioconductor. The Python implementation is available on PyPI.

13.2 Results

The Voyager framework

Voyager brings some of the traditions of geospatial ESDA to spatial -omics. In the R ecosystem, this is implemented in time honored packages, such as `spdep` [36] for the neighborhood view [14], of spatial autocorrelation analyses based on a spatial neighborhood graph, and `gstat` [37] for the distance view, in which spatial autocorrelation is modeled based on a continuous functions of distance between observations. Voyager wraps many of these methods from `spdep` and `gstat`. Besides the neighborhood or distance views, ESDA methods can be broadly categorized by the number of variables analyzed: univariate (e.g. Moran's I), bivariate (e.g. Lee's L), and multivariate (e.g. MULTISPATI PCA); each of these has a main function providing a uniform user interface to a variety of methods,

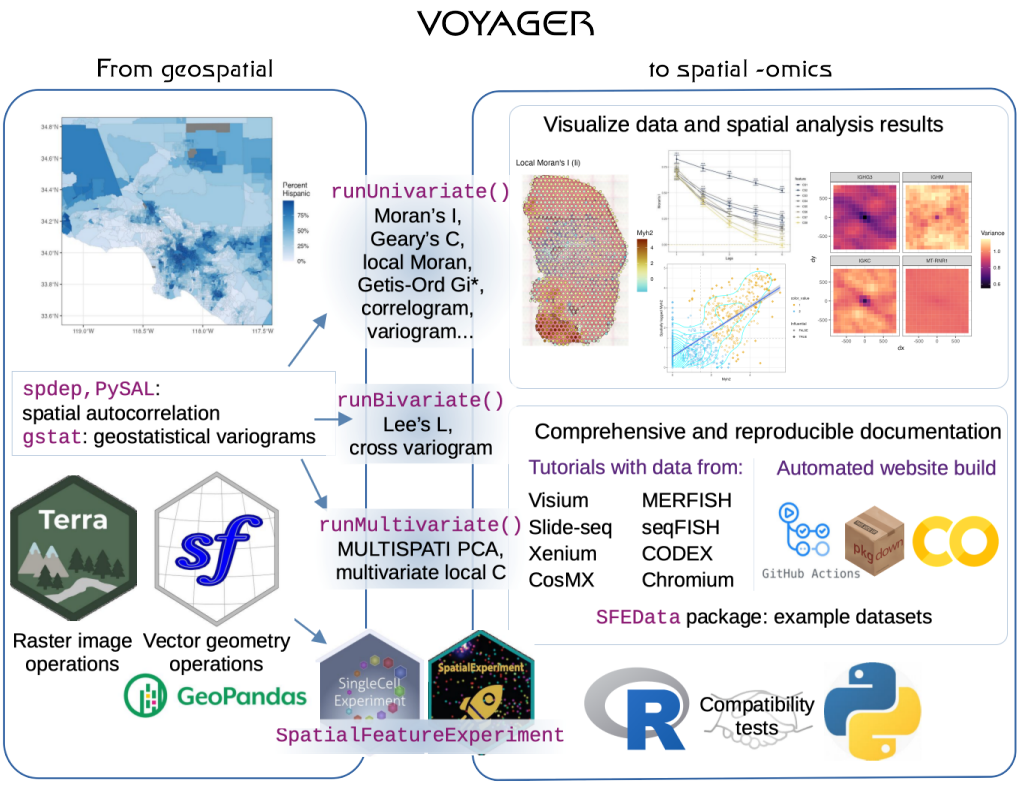

Figure 13.1: See Section 13.7 for caption.

minimizing user learning curve (Fig. 13.1), inspired by the Tidymodels machine learning framework [38]. These methods can also be categorized as global, where one set of results are returned for the entire dataset, or local, where each location or cell has its own set of results. The latter facilitate explorations of local heterogeneity in spatial relations. Users can extend Voyager and make the uniform user interface run custom ESDA methods to reduce redundant code and facilitate organization and visualization of results in Voyager.

Considering that the geospatial data for which `spdep` and `gstat` were developed tend to have a much smaller number of features and observations than modern single cell spatial -omics datasets, Voyager implements parallel processing when running a univariate or bivariate spatial method over a large number of genes, and reimplements methods whose original implementations don't scale to modern spatial -omics data to drastically speed up computation, including MULTISPATI PCA, Lee's L, finding bounds of Moran's I from spatial neighborhood graph, and distance based edge weighting of k nearest neighbors or distance based graphs.

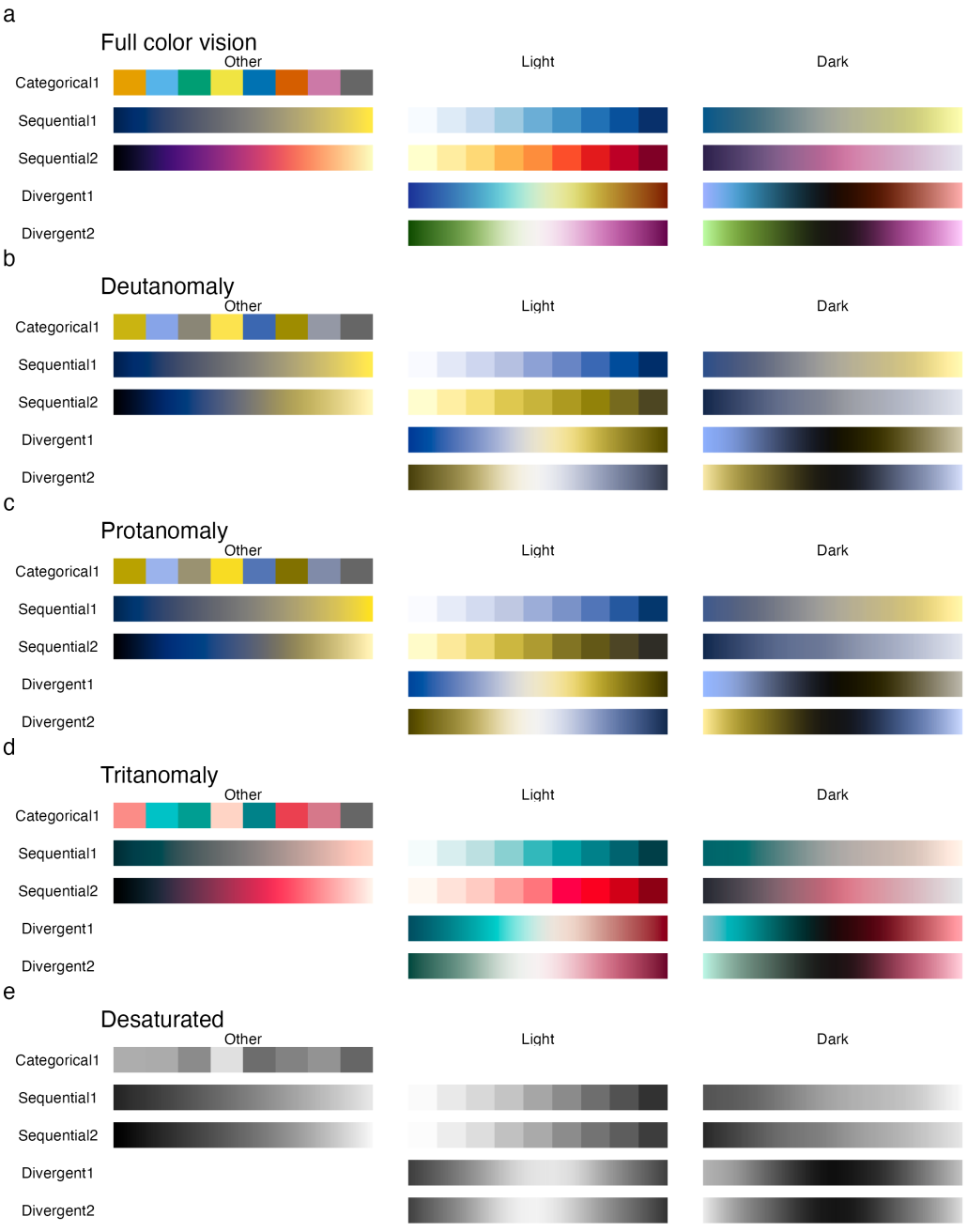

Figure 13.2: See Section 13.7 for caption.

Visualization is an essential part of the EDA process. As in Seurat, Voyager implements static plotting functions for gene expression, cell attributes, and cell projections along dimensions in dimension reduction plotted in histological space. The histology image can be optionally plotted behind the cells or Visium spots. Despite the proliferation of interactive visualization packages, static plots have their place in publications and reports. While most existing packages plot cells as points, Voyager can plot cell or nuclei segmentation polygons as well. For larger datasets, the users can specify a bounding box to zoom into a smaller area. The default palettes are designed to have color values discernible with color vision deficiencies (Fig. 13.2). Default continuous palettes come from ColorBrewer [35], scico [34], and viridis. A sequential palette is used by default, but a divergent palette is available when there is a meaningful center of divergence, such as 0 in local Moran's I and Lee's L; the palette is centered on the center of divergence of interest, so warm colors denote values above the center and cool colors denote values below (full color vision). The default categorical palette comes from dittoSeq [33], which is designed for colorblind friendly scRNA-seq data visualization.

In addition, Voyager implements plotting functions for spatial analysis results, such as the Moran scatter plot, correlogram, variogram, and local spatial statistics shown in histological space (Fig. 13.1). In contrast to the plotting functions in `spdep` and `gstat`, Voyager plotting functions are based on `ggplot2` [15] to be more visually appealing and customizable by users, and are designed to visualize results from multiple features and tissue sections at once.

Voyager is based on the data structure SFE, which inherits from SPE and SCE (Figs. 13.1,12.1). As a result, non-spatial scRNA-seq EDA methods and plotting functions implemented in `scater` [39], the package implementing scRNA-seq data preprocessing, quality control (QC), and EDA, can be applied as usual. We also try to make arguments of Voyager plotting functions consistent with their counterparts in `scater`, to maintain a more uniform user interface that is easier to learn.

Voyager has a comprehensive documentation website that features tutorials on applying EDA and ESDA to datasets from multiple spatial -omics technologies, including 10X Visium, Xenium [40], and Chromium, Nanostring CosMX [41], Vizgen MERFISH [42], Slide-seq [43], seqFISH [44], and CODEX [45] (Fig. 13.1). On the website, each technology has a landing page with an introduction of the technology and a table linking to vignettes using a dataset from the technology. For the most popular technology Visium [1], we have written introductory vignettes mostly per-

forming visualizations, as well as their advanced counterparts that delve more into spatial analyses. Each ESDA method has a similar landing page, with an introduction to the method and a similar table, linking to sections in each vignette using the ESDA method, some of which include further considerations on the ESDA methods and references to the geospatial ESDA literature. While Chromium is non-spatial, neighborhood view ESDA methods were applied to the k nearest graph in gene expression PCA space. Example datasets used in the vignettes are available as SFE objects in the `SFEData` package. In addition, there are vignettes instructing users on constructing an SFE object and extending Voyager for custom ESDA methods.

To make sure that the vignettes are reproducible, we build the documentation website on GitHub Actions, rendering all the vignettes on the cloud with a fresh machine. Because the GitHub Actions runner has less computational resources than a typical modern laptop, we can ensure that the vignettes are scalable to larger datasets, such as a MERFISH mouse liver dataset with almost 400,000 cells. The vignettes are also automatically converted into Google Colab notebooks to be run interactively and allow users to experiment with different parameters.

Because the field of single cell and spatial -omics data analysis is largely split between R and Python, we have developed a Python implementation of core functionalities and some basic vignettes of the Voyager R package, opening the way to deep learning and image analysis methods better supported in Python than R. At present, the Python implementation supports univariate spatial statistics, based on PySAL [46] and GeoPandas [47]. We have written "compatibility tests" to ensure that the R and Python implementations give the same results for core functionalities. Because Bioconductor requires software packages to have unit tests and pass an automated check run daily, whereas PyPI and conda do not perform automated tests, the Python implementation is indirectly held to the Bioconductor standard by the compatibility tests.

Compatibility tests

In scRNA-seq, Seurat and scanpy are both commonly used EDA frameworks. However, their default settings not only give different log fold changes but also significantly different PCA results as in a mouse olfactory bulb Visium dataset (Fig. 13.3a,c,d), which might lead to different biological conclusions. In contrast, in Voyager, there are no visible differences in Visium spot embeddings in the first two PCs (Fig. 13.3b). In Seurat vs. scanpy, the cosine difference (see Section 13.4)

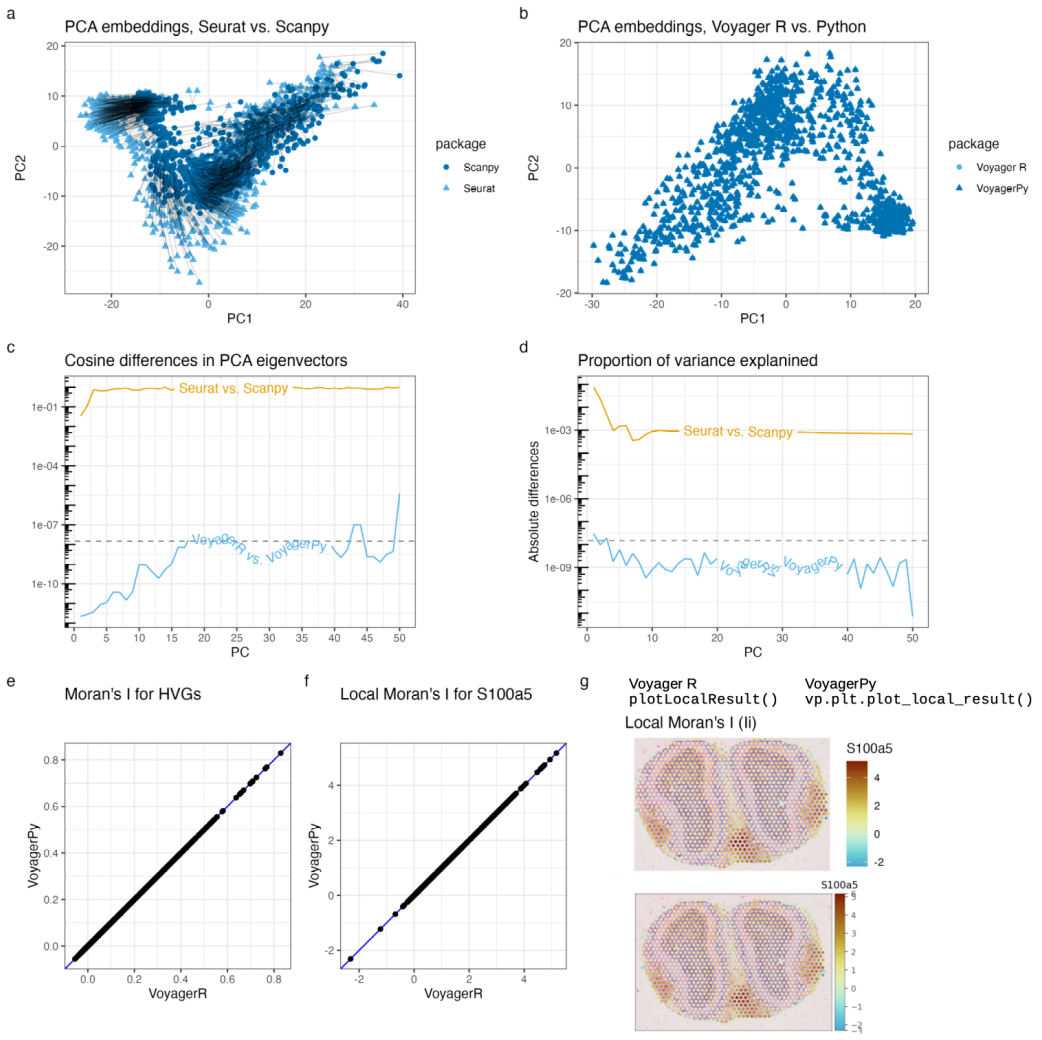

Figure 13.3: See Section 13.7 for caption.

between PCA eigenvectors (gene loadings) in each of the top 50 PCs is nearly 1 after the first few PCs, which means the angles between the corresponding eigenvectors are nearly 90 degrees (Fig 13.3c). In Voyager, this difference is much smaller, well below epsilon (dashed line, around 1.5e-8, see Section 13.4) until PC20 (Figure 2C). While the difference sometimes rises above epsilon after PC20, it does not exceed 1e-5. While Seurat and scanpy PCA with default parameters give sizable differences in proportion of variance explained by each PC, these differences in the R and Python implementations of Voyager are generally within epsilon (Fig. 13.3d).

For spatial statistics, the R and Python implementations of Voyager give consistent results for global Moran's I of the highly variable genes (Fig. 13.3e), with differences within epsilon. For local Moran's I (Fig. 13.3f-g), the results are the same

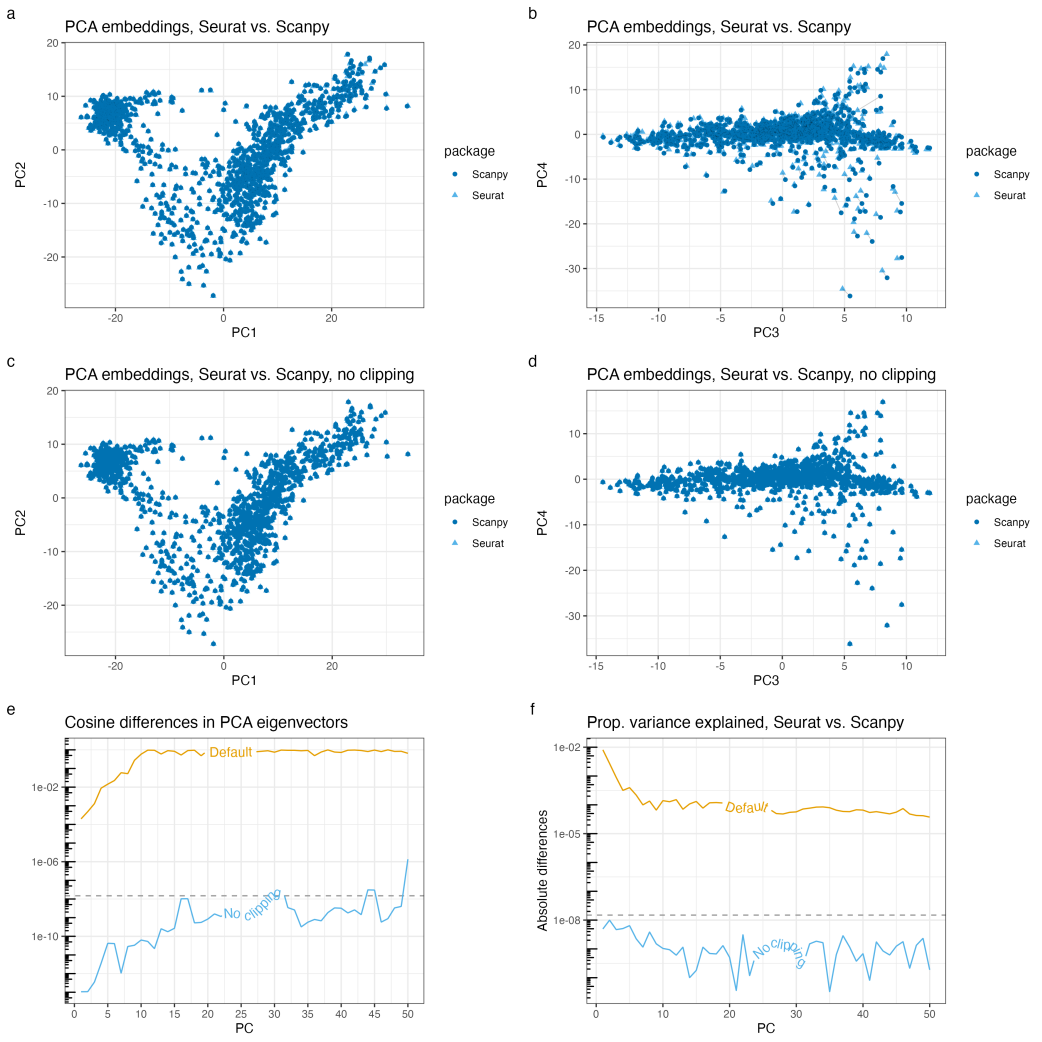

Figure 13.4: See Section 13.7 for caption.

(within epsilon) but with a non-default `spdep` parameter (see list below). In both implementations, besides the different defaults in `ggplot2` and `matplotlib`, the plotting functions give visually similar plots with the same palettes (Fig. 13.3g).

While the R and Python implementations of Voyager may eventually diverge in functionalities, as the two languages have better support for different types of analyses, we have written compatibility tests to make sure that the core functionalities and basic vignettes give the same results in R and Python, so language preference does not inadvertently lead to different biological conclusions.

That Seurat and scanpy give different PCA results is largely because of their different methods to find highly variable genes. The PCA results remain somewhat different

while using the same set of highly variable genes because of a hidden default in Seurat that clips scaled data to 10 while scanpy by default does not clip the scaled data (Fig. 13.4). Seurat and scanpy give different marker gene rankings from ostensibly the same differential expression method (e.g. t-test or Wilcoxon test) because they rank the genes differently and even give different log fold changes [28]; most users may be unaware of such differences. Therefore we strive to be transparent about the defaults and the reasons why we chose them.

At present, there are two vignettes subject to compatibility tests, one using a mouse olfactory bulb Visium dataset from 10X website, and the other applying univariate spatial statistics to the k nearest neighbor graph of a human peripheral blood mononuclear cells (PBMC) Chromium dataset. Below we list the defaults used in the basic vignettes covered by the compatibility tests; some of the defaults come from conventions and defaults in established packages, and are hence subject to further research and improvement:

1. Because SFE inherits from SCE, the R vignettes use `scater` and `scran` [48] from the SCE ecosystem for QC, data normalization, and non-spatial EDA. Data normalization in `scater` computes $\log_2\left(\frac{x}{N/\bar{N}}+1\right)$, where N denotes the total UMI count in one Visium spot, $\bar{N}$ is the average total UMI count in all spots in this dataset, and x is the UMI count of one gene in the Visium spot of interest. The pseudocount (default to 1), library size factors (default to $N/\bar{N}$), and transform (default to $\log_2$) can be changed. The size factor is centered on 1 to make it easier to translate log normalized counts back to raw counts. Log 2 is used because differences in values can be interpreted as log fold change. The Python implementation uses the same data normalization.

2. This is how `scran` finds highly variable genes (HVGs): with default parameters, a parametric non-linear curve $y=\frac{ax}{x^n+b}$ of variance vs. mean is fit for each gene of the log normalized data. Then the log ratio of the actual variance to the fitted variance from the curve is calculated, and a Lowess curve is fitted to this log ratio vs. mean scatter plot for each gene. The "technical" component of the variance is the fitted values from the Lowess curve. The "biological" component is the difference between the actual log ratio and the Lowess fitted log ratio. The top HVGs are genes with the largest biological component. The default parameters are used in R vignettes not covered by compatibility tests. For the basic vignettes covered by compatibility tests, because we did not find a

Python implementation of Lowess, the Python version reimplements `scran`'s HVG method when using parameter `lowess = FALSE` in `modelGeneVar()`, i.e. omitting the Lowess step, so the fitted, "technical" values come from the parametric curve and the "biological" component is the difference between the actual variance and the curve. This method assumes that most genes are not biologically interesting to the study of interest.

3. The number of top HVGs is 2000, from Seurat convention, and Seurat is the most popular scRNA-seq EDA package. This number is also not unreasonable.

4. PCA is performed on log normalized data, with the HVGs as in Seurat convention. Data is scaled before performing PCA, i.e. each gene from the log normalized data is scaled and centered to have mean 0 and variance 1, so genes that are more abundant but not necessarily more biologically variable don't drown out genes that are less abundant but more biologically variable in the top principal components. This can happen because gene expression data is overdispersed, so the variance not only increases with mean but also exceeds the mean, so the variance is greater than one would expect from a Poisson distribution. The data is scaled also because this is the Seurat convention.

5. When scaling the data, the variance is computed and the data for each gene is divided by the variance. One can either divide by n or (n-1); the former is the default in Numpy, as the maximum likelihood estimate of variance although it's biased, while the latter is the default in R and scanpy, as an unbiased estimate. We divide by (n-1) in both implementations, to be consistent with the R and scanpy convention.

6. The number of PCs used for non-spatial clustering is determined by the elbow plot per the Seurat convention. In the vignette using mouse olfactory bulb Visium data, the cell projections into the PCs are also visually inspected to exclude PCs that appear to pick up artifacts and outliers.

7. The k in k nearest neighbor graph used in Leiden clustering is the default in igraph, which is 10, not including self.

8. Leiden clustering is used because both Leiden and Louvain are conventionally used in scRNA-seq and Leiden has improved upon Louvain.

9. In Leiden clustering, the resolution parameter (0.5 in the basic vignettes) and objective function (modularity in the basic vignettes) are chosen to give a few

clusters that appear well-separated in the first few PCs but not so many that their colors are difficult to tell apart when plotting. This is for visualization purposes and may not be the best biological choice. However, we can't guarantee that the Leiden clustering results exactly match due to the random nature of the Leiden algorithm. Results can differ between R and Python and after software updates despite setting a random seed.

10. For the Chromium PBMC dataset, to find the cluster marker genes, findMarkers() in `scran` was used, with Wilcoxon rank sum tests, only testing up regulations. The most highly ranked genes are those differentially expressed in the current cluster and all other clusters (`pval.type = "all"`). The top marker gene for each cluster is the one with the smallest p-value; there were no ties in this case. The Python package reimplemented the `scran` method to match the results. These parameters were chosen because they are similar to Seurat conventions.

11. In spatial statistics performed on the k nearest graph in the Chromium PBMC dataset, $k = 10$ (not including self) was chosen to be consistent with Leiden, so the spatial statistics can be better compared to Leiden clustering.

12. "W" style edge weights are chosen for the spatial neighborhood graph, i.e. the rows of the binary adjacency matrix are normalized to sum up to 1. For a k nearest neighbor graph where all nodes have the same degree, all the edge weights are the same, so will not lead to a different Moran's I. However W style is chosen because it's preferable for the Moran scatter plot, which was performed in the vignette. Using W style, the slope of the line fitted to the Moran scatter plot is Moran's I [14]. W style is also the default across Voyager (except that binary style is recommended for Getis-Ord Gi*) because it is the default in `spdep`, it's not unreasonable, and it simplifies the math for some spatial statistics such as Moran's I, Lee's L, and MULTISPATI PCA.

13. For local Moran's I, such as shown in Fig. 13.3, `spdep` and the PySAL esda package have different defaults. In order for the results to perfectly match, set `mlvar = FALSE` in the R implementation of Voyager, which is passed to `spdep`, so both implementations divide by n-1 when computing the variance.

14. The R and Python implementations use the same palettes. Choice of the palettes has been explained in the previous section. See Fig. 13.2 on color-blindness simulations of the default palettes.

These are defaults specific to the R package and website at present:

1. By default, the image, if present, is not plotted behind the geometries of Visium spots or cells, because the geometries cover up much of the image, and the image can distort color perception of the geometries. However, plotting the image can be useful to visually relate a value in space to histology.

2. In neighborhood view ESDA, depending on how the spatial graph is defined, sometimes there are cells that don't have any spatial neighbors. The argument `zero.policy` in `spdep` determines what to do with these singleton cells. By default, in `spdep`, the global option is used, and when a global option is absent, `spdep` throws an error when there are singletons. Voyager generally follows the `spdep` default. However, in many examples, `zero.policy` is set to TRUE where singletons can occur, such as in the correlogram when some cells or spots don't have higher order neighbors. When `zero.policy = TRUE`, spatially lagged values of singletons are set to 0, while when it's `FALSE`, the spatially lagged values are `NA`. `TRUE` is chosen to silently drop the singletons when spatially lagged values are needed without stopping the computation for non-singletons.

3. In vignettes for image based single cell resolution datasets, we used k nearest neighbor graph with $k = 5$ for cell centroids, found with the KMKNN algorithm in `BiocNeighbors` (v1.18.0). We chose $k = 5$ because most real world tessellations of the 2D plane are somewhere between a square ($k = 4$ rook style) and hexagonal ($k = 6$) tessellation, and $k = 5$ seems reasonable based on visual inspection. Furthermore, for a larger dataset with over hundreds of thousands of cells, the k nearest neighbor graph is much faster to compute than most other types of neighbors, such as those that require triangulation. Although with GEOS spatial indexing, the polygon contiguity graph is fast to compute for larger datasets, given the imperfection of cell segmentation, many cells that don't appear contiguous in the polygons might in fact be physically touching, resulting into many false negatives in spatial neighbors and many singletons. As a result, we did not use the polygon contiguity graph. Inverse distance weighting is used since k nearest neighbors may or may not be physically interacting so further cells have less weight. W style edge weight normalization is used, for reasons mentioned above.

Spatial data analysis methods used in the ESDA case studies

Here we briefly introduce spatial data analysis methods necessary to understanding the ESDA case studies in Section 13.2 .

Moran's I

Tobler's first law of geography states that "Everything is related to everything else. But near things are more related than distant things." [49]. This observation motivates the examination of spatial autocorrelation. Positive spatial autocorrelation is evident when nearby things tend to be more similar, such as that weather in Pasadena and downtown Los Angeles (as opposed to the weather in Pasadena and San Francisco). Negative spatial autocorrelation is evident when nearby things tend to be more dissimilar, such as squares on a chessboard. Spatial autocorrelation can arise from an intrinsic process such as diffusion or communication by physical contact, or result from a covariate that has such an intrinsic process, or in areal data, when the areal units of observation are smaller than the scale of the spatial process [50].

Moran's I is one of the most commonly used statistic of spatial autocorrelation, defined as

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}, \tag{13.1}$$

where w is the number of spots or locations, i and j are different locations, or spots in the Visium context, x is a variable with values at each location, and w_{ij} is a spatial weight, which can be inversely proportional to distance between spots or an indicator of whether two spots are neighbors, subject to various definitions of neighborhood. Moran's I is similar to the Pearson correlation between the value at each location and the average value at its neighbors (but not identical, see [26]). Just like Pearson correlation, Moran's I is generally bound between -1 and 1, where positive value indicates positive spatial autocorrelation and negative value indicates negative spatial autocorrelation.

Local Moran's I is defined as

$$I_i = \frac{\left(x_i - \bar{x}\right)\sum_{j=1}^{n} w_{ij}\left(x_j - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2/(n-1)}, \tag{13.2}$$

an unnormalized and disaggregated form of Moran's I, as the contribution of each location to global Moran's I.

MULTISPATI PCA

The Moran's I expression above can be rearranged as

$$I = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 / n}, \tag{13.3}$$

where the denominator is the maximum likelihood estimate (MLE) of the variance of the data. Let z denote the scaled and centered data, whose mean is 0 and variance (using MLE, divide by n instead of $n - 1$) is 1. Also for simplicity, suppose the spatial weights matrix W is scaled so its rows sum to 1, so the denominator of the first term becomes n. Then Moran's I can be expressed as

$$I = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} z_i z_j / n, \tag{13.4}$$

or using matrix notation, $I = \mathbf{z}^T\mathbf{W}\mathbf{z}/n$. Let $\mathbf{Z}$ denote a matrix whose columns are scaled (divided by the standard deviation, making the variance 1) and centered (subtract the mean, thus summing to 0) variables and whose rows are observations such as cells in space. Then this expression of Moran's I can be generalized to multiple variables: $\mathbf{M} = \mathbf{Z}^T\mathbf{W}\mathbf{Z}/n$. The diagonal of $\mathbf{M}$ is Moran's I coefficients of the variables in $\mathbf{Z}$. We can find the eigenvalues and eigenvectors (i.e. diagonalize) of this matrix as a spatially informed form of PCA proposed by Wartenberg in 1985 [51].

This is analogous to PCA, in which the covariance matrix $\mathbf{X}^T\mathbf{X}/n$ is diagonalized where each variable (column) in $\mathbf{X}$ is centered. In non-spatial PCA, the first eigenvector (principal component, or PC), which has the largest eigenvalue, finds the linear combination of the original variables that explains the most variance. The eigenvalue is the amount of variance explained. The second PC (PC2) is found by maximizing the variance again provided that PC2 is orthogonal to PC1, and so on. Because the covariance matrix is symmetric and positive semidefinite, all of its eigenvalues are real and non-negative, and it has orthonormal eigenvectors, i.e.

orthogonal to each other and have length or more generally norm 1. However, the interpretation of the eigenvalues and eigenvector of $\mathbf{M}$ is not explored in reference number [51].

The 1985 Wartenberg method summarized above was generalized for the statistical triplet of multivariate data analysis in the duality diagram paradigm in reference number [27] and implemented in the `adespatial` R package; the interpretation of the eigenvalues described below was also derived. With `adespatial`, the spatial information can not only be used for PCA but also for other multivariate analyses with the duality diagram such as correspondence analysis. However, for now, Voyager's much faster implementation of MULTISPATI only applies to PCA. See [52] for an introduction to the duality diagram.

Let $\mathbf{X}$ be a data matrix with n rows and p columns with observations in rows and variables in columns, and assume that each variable is centered. MULTISPATI PCA seeks to find vector $\mathbf{u}_1$ of norm 1 that maximizes $Q(\mathbf{u}_1) = \mathbf{u}_1^T\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{u}_1/n$. Let $\mathbf{a}_1 = \mathbf{X}\mathbf{u}_1$. Then MULTISPATI PCA maximizes $Q(\mathbf{u}_1) = \mathbf{a}_1^T\mathbf{W}\mathbf{a}_1/n$. Remember the matrix expression of Moran's I, $I = \mathbf{z}^T\mathbf{W}\mathbf{z}/n$. Because $\mathbf{X}$ is centered, all columns sum to 0, so $\mathbf{1}^T\mathbf{X} = \mathbf{0}^T$, where $\mathbf{1}$ is a vector of n 1's and $\mathbf{0}$ is a vector of p 0's. Hence $\mathbf{1}^T\mathbf{a}_1 = \mathbf{1}^T\mathbf{X}\mathbf{u}_1 = 0$, meaning that $\mathbf{a}_1$ is also centered. Then we need to scale $\mathbf{a}_1$ so its variance is 1 by dividing it by its standard deviation, which is square root of the variance. The MLE of variance is $\sum_{i=1}^{n} a_{1i}^2/n = \|\mathbf{a}_1\|^2/n$, so $I(\mathbf{a}_1) = \mathbf{a}_1^T\mathbf{W}\mathbf{a}_1/\|\mathbf{a}_1\|^2$. Then $Q(\mathbf{u}_1) = I(\mathbf{a}_1)\|\mathbf{a}_1\|^2/n$. Hence for PC1, MULTISPATI PCA maximizes the product of Moran's I of the projection of the observations onto PC1 and variance explained by PC1.

Just like in non-spatial PCA, these maximizations are achieved by diagonalizing the spatially weighted covariance matrix; the eigenvalues are $Q(\mathbf{u}_i)$. Because $\mathbf{W}$ doesn't have to be symmetric, $\mathbf{X}^T\mathbf{W}\mathbf{X}$ doesn't have to be symmetric. So in practice, it's preferable to diagonalize the symmetric matrix $\mathbf{H} = \frac{1}{2n}\mathbf{X}^T(\mathbf{W}^T + \mathbf{W})\mathbf{X}$ instead, which gives the same eigenvalues as $\mathbf{X}^T\mathbf{W}\mathbf{X}/n$; the eigenvalues are guaranteed to be real when computed and the eigenvectors are orthonormal. However, since asymmetric real matrices don't have orthonormal eigenvectors, the eigenvectors of $\mathbf{H}$ are different from those of $\mathbf{X}^T\mathbf{W}\mathbf{X}/n$. The eigenvectors can be interpreted as if a symmetrized spatial weights matrix $(\mathbf{W}^T + \mathbf{W})/2$ is used for the spatially weighted covariance matrix. The effects of different choices of $\mathbf{W}$ on the results remains to be investigated.

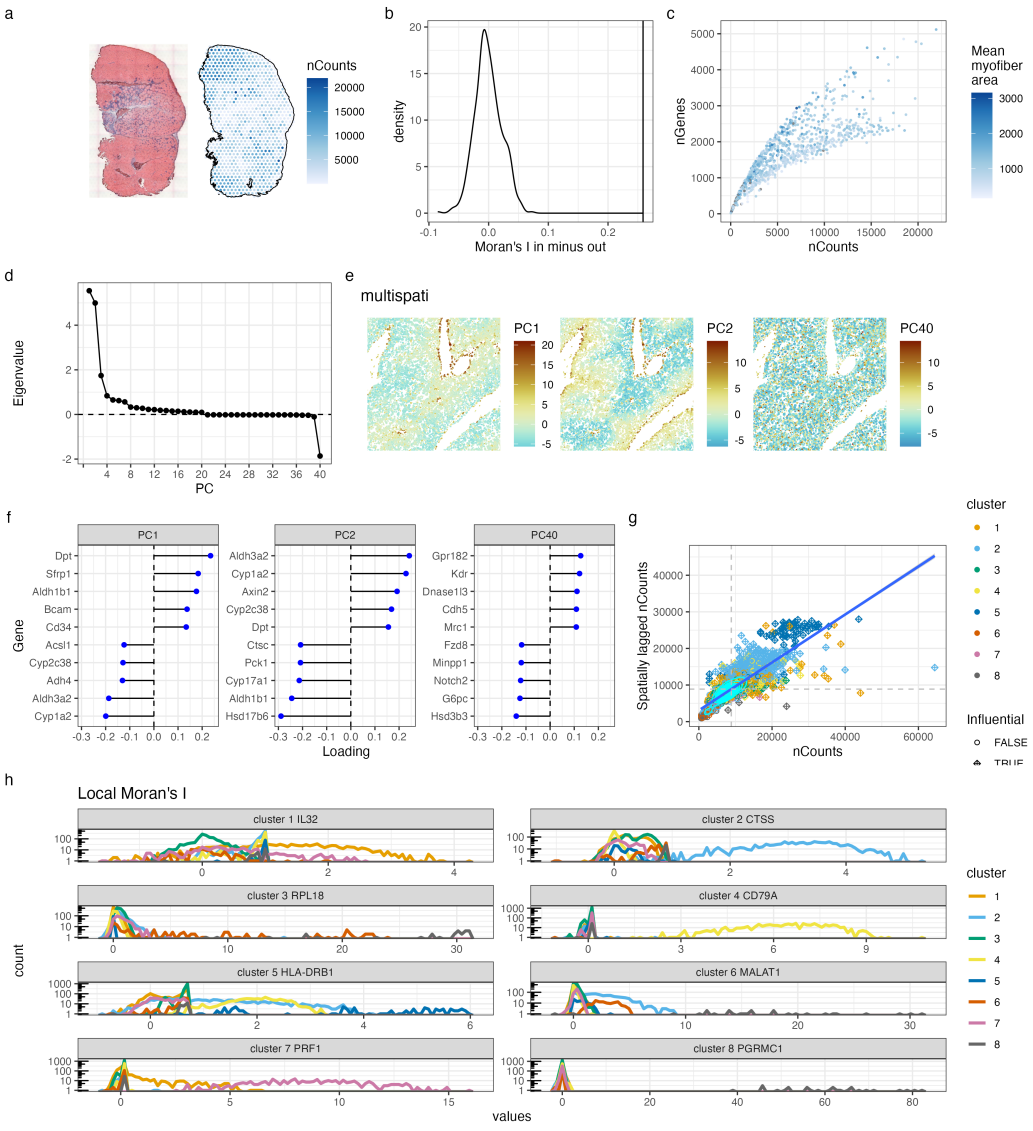

Figure 13.5: See Section 13.7 for caption.

ESDA case studies

Here we show some use cases of ESDA in spatial and non-spatial single cell transcriptomics and potential insights gained from the analyses. First, we consider a mouse skeletal muscle Visium dataset [53], two days after notexin injury (Fig. 13.5a-c). In the H&E image, the region with many leukocyte nuclei is the injury site, the darker red strip and blocks are muscle tendon junctions, and the remaining pink regions are myofibers (Fig. 13.5a). When the library size (nCounts) per spot among spots that intersect the tissue is plotted in space, we find that different histological regions have different library sizes. For example, the muscle tendon

junctions tend to have smaller library sizes than myofibers and some of the injury site, and regions with tightly packed myofibers (top and bottom left) tend to have larger library sizes than the region with larger myofibers surrounded by leukocytes (right). This confirms the finding that in spatial transcriptomics, library size confounds biology and should not be treated as a technical artifact as commonly done for scRNA-seq [54].

Further evidence is that library size in the tissue has stronger spatial autocorrelation than outside the tissue, as shown in a more positive Moran's I (Fig. 13.5b). The library size values are permuted in space for spots that intersect the tissue and those that don't, and Moran's I is computed for these permutations to estimate a null distribution. The density plot in Fig. 13.5b shows the null distribution of permuted Moran's I from spots intersecting tissue minus that from spots not intersecting tissue, and the vertical line shows the actual difference, which is much larger than all the 499 simulated values. While there is no one to one correspondence between Visium spots and myofibers, with a myofiber segmentation, geometric operations can find the myofibers that intersect each Visium spot and their areas. QC metrics library size and number of genes detected (nGenes) in this dataset are related to myofiber size; spots on larger myofibers tend to have more genes detected given the same library size than spots on smaller myofibers (Fig. 13.5c).

Next we demonstrate MULTISPATI PCA and biological relevance of negative spatial autocorrelation in a mouse liver MERFISH dataset from the Vizgen website (Fig. 13.5d-f). While non-spatial PCA maximizes variance explained by each principal component (PC) given that the PCs are orthogonal, MULTISPATI PCA maximizes the product of variance explained and Moran's I, which is the eigenvalues (Fig. 13.5d). Positive eigenvalues mean that the PCs not only explain more variance but also are spatially coherent (large positive Moran's I). Negative eigenvalues mean that the PCs not only explain more variance, which is non-negative, but also have negative spatial autocorrelation, i.e. nearby values tend to be more different. In this dataset, the positive eigenvalues show an elbow as in non-spatial PCA, and there is one substantial negative eigenvalue (Fig. 13.5d). In non-spatial PCA, the PCs are not spatially structured, until PC5 which picks up zoning (Fig. 13.6b). PC1 highlights Kupffer cells (Cdh5, Fig. 13.6a) and endothelial cells (Egfr, Fig. 13.6a), and PC2 also highlights endothelial cells. In contrast, because MULTISPATI PCA also maximizes Moran's I, zonation is picked up by the first 2 PCs. PC1 is periportal and PC2 is pericentral (Fig. 13.6e). This may complement existing

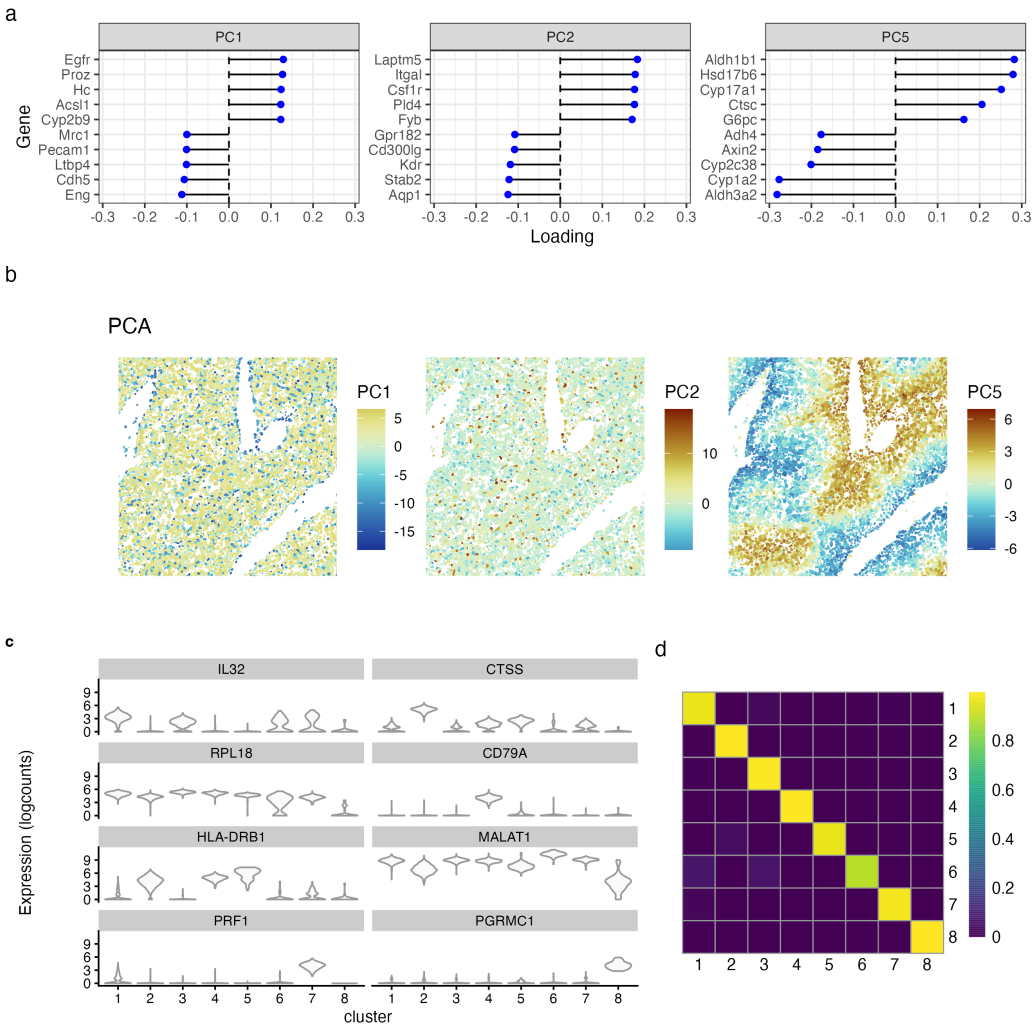

Figure 13.6: See Section 13.7 for caption.

methods to find spatially variable genes (maximize Moran's I) that are also more likely to be biologically relevant (maximize variance explained). Furthermore, the more spatially coherent PCs can be used for clustering to find more spatially coherent clusters, complementing clusters found with non-spatial PCs.

Negative spatial autocorrelation is one of the most neglected topics in spatial data analysis [55], as there are many more examples of positive than negative spatial autocorrelation. Negative spatial autocorrelation can arise from competition between neighbors (see [55]), or from functional roles played by spatial contacts between different types of entities. In this dataset, the latter seems to be the case; the PC with the most negative eigenvalue separates endothelial cells (Kdr) and Kupffer cells (Cdh5) from hepatocytes (Hsd3b3, Fig. 13.5e-f). Existing methods of spatially informed

dimension reduction [56, 57, 58] and methods to find spatially variable genes [59, 60, 61] tend to only consider positive spatial autocorrelation. This example here shows that at the single cell level, negative spatial autocorrelation is relevant to biology, and should be further investigated, as the cells—unlike Visium spots and administrative boundaries—are meaningful and non-arbitrary units of observations and of biological function.

Finally, we note that neighborhood view spatial methods can be applied to neighborhood graphs in gene expression space rather than histological or geographical space. We applied spatial statistics to QC metrics and gene expression in a 10X Chromium peripheral blood mononuclear cells (PBMC) dataset, using the k nearest neighbors graph in PCA gene expression space as the "spatial" neighborhood graph (Fig. 13.5g-h). Moran scatter plot of library size shows further evidence that library size is confounded by biology even in non-spatial data (Fig. 13.5g). In a Moran scatter plot, the x axis is the value of a variable at each cell, and the y axis is spatially lagged value (i.e. sum of values from spatial neighbors weighted by edge weights of the spatial neighborhood graph). When the adjacency matrix of the spatial neighborhood graph is row normalized, it is shown in14 that the slope of the line fitted to the scatter plot is global Moran's I, while the scatter plot shows local heterogeneity in spatial autocorrelation. Here, cluster 5 (activated T cells) tends to have larger library sizes and stronger spatial autocorrelation in library size than average (Fig. 13.5g).

Local Moran's I is a locally disaggregated form of Moran's I, showing contribution of each cell to Moran's I. More positive values indicate neighborhoods more homogeneous in the variable of interest, and more negative values indicate more heterogeneous neighborhoods. We computed local Moran's I (I_i) for the top marker gene of each cluster (Section 13.4). The marker gene has much higher I_i in cells in the cluster of interest than cells in other clusters, except for clusters 3 and 6 whose top marker genes don't clearly distinguish these clusters from most other clusters and don't seem to have cell type specific functions (Figs. 13.5h, 13.6c). When the marker gene is very specific, cells in other clusters have I_i tightly clustered around 0, such as for cluster 4 (B cells) and cluster 8 (platelets). When the marker gene is not very specific, cells in other clusters that also express the marker gene often also have I_i higher than clusters not expressing the genes (e.g. cluster 1 T cells and cluster 7 natural killer cells and cytotoxic T cells for IL32 and PRF1; cluster 5 activated T cells, cluster 2 monocytes, and cluster 4 B cells for HLA-DRB1; but this

doesn't apply to CTSS, marker gene of cluster 2 which is also somewhat expressed in clusters 4 and 5, see Fig. 13.6c). We also see a wide range of I_i values for cells in the clusters of interest (Fig. 13.5h). As Leiden clustering is also based on the k nearest neighbor graph and marker genes are typically found by performing a differential expression test on one gene at a time, I_i on cluster marker genes can shed light on how well Leiden clusters match the k nearest neighbor graph and quality of the marker genes. The downside is that I_i can be computed on only one gene at a time while a combination of multiple genes may better characterize the clusters.

13.3 Discussion

Spatial -omics has come of age. Software packages for data visualization, general EDA frameworks, and more specialized tasks have proliferated in this field, but much of the decades of ESDA research has not been systematically utilized. As an EDA framework, Voyager is somewhat akin to Seurat, squidpy, Giotto, and semla, but Voyager is unique in systematically bringing decades of ESDA research to spatial -omics, with a consistent user interface to reduce the learning curve. By reusing the SCE infrastructure and ecosystem, Voyager complements many other spatial and non-spatial data analysis methods. The SFE class extends SCE and SPE with efficient tools from the geospatial field, to represent and operate on vector geometries and raster images. Voyager has a comprehensive, reproducible, and easy to navigate documentation website with tutorials on data from various technologies and ESDA methods, with references for further reading and considerations from the ESDA tradition. With the compatibility tests to make sure the R and Python implementations give consistent results for core functionalities and transparency on defaults, the Voyager project also seeks to bridge the R vs. Python divide in the single cell and spatial -omics tool where hidden defaults and undocumented divergent implementations cause language preference to inadvertently lead to different results that may affect biological conclusions. The case studies show that ESDA can bring novel insights on biology and the process of data analysis. With the rich ESDA tradition at their fingertips, researchers in spatial -omics may know more about what they can do with the data and go further down the rabbit hole of curiosity in the spirit of EDA.

"Tradition is the living faith of the dead, traditionalism is the dead faith of the living."[62] While the ESDA tradition largely developed prior to the rise of spatial -omics can help us gain insights from the spatial aspects of the data, new methods that take into account the peculiarities of spatial -omics data, such as the larger

size, case and control and multiple biological replica, non-normal distribution of the data, and three-dimensional data from thick slices and multiple sections will take us even further. Moreover the ESDA tradition continues to evolve. Voyager is designed to be extensible by the user and developer to use the new methods with a consistent user interface. Future versions of Voyager will address these peculiarities of spatial -omics data, such as adapting some spatial methods to 3D and finding better ESDA methods for multiple biological replica and across case and control. While the SCE infrastructure already allows for on-disk gene count matrices with `DelayedArray`, the geometries and spatial analysis results are currently in memory but can get very large. A future version should also allow for on-disk geometries and spatial results. Also, while we have documented the defaults and reasons why we chose them so hidden defaults most users are unaware of don't inadvertently lead to different results in the R and Python implementations, the reasons are often convention in the field and `spdep` defaults. Further research should scrutinize effects of these parameters and find better defaults, such as the type of spatial neighborhood graph and edge weights. The problem of choosing a spatial neighborhood graph has long been studied and some methods to find a graph based on the data have been devised61, but they are not currently supported by `spdep` and may or may not be suitable for spatial -omics data. Finally, while we have chosen colorblind friendly default palettes to make Voyager a little more accessible, future research should be conducted on accessibility of spatial -omics data analysis, such as in data sonification.

13.4 Methods

All R plots in the figures were made with R 4.3.0 with Apple vecLib BLAS, Bioconductor 3.17, Voyager 1.2.3, `SpatialFeatureExperiment` 1.2.1, `scater` 1.28.0, `spdep` 1.2.8, `Seurat` 4.3.0, `sf` 1.0.12, and `ggplot2` 3.4.2, on MacOS Ventura 13.3.1, 2.3 GHz Dual-Core Intel Core i5, 8 GB RAM. R package profvis 0.3.7 was used to profile time and memory usage by lines of code in the benchmarks, and bench 1.1.2 was used for the benchmarks over different numbers of cells. When comparing Seurat vs. scanpy and the R and Python implementations of Voyager, the Python code was run through `reticulate` (v1.28) in RStudio. Python 3.10 and `scanpy` 1.9.3 were used. Multipanel plots were assembled with patchwork 1.1.2 when all panels are R plots, and were otherwise assembled in LibreOffice Draw.

Spatial methods

At present, all neighborhood view spatial methods are implemented in `spdep` and wrapped by Voyager, except for Lee's L which has a more efficient implementation in Voyager. Defaults follow thos in `spdep`. All distance view spatial methods are implemented in `gstat` and wrapped by Voyager. Variogram model fitting is implemented in `automap`, which is a user-friendly wrapper of `gstat` that tries a number of different models and selects the one with the best fit. Voyager has a more efficient implementation of MULTISPATI PCA which is originally implemented in `adespatial`, thus not importing `adespatial`.

Compatibility tests

Everything in the two core vignettes other than the plots themselves are subject to compatibility tests to see if the R and Python implementations of Voyager give the same results for core functionalities. This is typically anything that gives numeric output, such as PCA and Moran's I results. The plots can't be quantitatively and automatically compared because of the different default styles and mechanisms of `ggplot2` and `matplotlib`; visual similarity would suffice. The "epsilon", or numeric differences that can be accounted for by machine double precision, is `sqrt(.Machine$double.eps)` in R. To compare PCA eigenvectors (gene loadings), cosine difference is used to geometrically compare the vectors, which is implemented as the magnitude of difference between the cosine of the angle between the two vectors and 1, i.e. cosine of 0 and 180 degrees; 180 degrees because the eigenvectors can be flipped and remain equivalent PCA results. This comparison was performed to each of the first 50 PCs individually for Fig. 13.3. To compare the proportion of variance explained, the absolute value of the difference was used.

Website build

The R Voyager documentation website is built with pkgdown on GitHub Actions, which builds function references and vignettes from the R package source code. All imported and suggested packages are installed on a fresh machine on the cloud and all vignettes are run on the cloud to be rendered, to ensure that they are reproducible. The Google Colab notebooks are automatically generated from the R Markdown vignettes with another GitHub Action. Because Bioconductor limits the installed size of the package, which includes the rendered vignettes, the vignettes on the documentation website are in a separate documentation branch from the main and devel branches that sync with Bioconductor, while a shorter vignette is

on Bioconductor. Also, there are packages suggested in the documentation branch but not the main branch, as while they are used in vignettes on the website, they are not used on the Bioconductor vignette. The code in the documentation branch is synced with code from the main branch by merging from the main branch, but the documentation branch is never merged into the main branch.

Performance improvements

In the benchmarks, a mouse liver MERFISH dataset from the Vizgen website with over 390,000 cells after QC was used. After removing cells with high proportion of transcripts from blank barcodes and the blank barcodes, the dataset was subsetted with bounding boxes of different sizes to produce datasets of different sizes while preserving spatial relationships among cells. Then the datasets of different sizes were used in the benchmarks.

K nearest neighbors with inverse distance weighting

The `spdep` implementation of distance based edge weights is slow because while `spdep` uses an efficient implementation of k nearest neighbors and distance neighbors in dbscan, it discards the distances between neighbors returned by dbscan. As a result, `spdep` has to re-compute the distances to compute the edge weights (Fig. 13.8). The implementation in SFE uses BiocNeighbors to find the k nearest and distance based neighbors, allowing users to choose from a number of different algorithms. Then the distances are saved for edge weight computations, skipping the most time consuming step. While dbscan is not much slower than BiocNeighbors when finding the neighbors (Fig. 13.8a-b), we found that not recomputing the distances speeds up finding the spatial neighborhood graph from 8 to over 30 times and using over 25 times less memory (Fig. 13.7a-b).

MULTISPATI PCA

The `adespatial` implementation of MULTISPATI PCA uses base R eigen decomposition, which always computes all eigenvalues and eigenvectors. Then `adespatial` discards the remaining eigenvectors when the user specifies a small number of eigenvectors, which is typical in single cell and spatial -omics. Furthermore, `adespatial` in fact performs the eigen decomposition twice. The first time is in `dudi.pca`, which performs non-spatial PCA, whose results are passed to the `multispati` function, which takes some weights and the original data but not the eigenvalues or eigenvectors from the `dudi.pca` output, and then performs

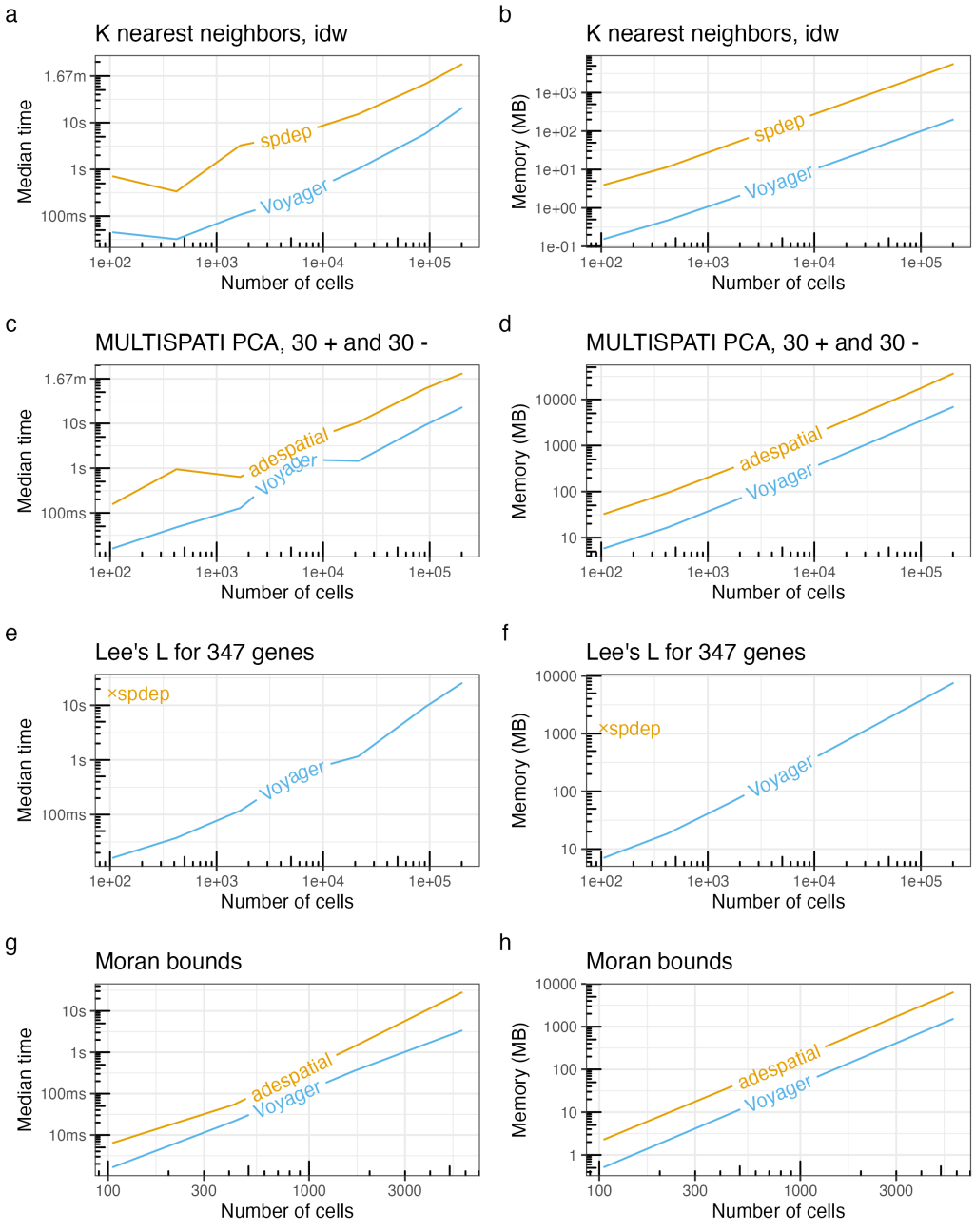

Figure 13.7: See Section 13.7 for caption.

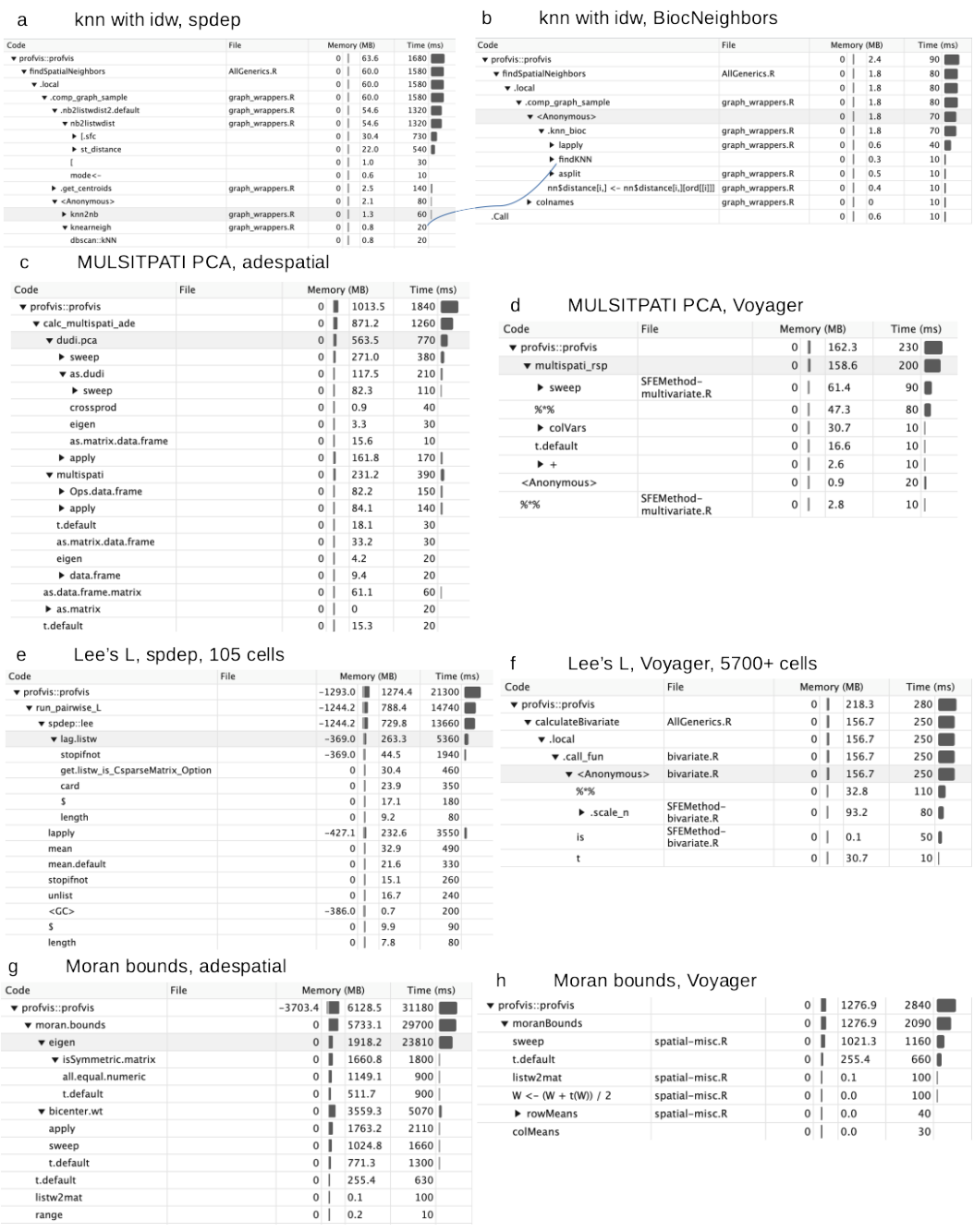

Figure 13.8: See Section 13.7 for caption.

eigen decomposition of the spatially weighted covariance matrix (Fig. 13.8c). The implementation in Voyager uses `RSpectra` for a more optimized implementation of partial eigen decomposition of the spatially weighted covariance matrix, only computing the eigenvectors the user requested and only once, hence avoiding a lot of unnecessary computation, speeding up computation by 2 to 20 folds, and using over 5 times less memory (Fig. 13.7c-d).

Lee's L

The `spdep` implementation of Lee's L computes both local and global Lee's L for one pair of genes at a time. As spatial transcriptomics data has hundreds (smFISH-based) to thousands of genes (sequencing based), Voyager's implementation uses matrix operations to make it more efficient to compute Lee's L for a large number of genes than iterating through each pair. This speeds up computation over 800 folds and using 100 times less memory, where one thread was used to iterate through all pairs of genes in the dataset for the `spdep` implementation (Fig. 13.7e-f). The inefficiency of `spdep`'s implementation is not only iterating through the pairwise combinations, but also a less efficient way to compute the spatially lagged values and the sum of edge weights (13.7e). Iterating through the pairs with the `spdep` implementation is so slow that it was only run for the smallest subset with 105 cells in the benchmark.

Moran bounds

Bounds of Moran's I given a spatial neighborhood graph can be computed from the largest and smallest eigenvalues of the double centered adjacency matrix of the graph [21]. The `adespatial` implementation of the function finding Moran bounds, all eigenvalues are computed. In Voyager's implementation, `RSpectra` is used to find only the largest and smallest eigenvalues of the matrix, without computing eigenvectors or the other irrelevant eigenvalues, speeding up computation by 4 folds and using 4 times less memory (Fig. 13.7g-h). While much of the time was spent in the eigen decomposition in the `adespatial` implementation, most of the time was spent on double centering in the Voyager implementation (Supplementary 13.7g-h). Due to double centering, a dense matrix with as many columns and rows as the number of cells is produced. Unless this can be avoided, this computation consumes a lot of memory for larger datasets. As a result, neither implementation scaled beyond around 6000 cells.

ESDA case studies

Mouse skeletal muscle Visium data

Space Ranger processed data was downloaded from GEO accession GSE161318, sample "Vis5A", 2 days after notexin injury. Myofiber segmentation was performed manually with the LabKit ImageJ plugin, exported as TIFF raster, and converted to polygons with `terra`. Redundant vertices of the polygons were removed when the polygons were simplified with `rmapshaper` (v0.4.5). Visium spot polygons were found from the centroids and spot diameters in pixels in the full resolution image from the Loupe image alignment JSON file. The tissue was segmented by thresholding the H&E image and removing small pieces. Then the thresholded mask was converted into polygon with `terra`, which was used to find spots intersecting the tissue with `sf`. The gene count matrix and polygons were made into an SFE object available to download from the `SFEData` package. The `sf` package was used to find myofiber areas and which myofibers intersect each Visium spot polygon. Only the full resolution H&E images were available on GEO; the image was downsampled to fit into a 1024×1024 pixel box for the vignettes and examples. Moran's I permutation tests were performed on nCounts for spots intersecting the tissue and spots not intersecting the tissue separately, with 499 permutations. The values are permuted in space. Then the simulated Moran's I's from spots not intersecting tissue were subtracted from those from spots intersecting tissue to form a null distribution of this difference. The observed value is the observed Moran's I from spots not intersecting tissue subtracted from that of spots intersecting tissue.

Mouse liver MERFISH data

The gene count matrix, cell metadata, and cell segmentation polygons were downloaded from Vizgen's website, and formed into an SFE object, which is available in the `SFEData` package. The scuttle package (1.10.0) was used to remove low quality cells. Proportion of transcripts attributed to blank barcodes was computed and log2 transformed, and cells with log proportion more than 3 median absolute deviations (MADs) higher than the median were deemed low quality and removed. Then the filtered gene count matrix was normalized by `logNormCounts()` in `scater`, and the genes are scaled and centered before performing non-spatial PCA with IRLBA through `scater`, and MULTISPATI PCA. MULTISPATI PCA requires a spatial neighborhood graph; see the Compatibility Tests section for reasons behind the parameters chosen.

Chromium PBMC data

The filtered 5k PBMC NextGem v3 data, processed with Cell Ranger 3.0.2, was downloaded from the 10X Genomics website and loaded into R as an SCE object, which is then converted to SFE for "spatial" analyses. Cells with at least 20% of UMIs from mitochondrially encoded genes were removed. Highly variable genes (HVGs) were found with the `scran` method, but without the Lowess fit. The 2000 genes with the highest biological component were used for PCA. The data was normalized with `logNormCounts()` in `scater`, and the genes were scaled and centered before performing PCA with the IRLBA algorithm. Based on the variance explained elbow plot, the 10 PCs were used to build a k nearest neighbor graph, with $k = 10$ (not including self). For Leiden clustering, $k = 10$ was also used so the clustering results can be compared to the "spatial" results. The objective function is "modularity" and the resolution parameter is 0.5. For the "spatial" neighborhood graph, inverse distance weighting and W style normalization were used, for reasons similar to those in k nearest neighbor graphs in histological space explained in the Compatibility Tests section. Differential expression is described in the Compatibility Tests section.

13.5 Data and code availability

The mouse skeletal muscle Visium dataset and the mouse liver MERFISH datasets are available in the `SFEData` R package. The GitHub repositories and websites related to this paper are linked below:

Voyager R package

SpatialFeatureExperiment

`SFEData`

Voyager Python package

Voyager R documentation website

Voyager Python documentation website

Code to reproduce figures

13.6 Acknowledgement

I thank Dario Righelli, author of SPE, for discussion of the SFE concept early in its development.

13.7 Figure legends

Figure 13.1: Schematic of the Voyager framework. Voyager brings ESDA methods initially developed for geospatial data to spatial -omics, with a consistent user interface for different methods. Voyager is based on the SpatialFeatureExperiment object, which uses `sf` and `terra` to extend SingleCellExperiment and SpatialExperiment. Voyager implements plotting functions for gene expression, cell attributes, and spatial analysis results. The documentation website has vignettes that demonstrate ESDA on data from multiple spatial -omics technologies, including Visium, Slide-seq, Xenium, CosMX, MERFISH, seqFISH, CODEX, and Chromium. The website is built automatically with GitHub Actions and pkgdown for reproducibility, and Google Colab notebooks are automatically generated from the vignettes. There is a Python implementation which uses PySAL and GeoPandas for the ESDA and geometry operations. Compatibility tests are used to make sure that the R and Python implementations give consistent results for core functionalities.

Figure 13.3: Comparisons between Seurat and scanpy PCA, and between Voyager R and Python for PCA and Moran's I. a, Comparison of Visium spot embeddings in the first 2 PCs from Seurat and scanpy with default parameters. The lines connect corresponding spots in Seurat and scanpy. **b,** As in A, but for Voyager R and VoyagerPy, with parameters stated in this section. **c,** Cosine distances between the first 20 PCA eigenvectors (gene loadings) from Seurat and scanpy (yellow), and from Voyager R and Python (blue). The dashed line is the magnitude that can be explained by machine double precision. The text part of the line is somewhat smoothed for readability but should not affect interpretation. **d,** Absolute values of differences in proportion of variance explained by each of the top 20 PCs. **e,** Moran's I from VoyagerPy vs. Voyager R. The blue line is y = x, showing that the results are consistent. **f,** Same as E but for local Moran's I for gene S100a5. **g,** Plotting the local Moran's I values in space, with the H&E image behind the spots, from Voyager R (top) and VoyagerPy (bottom).

Figure 13.2: a, All palettes used in the Voyager R package, full color vision. **b,** Deutanomaly perception of the palettes. **c,** Protanomaly perception of the palettes. **d,** Tritanomaly perception of the palettes. **e,** Desatruated palettes. A dark theme is implemented to better visualize data with a fluorescent image in the background. The light and dark themes have different default palettes; in the light theme, darker color denotes higher values as if staining, while in the dark theme, lighter color denotes higher values as if glowing, so higher values stand out from the background. It is

possible to simultaneously use two different palettes within the light or dark theme, such as to color Visium spots by one palette and cell segmentation from the same dataset with another in the same plot, but this should be used with caution. While the different palettes within one theme are chosen to avoid similar colors as much as possible, we do not suggest using two divergent palettes simultaneously, because doing so can distort color perception of either palette. We also do not suggest people with color vision deficiencies to use any two palettes simultaneously in one plot.

Figure 13.5: a, In a mouse skeletal muscle dataset, the total UMI counts, or library size per spot (nCounts), are plotted in space. Only spots that intersect tissue are plotted. The H&E image is plotted on the side as a reference. **b,** Simulated (density plot) and observed (vertical line) difference between Moran's I in nCounts of spots that intersect tissue (in) and that of spots that don't (out). **c,** Scatter plot of number of genes detected per spot (nGenes) vs. nCounts, colored by mean area of myofibers that intersect each spot. **d,** The 20 most positive and 20 most negative eigenvalues from MULTISPATI PCA of a mouse liver MERFISH dataset. As other eigenvalues were not computed, there's a break after PC20 in this plot. **e,** A subset of the MERFISH data showing a portal triad (near top right) and two central veins (left and bottom right), with cell polygons colored by their projections into 2 PCs with the most positive eigenvalues and the PC with the most negative eigenvalue ("PC40"). The first 2 PCs show zonation. **f,** The most positive and negative gene loadings for PCs 1, 2, and "40". **g,** Moran scatter plot of nCounts in a 10X Chromium human PBMC dataset. The spatial lags were computed with the k nearest neighbors graph in PCA gene expression space. The line is least square fitted to the scatter plot. The gray shade around the line is the 95% confidence interval of the fit. Contours show point density. **h,** Histograms of local Moran's I values per cell of top marker genes of each cluster in the PBMC dataset, colored by cell cluster. The y axis (number of cells per bin) is log transformed for better dynamic range. The histograms are plotted as lines instead of bars to avoid overlapping bars from different clusters.

Figure 13.7: Benchmarks of time and memory use of the original implementations and the more efficient implementations in Voyager. a,b, K nearest neighbor graph with inverse distance weighting (idw), with $k = 5$, W style. **c,d,** MULTISPATI PCA with 30 positive and 30 negative eigenvalues, using the same k nearest graphs from a,b. **e,f,** Lee's L for 347 genes. **g,h,** Finding bounds of Moran's I given spatial neighborhood graph (same as in a-b).

Figure 13.8: Screenshots from profiling original implementations (left column) and

the more efficient implementations in Voyager (right column). Both implementations were run on the same dataset unless otherwise noted. **a,** Most of the time is spent on re-finding distances between neighbors when using k nearest neighbors (knn) in `spdep` with inverse distance weighting (idw). **b,** Voyager avoids this time consuming step; the curve connects code in each implementation that finds the knn. **c,** Much of the time is spent on preprocessing the data frame input in the `adespatial` implementation of MULTISPATI PCA ("sweep" to scale and center, and "apply" to compute spatial lag). Also note that eigen was called twice. **d,** In the Voyager implementation, the direct input to the MULTISPATI function is a matrix. Much of the time was spent on scaling and centering the matrix (sweep) and computing the spatially weighted covariance matrix with matrix multiplication, which are much faster than the preprocessing steps in `adespatial`. For this dataset with around 1000 cells, time spent on partial eigen decomposition with RSpectra was negligible so it didn't show up in the profile. **e,** In `spdep`, much of the time was spent computing the spatially lagged values (lag.listw) and computing the sum of edge weights (lapply) when computing Lee's L for each pair of features. **f,** In Voyager, the listw spatial neighborhood graph is first converted to a sparse matrix so it's much faster to compute the sum of all non-zero entries and to compute the spatial lags for many features at once with matrix multiplication. For the same 347 genes, Voyager's implementation run on 5785 cells was much faster than `spdep`'s implementation run on 105 cells in e. **g,** When finding Moran bounds with `adespatial`, most of the time was taken up by finding all the eigenvalues even though only the largest and smallest ones are needed. **h,** Most of the time and memory was spent on creating the double centered matrix (sweep, transpose, and etc.) while time spent finding the largest and smallest eigenvalues with RSpectra was negligible in Voyager's implementation of Moran bounds.

Figure 13.4: Comparisons of PCA results from Seurat and scanpy using the same highly variable genes from Seurat. a, There's no visible difference between the Seurat and scanpy cell projections in the first 2 PCs. **b,** There are visible differences between Seurat and scanpy cell projections in PC3 and PC4. The lines connect corresponding cells from Seurat and scanpy. **c,d,** When not clipping the scaled data in Seurat, there's no visible difference between the Seurat and scanpy cell projections in the first 4 PCs. **e,** Cosine differences in each of the top 50 PCA eigenvectors between Seurat and scanpy. The dashed line is epsilon, or what can be accounted for by machine double precision. **f,** Absolute differences in proportion of variance explained by each PC in Seurat and scanpy. The differences are 5 orders of

magnitude smaller without clipping than with default parameters. Without clipping, the differences are also within epsilon after the first two PCs, indicating that Seurat's clipping default which differs from that of scanpy is causing the different PCA results. e and f were made with the geomtextpath package v0.1.1.

Figure 13.6: **a,** Gene loadings of PCs 1, 2, and 5 from non-spatial PCA. **b,** MERFISH cell polygons colored by cell projections in PCs 1, 2, and 5. **c,** Violin plots of log normalized counts of the top marker gene of each Leiden cluster in the PBMC dataset. **d,** Concordex heatmap for the PBMC Leiden clusters, made with concordexR v1.0.0 [63]. High diagonal and low off diagonal values indicate high clustering quality, or that the Leiden clusters reflect the k nearest neighbor graph well, but cluster 6 has somewhat lower quality.

References

1. Moses L and Pachter L. Museum of spatial transcriptomics. Nature Methods 2022; 19. DOI: 10.1038/s41592-022-01409-2

2. Hao M, Hua K, and Zhang X. SOMDE: A scalable method for identifying spatially variable genes with self-organizing map. bioRxiv 2021 Jan :2020.12.10.419549. DOI: 10.1101/2020.12.10.419549. Available from: http://biorxiv.org/content/early/2021/03/24/2020.12.10.419549.abstract

3. Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, Rybakov S, Ibarra IL, Holmberg O, Virshup I, Lotfollahi M, Richter S, and Theis FJ. Squidpy: a scalable framework for spatial omics analysis. en. Nat. Methods 2022 Feb; 19:171–8

4. Dries R, Zhu Q, Dong R, Eng Chee-Huat Linus, Li H, Liu K, Fu Y, Zhao Tianxiao, Sarkar A, Bao F, George RE, Pierson N, Cai L, and Yuan GC. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. en. Genome Biol. 2021 Mar; 22:78

5. Bergenstråhle J, Larsson L, and Lundeberg J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. BMC Genomics 2020 Dec; 21:482. DOI: 10.1186/s12864-020-06832-3. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-06832-3

6. Pechuan-Jorge X, Li X, Risom T, Zubkov A, Tabatsky E, Prilipko A, Ye X, Shi Z, Nowicka M, Peale F, Hibar D, Ziai J, Jesudason R, and Orlova D. SPEX: A modular end-to-end analytics tool for spatially resolved omics of tissues. bioRxiv 2022 Aug :2022.08.22.504841. DOI: 10.1101/2022.08.22.504841. Available from: https://www.biorxiv.org/content/

10.1101/2022.08.22.504841v1%20https://www.biorxiv.org/content/10.1101/2022.08.22.504841v1.abstract

7. Behanova A, Avenel C, Andersson A, Chelebian E, Klemm A, Wik L, Östman A, and Wählby C. Visualization and quality control tools for large-scale multiplex tissue analysis in TissUUmaps3. Biological Imaging 2023 Feb; 3:e6. DOI: 10.1017/S2633903X23000053. Available from: https://www.cambridge.org/core/product/identifier/S2633903X23000053/type/journal_article

8. Preibisch S, Karaiskos N, and Rajewsky N. Image-based representation of massive spatial transcriptomics datasets. bioRxiv 2022 Jun :2021.12.07.471629. DOI: 10.1101/2021.12.07.471629. Available from: https://www.biorxiv.org/content/10.1101/2021.12.07.471629v2%20https://www.biorxiv.org/content/10.1101/2021.12.07.471629v2.abstract

9. Sriworarat C, Nguyen A, Eagles NJ, Collado-Torres L, Martinowich K, Maynard KR, and Hicks SC. Performant web-based interactive visualization tool for spatially-resolved transcriptomics experiments. bioRxiv 2023 Feb :2023.01.28.525943. DOI: 10.1101/2023.01.28.525943. Available from: https://www.biorxiv.org/content/10.1101/2023.01.28.525943v2%20https://www.biorxiv.org/content/10.1101/2023.01.28.525943v2.abstract

10. Bienroth D, Charitakis N, Jaeger-Honz S, Garkov D, Elliott DA, Porrello ER, Klein K, Nim HT, Schreiber F, and Ramialison M. Spatially Resolved Transcriptomics Mining in 3D and Virtual Reality Environments with VR-Omics. bioRxiv 2023 Apr :2023.03.31.535025. DOI: 10.1101/2023.03.31.535025. Available from: https://www.biorxiv.org/content/10.1101/2023.03.31.535025v1%20https://www.biorxiv.org/content/10.1101/2023.03.31.535025v1.abstract

11. Kim EH, Howard D, Chen Y, Tripathy SJ, and French L. LaminaRGeneVis: A Tool to Visualize Gene Expression Across the Laminar Architecture of the Human Neocortex. Frontiers in Neuroinformatics 2022 Feb; 16:8. DOI: 10.3389/FNINF.2022.753770/BIBTEX

12. Guo B and Hicks SC. escheR: Unified multi-dimensional visualizations with Gestalt principles. bioRxiv 2023 Mar :2023.03.18.533302. DOI: 10.1101/2023.03.18.533302. Available from: https://www.biorxiv.org/content/10.1101/2023.03.18.533302v1%20https://www.biorxiv.org/content/10.1101/2023.03.18.533302v1.abstract

13. Bergenstråhle J, Bergenstråhle L, and Lundeberg J. SpatialCPie: an R/Bioconductor package for spatial transcriptomics cluster evaluation. BMC Bioinformatics 2020 Dec; 21:161. DOI: 10.1186/s12859-020-3489-7. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-3489-7

14. Anselin L. The Moran scatterplot as an ESDA tool to assess local instability in spatial association. Spatial Analytical Perspectives on GIS 1996. DOI: 10.1201/9780203739051-8

15. Wickham H and Grolemund G. R for Data Science: Visualize, Model, Transform, Tidy, and Import Data. 2017

16. Tukey JW. Exploratory Data Analysis. en. Addison-Wesley Publishing Company, 1977

17. Moran PAP. Notes on continuous stochastic phenomena. en. Biometrika 1950 Jun; 37:17–23

18. Geary RC. The Contiguity Ratio and Statistical Mapping. The Incorporated Statistician 1954; 5:115–46

19. Griffith DA and Chun Y. Some useful details about the Moran coefficient, the Geary ratio, and the join count indices of spatial autocorrelation. Journal of Spatial Econometrics 2022; 3. DOI: 10.1007/s43071-022-00031-w

20. Griffith DA. The Moran coefficient for non-normal data. Journal of Statistical Planning and Inference 2010; 140. DOI: 10.1016/j.jspi.2010.03.045

21. Jong P de, Sprenger C, and Veen F van. On Extreme Values of Moran's I and Geary's c. Geographical Analysis 1984; 16. DOI: 10.1111/j.1538-4632.1984.tb00797.x

22. Cliff AD and Ord JK. Spatial Processes: Models & Applications. en. Pion, 1981

23. Cressie N. Statistics for Spatial Data. en. Wiley, 1993 Sep

24. Anselin L. Local Indicators of Spatial Association—LISA. en. Geogr. Anal. 1995 Apr; 27:93–115

25. Ord JK and Getis A. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. en. Geogr. Anal. 1995 Oct; 27:286–306

26. Lee SI. Developing a bivariate spatial association measure: An integration of Pearson's r and Moran's I. J. Geogr. Syst. 2001 Dec; 3:369–85

27. Dray S, Sa\"\id S, and Débias F. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. en. J. Veg. Sci. 2008 Feb; 19:45–56

28. Pullin JM and McCarthy DJ. A comparison of marker gene selection methods for single-cell RNA sequencing data. en. bioRxiv 2022 Sep :2022.05.09.490241

29. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, Marini F, Rue-Albrecht K, Risso D, Soneson C, Waldron L, Pagès H, Smith ML, Huber W, Morgan M, Gottardo R, and Hicks SC. Orchestrating single-cell analysis with Bioconductor. en. Nat. Methods 2020 Feb; 17:137–45

30. Righelli D, Weber LM, Crowell HL, Pardo B, Collado-Torres L, Ghazanfar S, Lun AT, Hicks SC, and Risso D. SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor. Bioinformatics 2022 May; 38:3128–31. DOI: 10.1093/BIOINFORMATICS/BTAC299. Available from: https://academic.oup.com/bioinformatics/article/38/11/3128/6575443

31. Pebesma E. Simple features for R: Standardized support for spatial vector data. en. R J. 2018; 10:439

32. Zhao E, Stone MR, Ren X, Guenthoer J, Smythe KS, Pulliam T, Williams SR, Uytingco CR, Taylor SEB, Nghiem P, Bielas JH, and Gottardo R. Spatial transcriptomics at subspot resolution with BayesSpace. en. Nat. Biotechnol. 2021 Nov; 39:1375–84

33. Bunis DG, Andrews J, Fragiadakis Gabriela K, Burt TD, and Sirota M. dittoSeq: universal user-friendly single-cell and bulk RNA sequencing visualization toolkit. en. Bioinformatics 2021 Apr; 36:5535–6

34. Crameri F. Scientific colour maps. 2018 May

35. Harrower M and Brewer CA. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. Cartogr. J. 2003 Jun; 40:27–37

36. Bivand R. R packages for analyzing spatial data: A comparative case study with areal data. en. Geogr. Anal. 2022 Jul; 54:488–518

37. Pebesma EJ. Multivariable geostatistics in S: the gstat package. Computers & Geosciences 2004 Aug; 30:683–91. DOI: 10.1016/J.CAGEO.2004.03.012

38. Kuhn M and Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. 2020

39. McCarthy DJ, Campbell KR, Lun ATL, and Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. en. Bioinformatics 2017 Apr; 33:1179–86

40. Janesick A, Shelansky R, Gottscho AD, Wagner F, Rouault M, Beliakoff G, Oliveira MFd, Kohlway A, Abousoud J, Morrison CA, Drennon TY, Mohabbat SH, Williams SR, Teams 1D, and Taylor SE. High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue. bioRxiv 2022 Nov :2022.10.06.510405. DOI: 10.1101/2022.10.06.510405. Available from: https://www.biorxiv.org/content/10.1101/2022.10.06.510405v2%20https://www.biorxiv.org/content/10.1101/2022.10.06.510405v2.abstract

41. He S, Bhatt R, Brown C, Brown EA, Buhr DL, Chantranuvatana K, Danaher P, Dunaway D, Garrison RG, Geiss G, Gregory MT, Hoang ML, Khafizov R, Killingbeck EE, Kim D, Kim TK, Kim Y, Klock A, Korukonda M, Kutchma

A, Lewis ZR, Liang Y, Nelson JS, Ong GT, Perillo EP, Phan JC, Phan-Everson T, Piazza E, Rane T, Reitz Z, Rhodes M, Rosenbloom A, Ross D, Sato H, Wardhani AW, Williams-Wietzikoski CA, Wu L, and Beechem JM. High-plex Multiomic Analysis in FFPE at Subcellular Level by Spatial Molecular Imaging. bioRxiv 2022 Jul :2021.11.03.467020. DOI: 10.1101/2021.11.03.467020. Available from: https://www.biorxiv.org/content/10.1101/2021.11.03.467020v3%20https://www.biorxiv.org/content/10.1101/2021.11.03.467020v3.abstract

42. Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, and Zhuang X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. Proceedings of the National Academy of Sciences of the United States of America 2016. DOI: 10.1073/pnas.1612826113

43. Stickels RR, Murray E, Kumar P, Li J, Marshall JL, Di Bella DJ, Arlotta P, Macosko EZ, and Chen F. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. en. Nat. Biotechnol. 2021 Mar; 39:313–9

44. Lohoff T, Ghazanfar S, Missarova A, Koulena N, Pierson N, Griffiths JA, Bardot ES, Eng CH, Tyser RC, Argelaguet R, Guibentif C, Srinivas S, Briscoe J, Simons BD, Hadjantonakis AK, Göttgens B, Reik W, Nichols J, Cai L, and Marioni JC. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. Nature Biotechnology 2021 40:1 2021 Sep; 40:74–85. DOI: 10.1038/s41587-021-01006-2. Available from: https://www.nature.com/articles/s41587-021-01006-2

45. Black S, Phillips D, Hickey JW, Kennedy-Darling J, Venkataraaman VG, Samusik N, Goltsev Y, Schürch CM, and Nolan GP. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. en. Nat. Protoc. 2021 Aug; 16:3802–35

46. Rey SJ, Anselin L, Amaral P, Arribas-Bel D, Cortes RX, Gaboardi JD, Kang W, Knaap E, Li Z, Lumnitz S, Oshan TM, Shao H, and Wolf LJ. The PySAL ecosystem: Philosophy and implementation. en. Geogr. Anal. 2022 Jul; 54:467–87

47. Jordahl K, Bossche JV den, Fleischmann M, Wasserman J, McBride J, Gerard J, Tratner J, Perry M, Badaracco AG, Farmer C, Hjelle GA, Snow AD, Cochran M, Gillies S, Culbertson L, Bartos M, Eubank N, maxalbert, Bilogur A, Rey S, Ren C, Arribas-Bel D, Wasser L, Wolf LJ, Journois M, Wilson J, Greenhall A, Holdgraf C, Filipe, and Leblanc F. geopandas/geopandas: v0.8.1. 2020 Jul. DOI: 10.5281/zenodo.3946761. Available from: https://doi.org/10.5281/zenodo.3946761

48. Lun ATL, McCarthy DJ, and Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. en. F1000Res. 2016 Aug; 5:2122

49. Tobler WR. A Computer Movie Simulating Urban Growth in the Detroit Region. Econ. Geogr. 1970 Jun; 46:234–40

50. Pebesma E and Bivand R. Spatial Data Science: With Applications in R. en. CRC Press, 2023 May. Chap. 15

51. Wartenberg D. Multivariate spatial correlation: A method for exploratory geographical analysis. en. Geogr. Anal. 1985 Sep; 17:263–83

52. De la Cruz O and Holmes S. The duality diagram in data analysis: Examples of modern applications. https://doi.org/10.1214/10-AOAS408 2011 Dec; 5:2266–77. DOI: 10.1214/10-AOAS408. Available from: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-5/issue-4/The-duality-diagram-in-data-analysis--Examples-of-modern/10.1214/10-AOAS408.full%20https://projecteuclid.org/journals/annals-of-applied-statistics/volume-5/issue-4/The-duality-diagram-in-data-analysis--Examples-of-modern/10.1214/10-AOAS408.short

53. McKellar DW, Walter LD, Song LT, Mantri M, Wang MFZ, De Vlaminck I, and Cosgrove BD. Large-scale integration of single-cell transcriptomic data captures transitional progenitor states in mouse skeletal muscle regeneration. en. Commun Biol 2021 Nov; 4:1280

54. Bhuva DD, Tan CW, Marceaux C, Chen J, Kharbanda M, Jin X, Liu Ningand, Feher K, Putri G, Asselin-Labat ML, Phipson B, and Davis MJ. Library size confounds biology in spatial transcriptomics data. en. bioRxiv 2023 Mar :2023.03.15.532733

55. Griffith DA. Negative Spatial Autocorrelation: One of the Most Neglected Concepts in Spatial Statistics. en. Stats 2019 Aug; 2:388–415

56. Shang L and Zhou X. Spatially aware dimension reduction for spatial transcriptomics. en. Nat. Commun. 2022 Nov; 13:7203

57. Townes FW and Engelhardt BE. Nonnegative spatial factorization applied to spatial genomics. en. Nat. Methods 2023 Feb; 20:229–38

58. Velten B, Braunger JM, Argelaguet R, Arnol D, Wirbel J, Bredikhin D, Zeller G, and Stegle O. Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. en. Nat. Methods 2022 Feb; 19:179–86

59. Svensson V, Teichmann SA, and Stegle O. SpatialDE: Identification of spatially variable genes. Nature Methods 2018 Apr; 15:343–6. DOI: 10.1038/nmeth.4636. Available from: https://www.nature.com/articles/nmeth.4636

60. Sun S, Zhu J, and Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nature Methods 2020 Feb; 17:193–200. DOI: 10.1038/s41592-019-0701-7. Available from: https://doi.org/10.1038/s41592-019-0701-7

61. BinTayyash N, Georgaka S, John ST, Ahmed S, Boukouvalas A, Hensman J, and Rattray M. Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments. bioRxiv 2020 Jan :2020.07.29.227207. DOI: 10.1101/2020.07.29.227207. Available from: http://biorxiv.org/content/early/2020/07/30/2020.07.29.227207.abstract

62. Pelikan J. The Vindication of Tradition. en. Yale University Press, 1984 Jan

63. Jackson K, Booeshaghi AS, Galvez-Merchan A, Moses L, and Pachter L. concordexR: Calculate the concordex coefficient. 2023. DOI: 10.18129/B9.bioc.concordexR. Available from: https://bioconductor.org/packages/concordexR