Graph modeling for genomics and epidemiology

Thesis by
Kristján Eldjárn Hjörleifsson

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

Caltech

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2023
Defended December 16, 2022

ACKNOWLEDGMENTS

I am endlessly grateful to my advisor, Lior Pachter, and my surrogate advisor Páll Melsted, for being kind, patient, and inspiring mentors. I would also like to thank Barbara Wold and Adam Wierman for being a source of inspiration, and for serving on my committee. On a more personal note, I would like to thank my parents, Sigrún and Hjörleifur, for equipping me with all the qualities needed to undertake this journey. A Ph.D. can be a lonely experience, but the group of friends I made along the way more than made up for it. In particular, my cohort in the CMS department was an invaluable source of support for me, my lab mates in the Pachter lab some of the most talented scientists I've met, and my two co-first authors Sölvi and Delaney a constant source of inspiration and grit. The people are too many to name here, but you know who you are. Finally, to my partner Þóranna: Thank you for your unyielding patience and support.

ABSTRACT

The last decades have seen great leaps made in the development of RNA sequencing technologies, yielding lower cost and greater throughput of experiments, to the point where the scale of the data produced on a daily basis is staggering. While computational hardware is also continuously improving, famously (or perhaps infamously) described by Gordon Moore (Moore, 1965), the rate at which data are produced eclipses advances on the hardware front. Over the last few years, many new methods have been proposed for bridging that ever-widening chasm, more than a few of which harness the latent graphical structure of genomic data to reduce the number of calculations required and pack the data tighter in memory. This body of work continues this development on three different, but related, fronts. Firstly, I present developments that greatly improve upon the efficiency of state-of-the-art methods for the quantification of RNA-seq reads, and describe a method that improves the accuracy of quantification without substantially increasing the computational overhead. Secondly, I introduce a procedure for the discovery of associations between novel gene isoforms and phenotypes, without prior knowledge of those isoforms. Lastly, I present the largest reconstruction of the transmission tree of a viral outbreak to date, modeled from viral genome sequences, contact tracing, and symptom data. I then use the reconstructed transmission tree to assess the efficacy of different vaccination strategies.

PUBLISHED CONTENT AND CONTRIBUTIONS

Figure 6.1, figure 6.4, and figure 6.5 are reprinted with permission from the copyright holder, Elsevier.

Hjörleifsson, Kristján Eldjárn, Lior Pachter, and Páll Melsted (2022). "Annotation-agnostic discovery of associations between novel gene isoforms and phenotypes". In: *bioRxiv*. DOI: `10.1101/2022.12.02.518787`.
K.E.H., P.M., and L.P. designed the study and interpreted the results. **K.E.H.** implemented the method and the simulation framework. **K.E.H.**, P.M., and L.P. drafted the manuscript.

Hjörleifsson, Kristján Eldjárn, Solvi Rognvaldsson, et al. (June 2022). "Reconstruction of a large-scale outbreak of SARS-CoV-2 infection in Iceland informs vaccination strategies". In: *Clinical Microbiology and Infection* 28 (6), pp. 852–858. ISSN: 14690691. DOI: `10.1016/j.cmi.2022.02.012`.
K.E.H. and S.R. contributed equally to this article. **KEH**, SR, PM, and KS designed the study and interpreted the results. ABA, MA, KB, GH, AdJ, AsJ, NK, BK, DNM, LLR, GMS, AS, FJ, OTM, GLN, and JS planned and performed the laboratory work. ESE, RP, MIS, and MT performed the data collection. **KEH**, SR, HJ, ESE, RF, GG, KRG, AG, BOJ, KSJ, TK, RP, MIS, GS, EAT, BT, MT, AH, HH, IJ, GM, PS, UT, and PM performed the data curation. **KEH**, SR, HJ, OE, DFG, and PM performed the statistical and bioinformatics analyses. **KEH**, SR, PM, and KS drafted the manuscript. All authors contributed to the final version of the paper.

Hjörleifsson, Kristján Eldjárn, Delaney K Sullivan, et al. (2022). "Accurate quantification of single-nucleus and single-cell RNA-seq transcripts". In: *bioRxiv*. DOI: `10.1101/2022.12.02.518832`.
K.E.H. and D.K.S. contributed equally to this article. **KEH**, D.K.S, and L.P. designed the study and interpreted the results. **KEH**, D.K.S., and L.P. performed the statistical and bioinformatics analyses. K.E.H., D.K.S., G.H., and P.M. created the software implementation of the method. **KEH**, D.K.S., and L.P. drafted the manuscript.

Melsted, Páll et al. (July 2021). "Modular, efficient and constant-memory single-cell RNA-seq preprocessing". In: *Nature Biotechnology* 39 (7), pp. 813–818. ISSN: 15461696. DOI: `10.1038/s41587-021-00870-2`.
P.M., A.S.B., L. Liu and L.P. developed the algorithms for bustools and P.M., A.S.B. and L. Liu wrote the software. A.S.B. conceived of and performed the UMI and barcode calculations motivating the algorithms. F.G. implemented and performed the benchmarking procedure, and curated indices for the datasets. A.S.B. and E.d.V.B. designed and produced the comparisons between Cell Ranger and kallisto bustools. L. Lu investigated in detail the performance of different workflows on the "10k mouse neuron" data and produced the analysis of that

dataset. A.S.B. designed the RNA velocity workflow and performed the RNA velocity analyses. K.M.H contributed to the development of the reproducible workflow. **K.E.H.** developed and investigated the effect of reference transcriptome sequences for pseudoalignment. J.G. interpreted results and helped to supervise the research. A.S.B. planned, organized and prepared figures. A.S.B., E.d.V.B., P.M., and L.P. planned the manuscript. A.S.B. and L.P. wrote the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

The initial sequencing of the human genome in 2001 (Lander et al., 2001) heralded beginning of another momentous undertaking: the large-scale, genome-wide quantification of mRNA in human cells. In the intervening years, great leaps have been made in the technologies involved, at first with cDNA microarrays, but more recently with RNA-sequencing (RNA-seq), which has become the *de facto* standard method for sequencing expression data, offering lower cost, while yielding a higher resolution and sensitivity (Mortazavi et al., 2008).

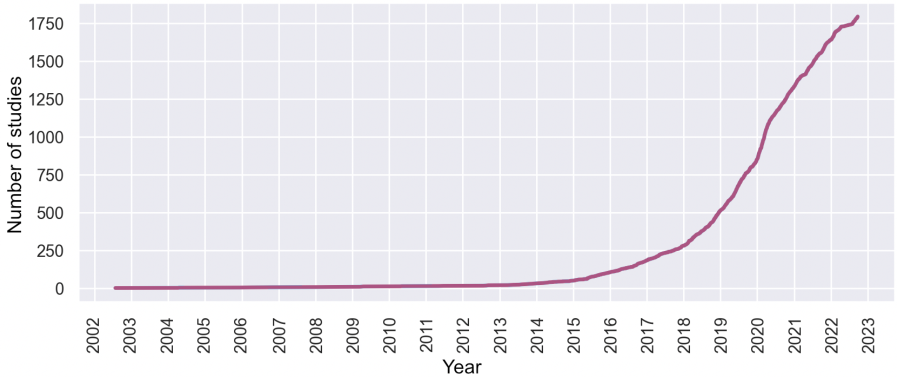


Figure 1.1: The number of published single-cell RNA sequencing studies over time. Data obtained from Svensson, Veiga Beltrame, and Pachter, 2020 on December 12, 2022.

Progress has been such, that the scale of data produced daily has become staggering. The associated computational challenges of processing the data, quantifying RNA abundances, and analyzing expression patterns, have been profound, and have spurred development of many algorithms (Conesa et al., 2016). The key step of RNA-seq *quantification* (Figure 1.2, figure 1.3), in which the raw RNA reads are aligned to known sequences in a reference genome, is particularly complex as it requires not just naive counting of molecules, but mapping of large number of reads and assignment of ambiguously mapping reads via the expectation-maximization and related algorithms (Nicolae et al., 2011). This is done in order to determine which gene or transcript the RNA read is a product of, and ultimately obtain a measure of the expression levels of different genes in the experiment. Many mapping algorithms have been proposed, most of which are computationally intensive, and new or updated software packages which implement novel algorithms and improve upon prior methods are regularly published.

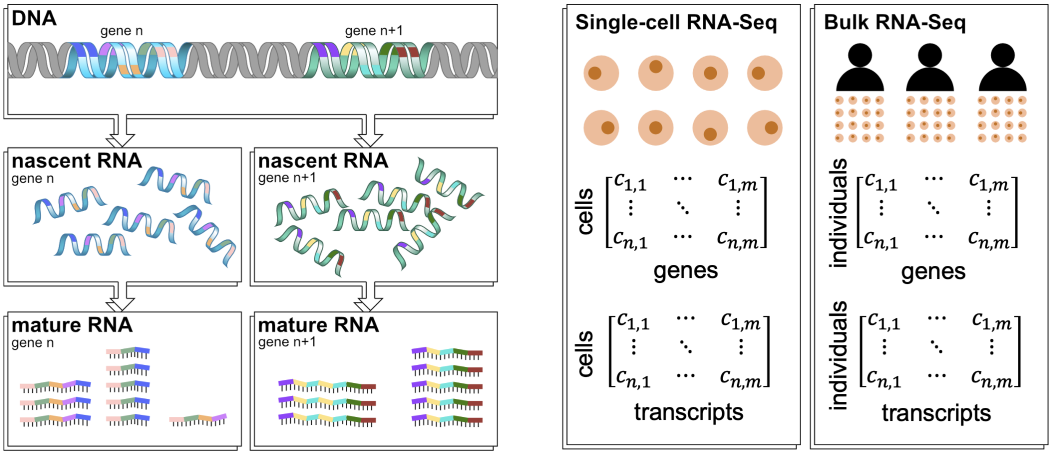


Figure 1.2: In any cell, there are strands of DNA, certain regions of which we refer to as genes. These genes consist mainly of two building blocks: exons, which play a role in the production of protein, and introns, which generally do not. Genes are constantly being *transcribed* into *nascent* RNA molecules, which are then in turn processed such that the non-protein-coding introns are spliced out. The remaining exons can be spliced together in different ways, meaning that each gene can have multiple different *mature* RNA products. For example, the Ensembl annotation 108 of the human transcriptome contains 33,548 genes, which can yield 236,233 different transcripts (Cunningham et al., 2022). We refer to the RNA molecules present in a cell at a given time as the *RNA expression* of the cell. RNA quantification involves sampling RNA molecules from multiple cells, and quantifying the gene or transcript abundances either per cell (in the case of single-cell RNA sequencing) or per individual in a cohort (in the case of bulk RNA sequencing). This results in a matrix where each element represents the abundance of a gene or transcript in a given cell or individual.

One such method is the RNA-seq quantification method kallisto (Bray et al., 2016), published in 2016, which pseudo-aligns RNA reads against a graphical representation of a *transcriptome*, which consists of the sequences of the known transcripts, or processed mature messenger RNA molecules, in the target genome. Previous methods align reads to individual transcripts in the transcriptome annotation, or to the genome (Figure 1.4). This method, which relies on the *de Bruijn* graph data structure for memory efficiency, has heralded the development of numerous fast and accurate methods that require minimal computational resources (Figure 1.5). In this body of work, I develop new algorithms for kallisto, which not only further reduce

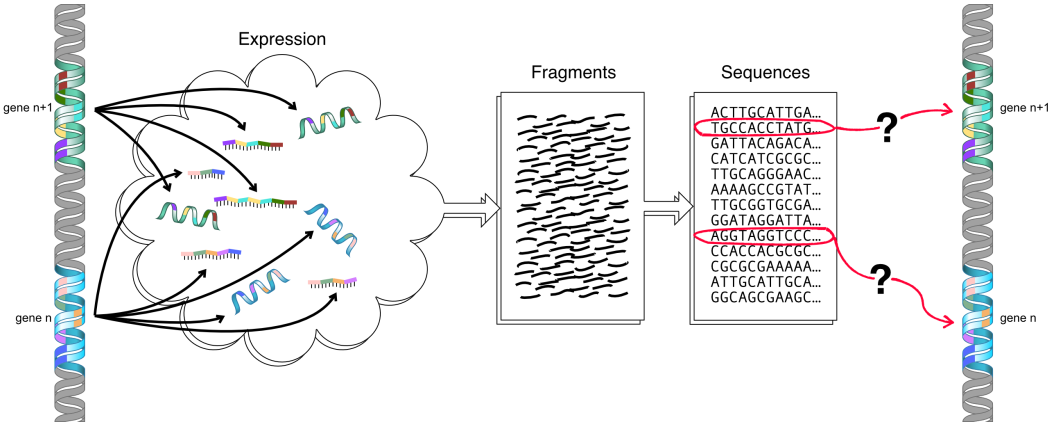


Figure 1.3: Both partially processed and completely processed transcripts are captured during single-cell RNA-seq. During sequencing, the molecules are broken up into smaller *fragments*, the sequences of which are digitized and represented as strings of As, Cs, Gs, and Ts. The quantification problem involves determining which gene or transcript each of those fragments originates from, in order to measure the relative expression levels of all genes in a sample.

the needed computational resources, but also enable more accurate and targeted quantification of RNA-seq data than previously possible. I define nomenclature for reasoning about the nuclear or cytoplasmic origins of RNA-seq reads and develop a provably optimal (under mild assumptions) algorithm of distinguishing between the two when they can be disambiguated.

Most quantification tools, kallisto included, have an oft-ignored blind spot when it comes to the quantification of novel transcripts. These methods rely on an annotation of the target transcriptome. However, annotations of mRNAs may be incomplete, or if the organism is not well studied they may be nonexistent (Zhang et al., 2020). This may result in data being discarded, and to erroneous quantifications (Kuo, Hansen, and Hicks, 2022). Furthermore, in downstream applications such as eQTL analysis, such errors can propagate and result in missed associations. In separate work, I present a novel algorithm for associating phenotypes with RNA expression, that can identify expression associations resulting from a wide variety of underlying transcriptional and post-transcriptional events, without prior knowledge of these events.

1.1 Thesis contribution and organization

Following is a summary of the main contributions of this thesis.

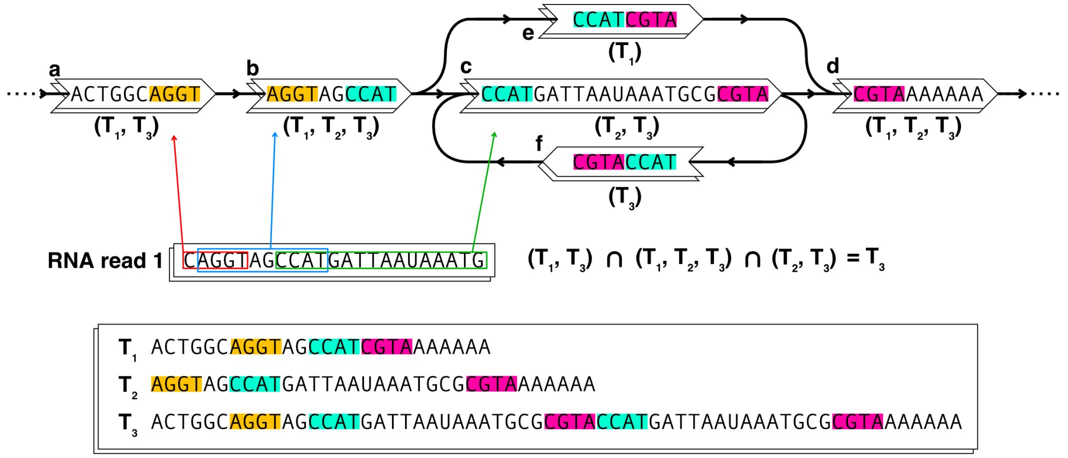

a
ACTGGCAGGT
(T_1, T_3)
b
AGGTAGCCAT
(T_1, T_2, T_3)
e
CCATCGTA
(T_1)
c
CCATGATTAAUAAATGCGCGTA
(T_2, T_3)
f
CGTACCAT
(T_3)
d
CGTAAAAAAA
(T_1, T_2, T_3)
RNA read 1
CAGGTAGCCATGATTAAUAAATG
$(T_1, T_3) \cap (T_1, T_2, T_3) \cap (T_2, T_3) = T_3$
T_1 ACTGGCAGGTAGCCATCGTAAAAAAA
T_2 AGGTAGCCATGATTAAUAAATGCGCGTAAAAAAA
T_3 ACTGGCAGGTAGCCATGATTAAUAAATGCGCGTACCATGATTAAUAAATGCGCGTAAAAAAA

Figure 1.4: RNA-seq quantification with kallisto. A *de Bruijn* graph is built from the reference transcriptome, which in this case consists of three transcripts, T_1, T_2, T_3. Each vertex in a k-dimensional *de Bruijn* graph represents a string of symbols of length at least k. In this toy example, $k = 5$. There is an edge between vertex a and vertex b if the $k - 1$ last symbols in the string represented by a are the same as the $k - 1$ first symbols in the string represented by node b. The set of transcript that uses each vertex is then used to color that vertex. I.e. both T_1 and T_3 use node a, so the color of node a is (T_1, T_3). During quantification, each read is broken down into its constituent k-mers, and the nodes they occur in are found in the graph. In order to determine which transcript the read corresponds to, the intersection of the colors is taken. If there are multiple transcripts in the intersection, the read is fractionally assigned to both.

- I develop an algorithm that increases the specificity of RNA-seq quantification, which under mild assumptions is provably correct, and requires only the minimal possible amount of data to do so. The algorithm leverages the efficient de Bruijn graph data structure to effectively mask out undesired sequences without minimal storage overhead, and retains its accuracy in terms of quantification of desired sequences. A key idea is the introduction of a new concept in sequence alignment, namely that of distinguishing flanking k-mers. I implement this algorithm in efficient software via a new version of kallisto that additionally leverages novel improvements I have developed that reduce memory requirements. I make use of k-mer minimizers (Roberts et al., 2004), minimal perfect hash functions (Limasset et al., 2017), memoization, and other graph theoertic principles. I show that quantifying RNA-seq reads using indices built with a *D-list* containing noncoding sequences, such as

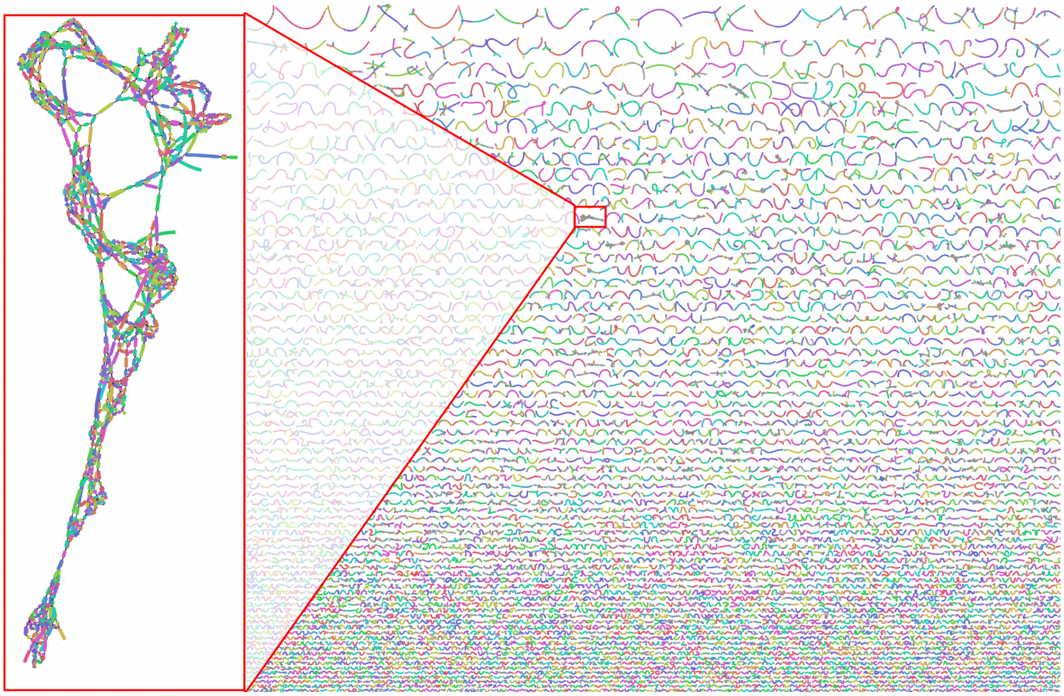


Figure 1.5: A *de Bruijn* graph of Ensembl annotation version 108 of the human transcriptome (Cunningham et al., 2022). The left-hand side shows one of many connected components. Each connected component mostly constitutes one gene, or a few genes, in case they share k-mers.

introns, intergenic regions, and transposable elements, increases the accuracy of quantification. Thus, the algorithmic ideas translate to biological insight by revealing how to improve the resolution of quantification with existing assays. Furthermore, I benchmark the performance against the previous version of kallisto and other widely used quantification tools. I show that my new version of kallisto outperforms other methods in terms of memory and CPU usage, and thereby show that accurate quantification can be done with modest hardware requirements.

- I describe the computational and graph theoretical improvements made to kallisto in order to enable the processing of ever larger data sets and lowering the barrier of entry for doing large-scale RNA-seq analyses.

- I present a novel method for RNA-seq quantification, which does not rely on an annotation of the underlying transcriptome. The non-annotated sequence abundances in my reference-free approach enable the direct association phenotypes with RNA expression, which can identify expression associations resulting from a wide variety of underlying transcriptional and post-transcriptional events, without prior knowledge of these events. I show that I can reliably reproduce known associations, and detect, *de novo*, phenotypically relevant transcriptional structures. Thus, by performing statistics tests directly on a useful data structure rather than on summarized data, I am able to extract associations that previously were not identifiable.

- I present a statistical model that, given viral genome sequences and extensive contact tracing data and medical data, reconstructs the latent transmission tree in a viral outbreak. The model implements an MCMC-algorithm (Metropolis et al., 1953) to sample from the distribution of all possible transmission trees using multiple custom likelihood functions to assess infector-infectee pairs in the tree. This method expands upon prior methods in terms of the types of data used (Campbell, Didelot, et al., 2018, Campbell, Cori, et al., 2019). Using this method, I recover a transmission tree for a large-scale outbreak of SARS-CoV-2 in Iceland. I quantify the effect that mandated quarantine has on the spread of the virus and I show a strong correlation between the age of patients and infection rate. Furthermore, I model different vaccination strategies using the transmission tree and compare their effectiveness in terms of the total number of cases, the number of ICU cases, and the number of deaths. A randomized algorithm for vaccination is revealed to be nearly as accurate as the optimal strategy, and practical to implement. This computational epidemiologic finding should be of use in the future as vaccination becomes more routine for a wide variety of infectious diseases.

Chapter 2

TRANSCRIPTION, PROCESSING, AND NOMENCLATURE

2.1 Introduction

In Chapter 3 we develop a method for the accurate quantification of single-cell RNA-seq reads. In particular, our method seeks to distinguish between completely spliced RNA molecules and those still containing introns. This chapter recapitulates the order of operations in the transcription and splicing processes as they are understood today. It then defines the terminology we will be using in chapter Chapter 3 and outlines the simplifying assumptions behind those terms.

2.2 The co-occurrence of transcription and splicing

The life cycle of an RNA molecule begins at the start of transcription, during which the molecule is considered to be nascent as it undergoes synthesis with RNA polymerase. Before their export from the nucleus, RNA molecules are processed, which entails removing the introns and splicing the remaining exons together, yielding what we term a completely processed RNA molecule. Although often studied separately, there is substantial evidence that splicing by the spliceosome is concurrent to the transcription (Y. Osheim, Miller, and Beyer, 1985; Beyer and Y. N. Osheim, 1988; Tennyson, Klamut, and Worton, 1995; Wuarin and Schibler, 1994; Neugebauer, 2002; Pandya-Jones and Black, 2009). In fact, the co-occurrence of transcription by RNA polymerase II (pol II) and processing has been shown to substantially increase the throughput of the processing, due to the C-terminal domain (CTD) of pol II, which facilitates attachment of RNA processing factors (Bentley, 2005). Thus, the notions of completely transcribed, but completely unprocessed molecules do not accurately reflect the molecular biology of the cell.

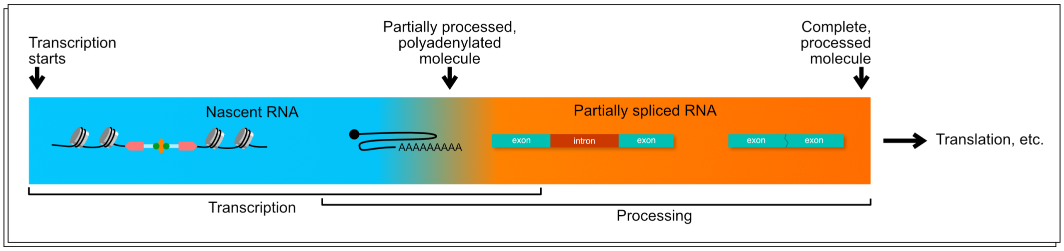


Figure 2.1: Overview of the order of operations in RNA transcription and splicing.

2.3 The order of splicing

Transcription culminates in the polyadenylation (poly(A)) of the molecule, in preparation for nuclear export (Stewart, 2019). Nascent pre-mRNAs are cleft 20-30 bp downstream of poly(A) sites, and poly(A) polymerase appends poly(A) tails to the 3' overhangs, after which they can be captured by current, 3' capturing RNA-seq methods. As with splicing, the polyadenylation process may begin co-transcriptionally, but continue after transcription has terminated (Neugebauer, 2002). Partially processed molecules therefore co-occur with completely processed molecules; in fact splicing can be a multistep process in which long introns are spliced recursively, yielding many partial products (Wuarin and Schibler, 1994; Tennyson, Klamut, and Worton, 1995). Per (Wetterberg, Baurén, and Wieslander, 1996) and (Pandya-Jones and Black, 2009) introns are generally removed in a 5'-to-3' order. Furthermore, (Wetterberg, Baurén, and Wieslander, 1996) also show that introns located in the 5' part of the gene are likely to be removed during transcription, whereas those located in the 3' part are more likely to be removed after transcription. Nevertheless, there are exceptions to this order of operations, and exon length appears to play a role in the order of excision (Tardiff, Lacadie, and Rosbash, 2006). However, this general 5'-to-3' order of intron removal means that with current, 3' capturing RNA-seq methods the presence of an exon-exon junction is a good indicator that most upstream introns have been removed, and the molecule completely or almost completely processed. Conversely, the presence of an intron-exon junction does not indicate that the molecule is completely unprocessed; only that it has not yet been completely processed, as introns upstream of it are likely to have been removed. Based on these known mechanics of the transcription and splicing processes, and state-of-the-art RNA sequencing methods, we can classify the molecules from which we obtain RNA reads in an experiment as completely processed (Ⓟ) and partially processed (Ⓤ).

Chapter 3

ACCURATE QUANTIFICATION OF SINGLE-CELL RNA-SEQ

The presence of both unprocessed and processed mRNA molecules in single-cell RNA-seq data leads to ambiguity in the notion of a "count matrix". Underlying this ambiguity is the challenging problem of separately quantifying completely processed and completely unprocessed mRNAs. In this chapter, we address this problem by relating k-mer assignment to read assignment in the context of different classes of molecules, and describe a unified approach to quantifying both single-cell and single-nucleus RNA-seq.

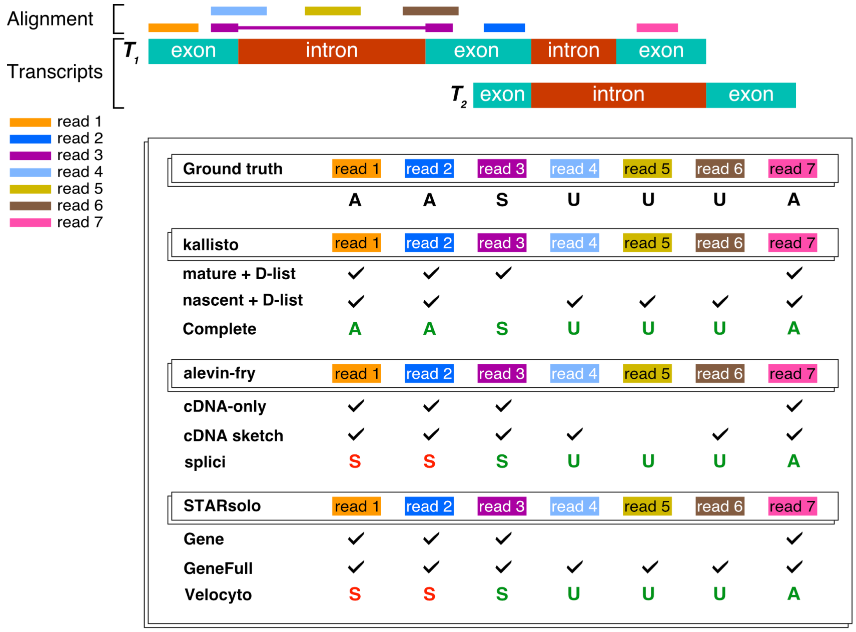
3.1 Introduction

The utility of single-cell RNA-seq measurements for defining cell types has represented a marked improvement over bulk RNA-seq, and is driving rapid adoption of single-cell RNA-seq assays (Zeng, 2022). Another application of single-cell RNA-seq that is not possible with bulk RNA-seq is the study of cell transitions and transcription dynamics, even via snapshot single-cell RNA-seq experiments (Gorin, Fang, et al., 2022). This novel application of single-cell RNA-seq is based on the quantification of both unprocessed and processed mRNAs (Figure 3.2.A), lending import to the computational problem of accurately quantifying these two modalities. The challenge of quantifying unprocessed mRNAs in addition to processed mRNAs has also been brought to the fore with single-nucleus RNA-seq (Kuo, Hansen, and Hicks, 2022).

The difficulty in quantifying both processed and unprocessed transcripts from single-cell RNA-seq data stems from the fact that sequenced reads are typically much shorter than transcripts, and therefore there can be ambiguity in classification of reads as originating from processed mRNAs vs. their unprocessed precursors (Figure 3.2.B). Reads that span a junction, i.e. cover two exons separated by an intron, must originate from a completely or partially processed mRNA (Ⓟ), whereas reads containing sequence unique to an intron must originate from an unprocessed or partially processed mRNA (Ⓤ), however there are many reads for which it is impossible to know whether they originated from an unprocessed or processed transcript (Ⓤ|Ⓟ). Furthermore, the way in which these cases are resolved depends on whether reads have been mapped to a whole genome, or directly to transcript sequences derived from annotations of the genome. The former approach lends itself well to identifying reads originating from completely unprocessed transcripts, as the genome includes all non-coding sequences, however the latter approach is superior for identifying spliced reads, because the sequence of processed transcripts is present in the reference being mapped to. Furthermore, methods that rely on k-mer matching to speed up alignment must account for the distinction between k-mer ambiguity and read ambiguity (Figure 3.2.C), and this distinction has not been carefully accounted for in existing k-mer based single-cell RNA-seq pre-processing workflows (Melsted et al., 2021; He et al., 2022).

As a result of these complexities, existing single-cell RNA-seq quantification algorithms provide a smorgasbord of options for users, but confusing, or at times contradictory, guidance on how to quantify single-cell RNA-seq, with unfortunate implications for analysis (Soneson et al., 2021). For example,the popular Cell Ranger software for quantifying single-cell RNA-seq generated with 10x Genomics machines, in its first six releases only quantified processed RNAs, and not unprocessed transcripts, and a separate program was required for generating unprocessed molecule counts (Manno et al., 2018). Moreover, the Cell Ranger quantification is based on an assumption that all reads that are ambiguous as to their origin from unprocessed or processed transcripts, are always counted as being derived from completely processed transcripts. Alevin-fry offers a large number of different quantification modes (He et al., 2022), and there are significant asymmetries in the quantification of single-cell vs. single-nucleus RNA-seq. For single-nucleus RNA-seq, typically all reads mapping to gene bodies are included in generation of a single count matrix, regardless of whether the reads originate from processed or unprocessed transcripts. For single-cell RNA-seq, great care is taken to avoid the counting of reads definitely originating from unprocessed transcripts, and only reads originating from processed transcripts, or ambiguous reads, are included in count matrices (Kaminow, Yunusov, and Dobin, 2021; He et al., 2022).

We address these shortcomings and inconsistencies based on a novel k-mer based method we develop for resolving reads as to their originating source. By utilizing k-mers, our approach has the benefit of being efficient as it is compatible with pseudoalignment, and we show via an implementation in the kallisto software (Bray et al., 2016; Melsted et al., 2021) that it yields a fast and efficient approach for quantification. Crucially, we introduce an approach to quantification of single-nucleus RNA-seq that focuses on the unprocessed transcripts, thereby mirroring the



Ground truth	read 1	read 2	read 3	read 4	read 5	read 6	read 7
	A	A	S	U	U	U	A
kallisto	read 1	read 2	read 3	read 4	read 5	read 6	read 7
mature + D-list	✓	✓	✓				✓
nascent + D-list	✓	✓		✓	✓	✓	✓
Complete	A	A	S	U	U	U	A
alevin-fry	read 1	read 2	read 3	read 4	read 5	read 6	read 7
cDNA-only	✓	✓	✓				✓
cDNA sketch	✓	✓	✓	✓		✓	✓
splici	S	S	S	U	U	U	A
STARsolo	read 1	read 2	read 3	read 4	read 5	read 6	read 7
Gene	✓	✓	✓				✓
GeneFull	✓	✓	✓	✓	✓	✓	✓
Velocyto	S	S	S	U	U	U	A

Figure 3.1: Comparison of alignment of Unprocessed, Processed, and Ambiguous reads between kallisto, STARsolo, and alevin-fry. Of the seven reads, three are Ⓤ, one is Ⓟ, and three are Ⓤ|Ⓟ. Note that alevin-fry and STARsolo attribute reads that are fully contained within a single exon to the "spliced" part of the count matrix. These sequences are however also present in the unprocessed transcripts, and should therefore be counted as ambiguous.

approach for quantifying single-cell RNA-seq that focuses on processed transcripts. This places the two assays on a (computationally) level playing field.

3.2 Results

To facilitate quantification of both unprocessed and processed mRNA transcripts, we distinguished k-mers into three categories, analogous to the three categories used for reads: $\boxed{\text{U}}$, $\boxed{\text{P}}$ or $\boxed{\text{U}}|\boxed{\text{P}}$ (Figure 3.3.B). A read can be classified as Ⓤ, Ⓟ or Ⓤ|Ⓟ based on the classification of its constituent k-mers (Methods). In order to classify k-mers without indexing all non-coding sequences, we utilized a D-list (Methods), which consists of distinguishing flanking k-mers (DFKs) that can be used to definitively assign a k-mer to a category (Methods). A read is classified as Ⓤ if it has at least one $\boxed{\text{U}}$ k-mer, as Ⓟ if it has at least one $\boxed{\text{P}}$ k-mer, and as Ⓤ|Ⓟ otherwise. Note that a read cannot have both $\boxed{\text{U}}$ and $\boxed{\text{P}}$ k-mers except in the rare cases where an exon is short enough that a read can span the junctions on either side of it, and that in the majority of casesthe DFKs suffice to classify all relevant k-mers for classifying a read (Methods).

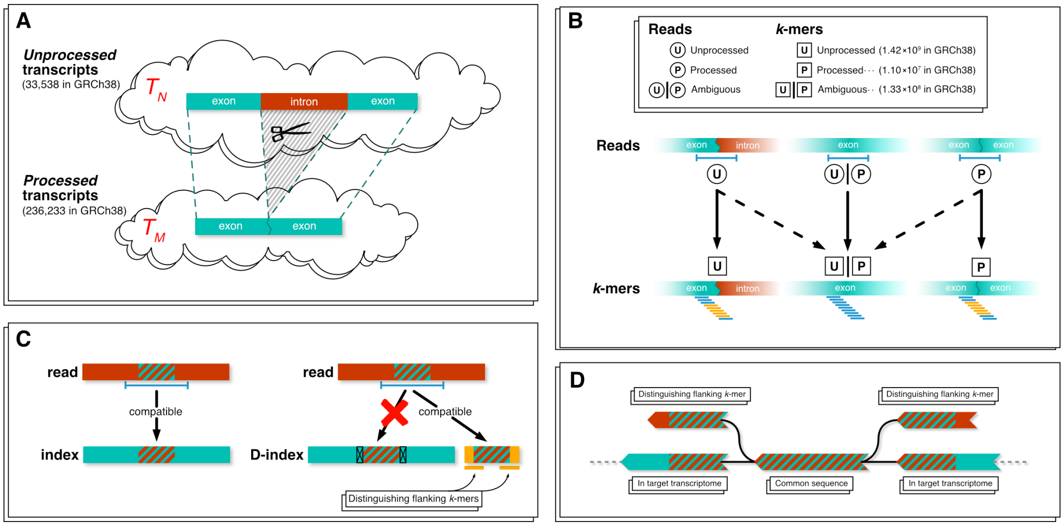


Figure 3.2: **A.** In any cell, there exist two sets of transcripts: the *unprocessed* set from which not all introns have been spliced, and the *processed* one, from which they have. The number of completely unprocessed and processed transcripts was obtained from version 104 of the Ensembl annotation of the human genome (Cunningham et al., 2022). **B.** Each read in an RNA-seq experiment is either explicitly Ⓤ, an expression of a unprocessed transcript, if it contains at least a part of a non-retained intronic sequence, Ⓟ, an expression of a mostly or completely processed transcript, if it contains an exon-exon boundary, or Ⓤ|Ⓟ, ambiguous, if it occurs in the interior of an exon. Furthermore, the k-mers that constitute an Ⓤ read may be $\boxed{\text{U}}$, nascent or $\boxed{\text{U}}\|\boxed{\text{P}}$, ambiguous, if they also occur in an exon, and the k-mers that constitute a Ⓟ, processed read may be $\boxed{\text{P}}$, processed if they contain the exon-exon junction or $\boxed{\text{U}}\|\boxed{\text{P}}$, ambiguous, if they are an interior k-mer in an exon. The number of unique $\boxed{\text{U}}$, $\boxed{\text{P}}$, and $\boxed{\text{U}}\|\boxed{\text{P}}$ k-mers in the human transcriptome was calculated from Ensembl version 104. **C.** A non-transcriptomic read containing a subsequence of length greater than k, which also occurs in a transcript in the target transcriptome will get attributed to that transcript. Distinguishing flanking k-mers (DFKs) can be used to determine whether a read compatible with a reference transcriptome may have originated from elsewhere in the genome. Using the entire set of GRCh38 scaffolds to construct a D-list for a kallisto index built from the completely unprocessed transcripts in version 104 of the Ensembl annotation yields 7,192,804 DFKs. **D.** A de Bruijn graph representation of DFKs

To validate our method for classifying reads, we generated 5,000,000 completely processed reads and 5,000,000 completely unprocessed reads directly from the respective transcripts without error, and assessed whether we could correctly assign reads when using kallisto with a D-list included in the index (Figure 3.3, Methods). We found that using a D-list, we can reliably classify reads as unprocessed, processed or ambiguous, the latter one by identifying reads that are classified as both unprocessed and processed. The alevin-fry method failed to classify all reads correctly. In particular, it assigned all Ⓤ|Ⓟ ambiguous reads, e.g. reads that are contained within a single exon, to the "spliced" count matrix corresponding which counts processed molecules. This leads to an asymmetry between quantifications of single-cell RNA-seq data and single-nucleus RNA-seq data, where for the latter, alevin-fry (He et al., 2022), STARsolo (Kaminow, Yunusov, and Dobin, 2021) and Cell Ranger (unpublished) always classify ambiguously mapped reads as processed regardless of assay. We benchmarked alevin-fry, and STARsolo in both "Gene/GeneFull"-mode and "Velocyto"-mode. STARsolo quantification in "Velocyto"-mode of the simulated data (Figure 3.3.C) deviated from the ground truth to such an extent that comparisons with kallisto and alevin-fry were meaningless. When processing reads from processed transcripts, STARsolo in "Velocyto"-mode only mapped 825,141 (82.5%) of the reads. When processing reads from unprocessed transcripts STARsolo only mapped 847,392 (84.7%) of the reads, and of those only 85.4% were mapped correctly.

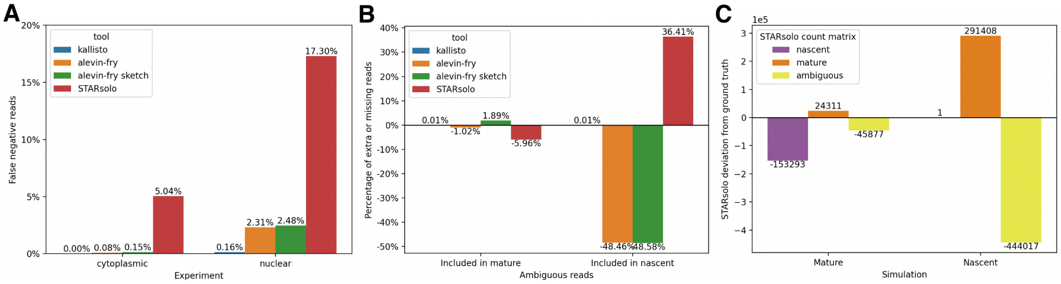


Figure 3.3: **A.** Left: the percentage of Ⓟ reads, not correctly identified as processed by kallisto, alevin-fry, and STARsolo in "Gene"-mode in a simulated single-cell RNA-seq experiment. Right: the percentage of Ⓤ reads, not correctly identified as unprocessed by kallisto, alevin-fry, and STARsolo in "GeneFull"-mode in a simulated single-nucleus RNA-seq experiment. **B.** Alevin-fry attributes ambiguous reads, e.g. reads that are contained within a single exon, to processed transcripts, leading to an asymmetry between quantifications of simulated single-cell, and single-nucleus RNA-seq reads. **C.** STARsolo quantification in "Velocyto"-mode of the simulated data. Left: simulated experiment with 1,000,000 reads containing no errors, from processed transcripts, 59,850 of which were ambiguous, occurring in both unprocessed and processed transcripts. Right: simulated experiment with 1,000,000 reads containing no errors, from unprocessed transcripts, of which 583,914 were ambiguous.

To better understand the performance of kallisto on data that includes errors, we assessed its performance using a simulation framework developed by the authors of STARsolo (Kaminow, Yunusov, and Dobin, 2021). In that simulation framework, errors were introduced into reads at 0.5% error rate, and reads were simulated from both coding and non-coding genomic sequence to mimic the presence of both unprocessed and processed transcripts in single-cell RNA-seq experiments. We also followed the assessment framework of (Kaminow, Yunusov, and Dobin, 2021), and compared the results of kallisto to STARsolo and alevin-fry. We found that kallisto performed similarly to STARsolo in a simulation containing multi-mapping reads, i.e. reads that align well to two or more distinct transcripts, and outperformed alevin-fry (Figure 3.4.B). In another simulation, containing no multi-mapping reads, STARsolo performed marginally better than both kallisto and alevin-fry (Figure 3.4.A). The poor performance of kallisto without the D-list is due to the unreasonable assessment procedure of (Kaminow, Yunusov, and Dobin, 2021), which omitted assessment of true negatives (Methods). Note that correct calculations of the Spearman coefficient yield a negligible difference between kallisto, STARsolo, and alevin-fry (Table 3.1, table 3.2).

Per-cell Spearman coefficient quantiles

Tool	**25%**	**50%**	**75%**
STARsolo	0.9893	0.9922	0.9948
kallisto (with D-list)	0.9678	0.9735	0.9789
alevin-fry splici cr-like	0.9665	0.9717	0.9763
alevin-fry splici cr-like-em	0.9208	0.9291	0.9384
alevin-fry cr-like	0.8604	0.8824	0.9157
alevin-fry cr-like-em	0.8294	0.8559	0.8966
kallisto	0.7594	0.7901	0.8381
alevin-fry sketch cr-like	0.7527	0.7830	0.8337
alevin-fry sketch cr-like-em	0.6932	0.7282	0.7891

Table 3.1: STARsolo, kallisto, and alevin-fry compared in the simulated single-cell RNA-seq experiment by Kaminow et al. containing no multimapping reads. The Spearman coefficient of the quantified abundances and the ground truth per cell was calculated and the quantiles reported. As per (Kaminow, Yunusov, and Dobin, 2021) the true negatives were left out of the abundance vectors used to calculate the Spearman coefficient, which artificially inflates the effect of false negatives and false positives on the coefficient.

To understand the implications of correct classification of reads into the unprocessed transcript, processed transcript, or ambiguous categories, we examined the correlation in gene counts with and without the use of a D-list on both single-cell RNA-seq and single-nucleus RNA-seq data (Methods). The overall result was not materially different for single-cell RNA-seq (Figure 3.5.A), with the Pearson correlation between kallisto and kallisto with the D-list at 0.99, corroborating the results of (Booeshaghi and Pachter, 2021). Similarly, kallisto with the D-list is highly similar to alevin-fry (Figure 3.5.B) and STARsolo (Figure 3.5.C). Even though quantifi-

Per-cell Spearman coefficient quantiles

Tool	25%	50%	75%
STARsolo	0.9961	0.9969	0.9977
kallisto (with D-list)	0.9885	0.9903	0.9919
alevin-fry splici cr-like	0.9909	0.9922	0.9933
alevin-fry splici cr-like-em	0.9740	0.9761	0.9782
alevin-fry cr-like	0.9468	0.9539	0.9635
alevin-fry cr-like-em	0.9427	0.9503	0.9610
kallisto	0.8792	0.8939	0.9139
alevin-fry sketch cr-like	0.8740	0.8893	0.9104
alevin-fry sketch cr-like-em	0.8564	0.8737	0.8988

Table 3.2: STARsolo, kallisto, and alevin-fry compared in the simulated single-cell RNA-seq experiment by Kaminow et al. containing no multimapping reads. The Spearman coefficient of the quantified abundances and the ground truth per cell was calculated and the quantiles reported. Unlike in (Kaminow, Yunusov, and Dobin, 2021), the true negatives were accounted for in the abundance vectors used to calculate the Spearman coefficient, which yields a meaningful measure of correlation.

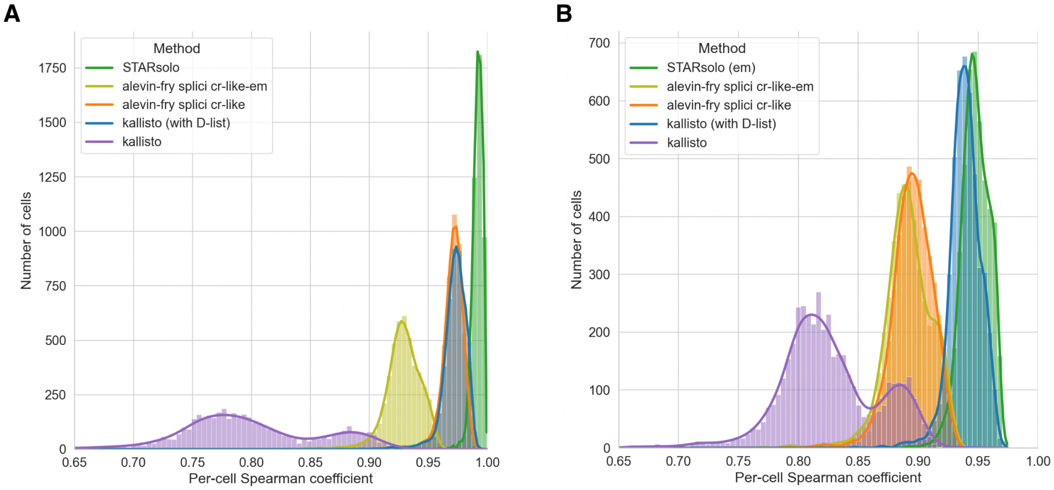


Figure 3.4: The Spearman coefficient for the correlation between the simulation ground truth expression and the expression quantified by kallisto, alevin-fry, and STARsolo in the simulation framework developed by Kaminow, Yunusov, and Dobin (2021). **A.** Reads from 4,548 cells, containing no multi-gene reads. **B.** Reads from 4543 cells, including multi-gene reads.

cation with the D-list does not affect quantifications much, its use does not affect running times much (Figure 3.6A) so it can be used routinely regardless.

More interesting, is the quantification of single-nucleus RNA-seq, for which unprocessed RNAs can be quantified by generating a D-list based on DFKs in processed transcripts. This biologically motivated approach to quantification of unprocessed RNAs, that counts only reads that are definitively unprocessed or ambiguous (but

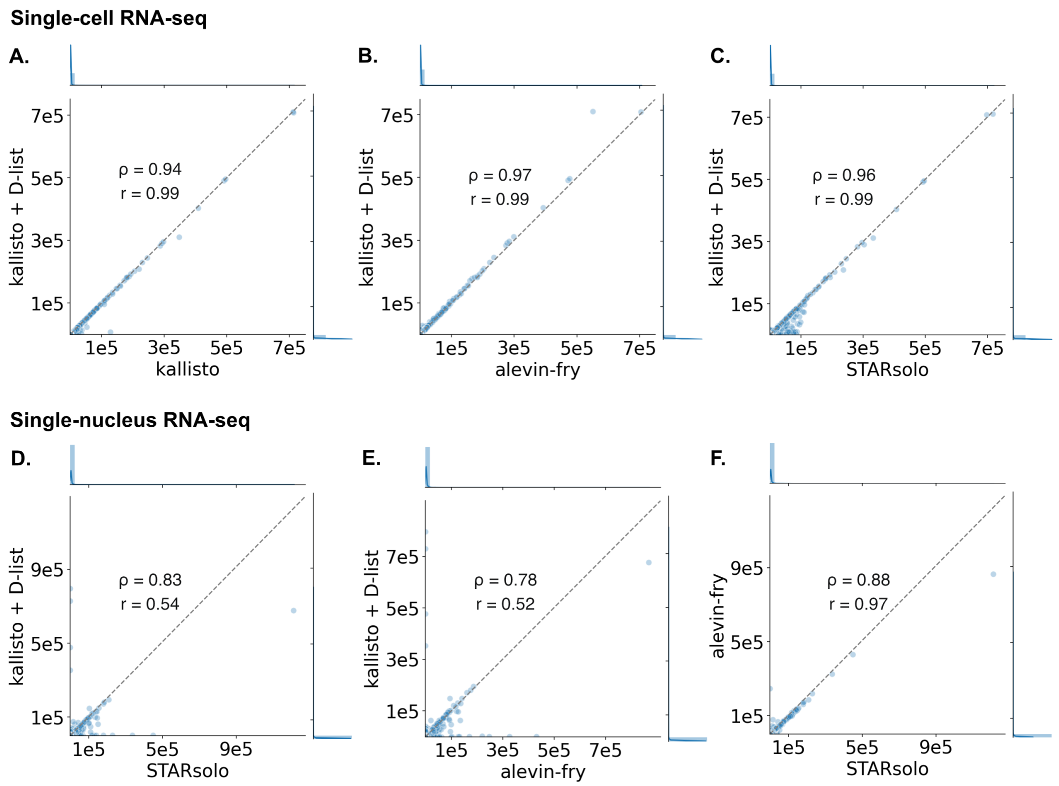


Figure 3.5: Comparison of kallisto with alevin-fry and STARsolo on 10x Genomics single-cell RNA-seq and single-nucleus data (Methods). **A.** Assessment of the effect of using the D-list with kallisto. **B.** Comparison of kallisto to alevin-fry on single-cell RNA-seq **C.** Comparison of kallisto to STARsolo on single-cell RNA-seq **D.** The difference between kallisto's quantification of unprocessed transcripts from single-nucleus RNA-seq, and alevin-fry's quantification of single-nucleus RNA-seq. **E.** The difference between kallisto's quantification of unprocessed transcripts from single-nucleus RNA-seq, and STARsolo's quantification of single-nucleus RNA-seq. **F.** The similarity of alevin-fry and STARsolo's quantification of single-nucleus RNA-seq. In all plots Spearman correlation is shown with ρ and Pearson correlation with r.

not processed mRNAs), is practical for biological data, and produces results that are significantly different from current approaches that quantify single-nucleus RNA-seq by agglomerating counts for unprocessed and processed transcripts together. This is reflected in a comparison of kallisto with the D-list to alevin-fry (Figure 3.5.C) and to STARsolo (Figure 3.5.D). STARsolo and alevin-fry, which both quantify single-nucleus by mixing up processed and unprocessed transcripts, are more similar to each other (Figure 3.5.E).

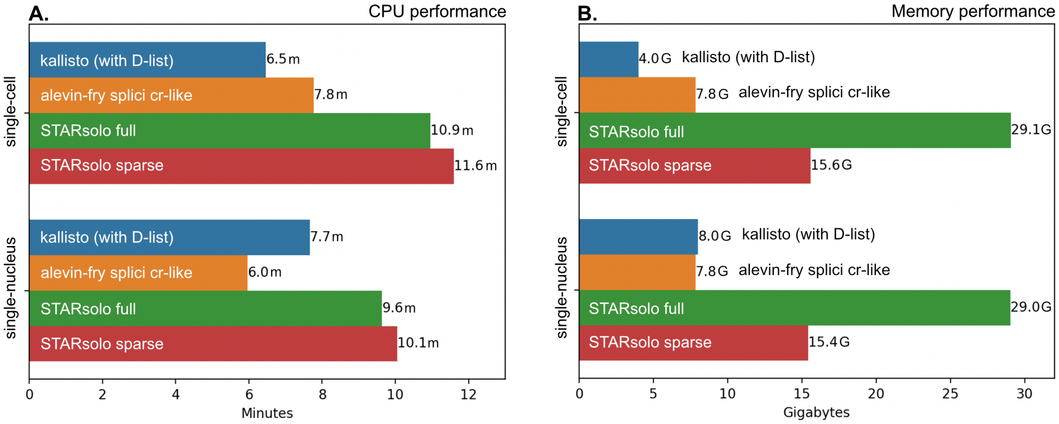


Figure 3.6: Running time and peak memory usage comparisons during quantification of 204,596,690 single-cell and single-nucleus RNA-seq reads from 7,377 mouse brain cells. All programs were allocated 20 threads for quantification (Methods).

3.3 Discussion

The use of single-nucleus RNA-seq in lieu of single-cell RNA-seq has been increasingly popular, primarily because nuclei don't need to be dissociated (Habib et al., 2017; Cervantes-Pérez et al., 2022; Liang et al., 2019; Al-Dalahmah et al., 2020). Yet, despite the increasing preponderance of single-nucleus RNA-seq data and fundamental differences between it and single-cell RNA-seq data, the two data types have not been treated similarly for quantification purposes. While processed transcripts are quantified for single-cell RNA-seq, single-nucleus RNA-seq is quantified by combining quantifications of both unprocessed and processed transcripts. Our approach to quantification offers flexibility for teasing apart the counts of unprocessed vs. processed mRNAs from such data, and highlights the possibility for quantifying unprocessed transcripts from single-nucleus RNA-seq, just as processed transcripts are quantified from single-cell RNA-seq. Our method is based on an efficient and accurate algorithm, and is implemented in software that can form part of reproducible workflows that have modest hardware requirements. Furthermore, the generation of unprocessed and processed transcripts counts will prove invaluable for methods that rely on such counts for downstream analysis. Arguably, all single-nucleus and single-cell RNA-seq should first involve integrating the unprocessed and processed transcript modalities using a biophysical model of transcription (Gorin and Pachter, 2022a), although investigation of that hypothesis is beyond the scope of this work.

There are several limitations to the quantification framework we have proposed. In a cell, the set of unprocessed mRNAs at any given time is likely to include partially processed molecules (Pai et al., 2018), and in principle the complete splicing cascade must be understood and known in order to accurately quantify single-nucleus or single-cell RNA-seq data (Gorin and Pachter, 2022b). Furthermore, the use of ambiguous reads both for single-cell and single-nucleus RNA-seq is unsatisfactory. Ideally reads should be longer so that they can be uniquely classified, or they should be fractionally classified according to probability estimates of the ratio between unprocessed and processed transcripts. The latter approach is non-trivial due to variation in effective transcript lengths that will depend on library preparation and must be accounted for (Pachter, 2011), but this is an interesting direction of study.

Finally, we believe the memory efficiencies introduced with the update to kallisto prepared for this work (Figure 3.6.B), will greatly extend its utility for a variety of other applications, such as single-cell genomics assays that generate reads that must be aligned to the genome (Gao and Pachter, n.d.) and metagenomics (Schaeffer et al., 2017).

3.4 Methods

The D-list

A D-list (distinguishing list) enables accurate quantification of RNA-seq reads in experiments where reads that are not an expression of the target transcriptome may still contain sequences, which do occur in the target transcriptome. Without the D-list, these reads may be erroneously quantified as transcripts in the target transcriptome, based on alignment of the common sequences. Thus, the D-list may contain any sequences that are not desired in the abundance matrix yielded by the quantification. They may be the transcriptomes of other organisms, in case the RNA sample is contaminated, they may consist of the unprocessed (unspliced) versions of the processed (spliced) transcripts in the target transcriptome, in case only the quantification of processed transcripts, or only the quantification of unprocessed transcripts is desired, or they may contain common transposable elements, such as Alu regions, which might otherwise introduce undesired noise. The D-list is incorporated into the index by finding all sequences, k base-pairs or longer, that occur in both the D-list and the target transcriptome. The first k-mer upstream and the first k-mer downstream of each such common sequence in the D-list are added to the index colored *de Bruijn* graph (Figure 3.2.C, Figure 3.2.B). We refer to these new vertices in the graph as distinguishing flanking k-mers (DFKs). The DFK vertices are left uncolored in the index, such that during quantification, reads that contain them will be masked out, and go unaligned. Note that there is no need for processing sequences that map to the DFK vertices in special cases downstream.

As an illustration, when mapping a Ⓟ read containing both $\boxed{\text{P}}$ and $\boxed{\text{U}}\|\boxed{\text{P}}$ k-mers to an index built from Ⓤ transcripts, the $\boxed{\text{U}}\|\boxed{\text{P}}$ k-mers will be found in the index, whereas the disambiguating $\boxed{\text{P}}$ k-mers will not. The whole read will be erroneously mapped based on the $\boxed{\text{U}}\|\boxed{\text{P}}$ ambiguous k-mers that are present in the index. By finding all $\boxed{\text{U}}\|\boxed{\text{P}}$ k-mers in the completely processed versions of the transcripts, and adding any distinguishing flanking $\boxed{\text{P}}$ k-mers to the index, the Ⓟ read will be masked from mapping to a Ⓤ transcript.

Recent papers have discussed various ways of reducing the number of false positives in RNA quantification. The simplest way of doing so is to align the RNA-seq reads against both the target transcriptome and a secondary transcriptome containing undesired transcripts. This has been explored in (Srivastava et al., 2020), and while it may yield fewer false positives than aligning against the target transcriptome alone, it is memory-intensive and, depending on the method, potentially CPU intensive. Another approach, also explored in (Srivastava et al., 20200) is to introduce a preprocessing step, wherein sequences that are similar to those in the target transcriptome are extracted, using some heuristic for similarity. These sequences are then added to the index, and handled in special cases downstream, during alignment and quantification. Most recently, alevin-fry introduced the splici index which reduces the number of false positives while controlling peak memory usage better than previous approaches. However, indexing intronic sequences, while enabling workflows like RNA velocity, still incurs a significantly larger memory cost than indexing just the target transcriptome.

Our method has the distinct advantage of incorporating only the minimum amount of data, required to disambiguate common sequences, into the index. Therefore, the memory usage and runtime of kallisto using a D-list are on par with the memory usage and runtime of kallisto, without the use of a D-list.

3.5 Validation

In addition to validating our results on the simulation framework developed by Kaminow, Yunusov, and Dobin (2021) we simulated two error-free experiments using reads generated by BBMap (Bushnell, 2014). One simulation represents a single-cell RNA-seq experiment consisting of 4,000,000 completely processed RNA reads and 1,000,000 completely unprocessed RNA reads. The other one represents a single-nucleus RNA-seq experiment consisting of 4,000,000 completely unprocessed RNA reads and 1,000,000 completely processed RNA reads. The respective ratios of processed to unprocessed reads was estimated based on (Gorin, Yoshida, and Pachter, 2022). The processed transcripts were obtained from version 104 of the Ensembl of GRCh38, and the unprocessed transcripts were taken to be the entire sequence from the start of the first exon through the end of the last exon. The kallisto quantification of the single-cell RNA-seq simulation was performed using an index built from the completely processed transcripts and a D-list constructed from all the GRCh38 scaffolds. For the single-nucleus RNA-seq simulation an index built from the completely unprocessed transcripts, with a D-list constructed from the completely processed transcripts was used.

The STARsolo simulation

We obtained the simulation framework developed by (Kaminow, Yunusov, and Dobin, 2021; He et al., 2022) from https://github.com/dobinlab/STARsoloManuscript/ and ran it as-is, substituting the deprecated “decoy”-mode of alevin-fry with its “splici”-mode replacement.

Correction of the STARsolo calculation of Spearman correlation

Kaminow, Yunusov, and Dobin (2021), calculate the correlation between quantification and the ground truth expression of a cell, only between the elements of the gene/cell count matrices which are expressed in either the simulation or tool quantification. Thus, genes for which there is no expression in either quantification or ground truth, i.e. true negatives, are ignored in the calculation of the Spearman correlation coefficient. This leads to an artificial inflation of the effect of false positives and false negatives on the coefficient. For example, consider a cell with very few genes expressed in the simulation, and as a result 0 counts for almost all genes. Now suppose a method reports only 7 genes with 1 count (out of thousands), and in the ground truth of the simulation there are also 7 genes with 1 count, with a disagreement on 2 genes, i.e. method = (1,1,1,1,0,1,1,1) and simulation = (1,1,1,0,1,1,1,1). (Kaminow, Yunusov, and Dobin, 2021) compute the Spearman correlation between these vectors, which is -0.1428571, for a cell where the method and simulation agree over thousands of genes. Table 3.1 shows the quantiles of the Spearman coefficient between the ground truth of the simulation developed by (Kaminow, Yunusov, and Dobin, 2021) and each of the quantification tool, calculated only from true positives, false positives, and false negatives. Table 3.2 shows the quantiles when the calculation attributes for true negatives. It is evident that when calculated correctly, there is negligible difference between kallisto, alevin-fry, and STARsolo with respect to the ground truth in this simulation.

Analysis of single-cell and single-nucleus RNA-seq

We quantified a 10xV3 dataset containing 204,596,690 single-cell, and single-nucleus RNA-seq reads from 7,377 adult mouse brain cells via 10x Genomics: https://www.10xgenomics.com/resources/datasets/5k-adult-mouse-brain-nuclei-isolated-with-chromium-nuclei-isolation-kit-3-1-standard, using kallisto, alevin-fry (splici em-like), and STARsolo. For quantification of the single-cell data, we constructed the kallisto index from version 108 of the Ensembl mouse annotation (Cunningham et al., 2022), using the entire mouse genome to construct the D-list. For quantification of the single-nucleus data, we constructed the kallisto index from completely unprocessed versions of the mouse transcripts (Validation), using the mature transcripts to construct the D-list. We ran STARsolo in “Gene”-mode for the single-cell data, and in “GeneFull”-mode for the single-nucleus data. All benchmarks were performed on a computer with two Intel(R) Xeon(R) Gold 6152 CPUs (a total of 44 cores), 768GiB of DDR4/2666MHz RAM, and twelve 12TB SATA hard drives. Programs were allocated 20 threads for quantifying the data, and CPU and peak memory usage were obtained via `/usr/bin/time -v`. Each program was run separately to minimize the likelihood of I/O bottlenecks.

Code availability

kallisto is available under the BSD-2-Clause license. The version used for this paper is available at `https://github.com/pachterlab/kallisto-D`. All code for simulations and downstream analyses is available at `https://github.com/pachterlab/HSHMP_2022`.

Chapter 4

MEMORY IMPROVEMENTS IN KALLISTO 0.49

Kallisto performs quantification of RNA-Seq data by building a *de Bruijn* graph (dBG) of the underlying transcriptome, and aligning the RNA-Seq reads to the graph. Eschewing the traditional alignment algorithms (Smith-Waterman, Needleman-Wunsch, Burrows-Wheeler, et cetera), kallisto instead exactly matches subsequences of length k, colloquially referred to as k-mers, to k-mers that occur in the graph, yielding what has been termed a *pseudoalignment* (Bray et al., 2016). Traversing the graph to find a vertex containing a specific k-mer would be glacially slow, so instead we keep track of which vertex in the graph contains which k-mers in an index. In tandem with the new, more targeted pseudoalignment paradigm we developed in the previous chapter, we undertook a major reimplementation of the main data structures used by kallisto, yielding a reduction in memory on the order of 80% while maintaining comparable running times. This chapter illustrates the key differences between the index used by kallisto up to and including version 0.48, and that used by kallisto as of version 0.49.

4.1 A primer on *de Bruijn* graphs

A k-dimenstional *de Bruijn* graph $G(V, E)$ on an alphabet Σ is a directed graph of overlapping strings of length k with $V \subseteq \Sigma^k$. For vertices $a = a_1 \cdots a_k \in V, b = b_1 \cdots b_k \in V$, the edge $E(a, b)$ exists if the last $k-1$ symbols in a are the first $k-1$ elements in b, i.e.

$$a_2 \cdots a_k = b_1 \cdots b_{k-1}.$$

For the sake of memory efficiency, if $E(a, b)$ exists, and the out-degree of a and in-degree of b are both 1, we may contract a and b into a single node of length $k+1$. We note that each k-mer is unique in the graph, which in turn means that no string of length $k+1$ is ever repeated in the graph. A *de Bruijn* graph is then a data structure which enables a compact representation of a set of strings that share common substrings. The amount of compression is a function of k, but also of the similarity between the strings in the set. An example of a simple *de Bruijn* graph can be found in Figure 1.4.

4.2 k-mer lookup vs. minimizer lookup

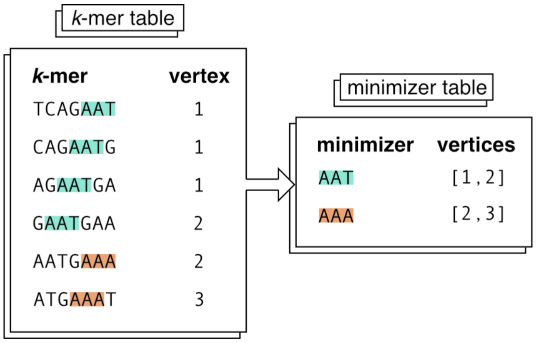
Prior to version 0.49, kallisto kept in its index a hash table that stored every single k-mer in the graph, along with a pointer to the (unique) vertex that contained it. For small graphs and/or small values of k this is convenient, but as the graph grows larger, this hash table becomes prohibitively large. As an illustration, $k = 31$ yields

$$4^{31} = 4.6 \times 10^{18} \text{ possible } k\text{-mers},$$

given an alphabet of size 4. In order to make large quantification problems tractable, the k-mers are replaced with *minimizers* (Roberts et al., 2004). This is similar to the concept of *winnowing*, introduced in Schleimer, Wilkerson, and Aiken, 2004. A minimizer is defined by three parameters $(g, k, \mathcal{O})$. Given an ordering $\mathcal{O}$ of minimizers in Σ^g, where Σ is an alphabet, a minimizer of length $g < k$ of a k-mer is the smallest g-mer occurring in the k-mer. Note that consecutive, overlapping k-mers have a substantial probability of sharing the same minimizer, if k and g are selected appropriately. Thus, the projection of a k-mer into minimizer space can be thought of a hash function, with an explicitly nonuniform collision rate. The more overlapping k-mers are, the more likely they are to share a minimizer, and therefore collide. Thus, consecutive k-mers in the graph are likely to share a minimizer. Intuitively, we might assume that lexicographical order of minimizers would be sufficient, but due to the structure of genetic sequences, this is problematic. As an example, sequences consisting of just `A` are abundant in genetic sequences, so the lexicographically smallest minimizer would occur in a large amount of k-mers, regardless of their locality. As of kallisto 0.49, a hash table mapping minimizers to vertices is stored in the index, instead of the hash table mapping k-mers to vertices in the index. The minimizer-based *de Bruijn* graph used by kallisto 0.49 is implemented in Bifrost (Holley and Melsted, 2020). Whereas k-mers are unique in the dBG, g-mers are not. I.e. k-mers that do not occur in the same vertex could share a minimizer and therefore two or more distinct vertices in the graph could contain the same minimizer. When we want to query which vertex in the dBG contains a given k-mer, we

1. Calculate the minimizer of the k-mer;
2. Look up the minimizer in the minimizer hash table;
3. For each vertex containing the minimizer, assess whether it contains the original k-mer (which is unique in the graph).

In theory, by indexing minimizers instead of k-mers, we therefore sacrifice lookup time in order to save memory, but in practice, the processing time is on par with earlier versions of kallisto, or better. In practice, for $k = 31$, $g = 23$ is a heuristic that yields a favorable trade-off between speed and memory. The largest memory savings from storing minimizers occurs when all k-mers in the graph share the same minimizer. This is not useful for our purposes, since the end goal is to be able to quickly determine which node in our graph contains the k-mer. If all k-mers share the same minimizer we have reverted back to a linear search over all vertices in the graph. The optimal memory reduction conditioned on maintaining close to constant time lookup is attained when all k-mers in each vertex share the same minimizer, and no two vertices share the same minimizer. Let n be the number of unique k-mers in the graph. This lower bound can then be expressed as



k-mer table

k-mer	vertex
TCAGAAT	1
CAGAATG	1
AGAATGA	1
GAATGAA	2
AATGAAA	2
ATGAAAT	3

minimizer table

minimizer	vertices
AAT	[1,2]
AAA	[2,3]

Figure 4.1: An example of a k-mer table and corresponding minimizer table for $k = 7, g = 3$ using lexicographical ordering for illustration purposes. Minimizers are means of hashing k-mers, such that collisions between k-mers that occur close to each other in a sequence are more likely than collisions between k-mers that are far apart.

$$b_g = \log_2(|\Sigma|) \cdot g \cdot |V| \text{ bits},$$

whereas the minimal number of bits required to store the n original k-mers is

$$b_k = \log_2(|\Sigma|) \cdot k \cdot n \text{ bits}.$$

The optimal memory reduction is then

$$r = 1 - \frac{b_g}{b_k} = 1 - \frac{|V|}{(k-g)n}.$$

In real-world experiments however, finding an order for the minimizers that satisfies the optimal case is at the best of times computationally intensive, but at the worst impossible. A random ordering of minimizers, e.g. by hashing, has been found to yield a good compromise.

4.3 Shrinking the hash table to size

Traditional (open accessing) hash tables require more space than they have keys. This is because two or more keys can hash to the same value, resulting in a collision which at lookup time needs to be resolved by direct comparison. When a collision occurs, we start at the hash value and search linearly for the key if performing a lookup or for the next empty slot if performing an insertion. As the hash table is populated with more and more values, the probability of a collision, and therefore a linear search, increases. As a heuristic, we may want to resize our hash table when we reach 80% capacity, in order to reduce the number of collisions. This comes with a penalty though, as it requires rehashing every key in the table with a new hash function. Chaining hash tables, which are generally implemented using linked lists, suffer from the same linear search fallback in the case of collisions. In our case, once a kallisto index has been built, the complete set of keys is known and no new minimizers will be added to the graph. This enables us to select a new hash function that solves two problems. Firstly, it removes the need for the extra capacity required to minimize the penalty for collisions, and secondly, it removes the penalty for a collision, namely the linear search.

We replace the old, dynamic hash table with a new, static one, using a minimal, perfect hash function (MPHF). Minimal, because it allows us to reduce the capacity to 100% of the number of keys and perfect because for the predefined set of keys in the table there are no collisions. More formally, given a set S of k keys, an MPHF is an injective function that maps each key in S to an integer in the range $[0, k - 1]$, thereby labeling each key with a unique integer, such that there are no collisions between labels of different keys, and the set of integers chosen is dense. The minimal memory to store such a function for a set of k keys has been shown to be $\log_2(e) \cdot k$ bits (Mehlhorn, 1982; Fredman and Komlós, 1984), but in practice, for large sets of keys the constant will be larger. The MPHF implementation used by kallisto 0.49 is provided by BBHash (Limasset et al., 2017).

4.4 Equivalence class preemption vs. dynamic equivalence classes

Kallisto assigns *equivalence classes* (ECs) to k-mers in the dBG. The equivalence class of a k-mer is the set of transcripts in which the k-mer occurs as a sub-sequence. It then colors each vertex in the dBG with the equivalence class that pertains to the set of transcripts that use that vertex. When attributing a read to a specific transcript, kallisto

1. Finds the vertices containing the k-mers that occur in the read;
2. Gathers the sets of transcripts that constitute the ECs of those vertices;
3. Takes the intersection of those sets.

If there is only one transcript in the intersection, the read will be attributed to that transcript, but if there are more it is fractionally assigned using an EM algorithm. Prior to kallisto 0.49, equivalence classes were preemptively generated for all of the vertices in the dBG, and stored in a data structure in memory. While easier to implement, this results in a very large number of equivalence classes being preempted, with proportionally few ever being used by any read. To add insult to injury, kallisto requires both a mapping from a set of transcripts to its EC *and* a mapping from an EC to its set of transcripts, resulting in an even larger memory footprint. As of kallisto 0.49, the vertices in the dBG are colored with the sets of transcripts that use them, rather than with the explicit EC. As a result, an EC only gets created once the k-mers in a read have been found in vertices in the dBG and the sets of transcripts associated with those vertices have been intersected. We store the equivalence classes in compressed bitmaps, implemented in Roaring (Lemire et al., 2017). Furthermore, we cap the size of equivalence classes at 250 transcripts, as very large equivalence classes are unlikely to affect the intersection of the transcript sets, while taking up a lot of space in memory.

Chapter 5

ANNOTATION-AGNOSTIC DISCOVERY OF ASSOCIATIONS BETWEEN NOVEL GENE ISOFORMS AND PHENOTYPES

In this chapter we present a novel method for associating phenotypes with RNA expression, that can identify expression associations resulting from a wide variety of underlying transcriptional and post-transcriptional events, without relying on annotations of these events. We show that we can reliably detect, de novo, phenotypically relevant transcriptional structures.

5.1 Introduction

The quantification of RNA reads is a key step in most analyses of RNA-seq data (Kukurba and Montgomery, 2015). Current quantification methods rely on annotations of the organisms' transcriptomes, which may be incomplete or nonexistent (Zhang et al., 2020). This may result in data being discarded and can lead to erroneous quantifications. In downstream applications such as eQTL analysis, such errors can propagate and result in missed, or erroneous associations (Saha and Battle, 2018). We present a novel method for associating phenotypes with RNA expression, that can identify expression associations resulting from a wide variety of underlying transcriptional and post-transcriptional events, without requiring a prior annotation of the transcriptome. By constructing a de Bruijn graph of all the reads overlapping a single gene, and pruning away nodes that are likely to be due to sequencing errors, we obtain a representation of the expression of the gene in our cohort. Each expressed isoform constitutes one path through the graph. We then run associations on the expression of each individual node and a phenotype. Should an isoform of the gene associate with the phenotype, there will be a set of nodes in the graph that uniquely identify the isoform, the expression of which also associates with expression of the phenotype. This method enables discovery of novel alternative polyadenylation, exon-skipping, duplications, insertions, deletions, and circular RNA, among other transcriptional and post-transcriptional variations, without prior knowledge of these events. We show that we can reliably reproduce known associations, and detect, de novo, phenotypically relevant transcriptional structures.

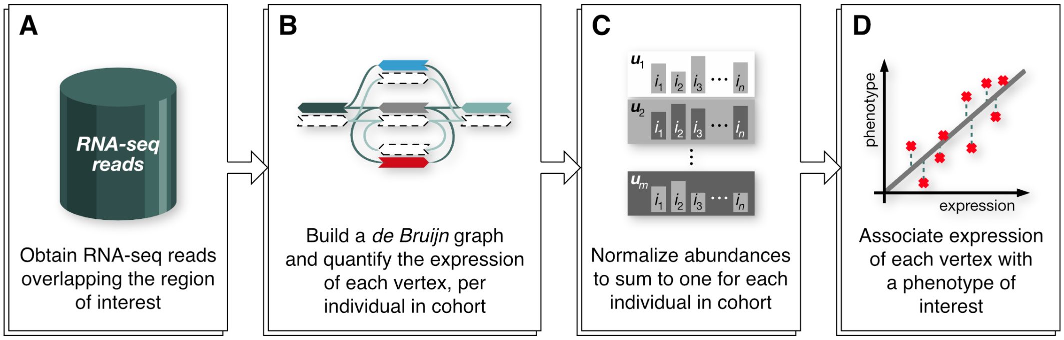

A
RNA-seq reads
Obtain RNA-seq reads overlapping the region of interest
B
Build a de Bruijn graph and quantify the expression of each vertex, per individual in cohort
C
u_1
i_1 i_2 i_3 ⋯ i_n
u_2
i_1 i_2 i_3 ⋯ i_n
⋮
u_m
i_1 i_2 i_3 ⋯ i_n
Normalize abundances to sum to one for each individual in cohort
D
phenotype
expression
Associate expression of each vertex with a phenotype of interest

Figure 5.1: The annotation-agnostic association process consists of four steps. **A.** We obtain a dataset of RNA-seq reads from a cohort of individuals, overlapping the genomic region of interest. **B.** A *de Bruijn* graph is constructed from the dataset, and the individuals' expression of each vertex is quantified. An optional pruning step can remove vertices that are likely to be erroneous (either due to genetic or technical noise) in order to reduce the number of association targets (See Methods). **C.** The vertex abundances are normalized such that the individuals' expression sums to one, respectively. **D.** The normalized expression of each vertex is associated with a phenotype of interest, and the p-values of the associations are aggregated using the harmonic mean p-value (See Methods)

5.2 Methods

RNA-Seq reads overlapping a genomic region of interest, e.g. a gene, are obtained from a cohort of people for which a phenotype of interest is available. A bi-directed de Bruijn Graph (dBG) is constructed, using Bifrost (Holley and Melsted, 2020), from these reads with k-mer size $k = 31$ and then compacted such that consecutive k-mers with out-degree 1 and in-degree 1 respectively are folded into a single, maximal unitig, which is a high-confidence contig. Each path between two unitigs represents distinct ways the corresponding part of the gene might be expressed in the cohort. Each individual's expression of each unitig in the graph is then quantified (Figure 5.2).

Quality filter

The size of the dBG is dependent on several factors: genetic variation, errors introduced in sequencing, the size and relatedness of the cohort, the length and number of different isoforms of the gene, the expression levels of that gene in the tissue of interest, etc. In order to reduce the scope of the problem and thereby reduce the number of association tests performed, we may prune from the graph nodes that are likely present due to noise, either genetic or technical. To that end, all k-mers that occur in common transposable elements, such as Arthrobacter luteus

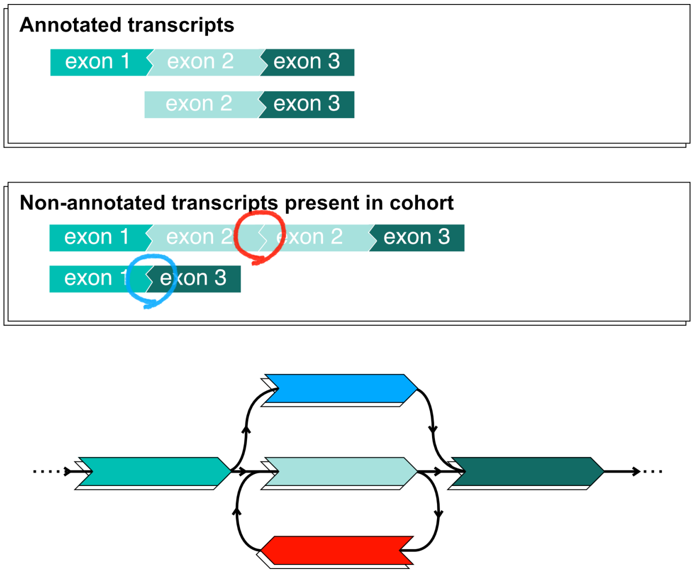


Figure 5.2: A *de Bruijn* graph representing annotated and non-annotated transcripts. The two annotated transcripts can be represented with the three green nodes. The first non-annotated transcript contains a duplication of exon 2. In order to represent that transcript we add a node whose sequence is the last $k - 1$ base pairs in exon 2 followed by the first $k - 1$ base pairs in exon 2. The second non-annotated transcript skips over exon 2. In order to represent that transcript, we add a node whose sequence is the last $k - 1$ basepairs in exon 1 followed by the first $k - 1$ base pairs in exon 3. In this example, the first non-annotated transcript can be uniquely identified by expression of the red vertex. If expression of the transcript correlates with a phenotype, then expression of the node will necessarily also correlate with that phenotype.

(Alu) regions are removed from the graph. Furthermore, the median abundance of isoforms of the gene that are present in an annotation of the target transcriptome are found for each individual in the cohort, and unitigs that are expressed less than 0.5% of the transcriptomic median abundance are deemed to be erroneous. Unitigs with less than 50 counts associated with them are taken to be erroneous if there is another unitig within Hamming distance 1, with more than four times their expression. Finally, tips, i.e. short chains of nodes that are disconnected on one end (Zerbino and Birney, 2008) are removed.

Associations

Unitig abundances are normalized to sum up to 1 for each individual, in order to capture associations between genotypes and phenotypes on one hand, and isoform-specific expression rather than individual coverage, on the other. The normalized abundances for each unitig respectively, are associated with a qualitative or a qualitative phenotype. Crucially, the expression of sequences from any isoforms containing structural variants, which are not part of any transcriptome annotation, are implicitly associated with the phenotype. For any such novel isoform, there will be a set of subsequences, the expression of which uniquely distinguishes the isoform from other transcripts. Since the relative expression levels of different isoforms of the same gene are not generally independent, the resulting p-values for the individual unitigs are aggregated using the Harmonic Mean P-value (HMP) (Wilson, 2019) with weights equal to the log-transformed mean counts normalized to 1, i.e. given a dBG with N unitigs with mean counts $u_1, \cdots, u_N$, the weight for the p-value of unitig i is $w_i = \frac{\log(ui+1)}{\sum_j \log(uj+1)}$ (Yi et al. 2018). Distinct genes are assumed to be expressed independently, and the aggregated p-value is Bonferroni corrected for the number of protein-coding genes in the target transcriptome.

Simulated differential transcript usage experiments

In order to assess the sensitivity of the method, we simulated bulk RNA-Seq datasets with novel structures and simulated phenotypes that correlated with those novel structures. While methods to generate signals that do not require theoretical models exist (Gerard, 2020), these generate their signal using reads from real RNA-seq datasets. In our case we need to simulate novel isoforms as well the signals correlated with them. Basing the signal simulation on real-world RNA-seq datasets is not tenable in our case, since they would not contain the novel isoform. A protein-coding gene was arbitrarily selected from version 108 of the Ensembl annotation of the GRCh38 assembly (Cunningham et al., 2022). An alternative transcriptome T_{alt} was generated from the reference transcriptome T_{ref} by adding a novel isoform based on an existing isoform, but containing a duplication, exon skipping, alternative polyadenylation, or circular structure not present in the original. A cohort of size 1,000 was created, 10 of which were chosen to express the novel structure, for a prevalence of 1%. Single-ended reads overlapping the chosen protein-coding gene were generated for all 1,000 individuals in the cohort using BBMap (Bushnell, 2014). For the affected cohort, the reads were generated from T_{alt}, whereas for the remaining 990 wild-type individuals the reads were generated from T_{ref}. Various different rates for SNPs and indels were used in order to assess robustness to noise. Qualitative phenotypes were obtained by assigning the affected individuals expressing the novel structure the phenotype 1 and others 0. Varying the allele frequency of SNPs and indels gives us an idea of the level of robustness to biological noise. A number of simulations were run, with the numbers of reads per individual varying from 1 to 100, in order to assess the method's sensitivity to coverage, with 25 reads per gene taken to be a reasonable coverage per (Svensson, Natarajan, et al., 2017), and assuming 19,116 protein-coding genes in the transcriptome (Piovesan et al., 2019). Simulating different levels of coverage shows us the sensitivity of the method as a function of the expected number of reads overlapping an identifying sequence of the novel isoform, given by

$$\mathbb{E}[\text{number of overlapping reads}] = C^* \cdot \frac{1}{|t|} \cdot \frac{L-1}{l(t^*)},$$

where C^* denotes the total number of reads in affected individuals, t denotes the set of isoforms, L denotes the expected read length, and $l(t^*)$ denotes the length of the novel isoform. Expected coverage of an identifying sequence in the range of $[1, 50]$ was simulated and quantified using the method [Figure 5.3.A]. Technical noise was simulated by adding or subtracting from each unitig for each individual a number of reads drawn from a Poisson distribution with parameter $\lambda = \mu * \xi$, where μ is the mean expression over all individuals and unitigs, and ξ is a scaling factor. Values of $\xi \in [0.01, 0.1]$ were simulated [Figure 5.3.C] to assess the robustness of the method w.r.t. technical noise. Genetic noise was simulated by varying the Single Nucleotide Polymorphism (SNP) rate r_{SNP}. Values of $r_{\text{SNP}} \in [0.05, 0.1, 0.5]$ were simulated [Figure 5.3.B] to assess the robustness of the method w.r.t. genetic noise. For each set of parameters, 1000 simulated experiments were run and the proportion of experiments where a signal was discovered was reported. Associations were performed using a Wilcoxon rank-sum test, and Bonferroni-corrected for 19,116 protein-coding genes (Piovesan et al., 2019), yielding a significance threshold of $p < 2.5\times10^{-6}$. The simulated reads were quantified using kallisto (Bray et al., 2016), using an index constructed from T_{ref}, in order to attempt to discover associations between those abundances and the target phenotype.

Code availability

The method was implemented in C++ using the Bifrost library (Holley and Melsted, 2020) for the construction and maintenance of de Bruijn Graphs. The source code is available under GPLv3 and can be downloaded from https://github.com/pachterlab/AAQuant. The simulation framework is available at https://github.com/pachterlab/HPM_2022.

5.3 Results

As evident in Figure 5.3.A, we can reliably recover associations between the expression of a novel isoform and a qualitative phenotype, even with low expected coverage of the identifying locus. These associations are not discovered when using transcript abundances from kallisto, using version 104 of the Ensembl annotation of GRCh38. Furthermore, Figure 5.3.B shows that even for high levels of genetic noise, we can still detect these associations, due to the pruning of low quality vertices from the graph. Lastly, per Figure 5.3.C, we can reliably recover the associations with high levels of instrumental noise, and the robustness to noise is relative to the expected number of reads overlapping a distinguishing region. Combining these three measures of robustness, we are able to detect associations 97.5% of all associations between the expression of a novel isoform and a qualitative phenotype, using reasonable parameters for a real-world experiment, e.g. 25 reads per individual, which yields an expected 13.9 reads overlapping a distinguishing region, SNP rate of 1%, and noise with a magnitude of 5% of the mean unitig expression.

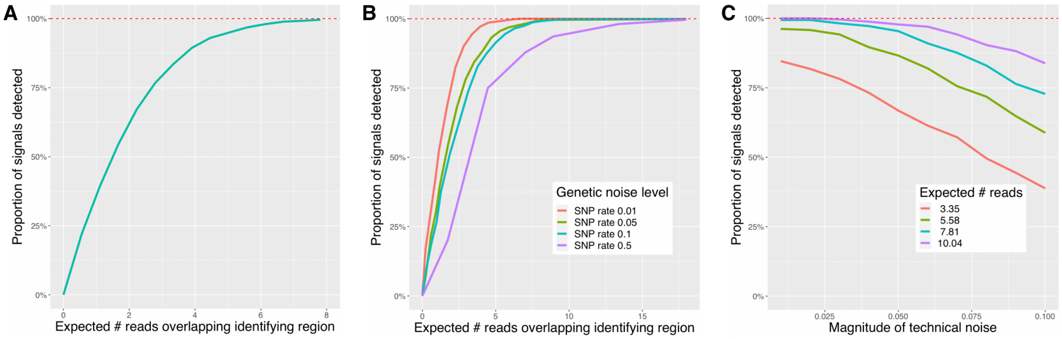


Figure 5.3: The proportion of 1,000 simulated annotation-agnostic association experiments in which a ground truth association between expression of a novel isoform, containing a duplication, and a phenotype, was detected. **A.** To detect a signal in an experiment with no genetic or technical noise, it was sufficient for an expected 5 reads to overlap an identifying region of the gene. **B.** Even with a SNP rate of 0.5 (See Methods), we detected over 95% of all signals, with an expected 11 reads overlapping an identifying region of the gene. Note however, that even with quality filters in place (See Methods]) the noise adds extraneous vertices to the graph, the expression of which must also be associated with the phenotype, resulting in a large number of computations. **C.** With an expected 10 reads overlapping an identifying region of the gene, we can reliably detect signals from simulations with technical noise (See Methods) of magnitude up to 7% of the mean expression in the cohort.

5.4 Discussion

AbundanceDBG enables annotation-agnostic discovery of associations between relative abundances of kmers in gene transcripts on one hand, and qualitative and quantitative phenotypes on the other. It does so in a memory and computationally efficient way by processing unambiguous, overlapping k-mers together, and by leveraging minimizers for lookup in the underlying graph. The ability to detect transcriptional and post-transcriptional events, without prior knowledge of those events is useful for discovering expression associations in instances where transcriptome annotations are incomplete or nonexistent. Firstly, we have shown that we can discover association between novel isoforms and a phenotype, without prior knowledge of the isoforms. Secondly, using simulated experiments, we have shown that we can reliably detect associations not found by state-of-the-art RNA-seq quantification methods. We have furthermore demonstrated that our discovery of the associations is robust to genetic and instrumental noise. However, the method does not attribute meaning to the associated sequence. Having discovered an association between a phenotype and a sequence, it must then be aligned against the genome to identify the transcriptional or post-transcriptional events that yielded the sequence. As an illustration, if the associated sequence was produced by a duplication in the gene, it remains to be determined where in the sequence the duplication splice junction is, and the loci of the sequences on either side of the junction.

Chapter 6

RECONSTRUCTION OF A LARGE-SCALE OUTBREAK OF SARS-COV-2 INFECTION IN ICELAND INFORMS VACCINATION STRATEGIES

The spread of SARS-CoV-2 is dependent on several factors, both biological and behavioral. The effectiveness of various non-pharmaceutical interventions can largely be attributed to changes in human behavior, but quantifying this effect remains challenging. In this chapter, we reconstruct the transmission tree of the third wave of SARS-CoV-2 infections in Iceland using contact tracing and viral sequence data from 2522 cases. This enables us to compare the infectiousness of distinct groups of persons directly. We find that people diagnosed outside of quarantine are 89% more infectious than those diagnosed while in quarantine, and infectiousness decreases as a function of the time spent in quarantine. Furthermore, we find that people of working age, 16-66 years old, are 47% more infectious than those outside that age range. Lastly, the transmission tree enables us to model the effect that given population prevalence of vaccination would have had on the third wave had they been administered before that time using several different strategies. We find that vaccinating in order of ascending age or uniformly at random would have prevented more infections per vaccination than vaccinating in order of descending age.

6.1 Introduction

Over 160 million cases of SARS-CoV-2 have been diagnosed globally, resulting in over 3.3 million deaths1. As the virus spreads, nations have invested heavily in monitoring the epidemic, including tracking the spread by various means. The first case of SARS-CoV-2 infection in Iceland was confirmed on February 28, 2020 and as of May 14, 2021 a total of 6,526 people have been diagnosed in the country. The first wave in Iceland was characterized by several persons introducing the virus from various countries (Gudbjartsson, Helgason, et al., 2020). Extensive sequencing of the viral genomes showed that it consisted of several genetically distinct outbreaks that were eliminated by May 2020 through non-pharmaceutical interventions, primarily isolation of cases, contact tracing, quarantine, social restrictions, and mandatory testing at the border (Gudbjartsson, Norddahl, et al., 2020). The second and third waves arose at the end of July and in mid-September 2020. Although these waves overlapped in time they can be distinguished with the sequences of the viral genomes. The third wave was considerably larger, consisting of 2,783 confirmed cases and was characterized by a single genetic clade, traced back to a person who entered the country in August 2020.

Being able to understand the differences between distinct groups of persons in epidemic outbreaks is a key to being able to employ targeted measures to contain them. By reconstructing the chain of events in an entire outbreak, we can observe these differences directly. Furthermore, having access to a case-by-case replay of the outbreak enables us to model the effect that vaccinations would have had on the third wave, had they been administered before that time. This constitutes the largest study to date investigating a single outbreak with complete contact tracing and sequence data.

6.2 Data collection and completeness

The outbreaks in Iceland are well characterized with good availability of data. Every diagnosed case was contact traced and recent contacts placed in quarantine. At the end of quarantine, all persons were tested and allowed to leave quarantine given a negative test result. This exit test allows for diagnosis of asymptomatic cases that otherwise could have gone undiagnosed3. Every PCR positive sample was sequenced within 36-48 hours of the sample collection and the results fed back to the contact tracing team to inform their inquiries. Furthermore, every infected person was enrolled in telehealth monitoring and received multiple structured phone calls to monitor symptoms after diagnosis and provide support regarding isolation practice3. This extensive data collection can provide insight into the effectiveness of non-pharmaceutical interventions.

Contact tracing of SARS-CoV-2 infections

Everyone who tested positive for SARS-CoV-2 was contacted by a team designated by the authorities to track their infection. They were required to isolate and everyone with whom they had been in contact, within 48 hours of the onset of symptoms, was required to quarantine. If the place of quarantine was shared with an infected person, the length of quarantine was two weeks, but otherwise a weeklong quarantine was sufficient. At the end of quarantine, they were tested for SARS-CoV-2.

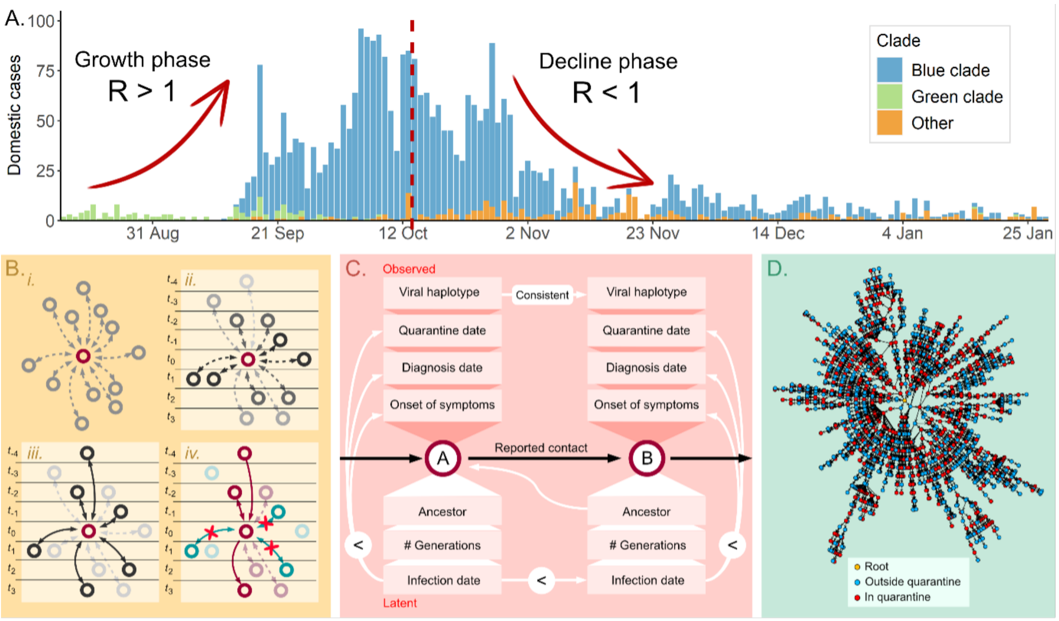


Figure 6.1: **A.** Daily cases during the third wave of SARS-CoV-2 infections in Iceland, excluding cases diagnosed at the border. On October 15, $\hat{R}_t$ t went below 1 outside of quarantine for the first time and stayed below 1 except for the time period covering the hospital outbreak. Based on this observation, we split the outbreak into a growth phase before October 15 (red dashed line) and a decline phase from then until the end of January 2021. **B.** *i.* When determining who infected a person, initially all diagnosed cases are equally likely. *ii.* Quarantine, diagnosis and dates of symptom onset make some people more likely than others, assuming specific incubation time and generation time distributions. *iii.* Contact tracing data make certain transmissions very likely but do not enable us to disregard others. *iv.* Given the viral haplotypes, we can disregard transmissions where the haplotypes are incompatible, i.e. neither is derived from the other, and in some cases determine the direction of the transmission, in cases where de novo mutations occur between generations. **C.** We use the real-world data and the tree structure to infer the latent data for each diagnosed case. The "<"-symbol represents that the date on the left needs to precede the date on the right. For each diagnosed person we infer the ancestor, i.e. the person who infected them, the date of infection, and the number of transmissions separating the ancestor and the person, κ. **D.** One instance of a reconstructed transmission tree for the third wave in Iceland.

Sequencing

The viral genomes of all positive PCR samples were sequenced at deCODE genetics. We performed reverse transcription and multiplex amplicon PCR on the basis of information provided by the Artic Network initiative (`https://artic.network/`) to generate complementary DNA and sequencing libraries. Samples were sequenced using either Illumina (n=1,939) or ONT (n=1,184) technologies. These numbers include cases diagnosed at the border. Illumina sequencing was performed using MiSeq sequencers (MiSeq v2 reagent kits) with 2 x 150 cycle paired-end reads, with up to 48 multiplexed samples per run. Samples for ONT were multiplexed using native barcodes and sequenced using either GridION or PromethION flowcells, version R.9.4.1.

Analysis of sequence data

We aligned amplicon sequences to the reference genome of the SARS-CoV-2 (GenBank number, NC_045512.2)[17]. For Illumina sequences we used the latest Burrows–Wheeler Aligner (BWA-MEM) and variants were called with sequencing utilities bcftools[18] as previously described[2]. For ONT sequences, sequence alignment and variant calling was performed using the Artic Network pipeline (https://artic.network/) with default parameters. Detailed description of sequencing methods is available in (Hjörleifsson, Pachter, and Melsted, 2022).

Data availability

All sequences used in this analysis are available in the European Nucleotide Archive (ENA) under accession number PRJEB44803 (`https://www.ebi.ac.uk/ena/browser/view/PRJEB44803`)

6.3 Reconstructing the transmission tree of a viral outbreak

In any outbreak of a viral disease there is a single progenitor, who infects a number of persons, each of whom in turn infects other persons and so forth until the disease is contained, or everyone has been infected. These transmissions from person to person form a tree of transmissions with the progenitor as its root. Since the third wave in Iceland originated with one infected person, it consisted mostly of a single subtree of the global transmission tree of the SARS-CoV-2 pandemic. Despite the extensive data collected on each diagnosed case, the true transmission tree of the third wave cannot be determined from them with certainty. In the contact tracing data, the contact resulting in a transmission may not be reported and there are reported contacts between persons where a transmission did not occur. Even in cases where an actual transmission occurred, the direction of transmission is unknown unless the virus accumulated a mutation at transmission (Figure 6.1.B).

In this study, we expand upon the size of transmission trees reconstructed in previous studies by analyzing an entire epidemic on a national scale (Kermack and Mckendrick, 1927, Aherfi et al., 2020). This enables us to quantify the efficacy of targeted interventions such as quarantine measures and compare the infectiousness of different age groups at different times. Current methods do not take into account much of the information we possess, such as quarantine times, household data and the fact that the third wave in Iceland was a single introduction outbreak. Therefore, we extended the Bayesian phylogenetic model Outbreaker2 (Campbell, Didelot, et al., 2018; Campbell, Cori, et al., 2019) to infer the most likely transmission trees using data from contact tracing, viral genome sequences, household membership, and times of onset of symptoms, quarantine, and diagnosis. This model infers the of the transmission tree with iterative MCMC sampling, generating 10,000 trees in each chain. Four MCMC chains were run, and every 50 samples were extracted with a burn-in of 5,000. The likelihood in the model is the product of the genetic, contact, generation time, incubation time and reporting likelihoods. In order to make use of our extensive data on each diagnosed case we implemented custom likelihood functions for the generation time likelihood and the incubation time likelihood, assuming the same distributions as in (Flaxman et al., 2020), but truncated them with respect to quarantine, onset of symptoms, and diagnosis dates.

The incubation time likelihood function

The incubation period of an infectious disease is the time from exposure to onset of symptoms. Because of the extensive follow-up of each person diagnosed with COVID-19 (Methods), the date of symptom onset in our data is well recorded. Furthermore, in instances where persons were in quarantine at the time of diagnosis, we can constrain a sample from the incubation time distribution to be upper bounded by the quarantine date, conditioned on that they are not in the same household as an infected person. For symptomatic persons, the incubation distribution, f^*, was chosen to be gamma distributed, like in Flaxman et al., 2020.

$$f^* \sim \Gamma(\alpha = 1.35, \beta = 0.27),$$

discretized by taking the density at each $t \in [1, \cdots, 100]$ and dividing by the sum over all t. For asymptomatic persons, a uniform distribution was chosen:

$$f(t) = \frac{1}{14}, \quad \text{for } t = 1, \cdots, 14.$$

For each person, the incubation distribution was constrained based on the available data in the following way. Let $t_{r,i} = \min(t_{d,i}, t_{s,i}$, where $t_{d,i}$ is the time of diagnosis, $t_{s,i}$ is the time of symptom onset (if not reported, $t_{s,i} = \infty$). We define an upper bound on the infection time of person i. to be $t_{u,i} = \min(t_{r,i}, t_{q,i} + 1)$, where $t_{q,i}$

is the quarantine time. If i was not quarantined or is member of a household with infected persons, then $t_{q,i} = \infty$. The likelihood $\mathcal{L}_{\text{TI}}$ of the infection time of person i, $t_{\text{inf},i}$, is then

$$\mathcal{L}_{\text{TI}}(t_{\text{inf},i}) = \begin{cases} 0, & \text{if } t_{\text{inf},i} > t_{u,i}, \\ \frac{f(t_{r,i} - t_{\text{inf},i})}{\sum_{k > t_{r,i} - t_{u,i}} f(k)}, & \text{otherwise.} \end{cases}$$

Finally, we lower bound the infection time of all persons at seven days before the first date of diagnosis in the third wave.

The generation time likelihood function

Generation time is the time from one transmission to the next in a chain of infections. Since the infection times are mostly unknown, this was approximated with the serial interval distribution, i.e. the time between onsets of symptoms of one to the next in a chain of infections. As with the incubation time distribution, we can place bounds on the infection time delay between two persons w.r.t. their respective quarantine dates, conditioned on them not being part of the same household. Like Flaxman et al., 2020, the serial interval distribution, s^*, was chosen to be gamma distributed with the parameters

$$s^* \sim \Gamma(\alpha = 2.6, \beta = 0.4),$$

discretized by taking the density at each $t \in [1, \cdots, 100]$ and dividing by the sum over all t.

For each person, the infection time delay distribution was constrained based on the available data in the following way. Conditioned on the sampled infection time of person i, $t_{\text{inf},i}$, we sample α_i, the person that infected i. We define an upper bound on the time of infection from α_i to i to be

$$t_{u,\alpha_i} = \begin{cases} \infty, & \text{if } \alpha_i \text{ was not quarantined at diagnosis,} \\ t_{r,i}, & \text{if } \alpha_i \text{ and } i \text{ share a household,} \\ \min\{t_{d,\alpha_i}, t_{q,\alpha_i}, t_{r,i}\}, & \text{otherwise.} \end{cases}$$

The likelihood $\mathcal{L}_{\text{TD}}$ of the delay $t = t_{\text{inf},i} - t_{\text{inf},\alpha_i}$ is as follows:

$$\mathcal{L}_{\text{TD}} = \begin{cases} 0, & \text{if } t_{\text{inf},i} > t_{u,\alpha_i}, \\ \frac{s(t)}{\sum_{k=1}^{t_{u,\alpha_i}} s(k)}, & \text{otherwise.} \end{cases}$$

Thus, we allow for the possibility that a quarantined person to infect other members of their household, but no one outside of their household during quarantine. Figure 6.2 shows the temporal distributions averaged over all transmission trees.

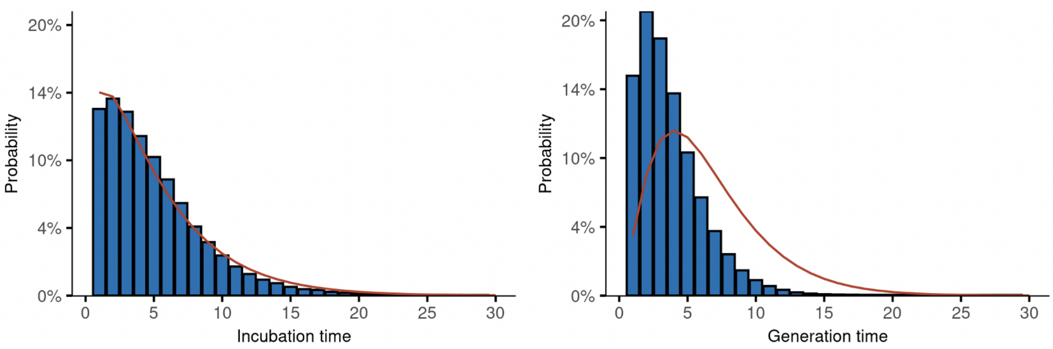


Figure 6.2: The incubation time distribution and generation time distribution averaged over all transmission trees in the final model. The red line shows the assumed distributions that were fed into the model. The incubation time distribution refers to only those who were symptomatic.

The genetic likelihood function

To estimate the likelihood of the viral haplotype of person i conditioned on α_i having infected them, we compare their set of mutations, M_i and M_{α_i}. Let $d_{m,i}$ be the sequencing depth over mutation m in person i, measured as the number of reads overlapping a 1 base pair window around the start (s_m) and end site (e_m) of the mutation ($s_m = e_m$ if m is a SNP or insertion). Further define $D_{i,\alpha_i} = \{m \in M_{\alpha_i} \setminus M_i : d_{m,i} < 5\}$ as the set of mutations in α_i with insufficient coverage in i. $D_{\alpha_i,i}$ is defined symmetrically to D_{i,α_i}. We define the accumulation of mutations $r_{\alpha_i,i}$ from person α_i to i as follows:

$$r_{\alpha_i,i} = \begin{cases} |(M_i \cup D_{i,\alpha_i} \setminus M_{\alpha_i}|, & \text{if } M_{\alpha_i} \subseteq M_i \cup D_{i,\alpha_i}, \\ -|(M_{\alpha_i} \cup D_{\alpha_i,i}) \setminus M_i|, & \text{if } M_i \subseteq M_{\alpha_i} \cup D_{\alpha_i,i}, \\ -\infty, & \text{otherwise.} \end{cases}$$

Let X be the random variable representing the accumulation of mutations between two persons separated by κ generations of infections. We model X with a Poisson random variable $X \sim \text{Poi}(\lambda = \kappa\mu)$, where μ is the mutation rate. We define the genetic likelihood of the viral genotype of person i, conditioned on α_i having infected them

$$\mathcal{L}_G(M_i | \alpha_i, M_{\alpha_i}, \kappa_i, \mu) = P(X = r_{\alpha_i,i}).$$

Note that for $r_{\alpha_i,i} < 0, \mathcal{L}_G(M_i | \alpha_i, M_{\alpha_i}, \kappa_i, \mu) = 0$, which disallows the occurrence of back mutations and incompatible haplotypes and ignores sequencing errors.

Other likelihoods

The default Outbreaker2 contact and reporting likelihoods were used as described in Campbell, Cori, et al., 2019.

Mutation rate and haplotype imputation

Furthermore, we implemented a custom genetic likelihood function in terms of variations from the blue clade. We estimated the accumulation of mutations in SARS-CoV-2 per transmission using pairs in the contact tracing network. We found 572 pairs of infected persons linked by contact tracing with a sequence coverage greater than 99%, that had their sampling dates separated by 2 days or more, where the receiving person in the pair shared haplotype or derived haplotype with the spreading person. The average number of mutations in the receiving persons that are not present in the spreading persons provides an estimate of the mutation rate per transmission. We found 158 such mutations among the 572 pairs, translating into a mutation rate of 0.28 per transmission. To correct for errors in the transmission chain inference and for false positive mutations, we reversed the role of the spreading and receiving person (459 pairs) and found 19 mutations (0.04 per transmission). This resulted in a corrected mutation rate of 0.23 per transmission (95%-CI: 0.19-0.28). We used the mutation rate as a fixed parameter in the model. This enabled us to control the likelihood of back mutations in the model and to impute missing or incomplete viral haplotypes. In the cases where an infected person had started producing antibodies at the time of sampling, the sequenced haplotype may have been incomplete or even missing altogether. If the sequence coverage was less than 95%, we chose a random person with whom they had reported contact and had them inherit their viral haplotype in the model.

Other hyperparameters

The default Outbreaker2 prior distributions were used for the remaining parameters as described by (Campbell, Cori, et al., 2019). The proportion of cases reported π, the proportion of contacts reported ϵ, and the probability of non-infectious contact between cases λ were all fitted by Outbreaker2. Posterior distributions of these hyperparameters are available in table 6.1. The mutation rate of the virus μ was statically estimated and fixed at 0.23 (See above). The probability of contact between transmission pairs η, and the probability of false-positive reporting of contact ζ were fixed to $\eta = 1$, $\zeta = 0$, as is the default in Outbreaker2.

Parameter	Mean	95%-PI
ϵ	0.74	(0.72 – 0.77)
λ	5.5×10^{-4}	$(5.3 \times 10^{-4} - 5.9 \times 10^{-4})$
π	0.87	(0.83 – 0.91)

Table 6.1: Posterior summary of the proportion of cases sampled π, the proportion of cases reported ϵ, and the probability of non-infectious contact between cases λ. PI refers to the posterior interval.

6.4 Estimating stratified reproduction number using transmission trees

The effective reproduction number R of a disease outbreak denotes how many persons each diagnosed person infects on average. It is useful for discerning whether an epidemic is in growth or successfully being contained. In order to contain an outbreak, R must stay below one. The R at a given time is denoted by R_t, the time-varying reproduction number. A variety of methods have been proposed to estimate R_t (Kermack and Mckendrick, 1927; Flaxman et al., 2020; Giordano, Blanchini, et al., 2020; Hethcote, 2000; Cori et al., 2013), all of which attribute the number of cases at time t to cases diagnosed in the preceding days weighted with the assumed generation time distribution. The idea of reconstructing the latent transmission tree of an outbreak has been explored in previous studies (Campbell, Cori, et al., 2019; Wallinga, 2004; Aherfi et al., 2020; James et al., 2021), most recently with the Outbreaker2 model which infers the transmission tree of an outbreak using contact data, sequence data and times of symptom onset. In comparison to the classical methods, in a transmission tree model, R_t is calculated by averaging the out-degree, i.e. the number of persons they infected, of everyone in the tree at time t. Since the data are available on an individual level, we can estimate the reproduction number for distinct groups of people, allowing us to compare their relative infectiousness in an outbreak.

6.5 Simulating the effects of vaccination on transmission trees

There are two distinct goals of the vaccination effort. Firstly, to protect those at risk, such as the elderly, those with underlying diseases, and front-line workers. Secondly, to obtain herd immunity to protect the community from future outbreaks. Once the first goal has been attained, the order in which vaccines should be distributed to the rest of the population needs to be decided. Some efforts have been made to simulate the effect vaccination has on the spread of the disease (Giordano, Colaneri, et al., 2021; Grauer, Löwen, and Liebchen, 2020; Huang et al., 2021). These models construct a theoretical wave of infections assuming a compartmental model (e.g. SIR models and variations thereof), and rely on multiple epidemiological constants. However, by using the transmission trees, we can use real-world data to reconstruct what would have happened if certain people in the tree had been immune at the time. By simulating immunity given a specific fraction of the adult population vaccinated, we can estimate what the size of the third wave in Iceland would have been, conditioned on all non-pharmaceutical interventions having been the same. Using this method, we can simulate the effect different vaccine distribution strategies would have had on the third wave, had they been employed before that time.

Simulating vaccination strategies

With our collection of likely infection trees, we simulated different vaccination strategies by selecting people 16 years and older to be immune in each respective infection tree and removing them and all their downstream transmissions from that tree. By counting the remaining persons in the tree, we obtained a measure of the size of the third wave in Iceland given that a particular fraction of the population would have been vaccinated at the time, and all non-pharmaceutical interventions being identical. This replay of the outbreak assumes that all transmissions remain the same, except some persons have been immunized and therefore break the chain of transmission, reducing the size of the outbreak. We considered three vaccination strategies: vaccinating in order of descending age, in order of ascending age, and uniformly at random.

The adult population was segmented into ten-year age brackets with each person in a given bracket assumed to be vaccinated with equal probability. We performed 1,000 simulations, sampling immune persons and calculating the average outbreak size over all transmission trees. The outbreak size point estimates were obtained by averaging the mean outbreak size over all simulations and 95% confidence intervals by taking the 2.5% and 97.5% quantiles. A person in a given simulation was selected to be vaccinated based on how large a proportion of their age group was vaccinated in the simulation, and a given vaccine efficacy, assumed to be 60% for the former dose and 90% for the latter, which is in line with reported efficacy of the mRNA vaccines (Dagan et al., 2021; Lipsitch and Kahn, 2021; Hall et al., 2021). The number of people in each age group was obtained from census data from the Registers Iceland and is accurate as of January 1st, 2021. We performed 1000 simulations. We also simulated the expected number of deaths, critical cases, and severe cases using the log-linear fits from (Herrera-Esposito and Campos, 2022) and (Levine-Tiefenbrun et al., 2021) (Figure 6.6, Table 6.5). These simulations were not sensitive to the initial cases in the tree.

We simulated the effect the actual distribution of vaccines in at-risk groups and would have had on the third wave. We obtained vaccination numbers for each age bracket from the Icelandic Directorate of Health. These consisted of the number of persons vaccinated per day from the first day of vaccination on December 28, 2020, until at-risk groups and front-line workers had been vaccinated, at which point 29% of the adult population had been vaccinated. For each day we modeled the third wave with the de facto number of accumulated vaccinations per age bracket.

In order to assess the sensitivity of the simulations to the initial cases in the tree, we repeated the simulations conditioned on the first 50 persons being unvaccinated, and furthermore, we ran the simulations on subtrees of size greater than 100 whose direct ancestor was one of the first 50 persons to be infected.

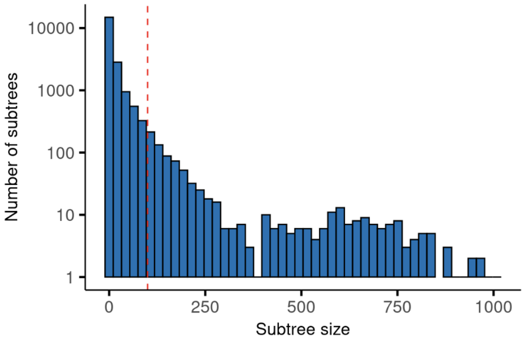


Figure 6.3: Log of the size distribution of the subtrees. The ones used in the simulations in Figure 6.7 are of size 100 or more (red, dotted line).

Statistical analysis

In order to estimate the R of a particular group of persons, we averaged the out-degree of everyone in the group over all transmission trees. The confidence intervals were obtained by iteratively calculating $\hat{R}$ with bootstrapping of the persons in the data set. To estimate the effect size of the difference in infectiousness between two distinct groups of persons, we calculated the ratio of $\hat{R}$ between the groups. Significance was tested by taking the difference of the logs of the bootstrapped $\hat{R}$ values for the two groups and performing a z-test, using the bootstrapped values to estimate the standard deviation. In addition to bootstrapping, we performed jackknife and permutation tests, with identical results.

We estimated the stratified time-varying reproduction number R_t to be the mean out-degree per group of everyone diagnosed in a four-day wliding window $(t-4, t]$, such that each person contributed to $\hat{R}_t$ on four days. We obtained the 95% confidence interval with bootstrapping.

Data and source code availability Source code for model construction and analysis is available at
https://github.com/DecodeGenetics/COVID19_reconstruction_iceland

6.6 Results

In the third wave of SARS-CoV-2 infections in Iceland, 89% of diagnosed cases had a single dominant haplotype, traced back to a person who entered the country in August 2020 (Figure 6.1.A). The third wave accumulated 2,783 cases of the same or derived haplotype (colloquially referred to as the blue clade, see Table 6.2) over a period of five months before being contained. Although cases of other clades were diagnosed during this period, we refer to the blue clade outbreak in Iceland as the third wave for the sake of brevity. Vaccinations against SARS-CoV-2 were initiated in Iceland on December 28th, 2020, and when the last blue clade case was diagnosed on January 28th, 2021, only 3.6% of the adult population (16 years and older) had received at least one vaccine dose. The success of the containment was largely due to non-pharmaceutical interventions, mass testing, and effective contact tracing measures.

Locus	241	445	3037	6023	6286	8017	13064	14408	18483	19999	20229	21255	22227	23403	25563	26801	28932	29645
Ref	C	T	C	T	C	A	C	C	T	G	C	G	C	A	G	C	C	G
Alt	T	C	T	C	T	G	T	T	C	T	T	C	T	G	T	G	T	T

Table 6.2: The mutations that make up the blue clade, representing the vast majority of diagnosed cases in the third wave.

We inferred a transmission tree using data on every person in the third wave diagnosed before December 1st, 2020, a total of 2,522 people (Figure 6.1.D). Of these, 2,431 had the blue clade and 91 had an incomplete or missing haplotype but were included because of contact tracing data. Although the third wave continued past December 1st, this dataset contains 91% of the 2,783 persons who were ever diagnosed with this clade. Contact tracing data, quarantine status, and onset of symptoms were available for everyone in the dataset. A total of 1,275 (51%) persons were diagnosed while in quarantine and 1,964 (78%) had reported contact with prior cases. 1,738 (69%) persons were symptomatic upon diagnosis and 187 (7%) never showed any symptoms. An average of 303 persons (12%) in the model had more than one transmission between them and their ancestor, indicating the presence of undiagnosed case. The Outbreaker2 model estimated the proportion of observed cases to be 87% of the total (95%-PI: 83%-91%) (Table 6.1).

Effect of contact tracing-informed quarantine on the effective reproduction number

Persons diagnosed outside of quarantine were 88.8% more infectious (95%-CI: 70.9%-109.2%, $p = 2.8 \times 10^{-32}$) than those diagnosed while in quarantine. The former had an $\hat{R}$ of 1.31 (95%-CI: 1.21-1.43) while the latter had an $\hat{R}$ of 0.70 (95%-CI: 0.66-0.73). Furthermore, the length of time from the start of quarantine to a positive PCR test had a significant effect on infectiousness. Persons diagnosed after a short quarantine, i.e. one or two days, were 66.6% more infectious (95%-CI: 49.3%-85.2%, $p = 4.0 \times 10^{-19}$) than those diagnosed after a long quarantine, i.e. three or more days, with an $\hat{R}$ of 0.89 (95%-CI: 0.83-0.96) and 0.54 (95%-CI: 0.50-0.58), respectively. This indicates that the sooner people were quarantined after exposure, the fewer opportunities they had to infect others, which shows that contact tracing is highly time critical. Additionally, those diagnosed outside quarantine were 144.4% more infectious (95%-CI: 116.8%-174.1%, p=2.5×10^{-50}) than those who were diagnosed after a long quarantine.

Effective reproduction number varies with age

We calculated the $\hat{R}$ of adults, 16 years and older, and children, 15 years and younger, demonstrating that adults were 59.5% more infectious (95%-CI: 41.6%-83.4%, $p = 1.7 \times 10^{-12}$), with an $\hat{R}$ of 1.06 (95%-CI: 0.99-1.12) compared to 0.66 (95%-CI: 0.59-0.73) for children. We also calculated the $\hat{R}$ of those of working age, 16 to 66 years old and found that they were 46.6% more infectious (95%-CI: 27.7%-65.4%, $p = 1.6 \times 10^{-8}$) than those outside that age range, with an $\hat{R}$ of 1.08 (95%-CI: 1.01-1.16) compared to 0.74 (95%-CI: 0.66-0.84) for children and the elderly. In addition to showing that adults were more infectious than children and the elderly in the third wave in Iceland, this indicates that people of working age in particular played a key role in the transmission of the virus.

Estimating the time-varying reproduction number

We calculated $\hat{R}_t$, stratified by whether or not people were in quarantine at the time of diagnosis (Figure 6.4.A). $\hat{R}_t$ outside quarantine is more variable than the relatively stable $\hat{R}_t$ in quarantine. Figure 6.4.A shows three peaks in $\hat{R}_t$ outside of quarantine, which correspond to three well characterized events: two superspreading events and one outbreak in a hospital. On October 15, $\hat{R}_t$ went below 1 outside of quarantine for the first time and stayed below 1 except for the time period covering the hospital outbreak. Based on this observation we split the outbreak into a growth and decline phase on October 15 (Figure 6.1.A).

An outbreak has (at least) two phases

Any outbreak has at least one growth phase and one decline phase. The mean out-degree of persons who get infected during a growth phase is strictly greater than one, and strictly less than one during a decline phase. For an entire outbreak, the mean out-degree is equal to $1 - {}^{1}/{}_{N}$, where N is the number of people in the transmission tree. This can be verified thus: Let N be the number of persons in a given transmission tree. Each person except for the root is infected by exactly one other person, who is also in the tree, thus has an in-degree of 1. The number of edges in the tree is then $N - 1$, and since the average R is the average out-degree of the transmission tree we have that $R = 1 - {}^{1}/{}_{N}$.

The $\hat{R}$ of different groups during the decline and growth phase of the third wave are shown in Table 1. All comparisons reported above remain significant in the growth phase and decline phase, except there is no significant difference between the infectiousness of those of working age and those outside working age in the decline phase (23.8%, 95%-CI: -3.3%-56.3%, $p = 0.08$).

Group	**Overall**		**Growth phase**		**Decline phase**	
	N	$\hat{R}$	N	$\hat{R}$	N	$\hat{R}$
Everyone	2522 (100%)	1.00	1442 (100%)	1.17 (1.09-1.27)	1080 (100%)	0.77 (0.70-0.85)
Outside quarantine	1247 (49%)	1.31 (1.21-1.43)	776 (54%)	1.45 (1.32-1.62)	471 (44%)	1.08 (0.93-1.25)
In quarantine	1275 (51%)	0.69 (0.66-0.73)	666 (46%)	0.84 (0.78-0.91)	609 (56%)	0.53 (0.49-0.57)
Short quarantine	564 (22%)	0.89 (0.83-0.96)	340 (24%)	1.02 (0.93-1.13)	224 (21%)	0.70 (0.62-0.78)
Long quarantine	711 (28%)	0.54(0.50-0.58)	326 (23%)	0.66 (0.58-0.74)	385 (36%)	0.43 (0.39-0.48)
Adults (16+ y.o.)	2164 (86%)	1.06 (0.98-1.12)	1269 (88%)	1.22 (1.13-1.32)	895 (83%)	0.82 (0.74-0.91)
Children (0-15 y.o.)	358 (14%)	0.66 (0.59-0.73)	173 (12%)	0.80 (0.69-0.93)	185 (17%)	0.53 (0.45-0.62)
Working age (16-66 y.o.)	1921 (76%)	1.08 (1.01-1.16)	1171 (81%)	1.25 (1.15-1.37)	750 (69%)	0.82 (0.73-0.92)
Outside working age	601 (24%)	0.74 (0.66-0.84)	271 (19%)	0.83 (0.74-0.92)	330 (31%)	0.66 (0.54-0.81)

Table 6.3: The number of people in different groups diagnosed in the growth phase and the decline phase of the third wave of SARS-CoV-2 infections in Iceland, and their estimated effective reproduction number $\hat{R}$.

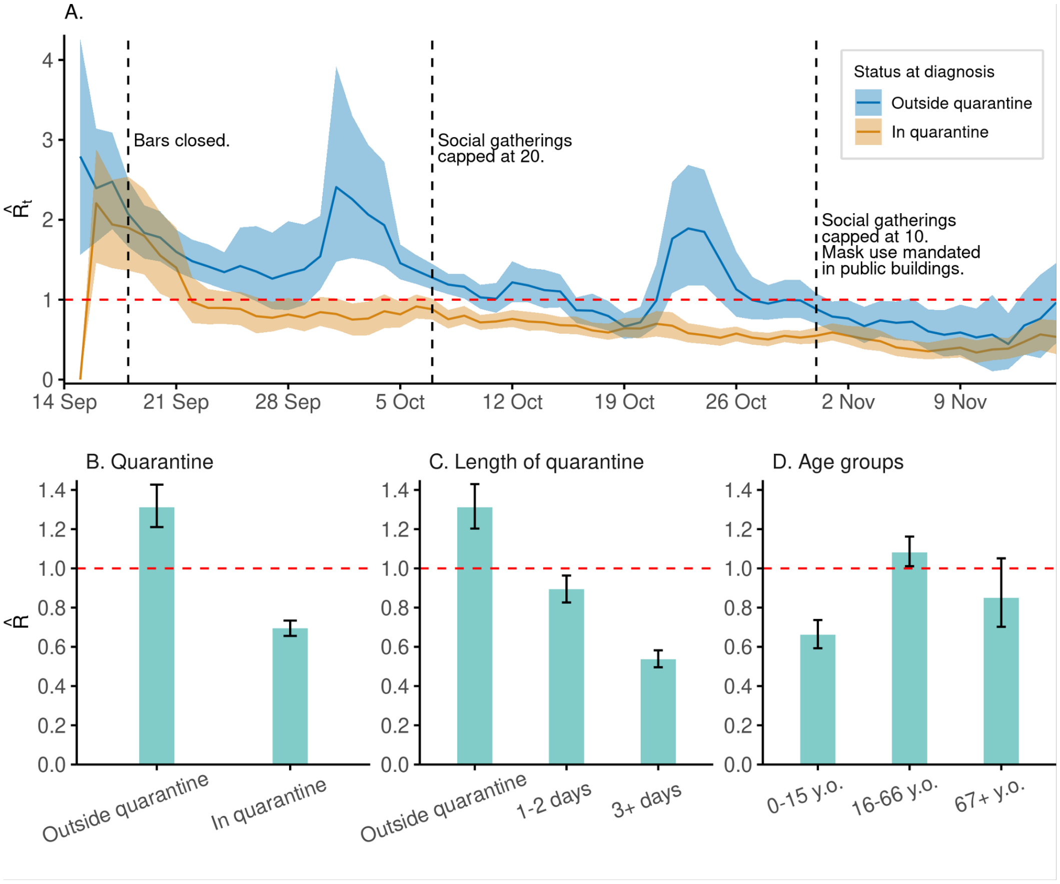


Figure 6.4: A. $\hat{R}_t$ for those diagnosed while in quarantine and those diagnosed outside quarantine, respectively. The shaded area represents the 95% confidence interval for the mean and dashed lines show dates of social restrictions imposed. B. Effective reproduction number of those diagnosed outside quarantine compared to those diagnosed in quarantine. Error bars reflect 95% CI of the mean. C. Effective reproduction number of those diagnosed outside quarantine, those diagnosed after 1-2 days in quarantine and those diagnosed after 3+ days in quarantine. D. Effective reproduction number stratified by age.

Simulating vaccination strategies

The effect of vaccination is not only determined by the proportion of persons vaccinated, but also who is vaccinated. As our results show, there was a significant difference in infectiousness between age groups in the third wave. To investigate this effect, we modeled three vaccination strategies on the adult population, 16 years and older: vaccinating by order of descending age, order of ascending age, and uniformly at random. We then compared these vaccination strategies in terms of the expected total number of cases, critical cases, severe cases, and deaths. For each strategy we iteratively increased the proportion of the adult population vaccinated, both starting at 0% and assuming a starting point of 29% as of April 28th, 2021. The second starting point reflects the actual vaccinations of persons at high risk and front-line workers in Iceland. We assumed that a single dose would lower the probability of being infected by 60% and that two doses would lower it by 90% (Dagan et al., 2021; Lipsitch and Kahn, 2021; Hall et al., 2021). We found that vaccinating the population uniformly at random or in order of ascending age would have yielded fewer overall cases than vaccinating in order of descending age. We found no significant difference in the number of deaths, critical cases, and severe cases between the vaccination strategies.

Figure 6.5 shows the mean size of the outbreak for the three vaccination strategies, assuming the first person in the transmission tree is unvaccinated. As a benchmark we compare the vaccination strategies by the lowest proportion of adults who would have needed to be vaccinated such that the final size of the third wave would have been 100 persons (4% of the observed outbreak) on average. These simulations are not sensitive to the initial cases in the transmission tree (Supplementary methods, supplementary figures 1-4).

All non-pharmaceutical interventions being the same as they were in the third wave, starting at 29% and vaccinating with a single dose in order of descending age would have yielded an outbreak with a mean size of 100 persons with 79% of adults vaccinated (95%-CI: 68%-89%). Vaccinating in order of ascending age would have yielded a 100-person outbreak with 64% of adults vaccinated (95%-CI: 54%-76%), and vaccinating uniformly at random with 72% vaccinated (95%-CI: 56%-85%). Table 2 shows comparisons between the different vaccination strategies.

	Proportion of adults vaccinated		
Model	**Age, descending**	**Age, ascending**	**Uniform at random**
Actual vaccinations/First dose	79.2% (67.6%-89.4%)	64.1% (53.7%-75.7%)	72.3% (56.1%-85.1%)
Actual vaccinations/Second dose	66.2% (57.1%-72.1%)	52.8% (42.4%-58.4%)	54.5% (43.7%-63.1%)
First dose	81.1% (71.8%-89.6%)	50.4% (38.5%-69.2%)	70.0% (50.0%-86.5%)
Second dose	66.8% (59.9%-72.4%)	35.0% (29.4%-40.1%)	47.0% (33.6%-57.5%)

Table 6.4: The lowest proportion of adults who would have needed to be vaccinated such that the final size of the third wave would have been 100 persons on average. The former two models use actual vaccination numbers up to the 29% mark and extrapolate from there using the three strategies. The latter two models start from zero.

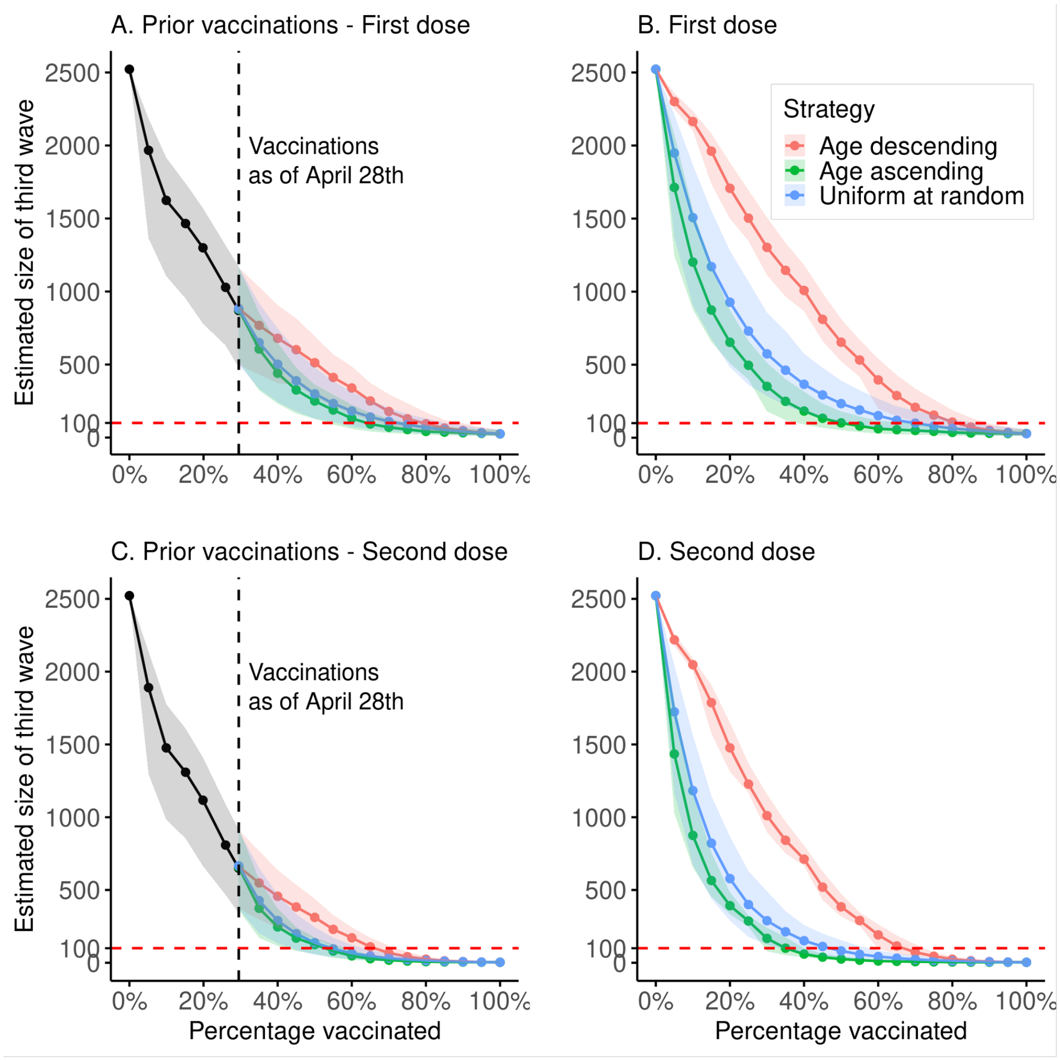


Figure 6.5: Simulations of the estimated final size of the third wave at a given population prevalence of vaccination. Solid lines show the mean size of the outbreak, shaded areas represent 2.5%-97.5% quantiles. A. Using the actual vaccination scheme for at-risk groups and front-line workers, up to 29% of the adult population, and using three separate vaccination strategies from 29% to 100%: age-descending, age-ascending and uniformly at random. Modeled vaccinations beyond the 29% mark are assumed to have an efficacy of 60%. B. Simulations of the size of the third wave, assuming 60% vaccine efficacy, under the three different vaccination strategies, starting with no vaccinations and concluding with 100% of the adult population vaccinated. C. Same simulation as in A, but all vaccinations are assumed to have an efficacy of 90% (both first and second dose administered). D. Same simulation as in C, but assuming 90% vaccine efficacy.

6.7 Discussion

Quarantine has been assumed to slow the spread of infectious diseases, but the extent to which it is effective has been difficult to quantify because doing so requires data

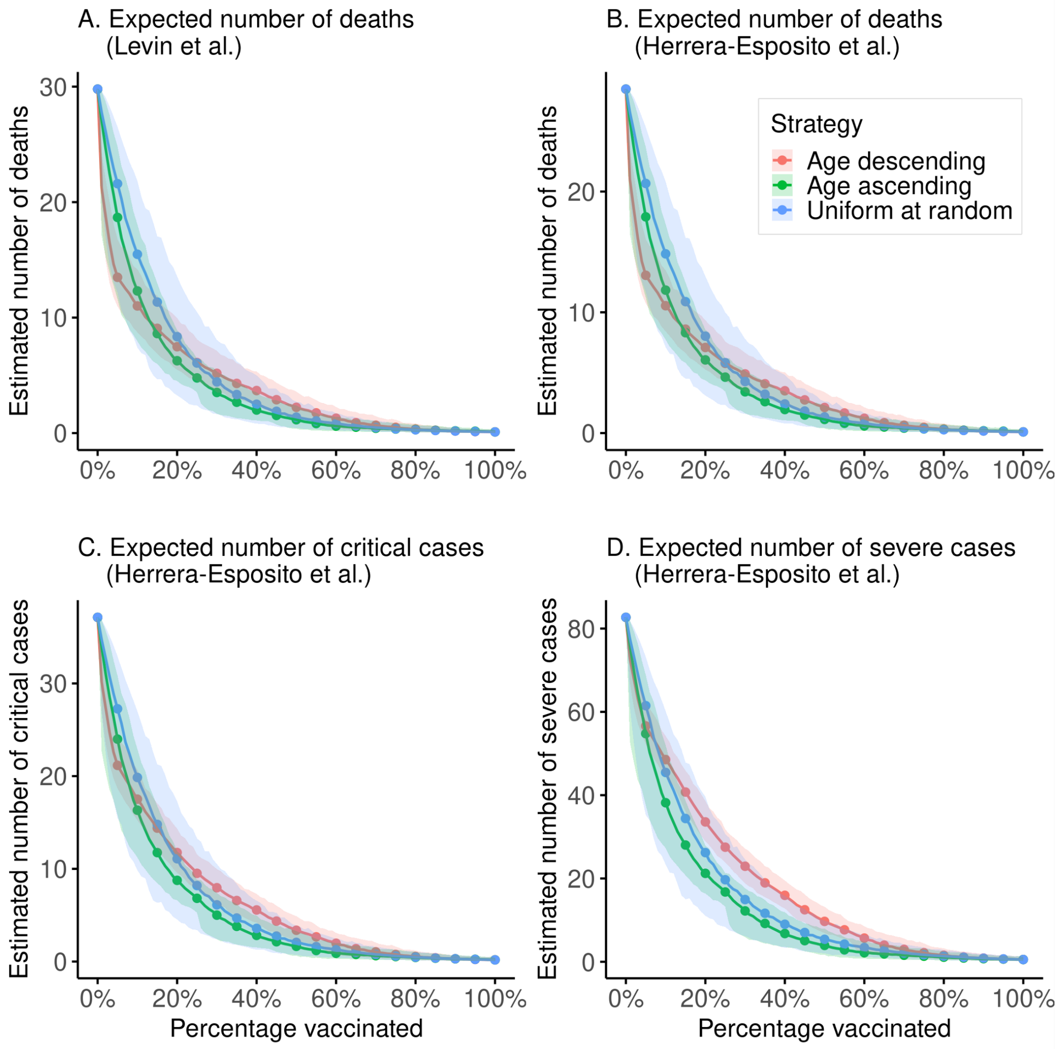


Figure 6.6: The expected number of deaths, critical cases, and severe cases as a function of the proportion of adults, 16 years and older vaccinated under the three different vaccination strategies modeled.

Proportion of adults vaccinated at crossover

Metric	**Age, ascending**	**Uniform at random**
Fatal cases (Levin et al.)	10.8% (0%-21.8%)	17.7% (0%-33.4%)
Fatal cases (Herrera-Esposito et al.)	10.9% (0%-22.1%)	18.0% (0%-33.8%)
Critical cases (Herrera-Esposito et al.)	6.9% (0%-15.3%)	12.9% (0%-26.8%)
Severe cases (Herrera-Esposito et al.)	3.4% (0%-10.7%)	7.4% (0%-17.8%)

Table 6.5: The crossover points at which vaccinating in order of ascending age and uniformly at random, respectively, yield fewer expected deaths, critical cases, and severe cases than vaccinating in order of descending age. There was no significant difference between the two vaccination paradigms.

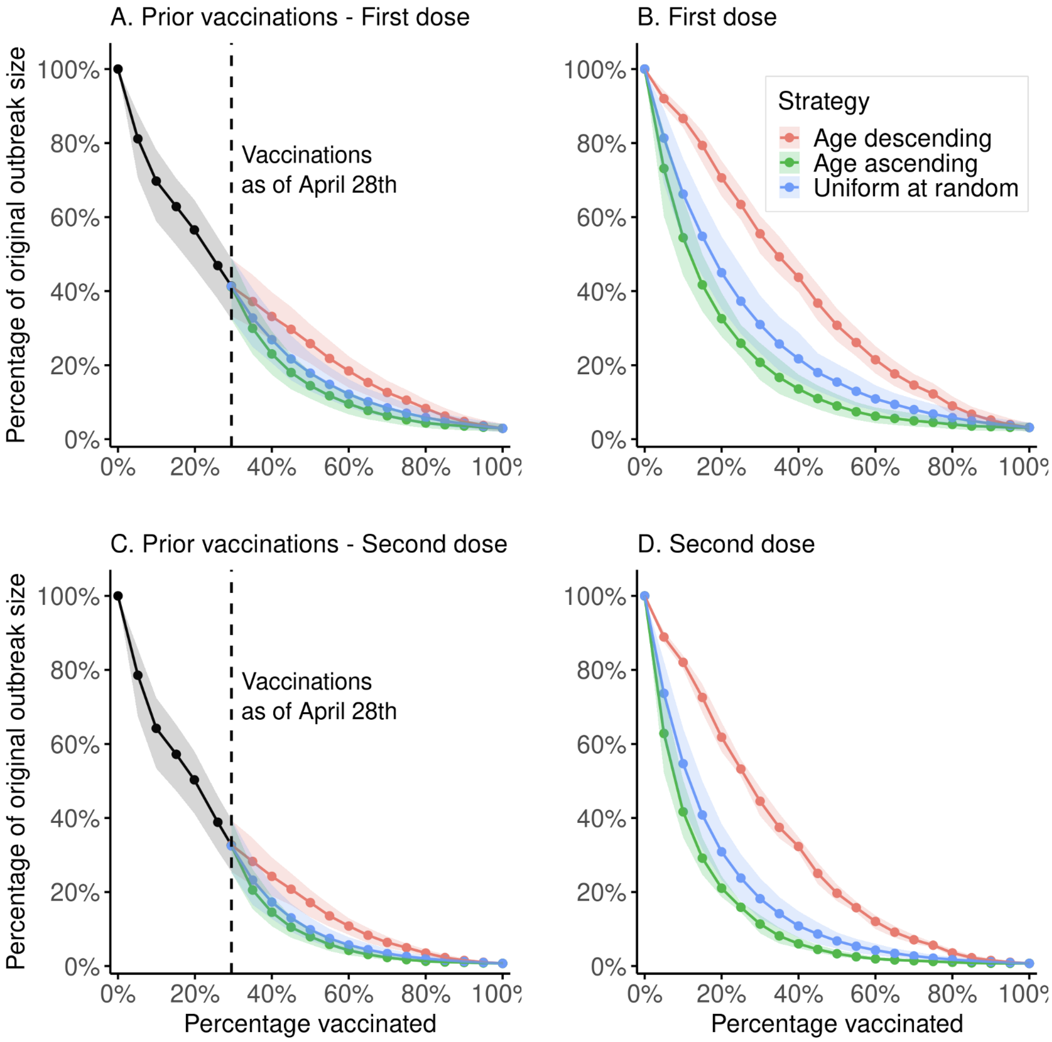


Figure 6.7: Simulations of the estimated size of the third wave for a given population prevalence of vaccination, using subtrees in order to remove the dependence on the age distribution in the initial cases. Each subtree is of size 100 and its root has a direct ancestor in the first 50 persons to be infected. Solid lines show the mean outbreak size, shaded areas represent 2.5%-97.5% quantiles. A. Using the de facto vaccination scheme for at-risk groups and front-line workers, up to 29% of the adult population, and using three separate vaccination strategies from 29% to 100%: age-descending, age-ascending and uniformly at random. Modeled vaccinations beyond the 29% mark are assumed to have an efficacy of 60%. B. Simulations of the size of the third wave, assuming 60% vaccine efficacy, under the three different vaccination strategies, starting with no vaccinations and concluding with 100% of the adult population vaccinated. C. Same simulation as in A, but all vaccinations are assumed to have an efficacy of 90% (both first and second dose administered). D. Same simulation as in B, but assuming 90% vaccine efficacy.

on the individual level. We found that mandated quarantine significantly decreased the spread of the third wave of SARS-CoV-2 infections in Iceland, with persons diagnosed outside of quarantine being 89% more infectious than those diagnosed while in quarantine. Furthermore, we observed that contact tracing is time critical by comparing the infectiousness of people diagnosed after a short quarantine to that of those diagnosed after a longer quarantine. Lastly, we found that people of working age played a key role in the generation of the third wave in Iceland, most likely resulting from more frequent contact among this age group, compared with older persons who may be retired.

We found that vaccinating persons in order of ascending age or uniformly at random would have prevented more transmissions per vaccination than vaccinating in descending order of age in the third wave in Iceland. Our estimates of the final size of the outbreak are sensitive to the assumed vaccine efficacy. However, the relative difference between the modelled vaccination strategies is independent of efficacy. Recent studies suggest that vaccinated persons who become infected have a lower viral load (Levine-Tiefenbrun et al., 2021) and may be less likely to infect others (Harris et al., 2021). This is not taken into account here.

The effect of vaccination on the spread of the disease has been studied with classical modelling approaches based on susceptible, infectious, and/or recovered models and variations thereof. These models can yield insights, but uncertainty remains as to how contacts, dependency between age of contacts, and variability due to superspreading events should be modelled. By reconstructing the third wave from real-world data, we circumvented these limitations by removing the behavioural modelling assumptions and simulating vaccinations directly on the transmission tree.

Our results show no significant difference in the expected number of deaths, critical cases, or severe cases between the modelled strategies. This implies that it is possible to minimize the number of cases without increasing the mortality or hospitalization rates. One possible explanation is that, although older persons are more likely to develop severe disease, the vaccination of younger persons prevents transmission to older people. Since the data were collected, SARS-CoV-2 variants have emerged that have been shown to be more infectious than previous ones, particularly the Delta variant (B.1.617.2). Like vaccine efficacy, this increased infectiousness would only affect the final size of the outbreak, but the relative difference between the strategies is independent of baseline infectiousness. Recent studies have considered vaccine efficacy (VE) against SARS-CoV-2 infection and whether VE against the Delta variant is reduced. A study from Qatar (Tang et al., 2021) based on convenience samples showed lower VE against all infections (symptomatic and asymptomatic) of the Delta variant for BNT162b2 (Pfizer; 53.5% VE), whereas no reduction was observed or mRNA-1273 (Moderna; 84.8% VE). Additionally, a recent survey-based study from the UK (Pouwels et al., 2021) revealed decreased VE against all infections of the Delta variant for ChAdOx1 nCoV-19 (Oxford- AstraZeneca; 67% VE), but no significant reduction for BNT162b2 (Pfizer; 80% VE). Based on these results, we believe that our findings would also apply to the Delta variant. The effectiveness of nonpharmaceutical interventions can largely be attributed to changes in human behaviour. Quantifying this effect remains challenging. By leveraging the extensive data collected for diagnosed persons in the third wave of SARS-CoV-2 infections in Iceland, we created a model that allowed us to observe the differences in infectiousness of distinct groups of people.

Although the data collected are extensive, some cases went undiagnosed. Serologic measurements after the first wave of SARS- CoV-2 infections in Iceland (Gudbjartsson, Helgason, et al., 2020) estimated that diagnosed cases were 56% of the total and another 14% were quarantined but undiagnosed. In addition, 95% of people quarantined in the third wave ($n = 21,225$) were PCR tested upon leaving quarantine. Due to this and the higher availability of PCR tests, we expect that at least 70% of cases in the third wave were diagnosed and therefore included in the transmission tree. Outbreaker2 estimated that 87% of cases were diagnosed, but this estimate does not include undiagnosed persons who did not infect others.

The vaccination of a population serves two distinct purposes: first, to prevent death and severe illness in groups at high risk, and second, to curb the spread of the virus in the population. We simulated the effect of three vaccination strategies using four different metrics: the number of infections, severe cases, critical cases, and deaths. Our results demonstrate a negligible difference between the vaccination strategies for the latter three metrics (Figure 6.6, Table 6.5), but a significant difference in the number of infections (Figure 6.5). Although our results for the third wave indicate that vaccinating in order of ascending age would have curtailed the outbreak sooner, this may reflect the age composition of this particular outbreak. Vaccinating the remaining adult population uniformly at random, once high-risk groups have been fully vaccinated, is a more robust strategy, because it removes the dependency between who is vaccinated and their age. When interpreting these results, it is important to keep in mind that they only provide a lower bound on the so-called herd immunity threshold.

BIBLIOGRAPHY

Aherfi, Sarah et al. (Nov. 2020). "Clusters of COVID-19 associated with Purim celebration in the Jewish community in Marseille, France, March 2020". In: *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases* 100, pp. 88–94. ISSN: 18783511. DOI: 10.1016/j.ijid.2020.08.049.

Bentley, David L (June 2005). "Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors". In: *Current Opinion in Cell Biology* 17 (3), pp. 251–256. ISSN: 09550674. DOI: 10.1016/j.ceb.2005.04.006.

Beyer, A L and Y N Osheim (June 1988). "Splice site selection, rate of splicing, and alternative splicing on nascent transcripts." In: *Genes & Development* 2 (6), pp. 754–765. ISSN: 0890-9369. DOI: 10.1101/gad.2.6.754.

Booeshaghi, A Sina and Lior Pachter (2021). "Benchmarking of lightweight-mapping based single-cell RNA-seq pre-processing". In: *bioRxiv*. DOI: 10.1101/2021.01.25.428188. URL: https://doi.org/10.1101/2021.01.25.428188.

Bray, Nicolas L. et al. (May 2016). "Near-optimal probabilistic RNA-seq quantification". In: *Nature Biotechnology* 34 (5), pp. 525–527. ISSN: 15461696. DOI: 10.1038/nbt.3519.

Bushnell, Brian (2014). "BBMap: A Fast, Accurate, Splice-Aware Aligner". In.

Campbell, Finlay, Anne Cori, et al. (Mar. 2019). "Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data". In: *PLoS Computational Biology* 15 (3). ISSN: 15537358. DOI: 10.1371/journal.pcbi.1006930.

Campbell, Finlay, Xavier Didelot, et al. (Oct. 2018). "outbreaker2: A modular platform for outbreak reconstruction". In: *BMC Bioinformatics* 19. ISSN: 14712105. DOI: 10.1186/s12859-018-2330-z.

Cervantes-Pérez, Sergio Alan et al. (Dec. 2022). "Review: Challenges and perspectives in applying single nuclei RNA-seq technology in plant biology". In: *Plant Science* 325, p. 111486. ISSN: 01689452. DOI: 10.1016/j.plantsci.2022.111486.

Conesa, Ana et al. (Jan. 2016). *A survey of best practices for RNA-seq data analysis*. DOI: 10.1186/s13059-016-0881-8.

Cori, Anne et al. (Nov. 2013). "A new framework and software to estimate time-varying reproduction numbers during epidemics". In: *American Journal of Epidemiology* 178 (9), pp. 1505–1512. ISSN: 00029262. DOI: 10.1093/aje/kwt133.

Cunningham, Fiona et al. (Jan. 2022). "Ensembl 2022". In: *Nucleic Acids Research* 50 (D1), pp. D988–D995. ISSN: 13624962. DOI: 10.1093/nar/gkab1049.

Dagan, Noa et al. (Apr. 2021). “BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting”. In: *New England Journal of Medicine* 384 (15), pp. 1412–1423. ISSN: 0028-4793. DOI: 10.1056/NEJMoa2101765.

Al-Dalahmah, Osama et al. (Feb. 2020). “Single-nucleus RNA-seq identifies Huntington disease astrocyte states”. In: *Acta Neuropathologica Communications* 8 (1). ISSN: 20515960. DOI: 10.1186/s40478-020-0880-6.

Flaxman, Seth et al. (Aug. 2020). “Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe”. In: *Nature* 584 (7820), pp. 257–261. ISSN: 14764687. DOI: 10.1038/s41586-020-2405-7.

Fredman, Michael L. and János Komlós (Mar. 1984). “On the Size of Separating Systems and Families of Perfect Hash Functions”. In: *SIAM Journal on Algebraic Discrete Methods* 5 (1), pp. 61–68. ISSN: 0196-5212. DOI: 10.1137/0605009.

Gao, Fan and Lior Pachter (n.d.). “Efficient pre-processing of Single-cell ATAC-seq data”. In: *bioRxiv* (). DOI: 10.1101/2021.12.08.471788. URL: https://doi.org/10.1101/2021.12.08.471788.

Giordano, Giulia, Franco Blanchini, et al. (June 2020). “Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy”. In: *Nature Medicine* 26 (6), pp. 855–860. ISSN: 1546170X. DOI: 10.1038/s41591-020-0883-7.

Giordano, Giulia, Marta Colaneri, et al. (June 2021). “Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in Italy”. In: *Nature Medicine* 27 (6), pp. 993–998. ISSN: 1546170X. DOI: 10.1038/s41591-021-01334-5.

Gorin, Gennady, Meichen Fang, et al. (Sept. 2022). “RNA velocity unraveled”. In: *PLoS Computational Biology* 18 (9). ISSN: 15537358. DOI: 10.1371/journal.pcbi.1010492.

Gorin, Gennady and Lior Pachter (Mar. 2022a). “Modeling bursty transcription and splicing with the chemical master equation”. In: *Biophysical Journal* 121 (6), pp. 1056–1069. ISSN: 00063495. DOI: 10.1016/j.bpj.2022.02.004.

– (2022b). “Monod : mechanistic analysis of single-cell RNA sequencing count data”. In: DOI: 10.1101/2022.06.11.495771. URL: https://doi.org/10.1101/2022.06.11.495771.

Gorin, Gennady, Shawn Yoshida, and Lior Pachter (2022). “Transient and delay chemical master equations”. In: DOI: 10.1101/2022.10.17.512599. URL: https://doi.org/10.1101/2022.10.17.512599.

Grauer, Jens, Hartmut Löwen, and Benno Liebchen (Dec. 2020). “Strategic spatiotemporal vaccine distribution increases the survival rate in an infectious disease like Covid-19”. In: *Scientific Reports* 10 (1). ISSN: 20452322. DOI: 10.1038/s41598-020-78447-3.

Gudbjartsson, Daniel F., Agnar Helgason, et al. (June 2020). “Spread of SARS-CoV-2 in the Icelandic Population”. In: *New England Journal of Medicine* 382 (24), pp. 2302–2315. ISSN: 0028-4793. DOI: `10.1056/NEJMoa2006100`.

Gudbjartsson, Daniel F., Gudmundur L. Norddahl, et al. (Oct. 2020). “Humoral Immune Response to SARS-CoV-2 in Iceland”. In: *New England Journal of Medicine* 383 (18), pp. 1724–1734. ISSN: 0028-4793. DOI: `10.1056/NEJMoa2026116`.

Habib, Naomi et al. (Oct. 2017). “Massively parallel single-nucleus RNA-seq with DroNc-seq”. In: *Nature Methods* 14 (10), pp. 955–958. ISSN: 15487105. DOI: `10.1038/nmeth.4407`.

Hall, Victoria Jane et al. (May 2021). “COVID-19 vaccine coverage in health-care workers in England and effectiveness of BNT162b2 mRNA vaccine against infection (SIREN): a prospective, multicentre, cohort study”. In: *The Lancet* 397 (10286), pp. 1725–1735. ISSN: 1474547X. DOI: `10.1016/S0140-6736(21)00790-X`.

Harris, Ross J. et al. (Aug. 2021). “Effect of Vaccination on Household Transmission of SARS-CoV-2 in England”. In: *New England Journal of Medicine* 385 (8), pp. 759–760. ISSN: 0028-4793. DOI: `10.1056/NEJMc2107717`.

He, Dongze et al. (Mar. 2022). “Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data”. In: *Nature Methods* 19 (3), pp. 316–322. ISSN: 15487105. DOI: `10.1038/s41592-022-01408-3`.

Herrera-Esposito, Daniel and Gustavo de los Campos (Dec. 2022). “Age-specific rate of severe and critical SARS-CoV-2 infections estimated with multi-country seroprevalence studies”. In: *BMC Infectious Diseases* 22 (1). ISSN: 14712334. DOI: `10.1186/s12879-022-07262-0`.

Hethcote, Herbert W (2000). *The Mathematics of Infectious Diseases* *, pp. 599–653. URL: `https://epubs.siam.org/terms-privacy`.

Hjörleifsson, Kristján Eldjárn, Lior Pachter, and Páll Melsted (2022). “Annotation-agnostic discovery of associations between novel gene isoforms and phenotypes”. In: *bioRxiv*. DOI: `10.1101/2022.12.02.518787`.

Holley, Guillaume and Páll Melsted (Sept. 2020). “Bifrost: Highly parallel construction and indexing of colored and compacted de Bruijn graphs”. In: *Genome Biology* 21 (1). ISSN: 1474760X. DOI: `10.1186/s13059-020-02135-8`.

Huang, Bo et al. (June 2021). “Integrated vaccination and physical distancing interventions to prevent future COVID-19 waves in Chinese cities”. In: *Nature Human Behaviour* 5 (6), pp. 695–705. ISSN: 23973374. DOI: `10.1038/s41562-021-01063-2`.

James, Alex et al. (Mar. 2021). “Model-free estimation of COVID-19 transmission dynamics from a complete outbreak”. In: *PLoS ONE* 16 (3 March). ISSN: 19326203. DOI: `10.1371/journal.pone.0238800`.

Kaminow, Benjamin, Dinar Yunusov, and Alexander Dobin (2021). “STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data”. In: DOI: 10.1101/2021.05.05.442755. URL: https://doi.org/10.1101/2021.05.05.442755.

Kermack, W 0 and A G Mckendrick (1927). “A Contribution to the Mathematical Theory of Epidemics”. In: URL: https://royalsocietypublishing.org/.

Kukurba, Kimberly R. and Stephen B. Montgomery (Nov. 2015). “RNA Sequencing and Analysis”. In: *Cold Spring Harbor Protocols* 2015 (11), pdb.top084970. ISSN: 1940-3402. DOI: 10.1101/pdb.top084970.

Kuo, Albert, Kasper D Hansen, and Stephanie C Hicks (2022). “Quantification and statistical modeling of Chromium-based single-nucleus RNA-sequencing data”. In: DOI: 10.1101/2022.05.20.492835. URL: https://doi.org/10.1101/2022.05.20.492835.

Lander, S et al. (2001). *Initial sequencing and analysis of the human genome International Human Genome Sequencing Consortium* The Sanger Centre: Beijing Genomics Institute/Human Genome Center*. URL: www.nature.com.

Lemire, Daniel et al. (Sept. 2017). “Roaring Bitmaps: Implementation of an Optimized Software Library”. In: DOI: 10.1002/spe.2560. URL: http://arxiv.org/abs/1709.07821%20http://dx.doi.org/10.1002/spe.2560.

Levine-Tiefenbrun, Matan et al. (May 2021). “Initial report of decreased SARS-CoV-2 viral load after inoculation with the BNT162b2 vaccine”. In: *Nature Medicine* 27 (5), pp. 790–792. ISSN: 1546170X. DOI: 10.1038/s41591-021-01316-7.

Liang, Qingnan et al. (Dec. 2019). “Single-nuclei RNA-seq on human retinal tissue provides improved transcriptome profiling”. In: *Nature Communications* 10 (1). ISSN: 20411723. DOI: 10.1038/s41467-019-12917-9.

Limasset, Antoine et al. (Feb. 2017). “Fast and scalable minimal perfect hashing for massive key sets”. In: URL: http://arxiv.org/abs/1702.03154.

Lipsitch, Marc and Rebecca Kahn (July 2021). “Interpreting vaccine efficacy trial results for infection and transmission”. In: *Vaccine* 39 (30), pp. 4082–4088. ISSN: 18732518. DOI: 10.1016/j.vaccine.2021.06.011.

Manno, Gioele La et al. (Aug. 2018). “RNA velocity of single cells”. In: *Nature* 560 (7719), pp. 494–498. ISSN: 14764687. DOI: 10.1038/s41586-018-0414-6.

Mehlhorn, Kurt (Nov. 1982). “On the program size of perfect and universal hash functions”. In: IEEE, pp. 170–175. DOI: 10.1109/SFCS.1982.80.

Melsted, Páll et al. (July 2021). “Modular, efficient and constant-memory single-cell RNA-seq preprocessing”. In: *Nature Biotechnology* 39 (7), pp. 813–818. ISSN: 15461696. DOI: 10.1038/s41587-021-00870-2.

Metropolis, Nicholas et al. (June 1953). "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21 (6), pp. 1087–1092. ISSN: 0021-9606. DOI: 10.1063/1.1699114.

Moore, Gordon (1965). In: *Electronics* 38 (8).

Mortazavi, Ali et al. (July 2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5 (7), pp. 621–628. ISSN: 15487091. DOI: 10.1038/nmeth.1226.

Neugebauer, Karla M. (Oct. 2002). "On the importance of being co-transcriptional". In: *Journal of Cell Science* 115 (20), pp. 3865–3871. ISSN: 1477-9137. DOI: 10.1242/jcs.00073.

Nicolae, Marius et al. (Apr. 2011). "Estimation of alternative splicing isoform frequencies from RNA-Seq data". In: *Algorithms for Molecular Biology* 6 (1). ISSN: 17487188. DOI: 10.1186/1748-7188-6-9.

Osheim, Y, O L Miller, and A L Beyer (Nov. 1985). "RNP particles at splice junction sequences on Drosophila chorion transcripts". In: *Cell* 43 (1), pp. 143–151. ISSN: 00928674. DOI: 10.1016/0092-8674(85)90019-4.

Pachter, Lior (Apr. 2011). "Models for transcript quantification from RNA-Seq". In: URL: http://arxiv.org/abs/1104.3889.

Pai, Athma A. et al. (Aug. 2018). "Numerous recursive sites contribute to accuracy of splicing in long introns in flies". In: *PLoS Genetics* 14 (8). ISSN: 15537404. DOI: 10.1371/journal.pgen.1007588.

Pandya-Jones, Amy and Douglas L. Black (Oct. 2009). "Co-transcriptional splicing of constitutive and alternative exons". In: *RNA* 15 (10), pp. 1896–1908. ISSN: 1355-8382. DOI: 10.1261/rna.1714509.

Piovesan, Allison et al. (June 2019). "Human protein-coding genes and gene feature statistics in 2019". In: *BMC Research Notes* 12 (1). ISSN: 17560500. DOI: 10.1186/s13104-019-4343-8.

Pouwels, Koen B. et al. (Dec. 2021). "Effect of Delta variant on viral burden and vaccine effectiveness against new SARS-CoV-2 infections in the UK". In: *Nature Medicine* 27 (12), pp. 2127–2135. ISSN: 1546170X. DOI: 10.1038/s41591-021-01548-7.

Roberts, Michael et al. (Dec. 2004). "Reducing storage requirements for biological sequence comparison". In: *Bioinformatics* 20 (18), pp. 3363–3369. ISSN: 13674803. DOI: 10.1093/bioinformatics/bth408.

Saha, Ashis and Alexis Battle (Nov. 2018). "False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors". In: *F1000Research* 7, p. 1860. DOI: 10.12688/f1000research.17145.1.

Schaeffer, L et al. (July 2017). “Pseudoalignment for metagenomic read assignment”. In: *Bioinformatics* 33 (14), pp. 2082–2088. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx106.

Schleimer, Saul, Daniel S. Wilkerson, and Alex Aiken (2004). “Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data : 2003, San Diego, California, June 09-12, 2003.” In: p. 702.

Soneson, Charlotte et al. (Jan. 2021). “Preprocessing choices affect RNA velocity results for droplet scRNA-seq data”. In: *PLoS Computational Biology* 17 (1). ISSN: 15537358. DOI: 10.1371/journal.pcbi.1008585.

Srivastava, Avi et al. (Sept. 2020). “Alignment and mapping methodology influence transcript abundance estimation”. In: *Genome Biology* 21 (1). ISSN: 1474760X. DOI: 10.1186/s13059-020-02151-8.

Stewart, Murray (Mar. 2019). “Polyadenylation and nuclear export of mRNAs”. In: *Journal of Biological Chemistry* 294 (9), pp. 2977–2987. ISSN: 00219258. DOI: 10.1074/jbc.REV118.005594.

Svensson, Valentine, Kedar Nath Natarajan, et al. (Mar. 2017). “Power analysis of single-cell rnA-sequencing experiments”. In: *Nature Methods* 14 (4), pp. 381–387. ISSN: 15487105. DOI: 10.1038/nmeth.4220.

Svensson, Valentine, Eduardo da Veiga Beltrame, and Lior Pachter (2020). “A curated database reveals trends in single-cell transcriptomics”. In: *Database* 2020. ISSN: 17580463. DOI: 10.1093/DATABASE/BAAA073.

Tang, Patrick et al. (2021). “BNT162b2 and mRNA-1273 COVID-19 vaccine effectiveness against the Delta (B.1.617.2) variant in Qatar”. In: DOI: 10.1101/2021.08.11.21261885. URL: https://doi.org/10.1101/2021.08.11.21261885.

Tardiff, Daniel F., Scott A. Lacadie, and Michael Rosbash (Dec. 2006). “A Genome-Wide Analysis Indicates that Yeast Pre-mRNA Splicing Is Predominantly Post-transcriptional”. In: *Molecular Cell* 24 (6), pp. 917–929. ISSN: 10972765. DOI: 10.1016/j.molcel.2006.12.002.

Tennyson, Christine N., Henry J. Klamut, and Ronald G. Worton (Feb. 1995). “The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced”. In: *Nature Genetics* 9 (2), pp. 184–190. ISSN: 1061-4036. DOI: 10.1038/ng0295-184.

Wallinga, J. (Sept. 2004). “Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures”. In: *American Journal of Epidemiology* 160 (6), pp. 509–516. ISSN: 0002-9262. DOI: 10.1093/aje/kwh255.

Wetterberg, I, G Baurén, and L Wieslander (July 1996). “The Intranuclear Site of Excision of Each Intron in Balbiani Ring 3 Pre-mRNA Is Influenced by the Time Remaining to Transcription Termination and Different Excision Efficiencies for the Various Introns”. In: *RNA* 2 (7), pp. 641–651.

Wilson, Daniel J. (Jan. 2019). “The harmonic mean <i>p</i> -value for combining dependent tests”. In: *Proceedings of the National Academy of Sciences* 116 (4), pp. 1195–1200. ISSN: 0027-8424. DOI: 10.1073/pnas.1814092116.

Wuarin, J and U Schibler (Nov. 1994). “Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing”. In: *Molecular and Cellular Biology* 14 (11), pp. 7219–7225. ISSN: 0270-7306. DOI: 10.1128/mcb.14.11.7219-7225.1994.

Zeng, Hongkui (July 2022). “What is a cell type and how to define it?” In: *Cell* 185 (15), pp. 2739–2755. ISSN: 10974172. DOI: 10.1016/j.cell.2022.06.031.

Zerbino, Daniel R. and Ewan Birney (May 2008). “Velvet: Algorithms for de novo short read assembly using de Bruijn graphs”. In: *Genome Research* 18 (5), pp. 821–829. ISSN: 1088-9051. DOI: 10.1101/gr.074492.107.

Zhang, David et al. (June 2020). “Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders”. In: *Science Advances* 6 (24). ISSN: 2375-2548. DOI: 10.1126/sciadv.aay8299.