The changing mouse embryo transcriptome at whole tissue and single-cell resolution

Thesis by

Peng He

In Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2019

(Defended Oct 29, 2018)

Peng He
ORCID: 0000-0002-2457-3554

To my mother, Suxue San, and my father, Gangyue He, for their support and trust, even without understanding exactly what I'm doing.

ACKNOWLEDGEMENTS

In my past attempts at a variety of projects at Caltech, I was fortunate enough to have colleagues who mentored, helped, and inspired me. Without them, I would not be able to reach where I am now.

First, I would like to thank my advisor, Barbara Wold, for her consistent patience and responsibility for my academic training. I especially feel grateful that she was able to discuss with me the details on research design and communication skills.

I would also like to thank my committee members for their support and guidance that complemented my advisor's input.

I am grateful for the mentorship I received from senior colleagues in the Wold lab. Brian Williams has been a major counselor for experimental choreography and literature mining. He has also significantly contributed to my projects. Daniel Kim introduced me into modern pluripotent stem cell biology and helped me initiate my work on functional studies of long-noncoding RNA in embryonic stem cells; Georgi Marinov shared with me his expertise in computational biology and provided a base set of scripts on top of which I built my own toolkit for bioinformatic analysis; Sreeram Balasubramanian guided me through molecular cloning.

I would also like to thank our computational team, Henry Amrhein, Diane Trout, and Sean Upchurch, without whom I would have drowned in the ocean of software and hardware bugs that were emerging throughout my bioinformatics endeavor. Also appreciated are other lab members (Say-Tar Goh, Katherine Fisher-Aylor, Ken McCue,

Libera Berghella, Gilberto DeSalvo, and Gordon Dan) who kindly provided their independent inputs in discussions with me. Gordon Dan also designed the illustrations of mouse histogenesis that have been used in this thesis. I would specially thank my undergraduate research assistant, Won Jun Kim, for working together with me on CRISPR knock-out constructions days and nights consistently for two years.

In addition to my labmates, I would also like to thank Alexei Aravin, Long Cai and Mitchell Guttman who hosted my rotations in my first year, as well as other fellow researchers and collaborators on campus from whom I learned a lot: Guttman lab members (Amy Chow, Sofia Quinodoz, Mario Blanco, Patrick McDonel, and Alexander Shishkin), Xun Wang of the Rothenberg lab, Shuai Wang of the Shan lab, Thomson lab members (Sisi Chen and Jeff Park), and Shirley Pease and Igor Antoshechkin for their support in molecular and cell biology experiments. Also thanked are Xun Wang (again) of the Rothenberg lab, Baltimore lab members (Guideng Li and Yapeng Su), Han Wang of the Sternberg lab, Miao Cui of the Davidson lab, Bronner lab members (Marcos Simoes Costa and Can Li), Theodora Koromila of the Stathopoulos lab, Ke Ding of the Anderson lab member and Gradinaru lab members (Ken Chan and Ben Deverman) for introducing their puzzles to me to solve computationally.

I would also like to thank our ENCODE collaborators at other institutes: Len Pennacchio, Diane Dickel, Axel Visel, Bing Ren, Ross Hardison, Yu Zhang, Zhiping Weng, Richard Myers, Ali Mortazavi, and their teams. Our work collectively provided a comprehensive

map of the inputs, outputs, and transient states of the dynamic genome, part of which is going to be discussed in the first chapter of this thesis.

ABSTRACT

Mammalian histogenesis is a sophisticated process of coordinated changes of cellular composition governed by selective gene expression. This thesis focuses on the systematic application of modern RNA-seq methods to histogenesis processes in developing mouse embryos. Most of the work presented here is conducted as part of the ENCODE (ENCyclopedia Of DNA Elements) Project. Chapter 1 introduces the current advances of transcriptome studies on tissue development. Chapter 2 discusses a large-scale study on the whole-tissue transcriptome of 12 embryonic tissues at up to 8 timepoints and 5 additional perinatal tissues. Coherent themes of biological function and underlying regulatory mechanisms are revealed from the large-scale analysis. Chapter 3 presents a high-resolution single-cell RNA-seq study focused on the developing forelimb of the mouse embryo. This approach enables the assignment of differential genes to corresponding lineages and provides an even more accurate picture of RNA level patterns and regulatory modes. Finally, whole-tissue and single-cell methods are compared, contrasted, and integrated in Chapter 4 to extrapolate from the main discoveries of this thesis.

PUBLISHED CONTENT AND CONTRIBUTIONS

Peng He, Georgi Marinov, Diane Trout, Henry Amrhein, Len Pennacchio, Diane Dickel, Axel Visel, Bing Ren, Yu Zhang, Barbara Wold. (2018). Submitted

P. H. participated in the design of the project, data analysis and manuscript writing

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

NOMENCLATURE

CCA. Canonical correlation analysis

cCRE. Candidate cis-regulatory element

CNS. Central nervous system

CPM. Counts per million

DHS. DNase hypersensitive site

EMP. erythro-myeloid progenitors

ENCODE. Encyclopedia of DNA Elements

FPKM. Fragments per kilobase of transcript per million mapped reads

GO. Gene Ontology

PCA. Principal component analysis

RPKM. Reads per kilobase of exon model per million mapped reads

RPM. Reads per million mapped reads

scRNA-seq. Single-cell RNA-seq

TF. Transcription factor

TPM. Transcripts per million

t-SNE. t-Distributed Stochastic Neighbor Embedding

TSS. Transcription start site

UMI. Unique molecular identifier

UTR. Untranslated region

Chapter 1

CURRENT ADVANCES IN TRANSCRIPTOME STUDIES

Abstract

In the past decade, second-generation sequencing has spurred a wave of more comprehensive and quantitative transcriptome assays. New technologies enabled accurate transcript counting with new isoform discovery, pushed sensitivity limitation towards single-cell standard, increased throughput to up to tens of thousands of cells per experiment, and are moving towards in situ measurement with comparable sensitivity and resolution. In this chapter, I am going to introduce the current advances in the field of transcriptomics with a focus on the application on development.

1.1 Introduction

The mammalian genome is estimated to contain more than 40,000 genes, responsible for the full variety of structures and functions in every cell of a tissue throughout the life of the animal. Each cell selectively expresses a subset of genes that defines its identity and function. To read out gene expression profiles, polyadenylated RNAs have been widely used as indicators of genome output. Thanks to their well-defined and coherent biochemical property, massive parallel quantification had been achieved as early as 1995 using microarray probe hybridization[1]. Although this method provided easy measurement for known transcripts, it was bounded by prior knowledge, could not discover novel sequences, and had trouble detecting splice isoforms and other more subtle sequence variations of RNAs.

Sequencing-based methods have been independently developed to characterize transcripts without prior knowledge[2,3,4]. A breakthrough was achieved that was especially important for the complex transcriptomes of large mammalian genomes[5,6]. RNA-seq, which is based on sequencing cDNA, provided the ability to discover new transcripts, and higher sensitivity to detect rare but meaningful transcripts from low amounts of input material. This triggered a revolution in transcriptome quantification[7] and helped prompt the trend to convert "omics" and screening studies into a count-by-sequencing problem[8] that could be routinely solved across virtually all organisms.

The technology boundary was further pushed to achieve single-cell resolution. This required optimizations in library construction[9,10] and manipulation of microfluidics[11,12,13] to enable increasingly large-scale single-cell readout. High-throughput single-cell measurements provide the power to deconvolute tissue complexity. A subset of these methods can also take on more demanding questions about the differences among similar cells of a given "type". With the ever-advancing technologies, numerous studies have

been carried out to understand tissue development at whole-tissue[14,15,16,17] and single-cell levels[18,19,20,21,22,23,24].

I am going to briefly introduce the development and most pertinent characteristics of current transcriptome methods and their application to tissue development, the subject of this thesis.

1.2 The molecular biology and bioinformatics of mRNA quantification

mRNA is a good proxy for gene expression due to the ease to transfer its sequence information into double-stranded DNAs for second-generation sequencing and its reasonable correlation with protein-abundance[25]. However, the majority of intracellular RNAs are ribosomal RNAs (rRNAs) that do not serve as templates for translation. To enrich signals, microarray-based methods cherry-picked known transcripts one by one through probe hybridization. Sequencing-based methods mainly use two different types of strategies to focus the sequencing resources on the RNAs of interest (typically aiming for mature mRNAs and lncRNAs). The first strategy is to remove ribosomal RNAs by probe hybridization before downstream library construction[26]. This method (variously called total RNA-seq[26] or Ribo-minus RNA-seq[27]) not only captures mature mRNAs but also other non-polyA RNAs such as histone mRNAs and immature RNAs[28]. Meanwhile, short RNAs such as snoRNAs, miRNAs, and tRNAs can also remain and still consume the majority of the reads. A pre-selection based on the sizes of intact total RNAs[29] is usually required before library construction to combat this problem. The relative success of these biochemical enrichments varies with protocol details and execution and is not equally reproducible for all practitioners and for all starting materials. These issues bring

extra technical factors that influence variation and subsequent quantifications. The second approach is to positively select polyA RNA either by Oligo(dT) selection[5] or Oligo(dT) priming[30]. This method focuses on polyA RNA only and has been the more widely used choice for mRNA sequencing. In the following part of this chapter, the word "RNA-seq" will only refer to the latter strategy.

RNA-seq makes use of second-generation sequencing methods to dilute cDNA libraries and count the sequencing reads derived that match the sequences of each gene. Unlike DNA mapping, RNAs may contain splice sites so that a read can "jump" over intron regions, adding to the difficulty of quantification. Methods that can map junction-spanning reads[31,32] and those that compare mapped coordinates with transcript annotations[33,34] have been developed and improved[35,36,37]. More recently, alignment-free methods directly look for reads that can match known transcript sequences without mapping to the whole genome and guessing how transcripts are spliced[38,39]. This recent kind of fast methods essentially works as "artificial microarrays" as they only look for counts of known transcripts, except that their "probes" have much lower cross-hybridization confusion than to the microarray's physical probes. On the other hand, for unknown transcript and isoform discovery, RNA-seq data can be fruitfully used to detect and reconstruct novel transcript structures to add to existing annotations for quantification of new isoforms[40,41,42].

Contemporary RNA-seq is a stochastic sampling process that does not give absolute copy numbers of transcripts, although well-executed spike-in controls can help with

calibration[43]. Normalization is required to correct for different sequencing depth when comparing between samples and to correct for transcript lengths when comparing between transcripts. Therefore, read counts are usually divided by sequencing depth and transcript length and adjusted in orders of magnitude to yield RPKM (reads per kilobase of exon model per million mapped reads) or FPKM (fragments per kilobase of transcript per million mapped reads) values[5]. However, in the past years, other metrics were proposed such as RPM (reads per million mapped reads) /CPM (counts per million) [44] and TPM (transcripts per million) [45] aimed at using units that can sum to a constant within every sample to try to achieve more faithful estimation of real physical parameters, which may not be as good as they claimed.

It has been a long journey to figure out what the most physiologically meaningful metric is for RNA. RT-qPCR methods used putative "house-keeping" genes as controls to get relative abundance[46,47]. However, as I learn more and survey a wider range of cell types and cell states, the notion of a very stable and reliable RNA level per cell has broken down. To my knowledge now, there is no house-keeping gene or gene set whose transcript abundance may be constant[48], and this is even more true at single-cell level. This means that historic studies need to be viewed with the knowledge that the amount of a housekeeping standard transcript is really a sampled distribution that is probably quite good for comparing multiple samples of a given tissue or cell type, but nevertheless is likely to have systematic issues across different tissues, cell types, and genotypes. Unlike qPCR, RNA-seq normalizes its signal against a whole-cell sum that is likely to be more robust. RPM/CPM divides the read count by sequencing depth and is thus proportional to

the mass of each transcript among **different transcripts in the same sample,** while FPKM and TPM consider transcript length and therefore are both proportional to the molar concentration and copy number of each transcript among **different transcripts in the same sample**. On the other hand, when comparing the **same transcript among different samples**, which is what RNA-seq is mostly aimed to do, CPM and FPKM for the same transcript would go hand-in-hand (if CPM of a transcript in Sample A is higher than that in Sample B, FPKM in A would also be higher than that of Sample B), both reflecting relative mass of that transcript, while TPM measures relative molar concentration. Therefore, comparing these units across samples depends on how a biologist defines the "fraction" (of mass or counts). It is not yet sure what is a more meaningful parameter for gene expression measurement as a predictor of protein output, since the mass (and thus length) of each transcript may be related to how much translational machinery it can take if longer transcripts can bind and use more ribosomes. But we also know that rates of ribosome loading and translational pauses (or lack thereof) vary for certain messages and entire systems, meaning that there are exceptions to overall length/loading generalization. On the other hand, relative molar concentration would be a better estimator if we believe all the transcripts share more or less the same translation rate per copy. We cannot know for sure which proposal is the ideal normalization method merely based on computation legitimacy without knowing how cells biochemically

interpret “concentration”. One experimental effort to better estimate gene expression using transcriptomic approach is to profile ribosomes on transcripts[49].

Another important task for RNA quantification, and the most central one presently in the areas of developmental biology, is differential RNA level analysis. To compare gene expressions between two samples, multiple computational methods have been developed, and negative binomial model-based algorithms proved to be able to capture read-count over dispersion compared to Poisson models[50,51]. However, due to the fact that biologically meaningful RNA levels often differ by orders of magnitudes and that their estimated abundance follows log-normal distribution, it has also become a common practice in the wider community to use t-test and ANOVA on log-transformed FPKM or TPM values.

Log-transformed FPKM values have also been frequently used when multiple samples get compared and grouped. Principal component analysis[52] (PCA) has been widely used to find “components” of co-expressed modules, classify samples, and reduce dimensionality to understand the main features of the data collection. Similarly used methods also include clustering analyses[53,54,55] and numerous other methods[56,57,58,59] for data mining among large-scale datasets. A more recent method[60,61] has been combining PCA and CCA (canonical correlation analysis)[62] to explain data structures based on metadata and was able to sensitively find new insights.

Applying RNA-seq quantification approaches to multiple samples have already showed off its power to quantitatively and comprehensively define the development of individual tissues. By comparing different tissues at different developmental stages, tissue-specific, age-dependent, and ubiquitous gene expression patterns have been revealed[17,63,64,65,66,67]. These studies formed some initial pictures of the structure of mouse transcriptome. Pervouchine et al. discovered that up to 40% of evolutionarily conserved genes are relatively constant across cell types in the mouse genome[17]. Lin et al.[66] did a joint RNA-seq analysis of human and mouse tissues and discovered that gene expression profiles are more dominated by tissue type instead of organism identity. Among these dominant tissue-specific signatures, those of testes, brain, liver, muscle, and kidney were the most prominent in terms of the number of tissue-specific genes. A similar conclusion was derived by Söllner et al.[68] independently. Additionally, in the latter study, thymus and pancreas turned out to be the top two outliers separated from other tissues and their PCA plot showed that the brain and liver mark the two extremes of the top principle component among mouse tissues in their collection. These discoveries show that the hierarchical mouse transcriptome contains dominant structures that can be reproducibly identified, as well as minor features that may be under-emphasized in different degrees due to different experimental and computational setup.

However, there has not been a large-scale systematic study of multiple samples spanning multiple time points assayed in the same way in embryonic mice to gain insights on gene regulation. To achieve this, on one hand, a larger-scale study at high depth elevates sensitivity and captures fine features of transcriptome for downstream studies; on the

other hand, emerging single-cell technology brings new possibilities to mine more deeply into the complexity of a given tissue development.

1.3 Single-cell resolution

Measuring the single-cell transcriptome has long been a goal for developmental biologists[69,70]. Improvement in RNA-seq library construction made it possible to measure a large number of transcripts in the same cell with reasonable sensitivity[9,10]. It was discovered that gene expression variations among single cells is largely attributed to intrinsic biological variations rather than merely technical stochasticity[71]. Integration of mRNA-seq library construction with FACS sorting[72], microfluidics[11,13,12], pool-split barcoding[73], and microwell handling[74,75] further increased the throughput of single-cell RNA-seq assays. These methods come in two major categories: full-length profiling or 3' tagging[76]. Full-length profiling methods such as MARS-seq[72] and C1 Fluidigm SMART-seq assays[11] provide good coverage over longer transcripts and thus have higher isoform resolution. Gene expression levels are usually measured in FPKM or TPM units. 3' tagging methods such as inDrop[13], Drop-seq[12], Seq-well[74], and Microwell-Seq[75] use 3' terminal sequence of transcripts as an anchor point for mRNA capture and gene identification. Although this method cannot resolve splice isoforms, it is easy to perform on beads that can greatly scale up parallel reactions and increase sample size at the sacrifice of per-cell quality. This class of methods usually quantifies RNA in digital counts.

Unlike traditional "whole-tissue" RNA-seq methods, single-cell RNA-seq meets the challenges, in variable degrees, of low input material and low signal-to-noise ratio. One

of the sources of noise in classical RNA-seq comes from differential PCR amplification when multiple PCR cycles are involved, which are part of the standard methods. To solve this problem, two major approaches have been developed. One way is to trace the original cDNA fragment by introducing a short random polymer DNA tag called unique molecular identifier (UMI)[77] during reverse transcription. This method keeps an ID of each individual starter cDNA fragment and helps collapsing PCR duplicates that have the same tag. UMI-based methods have been a routine and generally effective part of droplet-based 3' tagging methods except for occasional PCR-based errors that can produce nucleotide substitution and indels in the UMIs[78]. Another approach is to eliminate PCR-based exponential amplification steps. One such example is CEL-seq[79,80], which linearly amplifies cDNA sequences to minimize over-amplification noise and has demonstrated its power to build a cell atlas for early *Caenorhabditis elegans* embryo development.

As single-cell RNA-seq is gradually improved and democratized, development biology has been experiencing more cell type-level discoveries. So far, single-cell RNA-seq has been used most to discover novel cell types[23,24], compare and contrast cell compositions[75,81], and increasingly, to track cell lineage during migration[82], differentiation[83], and regeneration[19] in embryos and tissues.

The ability to infer cell lineage maps based on large numbers of single-cell snapshots of gene expression is especially intriguing for development biologists. Taking advantage of the fact that cell types can be robustly defined by a shallow sequencing[84] and that a large

sampling experiment usually captures cells at "intermediate" states in a lineage, multiple computational approaches have been developed to infer lineage relationships between cells based on their pair-wise similarity. A first group of methods such as Monocle[85,86] and Waterfall[87] reduces transcriptome dimensionality and then constructs a minimum spanning tree to define the backbone of the trajectory tree before filling in details with individual cells. A second group of methods such as Wanderlust[88] and Wishbone[89] uses k-nearest neighbor (knn) graphs to link similar cells together and then smoothen out a shortest path with or without bifurcation allowed. Unlike these methods, a more recent algorithm goes on a different path. This algorithm called RNA velocity[90] implements the idea that unspliced pre-mRNAs foreshadow cell fate transition[91] utilizing the fact that these pre-mRNA sequences are present in the single-cell RNA-seq[10] due to mis-priming. Therefore, based on these hitchhiked pre-mRNAs, RNA velocity algorithm constructs a hypothetical "extrapolated state" linked to each of the single cells. It not only predicts the direction a cell is moving in but also estimates how fast the transition will be. It would be interesting to compare these lineage inference algorithms among themselves and against experimental lineage tracing methods[92,93] for evaluation.

In terms of its application to embryo development, single-cell RNA-seq has so far been used to construct large-scale cell-type maps in multiple tissues by at least two independent teams[81,75]. Both of them were able to find distinct cell types, common or biased, between different organs. Single-cell RNA-seq has also been coupled with a lineage-tracing method to construct a lineage map of the zebrafish embryo[83]. Interestingly, this study discovered that a computationally inferred "lineage map" based

on cell states is not necessarily overlapping with the real lineage history[83], which raises the question about what a "lineage map" really means. The power of single-cell RNA-seq to define cell types has also been used to infer regulatory codes. Buenrostro et al.[94] integrated single-cell RNA-seq and single-cell ATAC-seq of the same sample and were able to match their subpopulations. By matching gene expression profiles with chromatin accessibility, they were able to infer transcription factors that potentially regulate cell state maintenance and transition. As more and more single-cell characterization studies are done, the structure of cell type-specific gene expression and regulation would become increasingly clear.

1.4 New directions

Transcriptome studies have evolved rapidly in the past decade, from whole-tissue to single-cell and from low-throughput to high-throughput. Their evolution is still going on.

Tissue development is a coordinated process where cells interact with each other in a limited space with carefully tuned distribution of chemical cues and mechanical parameters. Spatial information that RNA-seq can acquire is usually limited by the minimal dissection size of a tissue. Although laser-assisted microdissection techniques may accurately extract samples or cells at a defined spatial coordinate on a tissue[95,96], they cannot be done in both a precise and high-throughput way. Therefore, multiple teams have been developing high-throughput versions of single-cell RNA-seq that maintain spatial information. One established method is called "spatial transcriptomics" that aligns an array of barcoded reverse transcription primers against histological sections to record two-dimensional spatial coordinates followed by linear amplification and

library construction[97]. This method achieves a resolution at 100um (10-20 cells) but can be further deconvoluted by single-cell RNA-seq from a comparable sample[98]. New ways of introducing spatial barcodes before downstream RNA sequencing may give rise to better spatial transcriptomic methods. Recent prototypes of "probe hybridization"-based methods such as seqFISH[99] and MERFISH[100] show that spatial patterns of up to 1000 RNA species can be visualized with single-cell resolution. Intron seqFISH further demonstrated its polyA-independent advantage by labeling more than 10,000 nascent transcripts to study transcription dynamics[101]. Another hybridization-based method was recently introduced as "STARmap", which used in situ sequencing to amplify signals and claimed to be able to read out spatial transcriptome of 3D intact tissues[102]. These imaging methods may hold great promise for single-cell spatial transcriptomics and may be integrated with tissue clearing[103,104], super resolution microscopy[105,106] and expansion microscopy[107] for additional benefits. However, the methods mentioned above all depend on careful experimental setup, such as the choice of probes to label specific marker genes. Further efforts are still needed to transparentize, simplify, and standardize detailed protocols of existing methods for broader applications.

Although spatial transcriptome can be probed by hybridization, complex isoform specificity is hardly achieved, nor can short-read RNA-seq. But development of long-read sequencing methods such as nanopore[108] and PacBio[109] makes it possible to quantify transcripts with clear isoform specificity combined with full-length RNA-seq and single-cell RNA-seq methods. Long-read methods may also reveal allelic information when applied to human samples.

Transcriptomic measurements can also be integrated with other assays in the same cell to bona fide link transcription to other aspects of cellular physiology. Efforts are been made for single-cell RNA-seq to incorporate information of genomic DNA[110,111,112,113], chromatin accessibility[114], protein abundance[115,116,117,118], perturbation[119,120,121,122,83], and DNA methylation[123]. More multi-omics single-cell approaches will be or are already being developed.

With existing technology getting stabilized and new approaches emerging, large-scale transcriptome data integration and cell type taxonomy construction were made possible. Projects like the Human Cell Atlas[124] have thus been initiated, benefiting from existing technology and in turn spurring new technology. Through collaborative efforts and integrative analyses, transcriptomics studies in tissue development would rise to a new level.

Chapter 2

GLOBAL ANALYSIS OF WHOLE-TISSUE MOUSE EMBRYOGENESIS TIMECOURSE

Abstract

Mammalian development is driven by selective gene expression and cell population coordination. To understand the dynamic modules of the gene regulatory network governing histogenesis during embryo development, I present here a collection of embryonic mouse transcriptome using the whole-tissue RNA sequencing method (RNA-seq). Transcript diversity and dynamics were explored in 12 tissues from E10.5 to birth. Five additional tissues at P0 were added to facilitate comparative analysis. Overall, among the 24,832 genes expressed, 63% were strongly differential, revealing strong temporal and tissue-specific components with underlying cell-proliferation, characteristic function, specialized function, and body-axis signatures. These differential genes formed 34 major co-expression clusters with known or novel functions and putative promoter regulatory codes. Interestingly, the remaining ubiquitous genes were associated with potential post-transcriptional regulations in addition to their promoter signature. This study not only presents a broad view of mouse embryonic development as part of the ENCODE project but also provides a processed data framework for users to integrate their own input.

2.1 Introduction

Hierarchical transcription programs unfold to regulate mammalian development through a dynamic transcriptome[125]. Although modern methods are able to read out the whole

transcriptome with reasonable sensitivity, existing studies[126,17,127] were done only for selected cell lineages and subsets of tissues in the developing mouse with minimal temporal resolution. Efforts to assemble multiple sources of data and analyze a broad developmental atlas is precluded by different methodological details between studies. Here I report a systematic matrix of polyA-RNA-seq data that includes 12 tissues from E10.5 to P0 (Figure 2.1) and covers much of organogenesis and histogenesis. By comparing the transcriptome profiles along time axis and across tissue identity space, finer details on gene-regulatory network structures and mechanisms can be revealed.

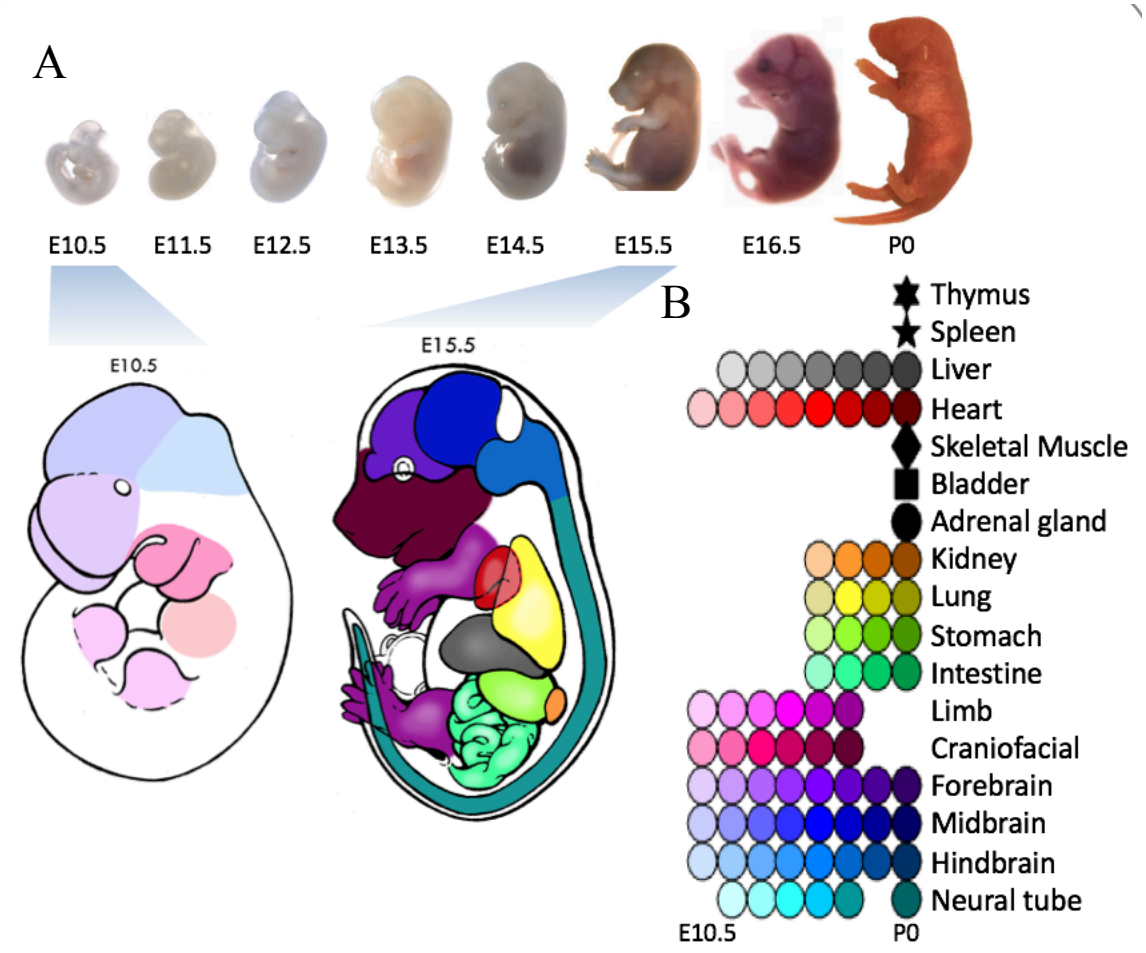


Figure 2.1: Tissues collected for whole-tissue RNA-seq assay. (A) Mouse embryo stages selected for RNA-seq. (B) Embryonic tissues assayed at E10.5 and E15.5 as examples to

be illustrated using the color key labeled on the right. Craniofacial: craniofacial prominence.

2.2 Results

2.2.1 Overall structure of the high dimensional transcriptome

As expected, the vast majority of protein-coding genes (84%) were significantly expressed (Figure 2.2), compared to a smaller fraction of lincRNAs (44%) and a minority of "other" annotated genes (26%) including many pseudogene, antisense, and small RNAs. This is probably due to the fact that protein-coding genes are better annotated while non-coding RNAs, especially those in "other" categories, are more likely to be mingled with pseudogenes and repetitive sequences. Current sequencing technology usually cannot perfectly quantify pseudogenes and repetitive sequences due to their lack of unique sequence information and the size limit of sequencing read length. lincRNAs stand in the middle of the spectrum, probably because they are longer and better characterized than, but not cleanly separated from, the "other" category. Additionally,

A B

lincRNAs and pseudogenes can be dynamically converted to each other during evolution[128], further blurring their boundary.

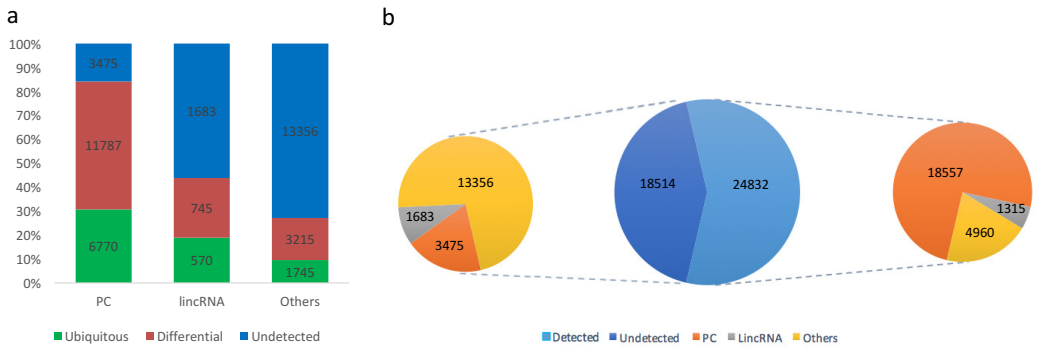


Figure 2.2: Number of genes detected by class. (A) percentages of ubiquitous, differential, and undetected genes in each of the three categories: PC (Protein-coding genes), lincRNA (long intergenic noncoding RNA), and others. (B) Percentages of PC, lincRNA, and other genes in detected and undetected categories.

The detected transcriptome can be further divided into two parts: ubiquitous and differential. The highly differential character of the transcriptome comes from 15,644 genes that differed by over 10-fold between at least two samples in the matrix, while

9085 genes were more uniformly expressed, forming the foundation of ubiquitous housekeeping activities and structures (Figure 2.3).

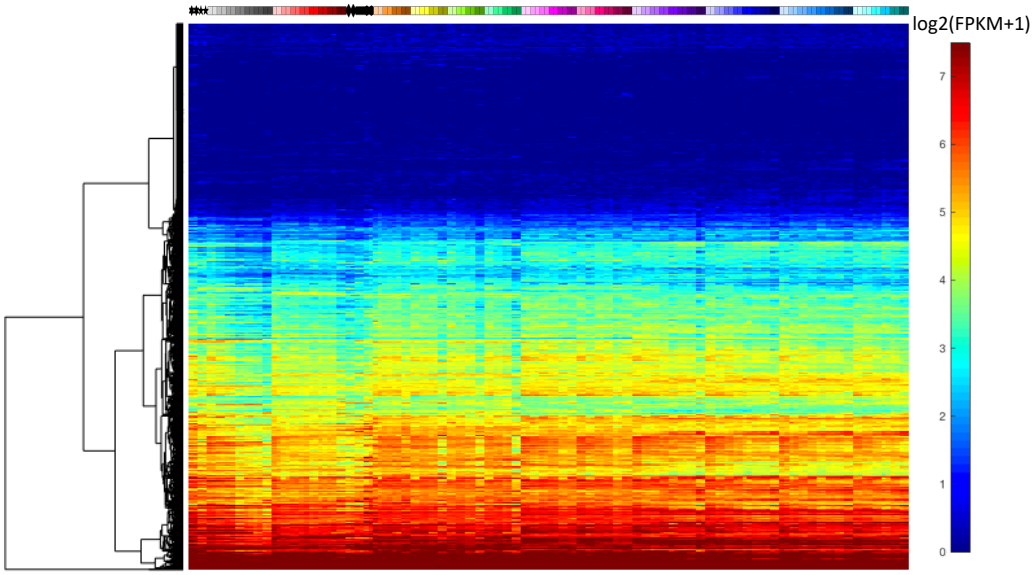


Figure 2.3: Hierarchical clustering of ubiquitous genes. Tissue identities and stages are labeled at the top using the color codes from Figure 2.1.

To understand the main themes of the transcriptome, I conducted Principle Component Analysis (PCA) for these whole-tissue data. Overarching organizational themes of this mouse developmental transcriptome are tissue identity and developmental time, visualized as color hue and intensity shown in Figure 2.4A. The top three principle components separated tissues into roughly three domains: hematopoietic (liver, spleen, and thymus) domain, neurogenic (all three brain regions and neural tube) domain, and a third domain with all the remaining tissues in it. These domains were mainly separated by PC1 and PC2, both of which were strongly associated with hematopoiesis and blood

component based on Gene Ontology (GO) analysis of top 100 loading genes (Table 2.1). Notably, the negative top-loading genes of PC1 were enriched with those associated with blood component, while the positive top-loading genes of PC1 yielded neurogenesis terms (Table 2.1). This indicates that the contrast between hematopoiesis and neurogenesis is the most outstanding feature in the differential transcriptome, confirmed by an independent t-SNE projection (Figure 2.4B). Interestingly, multiple tissues emanate from a "hub" and form "branches" (labeled with arrows in Figure 2.4A) and descend toward the negative direction of PC3 (to be discussed later in this chapter) with decreased similarity to each other, representing a gradual enhancement of their unique signatures.

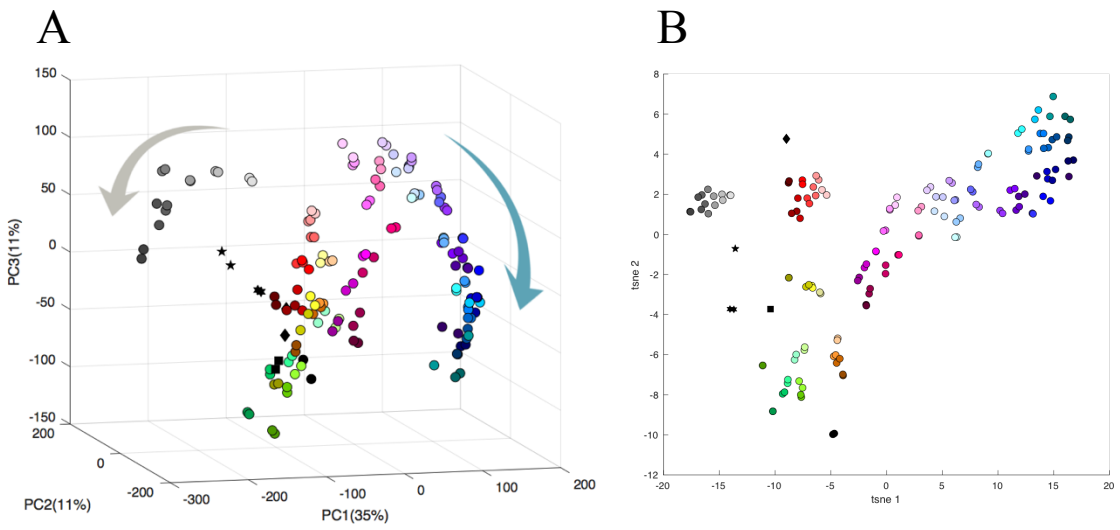


Figure 2.4: Whole-tissue transcriptome viewed at reduced dimensionality. Each tissue is labeled using the color code in Figure 2.1. The dimensions are respectively (A) the first three Principal Components from PCA and (B) the two axes of t-SNE two-dimensional plane.

PC	Negative loading	P-value	Positive loading	P-value
1	Blood microparticle	2e-22	Neuron part	2e-22
2	Embryonic morphogenesis	4e-22	Blood microparticle	3e-35
3	Extracellular region part	3e-20	Mitotic cell cycle process	3e-13

4	Digestive system process	1e-10	Contractile fiber part	1e-40
5	Intestinal epithelial cell differentiation	5e-9	Skeletal muscle contraction	3e-12
6	Muscle system process	3e-21	Apical plasma membrane	6e-14
7	Neuron fate commitment	3e-11	A/P patern specification	2e-33
8	Embryonic limb morphogenesis	1e-11	Hemoglobin complex	4e-14
9	A/P pattern specification	7e-11	Lung development	3e-8
10	Cerebral cortex neuron differentiation	1e-15	A/P pattern specification	5e-11
11	A/P pattern specification	5e-37	Neuron differentiation	7e-11
12	Urea cycle	1e-5	Cornified envelope	1e-11
13	Embryonic forelimb morphogenesis	8e-9	NA	>1e-5
14	Striated muscle contraction	7e-7	Alpha-amylase activity	2e-9
15	Embryonic morphogenesis	1e-11	A/P pattern specification	7e-22
16	Contractile fiber part	8e-7	Monooxygenase activity	6e-20
17	Neuron differentiation	1e-10	Blood microparticle	2e-22
18	Alpha-amylase activity	2e-9	Intermediate filament	2e-9
19	Monooxygenase activity	9e-11	Regionalization	3e-25
20	Midbrain development	2e-12	A/P pattern specification	3e-17

Table 2.1: Representative GO terms for top-loading genes of each PC for whole-tissue RNA-seq analysis. Negative PC's are on the left panel and positive on right. Representative GO terms are listed in the corresponding boxes.

In order to better understand the dynamics and specificity of the differential genes, I performed hierarchical analysis, identifying altogether 34 major clusters of co-expressed genes (Figure 2.5). These clusters mostly showed enriched biological themes based on Gene Ontology analysis of their gene members (Figure 2.6). Notably, most of these clusters have a trend of increasing abundance of genes, and upward genes are less likely to be shared among multiple tissues compared to downward genes. This pattern favors the model that embryonic histogenesis is mainly driven by differential activation of genes instead of differential repression of already active genes.

For validation purposes, I compared PCA with hierarchical clustering results by mapping PC scores (Figure 2.5C) and loadings (Figure 2.5B) onto the gene-by-sample matrix. Indeed, the two largest clusters (Cluster 10 and Cluster 34) respectively load to the negative and positive ends of PC1, which aligns with the contrast between hematopoiesis

and neurogenesis. In fact, nearly 1/5 (~5000 genes) of the expressed transcriptome defines this axis, perhaps reflecting the elaborate branching cell lineages that produce, within this timeframe, exceptionally large repertoires of distinctive neuronal and hematopoietic cell types. Not surprisingly, their feature tissues show negative (thymus, spleen, and liver for Cluster 10) and positive (brain regions and neural tube for Cluster 34) scores for PC1. Other PCs and clusters also showed a high degree of relatedness, robustly reflecting the underlying modular nature of the transcriptome.

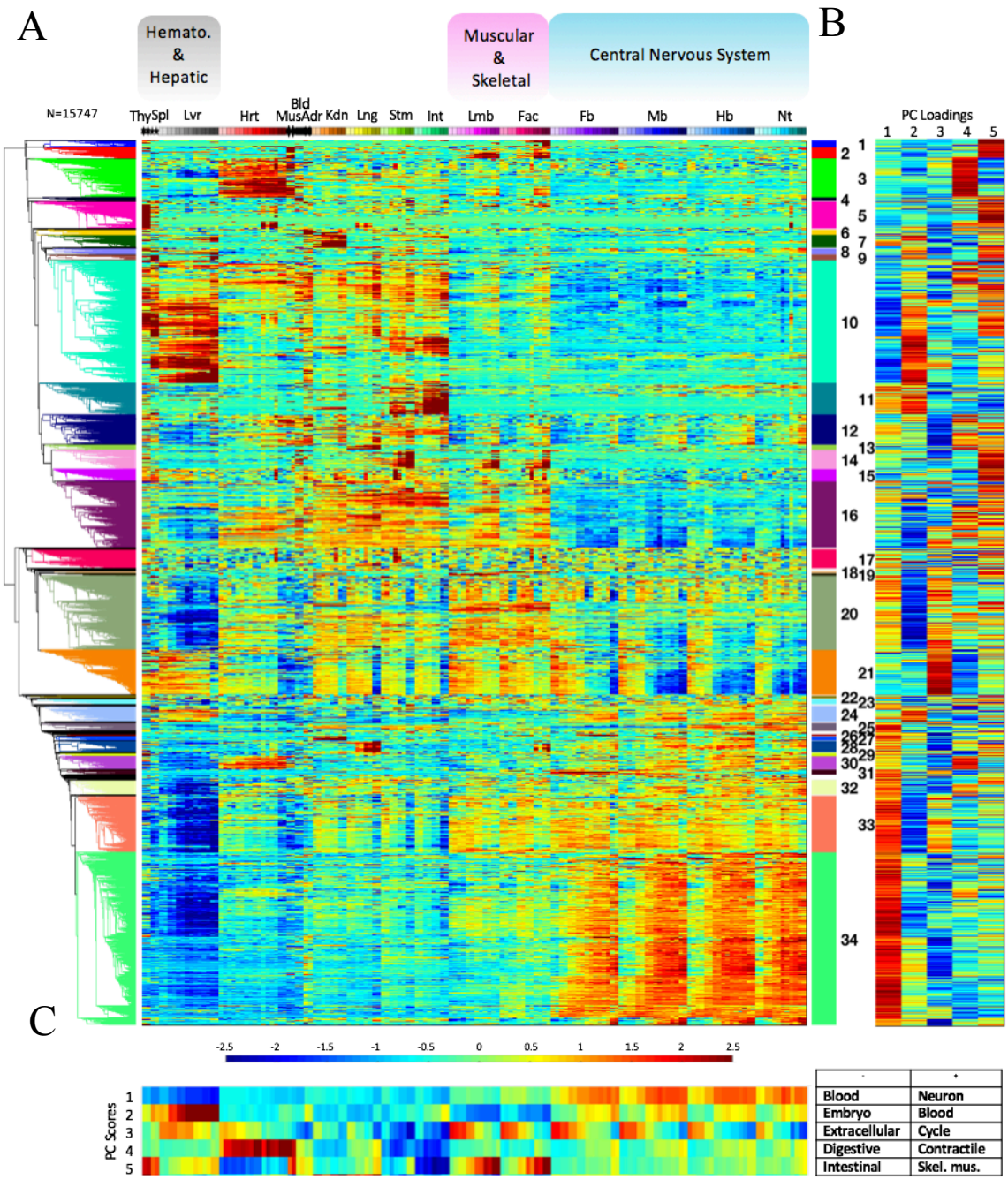


Figure 2.5: Hierarchical clustering of differential genes. (A) Gene expression profiles across tissues. Normalized log2 FPKM values are represented by the colormap (bottom), and tissue names are labeled on top: Thy, thymus; Spl, spleen; Lvr, liver; Hrt, heart; Mus,

skeletal muscle; Bld, bladder; Adr, adrenal gland; Kdn, kidney; Lng, lung; Stm, stomach; Int, intestine; Lmb, limb; Fac, craniofacial prominence; Fb, forebrain; Mb, midbrain; Hb, hindbrain; Nt, Neural tube. (B) PC loading coefficients of each gene in panel A. Normalized coefficients are shown using a heatmap with color scheme labeled in panel A. (C) PC scores of each sample in Panel A. Normalized coefficients are shown using a heatmap with color scheme labeled in panel A. Representative Gene Ontology terms are labeled on the right for negative and positive directions.

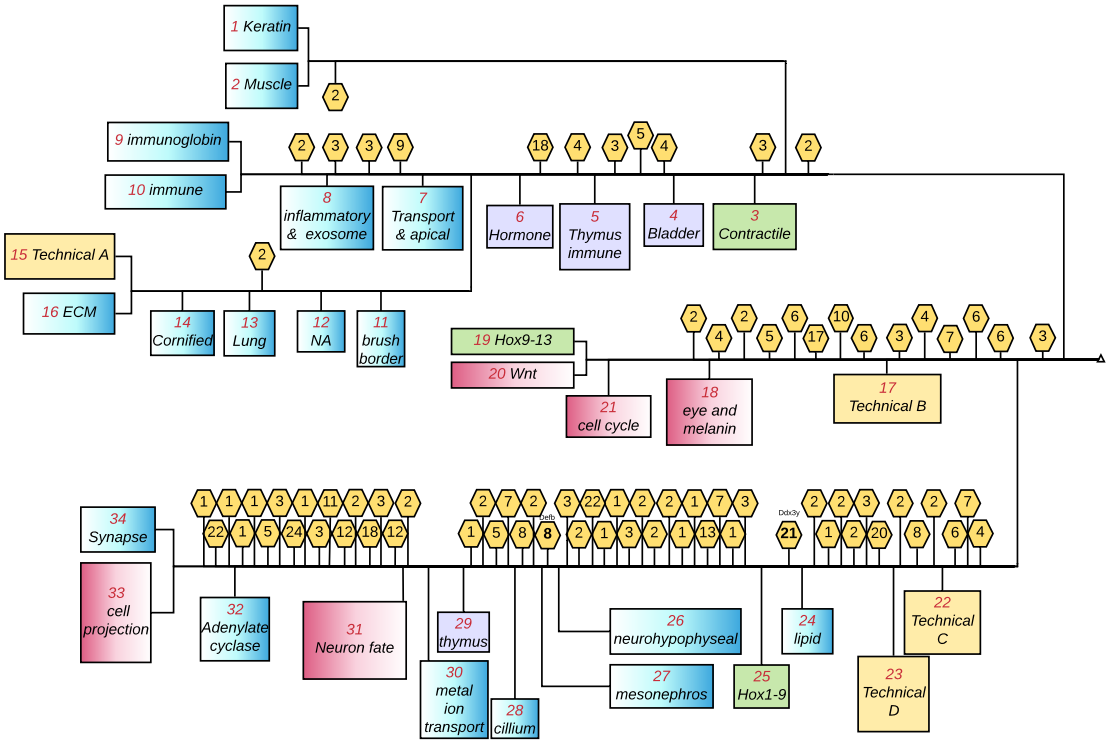


Figure 2.6: Schematic view of cluster relationships. Rectangle boxes represent the 34 major clusters, each of which contains more than 30 members. The text in each box labels the dominant features for each of these clusters based on Gene Ontology, tissue specificity, and gene class. The boxes in blue represent clusters with genes that mainly increase over time, while pink boxes are the opposite and green are constant; lavender indicates an unknown trend due to lack of time-course data; yellow boxes are genes without obvious dominant trends and are likely due to technical issues of the assays. The remaining dynamic genes fall into minor clusters with fewer than 30 genes per cluster, labeled as hexagons with the cluster size inside them.

To rule out the possibility that the global structure is highly biased towards the choice of tissue I have (there are way more brain samples than digestive tissue, for example), I performed similar analyses, but with brain tissues down-sampled (only midbrains kept). The results were barely changed: PC1 was still about hematopoiesis vs. neurogenesis, which was still contributed by a large number of genes enriched in their corresponding tissues (Figure 2.7). The fine structures revealed by individual clusters were also consistent with the original analysis, which is going to be discussed in the next section.

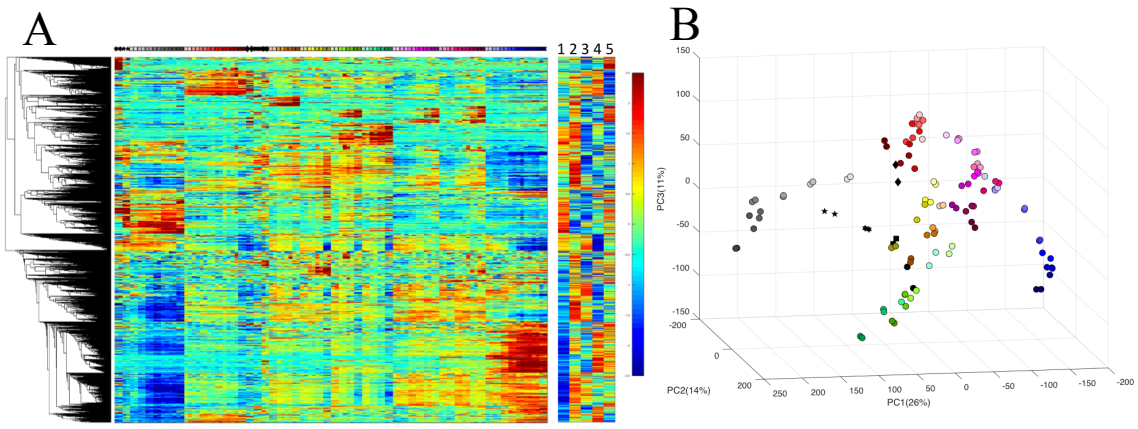


Figure 2.7: Validation of global transcriptomic structures by down-sampling. Results of hierarchical clustering (A) and PCA projection (B) of whole transcriptome with forebrains, hindbrains, and neural tubes removed are shown. Color codes follow those of Figure 2.1.

2.2.2 Diverse biological insights from co-expression clusters

Hierarchical clustering produced 34 major clusters with different tissue specificity and temporal patterns, summarized in Figure 2.6. Each of them has its own unique expression patterns, enriched Gene Ontology terms, and complexities.

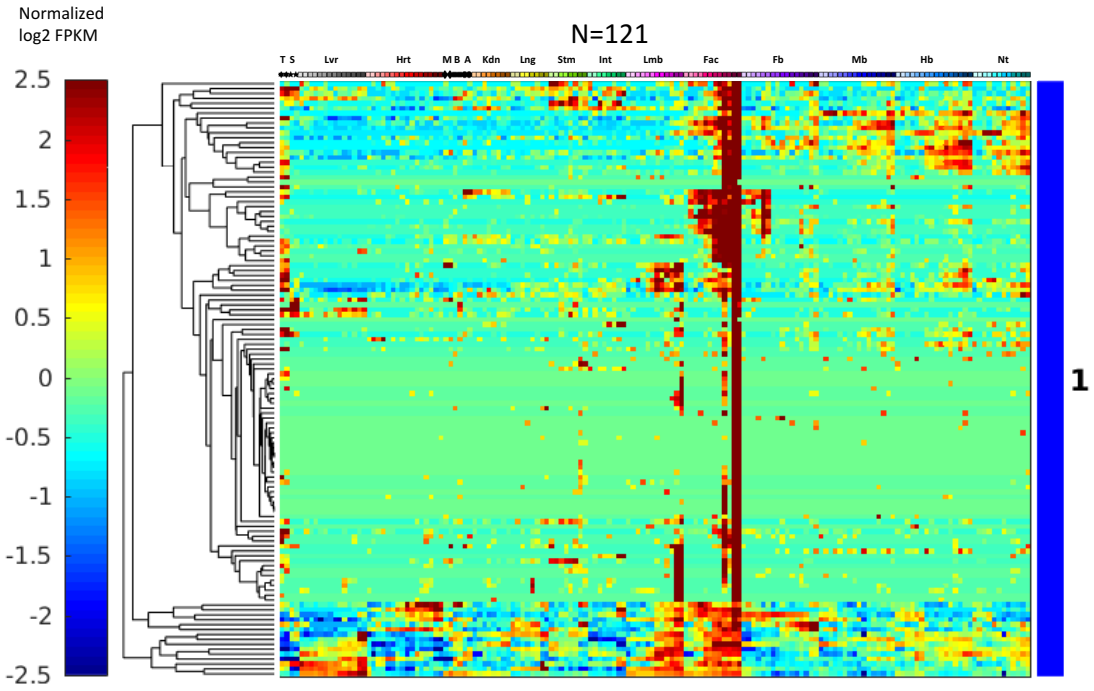


Figure 2.8: Cluster 1 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Cluster 1 (Figure 2.8) has prominent increasing transcript abundance in limb and craniofacial prominence. Over one third of the genes in this cluster are genes coding keratin and keratin-associated proteins, with top GO terms “intermediate filament” (p=3.7e-34) and “hair cycle” (p=2.4e-7), pointing to development of skin and hair. This

group of genes may be those specifically expressed in the skins of limb and craniofacial prominence that increase their proportion during development.

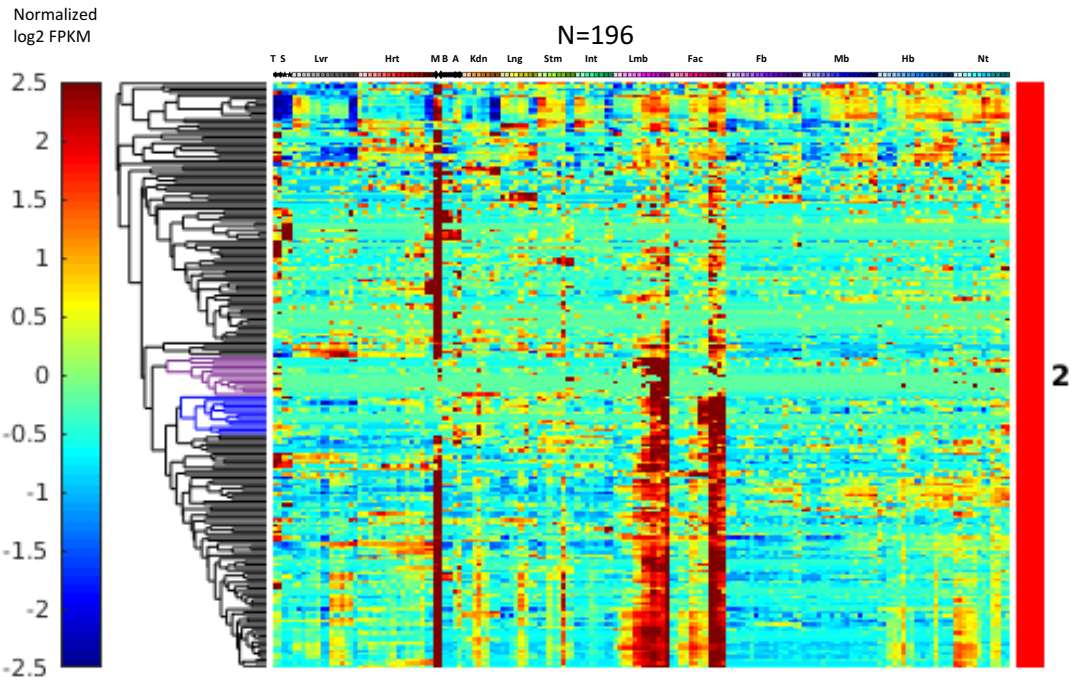


Figure 2.9: Cluster 2 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Cluster 2 (Figure 2.9) has prominent expression in skeletal muscle and an increasing trajectory in limb and craniofacial prominence. It contains multiple muscle regulators like Myod1 and Myog. Its top GO terms include “muscle system process” (4.5e-18) and “contractile fiber part” (9.9e-14). The increasing expression in limb and craniofacial prominence is likely due to differentiation of muscle precursors and to increasing relative muscle mass as a fraction of the total tissue. In addition to the dominant muscle-limb-face feature, there are two clades with different patterns that illustrate the informational

leverage that comes from the absence of expression in a more pure P0 dissected tissue (here muscle).

The clade of 13 genes labeled in blue has increasing expression in limb and craniofacial prominence but not in the P0 pure skeletal muscle sample. Among the 13 genes, five (Dcstamp, Mmp13, Bglap, Ifitm5 and Ibsp) are associated in prior work with osteogenesis. Another clade of 13 genes labeled in purple is biased for limb alone, and not cranioface. It includes four major urinary protein (MUP) genes at low but detectable abundance. The mouse genome has 21 annotated Mup genes in a 2Mb cluster on Chromosome 4. Although none have human orthologs, members of the family have known functions in mouse chemical communication and nutrient metabolism[129]. A recent study reported dramatic and unexpected upregulation of Mup1 in mouse embryos when Shox2[130], a transcription factor regulating proximal bone formation in limbs, is mutated. This raises the possibility that MUPs in this limb cluster play a role in limb development.

Technical issues were also identified: sporadic samples of adrenal gland, kidney, lung, stomach, hindbrain, and neural tube from this mouse embryo series show slight enrichments for genes from this cluster, implying variable minor tissue contamination during dissection.

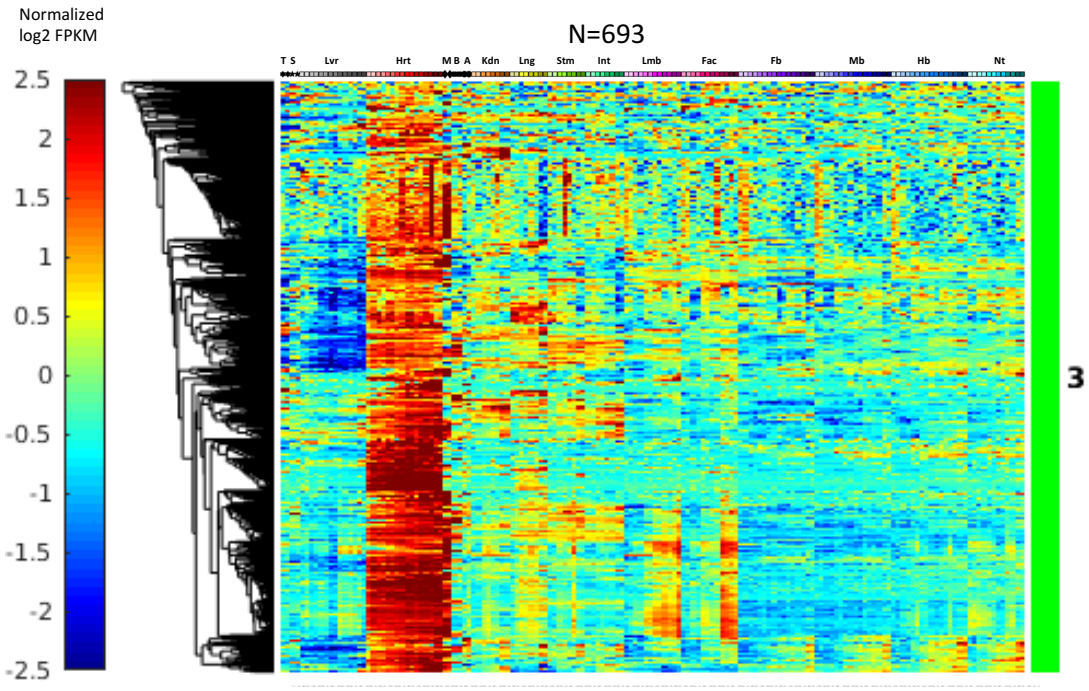


Figure 2.10: Cluster 3 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Most genes in Cluster 3 (Figure 2.10) have high and constant levels of expression in the heart. Roughly half of them also have substantial expression in skeletal muscle-containing samples, suggesting a program of genes shared by cardiac muscle and skeletal muscle. GO terms are mainly about muscle, including “contractile fiber part” (p=8.9e-47) and “regulation of heart contraction” (p=4.3e-21). Genes coding DMD, alpha-actins, Desmin, Leiomodins, myosin peptides, and troponins were found in this cluster. The clade in the upper half of the heatmap has narrow dark red bars, which indicate single replicate enrichment. This group of genes contains mostly pseudogenes. This may represent the limit of current sequencing methodology in quantifying the non-unique part of the transcriptome.

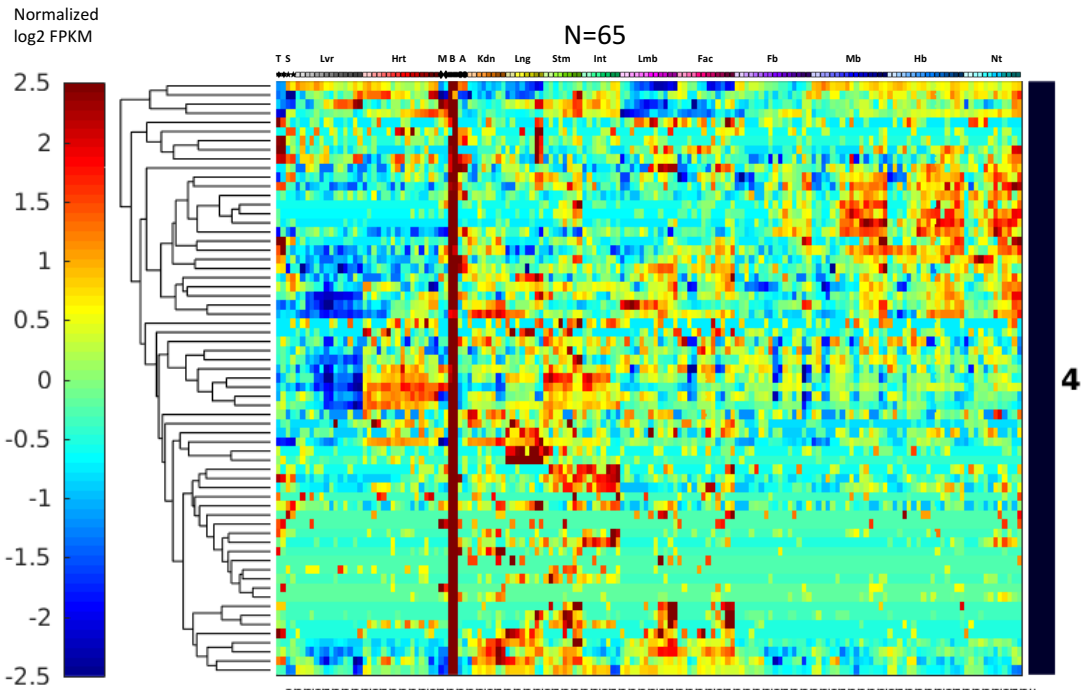


Figure 2.11: Cluster 4 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Genes in Cluster 4 (Figure 2.11) show differing degrees of bladder-specific expression, which may result from a bladder-specific cell type that has a unique transcriptome signature. GO analysis produced no terms. A possible reason is that the mouse bladder has not been extensively studied. Under-annotation may compromise the statistical power of GO analysis in this case.

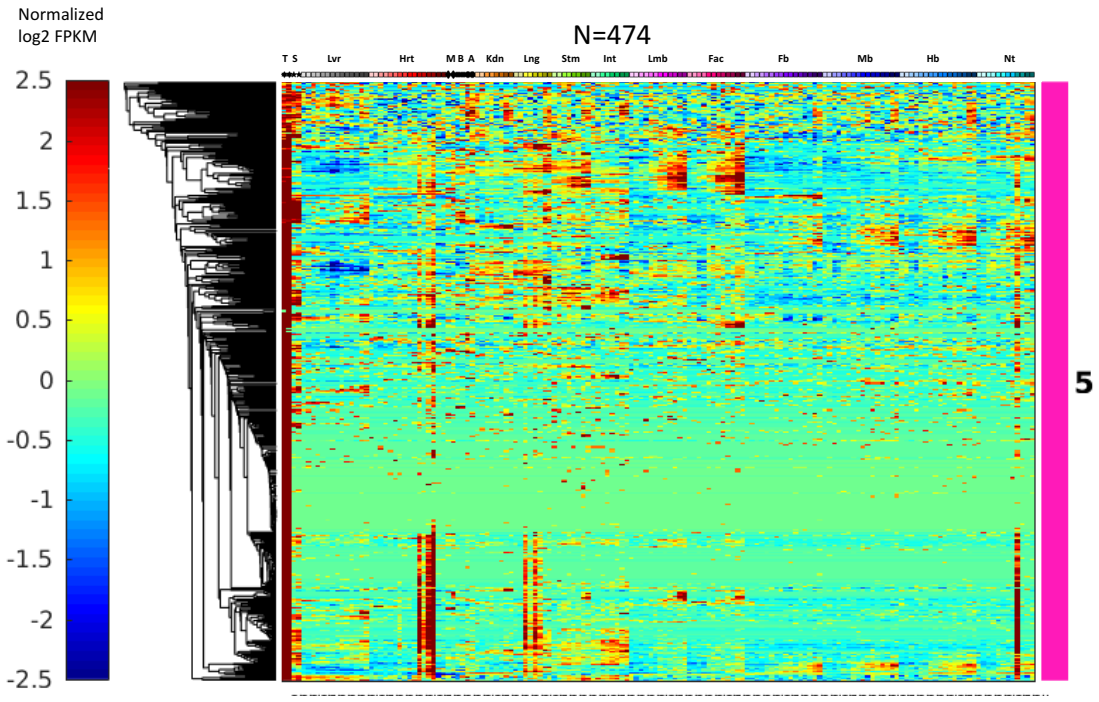


Figure 2.12: Cluster 5 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Genes in Cluster 5 (Figure 2.12) are very prominently expressed in the thymus, and most have minimal expression in other tissues. Highly expressed genes also have positive signals in several non-thymus samples (at least 4 heart samples, 2 lung samples, and one neural tube sample), with atypical irreproducibility between replicates. A candidate explanation is a batch-specific contamination of thymus-proximate tissues with thymus during dissection. While this kind of contamination does not greatly alter global QC scores, it is readily detectable in this clustering analysis. GO analysis revealed enrichment in later stage maturing immune components, especially T-cell terms. Top terms include "immune system process" (p=1.8e-18) and "regulation of T-cell activation" (p=3.0e-13). Roughly one quarter of the genes are T-cell receptor components (alpha chain, gamma

chain, and delta chain). Interestingly, the two recombinases Rag1 and Rag2 are also in this cluster, indicating a TCR VDJ theme for this cluster.

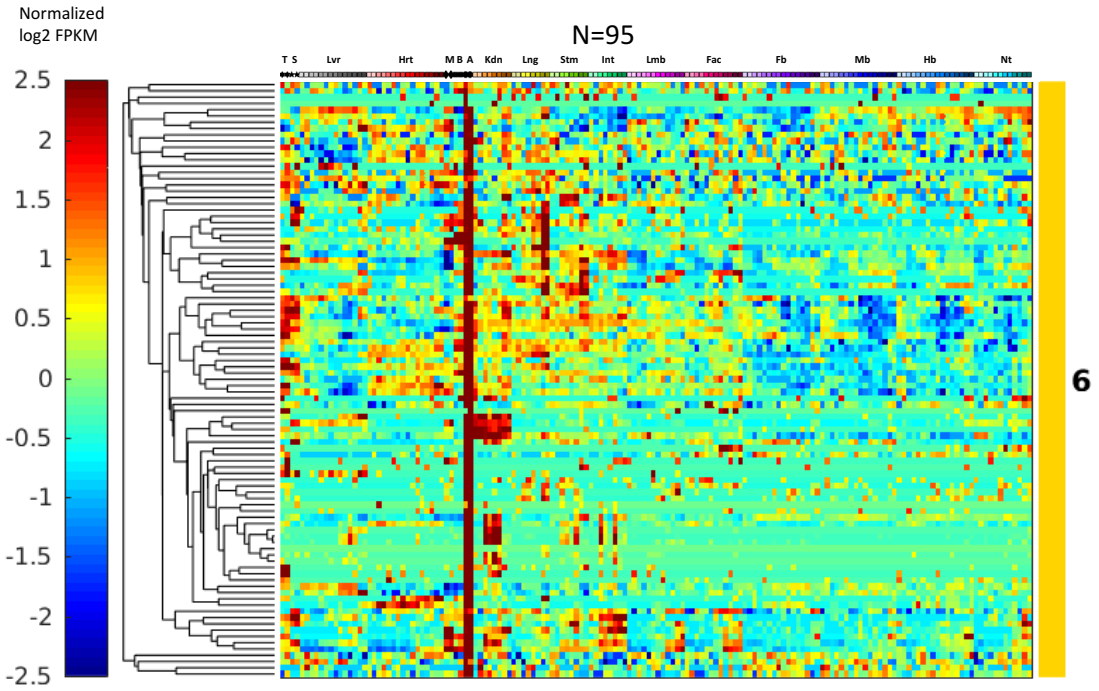


Figure 2.13: Cluster 6 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

The unifying theme of Cluster 6 (Figure 2.13) is high expression in the adrenal gland. Top GO terms include “hormone biosynthetic process” (p=1.5e-7) and “hormone metabolic process” (p=7.3e-7). More specifically, Cyp11b1, Cyp21a1, and Cyp11b2 contribute to the term “mineralocorticoid biosynthetic process” (p=2.8e-7). These cytochrome P450 genes are involved in biosynthesis of aldosterone which, unlike many other hormones, is produced only in the adrenal gland. However, all of Cyp11b1, Cyp21a1, and Cyp11b2 mentioned above have detectable expression signals in E15.5 and E16.5 samples of kidney. Their presence at E15.5 and E16.5 stages and absence in E14.5

and P0 may be due to the fact that E15.5 and E16.5 samples were pooled dissections, while E14.5 and P0 samples were dissected from individual embryos by another lab, which were more contamination-free. The reason why other adrenal genes did not show up in those contaminated kidney samples is probably that those genes do not have as high abundance as these.

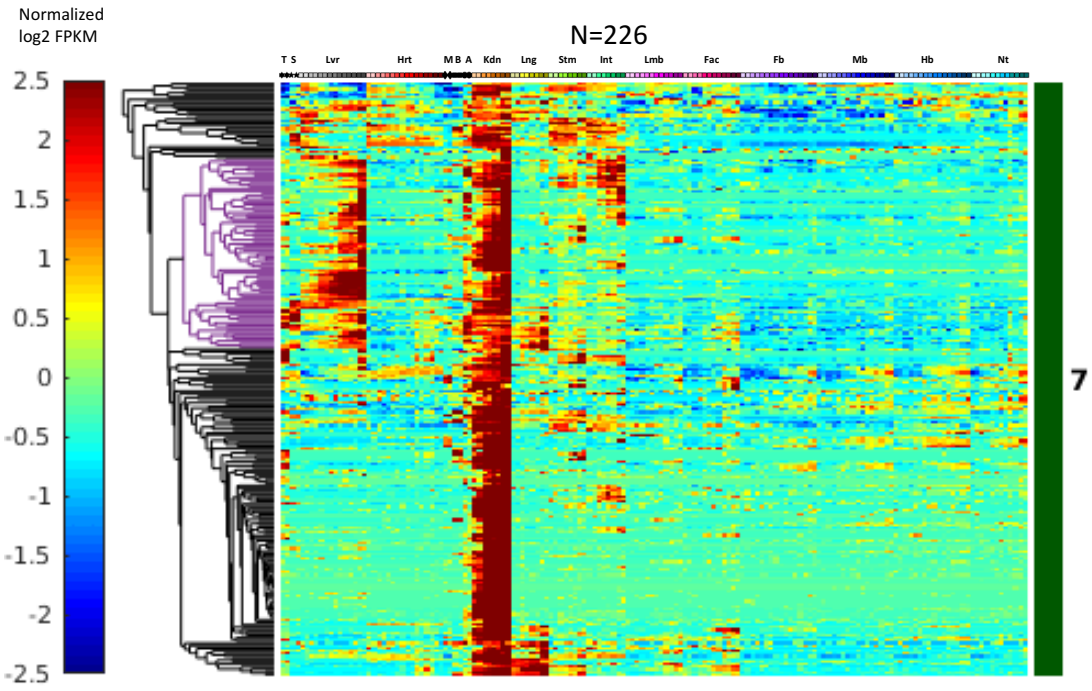


Figure 2.14: Cluster 7 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Cluster 7 (Figure 2.14) has a theme of kidney-specific expression, where transcript abundances increase over time. Roughly 40% of these genes are also expressed in the liver, again with increasing trajectories, plus some smaller subclades that are shared with gut or lung samples. Top GO terms of this cluster include transporter-related categories

such as “sodium ion transport” (p=2.0e-14) and “anion transport” (p=4.7e-9) and structural terms like “apical plasma membrane”. Therefore, this cluster seems to be dominated by genes responsible for transporter machinery and epithelial cell organization in the kidney. The clade of 72 genes labeled in purple contains genes enriched in both the liver and kidney. The top enriched GO terms for this group are for amino acid catabolic processes performed in both the liver and kidney (“organic acid metabolic process” (p=7.7e-12), “fatty acid metabolic process” (p=1.2e-7), and “alpha-amino acid catabolic process” (p=1.5e-6). 20 of these genes are enriched in kidney proximal tubule brush border cells, while 7 are enriched in hepatocytes[131].

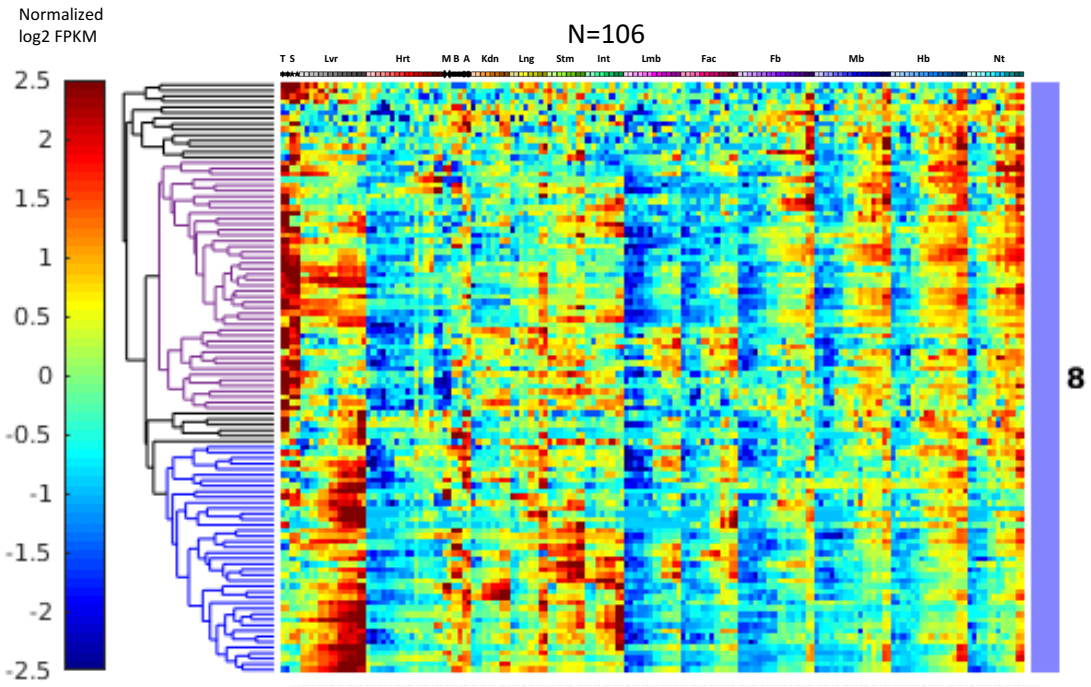


Figure 2.15: Cluster 8 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

The genes in Cluster 8 (Figure 2.15) have increasing expression patterns in almost all tissues, although the kinetics of increase differ. Top enriched GO terms include "inflammatory response" (p=1.0e-6) and "extracellular exosome" (p=1.5e-5). There are two major clades. The clade labeled in purple is consistent with genes marking the immune system, whose levels are highest in the thymus and spleen, but also include expression in the hematopoietic fetal liver. Subsets of these genes increase at later times in other tissues. GO analysis called terms including "regulation of T cell activation" (p=8.8e-5) and "inflammatory response" (p=2.2e-5). The second major clade, labeled in blue, is dominated by increasing expression in liver and gut tissues. Top GO terms included "extracellular exosome" (p=3.3e-9) for unknown reasons. This group of genes also highly overlaps with limb macrophage marker genes (defined in Chapter 3).

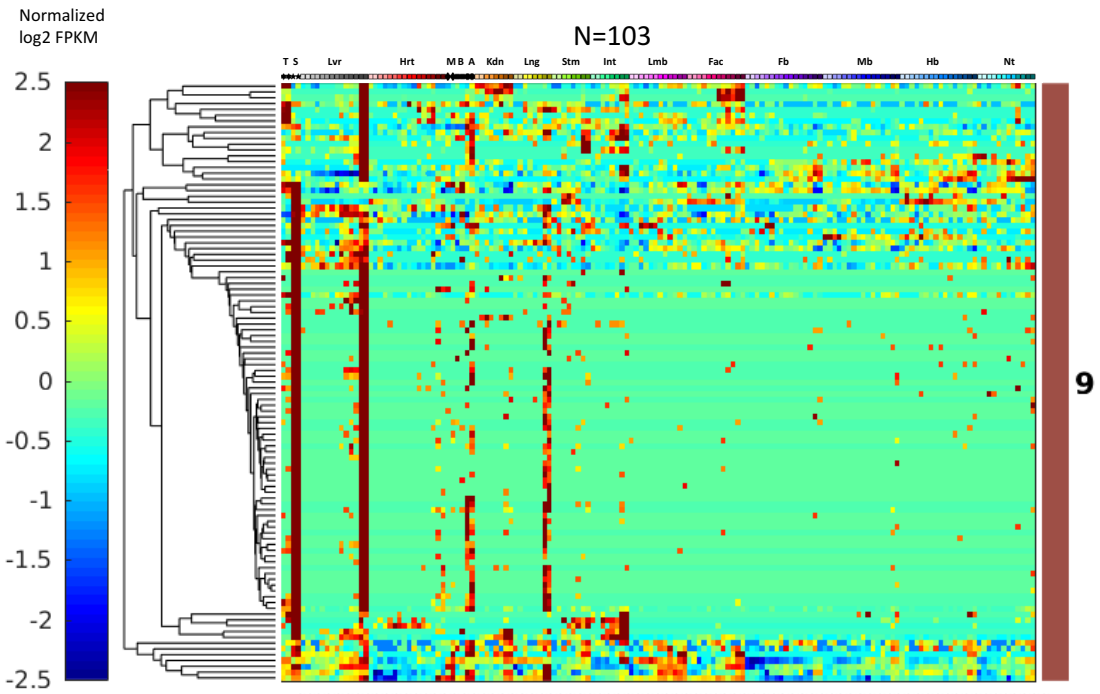


Figure 2.16: Cluster 9 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

The genes in Cluster 9 (Figure 2.16) have the highest enrichment by far in the spleen and on the P_0 liver (but not at earlier times). Moderate abundance is also seen in the adrenal gland and lung. There is minimal but detectable expression in all other tissues at P0, but very little at all times before birth. GO analysis did not yield significantly enriched terms, but more than half of the genes in this cluster are immunoglobulin components (kappa and lambda light chain variables, heavy chain variables, and constant regions), consistent with B-cell maturation, appearing in the liver, spleen, and in lesser proportions the lung and other lymphatic-containing dissections.

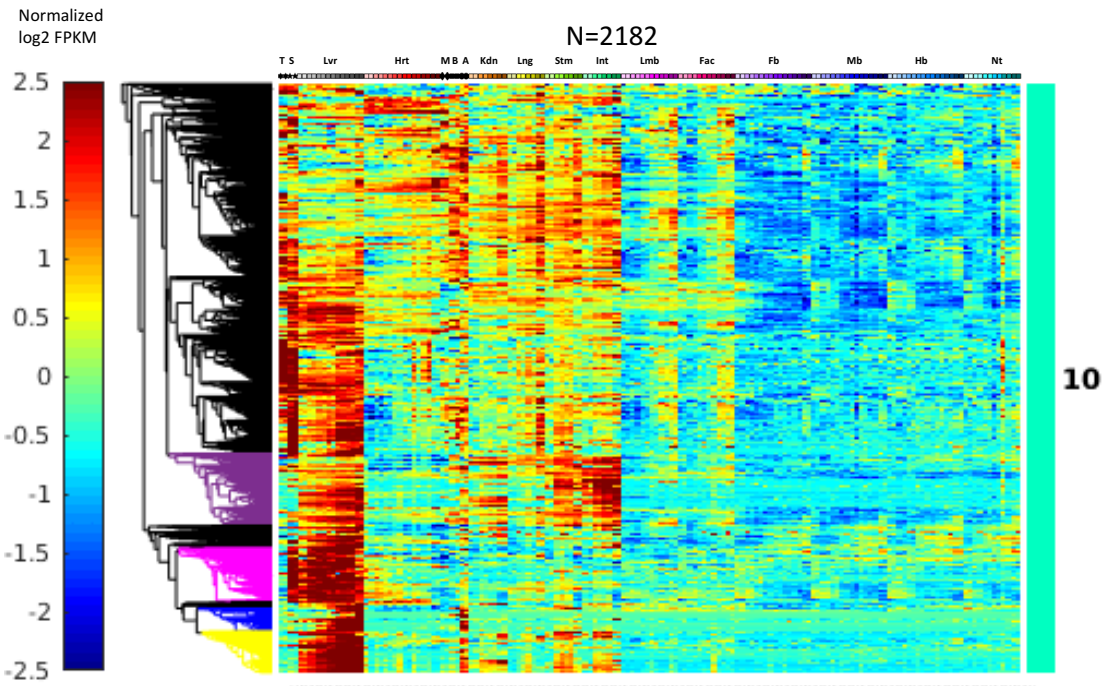


Figure 2.17: Cluster 10 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Over 60% of the genes in Cluster 10 (Figure 2.17) are preferentially expressed in the liver, lower in CNS tissues, and variously detected in other tissues. The RNA abundances mainly increase with time, but with differing kinetics. Top GO terms of Cluster 10 include the immune system, such as “immune system process” (p=4.8e-101) and “regulation of immune system process” (p=2.0e-62). The additional prominence of many genes in the P0 thymus and/or spleen, along with other non-CNS tissues, points to the lymphatic system.

In addition to the main immune theme, four clades with distinctions emerged. The one containing 267 genes labeled in purple are most enriched in the liver, as well as the stomach and intestine, increasing over time. Its top GO terms focus on lipids, including “lipid metabolic process” (p=3.8e-13) and “lipid transport” (p=4.1e-11), pointing to metabolic functions shared by hepatocytes and gut tissues.

The clade of 200 genes labeled in pink contains genes enriched in the spleen and liver only and points to erythropoiesis. Its top GO terms are mainly related to maturing red blood cells, such as “tetrapyrrole biosynthetic process” (p=1.2e-20) and “erythrocyte development” (p=3.9e-10). DNA motif analysis (to be discussed in detail in 2.2.5) of promoters in this clade revealed a significant enrichment of Tal1:Gata1, a known pair of regulators essential for hematopoiesis.

Members of the clade of 91 genes labeled in blue are mainly expressed in the late-stage liver and adrenal gland. Top GO terms include “monooxygenase activity” (p=1.1e-31),

"steroid hydroxylase activity" (1.7e-20) and "steroid metabolic process" (p=5.3e-9). More than a quarter of these are protein-coding components of cytochrome P450, which are involved in steroid and drug metabolism. Additionally, six sulfotransferase genes are also in this group. Sulfotransferase plays an important role in the metabolism of drugs, hormones, and bile acids.

Lastly, the clade of 155 genes labeled in yellow shows more constant levels through time in the liver, with additional expression detected in the adrenal gland, kidney, stomach, and intestine. Its top GO terms include "blood coagulation" (1.7e-29) and "alpha-amino acid metabolic process" (3.1e-12), with six coagulation factors, six complement factors, fibrinogens, and regulators (protein C and serpins)found in this clade.

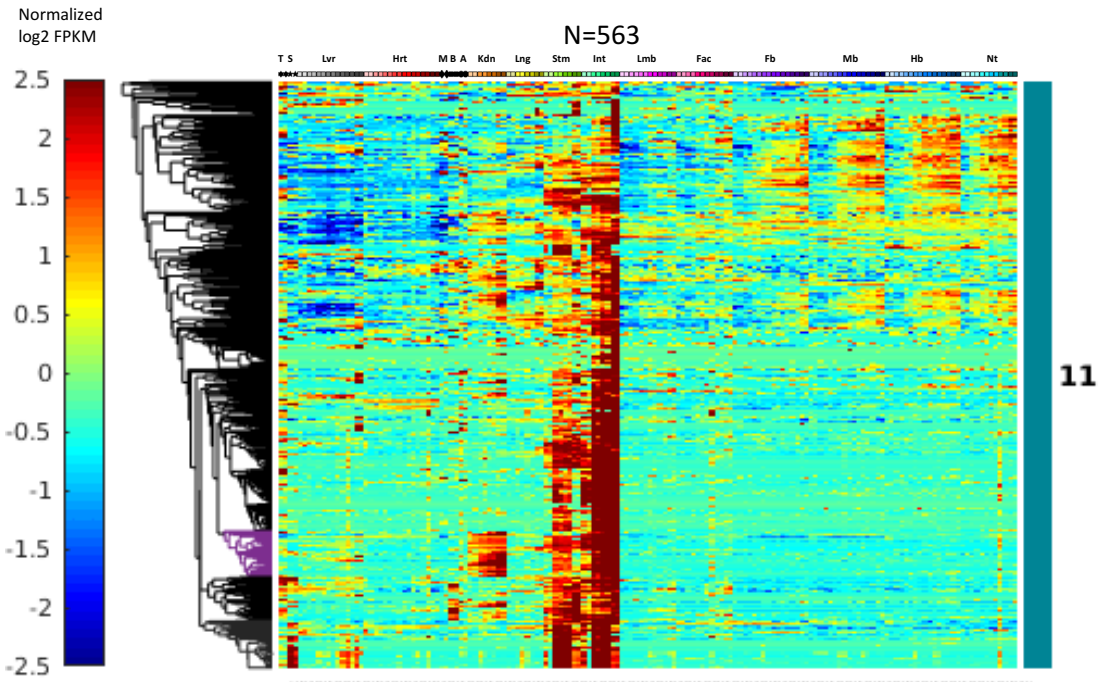


Figure 2.18: Cluster 11 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Genes in Cluster 11 (Figure 2.18) are most highly expressed in the intestine and are also enriched in the stomach, with sharing of genes with the kidney or CNS tissues. E14.5 and P0 timepoints show lower expression for multiple clades, which likely reflects systematic dissection differences at the boundaries between the two gut tissues. Top GO terms are mainly about intestine structure, including “brush border” (p=2.3e-11) and “brush border membrane” (p=3.1e-9). Interestingly, out of 16 genes contributing to the term “brush border”, 8 are in the small clade of 43 genes labeled in purple. This clade also has prominent increasing expression in the kidney, representing a shared program of brush border genes between the kidney and intestine. Other terms include “sodium ion transport” (p=1.0e-6), “digestive system process” (p=1.0e-6), and “alpha-amylase activity” (p=1.8e-6). Additionally, several gut hormones or peptides are found in this

group, such as cholecystokinin, gastrin, vasoactive intestinal polypeptide, ghrelin, glucagon, and insulin genes.

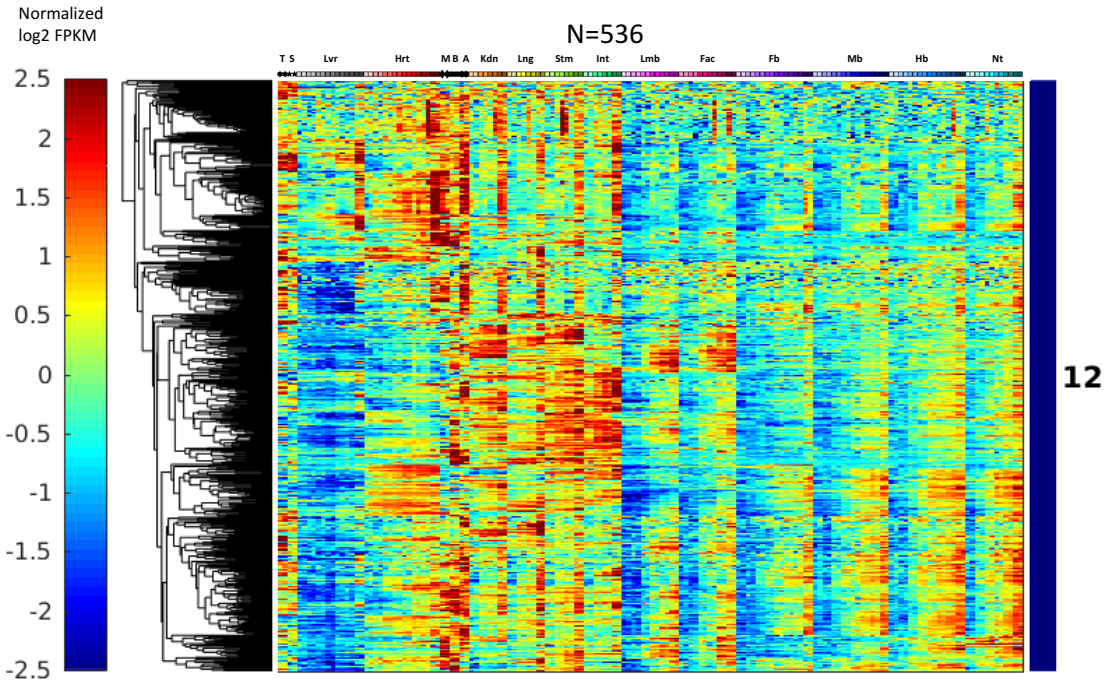


Figure 2.19: Cluster 12 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Most genes in Cluster 12 (Figure 2.19) are expressed widely and with an increasing trend, except in the liver, where most of the cluster is depleted at all times. The most prominent secondary theme is strong up-regulation at birth in multiple organs. Although no GO terms were significantly enriched, this cluster, and sub clusters within, are candidates for novel DNA sequence motif-derivation or for correlated microRNA signatures that could mediate the birth transition pattern and/or the liver suppression pattern.

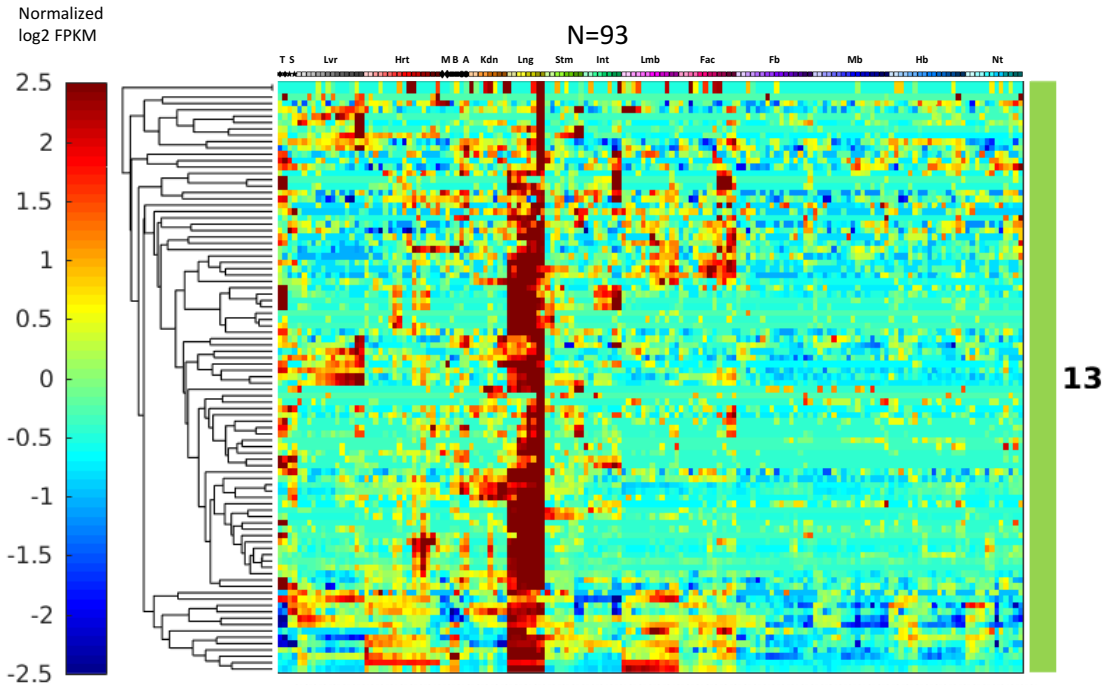


Figure 2.20: Cluster 13 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Gene expressions in Cluster 13 (Figure 2.20) are mostly enriched in the lung, especially at later stages. Partly because of the small cluster size, Gene Ontology did not provide highly significant terms. However, 4 surfactant-associated proteins contributing to the term “multi-vesicular body” (p=4.3e-6) are included in this cluster, indicating a possible link to Type II alveolar cells in the lung.

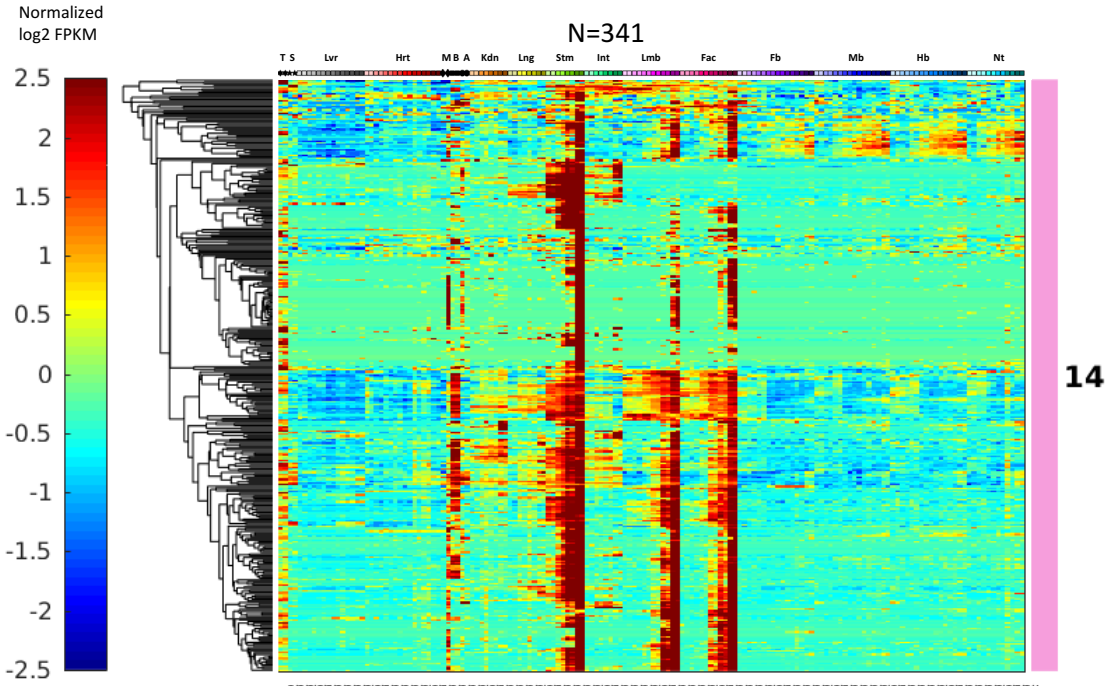


Figure 2.21: Cluster 14 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Cluster 14 (Figure 2.21) contains genes that are highly expressed in the stomach. Most are also highly expressed in limb and craniofacial prominence at very late stages. About a quarter of them are expressed in the P_0 bladder as well. The top GO terms are "cornified envelope" (p=6.7e-26) , "keratinization" (p=1.7e-27), "epidermis development", and "keratinocyte differentiation" (p=1.6e-16). The cornified envelope is composed of a layer of dead cells found in the skin epidermis and forestomach for protection against the environment. Its major components include loricrin, filaggrin, involucrin, keratins, and small proline-rich protein (SPR) genes that are all found in this cluster, together with the genes required for generating the cornified envelope, such as transglutaminase, cystatin, and envoplakin.

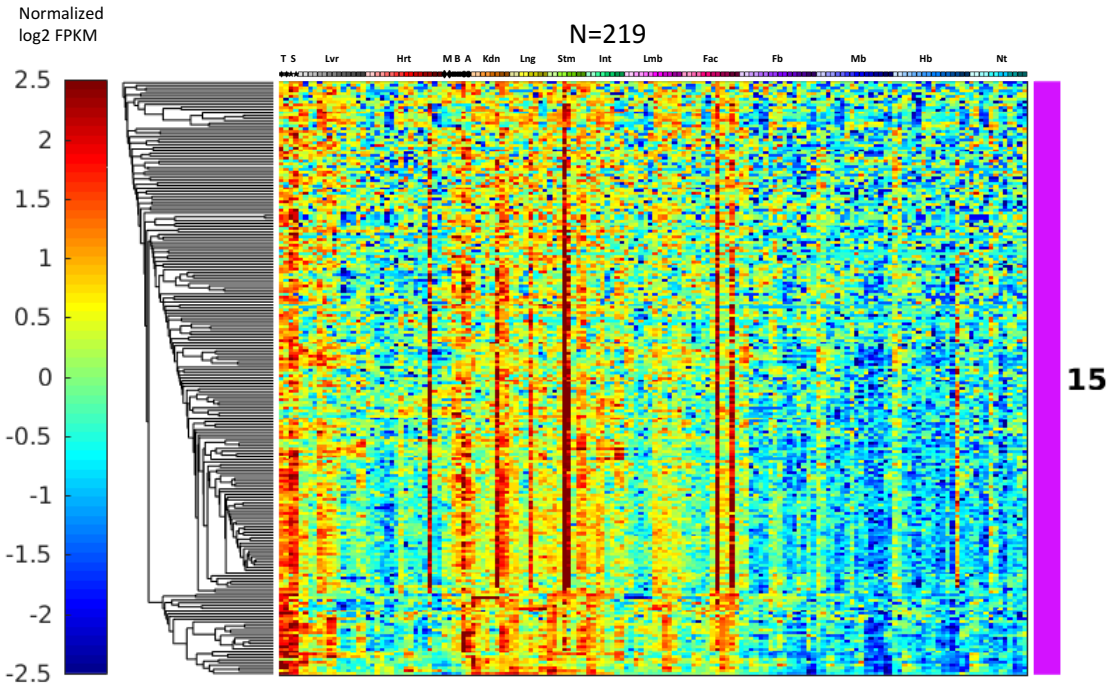


Figure 2.22: Cluster 15 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Genes in Cluster 15 (Figure 2.22) are coherently enriched in specific samples, but they do not reproduce between replicates or among related tissues. Almost all genes in this cluster are known pseudogenes or protein-coding genes with low mappability. These low-mappability genes' top-abundance mappable counterparts (their corresponding protein-coding genes or paralogs) do not display similarly variable enrichment, which argues that the differences are not due to the obvious source of biological variation (data not shown).

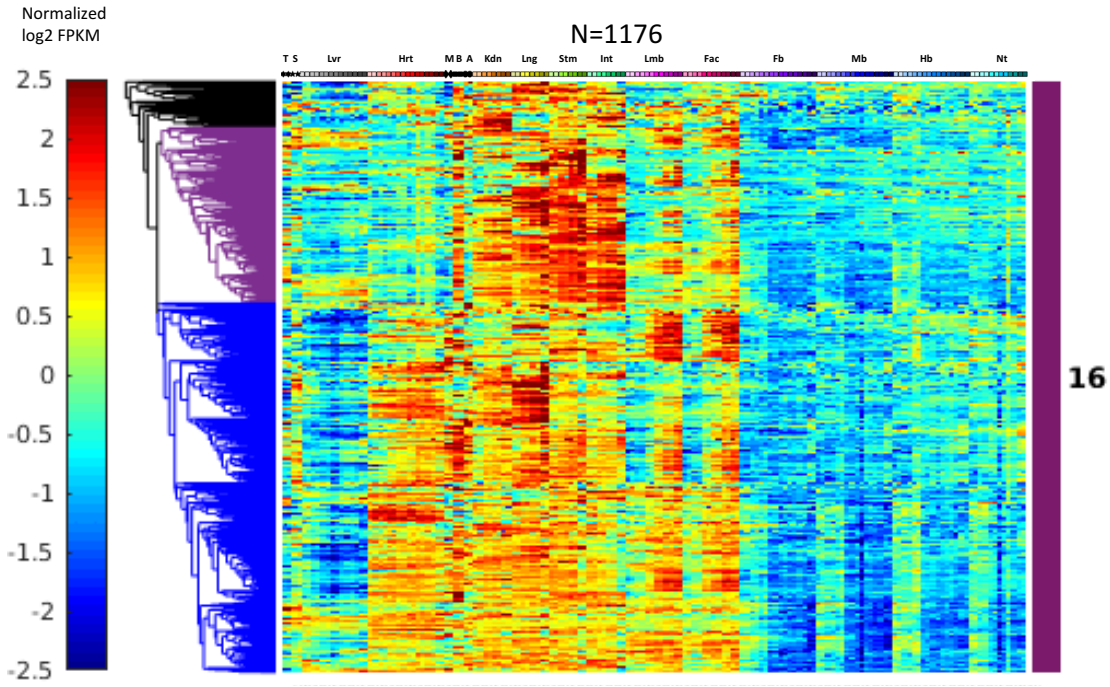


Figure 2.23: Cluster 16 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

The broad theme of Cluster 16 (Figure 2.23) is expression in most tissues and organs, with the exception of the CNS and liver, both of which show little expression. Over the developmental time course, most members increase in limb and craniofacial prominence but are relatively invariant or decreasing in other tissues. The top GO terms of Cluster 16 are dominated by extracellular matrix (ECM) components, such as “extracellular matrix” (p=8.7e-58), “extracellular region part” (p=6.0e-42), and ‘’basement membrane” (p=4.4e-

34). Other significant terms include “regulation of cell migration” (p=8.0e-26), “angiogenesis” (p=1.1e-25), and “cell junction” (p=2.3e-18).

This cluster contains two major clades, highlighted in purple and blue, that share expression in the bladder, kidney, lung, stomach, intestine, limb, and craniofacial prominence. The blue clade is distinct in also showing strong expression in the heart. Their GO terms identify different biases. The purple clade features “occluding junction” (p=7.6e-9) in addition to ECM terms, while “angiogenesis” is absent. The blue clade includes most of the Cluster 16-themed terms, but also emphasizes “anchoring junction” (p=7.3e-19) and particularly “adherens junction” (p=3.0e-18), consistent with epithelial/endothelial cell junction formation and tube morphogenesis. Thus the purple clade focuses on tight junctions that consist of an epithelial barrier and molecular gate between a cell mass and the environment, while the blue clade concentrates on angiogenesis and adherens junctions that link cells together and also carry cadherin receptors important for tissue morphogenesis.

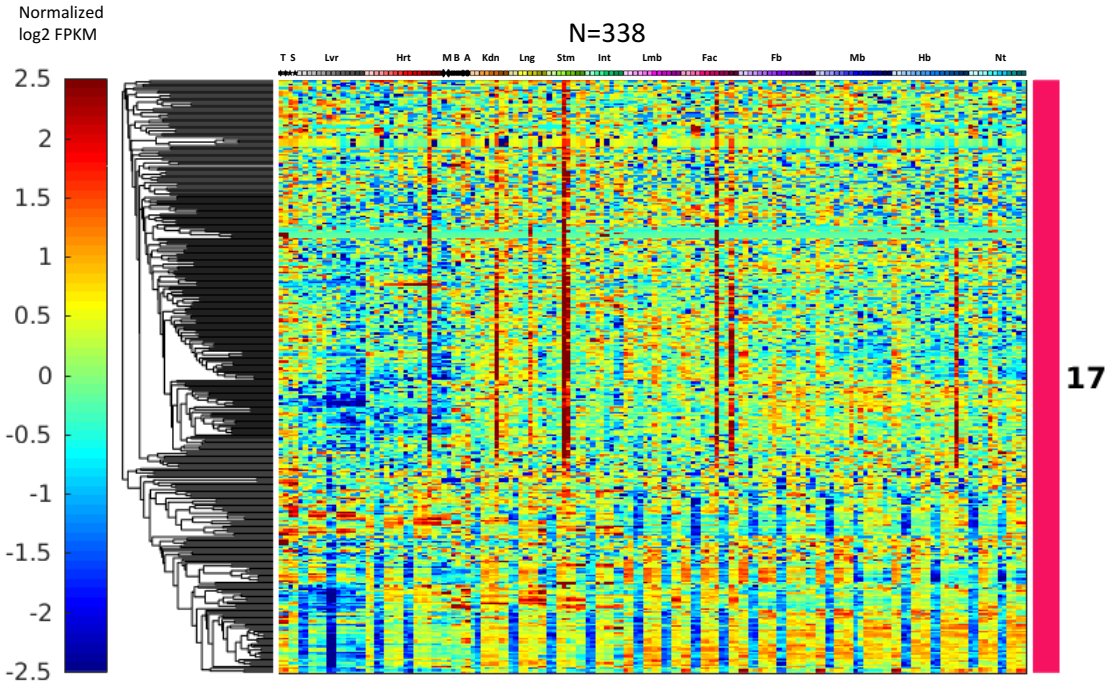


Figure 2.24: Cluster 17 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Cluster 17 (Figure 2.24) is divided into two major clades. The upper clade, similar to cluster 15, is coherently enriched in a few individual samples that do not replicate, nor do they reproduce among related tissues. Apart from these individual samples, the pattern is noisy across developmental time. This pattern does not correspond to any known dissection or global quality issue. Also similar to Cluster 15, no GO term enrichment was found. The lower clade contains genes that are widely expressed among different tissues

that are also systematically depleted in the E11.5 and E14.5 samples. This reflects known batch effects at tissue collection/dissection steps.

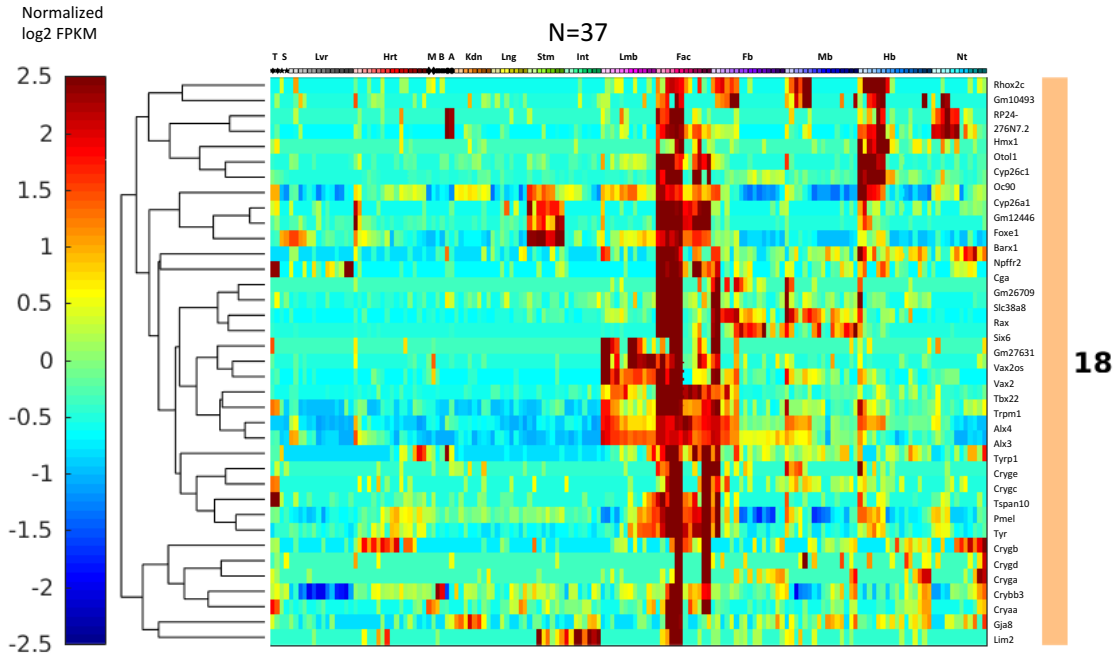


Figure 2.25: Cluster 18 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Most genes in Cluster 18 (Figure 2.25) are enriched in the craniofacial prominence at early stages but not later. Its top GO terms mainly concern eye development, including "structural constituent of eye lens" (p=1.6e-17) and "eye development" (p=1.1e-15). Genes include crystallins, retinoic acid-metabolizing enzymes (Cyp26a1 and Cyp26c1), lens membrane protein (Lim2), melanin regulators (Tyrp1, Tyr and Pmel), and one developmental regulator (Vax2). The dissection plan for cranioface was to exclude the eyes, but at earlier stages it appears not to have been fully successful. The expression pattern and Gene Ontology of Cluster 18 genes in the early craniofacial prominence

samples (E10.5, E11.5 and E12.5) including sharp transitions between adjacent timepoints, are likely due to imperfect removal of early eyes. Sporadic enrichment of these genes in later stage craniofacial prominence samples (E16.5) is likely due to a few imperfect dissections in a large embryo pool.

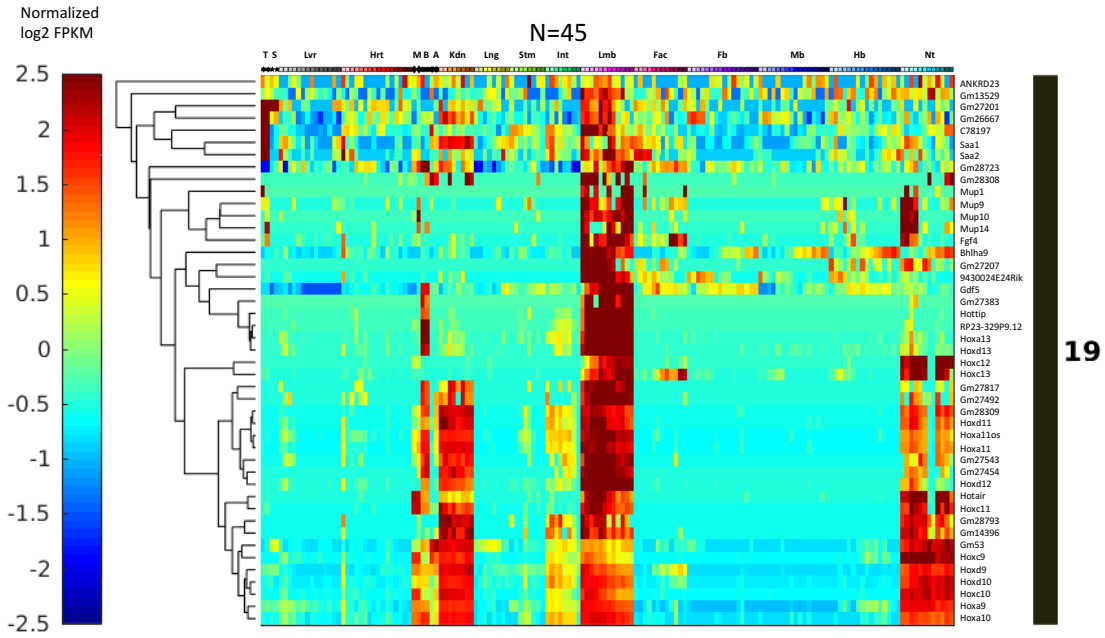


Figure 2.26: Cluster 19 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Genes in Cluster 19 (Figure 2.26) are mostly enriched in the bladder, kidney, limb, and neural tube. Within these tissues, expression levels are relatively constant over developmental time. The lower half of this cluster contains 5' Hox genes (9-13) and lincRNAs localized in the 5' region of Hox clusters (Hotair and Hottip). Genes with names beginning with "Gm" that are clustered together with Hox genes are also localized in 5' Hox gene regions, suggesting shared transcriptional regulatory elements or RNA

precursors. Among these Hox-cluster genes, there are distinctions, with 5' Hox expression being more abundant in posterior tissues (e.g. bladder, kidney, and intestine), consistent with previous findings. As the time course begins at E10.5, I could not follow the well-known upregulation sequence of Hox genes which displays "temporal co-linearity", except for a gradual increase in Hoxc12 and Hoxc13 in the limb, which represents the distal ends of limbs and whose upregulation pattern is late enough to be captured in our time window. In E14.5 neural tube samples, the 5' most Hox genes Hox11-13 are missing-because that batch of embryo dissections did not include the posterior tip of the tube.

Four major urinary protein (MUP) genes are enriched in the limb, similar to the MUP paralogs in Cluster 2. However, unlike those in Cluster 2, they are also enriched in early neural tube samples. Further studies are needed to understand their functions.

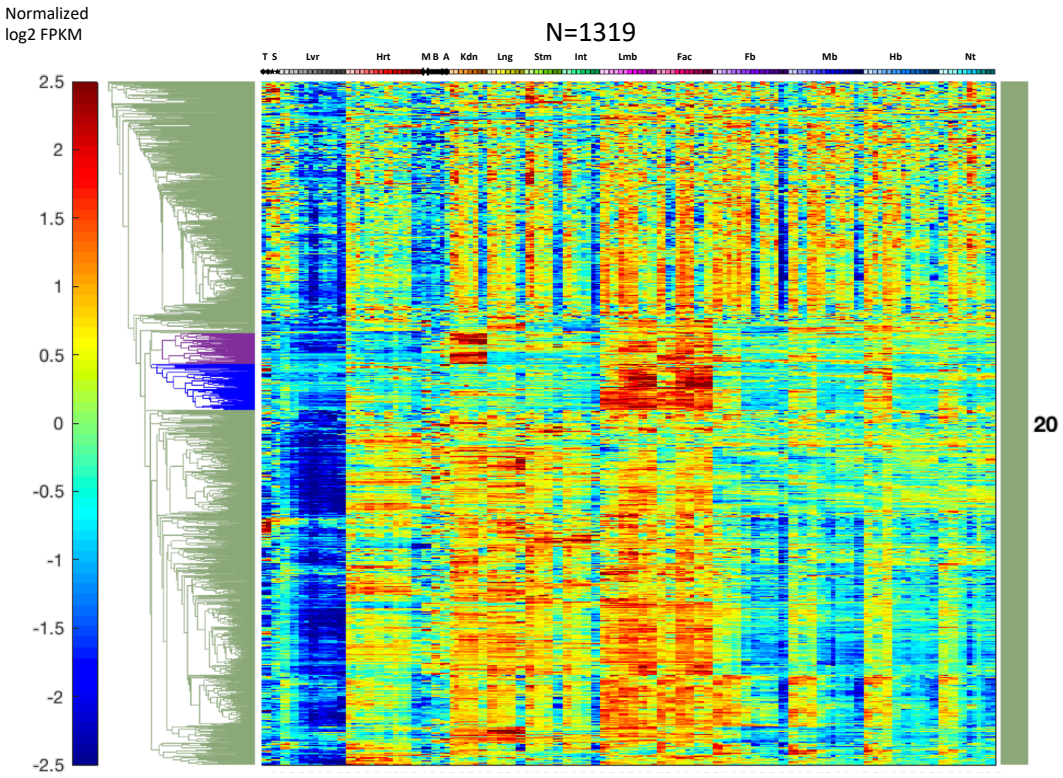


Figure 2.27: Cluster 20 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Overall, genes in Cluster 20 (Figure 2.27) are prominently absent from the liver at all stages and are absent or strongly reduced at P_0 in most tissues having P_0 data. The top ~1/3 of the cluster contributes little to the two major themes derived from expression and GO. It contains mainly pseudogenes and lncRNAs. In the limb, craniofacial prominence, and brain, depletion of some of these genes is evident at E11.5 and E14.5 similar to Cluster 17, and possibly related to the batch effect discussed before. In the remaining bottom 2/3, there is considerable substructure among expressing tissues due to the two major biological themes: the first GO enrichment theme is tissue morphogenesis and development, such as “skeletal system morphogenesis” (p=3.3e-13), “branching

morphogenesis of an epithelial tube" (p=5.5e-12), "sensory organ development" (p=7.7e-11), "odontogenesis" (p=1.3e-10), "gland development" (4.5e-10), "ossification" (p=9.1e-10), "limb morphogenesis" (p=2.1e-9), and "kidney development" (6.1e-9); the second GO theme is Wnt signaling, such as "regulation of Wnt signaling pathway" (8.2e-12), "Wnt signaling pathway" (2.0e-10), and "Wnt-protein binding" (3.2e-10). Although this cluster called terms covering a variety of different aspects of development demonstrated in the first theme, the driving genes are often shared among multiple terms referring to different tissues. This likely reflects the broad usage of these signaling pathways in patterning and morphogenesis. Moreover, roughly a quarter of the genes contributing to any morphogenesis-theme terms also contributes to the Wnt theme. Other morphogenesis-theme genes that do not currently contribute to the Wnt GO terms, such as Irx3, Runx2, TWIST, Bmp4, Tbx1, and Tbx3 have been independently associated with this signaling system. This is consistent with the current appreciation that Wnt signaling plays an important and widely distributed role in different individual anlage, including stem cell renewal[132].

The clade of genes colored in purple are highly enriched in kidney and moderately enriched in limb and craniofacial prominence. The themes are suggested by Gene Ontology with terms "skeletal system morphogenesis" (p=1.2e-6) and "branching morphogenesis of an epithelial tube" (2.3e-6). These terms of this clade are consistent with the overall theme of this big cluster, but the distinct gene expression pattern suggests intensive usage of this subprogram of genes in the kidney. The clade of genes labeled in blue has prominent enrichment in limbs and craniofacial prominence and is lower but still

detectable and decreasing in other tissues. The top GO terms are similar to those called from the whole cluster, but with much enhanced significance for “embryonic skeletal system morphogenesis” (p=1.8e-15) and “cartilage development” (p=1.5e-9).

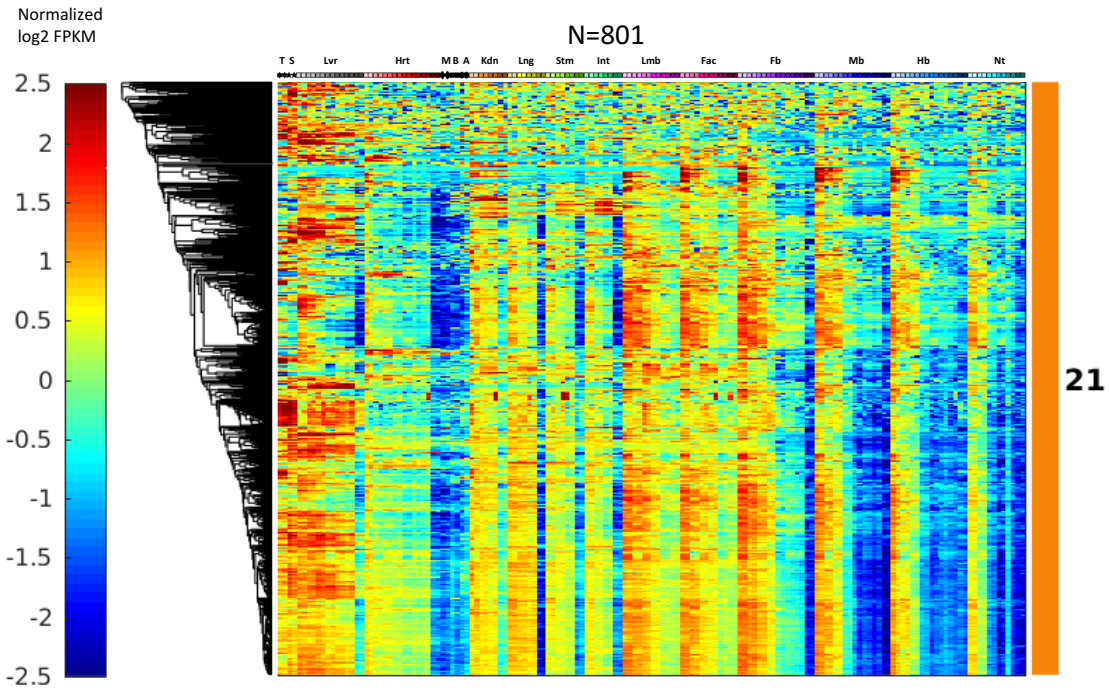


Figure 2.28: Cluster 21 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Genes in Cluster 21 (Figure 2.28) are expressed in all the fetal tissues and decrease over time in most. At P0, the majority are expressed in the thymus and spleen but are notably depleted elsewhere. Top GO terms are mainly about cell division and nucleus components, such as “chromosomal part” (p=1.2e-93) and “cell cycle process” (p=1.6e-87) consistent with genes involved in executing the cell cycle, especially components of the chromosome and its associated proteins. The global decrease in RNA levels from

these genes over time is consistent with shifting from fast growing proliferating cells to more differentiated ones. Fetal hemoglobins are also found in this cluster, such as Hbb-bh1, Hbb-y, and Hba-x.

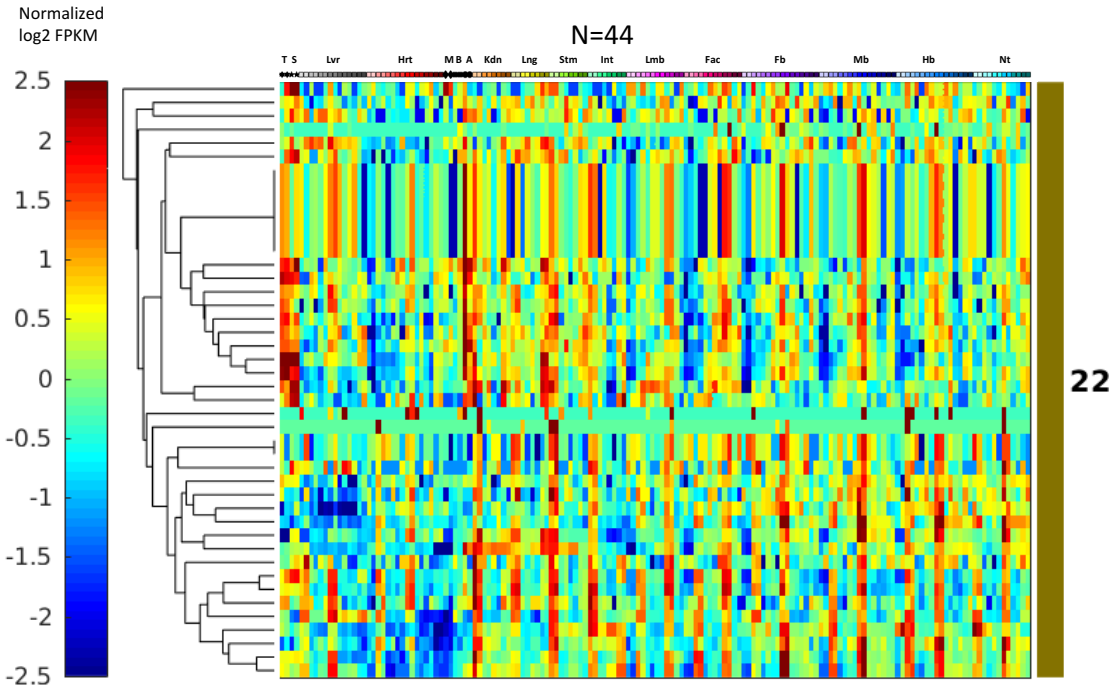


Figure 2.29: Cluster 22 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Genes in Cluster 22 (Figure 2.29) show distinct enrichment at E11.5 and E14.5 stages in some tissues, and they do so reproducibly among the replicates. Most of these genes are pseudogenes and low-mappability protein-coding genes. They are similar to the batch-effect heavy clades in Cluster 17 and Cluster 20 but display the inverse pattern trend. They may be artifacts from a similar type of batch effect.

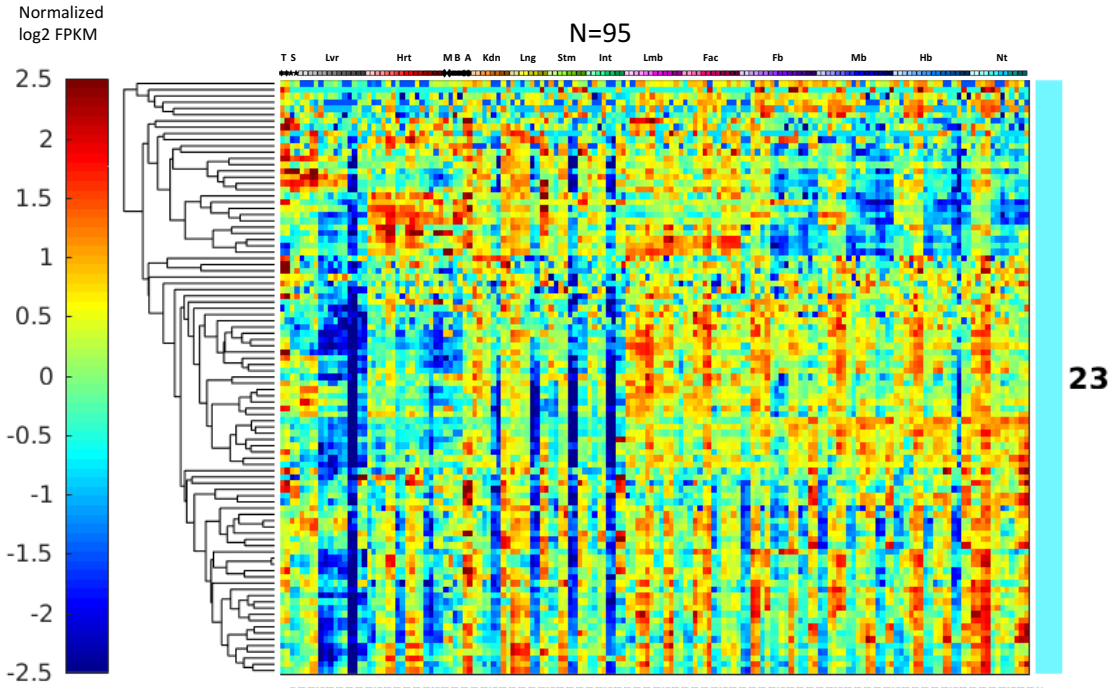


Figure 2.30: Cluster 23 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Cluster 23 (Figure 2.30) contains genes most prominently expressed at early times in CNS tissues. They are also depleted preferentially at E16.5 in many other tissues. Unlike the candidate batch effects of clusters like 15, 17, and 22 that are heavily enriched in pseudogenes, this cluster is not explained by annotated pseudogenes. There was no significant GO enrichment, and batch effect might be a cause of the expression pattern.

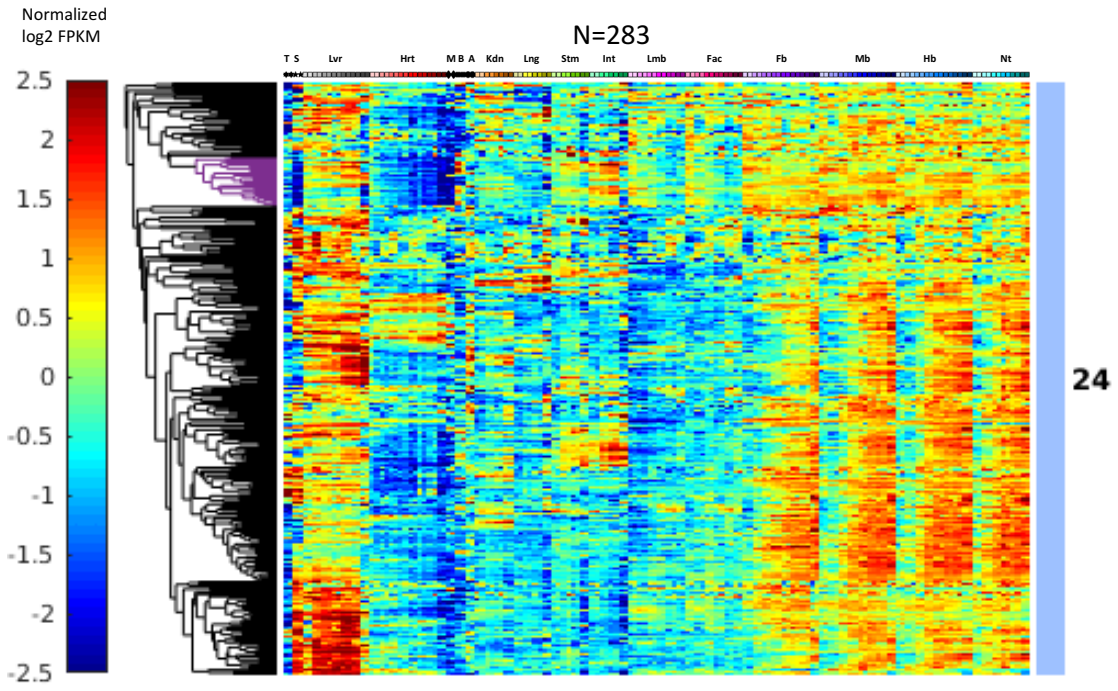


Figure 2.31: Cluster 24 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1

Genes in Cluster 24 (Figure 2.31) are widely expressed and are preferentially higher in the CNS regions and/or in the developing liver. Most, but not all, increase during development of these tissues. Top Gene Ontology terms are dominated by lipid metabolism, such as “lipid metabolic process” (p=2.7e-13) and “cholesterol biosynthetic process” (p=1.7e-11). Interestingly, all of the nine genes contributing to the term “cholesterol biosynthetic process” are localized in a tiny clade of 23 genes labeled in purple. These 23 genes are all very abundant and highly correlated among themselves. Further research may contribute to the shared metabolic mechanisms between liver and CNS tissues.

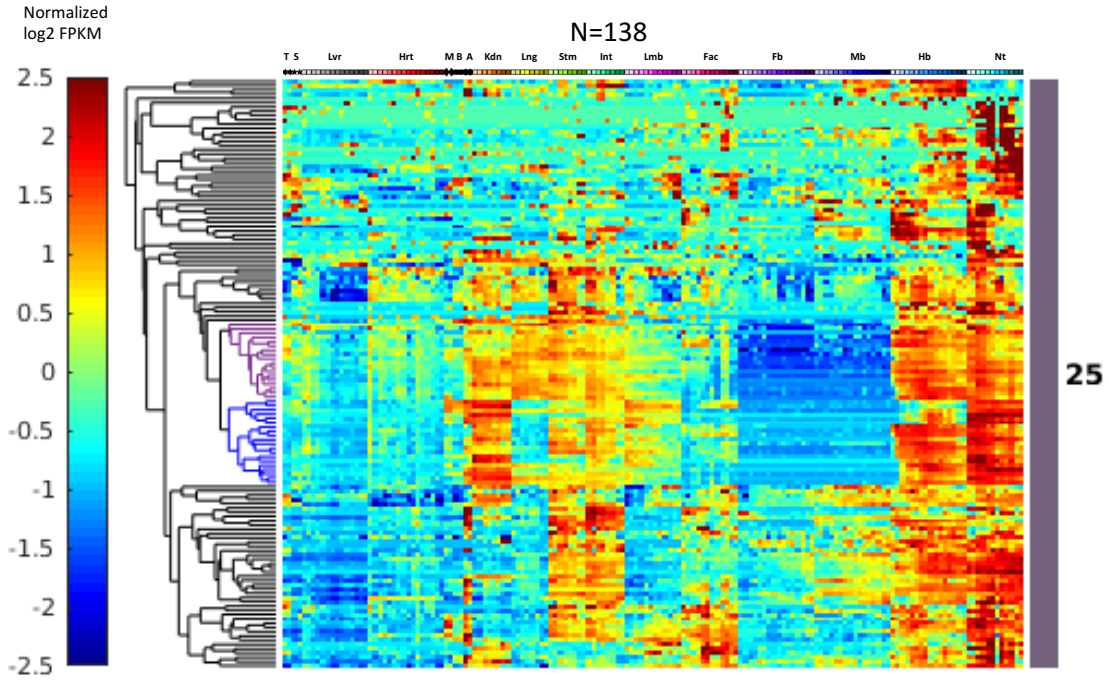


Figure 2.32: Cluster 25 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

More than half of the genes in Cluster 25 (Figure 2.32) are consistently and highly enriched in the hindbrain and neural tube plus the stomach, intestine and adrenal gland. The Kidney and lung also express distinct subsets of these genes. This cluster contains most of the 3' Hox genes, almost all located in the two clades labeled in purple and blue. The purple clade consists of the 3' most Hox genes and genes sitting in the 3' end of Hox gene clusters, while the blue clade is made of Hox genes and non-Hox genes in the center (less 3' but not 5') of Hox clusters. The purple-clade genes are expressed in the lung while the blue ones are mostly not. This is probably because the lung is relatively anterior to other endoderm tissues assayed, which correspond to the 3' end of the endoderm Hox A/P axis. For the genes outside the Hox gene clades combined, Gene Ontology generated

terms related to the neural system, such as “neuron differentiation” (p=1.4e-7), focusing on “enteric nervous system development” (p=2.4e-7). In E14.5 neural tube samples, some genes are more depleted compared to E13.5 and E15.5. I think this might result from a dissection protocol detail that produced shorter spinal cords and depleted the 3’-most Hox expressing tissue.

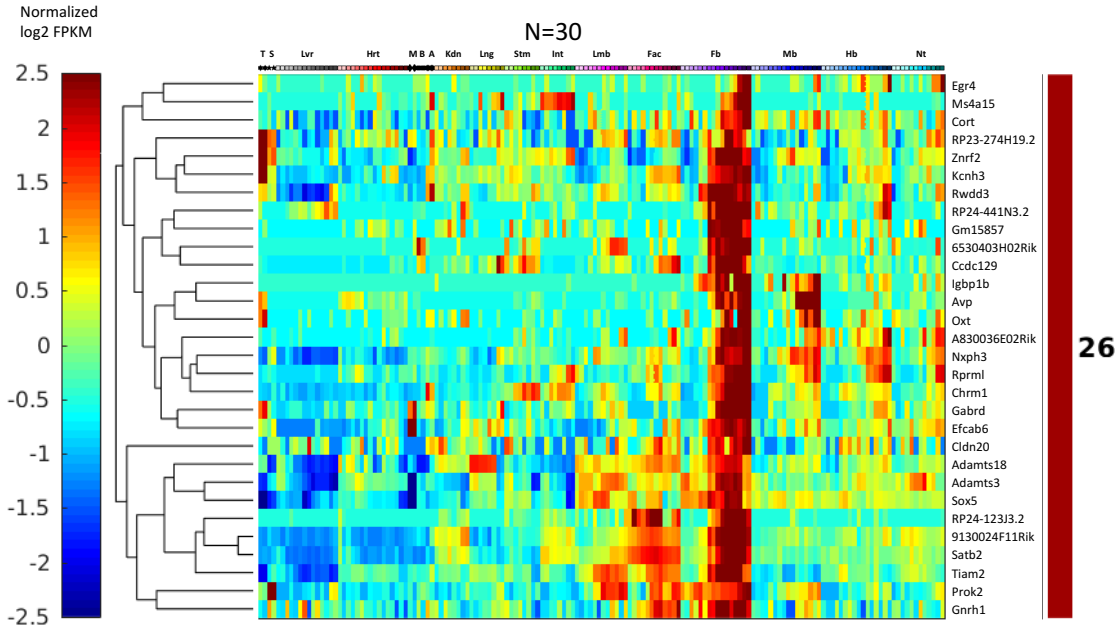


Figure 2.33: Cluster 26 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Genes in Cluster 26 (Figure 2.33) are mostly enriched in the forebrain at late stages. Avp and Oxt encode neuropeptides synthesized in the hypothalamus that regulate complex maternal and sexual behaviors[133]. They are clustered together within 6kb on chromosome 2. It would be interesting to further study the relationship between these two genes and others in this cluster.

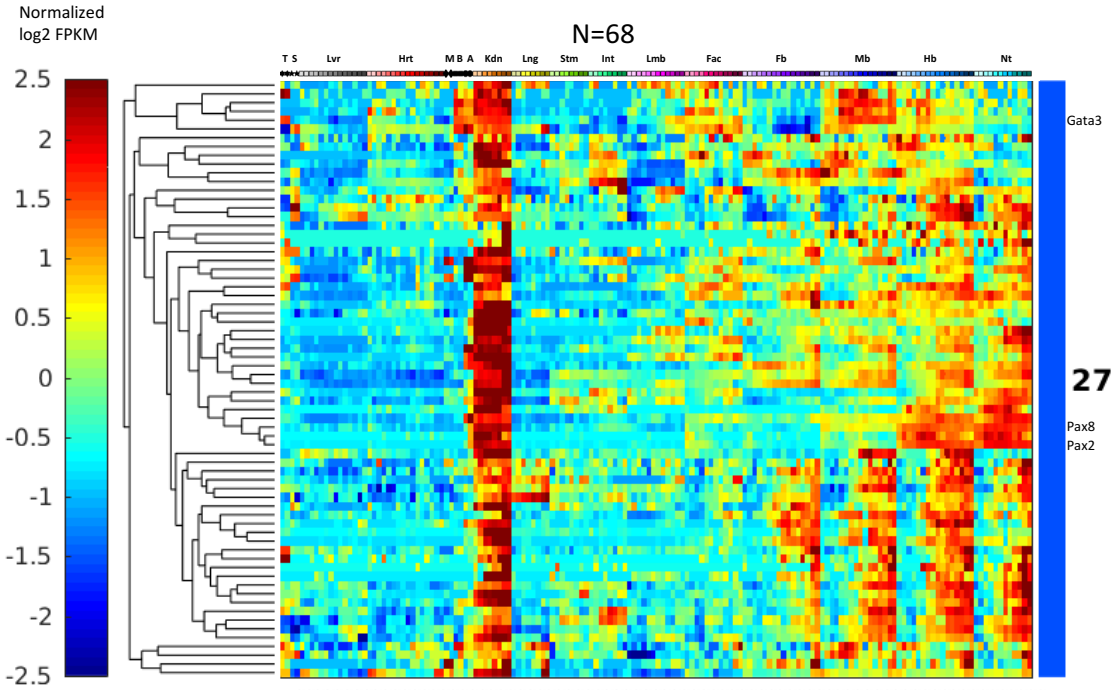


Figure 2.34: Cluster 27 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Cluster 27 (Figure 2.34) contains genes highly enriched in the kidney. Most are also expressed in the brain and neural tube at later stages but less abundantly than in kidney. Pax2, Pax8, and their target Gata3[134] are found in this cluster, which specify the nephric lineage and regulate branching morphogenesis in the developing kidney. Pax2 and Pax8 are also reported to specify GABAergic and glycinergic neuronal fates[135], partly explaining expression in the hindbrain and neural tube. It is possible that this cluster concerns two independent cell fate specification and morphogenesis programs that use overlapping regulatory factor sets, such as Pax2, Pax8, and Gata3.

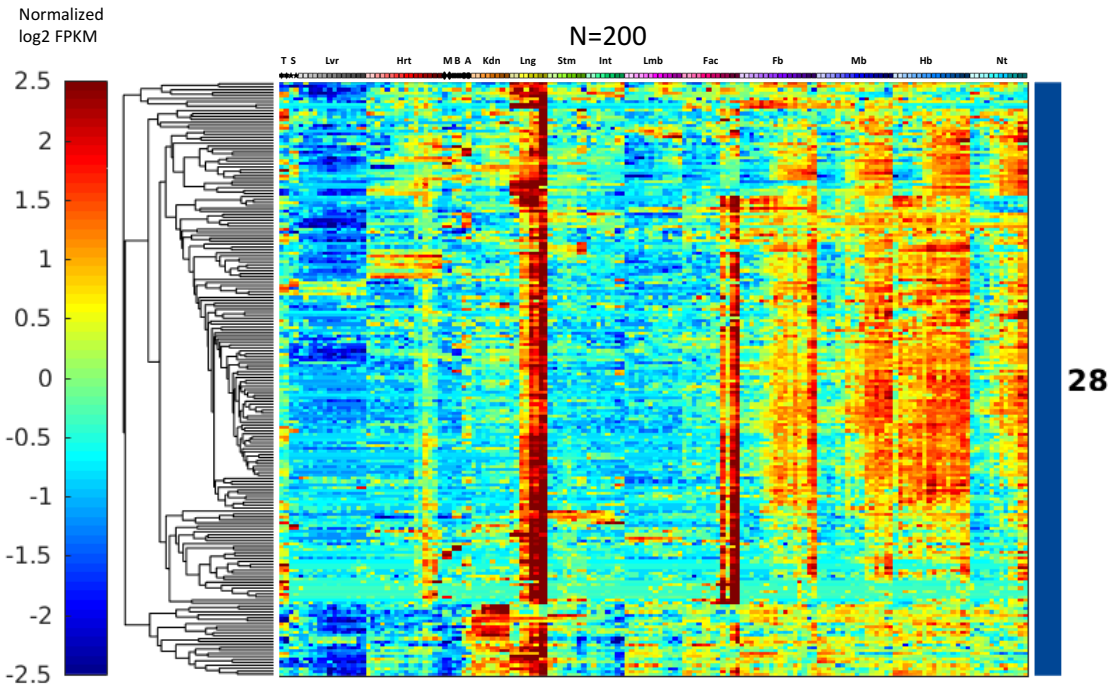


Figure 2.35: Cluster 28 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Genes in Cluster 28 (Figure 2.35) are expressed in many tissues, but with lung and craniofacial prominence being highest, followed by CNS regions. Almost all increase over time, but with differing kinetics in different tissues and brain regions. Most of the significant GO terms are about ciliogenesis, such as “cilium movement” (p=1.4e-19), “cilium” (1.2e-17), and “outer dynein arm assembly” (1.5e-14). The contributing genes include components of dynein arms and radial spokes, genes coding for assembly machinery such as dynein docking complex and the tubulin modifying enzyme, and the nexin-dynein regulatory complex. Two known cilium regulators, Foxj1 and Mcidas[136], are also in this cluster. The cilium is a fundamental structure, with primary cilia being ubiquitous while secondary and sensory cilia having more specialized distributions›[137]

that correspond well with the pattern for the majority of genes in Cluster 28. The pattern can be explained by the emergence of airway cilia in the lung, the airways of the craniofacial prominence, and the ependymal cilia of the CNS.

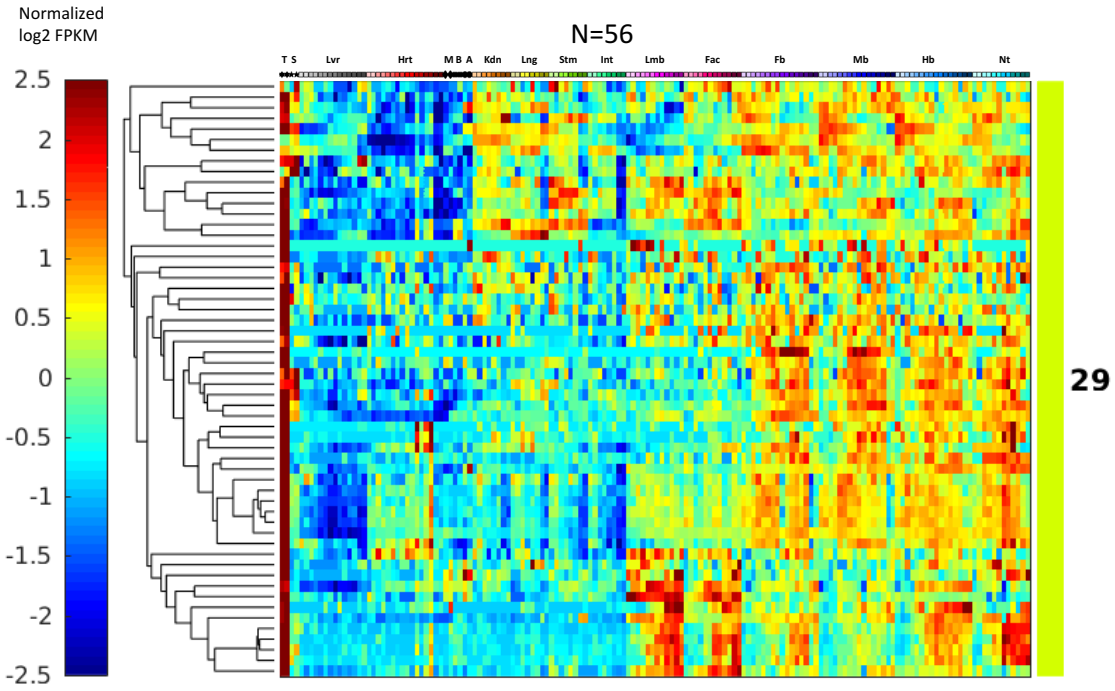


Figure 2.36: Cluster 29 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Most genes in Cluster 29 (Figure 2.36), a relatively small cluster, are distinguished by highest expression in the thymus, but more than half are also expressed substantially in the brain or in the face/limb or in the kidney/lung and gut. Gene Ontology failed to identify a significantly enriched term for this group, probably due to small sample size.

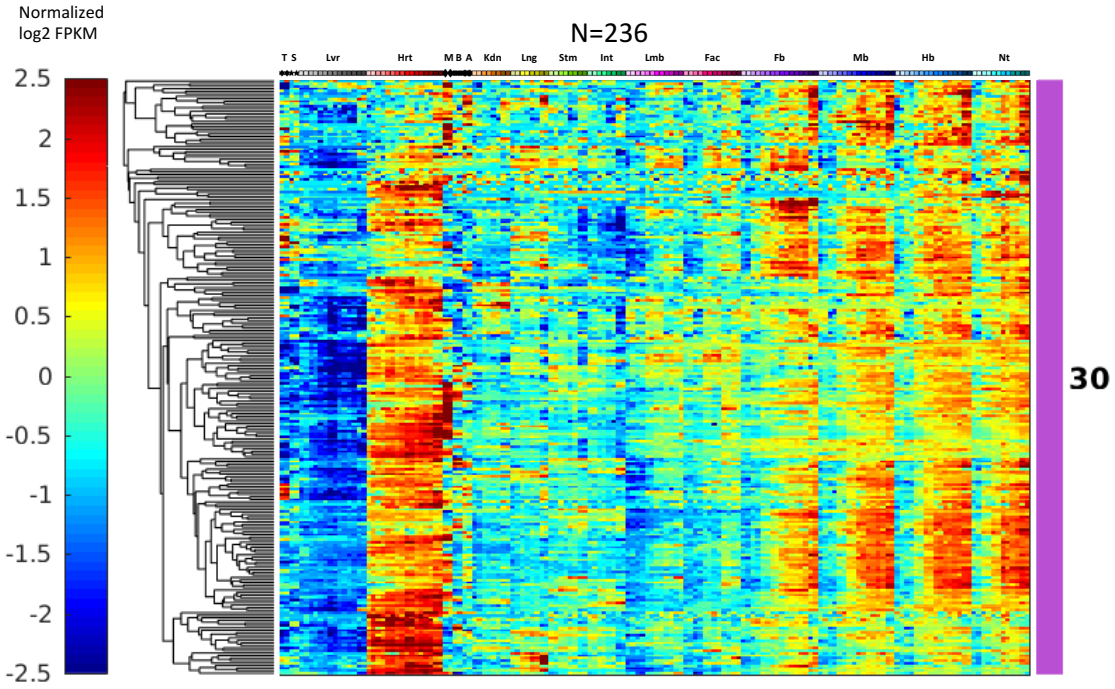


Figure 2.37: Cluster 30 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Cluster 30 (Figure 2.37) contains genes expressed most prominently in the heart and/or CNS samples, with the admixture among the tissues varying across different clades. Top enriched GO terms mainly identify transport of metal ions, such as “metal ion transport” (p=4.1e-8), “metal ion transmembrane transporter activity” (p=1.3e-7), and “potassium ion transmembrane transporter activity” (p=2.0e-7).

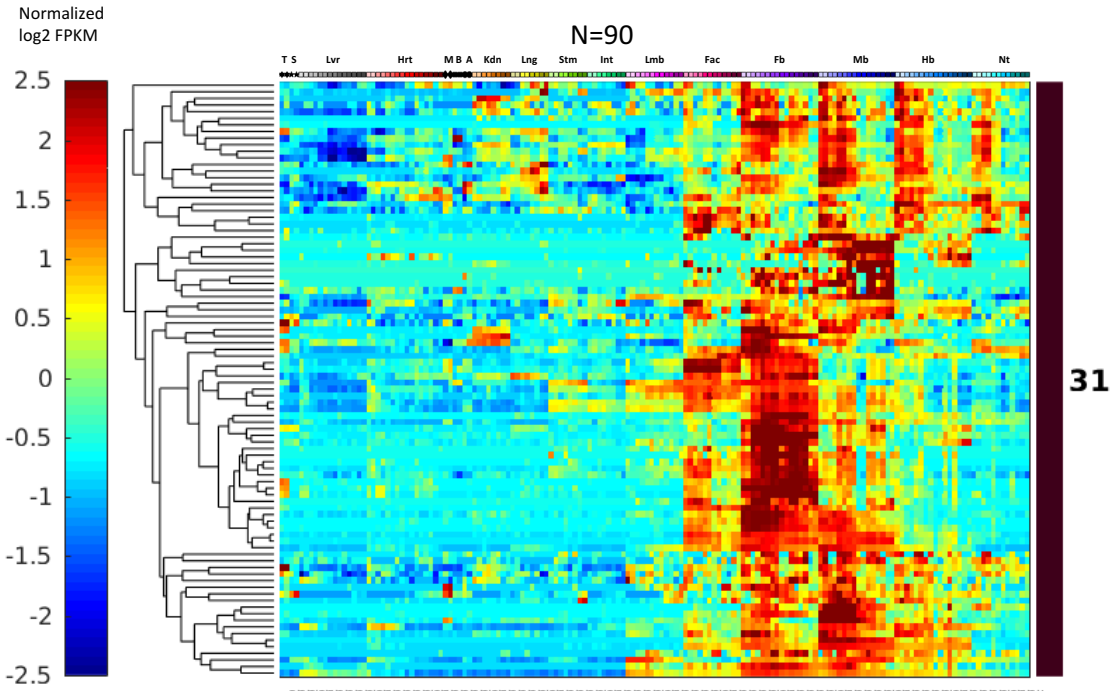


Figure 2.38: Cluster 31 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Genes in Cluster 31 (Figure 2.38) are mainly enriched in the brain and neural tube, with different regionalization for sub-clusters, plus facial prominence (perhaps partly driven by cross-contamination of face with forebrain dissection at early times). More than a third of the genes in this cluster are transcription factors ("sequence-specific DNA binding", p=1.0e-26), most of which also contribute to the GO term "neuron differentiation" (p=4.4e-18). It is likely that this group of genes is involved in neuron maturation, such as Dlx1, Dlx2[138], as well as Helt, which specifies GABAergic neuron differentiation[139]. Genes responsible for cerebral cortex GABAergic interneuron migration (Lhx6[140], Arx[141], and Fezf2[142]) are also found in this cluster.

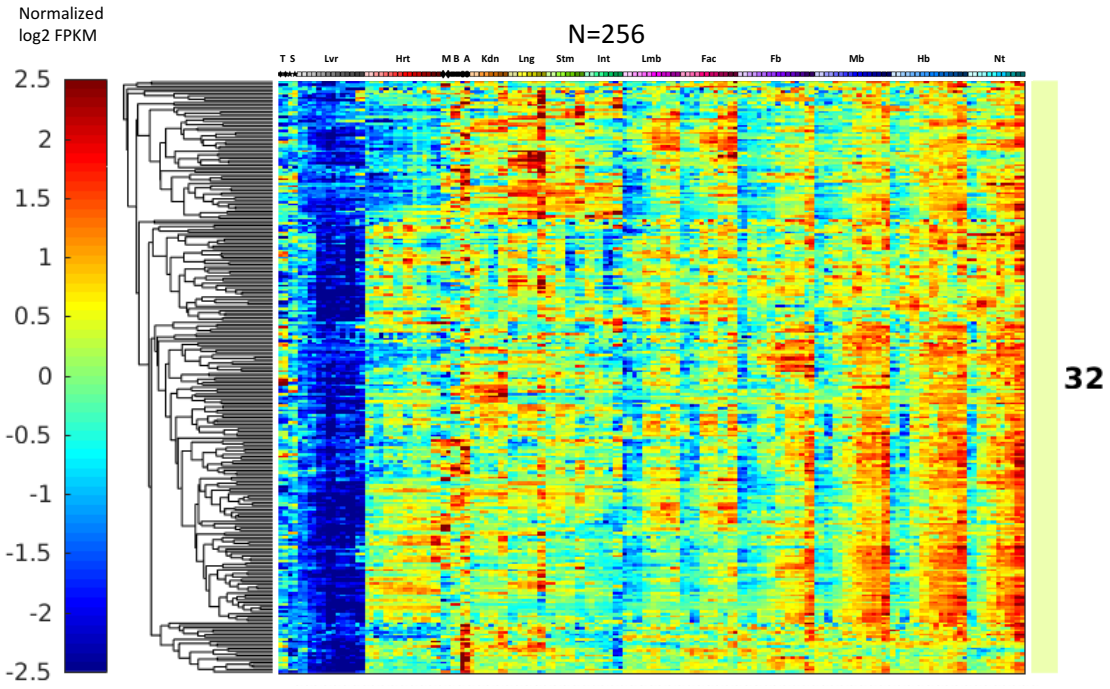


Figure 2.39: Cluster 32 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Genes in Cluster 32 (Figure 2.39) are expressed in nearly all the tissues with increasing trajectories over time, with the notable exception of the liver where they are expressed at very low levels and then decrease. This cluster contributes to the global separation of the CNS (where expression is strongest) from the developing liver. Gene Ontology offered little specific annotation, except “positive regulation of adenylate cyclase activity” (p=1.7e-5).

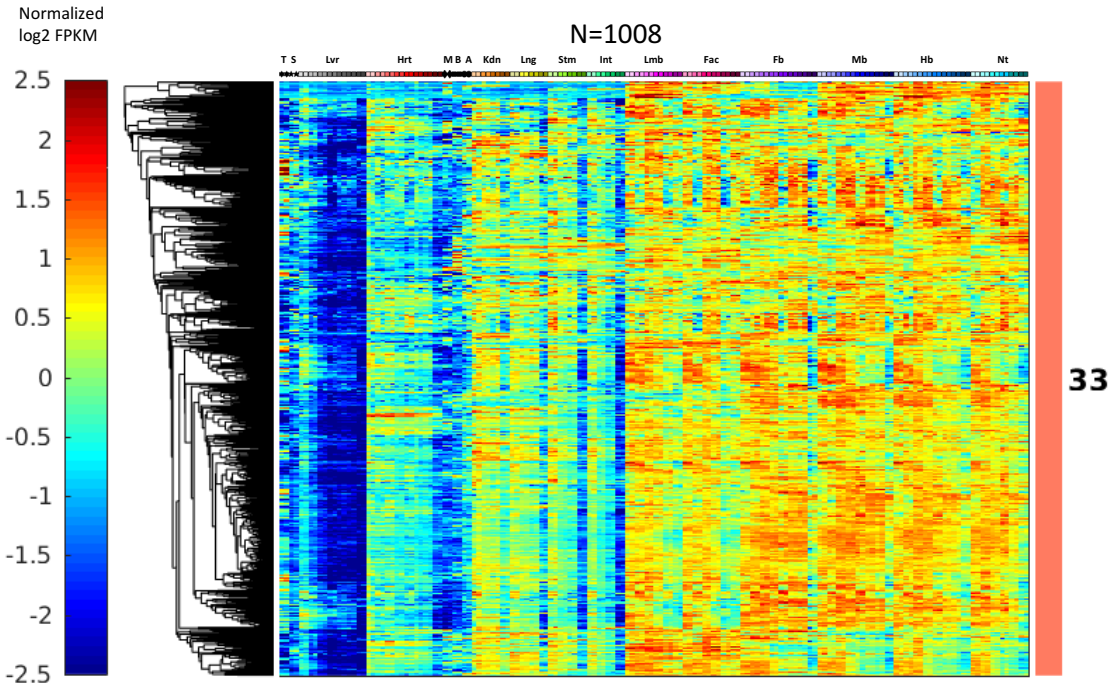


Figure 2.40: Cluster 32 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

Genes in the large Cluster 33 (Figure 2.40) are expressed in most tissues prior to P_0, except liver. CNS, face, and limb are by far the most prominent. Most of these genes are time-course variant. Time courses in different tissues display distinctive trajectories, with decreasing courses being more common, unlike most other major clusters. Thus, genes in this cluster vanish very early in the liver; decrease monotonically in the kidney, lung, stomach, and intestine; and remain constant early and slightly decrease at later stages in the heart, craniofacial prominence, and limb. Gene Ontology enrichment produced three major themes. First, 159 genes (16%) encode DNA binding proteins - especially transcription factors - contributing to “DNA binding” (p=1.7e-17) and “RNA biosynthetic process” (p=7.5e-14). Zinc finger presumptive repressors[143] are especially

prominent (Figure 2.47C). Second, this cluster contains genes regulating different aspects of morphogenetic processes, with enrichment in the term “embryonic morphogenesis” (p=1.9e-12). This is similar to Cluster 20, which also has a broadly decreasing pattern, though it features an emphasis on the Wnt pathway that does not apply to Cluster 33. Finally, significant overlaps of Cluster 33 with cell projection-related genes are called by terms “cell projection organization” (p=1.3e-15), “cilium assembly” (p=1.9e-14), “neuron projection guidance” (p=4.4e-11), and “regulation of nervous system development” (p=4.2e-12). This cluster of genes is different from the cilium-related Cluster 28 in expression dynamics, showing opposite temporal trajectories that argue strongly for distinct regulation. Further studies would be desired to dissect this complex cluster.

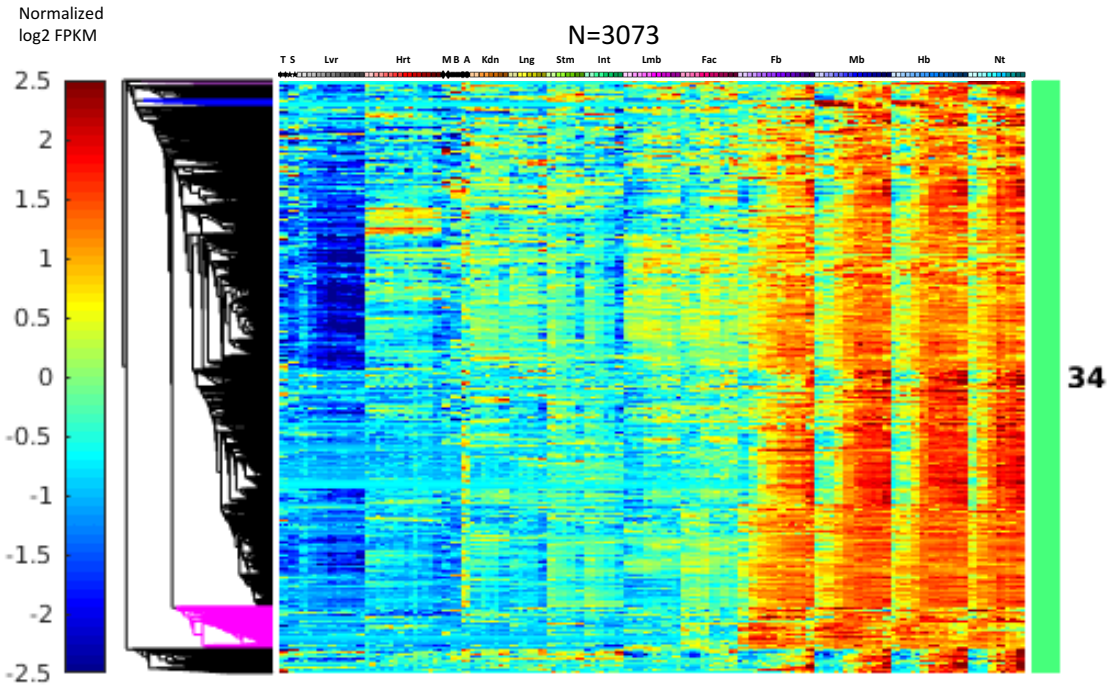


Figure 2.41: Cluster 34 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

The theme of Cluster 34 (Figure 2.41) is expression in all four CNS tissues, with a dominant upward temporal pattern. While most increase over time, they do so with varying kinetics among subclades and between brain regions. Gene Ontology revealed enrichment for a large number of neuron-identity and structure terms associated with neuronal differentiation and maturation, with the most dominant ones being “synapse” (p=1.0e-93), “neuron projection” (p=6.9e-55), “behavior” (p=2.3e-42), and “regulation of nervous system development” (p=2.9e-34). Apart from the central neuronal theme, subclades (colored purple, blue, and pink) differ from each other and from the major neural cluster. All three are significantly enriched with transcription factors and neural development regulators, and they display diverse tissue patterns relative to each other. The small purple clade at the top (next to the blue clade but too small to see clearly) features genes enriched caudally in neural tube and hindbrain. The blue clade below it is enriched in the midbrain and significantly but less so in the hindbrain and neural tube, with overall downward trajectories. The pink clade near the bottom features genes expressed earliest in all four CNS regions, diminishing in later stages.

2.2.3 Additional implications from other expression patterns

Although most of the genes are categorized as members of major clusters, 441 of 15747 differential genes cannot join any major clusters due to their eccentric expression patterns and thus join smaller clusters. Two models may explain this phenomenon. First, these transcripts may be produced due to stochasticity either from random transcription or unknown technical issues in library construction or quantification. Indeed, there are a lot of genes without biologically meaningful names in minor clusters, implying that they

might be biological or technical noise that never robustly showed insightful patterns in existing studies. Secondly, these genes may actually have functions but they do not have a multi-target regulator to bring in other co-regulated targets into the group (unlike Cluster 28 in which Foxj1 regulates multiple cilium-related genes to join that cluster). One interesting minor cluster between Cluster 26 and Cluster 27 contains three beta-defensins in it (Figure 2.42). Defb9, Defb10, and Defb 11 are prominently expressed in the hindbrain and are also detectable in the forebrain. A previous study found Defb10 and Defb11 expressed in adult and neonate brains but barely in other tissues like kidney, while Defb9 was found present in adult hippocampus[144]. Defb11 was also predicted to be related to Alzheimer's disease[145] and was believed to be positively regulated by Tau, a neuronal phosphoprotein responsible for neurofibrillary tangles formation[146]. Although mouse beta defensins had gone through their own recent evolutionary changes[144], it would still be interesting and promising to study how human beta defensins contribute to Alzheimer's disease in human brains[147].

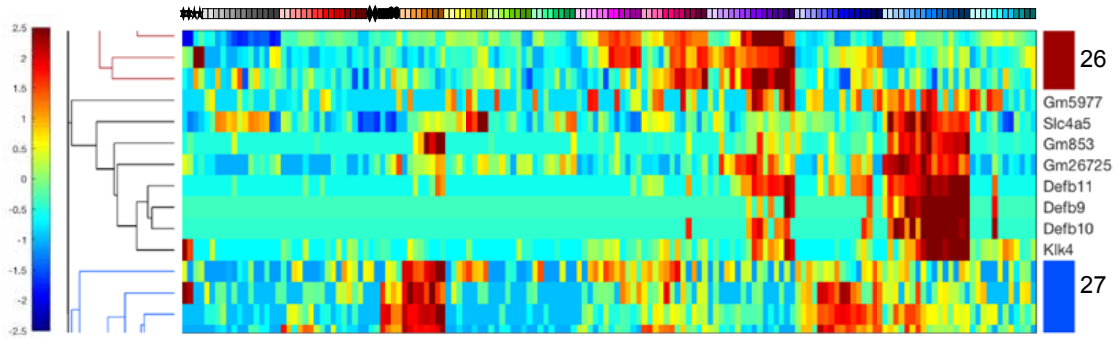


Figure 2.42: Genes between Cluster 26 and Cluster 27 from hierarchical clustering analysis. Sample identities are labeled at the top with the code specified in Figure 2.1.

In addition to these differential genes in or out of major clusters, 9085 genes were more uniformly expressed (Figure 2.3) across identity and temporal domains, which account for 36.6% of genes detected in this study. Although this is relatively lower than in most other studies, I believe it is because the data collection here covered many distinct developmental processes and cell types so that many more differential features were captured. Even so, ENCODE3's collection of mouse tissue samples is not complete either, meaning that more differential features might be discovered in the future. Our current estimation of the percentage of ubiquitous genes is only an upper-bound for the real fraction. Genes currently viewed as ubiquitous in our annotation may turn out to be highly enriched or depleted in a different tissue that was not assayed by ENCODE3. One such example might be mouse testis, since most of the orthologs of human testis-specific genes[16] were not prominently detected in ENCODE3 mouse samples (Figure 2.43).

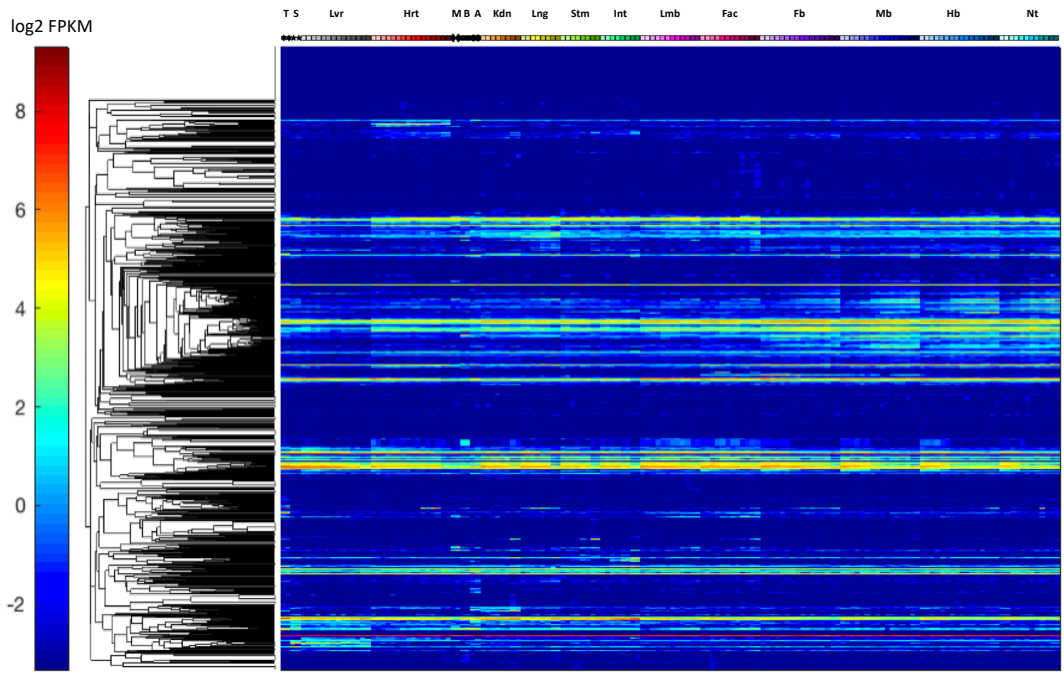


Figure 2.43: Mouse orthologs of human testis-specific genes. Sample identities are labeled at the top with the code specified in Figure 2.1. The heatmap shows Log2-transformed FPKM values based on the color scheme on the left.

Another notable feature of the ubiquitous genes I defined in this thesis is that they generally have a downward trend in terms of abundance (Figure 2.3), although the fold-change threshold was set to be symmetric. This indicates a systematic drift in their measurements. In fact, as mentioned before and visualized in Figure 2.6, much more differential gene clusters show increasing patterns than those with decreasing patterns. Therefore, the transcriptome might become more dominated by differential genes that are tuned up at later time points (Figure 2.43A) in terms of relative abundance. It is possible that overall more differential genes get transcribed as development progresses such that ubiquitous genes just simply get diluted. As a control, when constructing RNA-seq libraries, a constant amount of spike RNAs was added to total RNAs at a fixed ratio. If the dilution hypothesis is true, the spike fraction would have a similar decreasing pattern. However, it has a different pattern (Figure 2.43B). Tracing the quantification of individual spike level also failed to explain the decreasing trend of ubiquitous genes (data not shown). Therefore, the ubiquitous genes are probably not constant. They gradually decrease in multiple tissues. In fact, they also show weak tissue-specificities such as a slightly higher level of mitochondria genes in the heart and a slightly higher level of metabolism-related genes in the liver (data not shown). This challenges the existing view of ubiquitous genes which treats them as constantly expressed house-keeping units and thus calibration references. In fact, “house-keeping” genes may also have biases and drifts because there is no clear border between “house-keeping” jobs and specialized jobs

for a protein to do. But with the stringent 10-fold threshold defining differential genes in this thesis, the analyses of differential transcriptome should be minimally related to house-keeping functions.

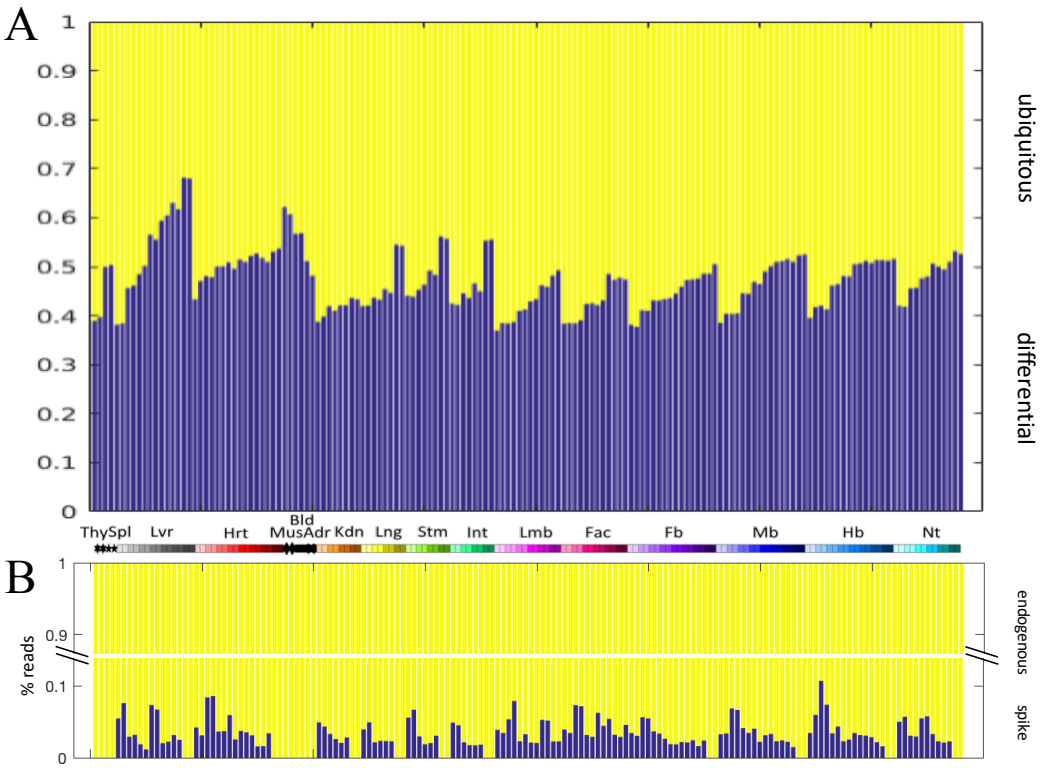


Figure 2.44: Proportions of reads for genes and spikes. (A) The ratio of reads mapped to differential genes over both ubiquitous and differential is visualized as a vertical bar in blue while the yellow part represents the remaining ubiquitous genes' reads. Tissue identities on x-axis are labeled at the bottom, using the color code in Figure 2.1. (B) The fraction of reads mapped to spikes among all the mapped reads is visualized as a vertical bar in blue while the yellow part represents the part of reads mapped to endogenous genes. Tissue identities on x-axis are matching those in (A).

2.2.4 Sources of technical artifacts.

In session 2.2.2, I briefly mentioned various artifacts due to dissection protocols, which can be summarized into six types, namely, muscle contamination (Figure 2.9), thymus contamination (Figure 2.12), incomplete removal of adrenal gland from kidney (Figure

2.13), differences in gut tissues' anatomical definition between labs (Figure 2.18), failure to remove the eyes from craniofacial prominence (Figure 2.25), and differences in neural tube dissection between labs (Figure 2.26 and Figure 2.32). In our ENCODE3 meetings, our lab was the only one that could identify these six types of artifacts, while labs producing chromatin data and DNA methylation data could not. This might reflect the superior sensitivity of RNA measurements compared to DNA and chromatin-based measurements where contamination signals could be diluted by numerous other nuclei with negative signals. This is because a contaminating cell may carry a large number of transcripts for a gene but can only carry two copies of DNA for it. Another explanation is that RNAs are much better annotated, so it is easier to explain strange patterns, while DNA loci are poorly understood, especially those in intergenic regions. But, my finding of experiment-based artifacts does provide important information for analyses of chromatin and DNA methylation data of the matching samples to reach careful conclusions.

A second group of artifacts are specific to either certain replicates in certain tissues (Figure 2.10, 1.22, and 1.24) or certain developmental time points (also correlated with production time) in certain tissues (Figure 2.24, 1.27, 1.29, and 1.30). These artifacts usually come in the form of inconsistent levels of pseudogenes and other low-mappability genes. Computational choreography (how the codes and their dependencies were run and whether the computational environment was consistent) combined with data quality

difference might be one reason, but the real cause is still unknown despite my efforts to try to figure it out before this thesis was due.

A third group of inconsistencies comes from the fact that tissues at E14.5 and P0 were dissected from individual embryos instead of pooled embryos by Wold lab, for numerous logistic reasons. That created a new layer of information: sex. Although it was difficult to determine the gender of early embryos, sex-specific genes still contain that information. Using Xist[148] and Ddx3y[149] as gender markers, I found that pooled samples always contained both genders while individual embryos always had a single clear gender (Figure 2.44).

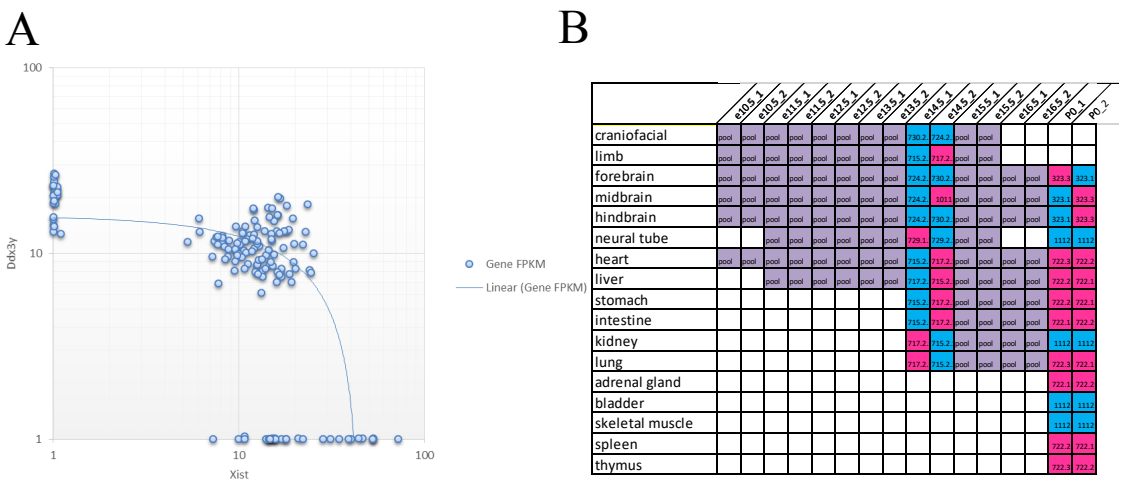


Figure 2.45: Inferred genders. (A) The FPKM values of Ddx3y and Xist in each sample are shown on the scatter plot at log scale with a linear model shown as a curved line. Data points sitting on X-axis are females while those on Y-axis are males. The remaining dots are mixed pools. (B) The gender of each sample is labeled using different colors (pink, female; blue, male; lavender, both). ID's of each individual embryo at E14.5 and P0 are filled in the corresponding boxes.

To further get a global view of batch effects, I used a principal component analysis – canonical correlation analysis (PCA-CCA) approach[61,60], to find a parsimonious relationship between major principal components and metadata. Developmental stages, tissue identities, gender identity based on gender-exclusive genes, and putative thymus contaminated samples defined by thymus-exclusive genes were marked down using Boolean variables (1 or 0). Scores of the top 20 principal components of log-transformed transcript FPKM values together with the Boolean variables were used for canonical correlation analysis (CCA). The results were summarized in Figure 2.46.

Similar to previous PCA results, CCA still ranks the contrast between liver and brains as the most important component (Figure 2.46A). In fact, PC1 strongly negatively contributes to the top 1 canonical variable (U1) (Figure 2.46C), reflecting a differential blood component.

Among the top canonical variables, the third pair corresponds to a batch effect specific to E11.5 and E14.5 samples across multiple tissues (Figure 2.46A and B). Based on the gene expression patterns of high loading genes on both ends of the V score spectrum, I identified a clear depletion pattern at both E11.5 and E14.5, corresponding to Cluster 17, which is likely a product of the early dissection batch effect. In fact, 16 of the top 100 negative loading genes are from Cluster 17 (P = 4.7e-13, Fold Enrichment = 11.75). Probably because the E14.5 specificity variable is also linked to the gender identity

variables that compromise the former, this batch effect isn't profound in terms of E14.5 loadings.

Among our ENCODE mouse tissue samples, subsets of heart, lung, and neural tube samples have shown thymus-specific genes expressed probably due to contamination. In my CCA, the fifth and eighth pairs have strong loadings from the variable labeling thymus contamination in Figure 2.46B. Among all the tissue identity variables, thymus identity is the only outlier that stands out in both pairs of canonical variables, confirming that contamination is not likely from another tissue, but instead, thymus. In fact, a large number of T cell receptor fragments are found in negative loading genes of U5 and positive loading genes of U8, corresponding to negative loading of thymus identity and thymus-contamination variables to V5 and positive loading of them to V8. Since tissues at E14.5 and P0 for RNA-seq assay were dissected from individual embryos instead of pools for samples at other timepoints, potential biases from gender and quality may be present in the samples. These biases are associated with the two variables labeling "male" and "female". I observed that both "male" and "female" variables highly and almost equally contribute to the sixth pair of canonical variables (Figure 2.46B). It is also evident that U6 shows high scores in samples at E14.5 and P0 (Figure 2.46A).

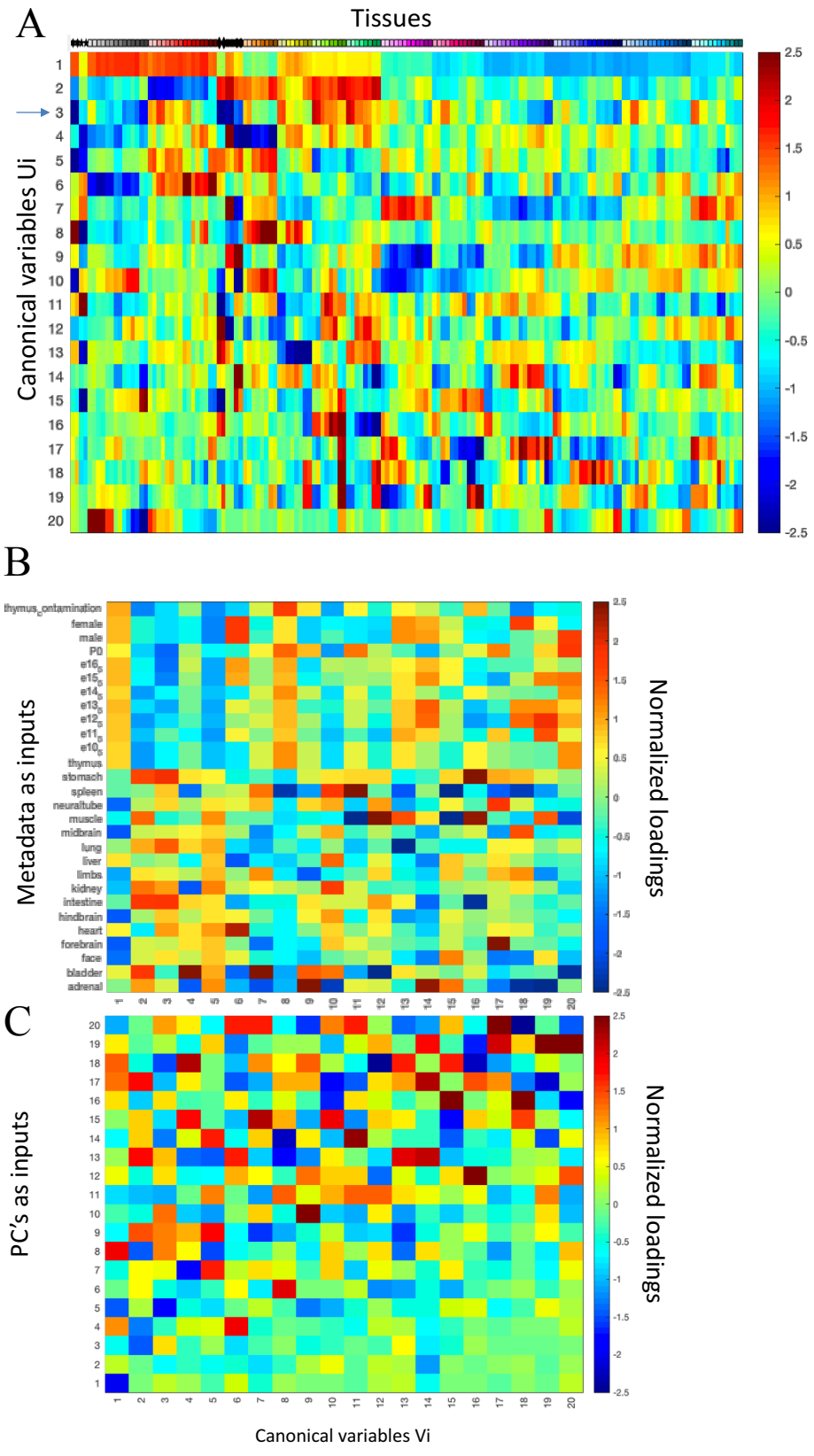

A
Tissues
Canonical variables Ui
B
Metadata as inputs
Normalized loadings
thymus_contamination
female
male
P0
e16.5
e15.5
e14.5
e13.5
e12.5
e11.5
e10.5
thymus
stomach
spleen
neuraltube
muscle
midbrain
lung
liver
limbs
kidney
intestine
hindbrain
heart
forebrain
face
bladder
adrenal
C
PC's as inputs
Normalized loadings
Canonical variables Vi

Figure 2.46: Canonical correlation analysis of whole-tissue RNA-seq samples. Principal component scores of log-transformed gene expressions and metadata information are used as two sets of variables for canonical correlation analysis. (A) Horizontally normalized scores of the set of canonical variables (U scores) corresponding to principal components are plotted based on the color scale shown on the right. Columns indicate tissue and stage identities matching the color code in Fig. 1.1. (B) Vertically normalized loadings of each metadata variable for their corresponding canonical variables (Vi) are plotted based on the color scale shown on the left. (C) Vertically normalized loadings of each PC for their corresponding canonical variables (Ui) are plotted based on the color scale shown on the right.

After scanning through the top loading genes for U6, I identified a strong depletion of negatively loading genes. Interestingly, these genes are highly enriched in blood-related GO terms, indicating that the individual embryo dissections (from Wold lab) may have carried over a smaller amount of blood than the pool dissections.

Among the 18 pairs of canonical variables that have high correlations (Correlation coefficient >0.7), the 18th pair showed a strong difference between the loadings of "male" and "female" variables (Figure 2.45B). By checking the high loading genes to each end, I noticed that male-specific genes (Eif2s3y, Ddx3y, etc.) and female-specific genes (Xist, Tsix, etc.) are respectively contributing to two opposite sides of the canonical variable. However, since the magnitude of loadings quickly drops off as I go down the list of the high-loading genes, and since the "male" and "female" variables are only contributing to the weakest pair (18th) of variables that sits at the edge of our detection range, I regard gender bias as very minimal and only limited to a small set of genes. In fact, the 18th pair is highly contributed by PC 16, which features gender identity.

Overall, the technical artifacts mainly come from dissection issues and unknown inconsistent quantifications of low-mappability genes. Gender difference in individual dissects is a very minor factor, and it only affects a small number of genes.

2.2.5. Mechanisms of transcriptional regulations of differential genes

Transcription factors play a very important role in the regulation of transcription. Based on animalTFDB 2.0[150], there are altogether 1485 transcription factors. These transcription factors also show tissue identity and temporal specificity, similar to but not necessarily in the same way as the global gene pattern (Figure 2.46A and B compared to Figure 2.4B and Figure 2.5A). Different families of transcription factors do not share the same distribution across differential gene clusters. Cluster 33 (Figure 2.40), which is associated with embryonic morphogenesis and neuron projection, is the cluster most significantly enriched with transcription factors, especially C2H2 zinc finger proteins. The biological function of many of these zinc finger factors are still poorly understood. Homeobox transcription factors are not only significantly enriched in Hox gene clusters (Cluster 19 and Cluster 25, Figure 2.26 and Figure 2.32) but also in clusters associated with eye development (Cluster 18, Figure 2.25) and neuron differentiation (Cluster 31, Figure 2.38). This may relate to their general role of axis definition and tissue regionalization, and imply more specialized roles of Homeobox factors. IRF factors are enriched in Cluster 10 that features liver expression and hematopoiesis. E2F factors are highly enriched in the cell-cycle associated cluster (Cluster 21). This is a family known to

regulate cell proliferations. Motif analysis to be mentioned later also confirmed this relationship.

Patterns of RNA co-expression are caused, at least in part, by transcriptional co-regulation. To systematically evaluate how transcription factor (TF) bindings affect gene expression patterns, I focused an initial global exploration on proximal promoters (500bp) by testing each cluster for enrichment of all 718 known consensus TF binding motifs (Figure 2.47A) from CIS-BP database.

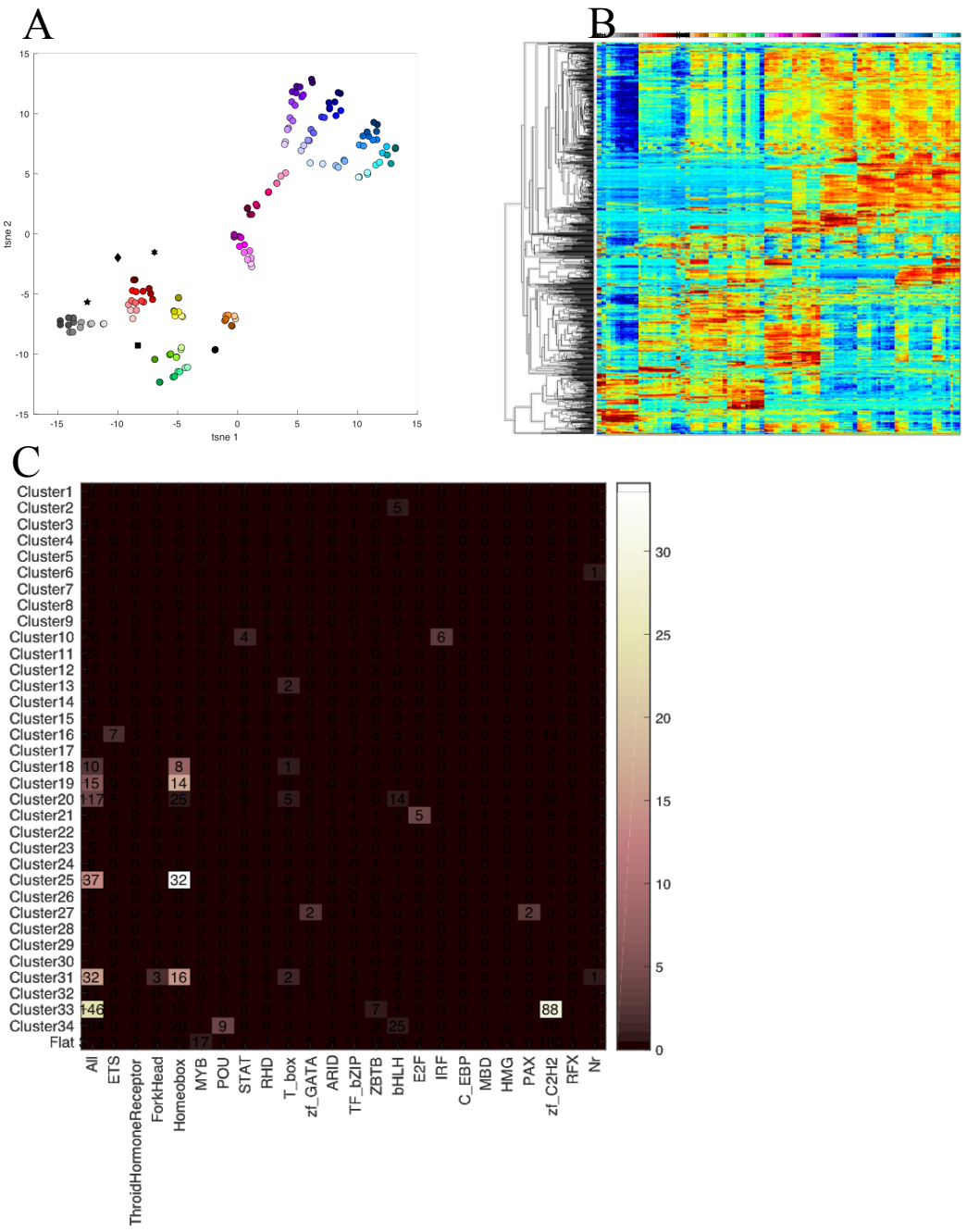


Figure 2.47: Transcription factor expressions in the whole-tissue RNA-seq data. (A) PCA projection of transcription factor expression profiles. (B) One-way hierarchical clustering of transcription factor expressions in whole-tissue data. Tissue identities are labeled with color codes in Fig. 1.1. (C) Counts of transcription factors from each family in individual whole-tissue expression clusters. Colors indicate corrected p-values from hypergeometric test.

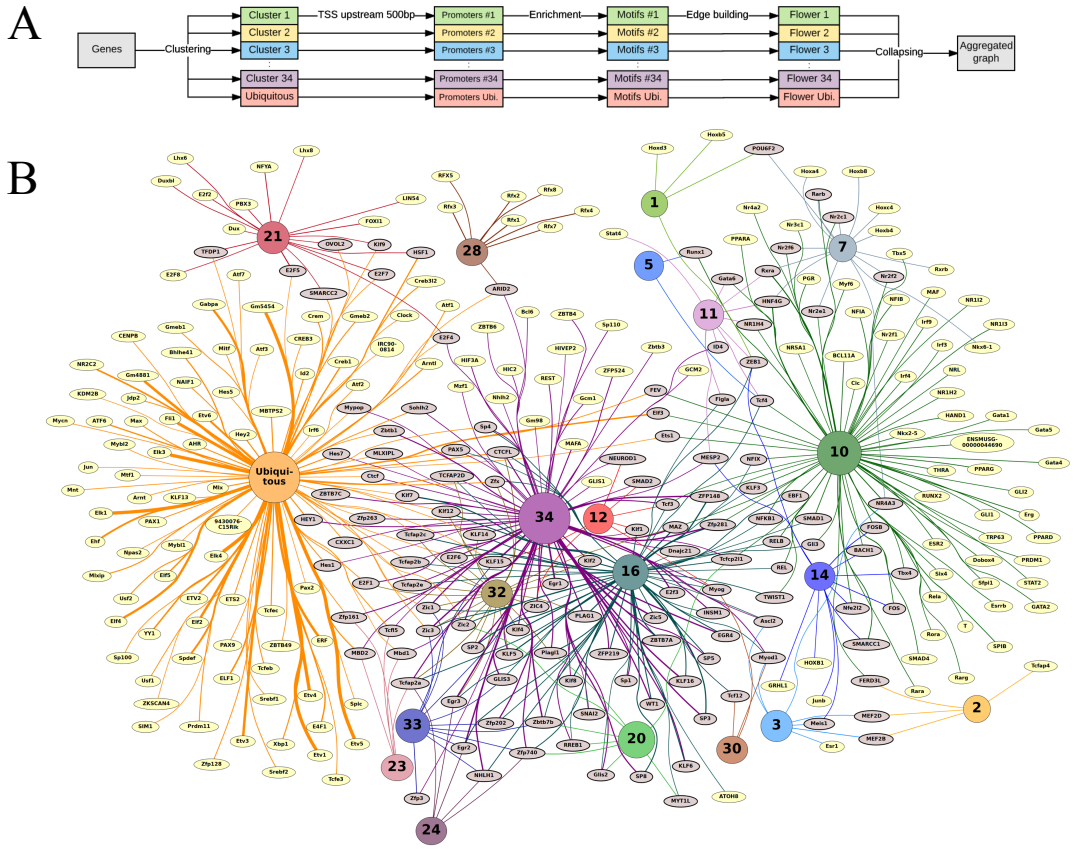


Figure 2.48: DNA Motif analysis of expression cluster proximal regulatory regions. (A) Flowchart of the analysis. TSS upstream sequences for each cluster were used to derive enriched motifs. (B) Bloom graph summary of unique and shared enrichment. Motifs and gene cluster identities are labeled as nodes. Enrichment is visualized as connecting lines between a motif and a gene cluster whose thickness indicates significance. Motifs enriched in more than one cluster are highlighted in gray instead of yellow. The sizes of gene cluster nodes represent the numbers of genes in the corresponding clusters.

The results were organized as a bloom graph, with patterns of motif-sharing between clusters shown as connected, shaded TF nodes or pedals (Figure 2.47B). Thus, one "flower" represents the collection of putative regulators of a single cluster. 307 motifs were significantly enriched in at least one cluster, including biologically sensible TF themes for both large and small gene clusters: fetal liver Cluster 10 (Figure 2.17) has

hematopoietic (GATA, Tal and IRF), hepatic (HNF4a, and Prox1), and bile duct factors (HNF6, HNF1b, Sox9); the highly specific Rfx factor family belongs to its cilium cluster (Cluster 28, Figure 2.35); and the E2F family is prominent in the previously discussed cell cycle-theme Cluster 21 (Figure 2.28).

The bloom graph also showed large-scale patterns: the prominent separation of neurogenesis (Cluster 34) from hematopoiesis (Cluster 10) seen at the RNA level (Figure 2.5) emerged independently with separation of their respective motif use. The most extensive code-sharing was centered on brain-specific Cluster 34, which radiates out shared motifs (shaded nodes) into numerous other clusters that contain genes expressed in both CNS and other tissues. A possible explanation for this CNS-centric pattern is that the involved TFs (and/or their paralogs) were recruited into new networks during evolution to support neuronal diversity. Finally, the ubiquitous-expression cluster produced by far the strongest motif enrichments in the entire transcriptome, mostly caused by ETS family motifs (Figure 2.48B and Figure 2.49D). ETS motif enrichment and occupancy has been associated previously with human housekeeping genes[151], along with several other factors whose motifs are also enriched here, though to lesser extent (e.g. ZFX and CRE).

To understand how ETS proteins are correlated with ubiquitous gene expression levels, I filtered out low abundance and high variance genes and divided the rest into three groups (high, medium, and low) by their abundances (Figure 2.49A). Surprisingly, these three abundance groups showed a similar degree of significance of ETS motif enrichment

(Figure 2.49D). It is possible that ETS proteins are only involved in activating or maintaining the ubiquitous expression pattern but do not fine-tune the absolute levels of each gene. In fact, when compared against GRO-seq data of C2C12 cells quantified in a similar way, these three groups of ubiquitous genes no longer have obvious differences in transcription rate (Figure 2.49B) compared to their mRNA abundances (Figure 2.49A). Notably, mRNA abundances of these genes are negatively related to their 3' UTR length (Figure 2.49C). It is likely that transcripts with longer 3' UTRs naturally harbor more binding sites for RNA degradation apparatus, such as RNAi pathway components and other inhibitory proteins, resulting in lower abundance of mature mRNAs (Figure 2.49B).

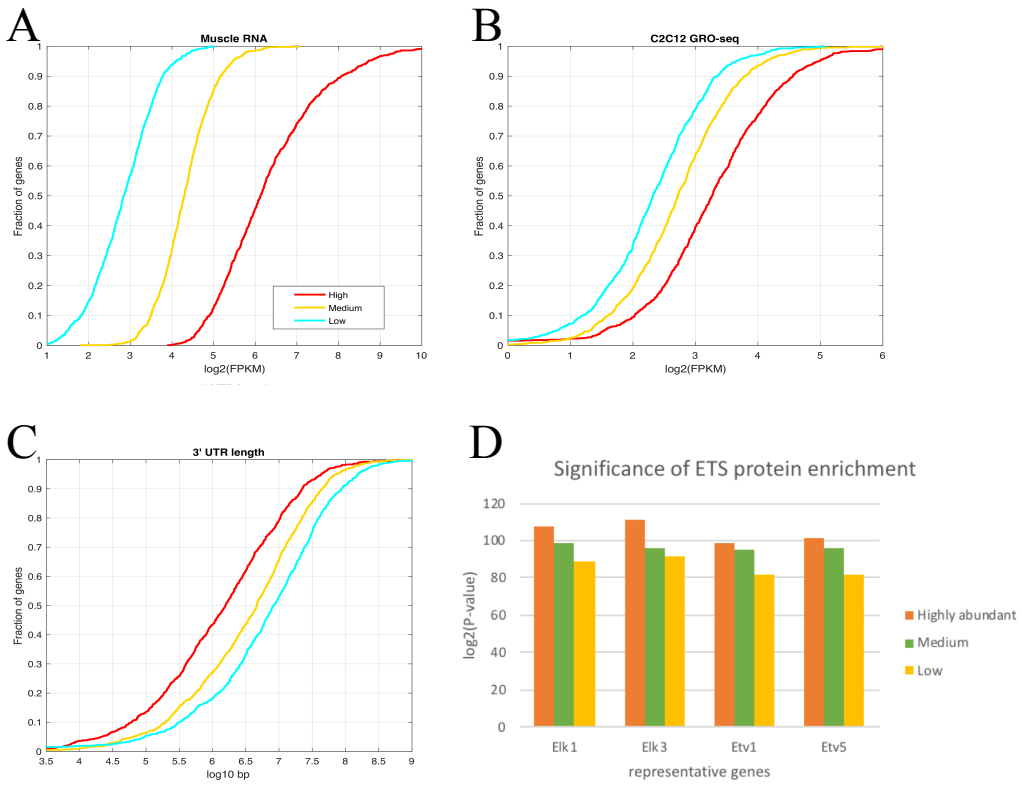


Figure 2.49: Comparison between three groups of ubiquitous genes. Poly-A RNA-seq measurements of skeletal muscle (A), C2C12 GRO-seq data (B), and average 3'UTR length (C), are compared among three groups defined by their abundance measured by RNA-seq in the cumulative distribution function plots. (D) Significance of ETS motif enrichment in the promoters of ubiquitous genes.

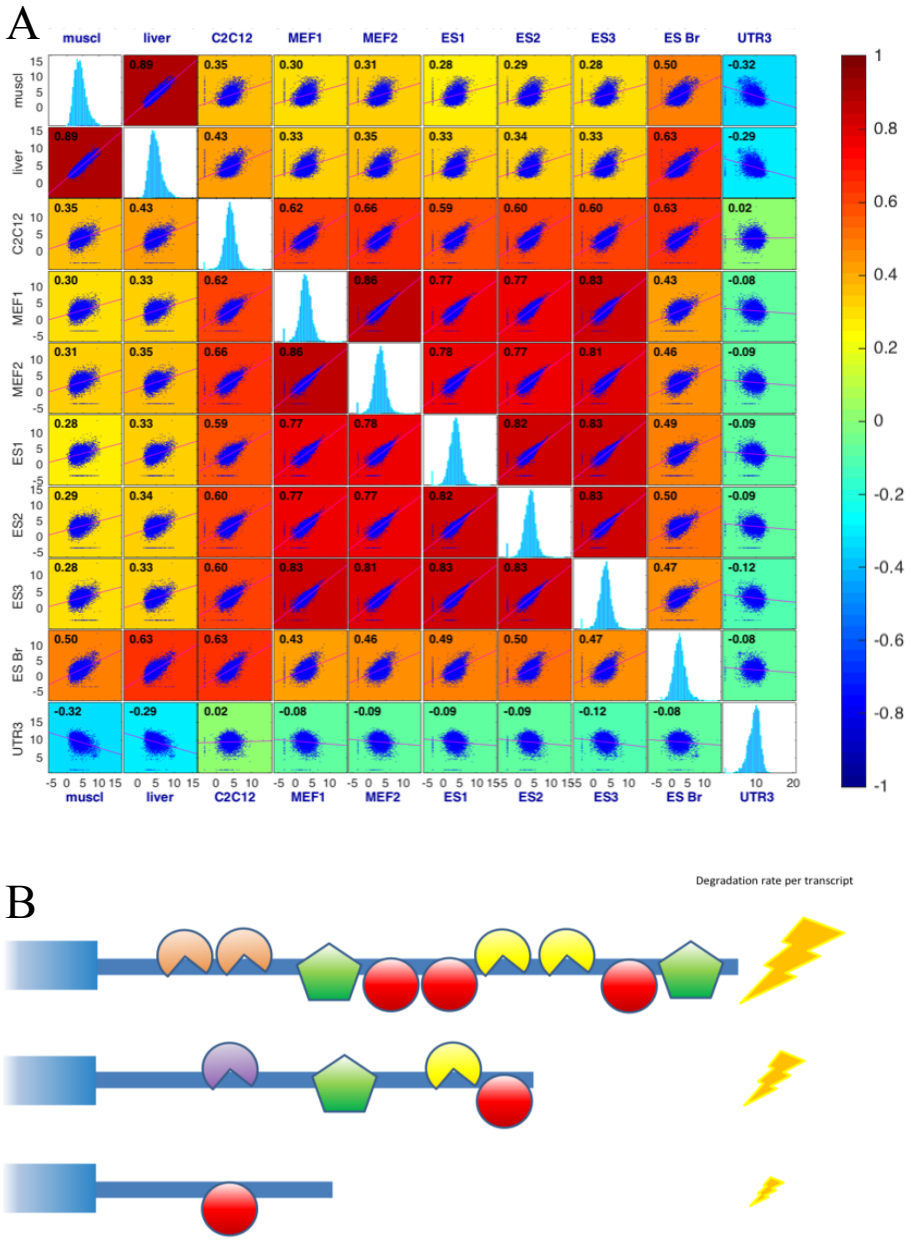


Figure 2.50: Validation of 3' UTR-mediated post-transcriptional regulatory mechanism in multiple samples. (A) Comparisons of 3'UTR length, GRO-seq, Bru-seq, and polyA RNA-seq assays among multiple different samples. Correlation scores between each pair of measurements on the columns and rows are visualized as a heatmap. In the corresponding cell of the comparison, a scatter plot is provided for the pair in comparison. On the diagonal, lined up are histograms of each individual measurement. (B) A model is proposed that longer 3'UTR may harbor more binding sites for RNA-decay apparatus, leading to lower abundance at steady states.

To test reproducibility of this phenomenon, I included more samples of RNA-seq, GRO-seq[152,153], and even Bru-seq[154] for comparison. mRNA abundances of ubiquitous genes are more consistent across different samples than compared to transcription rate measurements (GRO-seq and Bru-seq) and are always negatively associated with 3' UTR lengths, which are independent of transcription rate measurements (Figure 2.50A). These observations indicate that although transcription factors may set the ubiquitously expressed patterns of these groups of genes, the steady-state level of mature messenger RNAs are fine-tuned by their 3' UTR-associated post-transcriptional regulation and slight modulation of transcription rate, independently.

Another way to explore regulatory mechanisms for differential genes is to integrate histone mark distribution over differential gene clusters. Therefore, I processed histone mark ChIP-seq data from ENCODE[155] and calculated their average enrichment over the promoters of the 34 differential gene clusters. This analysis checked average signal profiles of 8 different histone marks (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac1, H3K27me3, and H3K36me3) in 12 tissues at up to 7 stages (E10.5 was removed due to inconsistent ChIP-seq protocol) against the promoters of all 34 clusters. In addition to expected and known trends weakly identified, I discovered a

strong decreasing trend of H3K27me3, a well-known repressive chromatin mark, which anti-correlates with the up-regulated levels of RNA in the CNS-specific cluster, namely, Cluster 34 (Figure 2.41). This strong anticorrelation (decrease in H3K27me3 signal and increase in RNA abundance) is not seen in other tissues with other gene clusters (Figure 2.51A and E), such as liver (Figure 2.51D), despite their similar RNA trajectories. I previously saw that the DNA binding motif for REST[156]/NRSF[157], a known zinc-finger repressor of neuronal genes that decreases in brains (Figure 2.51C), is highly enriched in Cluster 34 (Figure 2.48B). It is likely that REST binds and represses the promoters of genes in Cluster 34.

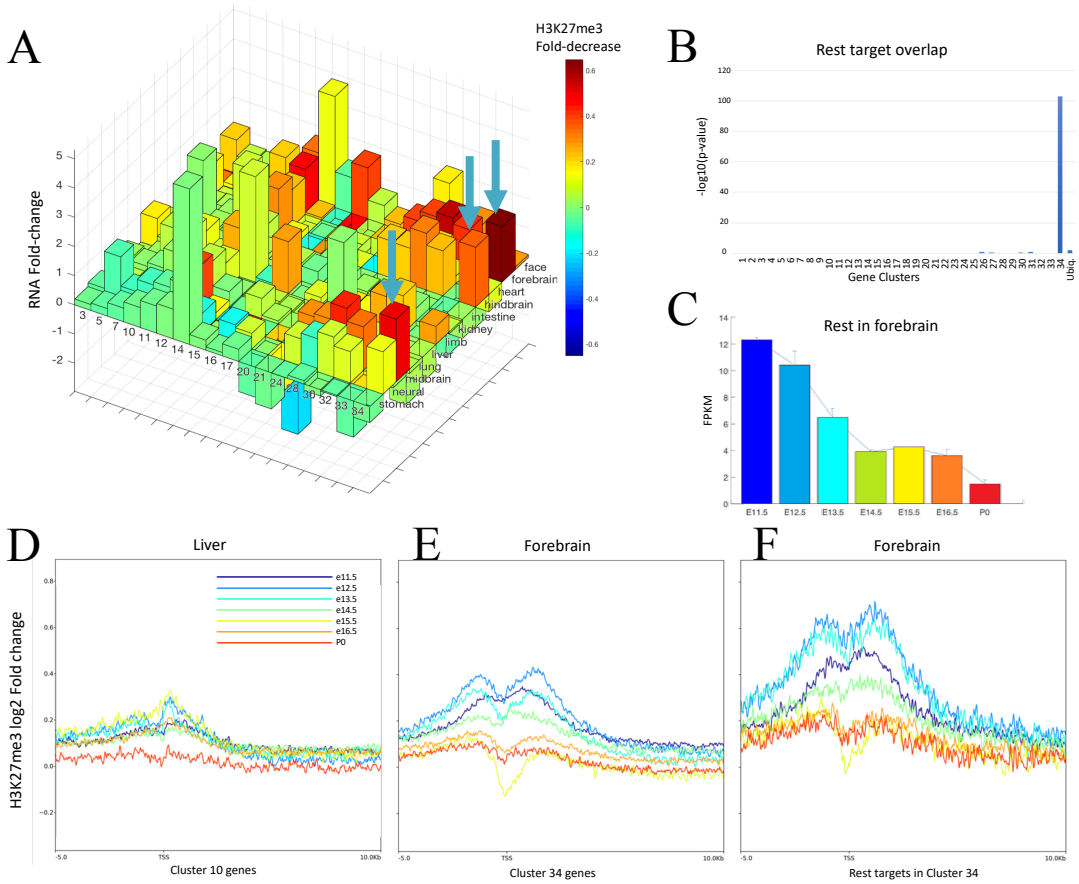


Figure 2.51: CNS-specific genes are associated with Rest binding and de-repression. (A) H3K27me3 fold-decrease and RNA fold-change. Each bar represents a cluster of genes in a tissue type. The height represents RNA increase fold between the earliest and latest time points, while the colors represent H3K27me3 ChIP signal fold decrease. The arrows point to the strongest decrease of H3K27me3 that happens in Cluster 34 in brain samples. (B) NRSF target enrichment in individual clusters. Corrected P values are calculated based on Hypergeometric tests. (C) Abundance of NRSF mRNA in forebrain. (D-F) Averaged H3K27me3 profiles near promoter regions for liver ChIP-seq signals over Cluster 10 genes (D), forebrain ChIP-seq signals over Cluster 34 genes (E), and forebrain ChIP-seq signals over REST-targeted genes in Cluster 34 (F).

Indeed, from ChIP-seq data in another study[158], Cluster 34 is much more significantly enriched with REST-occupied genes than other clusters (Figure 2.51B). Analysis of REST/NRSF-occupied promoters showed even greater early-time H3K27me3 signal (Figure 2.51F). This in vivo brain result contrasts with an embryonic stem cell study in which no H3K27me3 enrichment was found at REST locations[159] , but agrees with a culture-based neural progenitor study[160]. Although the genes in Cluster 34 that are not direct REST targets might be indirectly regulated by REST, it is possible that other repressors also play an important role in regulating CNS-specific genes. Additional repressive contributions might be mediated by other downward repressors such as Snai2, whose expression pattern is highly similar to Rest (they sit in the same clade of three genes in Cluster 16), possibly through a tandem negative feed-forward mechanism mediated by Mir124a[161,162,163,164].

2.3 Discussions

The mouse fetal poly-A RNA matrix captured themes of histogenesis at multiple levels of organization. At the whole-embryo level, universal temporal RNA signatures were evident, while at the individual tissue level, complex expression signatures arose from

shifting proportions of constituent cell lineages, with each lineage maturing at its own pace toward cytodifferentiation. Although using whole-tissue embryo data alone cannot directly deconvolute cell type and lineage contributions to the whole-tissue transcriptome, I was still able to parse the transcriptome by comparing and contrasting expression profiles of different tissues at a wide range of stages assayed by the same protocol. My analyses separated transcriptome into ubiquitous and differential parts, the latter of which was further annotated into 34 major co-expression clusters. These 34 clusters not only showed consistent ontology themes but also demonstrated interesting promoter regulatory codes. Ubiquitous genes possess coherent promoter codes, but post-transcriptional regulation seems to secondarily modulate mRNA levels. Finally, integrating chromatin ChIP-seq data with cluster annotation highlighted a unique de-repression mechanism of CNS-specific genes that are possibly mediated by the repressor REST/NRSF.

Based on transcription "output" measurements, this whole-tissue RNA-seq study uses a divide-and-conquer strategy to parse the transcriptome and to further gain insights on their regulation, guiding follow-up experiments for deeper understanding of the genome. Although this study tries to make the most of the whole-tissue resource and to indirectly infer cellular events, it will miss a lot of interesting details hidden in certain cell

populations with weak saliency or low abundance. In the following chapter, I am going to discuss my single-cell study of tissue development.

2.4 Materials and methods

2.4.1 Whole-tissue RNA-Seq from mouse embryo tissues

Pulverized pooled mouse embryo tissue replicates from timepoints E10.5, E11.5, E12.5, E13.5 E15.5, and E16.5 were received from the Ren lab which supplied these tissues for the entire mouse development project[155]. E14.5 and P0 tissues were dissected from single animals at Caltech. Replicate tissue samples were lysed and extracted using the Ambion mirVana protocol (AM1560). Residual genomic DNA was removed using the Ambion Turbo DNA-free kit (AM1907). Total RNA was quantified with Qubit, and RIN values were collected with the BioAnalyzer Pico RNA kit (5067-1513). The median RIN value was 9.7 (CV=4.4%). Each cDNA library was built using 10 ng total RNA spiked with ERCC spikes (AM4456740) diluted 1:5,000 in UltraPure H2O (InVitrogen 10977023) containing carrier tRNA (AM7119) at 100 ng/uL, RNAse inhibitor (Clontech 2313A) at 1 units/uL and DTT (Promega P1171) at 1mM. cDNA was reverse transcribed and amplified according to the protocol in the SMARTer UltraLow RNA kit for Illumina (634935) using Clontech SMARTScribe reverse transcriptase (639536), and TSO, dT priming, and amplification primers from the Smart-seq2 protocol[30] . The first strand product was cleaned up on Ampure XP beads, and then amplified using the Clontech Advantage 2 PCR kit (639207) with 13 PCR cycles and an extension time of 12 minutes. After a second round of Ampure XP cleanup, the amplified cDNA was quantified on Qubit and the size distribution was checked with the HS DNA BioAnalyzer kit (5067-

4626). cDNA libraries were then tagmented using the Illumina/Nextera DNA prep kit (FC 121-1030) with index tags from Illumina (FC 121-1031), cleaned up with Ampure XP beads, quantified on Qubit, and sized with the Agilent HS DNA kit. Libraries were sequenced on the Illumina HiSeq 2500 as 100 bp single-end reads to 30M aligned reads depth. Inclusion for ENCODE submission required replicate concordance scores by Spearman correlation of FPKM values > 0.9.

2.4.2 Reads mapping and quantification

All the whole-tissue RNA-seq data were processed through the standard ENCODE pipeline (https://www.encodeproject.org/pipelines/ENCPL002LSE/).Downstream analyses were mainly done using MATLAB scripts (https://github.com/brianpenghe/MATLAB-genomics).

2.4.3 Whole-Tissue RNA-seq quality control, PCA, CCA, and Hierarchical clustering

tRNA genes and genes covered by fewer than 10 reads in all tissues were removed. PCA was performed over the log2-transformed FPKM values with 0.1 added as pseudo-counts. CCA was performed on the top 20 PCs' and Boolean variables for tissue identities, thymus contamination, gender identities and stages (17 + 1 + 2 + 8 variables). Dynamic genes were defined as those with at least 10-fold difference (otherwise defined as flat, or ubiquitous) in FPKM values between the most and least abundant RNA samples. Dynamic genes and ubiquitous genes are further categorized into different classes based on gene types (protein-coding, lincRNA, etc.) annotated by GENCODE M4. One-way and two-way hierarchical clustering were done using Pearson correlation coefficient and

average linkage. Clusters were defined by traversing from the root of the tree towards the leaves and splitting out clades with different tissue dominance and GO terms, recognized manually, until no more major clusters could be split out. Clades with at least 30 nodes were defined as major clusters.

2.4.4 Motif analysis

The analysis of transcription factor recognition motifs was carried out using version 4.11.2 of the MEME-SUITE[165]. Motifs annotated in the CIS-BP database[4] (http://cisbp.ccbr.utoronto.ca/) were used to evaluate motif enrichment around the TSSs (500bp stream) of each cluster of similarly expressed genes; enrichment was scored by the AME program in the MEME-SUITE[165]. The analysis was carried out twice based on UCSC mm10 refFlat and GENCODE M4 separately, and only motifs with corrected p-values smaller than 0.01 in both analyses were called significant.

2.4.5 Ubiquitous gene analysis

Among the genes defined ubiquitous by the whole-tissue RNA-seq analysis, those with log2(FPKM+0.1) values no higher than 2 were removed. The 3000 genes with smallest sample variance were equally assigned into high, medium, and low groups based on their average FPKM values.

GRO-seq and Bru-seq reads were mapped and quantified using the ENCODE standard pipeline for computational consistency. Average 3’ UTR lengths for each gene were

extracted from GENCODE M4 annotation. The log2(FPKM+0.1) values and log10(3' UTR length) were used for comparisons.

2.4.6 Histone modification analysis

Histone modification ChIP-seq data were processed using the ENCODE ChIP-seq pipeline (https://www.encodeproject.org/pipelines/ENCPL220NBH/) and Log2 fold change for ChIP-seq samples over input controls were calculated and plotted using Deeptools2.4.1[166] (https://github.com/fidelram/deepTools/tree/2.4.1). To summarize fold decrease of histone modification signals in a specific sample among a specific cluster of genes, a 4kb window enclosing TSS at the center was used and average log2 fold change against input samples were calculated. The fold decrease was the difference between the fold changes of the earliest and latest timepoint. Rest target overlap p value was calculated based on hypergeometric test using iQNP Rest ChIP-seq target list[158].

2.4.7 Gene Ontology Analysis

FuncAssociate[167] 3.0 (http://llama.mshri.on.ca/funcassociate/) was used at its default settings for terms calling.

Chapter 3

SINGLE-CELL TRANSCRIPTOMIC STUDY OF EMBRYONIC FORELIMB

Abstract

Mammalian tissue development is a complex process coordinating dynamic changes in multiple cell lineages. These cell lineages differ in their own ways of differentiation, migration, and turnover pace. To understand mechanisms of these processes, I present here a single-cell transcriptomic study of mouse embryonic forelimb from E10.5 to beyond E13.5. Based on the most variable features of the transcriptome, major cell lineages were identified and tracked, whose differential transcription factor networks highlighted both gradual and Boolean transitions between stages. My result presents the first high-resolution transcriptome atlas of the developing forelimb and provides new insights in regulation of cell state transition in limb development.

3.1 Introduction

The mouse forelimb is a classic model to study tissue development. It is composed of multiple cell lineages including muscle, skin, skeletal, endothelial, and immune lineages. Its rich existing knowledge base and ease to dissect make it an ideal tissue to perform modern high-throughput high-dimensional transcriptom measurements. Single-cell transcriptome data can identify constituent cell-types and states that comprise a complex tissue[168,21,169,170]. For embryogenesis, scRNA-seq further promises to address long-standing questions about the number of transcription states that comprise a given developmental lineage and the nature of transitions between stages. Finally, scRNA-seq

data offer a critical input for gene network modeling by unambiguously assigning to an individual cell (or defined cell group) its transcription factor repertoire and candidate downstream target genes. Major contemporary methods of scRNA-seq have complementary strengths for these purposes, with some assaying modest numbers of cells at relatively high transcript detection efficiency and RNA isoform coverage, while others capture larger cell numbers at lower transcript detection efficiency and without isoform or promoter use information[81]. I present an analysis of ENCODE single-cell RNA-seq resource of the former type for the developing limb. These data identified and parsed major known limb cell lineages and stages within them, revealed how relative contributions change over developmental time, and enabled inference of corresponding differential TF networks.

3.2 Results

3.2.1 Resident and immigrating cell types identified

The whole-tissue transcriptome analysis implied three biological themes of histogenesis that could be further illuminated by high quality single cell data: 1) cell lineage specification and progression to differentiation, 2) immigration of new cell types, and 3) differential proliferation and/or cell death between cell types (Figure 3.1). To capture their transcriptome signatures and assemble underlying regulatory networks, our lab produced 920 high-quality single-cell transcriptome profiles distributed from E10.5 to E15 beyond (Figure 3.3A). Fluidigm C1 single-cell platform was chosen to maximize compatibility with whole-tissue data by using the same RNA-seq biochemistry, sequencing to relatively high depth (~ 1 million reads per cell) to support transcript

detection (Figure 3.2B and C) and coverage. In aggregate, we detected 17,987 protein coding and 1562 lincRNAs, of which 98% and 90% respectively overlap with the limb whole-tissue RNA-seq time-course (Figure 3.2).

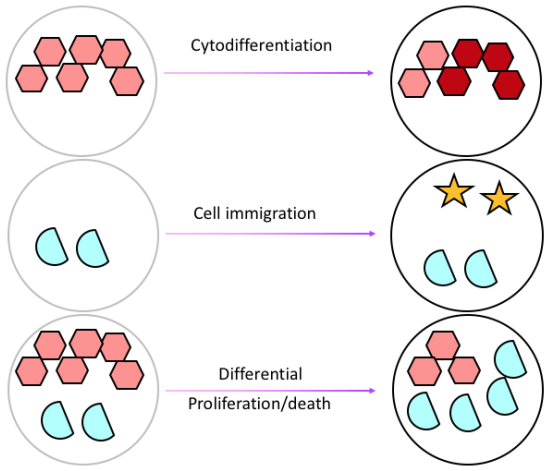
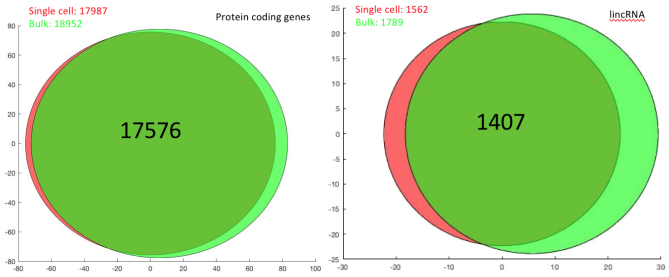


Figure 3.1: Three principles underlying dynamic regulation of gene expressions in whole tissues. Tissues at earlier stages are the circles on the left while later ones are on the right. Smaller shapes represent cells and those in the same color belong to the same type.

To separate putative cell types, I selected the top 1500 high dispersion (variance divided by mean of log2-transformed FPKM values) genes and only kept protein-coding genes coded in the nucleus (rather than the mitochondrium). Two-way clustering of these 1269 remaining genes across all cells (Figure 3.3B) identified 11 cell types of different origins: the chondrogenic/osteogenic lineages and the resident limb mesenchyme from which they arise, the independently immigrating myogenic, monocyte/macrophage, endothelial, and neural crest lineages. Only a few epithelial cells were sampled due to our intentional removal during tissue dissection (see 3.4 Materials and Methods). Cell identity assignments were based on GO enrichment analysis and the developmental literature for

marker genes (Figure 3.4F) of each lineage. In addition to these marker genes, genes involved in proximal-distal patterning of the limb (Hox genes), cell-cycle regulation (such as Ccnb1), and cell signaling (such as Pdgfa and Pdgfra) were also found to be heterogeneously expressed in the forelimb.

A



B

	Psmc
run3 e10.5 median	0.46
run4 e13.5 median	0.45
run5 e11.0 median	0.44
run6 e11.5 median	0.53
run7 e12.5 median	0.64
run8 e13.5 median	0.395
run10 e11.5 median	0.56
run15 e10.5 median	0.585
run16 e11.0 median	0.41
run17 e14.5 median	0.525

C

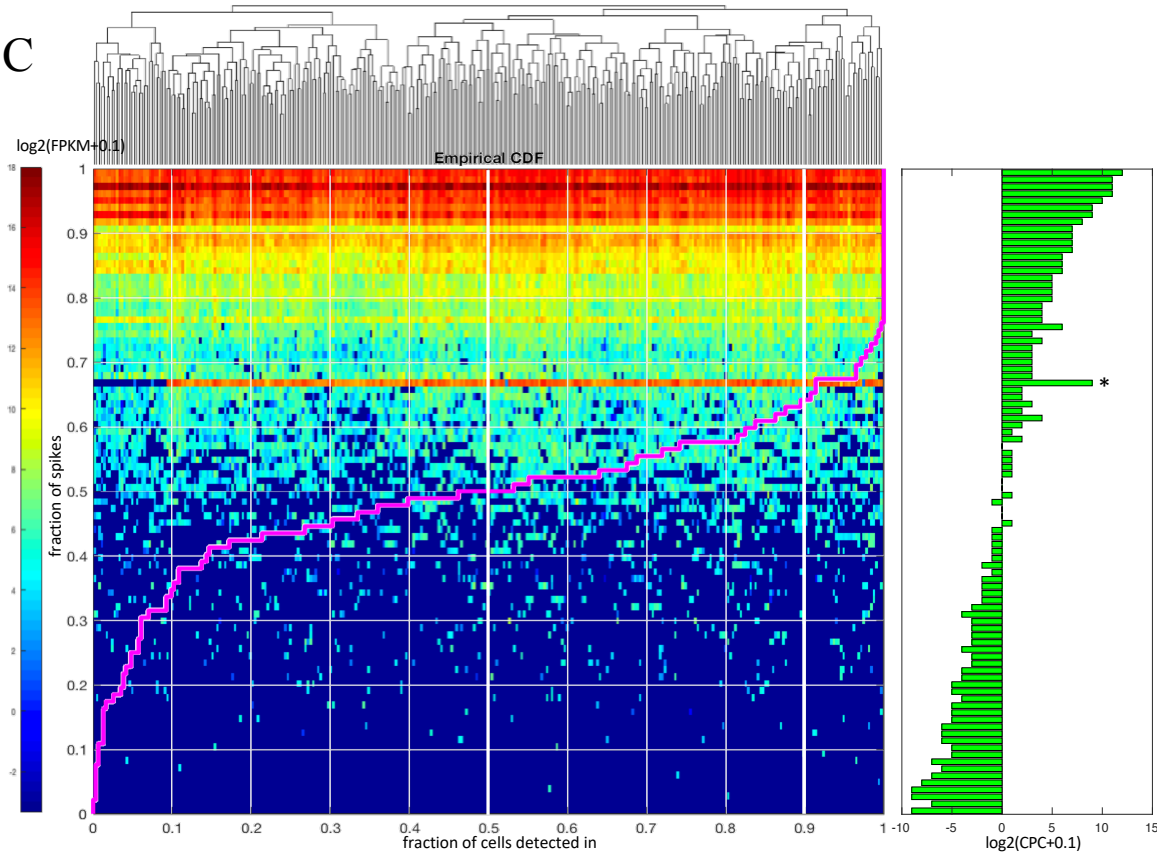


Figure 3.2 Quality metrics of whole-tissue RNA-seq and single-cell RNA-seq. (A) Numbers of protein-coding genes and lincRNAs detected in ENCODE whole limb RNA-seq and single-cell forelimb RNA-seq data. (B) Average single molecule capture probability (psmc) in single-cell limb experiments. (C) Quantification of spike-in

controls. The heatmap shows the RNA-seq FPKM values of spike-in transcripts. Each row represents a spike-in transcript and each column represents one cell. Superimposed on the heatmap is a cumulative curve (magenta) of detection rate for each spike-in transcript. 50% and 90% detection rates are highlighted by white vertical bars on the plot. On the right are the expected number of copies of transcripts of the spike-in controls matching the rows on the left panel.

Multiple replicates confirmed the reproducibility of cell type classifications (Figure 3.3B). Additionally, an initial 10x Genomics platform was used to do 3'-transcript counting assays. Around 50,000 cells were sequenced, and the result was very similar (Figure 3.4A).

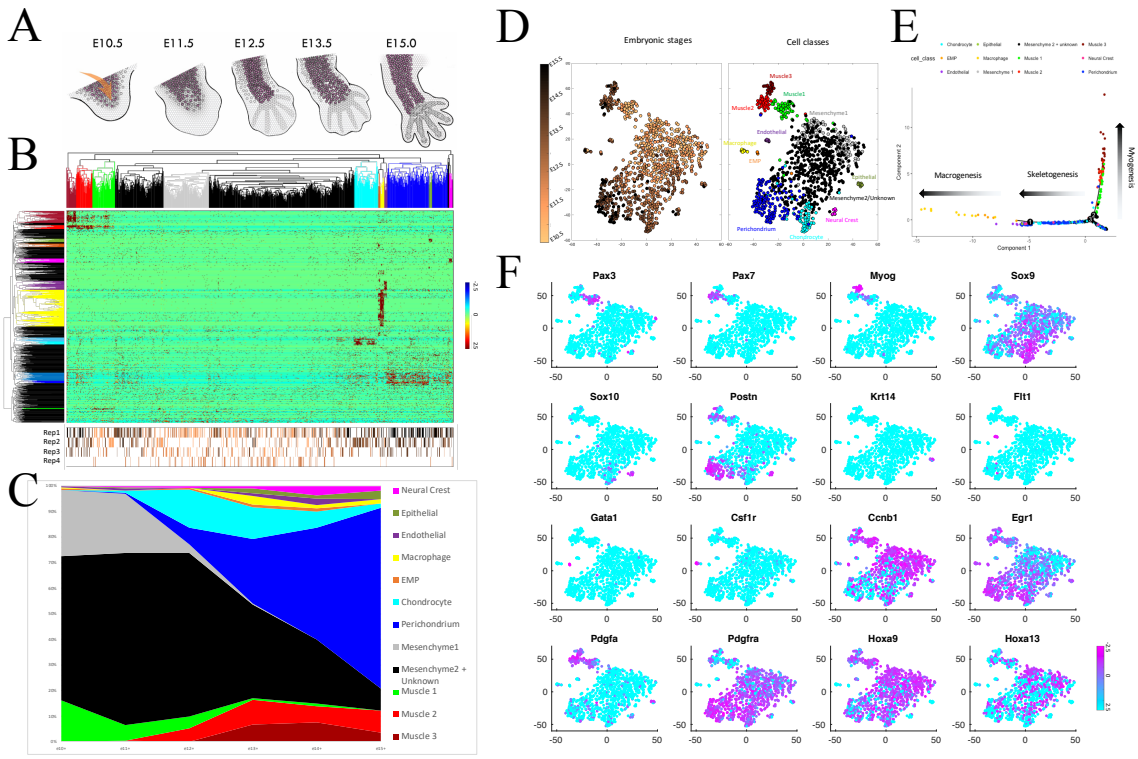


Figure 3.3 Single cell analysis reveals multiple subpopulations bearing distinct gene signatures. (A) Limb development schematic. Arrow indicates immigrating lineages that enter the resident limb bud mesenchyme. (B) Hierarchical clustering of single-cells (horizontal axis) and genes (vertical) performed based on the highest dispersion genes (n=1269). Corresponding tissue batch IDs are labeled below the heatmap with different

colors representing time points using the same color code applied in (D). (C) Cell type composition is plotted as a time series with time as x-axis and percentages as y-axis. The color code corresponds to highlighted cell clusters in (B) and gives provisional cell identity (see text). (D) Individual cells are laid out on 2D t-SNE planes, with the embryonic stages of tissue origin (left) and cell cluster membership color-coded from (B). (E) Monocle lineage inference. (F) Normalized log-transformed expression levels of selected genes are visualized as heatmaps on the t-SNE plane.

The only differences are probably the additional clusters seen in 10X data. An additional muscle cell cluster with exclusive Myot expression was split out from the Muscle 3 cluster (Figure 3.4B) that was not clearly seen in C1 data, probably due to under-sampling of late time points after e14 in C1 experiments. Indeed 4 cells in C1 collection showed Myot expression but ended up in Muscle 3 cluster probably due to low cell number. Epithelial cell type splitted into two clusters in 10X data while equivalent cells in C1 dataset only formed one cluster. A possible explanation is that the ectoderm could not be well removed at later stages and that under-sampling of it in C1 experiments made it hard to identify. Even so, the epithelial cluster of C1 data still contains both subpopulations. Finally, red blood cells formed a new cluster on 10X t-SNE plot. Gene signatures of these cells were only seen in 4 cells in C1 data labeled as different cell types. This means that C1 experiments could not perfectly isolate and capture red blood cells unless they stuck to other cells. Cell size and shape might have been the underlying reason. In all, cell-type classification by C1 data was well reproduced by 10X 3’ tagging.

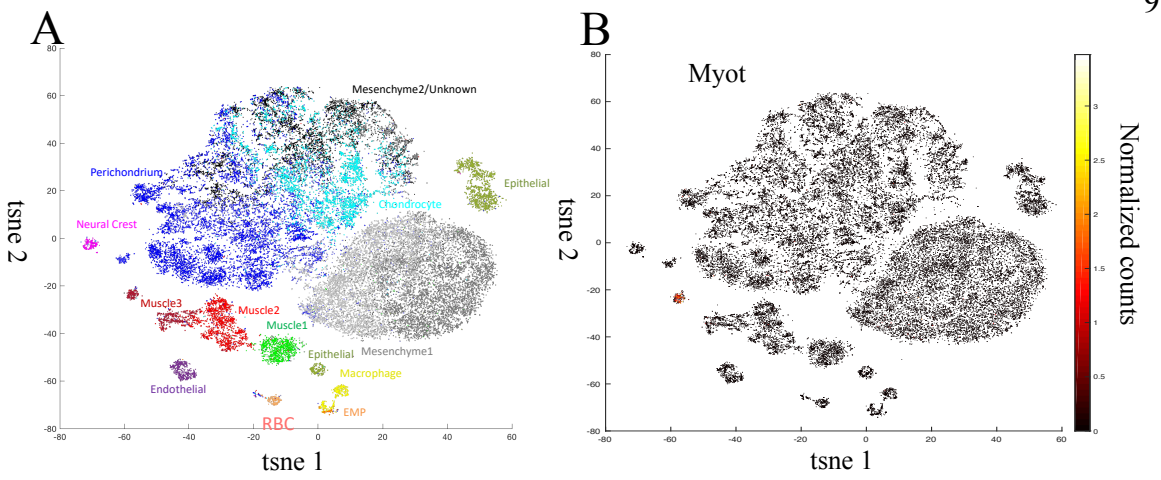


Figure 3.4 3'-transcript tagging method reveals discrete cell types. (A) Individual cells visualized on 2D t-SNE planes. Colors represent cell types. (B) Counts of Myot transcripts. Normalized log-transformed Myot counts are visualized using the color scheme on the right. Cell coordinates on the 2D t-SNE plane are the same as (A).

3.2.2 Developmental lineage progression and gradients of cell-signature salience

Using algorithms such as t-SNE (Figure 3.3D) and Monocle 2 (Figure 3.3E), major developmental trajectories of the limb are highlighted. Early stage samples (E10.5-11.5) are dominated by resident limb mesenchymal cells and early muscle precursors (eMPC) (Mesenchyme1 and Muscle1 cell clusters); while at later timepoints, immature chondrocytes, perichondrial cells, myoblasts (MB), and early myocytes (eMC) (Muscle 2 and Muscle 3 cell clusters, respectively) emerge (Figure 3.3C). After E14.5, perichondrial cells dominate the chondrogenic/osteogenic lineage while myocytes increasingly represent the myogenic lineage (Figure 4C and 3D), although they are likely under-

represented because they fuse into large multinucleate myotubes that are not successfully recovered into the single cell format.

Overall, the inferred lineage relationships were consistent with classical and genetic knockout and modern tracing studies. For example, early myogenic progenitors enter the limb bud from adjacent axial somites, requiring the TF Pax3 for cell migration[171,172,173], and Pax3 emerged as the strongest differential gene defining the Muscle1 cell cluster (p=4.7e-12 by t-test). Together, Monocle and t-SNE defined a tri-partite myogenic progression (Figure 3.3D and 3.3E), whose subdivision was informed by 1) sharply cluster-restricted genes (Figure 3.6A) that split stages (e.g. regulators Pax3, Pax7, Msc, Myog) and by 2) pan-lineage and multi-cluster regulators that joined stages together (e.g. Pitx2/Pitx3, Myod1) (Fig. 3.6A and B). These results confirm at single-cell resolution the prior microarray results from FACS analysis of GFP targeted muscle precursor cells[174,175]. My model with respect to myogenesis shares some overall characteristics with one constructed for in vitro differentiation of human adult muscle[85], but it differs substantially in the stages represented and in many pertinent regulators, reflecting both known and newly highlighted differences between adult in vitro and embryo in vivo systems. The EMP to macrophage relationship is also clearly evident in the data. Skeletogenesis is also prominent, comprising the largest lineage fraction. Condensation, ongoing expansion, and differentiation into cartilage and bone are the primary fate of the resident limb mesenchyme[176,177]. It is represented here in the global models that focus on putative chondrocytes and perichondrial cells, though the lineage structure is less clearly

partitioned and ordered than myogenesis. Substantially deeper cell sampling will likely be needed to develop strong lineage models.

Although cell type and lineage classification can be done based on their most prominent and specific gene expression signatures, a spectrum of different salience was observed. In the limb system, cell-specific signatures increased in complexity as lineages progressed (Figure 3.5). More mature types (e.g. Muscle 3 or Macrophage) were easier to define and detect because their highly salient signatures discriminated both inter- and intra-lineage distinctions (Figure 3.5). However, early cell types and progenitors (Muscle1, EMP, and Mesenchyme1) displayed lower salience, with very few progenitor-unique genes compared to their more differentiated counterparts. Defining such progenitor types was bolstered by their early, and sometimes very low-level, expression of multi-stage and pan-lineage genes (Figure 3.5B).

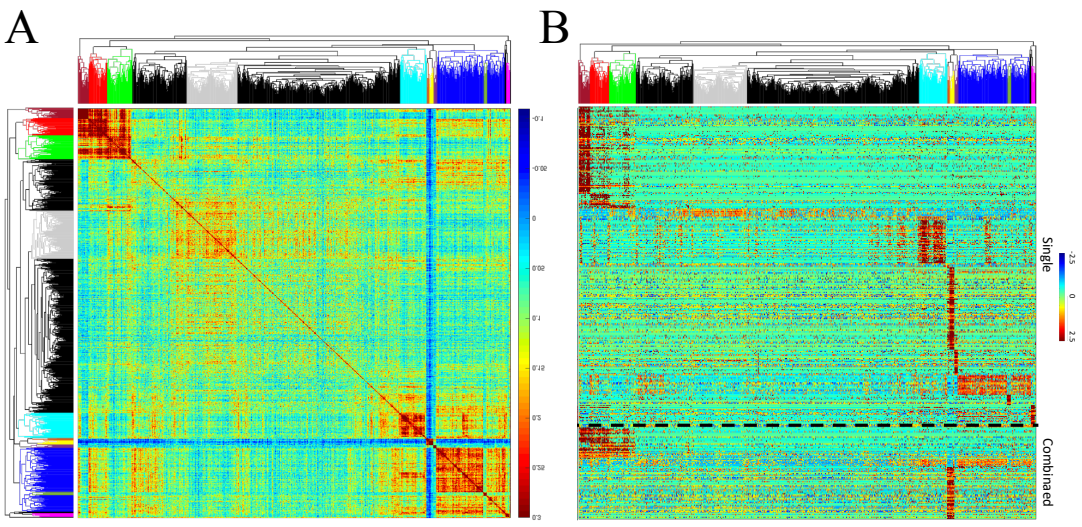


Figure 3.5 Cell-type relationships and salience. (A) Spearman coefficients between different single-cells. Spearman coefficients were calculated from high-dispersion genes

and are visualized by a heatmap. Colored dendrograms from Figure 3.3B are used to label cell types. (B) Boolean markers (not expressed elsewhere) for single and combined cell type(s). Genes specifically enriched in one or selected combinations of multiple cell types are included. Normalized log-transformed FPKM values are visualized as heatmaps.

3.2.3 Lineage networks integrate Boolean versus quantitative regulators

Differentially expressed TF sets for each major cell type and stage were extracted and their known protein and genetic relations used to organize them into interaction networks (Figure 3.6A and C). In the myogenic lineage, some regulators show strong contrast between the cell cluster stages. Pax3 uniquely marked muscle precursor identity (Muscle1) (Figure 3.6B and F) and is replaced by Pax7 which is positive only in the intermediate myoblast-like cell state and then turns off (Figure 3.3F). Myog and Zbtb18 are turned on late in differentiating myocytes (Muscle 3)(Figure 3.6B and F). In contrast to these stark Boolean regulators, more gradual quantitative regulators like Pitx3, Hes6, Myod1, and Ybx3 can be detected early in precursor cells and ramp up gradually through later stages, while Hmgb3 and Hmga1 are gradually reduced (Fig. 3.6A). Boolean-type TF switches were also found in the progression from the EMP state to the pre-macrophage state[82], with Gata1 high in EMPs and undetectable in pre-macrophage cells,

whereas Emr1 showed the opposite pattern, and Runx1 joined the two cell stages while discriminating them from others in the limb.

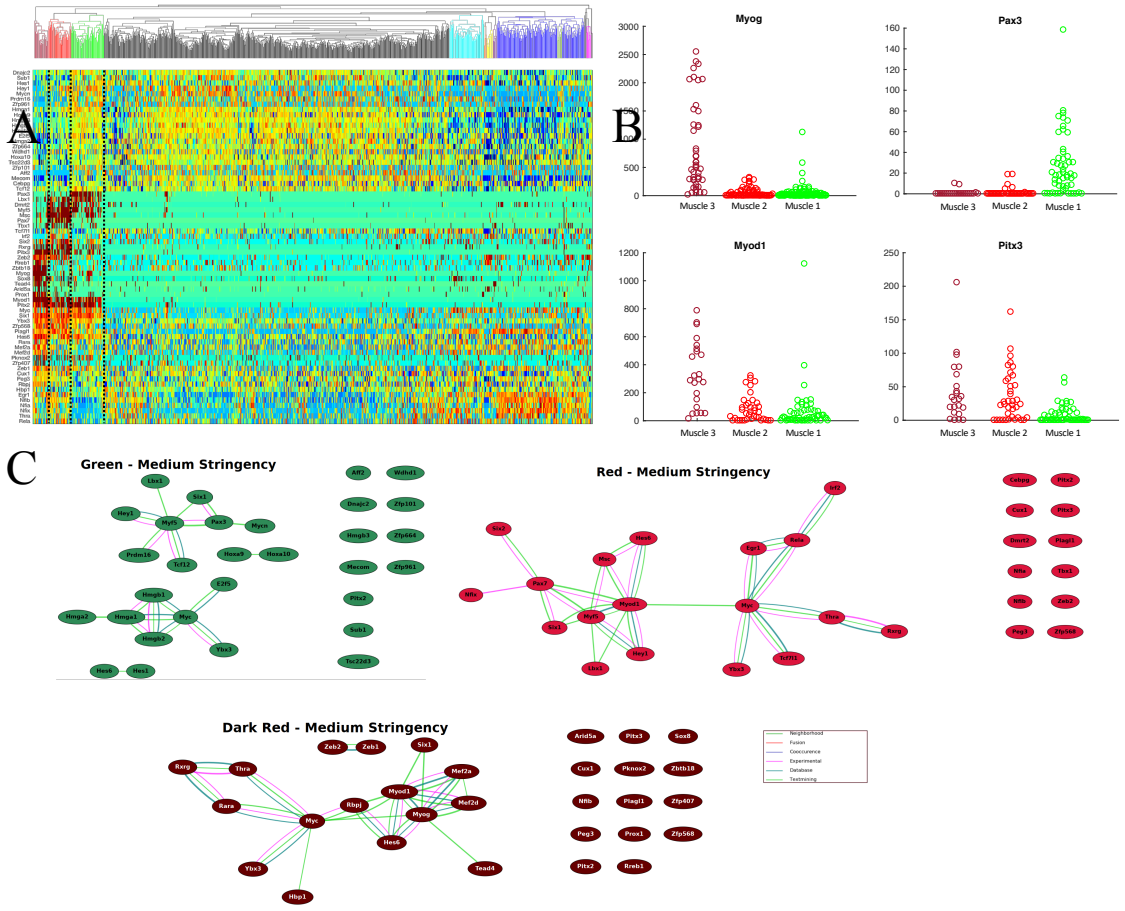


Figure 3.6 Differentially expressed transcription factors. (A) Heatmap of expression levels of the differentially expressed transcription factors. Cell identities labeled on top match those in Figure 3.3B. (B) Distributions of expression levels in distinct muscle states. (C) STRING networks of three muscle states based on enriched differential transcription factors.

The distinction between Boolean and quantitative regulators cannot be seen in a whole-tissue time course because the average profiles of both Boolean and quantitative transitions will look quantitative. Single-cell data here revealed that transcription factors

actually do have different transition patterns. It is possible that downstream targets of quantitative regulators have intrinsic thresholds to translate quantitative regulator concentration into discrete ways of actions and thus lead to discrete cell fates. More work is demanded to better understand this already complex network since Boolean and quantitative factors do not act independently but actually interact with each other based on STRING database[178] (Figure 3.6C).

3.3 Discussions

While the large-scale whole-tissue RNA-seq measures the average profile of mixed cell types that go through differentiation, migration, and unsynchronized turn-over, I demonstrated the power of single-cell RNA-seq analysis that directly decomposed much of this complexity by separating multiple cell lineages and stages for limb development, revealing distinct graded versus Boolean patterns of TF change along individual lineage axes. Lessons about complexity and salience of gene signatures at different developmental stages were also learned from this limb single-cell data, which can be useful for future experiment design and data interpretation.

3.4 Materials and methods

3.4.1 Single-cell transcriptome measurements using the Fluidigm C1

One pair of embryonic forelimbs from a single mouse was used at each timepoint (E10.5, E11.5, E12.5, and E13.5). After dissection from the carcass, limbs were incubated in a 50 uL droplet of a 10% collagenase solution (Worthington LS004202) for 5 minutes at 37 ℃. The limbs were then visualized under a dissecting scope and the ectoderm was removed manually with a pair of #5 Dumont forceps. The mesenchymal core of the limb

bud was then transferred to a 200 uL droplet of Accumax (AM105), and the dish was reincubated for 15 minutes at room temperature. The cells were then manually triturated once with a P200 tip to suspend them, and pipetted into 500 uLs of DMEM + 10% FBS. Limb cells were spun at 500g for 5 minutes at 4℃, resuspended in 500 uLs fresh DMEM + 10% FBS, and passed over a 20 micron mesh (Miltenyi 130-101-812). They were then counted and diluted in DMEM + 10% FBS to achieve a final concentration of 250,000 cells/mL. 12 uLs of this suspension was added to 8 uLs of Fluidigm Cell Suspension Reagent for loading on the Fluidigm IFC (10-17 micron size). Cells were then visually inventoried for doublets and empty chambers, and returned to the C1 for lysis, reverse transcription, and amplification using the SMART-Seq v4 protocol. Lysis buffer: 8.6 uLs water, 1 uLs C1 loading buffer, 2.4 uLs Smart-seq2 oligo(dT) primer (10mM), 2.4 uLs Clontech 10mM dNTPs, 2 uLs ERCC spikes (AM4456740) (diluted 1:40,000 in UltraPure H2O (InVitrogen 10977023) containing carrier tRNA (AM7119) at 200 pg/uL, RNAse inhibitor (Clontech 2313A) at 1 units/uL and DTT (Promega P1171) at 1mM), 0.5 uLs 100mM DTT, 2.6 uLs Clontech single-cell reaction buffer. Reverse transcription reaction: 5.6 uLs Clontech 10X transcription buffer, 0.6 uLs C1 loading buffer, 5.6 uLs Smart-seq2 TSO (10mM), 0.4. uLs Clontech RNAse inhibitor, 2.8 uLs Clontech SMARTScribe. PCR reaction: 4.4 uLs water, 4.5 uLs C1 loading buffer, 75.2 uLs Clontech SeqAmp buffer, 3 uLs Smart-seq2 amplification primers (10 mM) and 2.9 uLs Clontech SeqAmp polymerase.

Amplified cDNA samples were diluted in 10 uLs of C1 DNA dilution reagent, and a 1 uL aliquot of each was quantified on Qubit. 11 samples from the IFC were selected for

BioAnalyzer sizing based on yield and chamber occupancy. An aliquot of the cDNA libraries was diluted to 0.1 – 0.3 ng/uL using C1 Harvest reagent, and the libraries were then tagmented using the Nextera XT DNA sample prep kit (FC 131-1096) and Nextera XT indices (FC 131-1002). After tagmentation and amplification, libraries were pooled, cleaned up twice with Ampure XP beads (0.9X volume), quantified on Qubit, and sized on the BioAnalyzer using the HS DNA kit. The libraries were sequenced as 50 bp single reads to a depth of about 1M aligned reads.

3.4.2 Reads mapping and quantification

Single-cell C1 RNA-seq data were processed through the standard ENCODE pipeline (https://www.encodeproject.org/pipelines/ENCPL002LSE/). Single-cell 10X data were processed by Cell Ranger (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger). Downstream analyses were mainly done using MATLAB scripts (https://github.com/brianpenghe/MATLAB-genomics).

3.4.3 Cell-type and lineage-specific marker genes identification.

To identify cell types, 1500 genes with highest dispersion (variance over mean) of log-transformed FPKM values were selected, and only nuclear protein-coding genes were kept. Then these cells were hierarchically clustered using Spearman correlation. Cell clusters were then manually picked based on enriched genes. Marker genes were defined as those that were significantly enriched against background cell type (Mesenchyme 2) based on one-tailed t-test ($p < 0.00001$) using log-transformed FPKM values. Genes with larger than 3 fold enrichment compared to any other group-average FPKM values were

removed due to low effect size. I did this analysis for each of the cell type other than the background cell type and for combinations of related cell types (Muscle 1+Muscle 2, Muscle 2+Muscle 3, Muscle1-3, Chondrocyte+Perichondrium, EMP+Macrophage). Mesenchyme 2 markers were defined to be those with more than 3 fold enrichment compared to any other group.

Chapter 4

INTEGRATING ENCODE WHOLE-TISSUE DATA AND SINGLE-CELL RNA PROFILES

Abstract

The mammalian genome undergoes robust regulation during development in a cell type-specific manner. Here I present analyses combining single-cell information with ENCODE whole-tissue measurements to gain insights on cell type-specific gene signatures and regulatory architecture. Single-cell measurements captured a lot of features that whole-tissue measurements could not discover even with large-scale combinatorics. Whole-tissue data combined with single-cell RNA-seq also parsed chromatin signatures contributed by high-salience cell types and provided valuable maps of candidate cis-regulatory elements. This chapter demonstrates the broader use of single-cell data for deeper mining of regulatory mechanisms.

4.1 Introduction

The mammalian genome undergoes robust regulation during development. Numerous high-throughput biochemical assays have been developed in the past decade to characterize the input and output of the dynamic genome, most of which are limited to whole-tissue level without single-cell resolution. One exception is RNA-seq, which has been advanced to robust high-throughput and automated fashions with single-cell resolution and has been widely used and understood in recent years. To comment on the advantage of using single-cell RNA-seq measurements compared to traditional whole-

tissue RNA-seq methods, and to leverage the ENCODE resources, I am providing integrative analyses of single-cell RNA-seq data and ENCODE whole-tissue data.

4.2 Results

4.2.1 Single-cell limb signatures parse whole-tissue limb cell compositions

Since the whole-tissue RNA-seq data in Chapter 2 and single-cell RNA-seq C1 data in Chapter 3 were both generated using protocols based on SMART-seq2[10], a direct comparison between them would be reasonable. Due to the fact that ENCODE limb samples contain both the forelimb and hindlimb and do not match our single-cell data which were derived only from forelimb, our lab also built whole-forelimb RNA-seq data to assess the difference.

Using the high-dispersion nuclear protein-coding genes derived in Chapter 3 as an input, CIBERSORT[179], a method for estimating cell composition of complex tissues, was able to infer the fraction of each known cell type in the mixed-cell tissues (Figure 4.1). In Chapter 3, I talked about the fact that late-stage muscle cells might be under-represented in single-cell data because they tend to fuse into large multinucleate myotubes that are hard to dissociate.

A B

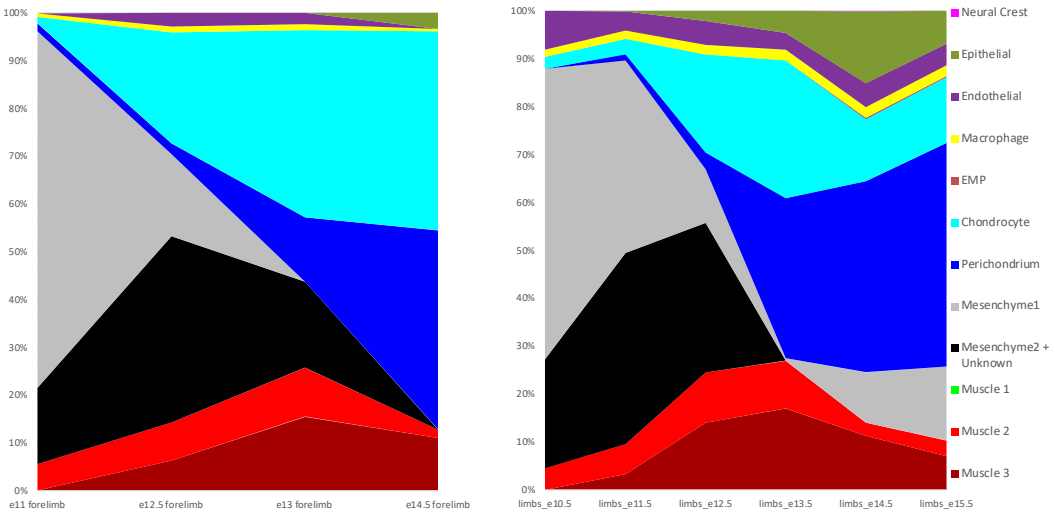


Figure 4.1 Cell composition of limb tissues inferred by CIBERSORT. (A) Forelimb whole-tissue cell composition inferred from single-cell signatures. X-axis denotes development stages while the height along Y-axis represents percentage of cells for certain cell types. (B) Cell composition of ENCODE mixed limb tissues inferred from single-cell signatures visualized in a similar way to (A).

That might partially explain the drop in empirical fraction of myoblast and myocyte (Figure 3.3C). However, based on CIBERSORT estimation, even in whole-tissue data myoblast and myocyte also drop after E13.5 (Figure 4.1). This implies that the decrease fraction of mature cell types in late-stage forelimb tissue may not be an artifact of single-cell isolation protocol but is probably due to differential turn-over rate between lineages. Muscle lineage propagation may not catch up with the rapid production of cell types in other lineages, such as perichondrial cells from mesenchyme differentiation. Therefore, the relative abundance of muscles cells gradually decreases, although the absolute quantity is probably increasing as the forelimb grows bigger.

Based on CIBERSORT's estimation, the single-cell dataset has roughly the same stoichiometry of cell types compared to the forelimb whole-tissue dataset, indicating

good sampling of major types of cells. The only discrepancy lies in minor cell type identification, especially those with weak gene signature salience (Muscle 1 and EMP cells). Those cell types do not have enough exclusive marker genes (Figure 3.5B) for recognition of their presence and have only a small number of samples (cells).

Furthermore, I performed a similar analysis for ENCODE mixed limb RNA-seq data. Overall, mixed forelimbs and hindlimbs did not show a dramatically different composition. Major differences include a higher percentage of "Epithelial" and "Mesenchyme 1" cells at late stages for ENCODE mixed dissections, partially due to their different dissection method.

Overall, using independent single-cell data, the cell composition of a tissue can be estimated from its whole-tissue RNA profile. But it has to be kept in mind that, although transcripts from individual cells are always extracted and processed together for whole-tissue assays, which assumes that a linear average of them are finally sampled in the whole-tissue RNA-seq library, the library construction process amplifies them with exponential noise, random dropout, and other types of biases, making it hard to robustly infer cell composition, especially for low-salience cell types.

4.2.2 Local versus global identities

Although limb tissue has a unique structure and function compared to other parts of the body, its cell types as building blocks are mostly also seen in other tissues. This implies that if enough RNA-seq data from different tissues at different stages have been

collected, by comparing them to each other, it might be possible to identify a lot of cell type-signatures as minimal modules. That has been the goal of the study in Chapter 2. Indeed multiple modules were identified. These modules contain genes with similar expression patterns, corresponding to collinearity patterns in the high dimensional space, which can thus be transformed by PCA to capture major features (principal components).

To understand whether the single-cell limb gene signatures were already identified in the whole-tissue data, I mapped the marker genes of each cell type onto the differential cluster map of whole-tissue data, and plotted the distribution of their coordinates on the map as violin plots (Figure 4.2). Marker genes of EMP and macrophage cells mainly concentrate on Cluster 10, which features hematopoiesis. Although these cells have a low fraction in the limb tissue, they have a large number of prominent markers and broad distribution across multiple tissues. Therefore, whole-tissue RNA-seq data already identified Cluster 10 to represent this signature. Similarly but less prominent, Muscle 3 signature overlaps with Cluster 2 (skeletal muscle) and Cluster 3 (cardiac muscle) featuring myogenesis, while the perichondrial signature overlaps with Cluster 16 featuring the extracellular matrix. Interestingly, this analysis also captured the fact that Mesenchyme 1, Muscle 1, and EMP cells as progenitor cell types have a slight overlap with Cluster 21 which represents cell cycle regulation, consistent with their active proliferating feature. However, none of the low-salience cell types had their cell type-exclusive signature discovered by the whole-tissue cluster analysis. Their marker genes are usually dispersed across multiple clusters due to the fact of the multi-tasking nature of many of these genes, that is, they also have a variety of expression and functional

features in other tissues and thus get assigned to different places. An independent CCA was also performed to look for linear combinations of whole-tissue profiles that can explain cell type-specific signatures in the limb, but did not find any evidence for low-salience signatures.

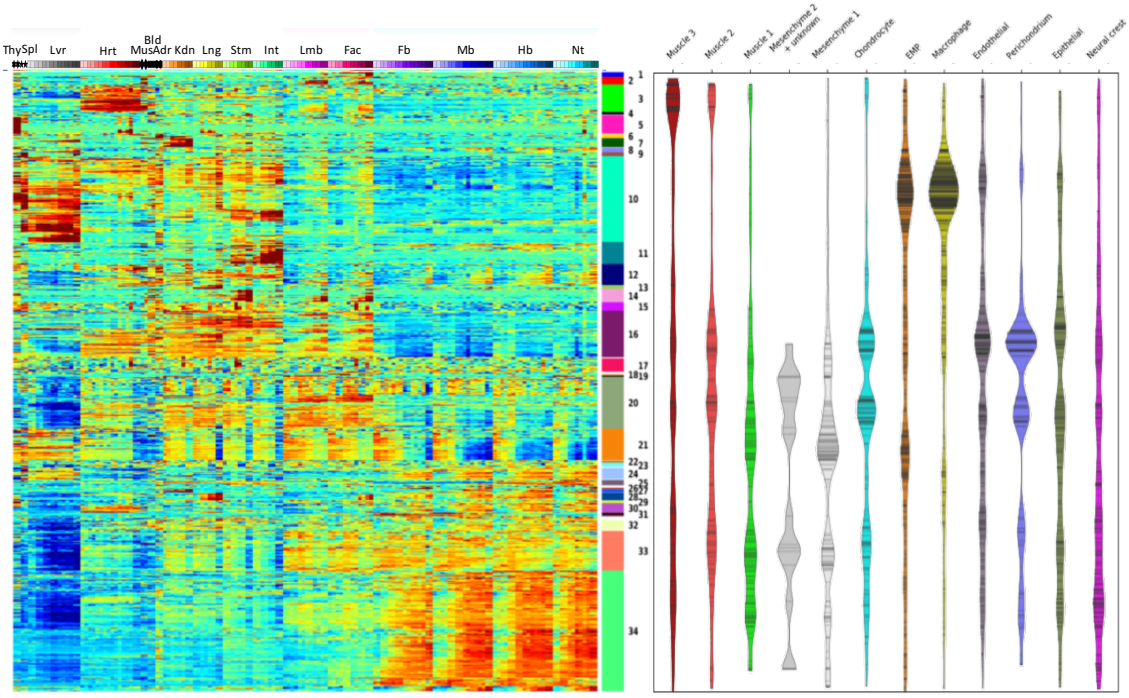


Figure 4.2 Cell type-specific markers mapped onto the bulk clustergram. Cell type-specific markers mapped onto whole-tissue clustergram. Each cell type has its markers positioned at corresponding y-axis position on the right, aligned with genes in the whole-tissue clustergram on the left. The distributions of these positions are smoothed into violin plots for better visualization.

These analyses show that single-cell RNA-seq has irreplaceable power to find features of certain cell populations that cannot be recognized by whole-tissue measurements, due to representation and salience issues. It also showed that defining a “marker gene” or a multi-gene cell-type signature is inevitably context dependent. Context ambiguity also

affected cell-type markers first defined elsewhere. For example, fate mapping studies have shown that some early mesenchyme, including a subset of Myf5 expressing cells, become fat rather than muscle[180], but expression of a prominent brown fat discriminator from that study (Prdm16) was widespread across all cell types in the developing limb context (data not shown).

4.2.3 scRNA-seq data deconvolve candidate cis-regulatory elements by cell type

The ENCODE histone mark, chromatin accessibility, and DNA methylation data provide rich biochemical signatures from which candidate cis-acting regulatory elements (cCRE) have been computationally inferred at the whole-tissue level[181,182]. However, the resulting cCREs are, for each tissue, an unresolved admixture of chromatin from all cell types of the tissue. To begin to parse cis-elements by cell type, as illustrated by the diagram in Figure 4.3, I merged limb DNase hypersensitive sites (DHSs) from ENCODE as candidate regulatory elements, associated each of them with the nearest promoter of an expressed gene, and then integrated them with a whole-tissue epigenomic regulatory state model[181,182] (IDEAS, see Methods) to identify active elements. Finally, the active elements that associate with cell type or lineage-specific marker genes were extracted. Overall, I was able to assign 4631 candidate active and poised elements to individual limb cell types and lineages. It was encouraging that DHS and chromatin marks from whole limb tissue had the sensitivity to identify known validated enhancers specific for cell clusters comprising less than 10% of the starting tissue (notably Muscle 3 and Macrophage, Figure 4.4A and C). Previously validated cCREs were affiliated with these and other cell types (Figure 4.4A and B). Of 405 cell-type and lineage-specific genes,

385 (95%) had at least one affiliated active or poised element (median 6 elements per gene), and all limb lineages and clusters had detectable elements. Modeling cell-type/state-preferential cCRE in this way should now be possible for any ENCODE matrix tissue with the addition of its corresponding sc-RNA data and its integration with IDEAS or other epigenomic state models.

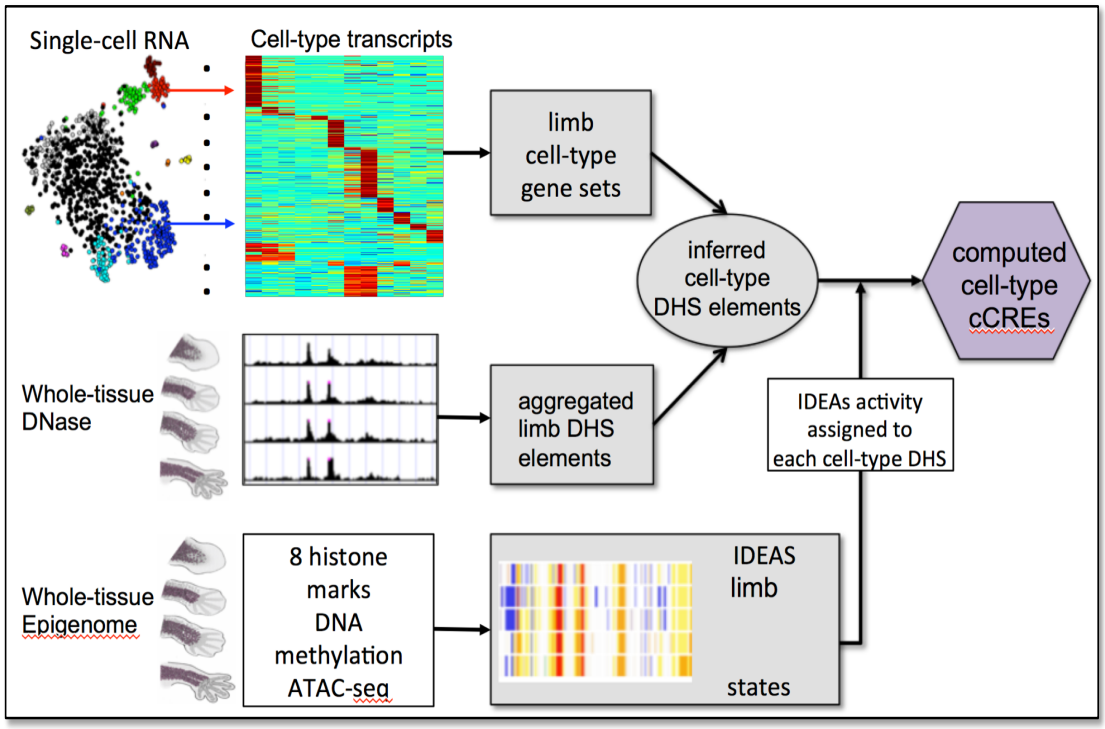


Figure 4.3 Assigning candidate cis-regulatory elements to cell types. Three types of data on the left are used to derive the final model of cell-type cCREs on the right. Arrows represent data processing flow.

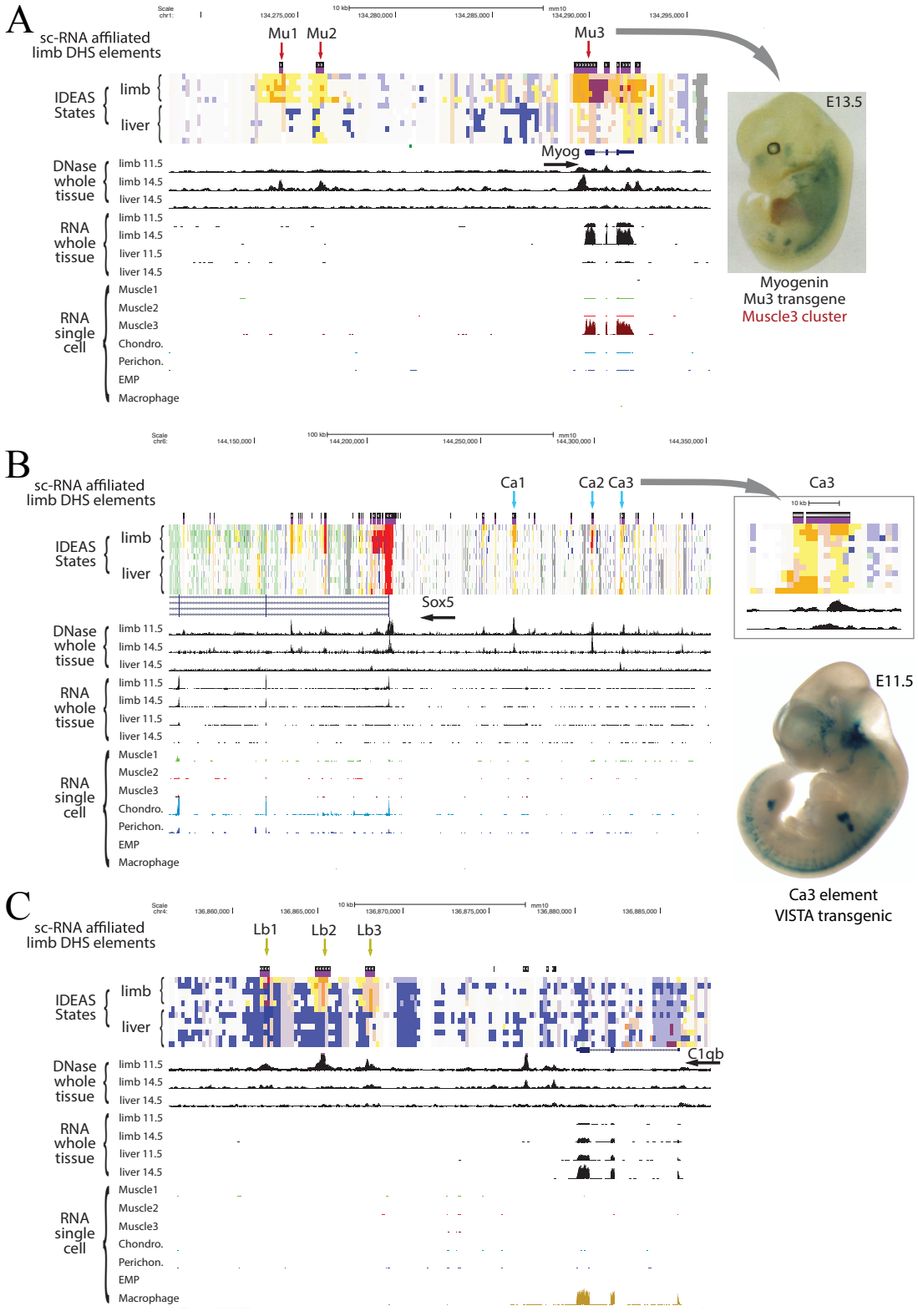

A
sc-RNA affiliated limb DHS elements
Mu1
Mu2
Mu3
IDEAS States
limb
liver
Myog
DNase whole tissue
limb 11.5
limb 14.5
liver 14.5
RNA whole tissue
limb 11.5
limb 14.5
liver 11.5
liver 14.5
RNA single cell
Muscle1
Muscle2
Muscle3
Chondro.
Perichon.
EMP
Macrophage
E13.5
Myogenin
Mu3 transgene
Muscle3 cluster
B
sc-RNA affiliated limb DHS elements
Ca1
Ca2 Ca3
Ca3
Sox5
E11.5
Ca3 element
VISTA transgenic
C
sc-RNA affiliated limb DHS elements
Lb1
Lb2
Lb3
C1qb

Figure 4.4 Representative cell type-specific cis-regulatory elements. (A) Computationally predicted limb myocyte cCREs. Active or bivalent cCREs are labeled with horizontal black bars on top, the cell-type specific (exclusive) ones of which are labeled in purple below. IDEAS states are shown below the cCREs bars. Red represents active promoters. Orange represents active distal enhancers. Blue and purple represent inactive regions. Mixed colors represent intermediate states. Bulk DNase and RNA-seq data tracks as well aggregated single-cell RNA-seq data tracks are shown at the bottom. An image of enhancer assay result of an element overlapping Mu3 is shown on the right, taken from a published paper by Yee et al[183]. (B) Skeletal enhancers (Ca1-3) shown with in vivo enhancer validation from the VISTA resource for Ca3 (https://enhancer.lbl.gov/frnt_page_n.shtml) at the Sox5 locus. (C) The C1qb locus that is macrophage-specific. Three cCREs (Lb1-3) for limb-specific expression of this macrophage gene were identified. The contrasting promoter proximal element (Lv1) is active in liver and has a corresponding DHS profile.

4.3 Discussions

Whole-tissue biochemical measurements present a high degree of complexity that arises from shifting cell composition and unsynchronized maturation pace. Single-cell RNA-seq proved to be able to not only separate cell type-specific gene signatures that whole-tissue data could not identify, but also assess the composition of known cell types in whole-tissue data. This latter power will be even more profound when the cell atlas of individual tissues are complete. Before single-cell epigenomic assays are ready for reliable and wide use, single-cell RNA-seq measurements could be integrated with whole-tissue measurements to infer the most prominent cell type-specific and lineage-specific regulatory codes. My analyses demonstrate how a simple integration algorithm can infer regulatory mechanisms at cell-type level and show the power of single-cell RNA-seq profiling beyond measuring RNA itself.

4.4 Materials and methods

4.4.1 Bulk forelimb RNA-Seq from mouse embryo tissues

Forelimbs were individually dissected and processed using exactly the same protocol as in 2.4.1.

4.4.2 Comparing whole-tissue RNA-seq and single-cell RNA-seq.

Filtered top 1500 high-dispersion nuclear protein coding genes (See 3.4.3) were used as input for CIBERSORT[179] to compare against whole limb RNA-seq data. To compare marker genes defined from 3.4.3 against ENCODE whole-tissue RNA-seq clusters, marker genes defined for each cell type were matched to the ordered heatmap of the whole-tissue clustergram (Figure 2.5A). For better visualization, I not only showed individual locations as horizontal bars but also borrowed violin plots to show smoothed distribution of these marker genes.

4.4.3 IDEAS states

The IDEAS epigenetic states on the ENCODE3 mouse developmental data were generated by the IDEAS software[181,182] using 10 epigenetic marks: H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, ATAC-Seq, and DNAse methylation data. We first converted the raw data in each sample to –log10 p-values using a Negative Binomial model. The mean and variance parameters of the model for each sample were calculated using the bottom 99% of the data. We then adjusted the mean parameters at each genomic position from the input data to account for local genomic variations. Specifically, we downloaded the input data for each tissue (see list of data sets), and we calculated rolling means per genomic position using a 20kb window centered at the position, for both signals and the input. The ratio between the two means at each position was multiplied to the overall mean estimate of the sample, and we

normalized the ratios across the genome to have mean 1. We treated the –log10 p-value as input data for IDEAS, capped at 16, and we ran the program in its default setting. The output from IDEAS is a set of genome tracks to display in the genome browser, where each epigenetic state is assigned a color as a weighted mixture of colors pre-assigned by the program to each epigenetic mark. The IDEAS segmentation can be accessed by the Hub link (http://bx.psu.edu/~yuzhang/me66/hub_me66n_org.txt).

BIBLIOGRAPHY

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **270**, 467–470 (1995).
2. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
3. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630 (2000).
4. Ruby, J. G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell* **127**, 1193–1207 (2006).
5. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
6. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
7. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57 (2009).
8. Wold, B. & Myers, R. M. Sequence census methods for functional genomics. *Nat. Methods* **5**, 19 (2007).
9. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377 (2009).

10. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096 (2013).
11. Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11**, 41–46 (2014).
12. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
13. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
14. Zhang, J. A., Mortazavi, A., Williams, B. A., Wold, B. J. & Rothenberg, E. V. Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. *Cell* **149**, 467–482 (2012).
15. Han, X. *et al.* Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. *Proc. Natl. Acad. Sci.* **106**, 12741–12746 (2009).
16. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
17. Pervouchine, D. D. *et al.* Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nat. Commun.* **6**, 5903 (2015).
18. Sebé-Pedrós, A. *et al.* Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq. *Cell* **173**, 1520–1534 (2018).
19. Gerber, T. *et al.* Single-cell analysis uncovers convergence of cell identities during axolotl limb regeneration. *Science* eaaq0681 (2018).

20. Nestorowa, S. *et al.* A single cell resolution map of mouse haematopoietic stem and progenitor cell differentiation. *Blood* blood–2016 (2016).

21. La Manno, G. *et al.* Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* **167**, 566–580 (2016).

22. Kim, D. H. *et al.* Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming. *Cell Stem Cell* **16**, 88–101 (2015).

23. Montoro, D. T. *et al.* A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).

24. Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).

25. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337 (2011).

26. Barski, A. *et al.* Pol II and its associated epigenetic marks are present at Pol III–transcribed noncoding RNA genes. *Nat. Struct. Mol. Biol.* **17**, 629–634 (2010).

27. Cui, P. *et al.* A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics* **96**, 259–265 (2010).

28. Sultan, M. *et al.* Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* **15**, 675 (2014).

29. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).

30. Ramsköld, D. *et al.* Full-Length mRNA-Seq from single cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).

31. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

32. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

33. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).

34. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2017).

35. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

36. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

37. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

38. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32**, 462 (2014).

39. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525 (2016).

40. Guttman, M. *et al.* Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503 (2010).

41. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644 (2011).

42. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290 (2015).

43. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).

44. Rau, A., Gallopin, M., Celeux, G. & Jaffrézic, F. Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics* **29**, 2146–2152 (2013).

45. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).

46. Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. Real time quantitative PCR. *Genome Res.* **6**, 986–994 (1996).

47. Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the comparative C_T method. *Nat. Protoc.* **3**, 1101–1108 (2008).

48. Rebouças, E. de L. *et al.* Real time PCR and importance of housekeepings genes for normalization and quantification of mRNA expression in different tissues. *Braz. Arch. Biol. Technol.* **56**, 143–154 (2013).

49. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**, 218–223 (2009).

50. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

51. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

52. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).

53. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95**, 14863–14868 (1998).

54. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).

55. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. From data mining to knowledge discovery in databases. *AI Mag.* **17**, 37 (1996).

56. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

57. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2007).

58. Comon, P. Independent component analysis, a new concept? *Signal Process.* **36**, 287–314 (1994).

59. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

60. Brown, B. C., Bray, N. L. & Pachter, L. Expression reflects population structure. *PLOS Genet.* **14**, e1007841 (2018).

61. Soneson, C., Lilljebjörn, H., Fioretos, T. & Fontes, M. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics* **11**, 191 (2010).

62. Hotelling, H. Relations between two sets of variates. *Biometrika* **28**, 321–377 (1936).

63. Yu, Y. *et al.* A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. *Nat. Commun.* **5**, 3230 (2014).

64. White, R. J. *et al.* A high-resolution mRNA expression time course of embryonic development in zebrafish. *eLife* **6**, e30860 (2017).

65. Bush, S. J. *et al.* Combination of novel and public RNA-seq datasets to generate an mRNA expression atlas for the domestic chicken. *BMC Genomics* **19**, 594 (2018).

66. Lin, S. *et al.* Comparison of the transcriptional landscapes between human and mouse tissues. *Proc. Natl. Acad. Sci.* **111**, 17224–17229 (2014).

67. Khan, S. Y. *et al.* Transcriptome Profiling of Developing Murine Lens Through RNA Sequencing. *Invest. Ophthalmol. Vis. Sci.* **56**, 4919–4926 (2015).

68. Söllner, J. F. *et al.* An RNA-Seq atlas of gene expression in mouse and rat normal tissues. *Sci. Data* **4**, 170185 (2017).

69. Cornelison, D. D. & Wold, B. J. Single-cell analysis of regulatory gene expression in quiescent and activated mouse skeletal muscle satellite cells. *Dev. Biol.* **191**, 270–283 (1997).

70. Pockwinse, S. M., Wilming, L. G., Conlon, D. M., Stein, G. S. & Lian, J. B. Expression of cell growth and bone specific genes at single cell resolution during development of bone tissue-like organization in primary osteoblast cultures. *J. Cell. Biochem.* **49**, 310–323 (1992).

71. Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.* **24**, 496–510 (2014).

72. Jaitin, D. A. *et al.* Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).

73. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).

74. Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395 (2017).

75. Han, X. *et al.* Mapping the mouse cell atlas by Microwell-seq. *Cell* **172**, 1091–1107 (2018).

76. Audic, S. & Claverie, J.-M. The significance of digital gene expression profiles. *Genome Res.* **7**, 986–995 (1997).

77. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).

78. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).

79. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).

80. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).

81. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367 (2018).

82. Mass, E. *et al.* Specification of tissue-resident macrophages during organogenesis. *Science* **353**, aaf4238 (2016).

83. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).

84. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst.* **2**, 239–250 (2016).

85. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381 (2014).

86. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979 (2017).

87. Shin, J. *et al.* Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. *Cell Stem Cell* **17**, 360–372 (2015).

88. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).

89. Setty, M. *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637 (2016).

90. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494 (2018).

91. Zeisel, A. *et al.* Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Mol. Syst. Biol.* **7**, 529 (2011).

92. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).

93. Frieda, K. L. *et al.* Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107 (2017).

94. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).

95. Schmid, M. W. *et al.* A Powerful Method for Transcriptional Profiling of Specific Cell Types in Eukaryotes: Laser-Assisted Microdissection and RNA Sequencing. *PLOS ONE* **7**, e29685 (2012).

96. Peng, G. *et al.* Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. *Dev. Cell* **36**, 681–697 (2016).

97. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).

98. Moncada, R. *et al.* Building a tumor atlas: integrating single-cell RNA-Seq data with spatial transcriptomics in pancreatic ductal adenocarcinoma. *bioRxiv* 254375 (2018). doi:10.1101/254375

99. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342–357 (2016).

100. Moffitt, J. R. & Zhuang, X. RNA Imaging with Multiplexed Error Robust Fluorescence in situ Hybridization. *Methods Enzymol.* **572**, 1–49 (2016).

101. Shah, S. *et al.* Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. *Cell* **174**, 363–376 (2018).

102. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* eaat5691 (2018). doi:10.1126/science.aat5691

103. Chung, K. & Deisseroth, K. CLARITY for mapping the nervous system. *Nat. Methods* **10**, 508–513 (2013).

104. Yang, B. *et al.* Single-Cell Phenotyping within Transparent Intact Tissue through Whole-Body Clearing. *Cell* **158**, 945–958 (2014).

105. Betzig, E. *et al.* Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. *Science* **313**, 1642–1645 (2006).

106. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3**, 793–795 (2006).

107. Chen, F. *et al.* Nanoscale imaging of RNA with expansion microscopy. *Nat. Methods* **13**, 679–684 (2016).

108. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).

109. Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum. Mol. Genet.* **19**, R227–R240 (2010).

110. Han, K. Y. *et al.* SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome Res.* **28**, 75–87 (2018).

111. Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & Van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* **33**, 285 (2015).

112. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519 (2015).

113. Macaulay, I. C. *et al.* Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. *Nat. Protoc.* **11**, 2081–2103 (2016).

114. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* eaau0730 (2018). doi:10.1126/science.aau0730

115. Lane, K. *et al.* Measuring signaling and RNA-seq in the same cell links gene expression to dynamic patterns of NF-κB activation. *Cell Syst.* **4**, 458–469 (2017).

116. Genshaft, A. S. *et al.* Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome Biol.* **17**, 188 (2016).

117. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865 (2017).

118. Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **35**, 936 (2017).

119. Dixit, A. *et al.* Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).

120. Hill, A. J. *et al.* On the design of CRISPR-based single cell molecular screens. *Nat. Methods* **15**, 271–274 (2018).

121. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol. Cell* **66**, 285-299.e5 (2017).

122. Jaitin, D. A. *et al.* Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896.e15 (2016).

123. Clark, S. J. *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).

124.Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041 (2017).

125. Peter, I. & Davidson, E. H. *Genomic control process: development and evolution*. (Academic Press, 2015).

126. Kobayashi, A. *et al.* Six2 defines and regulates a multipotent self-renewing nephron progenitor population throughout mammalian kidney development. *Cell Stem Cell* **3**, 169–181 (2008).

127. Schwickert, T. A. *et al.* Stage-specific control of early B cell development by the transcription factor Ikaros. *Nat. Immunol.* **15**, 283 (2014).

128.Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol.* **13**, R51 (2012).

129. Zhou, Y. & Rui, L. Major Urinary Protein Regulation of Chemical Communication and Nutrient Metabolism. *Vitam. Horm.* **83**, 151–163 (2010).

130. Vickerman, L., Neufeld, S. & Cobb, J. Shox2 function couples neural, muscular and skeletal development in the proximal forelimb. *Dev. Biol.* **350**, 323–336 (2011).

131. Han, S. J., Li, H., Kim, M., Shlomchik, M. J. & Lee, H. T. Kidney Proximal Tubular TLR9 Exacerbates Ischemic Acute Kidney Injury. *J. Immunol.* ji1800211 (2018). doi:10.4049/jimmunol.1800211

132.Nusse, R. Wnt signaling and stem cell control. *Cell Res.* **18**, 523–527 (2008).

133. Carter, C. S., Grippo, A. J., Pournajafi-Nazarloo, H., Ruscio, M. G. & Porges, S. W. Oxytocin, vasopressin and sociality. in *Progress in Brain Research* (eds. Neumann, I. D. & Landgraf, R.) **170**, 331–336 (Elsevier, 2008).

134. Grote, D., Souabni, A., Busslinger, M. & Bouchard, M. Pax 2/8-regulated Gata 3 expression is necessary for morphogenesis and guidance of the nephric duct in the developing kidney. *Dev. Camb. Engl.* **133**, 53–61 (2006).

135. Batista, M. F. & Lewis, K. E. Pax2/8 act redundantly to specify glycinergic and GABAergic fates of multiple spinal interneurons. *Dev. Biol.* **323**, 88–97 (2008).

136. Choksi, S. P., Lauter, G., Swoboda, P. & Roy, S. Switching on cilia: transcriptional networks regulating ciliogenesis. *Development* **141**, 1427–1441 (2014).

137. Ibañez-Tallon, I., Heintz, N. & Omran, H. To beat or not to beat: roles of cilia in development and disease. *Hum. Mol. Genet.* **12**, R27–R35 (2003).

138. Petryniak, M. A., Potter, G. B., Rowitch, D. H. & Rubenstein, J. L. R. Dlx1 and Dlx2 Control Neuronal versus Oligodendroglial Cell Fate Acquisition in the Developing Forebrain. *Neuron* **55**, 417–433 (2007).

139. Nakatani, T., Minaki, Y., Kumai, M. & Ono, Y. Helt determines GABAergic over glutamatergic neuronal fate by repressing Ngn genes in the developing mesencephalon. *Development* **134**, 2783–2793 (2007).

140. Liodis, P. *et al.* Lhx6 activity is required for the normal migration and specification of cortical interneuron subtypes. *J. Neurosci. Off. J. Soc. Neurosci.* **27**, 3078–3089 (2007).

141. Colasante, G. *et al.* Arx is a direct target of Dlx2 and thereby contributes to the tangential migration of GABAergic interneurons. *J. Neurosci. Off. J. Soc. Neurosci.* **28**, 10674–10686 (2008).

142. Lodato, S. *et al.* Excitatory Projection Neuron Subtypes Differentially Control the Distribution of Local Inhibitory Interneurons in the Cerebral Cortex. *Neuron* **69**, 763–779 (2011).

143. Emerson, R. O. & Thomas, J. H. Adaptive evolution in zinc finger transcription factors. *PLoS Genet.* **5**, e1000325 (2009).

144. Maxwell, A. I., Morrison, G. M. & Dorin, J. R. Rapid sequence divergence in mammalian beta-defensins by adaptive evolution. *Mol. Immunol.* **40**, 413–421 (2003).

145. Zhang, L., Ju, X., Cheng, Y., Guo, X. & Wen, T. Identifying Tmem59 related gene regulatory network of mouse neural stem cell from a compendium of expression profiles. *BMC Syst. Biol.* **5**, 152 (2011).

146. Woo, J.-M. *et al.* Characterization of changes in global gene expression in the brain of neuron-specific enolase/human Tau23 transgenic mice in response to overexpression of Tau protein. *Int. J. Mol. Med.* **25**, 667–675 (2010).

147. Williams, W. M. *et al.* Antimicrobial peptide β-defensin-1 expression is upregulated in Alzheimer's brain. *J. Neuroinflammation* **10**, 127 (2013).

148. Brown, C. J. *et al.* A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**, 38–44 (1991).

149. Yang, X. *et al.* Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res.* **16**, 995–1004 (2006).

150. Zhang, H.-M. *et al.* AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.* **43**, D76–D81 (2015).

151. Hollenhorst, P. C., Shah, A. A., Hopkins, C. & Graves, B. J. Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ETS gene family. *Genes Dev.* **21**, 1882–1894 (2007).

152. Ji, Z. *et al.* Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol. Syst. Biol.* **7**, 534 (2011).

153. Min, I. M. *et al.* Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev.* **25**, 742–754 (2011).

154. Paulsen, M. T. *et al.* Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. *Methods San Diego Calif* **67**, 45–54 (2014).

155. Gorkin, D. U., Williams, B. A., Trout, D. & Amrhein, H. Systematic mapping of chromatin state landscapes during mouse development. (2017).

156. Chong, J. A. *et al.* REST: A mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* **80**, 949–957 (1995).

157. Schoenherr, C. J. & Anderson, D. J. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science* **267**, 1360–1363 (1995).

158. Mukherjee, S., Brulet, R., Zhang, L. & Hsieh, J. REST regulation of gene networks in adult neural stem cells. *Nat. Commun.* **7**, 13360 (2016).

159. McGann, J. C. *et al.* Polycomb- and REST-associated histone deacetylases are independent pathways toward a mature neuronal phenotype. *eLife* **3**, e04235 (2014).

160. Arnold, P. *et al.* Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. *Genome Res.* **23**, 60–73 (2013).

161. Xia, H. *et al.* Loss of brain-enriched miR-124 microRNA enhances stem-like traits and invasiveness of glioma cells. *J. Biol. Chem.* **287**, 9962–9971 (2012).

162. Du, S. *et al.* MicroRNA-124 inhibits cell proliferation and migration by regulating SNAI2 in breast cancer. *Oncol. Rep.* **36**, 3259–3266 (2016).

163. Hu, F.-Y. *et al.* miR-124 modulates gefitinib resistance through SNAI2 and STAT3 in non-small cell lung cancer. *J. Huazhong Univ. Sci. Technol. Med. Sci. Hua Zhong*

Ke Ji Xue Xue Bao Yi Xue Ying Wen Ban Huazhong Keji Daxue Xuebao Yixue Yingdewen Ban **36**, 839–845 (2016).

164. Conaco, C., Otto, S., Han, J.-J. & Mandel, G. Reciprocal actions of REST and a microRNA promote neuronal identity. *Proc. Natl. Acad. Sci.* **103**, 2422–2427 (2006).

165. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME suite. *Nucleic Acids Res.* **43**, W39–W49 (2015).

166. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).

167. Berriz, G. F., King, O. D., Bryant, B., Sander, C. & Roth, F. P. Characterizing gene sets with FuncAssociate. *Bioinformatics* **19**, 2502–2504 (2003).

168. Scialdone, A. *et al.* Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535**, 289 (2016).

169. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).

170. Cadwell, C. R. *et al.* Multimodal profiling of single-cell morphology, electrophysiology, and gene expression using Patch-seq. *Nat. Protoc.* **12**, 2531 (2017).

171. Buckingham, M. & Relaix, F. PAX3 and PAX7 as upstream regulators of myogenesis. *Semin. Cell Dev. Biol.* **44**, 115–125 (2015).

172. Goulding, M., Lumsden, A. & Paquette, A. J. Regulation of Pax-3 expression in the dermomyotome and its role in muscle development. *Development* **120**, 957–971 (1994).

173. Williams, B. A. & Ordahl, C. P. Pax-3 expression in segmental mesoderm marks early stages in myogenic cell specification. *Development* **120**, 785–796 (1994).

174. Biressi, S. *et al.* Intrinsic phenotypic diversity of embryonic and fetal myoblasts is revealed by genome-wide gene expression analysis on purified cells. *Dev. Biol.* **304**, 633–651 (2007).

175. Lagha, M. *et al.* Transcriptome analyses based on genetic screens for Pax3 myogenic targets in the mouse embryo. *BMC Genomics* **11**, 696 (2010).

176. Kozhemyakina, E., Lassar, A. B. & Zelzer, E. A pathway to bone: signaling molecules and transcription factors involved in chondrocyte development and maturation. *Development* **142**, 817–831 (2015).

177. Hartmann, C. Transcriptional networks controlling skeletal development. *Curr. Opin. Genet. Dev.* **19**, 437–443 (2009).

178. Szklarczyk, D. *et al.* STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).

179. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453 (2015).

180. Sanchez-Gurmaches, J. *et al.* PTEN loss in the Myf5 lineage redistributes body fat and reveals subsets of white adipocytes that arise from Myf5 precursors. *Cell Metab.* **16**, 348–362 (2012).

181. Zhang, Y., An, L., Yue, F. & Hardison, R. C. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.* **44**, 6721–6731 (2016).

182. Zhang, Y. & Hardison, R. C. Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. *Nucleic Acids Res.* **45**, 9823–9836 (2017).

183. Yee, S.-P. & Rigby, P. W. The regulation of myogenin gene expression during the embryonic development of the mouse. *Genes Dev.* **7**, 1277–1289 (1993).