Computational Methods for Resolving Heterogeneity in Biological Data

by

Hongjiu Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2019

Doctoral Committee:

Professor Margit Burmeister, Co-chair
Associate Professor Yuanfang Guan, Co-chair
Professor Jun Li
Professor Kayvan Najarian
Professor Gilbert S. Omenn
Assistant Professor Stephen Parker

Hongjiu Zhang
zhanghj@umich.edu
ORCID iD: 0000-0003-0545-5613

© Hongjiu Zhang 2019

Dedication

For you who enlighten my soul. For you who guide my feet.

Acknowledgments

I would like to thank my advisor Dr. Yuanfang Guan's support for her help in the past four years. The pragmatic aspect of algorithm design is often overlooked yet essential in real-world data analysis. Her advices regarding the issue, both during competitions and in other projects on method development, have profoundly changed my approach towards computational challenges. This has to be the most precious gift I have received through my PhD study and would surely make a difference in the future. I would like to thank my fellows in the Guan Lab, specifcially Dr. Ridvan Eksi, Zhengnan Huang, Hongyang Li, Zhengning Zhang, Shaocheng Wu, and Kaiwen Deng. The enormous support from my collegues during my PhD studies means a lot to me.

I would also like to thank Dr. Margit Burmeister, as my co-mentor, our graduate program director, and my friend, for her support during my PhD study. Without her support, I would not have met my mentor, decided to move forward to my PhD study, or have the chance to work with the top tier scientists at this wonderful place. I am also grateful to the support I received from my dissertation committee. Dr. Gilbert S. Omenn has co-mentored my research and exposed me to a variety of biological and medicinal topics. Through him, I have the chance to meet and work with the best researchers across the globe on the most challenging topics in different areas. Dr. Jun Li provided great support to many of my projects, from tumour heterogeneity to single-cell sequencing analysis. His broad knowledge of cutting-edge researches in the fields allows me to avoid many pitfalls in my work. Dr. Kayvan Najarian and Dr. Stephen Parker have been very generous to share their wisdom with me throughout my struggling journey in this interdisciplinary

field.

Throughout my PhD studies, I have received help from many of my collaborators. I want to thank Dr. Marcin Cieslik and Dr. Arul M. Chinnaiyan in their help in tumour heterogeneity projects, Dr. Henry Paulson and Dr. Hiroko Dodge for their help in the Alzheimer's disease challenge, Dr. Stephen Goutman, Dr. Bhramar Mukherjee, and Jonathan Boss for their help in the survival model analysis, Dr. Ebrahim Azizi, Dr. Sue Hammoud, Adrienne Shami, Dr. Anna Gilbert, Dr. Xiang Zhou, Dr. Joshua Welch for their help in single-cell sequencing projects, Dr. Hayley McLoughlin for her help in the ataxia project, Dr. Ryan E. Mills and Yifan Wang for their help in the transcript quantification project, Dr. Jianzhi Zhang and Dr. Zhengting Zou for their support in the phylogenetic inference project, and many others for their precious collaborative efforts. My PhD study is partially funded and supported by Ryss Tech, and I am grateful for their support as well.

I have received generous support from the staff in the department. Jonathan Poisson, Ken Weiss, and Aaron Bookvich have kept the computational infrastructure in a great shape. Julia Eussen and Amy Koger have been helping me on many administrative issues. Alex Terzian have been helping the audio-video settings and been always kind to me. And many more helped me in the past few years, including Suzanne Stevenson-Howard who is unfortunately no longer with us.

I would like to thank my friends and peers within the department, at the campus, and from local communities for making my past years the most splendid experience in my life. And last but not the least, I want to thank my families for their support to my life. I have not been back with my families for years, and I was unable to see my grandfather in his final days. He is the inspiration of my academic pursuit. My families have dealt with no less challenging situations than mine during the years of my absence, and they still fully supported me with no reservation. Without them, I can in no way reach this far.

Table of Contents

Dedication ii

Acknowledgments iii

List of Figures viii

List of Acronyms x

Abstract xi

Chapter 1. General introduction 1

1.1 Heterogeneity in protein evolution 3

1.2 Clonal heterogeneity in tumours 6

1.3 Transcriptomic heterogeneity at the molecular level 9

1.4 Preview 13

1.5 Reference 15

Chapter 2. Inferring quartet phylogeny using a convolutional neural network 28

2.1 Introduction 28

2.2 Methods 30

2.2.1 Deep neural network 30

2.2.2 Simulated phylogenetic datasets 32

2.2.3 Training and evaluation . . . 33
2.3 Results . . . 34
2.4 Discussion . . . 36
2.5 Conclusion . . . 38
2.6 Figures and Tables . . . 39
2.7 Reference . . . 42

Chapter 3. Resolving tumour heterogeneity through density-hinted optimization of mixture models **45**

3.1 Introduction . . . 45
3.2 Methods . . . 47
3.2.1 FastClone algorithm . . . 47
3.2.2 Benchmark . . . 53
3.3 Results . . . 54
3.4 Discussion . . . 57
3.5 Conclusion . . . 59
3.6 Figures and Tables . . . 60
3.7 Reference . . . 62

Chapter 4. Recovering single-cell transcriptomic expression profiles through weighted pooling **68**

4.1 Introduction . . . 68
4.2 Methods . . . 70
4.2.1 Seekmer algorithm . . . 70
4.2.2 Real-world sequencing data . . . 73
4.2.3 Simulated sequencing data . . . 74

4.2.4 Benchmarks 74
4.3 Results 75
4.4 Discussion 80
4.5 Conclusion 82
4.6 Figures and Tables 84
4.7 Reference 107

Chapter 5. Conclusion **111**

5.1 Reference 115

List of Figures

2.1 The prediction model for quartet phylogeny inference. 39

2.2 The validation performance of the deep neural network model along the training process. 40

2.3 The schematic of the Long branch attraction (LBA) problem. 41

3.1 An example of the density-hinted inference. 60

3.2 Conservative estimation of subclones yields better results. 61

4.1 Low read counts compromises the accuracy of transcript abundance estimation. 84

4.2 Low read counts hurts the estimation of transcripts of intermediate expression. . 85

4.3 Seekmer transcript quantification approach. 86

4.4 Seekmer read pooling approach. 87

4.5 Seekmer performs well on simulated single cell RNA sequencing data. 88

4.6 Seekmer is not sensitive to the weighting power parameter. 89

4.7 Seekmer greatly reduces the expression variation due to down sampling. 90

4.8 Seekmer preserves the cluster structure in the simulated data. 91

4.9 Seekmer removes biases of read counts from clustering. 92

4.10 Seekmer is not sensitive to cell counts (Log-Pearson scores). 93

4.11 Seekmer is not sensitive to cell counts (Spearman scores). 94

4.12 Seekmer can impute small cell clusters accurately (UHRR cells). 95

4.13 Seekmer can impute small cell clusters accurately (HBRR cells). 96
4.14 Seekmer imputation improves the quantification of SIRV spike-in sequences. . 97
4.15 Seekmer preserved intra-cluster variation. 98
4.16 Seekmer preserved cluster structure. 99
4.17 Seekmer clustering is not affect by imbalanced read counts. 100
4.18 Clusters of Fluidigm Polaris data in both raw quantification (left) and imputed results (right) are similar. 101
4.19 Pooly sequenced cells in Fluidigm Polaris data forms a separate cluster. 102
4.20 The t-SNE analysis of the Seekmer imputation over the Fluidigm Polaris dataset. 103
4.21 GYPA and GYPB genes are well expressed in K562 cells. 104
4.22 The SUM149 cells express the canonical transcript of EGFR (EGFR-201). . . . 105
4.23 The SUM149 cells express both the canonical transcript of CD44 and the transcript with v6 exon. 106

List of Acronyms

CNA copy number aberration.

DREAM SMC-Het Challenge DREAM Somatic Mutation Calling–Heterogeneity Challenge.

HBRR Human Brain Reference RNA.

LBA Long branch attraction.

qRT-PCR quantitative reverse-transcription polymerase chain reaction.

RNA-seq RNA sequencing.

scRNA-seq single-cell RNA sequencing.

SNV single-nucleotide variant.

t-SNE t-distributed stochastic neighbour embedding.

UHRR Universal Human Reference RNA.

Abstract

The complexity in biological data reflects the heterogeneous nature of biological processes. Computational methods need to preserve as much information regarding the biological process of interest as possible. In this work, we explore three specific tasks about resolving biological heterogeneity.

The first task is to infer heterogeneous phylogenetic relationship using molecular data. The common likelihood models for phylogenetic inference often makes strong assumptions about the evolution process across different lineages and different mutation sites. We use convolutional neural network to infer phylogenies instead, allowing the model to describe more heterogeneous evolution process. The model outperformes commonly used algorithms on diverse simulation datasets.

The second task is to infer the clonal composition and phylogeny from bulk DNA sequencing data of tumour samples. Estimating clonal information from bulk data often involves resolving mixture models. Unfortunately, simpler models are often unable to capture complex genetic alteration events in tumour cells, while more sophisticated models incur heavy computational burdens and are hard to converge. We solve the challenge through density-hinted optimization with *post hoc* adjustment. The model makes conservative predications but yields better accuracy in assessing co-clustering relationship among the somatic mutations.

The third task is to estimate the abundance of splicing transcripts from full-length single-cell RNA sequencing data. Transcript inference from RNA sequencing data needs a plethora of reads

for accurate abundance estimation. Yet single-cell sequencing yields much fewer reads than bulk sequencing. To recover transcripts from full-length single-cell RNA sequencing data, we pool reads from similar cells to help assign transcripts without disrupting the cluster structures.

These methods describe complex biological processes with minimal runtime overhead. Taking these methods as examples, we will briefly discuss the rationale and some general principals in designing these methods.

Chapter 1

General introduction

The heterogeneous nature of various biological processes is often portrayed through the complexity in various biological data [1]. The complexity may manifest in different forms. Studies on evolution processes of cells and species often encounter scenarios where the ever-changing environment leads to changes in selective pressure and thus in genotypes descendant cells or organisms, the complexity of which yields the heterogeneity of the observed population [2, 3]. Single-cell studies presents the complicated changes and regulations among different cell types, the complexity of which provides insights to heterogeneous behaviour of cells in multicellular tissue development [4, 5]. Multi-omics studies show drastic difference in the abundance of different gene products, the complexity of which reflects the heterogeneity in the molecular machinary [6–8]. The wide spectra of temperal, spatial, taxonomical, and individual differences in biological phenomena depicts the complexity of biological processes. These factors are not exclusive against each other. Therefore, fully understanding these processes would require accurate char-

acterization of biological heterogeneity.

Unfortunately, currently available experiment techniques rarely allow direct and complete determination of the complexity of these biological processes. Existing experiment protocols can measure just a limited portion of the molecular products that are present in the cells, track limited number of cells for various profiling, or observe only the final outcomes of a long-term ever-changing process [9–11]. Yet these observations capture important information useful for inference of missing elements that are not directly observable. The observed molecular sequences resulted from a long term evolution process provide information regarding distances between the species, which are helpful in phylogeny reconstruction [12]. Molecular products captured from a cell can be used to infer their relative abundance [13, 14]. Adequately sampled cells allow estimation of cell-type composition in a tissue and reveal the transitions between different cell states [15–17]. In ideal cases where no confounding factors are involved, the observation should provide enough information to recover the biological processes accurately.

However, biological signals are hidden inside a mixture of real signals, batch effects, technical biases, and random errors, more often than otherwise [18]. These factors contribute to the variation in the observation. If not handled correctly, such variation may drive the inference to generate artifacts or even completely wrong conclusions [19, 20]. To resolve the biological heterogeneity correctly, computational models need to discriminate the unwanted factors against the real signals [21–23]. The noise and biases in the observation add another layer of complication and hinder accurate determination of the original biological processes.

Therefore, good computational methods obviously need to achieve two goals: on one hand,

it should recover as much information as possible regarding the biological process of interest; on the other hand, from limited observed samples, the model needs to rule out confounding factors. These goals are challenging due to the "bias–variance tradeoff", a well-known dilemma in statistical learning that renders balancing these two targets difficult [24]. Thanks to the "no free lunch theorum" in statistical learning, the specific methods to optimise models vary for different data and tasks [25]. In this work, we will examine three specific challenging questions regarding biological heterogeneity in molecular evolution of species, tumour development, and multicellular tissue development. Here, we will review the biological background of these tasks in the following sections.

1.1 Heterogeneity in protein evolution

The sole illustrative figure in Charles Darwin's *On the Origin of Species* demonstrates the author's idea of the origin and evolutionary relationship of all species [26]. The tree diagram present in the figure, known as the (hypothetical) tree of life, depicts the scenario where all species share a common ancestor, i.e. the "Origin of Species", and species diverge from their ancestors to form new species and genera along time. The very idea of the tree of life, now reverred as the inception of evolutionary biology, is a dream of generations of biologists [27]. Reliable phylogenies that describes the evolutionary relationship of all living species provide a grand panorama of evolutionary histories and mechanisms, which is in turn the key to understanding virtually all biological phenomena.

In early years of evolutionary biology, the work of phylogeny reconstruction heavily relies on organismic and mechanistic approaches, as well as subjective judgement [28]. The situation lasted until last century, when advances in protein chemistry and molecular genetics changed the face of evolutionary biology [29]. Since then, protein and DNA sequencing techniques have revolutionised comparative biology [30]. Earliest compilation of partial protein and nucleotide sequences led by Margaret O. Dayhoff and her colleagues initiated the subsequent nonstoppable accumulation of sequence data for comparative studies [31–35]. These data provides an opportunity to establish an objective method for classification and phylogenetic analysis of different species [36]. Nowadays, DNA and protein sequences are the primary data used for phylogenetic inference in modern molecular evolution studies [37]. In this work, we will be focusing on phylogenetic reconstruction using protein sequences, but the principle can be easily extended to nucleic acid sequences with minimal changes.

Molecular evolution refers to the change in the sequences of biological molecules such as nucleic acids or proteins over time [38]. In the context of phylogeny analysis of protein sequences, we will discuss specifically missense point mutations. The molecular clock hypothesis suggests mutations occur at an steady rate [39–41] Since proteins are marginally stable, replacing amino acids with ones of different charges, side-chain sizes, hydrophobicity, or other biophysical properties can disrupt the structure and function of the protein [42]. These mutations are then fixed by random drift, purifying selection, or positive selection. The neutral theory of molecular evolution, which is now widely accepted, states that majority of mutations affect little on species' survivability and reproductivity, and that the random drift is the main reason behind the variations

between species at the molecular level [43]. Over time, new species emerge from accumulation of mutations.

Yet the above description oversimplifies the real scenario [44]. While numerous studies reported the overall mutation rates measured across species seem similar, the measurements disagree across different panels of species [45]. When exposed to different environments, different branches of the phylogeny may face different selection pressure and thus follow different mutation patterns [46]. Also, the functions of the same genes may be of different importance for different species, and thus they face different selection pressure [47]. Within a gene, the mutation rates of different mutation sites vary [48]. Different sites may also specify amino acid composition, which may also vary across branches [49, 50]. Furthermore, different sites may not evolve independently [51]. The branch-level and site-level variability intermingle and lead to heterogeneity in protein evolution.

Phylogeny reconstruction is to infer the phylogenetic relationship among species [52]. Nowadays, inference algorithms rely on the accumulation of mutation data to estimate the original evolutionary process. Inference rely on pre-defined models [3]. These methods often formulate molecular evolution as a continuous time Markov process, at each step of which a substitution model determines the probability of amino acid or nucleotide substitution [12]. Commonly used substitution models for protein sequence modelling specify transition matrices that are compiled statistically from aligned orthologous protein sequences, such as the Dayhoff matrix [53], the JTT matirx [54], and the LG matrix [55]. These matrices can be expressed as the product of the amino acid exchangability and the equilibrium frequencies of amino acids [12]. Methods like

maximum parsimony and neighbour joining calculate distances between species based on the substitution models and construct phylogenetic trees based on the distance matrices [31, 56]. On top of that, recent statistical inference model incorporates multiple parameters to address more heterogeneous scenarios [57–60]. Later in this work, we will explore further extension of phylogenetic inference modelling to handle more heterogeneous scenarios in protein evolution.

1.2 Clonal heterogeneity in tumours

Cancers are the second leading cause of death worldwide right after cadiovascular diseases [61]. Since the first case report of cancer in 1507, our understanding of cancers has expanded vastly [62]. Cancers results from uncontrolled proliferation of cells in the body [63]. Traditional classification of cancers is often based on the origin organs and tissues [64]. Whether cancer development starts from a single cell-of-origin or multiclonal tumour origin is still under debate [65, 66]. Yet both hypotheses admit that the development of cancers at the cellular level occurs in multiple steps involving somatic mutations and selections [67]. Tumour growth is also associated with changes in the microenvironment of the cells [68]. Newly grown vascular network allows the tumour cells to pick up glucose, amino acids and other nutrient molecules more rapidly [69, 70]. The tumour cells can also penetrate blood vessels and spread to other organs [71]. The migrated tumour cells can proliferate at different sites, leading to life-threatening metastatic spread of cancer tissue [72]. The complicated multi-stage process of cancer development involves multiple biological processes, and their underlying mechanisms are not fully understood.

Rapid advances in DNA/RNA sequencing techniques have enabled in-depth research of the molecular mechanisms of cancers. The view that cancers emerge from a series of mutations in genes and other variations eventually leading to changes in cell function is now widely accepted [73]. These mutations and variations may be germline that predispose individuals to higher risks [74], or somatic ones that eventually lead to abnormalities in cell functions [67]. Some cancer cells greatly rely on activation of a single oncogene to grow and survive [75], while others might have more complicated genetic mechanisms [76]. The accumulation of mutations and the selection of advantageous ones go together with the development of the tumour tissues [77]. Interestingly, the process resembles the evolution processes of asexually reproducing species [78]. As the principles of molecular evolution suggest and published studies show, cancer evolution processes also show branched clonal evolution, neutral evolution associated with non-adaptive mutations, and punctuated evolution associated with adaptive mutations [79]. Therefore, characterizing the mutations and variations involved in cancer evolution lies at the core of molecular pathology of cancers.

However, characterization of mutations and variations in tumour tissues is difficult due to their heterogeneous nature [80]. The heterogeneity occurs at different levels. First, the accumulation and selection of these variations within a single tumour tissue render its development a rapid evolutionary process [79]. Like any evolutionary process, tumour evolution also involves branching and clonal expansion [81]. This leads to genetic heterogeneity within tumour tissues. Second, tumour cells also live in a microenvironment that allows these cells to proliferate [82]. The supporting cells in the environment also undergo drastic changes, such as angiogenesis [69, 71] This

gives birth to a mixture of tumour cells whose expansion is uncontrolled and recruited surrounding cells that support the tumour development. Third, tumour cells can penetrate the vessels and migrate to other tissues or organs [83]. The migration, called metastasis, leads to another level of complexity. The migrated tumour cells adapt to the new environment and face different selection pressure [84]. These cells carry genetic features of their primary tumours of origin, but also acquires new ones that allow them to grow in the new site [82]. Also, the supporting cells in its new environment would be different from those at the original site [85]. The metastatic tumour tissues eventually become different from the primary tumours [86]. Fourth, different tumour types carry different genetic variations. Tumours of different tissue or organ origins carry distinct sets of genetic markers [87]. Even tumours emerged from the same organs may be different in their genetic features and thus can be further classified into subtyped [88]. Finally, the genetic difference among individuals, as well as the stochastic nature of somatic mutations and variations, determines the differences among different patients [89]. Adhering to conventions, we refer to the first two levels of the heterogeneity as "intratumour heterogeneity" and the remainders "intertumour heterogeneity" [90]. These levels of complexity underscore the challenge in characterizing cancers at the molecular level.

Both intertumour and intratumour heterogeneity cause phenotypic diversity and hinder accurate characterization and effective treatment of the disease [90]. Especially, intratumour genetic heterogeneity has been found in various tumour types, affecting response to therapy and relapse risks [91, 92]. The same tumour tissue may present different transcriptomic subtypes together [93], and possible codependencies of different subclones further obstruct the design of effec-

tive therapy [92]. Different treatment agents also select pre-existing drug-resistant clones in the tumour, eventually leading to polyclonal multi-drug resistant and therapy failure [94]. Deciphering tumour heterogeneity would be a paramount task in future development and improvement of minimally invasive therapeutic strategies.

Next-generation sequencing studies have driven the most recent research in tumour heterogeneity, especially intratumour heterogeneity, and revealed patterns of evolution that are clinically relevant [90]. Whole-genome sequencing of bulk tumour tissues reveals common genetic alterations such as single-nucleotide variants (SNVs), copy number aberrations (CNAs), and structural variations [95]. Single-nucleotide polymorphism arrays provide a more economic method to collect allelic frequencies over predefined loci and copy number profiles [96, 97]. In recent years, single-cell sequencing methods have become an exciting alternative approach to the bulk sequencing [98, 99]. These experiment methods collect genetic alterations that are informative of the cancer evolution process and allow computational algorithms to infer the cancer evolutionary processes. Later in this work, we will focus on inference using bulk sequencing data.

1.3 Transcriptomic heterogeneity at the molecular level

The central dogma of molecular biology, proposed by Crick in 1958, states that the genetic information encoded in the DNAs is transcribed into RNAs, and RNAs are then translated into proteins to carry out biological functions [100]. Since then, many discoveries about the molecular mechanisms in living cells have shown that the real biological process is far more complicated

and more finely regulated [101]. Among these discoveries, alternative splicing is among the most unexpected findings in molecular biology [102]. While splicing events happen in prokaryotic organisms like bacteria and archaea, eukaryotic splicing is far more frequent [103]. The mechanism behind the splicing events, the spliceosomal pathway, is ubiquitous in (examined) eukaryotes [104]. The discovery of alternative splicing also changed the general view of the eukaryotic gene model. Unlike single-exon protein-coding genes in many prokaryotes, eukaryotic genes are almost always discontinuous and consist of multiple protein-coding exonic segments and noncoding intronic segments [102]. Alternative splicing serves as a major mechanism to allow eukaryotic genes to generate diverse transcriptomic and protein products from much smaller number of genes [105]. Take human as example. While the human genome contains about 20,000 protein-coding genes, human produces over 90,000 different types of proteins [106]. The mechanisms allows more fine-grained regulation of products of gene expression, adding complexity to already intricate eukaryotic gene expression system [107].

Recent advances of the next-generation sequencing techniques revolutionised our way to study genetics [108]. Succeeding microarray as the major techniques in transcriptomics in the past decade, RNA sequencing (RNA-seq) has become a powerful tool to study gene expression, splicing, allelic-specific expression, and RNA editing [109]. A RNA-seq experiment, based on Illumina sequencing technology, starts with preparing reverse-transcribed full-length cDNAs from RNA molecules in the cells [110–112]. The cDNAs are then amplified, fragmented, and sequenced. In the end, RNA-seq yields paired sequences of both ends of short fragments of the transcript molecules. The sequences are then mapped against a reference genome or transcrip-

tome to assign their identities [113]. The alignment, and sometimes unaligned reads, are then used in subsequent computational analysis.

Specifically for splicing isoform analysis, a common task is to estimate the abundance of different transcripts, especially splicing isoforms [114]. A major challenge in transcrpt quantification is that the paired sequences of a short fragment from RNA-seq hardly allow unambiguous identification of the source transcript of the fragment [115]. Sequences of isoforms from the same gene are highly similar, and the sequence around the splicing sites are the only clue to differentiate different isoforms. Also, many isoforms share some splicing patterns. For these isoforms, only simultaneous capture of sequences around multiple splicing sites allows assigning read uniquely to their source isoforms. These requirements are almost impossible for short-read sequencing platforms like Illumina, as intervals between splicing sites often stretch much longer than fragments in illumina sequencing. Therefore, it is impossible to solve transcript quantification by unique transcript mapping.

Current algorithms for isoform quantification take a different approach. Programs like Cufflinks and RSEM infer the abundance of transcripts by optimizing a mixture model [114, 116, 117]. The optimised models proportionally assign reads to transcripts to their abundance and the transcript sequence lengths. Accurate inference requires adequate reads from RNA-seq. The inference is considered to be accurate enough for transcripts of intermediate and high expression levels [118]. Alternative methods like PacBio Iso-seq provides full-length sequencing of isoforms [119]. However, the transcripts detected from Iso-seq often conflicts with RNA-seq results and the performance is not well calibrated [120]. The experiment is more expensive than

RNA-seq [121]. Therefore, RNA-seq and computational abundance inference become the most popular approach to estimate the transcript abundance.

In the past few years, single-cell sequencing became an exciting technique to further improve the resolution of transcriptomic studies [5]. These methods aim at detecting the difference of transcripts within individual cells and are useful in studying transcriptomic heterogeneity among cells. Most single-cell sequencing techniques are also based on Illumina sequencing platform [122]. Methods like Drop-seq and 10x barcode the origin transcript molecules, and amplify the barcoded terminal of the molecules [123, 124]. The sequences are aligned and deduplicated based on the barcode sequences. Because the sequences are highly biased towards to the barcode terminal of the original molecule, these methods are used for gene-level single-cell transcriptomic analysis instead of isoform-level [5]. Other methods like Fluidigm Polaris and Smart-seq are much closer to traditional bulk RNA-seq [125, 126]. They generate full-length coverage over the original transcript molecules. Ideally, they can be combined with aforementioned quantification algorithms to study single-cell splicing events. Yet almost all single-cell sequencing methods nowaday generate much less reads per cell than the total reads in a bulk RNA-seq experiment [127]. Current available single-cell isoform pipelines often group transcripts in such analysis [128, 129]. Later in this work, we will discuss how to infer single-cell transcript abundance from limited number of reads.

1.4 Preview

We have introduced the biological background of the three tasks in solving heterogeneity in biological data. In all three cases, the observations result from random sampling. (The observed sequences of species are selected from successive random mutations. The detected allelic frequencies in tumours and The RNA sequencing reads for transcript quantification are sampled from amplified nucleic acid molecules.) The observations are all ambiguous and can be attributes to multiple possible events. (There can be multiple possible evolution paths leading to the same genomic sequences, each associated with different selection pressure. The allelic frequencies can be classified into different subclones with different likelihood, and the phylogenetic relationship among subclones is often ambiguous. The short read sequencing of transcriptomic can be mapped to different transcript splicing isoforms.) The observations are potentially associated with confounding factors that may mislead inference. (There may be unseen species that are informative of extra speciation events. The low allelic frequencies may be caused by biases of mutation callers or other technical factors. The single-cell RNA-seq reads are subject to amplification biases and potential drop-out events.) These tasks are computationally challenging.

Other than these three scenarios, there are many other types of heterogeneity that may manifests in biological data analysis. Each may presents a different kind of challenge to mathematical modelling and biological interpretation. In this following chapters, we will focus on the aforementioned three scenarios, discuss the merits and disadvantages of existing solutions, and proposes improved computational methods to address the issues. In the end, without losing generality, we will discuss some principles in algorithm designs for solving such challenges.

Chapter 2 focuses on the heterogeneity in the evolution process and presents a deep neural network model that estimates the phylogeny relationship among four species. The model addresses branch-level and site-level heterogeneity that existing solutions have not covered. It also improves the computational efficiency. The model is evaluated on various simulated models that are previously published and accepted.

Chapter 3 focuses on the tumour heterogeneity inference and presents a density-hinted mixture model that estimates the subclones inside a tumour sample. The model targets at recovering the co-occurrence relationship among mutations. To evaluate the performance of such model, the chapter discusses the pros and cons of different evaluation metrics and their impact on biological interpretation.

Chapter 4 focuses on the transcriptomic heterogeneity at the cellular level and presents a read-pooling method to recover the transcriptomic profiles from single-cell sequencing data. The method works at both gene and transcript levels and infer the abundance of the transcripts based on reads from cells that have similar gene expression levels. An important aspect of this study is to examine the contribution of biological and technical factors to the observed variations among the cells.

Chapter 5 summarises the methods above. While each previous chapter explains the biological merits of the individual model, this chapter briefly discusses the rationales and principles of algorithms for real-world biological questions.

1.5 Reference

1. Li, Y. & Chen, L. Big Biological Data: Challenges and Opportunities. *Genomics, Proteomics & Bioinformatics* **12,** 187–189. ISSN: 1672-0229 (Oct. 2014).

2. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168,** 613–628. ISSN: 0092-8674 (Feb. 2017).

3. Liò, P. & Goldman, N. Models of Molecular Evolution and Phylogeny. en. *Genome Research* **8,** 1233–1244. ISSN: 1088-9051, 1549-5469 (Dec. 1998).

4. Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N. & Werb, Z. Tumour heterogeneity and metastasis at single-cell resolution. En. *Nature Cell Biology* **20,** 1349. ISSN: 1476-4679 (Dec. 2018).

5. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. En. *Experimental & Molecular Medicine* **50,** 96. ISSN: 2092-6413 (Aug. 2018).

6. Genuth, N. R. & Barna, M. Heterogeneity and specialized functions of translation machinery: from genes to organisms. En. *Nature Reviews Genetics* **19,** 431. ISSN: 1471-0064 (July 2018).

7. Hernandez-Segura, A., Jong, T. V. d., Melov, S., Guryev, V., Campisi, J. & Demaria, M. Unmasking Transcriptional Heterogeneity in Senescent Cells. English. *Current Biology* **27,** 2652–2660.e4. ISSN: 0960-9822 (Sept. 2017).

8. Tanaka, T. S. Transcriptional heterogeneity in mouse embryonic stem cells. en. *Reproduction, Fertility and Development* **21,** 67–75. ISSN: 1448-5990 (Dec. 2008).

9. Ren, X., Kang, B. & Zhang, Z. Understanding tumor ecosystems by single-cell sequencing: promises and limitations. *Genome Biology* **19,** 211. ISSN: 1474-760X (Dec. 2018).

10. Nabhan, A. R. & Sarkar, I. N. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in Bioinformatics* **13,** 122–134. ISSN: 1467-5463 (Jan. 2012).

11. Ibrahim, J. G. & Molenberghs, G. Missing data methods in longitudinal studies: a review. *Test (Madrid, Spain)* **18,** 1–43. ISSN: 1133-0686 (May 2009).

12. Yang, Z. *Computational Molecular Evolution* en. ISBN: 978-0-19-856699-1 (OUP Oxford, Oct. 2006).

13. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4.** ISSN: 2046-1402. doi:10.12688/f1000research.7563.2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4712774/> (2019) (Feb. 2016).

14. Kanitz, A., Gypas, F., Gruber, A. J., Gruber, A. R., Martin, G. & Zavolan, M. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biology* **16,** 150. ISSN: 1465-6906 (July 2015).

15. Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H. & van Oudenaarden, A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. en. *Nature* **525,** 251–255. ISSN: 1476-4687 (Sept. 2015).

16. Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J. & Wigler, M. Tumour evolution inferred by single-cell sequencing. en. *Nature* **472,** 90–94. ISSN: 1476-4687 (Apr. 2011).

17. Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C. & Stegle, O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. en. *Nature Biotechnology* **33,** 155–160. ISSN: 1546-1696 (Feb. 2015).

18. Sloutsky, R., Jimenez, N., Swamidass, S. J. & Naegle, K. M. Accounting for noise when clustering biological data. en. *Briefings in Bioinformatics* **14,** 423–436. ISSN: 1467-5463 (July 2013).

19. Bergsten, J. A review of long-branch attraction. en. *Cladistics* **21,** 163–193. ISSN: 1096-0031 (2005).

20. Liu, S. & Trapnell, C. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research* **5.** ISSN: 2046-1402. doi:10.12688/f1000research.7223.1. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4758375/> (2019) (Feb. 2016).

21. Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. & Delsuc, F. Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology* **5,** 50. ISSN: 1471-2148 (Oct. 2005).

22. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. en. *Nature Methods* **14,** 417–419. ISSN: 1548-7105 (Apr. 2017).

23. Grün, D., Muraro, M. J., Boisset, J.-C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., van Es, J., Jansen, E., Clevers, H., de Koning, E. J. P. & van Oudenaarden, A. De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell* **19,** 266–277. ISSN: 1934-5909 (Aug. 2016).

24. Geman, S., Bienenstock, E. & Doursat, R. Neural Networks and the Bias/Variance Dilemma. *Neural Computation* **4,** 1–58. ISSN: 0899-7667 (Jan. 1992).

25. Wolpert, D. H. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation* **8,** 1341–1390. ISSN: 0899-7667 (Oct. 1996).

26. Darwin, C. R. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* eng (Apr. 1860).

27. Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., Crandall, K. A., Deng, J., Drew, B. T., Gazis, R., Gude, K., Hibbett, D. S., Katz, L. A., Laughinghouse, H. D., McTavish, E. J., Midford, P. E., Owen, C. L., Ree, R. H., Rees, J. A., Soltis, D. E., Williams, T. & Cranston, K. A. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. en. *Proceedings of the National Academy of Sciences* **112,** 12764–12769. ISSN: 0027-8424, 1091-6490 (Oct. 2015).

28. Laurin, M. The subjective nature of Linnaean categories and its impact in evolutionary biology and biodiversity studies. en. *Contributions to Zoology* **79,** 131–146. ISSN: 1875-9866, 1383-4517 (Oct. 2010).

29. Suárez-Díaz, E. Molecular Evolution in Historical Perspective. eng. *Journal of Molecular Evolution* **83,** 204–213. ISSN: 1432-1432 (Dec. 2016).

30. Anfinsen, C. B. *The molecular basis of evolution* (Wiley, New York, NY, 1959).

31. Dayhoff, M. O., Eck, R. V., Chang, M. A. & Sochard, M. *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Silver Spring, MD, 1965).

32. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. eng. *Nucleic Acids Research* **33,** D34–38. ISSN: 1362-4962 (Jan. 2005).

33. Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F. G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr, P., Tuli, M. A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W. & Apweiler, R. The EMBL Nucleotide Sequence Database. eng. *Nucleic Acids Research* **33,** D29–33. ISSN: 1362-4962 (Jan. 2005).

34. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. eng. *Nucleic Acids Research* **28,** 45–48. ISSN: 0305-1048 (Jan. 2000).

35. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. eng. *Methods in Molecular Biology (Clifton, N.J.)* **406,** 89–112. ISSN: 1064-3745 (2007).

36. Hagen, J. B. Naturalists, Molecular Biologists, and the Challenges of Molecular Evolution. *Journal of the History of Biology* **32,** 321–341. ISSN: 0022-5010 (1999).

37. Suárez-Díaz, E. & Anaya-Muñoz, V. H. History, objectivity, and the construction of molecular phylogenies. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **39,** 451–468. ISSN: 1369-8486 (Dec. 2008).

38. Nei, M. *Molecular Evolutionary Genetics* en. ISBN: 978-0-231-06321-0 (Columbia University Press, 1987).

39. Zuckerkandl, E. & Pauling, L. in *Evolving Genes and Proteins* (eds Bryson, V. & Vogel, H. J.) 97–166 (Academic Press, Jan. 1965). ISBN: 978-1-4832-2734-4. doi:10.1016/B978-1-4832-2734-4.50017-6. <http://www.sciencedirect.com/science/article/pii/B9781483227344500176> (2019).

40. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. eng. *Nature Reviews. Genetics* **6,** 678–687. ISSN: 1471-0056 (Sept. 2005).

41. Drake, J. W. A constant rate of spontaneous mutation in DNA-based microbes. en. *Proceedings of the National Academy of Sciences* **88,** 7160–7164. ISSN: 0027-8424, 1091-6490 (Aug. 1991).

42. Taverna, D. M. & Goldstein, R. A. Why are proteins marginally stable? en. *Proteins: Structure, Function, and Bioinformatics* **46,** 105–109. ISSN: 1097-0134 (2002).

43. Kimura, M. Evolutionary Rate at the Molecular Level. En. *Nature* **217,** 624. ISSN: 1476-4687 (Feb. 1968).

44. Camps, M., Herman, A., Loh, E. & Loeb, L. A. Genetic Constraints on Protein Evolution. *Critical reviews in biochemistry and molecular biology* **42.** ISSN: 1040-9238. doi:10.1080/10409230701597642. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3825456/> (2019) (2007).

45. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. Rates of Spontaneous Mutation. en. *Genetics* **148,** 1667–1686. ISSN: 0016-6731, 1943-2631 (Apr. 1998).

46. Baquero, F., Negri, M.-C., Morosini, M.-I. & Blázquez, J. Antibiotic-Selective Environments. en. *Clinical Infectious Diseases* **27,** S5–S11. ISSN: 1058-4838 (Aug. 1998).

47. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. en. *Proceedings of the National Academy of Sciences* **102,** 14338–14343. ISSN: 0027-8424, 1091-6490 (Oct. 2005).

48. Tajima, F. The Amount of DNA Polymorphism Maintained in a Finite Population When the Neutral Mutation Rate Varies Among Sites. en. *Genetics* **143,** 1457–1465. ISSN: 0016-6731, 1943-2631 (July 1996).

49. Foster, P. G. Modeling Compositional Heterogeneity. en. *Systematic Biology* **53,** 485–495. ISSN: 1063-5157 (June 2004).

50. Foster Peter G., Cox Cymon J. & Embley T. Martin. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364,** 2197–2207 (Aug. 2009).

51. Phillips, P. C. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. en. *Nature Reviews Genetics* **9,** 855–867. ISSN: 1471-0064 (Nov. 2008).

52. De Bruyn, A., Martin, D. P. & Lefeuvre, P. en. in *Molecular Plant Taxonomy: Methods and Protocols* (ed Besse, P.) 257–277 (Humana Press, Totowa, NJ, 2014). ISBN: 978-1-62703-767-9. doi:10.1007/978-1-62703-767-9_13. <https://doi.org/10.1007/978-1-62703-767-9_13> (2019).

53. *Atlas of Protein Sequence and Structure: Supplement* en. ISBN: 978-0-912466-07-1 (National Biomedical Research Foundation, 1978).

54. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. en. *Bioinformatics* **8,** 275–282. ISSN: 1367-4803 (June 1992).

55. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. eng. *Molecular Biology and Evolution* **25,** 1307–1320. ISSN: 1537-1719 (July 2008).

56. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. eng. *Molecular Biology and Evolution* **4,** 406–425. ISSN: 0737-4038 (July 1987).

57. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. en. *Systematic Biology* **59,** 307–321. ISSN: 1063-5157 (May 2010).

58. Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. en. *Systematic Biology* **61,** 539–542. ISSN: 1063-5157 (May 2012).

59. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. en. *Bioinformatics* **30,** 1312–1313. ISSN: 1367-4803 (May 2014).

60. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. en. *Molecular Biology and Evolution* **35,** 1547–1549. ISSN: 0737-4038 (June 2018).

61. Hassanpour, S. H. & Dehghani, M. Review of cancer from perspective of molecular. *Journal of Cancer Research and Practice* **4,** 127–129. ISSN: 2311-3006 (Dec. 2017).

62. Hajdu, S. I. A note from history. en. *Cancer* **116,** 2493–2498. ISSN: 1097-0142 (2010).

63. Cooper, G. M. en. in *The Cell: A Molecular Approach. 2nd edition* (2000). <https://www.ncbi.nlm.nih.gov/books/NBK9963/> (2019).

64. Schneider, G., Schmidt-Supprian, M., Rad, R. & Saur, D. Tissue-specific tumorigenesis – Context matters. *Nature reviews. Cancer* **17,** 239–253. ISSN: 1474-175X (Apr. 2017).

65. Rycaj, K. & Tang, D. G. Cell-of-Origin of Cancer versus Cancer Stem Cells: Assays and Interpretations. en. *Cancer Research* **75,** 4003–4011. ISSN: 0008-5472, 1538-7445 (Oct. 2015).

66. Parsons, B. L. Multiclonal tumor origin: Evidence and implications. *Mutation Research/Reviews in Mutation Research* **777,** 1–18. ISSN: 1383-5742 (July 2018).

67. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. en. *Science* **349,** 1483–1489. ISSN: 0036-8075, 1095-9203 (Sept. 2015).

68. Finicle, B. T., Jayashankar, V. & Edinger, A. L. Nutrient scavenging in cancer. En. *Nature Reviews Cancer* **18,** 619. ISSN: 1474-1768 (Oct. 2018).

69. Folkman, J. Tumor Angiogenesis: Therapeutic Implications. *New England Journal of Medicine* **285,** 1182–1186. ISSN: 0028-4793 (Nov. 1971).

70. Selwan, E. M., Finicle, B. T., Kim, S. M. & Edinger, A. L. Attacking the supply wagons to starve cancer cells to death. eng. *FEBS letters* **590,** 885–907. ISSN: 1873-3468 (Apr. 2016).

71. Nishida, N., Yano, H., Nishida, T., Kamura, T. & Kojiro, M. Angiogenesis in Cancer. *Vascular Health and Risk Management* **2,** 213–219. ISSN: 1176-6344 (Sept. 2006).

72. Lambert, A. W., Pattabiraman, D. R. & Weinberg, R. A. Emerging Biological Principles of Metastasis. English. *Cell* **168,** 670–691. ISSN: 0092-8674, 1097-4172 (Feb. 2017).

73. Loeb, L. A., Loeb, K. R. & Anderson, J. P. Multiple mutations and cancer. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 776–781. ISSN: 0027-8424 (Feb. 2003).

74. Ilyas, M., Straub, J., Tomlinson, I. P. M. & Bodmer, W. F. Genetic pathways in colorectal and other cancers. *European Journal of Cancer* **35,** 335–351. ISSN: 0959-8049 (Mar. 1999).

75. Weinstein, I. B. & Joe, A. Oncogene Addiction. en. *Cancer Research* **68,** 3077–3080. ISSN: 0008-5472, 1538-7445 (May 2008).

76. Sugimura, T., Terada, M., Yokota, J., Hirohashi, S. & Wakabayashi, K. Multiple genetic alterations in human carcinogenesis. *Environmental Health Perspectives* **98,** 5–12. ISSN: 0091-6765 (Nov. 1992).

77. Tomlinson, I. P. M., Novelli, M. R. & Bodmer, W. F. The mutation rate and cancer. *Proceedings of the National Academy of Sciences of the United States of America* **93,** 14800–14803. ISSN: 0027-8424 (Dec. 1996).

78. Tibayrenc, M., Avise, J. C. & Ayala, F. J. In the light of evolution IX: Clonal reproduction: Alternatives to sex. en. *Proceedings of the National Academy of Sciences* **112,** 8824–8826. ISSN: 0027-8424, 1091-6490 (July 2015).

79. Davis, A., Gao, R. & Navin, N. Tumor evolution: Linear, branching, neutral or punctuated? *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. Evolutionary principles - heterogeneity in cancer?* **1867,** 151–161. ISSN: 0304-419X (Apr. 2017).

80. Cyll, K., Ersvær, E., Vlatkovic, L., Pradhan, M., Kildal, W., Avranden Kjær, M., Kleppe, A., Hveem, T. S., Carlsen, B., Gill, S., Löffeler, S., Haug, E. S., Wæhre, H., Sooriakumaran, P. & Danielsen, H. E. Tumour heterogeneity poses a significant challenge to cancer biomarker research. *British Journal of Cancer* **117,** 367–375. ISSN: 0007-0920 (July 2017).

81. Gerlinger, M., Rowan, A. J., Horswell, S., Math, M., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A. & Swanton, C. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *The New England journal of medicine* **366,** 883–892. ISSN: 0028-4793 (Mar. 2012).

82. Joyce, J. A. & Pollard, J. W. Microenvironmental regulation of metastasis. en. *Nature Reviews Cancer* **9,** 239–252. ISSN: 1474-1768 (Apr. 2009).

83. Chambers, A. F., Groom, A. C. & MacDonald, I. C. Metastasis: Dissemination and growth of cancer cells in metastatic sites. en. *Nature Reviews Cancer* **2,** 563–572. ISSN: 1474-1768 (Aug. 2002).

84. Quail, D. & Joyce, J. Microenvironmental regulation of tumor progression and metastasis. *Nature medicine* **19,** 1423–1437. ISSN: 1078-8956 (Nov. 2013).

85. Pickup, M., Novitskiy, S. & Moses, H. L. The roles of TGFβ in the tumour microenvironment. en. *Nature Reviews Cancer* **13,** 788–799. ISSN: 1474-1768 (Nov. 2013).

86. Szekely, B., Bossuyt, V., Li, X., Wali, V. B., Patwardhan, G. A., Frederick, C., Silber, A., Park, T., Harigopal, M., Pelekanou, V., Zhang, M., Yan, Q., Rimm, D. L., Bianchini, G., Hatzis, C. & Pusztai, L. Immunological differences between primary and metastatic breast cancer. eng. *Annals of Oncology: Official Journal of the European Society for Medical Oncology* **29,** 2232–2239. ISSN: 1569-8041 (Nov. 2018).

87. Vaidyanathan, K. & Vasudevan, D. M. Organ Specific Tumor Markers: What's New? *Indian Journal of Clinical Biochemistry* **27,** 110–120. ISSN: 0970-1915 (Apr. 2012).

88. Dai, X., Xiang, L., Li, T. & Bai, Z. Cancer Hallmarks, Biomarkers and Breast Cancer Molecular Subtypes. *Journal of Cancer* **7,** 1281–1294. ISSN: 1837-9664 (June 2016).

89. Urbach, D., Lupien, M., Karagas, M. R. & Moore, J. H. Cancer heterogeneity: origins and implications for genetic association studies. *Trends in genetics : TIG* **28,** 538–543. ISSN: 0168-9525 (Nov. 2012).

90. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. en. *Nature* **501,** 338–345. ISSN: 1476-4687 (Sept. 2013).

91. Szerlip, N. J., Pedraza, A., Chakravarty, D., Azim, M., McGuire, J., Fang, Y., Ozawa, T., Holland, E. C., Huse, J. T., Jhanwar, S., Leversha, M. A., Mikkelsen, T. & Brennan, C. W. Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response. en. *Proceedings of the National Academy of Sciences* **109,** 3041–3046. ISSN: 0027-8424, 1091-6490 (Feb. 2012).

92. Inda, M.-d.-M., Bonavia, R., Mukasa, A., Narita, Y., Sah, D. W. Y., Vandenberg, S., Brennan, C., Johns, T. G., Bachoo, R., Hadwiger, P., Tan, P., DePinho, R. A., Cavenee, W. & Furnari, F. Tumor heterogeneity is an active process maintained by a mutant EGFR-induced cytokine circuit in glioblastoma. en. *Genes & Development* **24,** 1731–1745. ISSN: 0890-9369, 1549-5477 (Aug. 2010).

93. Sottoriva, A., Spiteri, I., Piccirillo, S. G. M., Touloumis, A., Collins, V. P., Marioni, J. C., Curtis, C., Watts, C. & Tavaré, S. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. en. *Proceedings of the National Academy of Sciences* **110,** 4009–4014. ISSN: 0027-8424, 1091-6490 (Mar. 2013).

94. Burrell, R. A. & Swanton, C. Tumour heterogeneity and the evolution of polyclonal drug resistance. *Molecular Oncology,* 1095–1111. ISSN: 1574-7891 (Mar. 2019).

95. Nakagawa, H. & Fujita, M. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Science* **109,** 513–522. ISSN: 1347-9032 (Mar. 2018).

96. Martinez, P., Kimberley, C., BirkBak, N. J., Marquard, A., Szallasi, Z. & Graham, T. A. Quantification of within-sample genetic heterogeneity from SNP-array data. En. *Scientific Reports* **7,** 3248. ISSN: 2045-2322 (June 2017).

97. Li, B. & Li, J. Z. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biology* **15,** 473. ISSN: 1474-760X (Sept. 2014).

98. Karaayvaz, M., Cristea, S., Gillespie, S. M., Patel, A. P., Mylvaganam, R., Luo, C. C., Specht, M. C., Bernstein, B. E., Michor, F. & Ellisen, L. W. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. En. *Nature Communications* **9,** 3588. ISSN: 2041-1723 (Sept. 2018).

99. Fan, J., Lee, H.-O., Lee, S., Ryu, D.-E., Lee, S., Xue, C., Kim, S. J., Kim, K., Barkas, N., Park, P. J., Park, W.-Y. & Kharchenko, P. V. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. eng. *Genome Research* **28,** 1217–1227. ISSN: 1549-5469 (2018).

100. Crick, F. H. On protein synthesis. eng. *Symposia of the Society for Experimental Biology* **12,** 138–163. ISSN: 0081-1386 (1958).

101. Thieffry, D. & Sarkar, S. Forty years under the central dogma. English. *Trends in Biochemical Sciences* **23,** 312–316. ISSN: 0968-0004 (Aug. 1998).

102. Lee, Y. & Rio, D. C. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual review of biochemistry* **84,** 291–323. ISSN: 0066-4154 (2015).

103. McGuire, A. M., Pearson, M. D., Neafsey, D. E. & Galagan, J. E. Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biology* **9,** R50. ISSN: 1465-6906 (2008).

104. Will, C. L. & Lührmann, R. Spliceosome Structure and Function. *Cold Spring Harbor Perspectives in Biology* **3.** ISSN: 1943-0264. doi:10.1101/cshperspect.a003707. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3119917/> (2019) (July 2011).

105. Roy, B., Haupt, L. M. & Griffiths, L. R. Review: Alternative Splicing (AS) of Genes As An Approach for Generating Protein Complexity. *Current Genomics* **14,** 182–194. ISSN: 1389-2029 (May 2013).

106. Valdivia, H. H. One gene, many proteins: alternative splicing of the ryanodine receptor gene adds novel functions to an already complex channel protein. eng. *Circulation Research* **100,** 761–763. ISSN: 1524-4571 (Mar. 2007).

107. Wang, Z. & Burge, C. B. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA* **14,** 802–813. ISSN: 1355-8382 (May 2008).

108. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. en. *Nature Reviews Genetics* **10,** 57–63. ISSN: 1471-0064 (Jan. 2009).

109. Costa, V., Angelini, C., De Feis, I. & Ciccodicola, A. Uncovering the Complexity of Transcriptomes with RNA-Seq. *Journal of Biomedicine and Biotechnology* **2010.** ISSN: 1110-7243. doi:10.1155/2010/853916. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896904/> (2019) (2010).

110. Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H. & Ecker, J. R. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133,** 523–536. ISSN: 0092-8674 (May 2008).

111. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. en. *Nature Methods* **5,** 621–628. ISSN: 1548-7105 (July 2008).

112. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. & Snyder, M. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. en. *Science* **320,** 1344–1349. ISSN: 0036-8075, 1095-9203 (June 2008).

113. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. & Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. eng. *Nature Protocols* **7,** 562–578. ISSN: 1750-2799 (Mar. 2012).

114. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. & Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. en. *Nature Biotechnology* **28,** 511–515. ISSN: 1546-1696 (May 2010).

115. Feng, J., Li, W. & Jiang, T. Inference of Isoforms from Short Sequence Reads. *Journal of Computational Biology* **18,** 305–321. ISSN: 1066-5277 (Mar. 2011).

116. Pachter, L. Models for transcript quantification from RNA-Seq. *arXiv:1104.3889 [q-bio, stat]*. arXiv: 1104.3889. <http://arxiv.org/abs/1104.3889> (2019) (Apr. 2011).

117. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. eng. *BMC bioinformatics* **12,** 323. ISSN: 1471-2105 (Aug. 2011).

118. Zhang, C., Zhang, B., Lin, L.-L. & Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18,** 583. ISSN: 1471-2164 (Aug. 2017).

119. Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J. C. & Ware, D. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. en. *Nature Communications* **7,** 11708. ISSN: 2041-1723 (June 2016).

120. Eksi, R. *Identification and Functional Annotation of Alternatively Spliced Isoforms* (PhD dissertation, University of Michigan, Ann Arbor, MI, 2017).

121. Akhter, S., Kretzschmar, W. W., Nordal, V., Delhomme, N., Street, N. R., Nilsson, O., Emanuelsson, O. & Sundström, J. F. Integrative Analysis of Three RNA Sequencing Meth-

ods Identifies Mutually Exclusive Exons of MADS-Box Isoforms During Early Bud Development in Picea abies. *Frontiers in Plant Science* **9.** ISSN: 1664-462X. doi:10.3389/fpls.2018.01625. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6243048/> (2019) (Nov. 2018).

122. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* **9,** 75. ISSN: 1756-994X (Aug. 2017).

123. Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A. & Kirschner, M. W. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161,** 1187–1201. ISSN: 0092-8674 (May 2015).

124. Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J. & Bielas, J. H. Massively parallel digital transcriptional profiling of single cells. en. *Nature Communications* **8,** 14049. ISSN: 2041-1723 (Jan. 2017).

125. Durruthy-Durruthy, R. & Ray, M. Using Fluidigm C1 to Generate Single-Cell Full-Length cDNA Libraries for mRNA Sequencing. eng. *Methods in Molecular Biology (Clifton, N.J.)* **1706,** 199–221. ISSN: 1940-6029 (2018).

126. Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S. & Sandberg, R. Full-length RNA-seq from single cells using Smart-seq2. en. *Nature Protocols* **9,** 171–181. ISSN: 1750-2799 (Jan. 2014).

127. Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I. & Enard, W. Quantitative single-cell transcriptomics. en. *Briefings in Functional Genomics* **17,** 220–232 (July 2018).

128. Ntranos, V., Kamath, G. M., Zhang, J. M., Pachter, L. & Tse, D. N. Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biology* **17,** 112. ISSN: 1474-760X (May 2016).

129. Ntranos, V., Yi, L., Melsted, P. & Pachter, L. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. En. *Nature Methods* **16,** 163. ISSN: 1548-7105 (Feb. 2019).

Chapter 2

Inferring quartet phylogeny using a convolutional neural network

2.1 Introduction

Vastly increasing genomic sequences of growing numbers of species allow the development of more advanced and realistic models for phylogenetic inference. While phylogenetic inference algorithms have evolved in the past decades and been successfully applied in many studies, they have been challenged by more complicated cases that demands more sophisticated inference strategies as well [1, 2]. Many recent statistical inference methods rely on likelihood optimization through heuristic search or Markov chain Monte Carlo sampling [3, 4]. The search algorithms, heuristics or statistical samplers, that scan over the parameter space, do not have the gaurantee to find the optimal solution in finite runtime, but work decently well for most real-world cases

[5–7]. The target likelihood models have incorporate many parameters, allowing more flexible and realistics modelling of the molecular evolution process [8–10]. Yet numerous cases of failed phylogenetic inference indicated that the current models may be inprecises and inadequate to model the molecular evolution [11, 12].

Expanding the likelihood model to infer more heterogeneous evolutionary process can be quite challenging. The current commonly used methods either assume these parameters are homogeneous across all branches, or require manual specification [13]. However, parameters such as amino acid compositions and other lineage-specific parameters vary across different species [14, 15]. Parameters may also show dependency among different mutation sites [16]. Modelling a panel of such species would require further extending the current statistical models. Several extensions are available for hard cases such as lineage-specific compositional heterogeneity [17, 18]. However, To infer the heterogeneity across the branch alongside tree searching can make the already computationally expensive inference process even less impractical [13, 19, 20].

We propose to use deep neural network to infer phylogeny for such case. Deep neural network is a machine learning method that allows fitting against almost arbitrary nonlinear functions. For our initial attempts, we focus on inference of quartet phylogeny, the evolutionary relationship of four species. Quartet phylogeny is the smallest possible question of phylogenetic inference in terms of the number of species. There can be three topologies for unrooted phylogenetic trees of four species. The problem can be formulated as a three-way classification problem, a well studied case in deep neural network. In this work, we will show the design of the neural network structures and benchmark the performance of the method on simulated datasets.

2.2 Methods

2.2.1 Deep neural network

The neural network structure is largely based on the design of the residual network. The input of the network is a matrix of encoded protein sequences of the target four species for prediction. The input sequences should be aligned, trimmed to the same length, and one-hot encoded. If the sequences are not aligned, gap positions are discarded. The size of the matrix would be $80 \times L$, where L is the length of the aligned sequences, i.e. the number of mutation sites of interest. The first dimension of 80 elements correspond to 20 amino acids for 4 species. For each site of the species, the amino acid sequence is one-hot encoded. The element in the tensor that corresponds to the amino acid at the site for the species will be labelled as 1, and the others 0. Therefore, for 80 elements in the matrix that correspond to a mutation site, there can be at most 4 ones (which label 4 amino acids for 4 species). The remaining elements are zeros.

The first layer of the network is a grouped convolutional layer that accepts 80 channels. The 80 channels correspond to the first dimension of the input matrix. The grouped convolutional layer splits the 80 channels into 20 groups, each group corresponds to one amino acid across four species. The layer convolves across four taxa for each amino acid at each site, with a kernel size of one. The output of the layer is a $80 \times L$ matrix, later batch normalized and then activated by a ReLU operator. The matrix is then send to a convolutional layer with a kernel size of one, batch normalization, and a ReLU activator. The output of the second convolutional layer has only 32 channels. The size of the output matrix becomes $32 \times L$. The matrix is then fed into an average

pooling layer to reduce the length of the sequences by half. The pooling layer average every two elements along the second dimension, keeping the first dimension intact. The output of the matrix is then $32 \times \frac{L}{2}$.

After the first two convolution layers and average-pooling layer, the resulting matrix is then fed into a series of four residual convolutional modules. Each module contains two groups of convolutional layers and one average pooling layer. Each group of convolutional layers have one convolutional layer with a kernel size of three and a padding size of one, one batch normalization layer, and a ReLU operator. All convolutional layers output 32 channels. Each residual module has a skip connection, adding the input matrix on top of the output of the last convolutional layer in the module. The summed output matrix has the same dimension of the input matrix and is then fed into the average pooling layer to reduce its second dimension size. The average pooling layer in the first three residual modules reduce the second dimension size of the matrix by half. The average pooling layer in the last residual module calculates the mean values across all elements along the second dimension, effectively reducing the size of the second dimension down to one. The output of all four rounds of residual convolutional operations is a vector of 32 elements. The vector is then fed into a fully connected layer which gives three output values, activated by a softmax operator. The final output is three values, corresponding to the probabilities of three different topologies of unrooted quartet phylogenetic trees.

2.2.2 Simulated phylogenetic datasets

The training and validation data are simulated using a likelihood model. The simulation first populates a large tree consists of 5 to 50 leaf taxa. For the root node of the tree, the simulator samples an initial amino acid equilibrium frequency profile. The profile is sampled, at each site, from known amino acid frequencies compiled from proteins found in 16 genes from thousands of species. Among all 16 protein groups, 13 are encoded in mitochondrial genomes, one is encoded in chloroplast genomes, and the remaining two are encoded in nuclear genomes. The simulator then sample the sequences at the root node of the tree following a multinomial distribution. Then the simulator iterates through all the branches of the tree. For each branch, the simulator first samples a branch length from a gamma distribution and mutation rates across all mutation sites from a gamma distribution as well. The branch length becomes the pseudo evolution time. Then the simulator decides whether a change in the mutation rate or the equilibrium frequency profile would occur. The probabilities of the changes are sampled from beta distributions. In case a change of mutation rates occur, the simulator swaps the mutation rates between randomly selected sites. In case a change of the equilibrium frequency profile occurs, the simulator swaps equilibrium frequencies of two randomly chosen amino acids. These two changes simulate the branch-wise heterogeneity and site-wise compositional heterogeneity. After that, the simulator mutates the sequence from the parent node of the branch following a continuous-time Markov process. The simulator samples a mutation time from an exponential distribution determined by the mutation rate for each site. If the mutation time at the site is longer than the remaining pseudo evolution time, the amino acid residue at the site is fixed. Otherwise, a new amino acid is sam-

pled based on a randomly chosen substition model. After sequences for all the nodes in the tree are available, the simulator randomly chooses four species from the tree and prune the tree to a quartet tree.

We added a LBA contaminated training set in ths study as well. In this dataset, the full tree simulation remains the same. Yet at the prune step, the four species are not randomly chosen. For a predifined ratio of the samples in the dataset, the simulator will try to find a LBA quartet tree. In this case, all quartet trees are enumerated and their LBA score is calculated. The score s is defined as

$$s = \frac{d_{\text{internal}} + d_{\text{short}}}{d_{\text{long}}} \tag{2.1}$$

where d_{internal} is the internal branch length, d_{short} is the length of the third longest leaf branch, and d_{long} is the length of the second longest leaf branch. The quartet tree with the smallest score is chosen.

2.2.3 Training and evaluation

The training of the deep neural network model uses cross-entropy as the final loss function and AdaBound as the optimizer. Each epoch of training involes 2000 simulated quartet trees, mutation sites of which are bootstrapped. The order of the four species are shuffled 24 times to enumerate all possible permutations. At the end of each training epoch, the model's performance is evaluated over an independent validation dataset simulated under the same parameter settings as the training dataset. The validation performance, in terms of cross-entropy as well, is used to select the best

epoch along the training process.

The evaluation is done over a series of simulated datasets using different simulation parameters. The performance is assessed in terms of the number of correctly predictedtrees. The performance is compared against MEGA X maximum parsimony, MEGA X neighbour joining, RAxML and Mr. Bayes. The evaluation used multiple simulated datasets with different parameters. The first is a simple quartet tree set with no heterogeneity introduced. The second is a general quartet tree set with all above mentioned heterogeneity activated. The third is a LBA set with all above mentioned heterogeneity activated and only trees with LBA.

2.3 Results

We formulated the inference of a quartet phylogeny into a three-way classification problem (Figure 2.1). The algorithm we use is a residual convolutional neural network. To brief the algorithm, the input of the algorithm is encoded protein sequences of the target four species for prediction. The encoded sequences is grouped by 20 amino acids at each mutation site and convolved across 4 species first. The resulting matrix goes through a series of convolutional operation and average pooling, reducing to a 32-element vector. The vector is then fed into a fully connected layer and a softmax operator. The final output of the network is three values, corresponding to the probabilities of three topologies of quartet phylogenetic trees.

The model is then evaluated over simulated dataset. The simulator uses a likelihood model to generate a phylogenetic tree of more than 4 species. The sequences are simulated following

a continuous-time Markov process. There are two aspects in the simuator that reflects the heterogeneity in evolution pressure. First of all, the simulator occasionally swap the mutation rates among random selected sites. This simulates the changing selection pressure over different sites in different lineage. Secondly, the simulators occasionally swap the equilibrium frequency profiles among random selected sites. This simulates the compositional heterogeneity in different lineage. The model is then trained and validated over datasets that are simulated in this way.

We first tested the model on a non-heterogeneous dataset, where the simulator does not include branch-wise or compositional heterogeneity. Out of 1000 test trees, the deep neural network model correctly predicted 992 trees. In comparison, MEGA maximum parsimony, MEGA neighbour joining, and RAxML correctly predicted 996, 994, and 992 trees, respectively. It is obvious that there is little difference betweem the performance of these tools over the non-heterogeneous dataset. While trained on a much more complicated dataset, the deep neural network model makes sane prediction for easy samples.

We then tested the model on heterogeneous datasets, where the heterogeneity parameters are the same as the training dataset. Figure 2.2 showed the performance of the deep neural network model over the two heterogeneous datasets and their comparison against other tools. Over a heterogeneous dataset with random sampled tree branches of 5000 quartet trees, the deep neural network model correctly predicted 4843 trees. In comparison, MEGA neighbour joining and PhyML also predicted correctly 4801 and 4791 trees, respectively. The difference among their performance is marginal. However, over a LBA heterogeneous tree sets, the difference became obvious. The LBA cases are often inferred wrongly due to their short internal branch length

(Figure 2.3). Methods such as MEGA maximum parsimony correctly inferred 2669 samples. Neighbour joining correctly predicted 3517 samples. The statistical models performed much better. The best is RAxML, which correctly recovered 3955 trees. Yet the fully trained deep neural network model performed even better. Epoch 900 predicted 4098 trees. Comparing the performance curve over the random sampled and LBA datasets, we noticed that the model very quickly learned to predict non-LBA samples, and its performance over non-LBA samples stays the same along the training process. The model gradually improves its performance over LBA samples along the training process.

2.4 Discussion

Current algorithms for phylogenetic inference make assumptions about the evolutionary process [12]. Maximum parsimony, a nonparametric model, assumes the observed difference in sequences results from minimal numbers of mutations. Distance-based methods such as neighbour joining not only specifies a model for the genetic distance between two species, also assumes that the final phylogenetic tree should be the balanced minimal spanning tree. Most likelihood models specifies the statistical model of various parameters that describe the evolutionary process [21]. When exposed to different environments, different branches of the phylogeny may face different selection pressure and thus follow different mutation patterns [22]. Also, the functions of the same genes may be of different importance for different species, and thus they face different selection pressure [23]. Within a gene, the mutation rates of different mutation sites vary [24].

Different sites may also specify amino acid composition, which may also vary across branches [25, 26]. Furthermore, different sites may not evolve independently [27]. The branch-level and site-level variability intermingle and lead to heterogeneity in protein evolution.

In the past decade, likelihood models have evolved to incorporate a lot more variables and became more realistic. Yet the branch-wise and site-wise heterogeneity in real-world evolution process dwarves our models. Most likelihood methods either ask the researchers to specify branch-wise substitution models or use the same substitution model for full tree inference [13]. The dynamics of the mutation rates is also often ignored in many current models [28]. Compositional heterogeneity also causes troubles for such analysis [18, 25]. With growing numbers of parameters in the models, the inference and optimization process becomes much more time consuming. Our work attempts to solve the problem by using an powerful approximator function to describe the process. The training process exposes the function to large number of tree samples, allowing it to approximate the heterogeneity without explicit specification of the parameters. The inference process, which involves simple calculation of feed-forward neural network, becomes much faster. The computational burden is shifted to the training process. The resulting model can be much more flexible to handle the heterogeneity in molecular evolution.

The current model demonstrated its power in capturing some of the hard cases in resolving heterogeneous phylogeny. Yet the limitation of the model is also obvious. The current structure of the deep neural network has a specified input dimension, which accepts only four species. Also, for more than four species, the phylogeny inference cannot be formulated as a classification problem with a definite number of classes. Therefore, the model does not capture more than

four species. Yet through large number of species subsampling, the model can generate multiple quartet trees and can be merged to infer phylogenies for more than four species. The current model does not fully utilise its convolution layers as no site-wise dependency is present in the training data. The data augmentation even relies on this feature to bootstrap sites. To include a proper model for site-wise dependency would allow the model to be more realistic and useful for exploring more interesting cases such as epistasis.

2.5 Conclusion

This work presents a deep neural network approach to predict quartet phylogeny. This approach relies on an approximate function to describe the evolutionary process, with no explicit parameters to specify during the inference process. The model uses simulated data for training. The simulation model can be extended without incurring major computational overhead. The model gives a flexible framework for phylogenetic researchers to study the lineage-wise and compositional heterogeneity in molecular evolution.

2.6 Figures and Tables

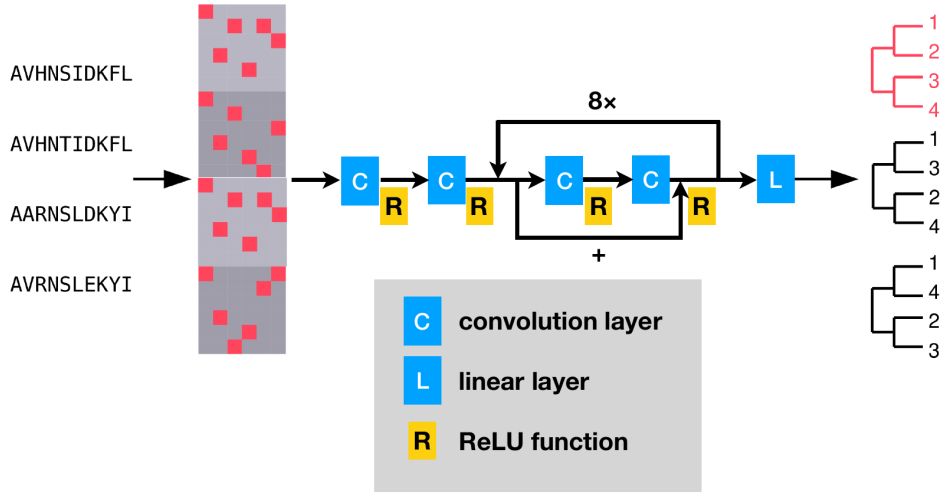


Figure 2.1: The prediction model for quartet phylogeny inference.
The neural network structure is largely based on the design of the residual network. The input of the network is a matrix of encoded protein sequences of the target four species for prediction. The first layer of the network is a grouped convolutional layer (with batch normalization and ReLU activation), followed by another convolution layer (with batch normalization and ReLU activation). Following these layers are eight residual modules, each has two convolutional layers. Finally the model uses a fully connected layer with softmax activation to predict the probabilities of three topologies of quartet trees. The model outputs the topology with the highest probabilities, highlighted in red.

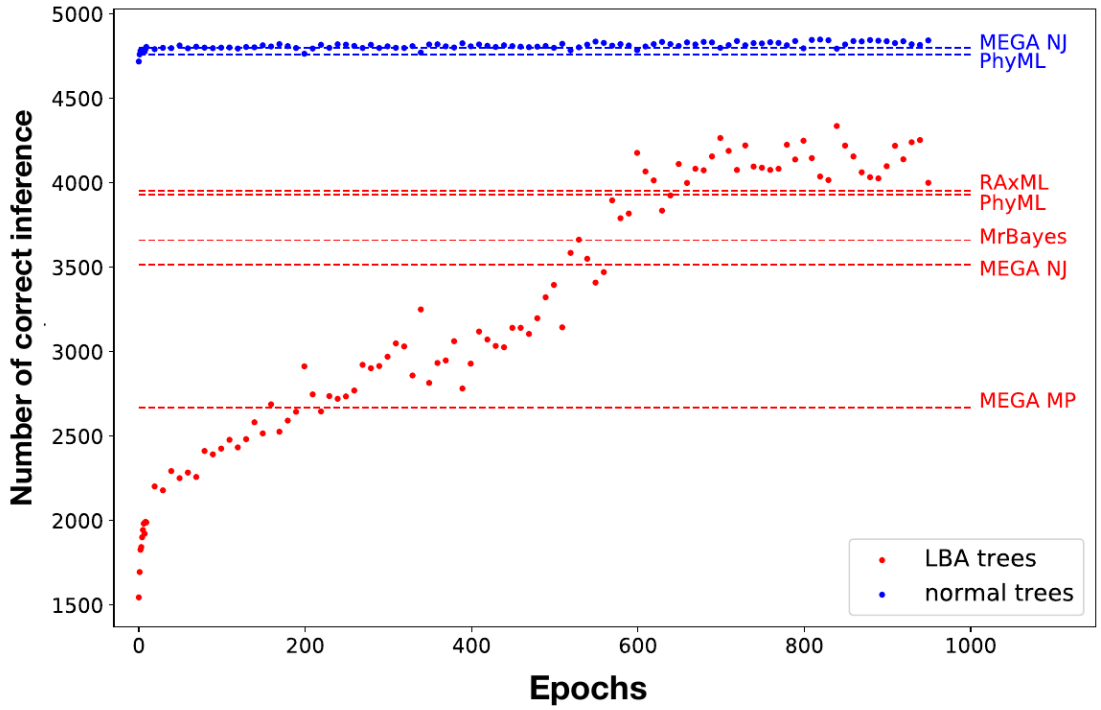


Figure 2.2: The validation performance of the deep neural network model along the training process.

The blue curve is the performance of the model over non-LBA trees, while the read are that over LBA trees. The performance is compared against MEGA maximal parsimony (MEGA MP), MEGA neighbour joining (MEGA NJ), RAxML, MrBayes, and PhyML.

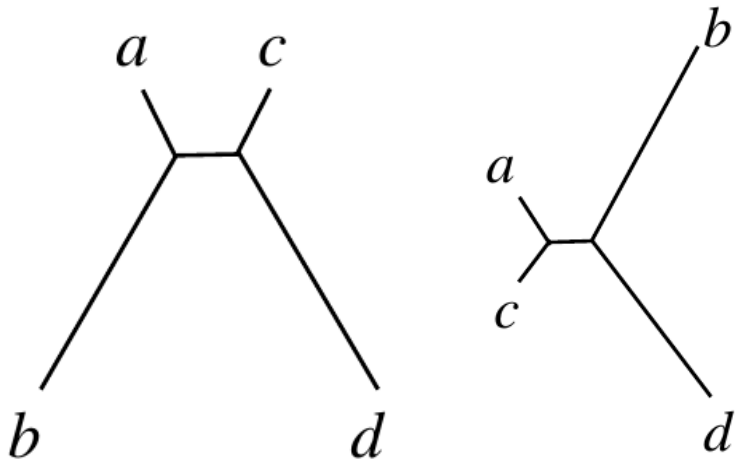


Figure 2.3: The schematic of the LBA problem.
The left unrooted tree shows a phylogeny with a short internal branch and unbalanced leaf branches. The evolution along the longer branches may introduce convergence in the molecular sequences. Misled by these wrong signals, the phylogeny inference methods might predict the topology of the right unrooted tree, which is wrong. This is the LBA.

2.7 Reference

1. Nei, M. *Molecular Evolutionary Genetics* en. ISBN: 978-0-231-06321-0 (Columbia University Press, 1987).

2. Felsenstein, J. *Inferring Phylogenies* en. Google-Books-ID: GI6PQgAACAAJ. ISBN: 978-0-87893-177-4 (Sinauer, Oct. 2003).

3. Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. en. *Systematic Biology* **61,** 539–542. ISSN: 1063-5157 (May 2012).

4. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. en. *Bioinformatics* **30,** 1312–1313. ISSN: 1367-4803 (May 2014).

5. Whelan, S. New Approaches to Phylogenetic Tree Search and Their Application to Large Numbers of Protein Alignments. en. *Systematic Biology* **56,** 727–740. ISSN: 1063-5157 (Oct. 2007).

6. Andreatta, A. A. & Ribeiro, C. C. Heuristics for the Phylogeny Problem. en. *Journal of Heuristics* **8,** 429–447. ISSN: 1572-9397 (July 2002).

7. Katoh, K. & Miyata, T. A heuristic approach of maximum likelihood method for inferring phylogenetic tree and an application to the mammalian SOX-3 origin of the testis-determining gene SRY. *FEBS Letters* **463,** 129–132. ISSN: 0014-5793 (Dec. 1999).

8. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. en. *Journal of Molecular Evolution* **17,** 368–376. ISSN: 1432-1432 (Nov. 1981).

9. Spencer, M., Susko, E. & Roger, A. J. Likelihood, Parsimony, and Heterogeneous Evolution. en. *Molecular Biology and Evolution* **22,** 1161–1164. ISSN: 0737-4038 (May 2005).

10. Rodrigue, N. On the Statistical Interpretation of Site-Specific Variables in Phylogeny-Based Substitution Models. *Genetics* **193,** 557–564. ISSN: 0016-6731 (Feb. 2013).

11. Yang, Z. & Zhu, T. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. en. *Proceedings of the National Academy of Sciences* **115,** 1854–1859. ISSN: 0027-8424, 1091-6490 (Feb. 2018).

12. Kolaczkowski, B. & Thornton, J. W. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. En. *Nature* **431,** 980. ISSN: 1476-4687 (Oct. 2004).

13. Groussin, M., Boussau, B. & Gouy, M. A Branch-Heterogeneous Model of Protein Evolution for Efficient Inference of Ancestral Sequences. en. *Systematic Biology* **62,** 523–538. ISSN: 1063-5157 (July 2013).

14. Zhang, J., Wang, X., Cheng, F., Wu, J., Liang, J., Yang, W. & Wang, X. Lineage-specific evolution of Methylthioalkylmalate synthases (MAMs) involved in glucosinolates biosynthesis. *Frontiers in Plant Science* **6.** ISSN: 1664-462X. doi:10.3389/fpls.2015.00018. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4315028/> (2019) (Feb. 2015).

15. Rosenberg, M. S. & Kumar, S. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. eng. *Molecular Biology and Evolution* **20,** 610–621. ISSN: 0737-4038 (Apr. 2003).

16. Fernandes, A. D. & Atchley, W. R. Site-specific evolutionary rates in proteins are better modeled as non-independent and strictly relative. eng. *Bioinformatics (Oxford, England)* **24,** 2177–2183. ISSN: 1367-4811 (Oct. 2008).

17. Rivera-Rivera, C. J. & Montoya-Burgos, J. I. LS^3: A Method for Improving Phylogenomic Inferences When Evolutionary Rates Are Heterogeneous among Taxa. *Molecular Biology and Evolution* **33,** 1625–1634. ISSN: 0737-4038 (June 2016).

18. Feuda, R., Dohrmann, M., Pett, W., Philippe, H., Rota-Stabelli, O., Lartillot, N., Wörheide, G. & Pisani, D. Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. *Current Biology* **27,** 3864–3870.e4. ISSN: 0960-9822 (Dec. 2017).

19. Feng, X., Buell, D. A., Rose, J. R. & Waddell, P. J. Parallel Algorithms for Bayesian Phylogenetic Inference. *J. Parallel Distrib. Comput.* **63,** 707–718. ISSN: 0743-7315 (July 2003).

20. Wen, D., Yu, Y. & Nakhleh, L. Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent. en. *PLOS Genetics* **12,** e1006006. ISSN: 1553-7404 (May 2016).

21. Yang, Z. *Computational Molecular Evolution* en. ISBN: 978-0-19-856699-1 (OUP Oxford, Oct. 2006).

22. Baquero, F., Negri, M.-C., Morosini, M.-I. & Blázquez, J. Antibiotic-Selective Environments. en. *Clinical Infectious Diseases* **27,** S5–S11. ISSN: 1058-4838 (Aug. 1998).

23. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. en. *Proceedings of the National Academy of Sciences* **102,** 14338–14343. ISSN: 0027-8424, 1091-6490 (Oct. 2005).

24. Tajima, F. The Amount of DNA Polymorphism Maintained in a Finite Population When the Neutral Mutation Rate Varies Among Sites. en. *Genetics* **143,** 1457–1465. ISSN: 0016-6731, 1943-2631 (July 1996).

25. Foster, P. G. Modeling Compositional Heterogeneity. en. *Systematic Biology* **53,** 485–495. ISSN: 1063-5157 (June 2004).

26. Foster Peter G., Cox Cymon J. & Embley T. Martin. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364,** 2197–2207 (Aug. 2009).

27. Phillips, P. C. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. en. *Nature Reviews Genetics* **9,** 855–867. ISSN: 1471-0064 (Nov. 2008).

28. Lynch, M. Evolution of the mutation rate. *Trends in genetics : TIG* **26,** 345–352. ISSN: 0168-9525 (Aug. 2010).

Chapter 3

Resolving tumour heterogeneity through density-hinted optimization of mixture models

3.1 Introduction

Tumour development is a rapid and heterogeneous evolutionary process [1]. Resolving tumour heterogeneity is helpful to therapeutic purposes yet challenging. Tumours are results of abnormal proliferation of cells [2, 3]. Along its development, cells accumulate mutations that eventually change the functions of the cells [4]. The development shows traits of linear evolution, branching

The chapter is adapted from the manuscript "FastClone infers subclonal composition and phylogeny: a DREAM Challenge benchmark study" by Hongjiu Zhang, Peter Ulintz, and Yuanfang Guan (in preparation). I co-designed and co-implemented the algorithm and conducted the benchmark evaluation.

evolution, neutral evolution, and puctuated evolution [5]. The result of the evolution process is heterogeneous tumour cells that are adapted to the local environment, rapidly proliferated, and resistant to immune responses and external treatments [6, 7]. Also, the tumour tissues recruit supporting cells from the host tissues and develop their microenvironment to facilitate their growth and metastatic migration to other tissues [8]. Tumour evolution presents a challenge to cancer treatment. Tumour cells that survived treatment and developed drug resistance would cause relapse and therapy failure [9]. Therefore, resolving tumour evolution is important to effective cancer treatment.

DNA sequencing techniques have empowered modern studies in tumour heterogeneity [10]. Deep whole-genome sequencing of bulk tumours provide a plethora of genetic alteration data to track tumour development [11]. Existing bioinformatics methods often use SNVs and CNAs to estimate the prevalence, the genetic alterations, and the phylogeny of the subclones in the tumour tissues [12–17]. While bulk sequencing data only provide averaged metrics of genetic alterations over multiple subclones, algorithms can statistically infer the subclones [18]. The algorithms attempt to find the most likely evolution process resulting sublones from which match the observed the genetic alteration data the best. Algorithms can even combine multiple, sometimes longitudinal, data of the same tumour tissues to further refine their inference [19]. However, there is not a consensus regarding the specific statistic models for such inference. Parsimony models can be computationally efficient, but more sophisticated models realistically describe complex genetic alteration events [20]. The challenge to infer heterogeneous tumour samples with low runtime overhead remains unsolved.

Here we present FastClone, a density-hinted mixture model to infer the heterogeneity of tumour tissues. FastClone describes the genetic alteration data of a tumour sample using a mixture model. Each component of the mixture model describes the likelihood of all SNVs belonging to a subclone. The model corrects the frequencies of the SNVs using CNAs information. FastClone takes the density of the allelic frequencies of somatic mutations as a hint to optimize a mixture model. The inference process takes only seconds for even tens of thousands of mutations. It reconstruct the phylogeny of the subclones through a post-hoc assignment. The algorithm won first place in DREAM Somatic Mutation Calling–Heterogeneity Challenge (DREAM SMC-Het Challenge) [21].

3.2 Methods

3.2.1 FastClone algorithm

FastClone takes both a copy number profile and somatic mutation frequencies as input. It starts by inferring the proportion of tumour cells in the sample. For each segment reported in the copy number profile, FastClone performs an initial estimation of the proportation of the tumour cells ρ. The estimation is based on B-allele frequencies (b) and the major and minor copy numbers of CNA (N_{major} and N_{minor} respectively). The major copy is the alleles that have a higher copy number, while the minor copy is the one with a lower copy number. Take CNA data from the autosomes and female X chromosomes first. If a segment has no CNA, or its major copy and minor copy in cancer cells are the same, the segment is not used in estimating cellularity. Otherwise,

the cellularity ρ can be inferred with

$$b = \frac{N_{\text{major}}\rho}{\left(N_{\text{major}} + N_{\text{minor}}\right)\rho + 2\left(1-\rho\right)}; \tag{3.1}$$

or if there is a 2-state CNA occur in the segment, which means that not all tumour cells carry the same CNA,

$$b = \frac{N'_{\text{major}}\rho f + N_{\text{major}}\rho\left(1-f\right) + \left(1-\rho\right)}{\left(N'_{\text{major}} + N'_{\text{minor}}\right) f\rho + \left(N_{\text{major}} + N_{\text{minor}}\right)\left(1-f\right)\rho + 2\left(1-\rho\right)} \tag{3.2}$$

where a proportion f of the tumour cells have different major and minor copy numbers of N'_{major} and N_{minor}, respectively. FastClone then estimates the proportion of tumour cells in the sample by averaging estimates over all segments without weighting.

After the initial purity estimation, FastClone clusters the mutations to identify subclones. FastClone first tries to calculate the prevalence of cells carrying individual SNVs based on allelic frequencies. If there are more than 50 SNVs within regions without CNAs, FastClone only calculates the prevalence for these SNVs. Given an autosomal SNV (or one on female X chromosomes) with an allelic frequency x, its prevalance p can be calculated from

$$p = 2x \tag{3.3}$$

For a SNV on male sex chromosomes, it is

$$p = x \tag{3.4}$$

If there are less than 50 SNVs within regions without CNAs, FastClone will also calculate the prevalence for those SNVs in regions of 1-state CNAs. Given an autosomal SNV (or one on female X chromosomes) with an allelic frequency x, its prevalance p can one out of three values. If the mutation occurs on the major copy before the CNA event,

$$p = \frac{x}{N_{\text{major}}} \left[\left(N_{\text{major}} + N_{\text{minor}} \right) \rho + 2 \left(1 - \rho \right) \right]; \tag{3.5}$$

if the mutation occurs on the minor copy before the CNA event,

$$p = \frac{x}{N_{\text{minor}}} \left[\left(N_{\text{major}} + N_{\text{minor}} \right) \rho + 2 \left(1 - \rho \right) \right]; \tag{3.6}$$

if the mutation occurs on either copy after the CNA event,

$$p = x \left[\left(N_{\text{major}} + N_{\text{minor}} \right) \rho + 2 \left(1 - \rho \right) \right]. \tag{3.7}$$

For a SNV on male sex chromosomes (therefore $N_{\text{minor}} = 0$), then its prevalence p is

$$p = \frac{x}{N_{\text{major}}} \left[N_{\text{major}} \rho + 2 \left(1 - \rho \right) \right] \tag{3.8}$$

if it occurs before the CNA event, or

$$p = x\left[N_{\text{major}}\rho + 2\left(1-\rho\right)\right] \tag{3.9}$$

if it occurs after the CNA event. FastClone then calculates the kernel density estimation function of the SNV prevalence. The density function $\hat{f}\left(x\right)$ is expressed as

$$\hat{f}\left(x\right) = \frac{1}{nZ}\sum_{j=1}^{n} \mathrm{e}^{\left(\frac{x-x_j}{h}\right)^2} \tag{3.10}$$

where n is the total number of SNV, x_j is the allelic frequency of i-th SNV, Z is the normalization constant to ensure $\int_{-\infty}^{+\infty} \hat{f}\left(x\right)\,\mathrm{d}x = 1$, and h is the bandwidth. We use Scott method to estimate the bandwidth h, that is

$$h = n^{-\frac{1}{d+4}} \tag{3.11}$$

for d-dimensional data (d samples). FastClone then enumerates all local maxima of the density function $\hat{f}$. It creates a grid over the real space with a resolution of 0.001, picks up any local maxima, and further optimises the values through gradient descent. FastClone uses these local maxima to specify clusters.

Then FastClone optimizes the mixture model

$$\lambda = \sum_{j=1}^{C} \ln \sum_{k=1}^{n} w_j L_{jk} \tag{3.12}$$

where C is the total number of maxima or clusters, w_j is the prevalence of the j-th cluster, and

L_{jk} is the probability of k-th mutation associated with the j-th subclone. For each mutation, its probability of being associated with a subclone is modelled as several modified binomial distributions

$$L_{jk} = \text{Binom}\left(x_k, r_k; \hat{x}_{jk}\right) \tag{3.13}$$

where L_{jk} is the probability of k-th mutation associated with the j-th subclone, x_k is the observed allelic frequency of the k-th mutation, $\hat{x}_{jk}$ is the expected allelic frequency of the k-th mutation if associated with the j-th subclone, and r_k is the total number of reads that cover the locus of the k-th mutations and passed the quality filter. Given the prevalence p_j of the j-th subclone, the expected allelic frequency of a mutation can be from following cases:

1. If the SNV is located in an autosomal or female X-chromosome region, there are three cases. If the mutation occurs on the major copy before the CNA event,

$$p_j = \frac{\hat{x}_{jk,\text{major}}}{N_{\text{major}}}\left[\left(N_{\text{major}} + N_{\text{minor}}\right)\rho + 2\left(1-\rho\right)\right]; \tag{3.14}$$

   if the mutation occurs on the minor copy before the CNA event,

$$p_j = \frac{\hat{x}_{jk,\text{minor}}}{N_{\text{minor}}}\left[\left(N_{\text{major}} + N_{\text{minor}}\right)\rho + 2\left(1-\rho\right)\right]; \tag{3.15}$$

   if the mutation occurs on either copy after the CNA event,

$$p_j = \hat{x}_{jk,\text{after}}\left[\left(N_{\text{major}} + N_{\text{minor}}\right)\rho + 2\left(1-\rho\right)\right]. \tag{3.16}$$

Then, $\hat{x}_{jk}$ will be the most likely value among these three, that is

$$\hat{x}_{jk} = \underset{\hat{x}\in\{\hat{x}_{jk,\text{major}},\hat{x}_{jk,\text{minor}},\hat{x}_{jk,\text{after}}\}}{\arg\max} \text{Binom}\left(x_k, r_k; \hat{x}\right). \tag{3.17}$$

2. If the SNV is located in a male sex chromosome region, there are two cases. Similar to the calculation listed above but with $N_{\text{minor}} = 0$. There is no case for mutation occurs on the minor copy, so $\hat{x}_{jk}$ will be the most likely value between the major-copy case and the after-copy case.

The probability is not related to the variables w_j. Therefore, the model can be optimized through usual expectation-maximization. In each iteration,

$$\hat{w}_j = \frac{1}{n}\sum_{k=1}^{n}\frac{w_j L_{jk}}{\sum_{k=1}^{n} w_j L_{jk}} \tag{3.18}$$

where $\hat{w}_j$ is the new value for the next round of the iteration. The weights are initialized with 1 at the beginning. With w_j and p_j fixed, all mutations can be assigned to a subclone with which the mutation has the maximum likelihood.

The phylogeny construction is done by iterating all possible tree structures and calculating the likelihood of the tree. There are two rules in iterating the tree structures:

1. A subclone with a smaller prevalence cannot be the parent of a subclone with a larger prevalance, the tree is not valid.

2. If the prevalence of a subclone is smaller than another in one sample but larger in another

sample, then these two subclones cannot be the parent of each other.

For the remaining trees, a likelihood is calculated by assuming that the prevalence ratios of the parents and the children follow a beta distribution. This assumption is consistent with the tree-structured stick-breaking process with the component distribution parameters fixed. The tree structure with the highest likelihood becomes the final result.

3.2.2 Benchmark

The benchmark evaluation is performed using data from the DREAM SMC-Het Challenge. The challenge organizers specified several tree structures and their subclonal compositions manually. To simulate real-world scenario, the challenge organizers used real-world sequencing data as the basis for the read generation. The reads were first aligned using BWA [22]. They were then analyzed using Battenberg to extract the copy number profile [12]. The challenge organizers specified a panel of mutations for each simulated dataset and inserted them into the sequencing reads using BAMSurgeon [23]. The modified reads were then processed using MuTect to call somatic mutations [15]. Participants of the challenge receive MuTect reports and copy number profiles and predict the subclonal structures.

The predictions are evaluated using multiple metrics. The number of subclones is evaluated based on relative error s as

$$s = 1 - \frac{\left| n_{\text{truth}} - n_{\text{prediction}} \right|}{n_{\text{truth}}} \tag{3.19}$$

where n_{truth} and $n_{\text{prediction}}$ are the simulated and predicted number of subclones, respectively. The

mutation assignment is evaluated based on the correlation of the co-clustering matrices. The false positive mutations called by MuTect are excluded from the assessment. The co-clustering matrix S of the remaining mutations are calculated as

$$S = PP^T \tag{3.20}$$

where P is the probability matrix of each mutation associated with each subclone ($n \times C$). The correlation between $S_{\text{prediction}}$ and S_{truth} is the score. The subclonal composition is evaluated based on the correlation of the predicted and simulated prevalences of cells carrying each mutations. The phylogeny is evaluated based on the correlation of the ancestor matrices. Again, the false positive mutations called by MuTect are excluded from the assessment. The ancestry matrix M of the remaining mutations are calculated as

$$M = PAP^T \tag{3.21}$$

where A is the asymmetric ancetor matrix of subclones ($C \times C$).

3.3 Results

FastClone applies several strategies to simplify the subclonal inference. FastClone takes both a copy number profile and somatic mutation frequencies as input. To find the subclones, FastClone first tries to identify subclones through kernel density estimation (Figure 3.1). FastClone

converts the allelic frequencies somatic mutation to the prevalences of the cells carrying these mutations. For mutations that are located in the regions of CNAs, FastClone simply iterate all possible solutions for cases in which the mutations occur on the major copy before CNAs occur, on the minor copy before CNA occur, or on either copy after CNAs occur. In the kernel density estimation, mutations with multiple possible solutions are weighted down equally. FastClone then calls all local maxima in the estimated distribution and uses them as the prevalences of sublcones. With the prevalences of subclones fixed, FastClone calculates the probabilities of all mutations associated with all subclones. It then infers the subclone composition using a mixture model. With the probabilities calculated, the mixture model can be optimised through expectation-maximization. FastClone then infer the phylogeny based on either the shallowest tree or the tree-structured stick-breaking process. Since the subclonal composition is solved, inference of tree-structured stick-breaking process becomes finding the solution with the highest likelihood of a beta distribution, which is much less time-consuming.

The model is then submitted to DREAM SMC-Het Challenge for independent evaluation. DREAM SMC-Het Challenge uses simulated data for assessment. To brief the simulation process, the assessment team specify several tree structures and their subclonal compositions. They also take real-world sequencing data as the basis of the simulation. Mutations are then inserted into the sequencing reads using BAMSurgeon [23]. The reads are then processed through BWA, Battenberg and MuTect to call somatic mutations [12, 15, 22]. Participants of the challenge receive MuTect reports and copy number profiles and predict the subclonal structures. Here the assessment focuses on two aspects: the subclonal composition and phylogeny and the assign-

ment of mutations to the subclones. We stick with the metrics used in the DREAM SMC-Het Challenge.

The challenge first evaluated the predicted cellularity and number of subclones of the tumour tissue. Since the monoclonal origin assumption of the simulation, the cellularity of the tumour tissue equals to the cellularity of the biggest subclone. Among all samples tested in the DREAM SMC-Het Challenge leaderboard, FastClone achieved a median relative error of 0.99 in cellularity estimation. It achieved an median absolute error of 1 in subclone number estimation. Peak identification approach tend to underestimate the number of subclones, as subclones with fewer mutations may not manifest as a subclone.

The challenge then evaluated the accuracy of assigning mutations to the subclones. The accuracy was measured in terms of the correlation of predicted and simulation co-clustering matrices. FastClone achieved the median correlation coefficient of 0.47. Major misassignment occurred due to missing subclones in the peak identification stage. To see whether this can be avoided, we conducted an additional test by enforcing the correct numbers and cellularity values of subclones and let FastClone assign the mutations to subclones using the same likelihood models. We do not see improvement in assignment. Figure 3.2 shows the two cases on one of the leaderboard sample. The nature of the likelihood model determines that the mutations that have the closest prevalence to the cellularity of the subclone are assigned together. This is not true in the simulation data and very unlikely in the real-world data. With one-dimensional data, the statistical models are unable to handle the dispersion of binomially distributed allelic frequencies.

Finally the challenge evaluated the accuracy of phylogeny inference, which is scored based on

the correlation of the predicted and simulated ancestry matrices. Because there are many possible phylogenies that yields the same subclonal composition, FastClone chooses a beta distribution model to estimate the tree structure. The model achieved a median correlation score of 0.69.

3.4 Discussion

Fast evolution of cancer cells is a great threat to cancer treatment [24]. Tumour cells that carry drug-resistant mutations may evade treatment and cause relapse [25]. Deep bulk DNA sequencing is a common solution to study tumour heterogeneity [26]. Through profiling CNAs and SNVs, researchers can infer the subpopulations present in the tumour sample and their genetypic information. There have been more than dozens of algorithms developed for tumour heterogeneity inference [16, 17, 27–42]. Many combines CNAs and SNVs data together to get more accurate results. The performance of these tools varies under different benchmark tests and different metrics. DREAM SMC-Het Challenge provides a great third-party independent assessment of different algorithms [21].

FastClone uses density-hinted optimization of mixture models to solve the subclonal decomposition problem. Without CNA, subclonal decomposition can be modelled as a binomial mixture, each component of which is a subclone. With a Dirichlet prior, such model can be easily optimized using variational inference. With CNA, the probability of SNVs vary and the distribution of a mutation belonging to a subclone no longer belongs to the exponential family. The optimization of such a model often involves Gibbs sampling, which is time consuming and hard to

converge on large dataset [43]. In our test, we show that almost no sampling-based software can handle more than thousands of mutations within reasonable runtime. Using density to specify the subclones greatly accelerates the inference. This approach does not gaurantee the mathematical optimal solution to the likelihood model, but works well on both simulated and real-world datasets. Also, sampling-based optimization over large dataset is hard to converge [43]. While theoretically these models eventually converge to the best solution in probability, it is quite time consuming in the real world. In the end, the extra computational burden does not pay off, and these methods may not perform well on simulated data from even the same models.

Evaluation of tumour heterogeneity inference needs well designed metrics. And in this challenge, we see a difficult case where different metrics drastically contradict with each other in some test samples. FastClone shows a tendency of under-clustering. In terms of the number of subclones, FastClone's prediction is biased. However, this is justified in our co-clustering analysis. Given the highly variable distribution of allelic frequencies, subclones with small numbers of mutations are much harder to detect. Forcefully calling these mutations leads to wrong assignment of mutations and can be quite misleading.

It is extremely hard to recover ancestry relationship from mutations, more so from a single sample. Many algorithms for heterogeneity inference chose to build the shallowest tree possible [44]. Recent statistical models rely on tree-structured stick-breaking process [16]. Neither assumption is supported by the statistical description of the evolution process. Yet, in real-world analysis, these methods work well for different cases. The performance of either methods is greatly affected by the depth and complexity of the tumour phylogeny. However, mutations and

copy number profiles are not enough to fully determine the evolution process of the tumour [35].

FastClone provides efficient inference of the subclonal composition and phylogeny from tumour sequencing data. There are multiple aspects of the tool can be further explored or improved. The current implementation handles multiple samples independently and merges the cluster from different samples to form concensus clustering. An alternative approach is to look for maxima from joint distribution of the recovered prevalences of mutations. Which performs better needs further tests. The algorithm currently accepts MuTect format only as its mutation data input. Considering Strelka is another popular somatic mutation calling tool, future development of FastClone will include an additional input wrapper for other file formats.

3.5 Conclusion

This work presents FastClone, an ultra-fast algorithm to infer subclonal composition and phylogeny from tumour DNA sequencing data. The density-hinted optimization of the mixture model greatly accelerates the inference process. The method avoids overspecifies the subclones and preserves the co-clustering relationship among mutations. It performed well on both simulated datsets and the real-world tumour data. In sum, FastClone empowers researchers to explore the heterogeneity of tumour tissues with minimal computational overhead.

3.6 Figures and Tables

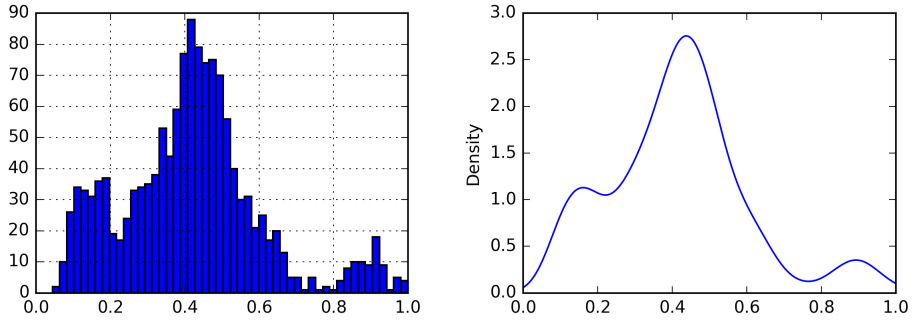


Figure 3.1: An example of the density-hinted inference.
FastClone first recover the prevalence for the mutations (left). It constructs a density function for the recovered prevalence (right). By identifying the local maxima, or peaks, in the density functions, FastClone determines the number of subclones for the following inference

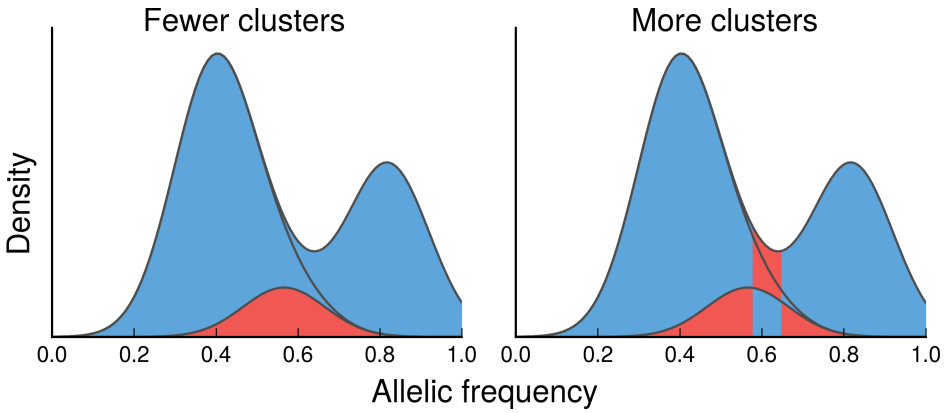


Figure 3.2: Conservative estimation of subclones yields better results.
The area of the red region is proportional to the number of misclassified mutations in each case. Taking a test sample from the DREAM SMC-Het Challenge as an example. On the left side, FastClone by default identifies two subclones from the two peaks from the density, though the simulation truth specified three samples. The cluster with fewer mutations cannot be inferred from the density. Yet enforcing the existence of the mutation, as shown on the right side, does not give a better prediction.

3.7 Reference

1. Ben-David, U., Beroukhim, R. & Golub, T. R. Genomic evolution of cancer models: perils and opportunities. En. *Nature Reviews Cancer* **19,** 97. ISSN: 1474-1768 (Feb. 2019).

2. Cooper, G. M. en. in *The Cell: A Molecular Approach. 2nd edition* (2000). <https://www.ncbi.nlm.nih.gov/books/NBK9963/> (2019).

3. López-Sáez, J. F., de la Torre, C., Pincheira, J. & Giménez-Martín, G. Cell proliferation and cancer. eng. *Histology and Histopathology* **13,** 1197–1214. ISSN: 0213-3911 (1998).

4. Loeb, L. A., Loeb, K. R. & Anderson, J. P. Multiple mutations and cancer. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 776–781. ISSN: 0027-8424 (Feb. 2003).

5. Davis, A., Gao, R. & Navin, N. Tumor evolution: Linear, branching, neutral or punctuated? *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. Evolutionary principles - heterogeneity in cancer?* **1867,** 151–161. ISSN: 0304-419X (Apr. 2017).

6. Bickel, S. T., Juliano, J. D. & Nagy, J. D. Evolution of Proliferation and the Angiogenic Switch in Tumors with High Clonal Diversity. en. *PLOS ONE* **9,** e91992. ISSN: 1932-6203 (Apr. 2014).

7. Casás-Selves, M. & DeGregori, J. How cancer shapes evolution, and how evolution shapes cancer. en. *Evolution* **4,** 624 (Dec. 2011).

8. Bussard, K. M., Mutkus, L., Stumpf, K., Gomez-Manzano, C. & Marini, F. C. Tumor-associated stromal cells as key contributors to the tumor microenvironment. en. *Breast Cancer Research : BCR* **18.** doi:10.1186/s13058-016-0740-2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4982339/> (2019) (2016).

9. Housman, G., Byler, S., Heerboth, S., Lapinska, K., Longacre, M., Snyder, N. & Sarkar, S. Drug resistance in cancer: an overview. eng. *Cancers* **6,** 1769–1792. ISSN: 2072-6694 (Sept. 2014).

10. Alizadeh, A. A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., Caldas, C., Califano, A., Doherty, M., Elsner, M., Esteller, M., Fitzgerald, R., Korbel, J. O., Lichter, P., Mason, C. E., Navin, N., Pe'er, D., Polyak, K., Roberts, C. W. M., Siu, L., Snyder, A., Stower, H., Swanton, C., Verhaak, R. G. W., Zenklusen, J. C., Zuber, J. & Zucman-Rossi,

J. Toward understanding and exploiting tumor heterogeneity. *Nature medicine* **21,** 846–853. ISSN: 1078-8956 (Aug. 2015).

11. Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., Dahlman, K. B., Pao, W. & Zhao, Z. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Medicine* **5,** 91. ISSN: 1756-994X (Oct. 2013).

12. Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S. L., Hinton, J., Menzies, A., Stebbings, L. A., Leroy, C., Jia, M., Rance, R., Mudie, L. J., Gamble, S. J., Stephens, P. J., McLaren, S., Tarpey, P. S., Papaemmanuil, E., Davies, H. R., Varela, I., McBride, D. J., Bignell, G. R., Leung, K., Butler, A. P., Teague, J. W., Martin, S., Jönsson, G., Mariani, O., Boyault, S., Miron, P., Fatima, A., Langerød, A., Aparicio, S. A. J. R., Tutt, A., Sieuwerts, A. M., Borg, Å., Thomas, G., Salomon, A. V., Richardson, A. L., Børresen-Dale, A.-L., Futreal, P. A., Stratton, M. R., Campbell, P. J. & Breast Cancer Working Group of the International Cancer Genome Consortium. The life history of 21 breast cancers. eng. *Cell* **149,** 994–1007. ISSN: 1097-4172 (May 2012).

13. Afyounian, E., Annala, M. & Nykter, M. Segmentum: a tool for copy number analysis of cancer genomes. eng. *BMC bioinformatics* **18,** 215. ISSN: 1471-2105 (Apr. 2017).

14. Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J. & Cheetham, R. K. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. eng. *Bioinformatics (Oxford, England)* **28,** 1811–1817. ISSN: 1367-4811 (July 2012).

15. Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S. & Getz, G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. eng. *Nature Biotechnology* **31,** 213–219. ISSN: 1546-1696 (Mar. 2013).

16. Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L. & Morris, Q. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology* **16,** 35. ISSN: 1465-6906 (Feb. 2015).

17. Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A. & Shah, S. P. PyClone: statistical inference of clonal population structure in cancer. en. *Nature Methods* **11,** 396–398. ISSN: 1548-7105 (Apr. 2014).

18. Liu, J., Halloran, J. T., Bilmes, J. A., Daza, R. M., Lee, C., Mahen, E. M., Prunkard, D., Song, C., Blau, S., Dorschner, M. O., Gadi, V. K., Shendure, J., Blau, C. A. & Noble, W. S.

Comprehensive statistical inference of the clonal structure of cancer from multiple biopsies. *Scientific Reports* **7.** ISSN: 2045-2322. doi:10.1038/s41598-017-16813-4. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5717219/> (2019) (Dec. 2017).

19. Jiang, Y., Qiu, Y., Minn, A. J. & Zhang, N. R. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **113,** E5528–E5537. ISSN: 0027-8424 (Sept. 2016).

20. Alves, J. M., Prieto, T. & Posada, D. Multiregional Tumor Trees Are Not Phylogenies. English. *Trends in Cancer* **3,** 546–550. ISSN: 2405-8033 (Aug. 2017).

21. Salcedo, A., Tarabichi, M., Espiritu, S. M. G., Deshwar, A. G., David, M., Wilson, N. M., Dentro, S., Wintersinger, J. A., Liu, L. Y., Ko, M., Sivanandan, S., Zhang, H., Zhu, K., Yang, T.-H. O., Chilton, J. M., Buchanan, A., Lalansingh, C. M., P'ng, C., Anghel, C. V., Umar, I., Lo, B., Participants, D. S.-H., Simpson, J. T., Stuart, J. M., Anastassiou, D., Guan, Y., Ewing, A. D., Ellrott, K., Wedge, D. C., Morris, Q. D., Loo, P. V. & Boutros, P. C. Creating Standards for Evaluating Tumour Subclonal Reconstruction. en. *bioRxiv,* 310425 (July 2018).

22. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. eng. *Bioinformatics (Oxford, England)* **25,** 1754–1760. ISSN: 1367-4811 (July 2009).

23. Ewing, A. D., Houlahan, K. E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T. N., Bare, J. C., P'ng, C., Waggott, D., Sabelnykova, V. Y., ICGC-TCGA DREAM Somatic Mutation Calling Challenge participants, Kellen, M. R., Norman, T. C., Haussler, D., Friend, S. H., Stolovitzky, G., Margolin, A. A., Stuart, J. M. & Boutros, P. C. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. eng. *Nature Methods* **12,** 623–630. ISSN: 1548-7105 (July 2015).

24. Enriquez-Navas, P. M., Wojtkowiak, J. W. & Gatenby, R. A. Application of Evolutionary Principles to Cancer Therapy. eng. *Cancer Research* **75,** 4675–4680. ISSN: 1538-7445 (Nov. 2015).

25. Bozic, I. & Nowak, M. A. Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers. eng. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 15964–15968. ISSN: 1091-6490 (Nov. 2014).

26. Aparicio, S. & Mardis, E. Tumor heterogeneity: next-generation sequencing enhances the view from the pathologist's microscope. *Genome Biology* **15.** ISSN: 1465-6906. doi:10.1186/s13059-014-0463-6. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4318188/> (2019) (2014).

27. Purdom, E., Ho, C., Grasso, C. S., Quist, M. J., Cho, R. J. & Spellman, P. Methods and challenges in timing chromosomal abnormalities within cancer samples. eng. *Bioinformatics (Oxford, England)* **29,** 3113–3120. ISSN: 1367-4811 (Dec. 2013).

28. Yau, C. OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. eng. *Bioinformatics (Oxford, England)* **29,** 2482–2484. ISSN: 1367-4811 (Oct. 2013).

29. Miller, C. A., White, B. S., Dees, N. D., Griffith, M., Welch, J. S., Griffith, O. L., Vij, R., Tomasson, M. H., Graubert, T. A., Walter, M. J., Ellis, M. J., Schierding, W., DiPersio, J. F., Ley, T. J., Mardis, E. R., Wilson, R. K. & Ding, L. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. eng. *PLoS computational biology* **10,** e1003665. ISSN: 1553-7358 (Aug. 2014).

30. Zare, H., Wang, J., Hu, A., Weber, K., Smith, J., Nickerson, D., Song, C., Witten, D., Blau, C. A. & Noble, W. S. Inferring clonal composition from multiple sections of a breast cancer. eng. *PLoS computational biology* **10,** e1003703. ISSN: 1553-7358 (July 2014).

31. Schwarz, R. F., Trinh, A., Sipos, B., Brenton, J. D., Goldman, N. & Markowetz, F. Phylogenetic quantification of intra-tumour heterogeneity. eng. *PLoS computational biology* **10,** e1003535. ISSN: 1553-7358 (Apr. 2014).

32. Qiao, Y., Quinlan, A. R., Jazaeri, A. A., Verhaak, R. G., Wheeler, D. A. & Marth, G. T. SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. eng. *Genome Biology* **15,** 443. ISSN: 1474-760X (Aug. 2014).

33. Hajirasouliha, I., Mahmoody, A. & Raphael, B. J. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. eng. *Bioinformatics (Oxford, England)* **30,** i78–86. ISSN: 1367-4811 (June 2014).

34. Fan, X., Zhou, W., Chong, Z., Nakhleh, L. & Chen, K. Towards accurate characterization of clonal heterogeneity based on structural variation. eng. *BMC bioinformatics* **15,** 299. ISSN: 1471-2105 (Sept. 2014).

35. Jiao, W., Vembu, S., Deshwar, A. G., Stein, L. & Morris, Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. eng. *BMC bioinformatics* **15,** 35. ISSN: 1471-2105 (Feb. 2014).

36. Malikic, S., McPherson, A. W., Donmez, N. & Sahinalp, C. S. Clonality inference in multiple tumor samples using phylogeny. eng. *Bioinformatics (Oxford, England)* **31,** 1349–1356. ISSN: 1367-4811 (May 2015).

37. El-Kebir, M., Oesper, L., Acheson-Field, H. & Raphael, B. J. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. eng. *Bioinformatics (Oxford, England)* **31,** i62–70. ISSN: 1367-4811 (June 2015).

38. Yau, C., Mouradov, D., Jorissen, R. N., Colella, S., Mirza, G., Steers, G., Harris, A., Ragoussis, J., Sieber, O. & Holmes, C. C. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. eng. *Genome Biology* **11,** R92. ISSN: 1474-760X (2010).

39. Letouzé, E., Allory, Y., Bollet, M. A., Radvanyi, F. & Guyon, F. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biology* **11,** R76. ISSN: 1474-760X (July 2010).

40. Greenman, C. D., Pleasance, E. D., Newman, S., Yang, F., Fu, B., Nik-Zainal, S., Jones, D., Lau, K. W., Carter, N., Edwards, P. A. W., Futreal, P. A., Stratton, M. R. & Campbell, P. J. Estimation of rearrangement phylogeny for cancer genomes. eng. *Genome Research* **22,** 346–361. ISSN: 1549-5469 (Feb. 2012).

41. Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M. & Getz, G. Absolute quantification of somatic DNA alterations in human cancer. eng. *Nature Biotechnology* **30,** 413–421. ISSN: 1546-1696 (May 2012).

42. Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. eng. *Genome Biology* **14,** R80. ISSN: 1474-760X (July 2013).

43. Turek, D., de Valpine, P. & Paciorek, C. J. Efficient Markov Chain Monte Carlo Sampling for Hierarchical Hidden Markov Models. en. <`https://arxiv.org/abs/1601.02698v1`> (2019) (Jan. 2016).

44. Strino, F., Parisi, F., Micsinai, M. & Kluger, Y. TrAp: a tree approach for fingerprinting subclonal tumor composition. eng. *Nucleic Acids Research* **41,** e165. ISSN: 1362-4962 (Sept. 2013).

Chapter 4

Recovering single-cell transcriptomic expression profiles through weighted pooling

4.1 Introduction

Alternative splicing, as the key mechanisms for diverse transcript and protein production in eukaryotic organisms, plays an important role in regulating cell proliferation, differentiation, and development [1]. Current sequencing techniques allows accurate transcript-level profiling at the bulk tissue level, but not yet at the cellular level [2]. Most widely used single-cell transcriptome

The chapter is adapted from the manuscript "Seekmer imputes transcript abundance from single-cell RNA sequencing data" by Hongjiu Zhang, Yifan Wang, Ebrahim Azizi, Gilbert S. Omenn, Ryan E. Mills, and Yuanfang Guan (in preparation). I designed and implemented the algorithm and co-conducted the benchmark evaluation.

sequencing methods are based on short-read sequencing platforms [3]. These platforms do not yield the full transcript sequences in single reads, but produce many fragments that cover different segments of the transcripts [4]. To estimate the transcript abundance, computational methods needs adequate fragment sampling to infer the relative abundance between different isoforms from the same genes [5]. Unfortunately, while the current single-cell sequencing methods yield decent amount of fragments per cell for gene-level expression analysis, accurate transcript-level profiling demands much more [6]. Before engineering solution is available, current bulk RNA-seq quantification algorithms might not be suitable for isoform quantification from single-cell RNA sequencing (scRNA-seq) data.

The limited number of reads in scRNA-seq is troublesome not only to isoform quantification, but also to many other analysis in scRNA-seq data analysis. Due to limited sampling of the transcript molecules, many cells often have only a few thousands of genes being sampled [7]. The lowly expressed genes are often missed in such sampling process, more often than what is specified in a naïve multinomial model, which is called dropout effect [8]. Due to this, scRNA-seq data are often inflated with zeros for genes that are expected to be found. For gene-level analysis, many imputation algorithms have been proposed to deal with such issues [9]. While the detailed algorithms differ, the basic idea is to borrow information from similar cells to help impute the missing values in the target cells [10–12]. Because such algorithms are often designed to deal with read count matrices, which are integer matrices, and not raw read data, they cannot deal with transcript imputation. Combining the likelihood models to model reads and the imputation approach to borrow information from similar cells might be key to estimate transcript-level

abundance from scRNA-seq data.

In this work, we present Seekmer, an algorithm to impute transcript abundance from full-length scRNA-seq data. The algorithm finds similar cells based on their gene expression profiles and pools reads from similar cells to facilitate the transcript abundance estimation. We applied the algorithm to a series of simulated data to test its robustness against various technical factors such as read counts and number of cells of same cell types. We found that Seekmer performs well over a wide range of read depths and cell cluster sizes. We further validated the algorithm on real-world spike-in data and cell line data. The imputation results preserved the clustering patterns in the original datasets while differentiates different splicing isoforms as well. The source code of Seekmer is hosted at `https://github.com/guanlab/seekmer`, and the software is available through Anaconda.

4.2 Methods

4.2.1 Seekmer algorithm

Seekmer consists of two part, an alignment-free mapper and an abundance estimator. The alignment-free mapper is similar to existing alignment-free transcript quantification tools such as Kallisto. The mapper needs to generate an index of reference transcriptomic sequences first. To build an index, Seekmer first collects all possible K-mers from reference transcriptomic sequences. Each K-mer is associated with a set of transcripts to which the K-mer can be mapped. Due to the high similarities between sequences of splicing isoforms from same genes, many K-mers can be

mapped to more than one transcripts. Seekmer groups together K-mers that are contiguous on the transcript sequences and share same sets of mappable transcripts. These grouped K-mers form contigs, as they occupy continuous segment of transcript sequences. Each contig has two terminal K-mers (and their reverse complements). These grouped K-mers, together with their sets of mappable transcripts, are the index for later mapping process.

When mapping a read, the mapper looks for a K-mer in the read that is also present in the Seekmer index. if a read has no such K-mer, the read is discarded. Starting from the matched K-mer, the mapper keep extending the match by iteratively jumping over the terminal K-mers of the contigs and mapping K-mers. By jumping over the contig, the mapper skips over K-mers that have the same mappable target transcripts, both speeding up the process and avoiding potential sequencing errors or mutations in the middle of the contig. The output of the mapper for each cell is a long list of transcript sets and how many reads can be mapped to these sets. These data are the input for the abundance estimator.

The abundance estimator of Seekmer takes the read mapping data and performs an initial estimatation the abundance of genes. The initial estimation optimizes a uniform mixture model, the log-likelihood function is expressed as:

$$\log L = \sum_{i=1}^{N_{\text{read}}} \log \sum_{j=1}^{N_{\text{transcript}}} \frac{\alpha_j}{l_j} \tag{4.1}$$

where l_j is the effective length of the j-th transcript, and $\frac{\alpha_j}{l_j}$ is the (length-normalized) abundance of the j-th transcript. The estimator estimates the gene abundance of each cell by adding the

abundance of all transcripts from same genes. Based on the initial estimation, the abundance estimator then calculates the Pearson correlation coefficients of gene expression for all pairs of cells as

$$\bar{g}_i = \frac{1}{N_{\text{gene}}} \sum_{k=1}^{N_{\text{gene}}} g_{ik} \tag{4.2}$$

$$r_{ij} = \frac{\sum_{k=1}^{N_{\text{gene}}} \left(g_{ik} - \hat{g}_i\right)\left(g_{jk} - \hat{g}_j\right)}{\sqrt{\sum_{k=1}^{N_{\text{gene}}} \left(g_{ik} - \hat{g}_i\right)^2} \sqrt{\sum_{k=1}^{N_{\text{gene}}} \left(g_{jk} - \hat{g}_j\right)^2}} \tag{4.3}$$

where g_{ik} is the expression level of k-th gene in i-th cell, and r_{ij} is the Pearson correlation coefficient between the gene expression profile of i-th cell and j-th cell. The estimator then applies K-mean clustering ($K = 2$) on the coefficients in the matrix and zeros out. The cluster of the lower coefficients in the correlation matrix are zero out. Seekmer then raise all elements in the processed correlation matrix to higher power to get the weight matrix W for all cells, that is

$$w_{ij} = r_{ij}^{p} \tag{4.4}$$

for unfiltered cell pairs, where w_{ij} is the contributing weight between i-th and j-th cells, and p is the power parameter. Given a target cell to impute, Seekmer build a model similarly to the uniform expression cells to the same number of the target cell. Then the read counts are multiplied by the precalculated weights. The new model for the estimation can be expressed as

$$\log L = \sum_{i=1}^{N_{\text{read}}} \log \sum_{j=1}^{N_{\text{transcript}}} \frac{\alpha_j}{l_j} \tag{4.5}$$

By optimizing the new scoring function, Seekmer estimates the transcript-level abundance of the cells.

4.2.2 Real-world sequencing data

Public available real sequencing data in this work includes Universal Human Reference RNA (UHRR), Human Brain Reference RNA (HBRR), and SMRT-seq sequencing of mouse embryonic cells with ERCC and SIRV spike-ins (Sample accession ID: SRR950078–SRR950087, E-MTAB-5481). The quantitative reverse-transcription polymerase chain reaction (qRT-PCR) assay results for UHRR and HBRR were generated from SEQC project as a 'gold standard' for RNA-seq quantification (GEO Accession: GSM1361812, GSM1361813). The SIRV spike-in reference sequences and annotations are available at `https://www.lexogen.com/sirvs/downloads/`https://www.lexogen.com/sirvs/downloads/.

The case study in this work uses a single-cell sequencing dataset on K562 and SUM149 cell lines. Breast cancer SUM149 and erythroleukemia K562 control cell lines were cultured in F12 and RPMI media, respectively. At about 80% confluency, cells were harvested from the culture flasks and diluted to about 300 cells/ul in PBS. Cell suspensions of SUM149 and K562 were separately processed using the Polaris instrument (Fluidigm, USA), 48-well full-length RNA-seq chip and reagents (Clontech and Fluidigm, USA). Captured single cells (48 breast cancer cells and 48 leukemia cells) were separately lysed to release total RNA and converted to cDNA libraries followed by pre-amplification of cDNAs all on the chip according to the manufacturer's protocol. The product of every single cell was then transferred to a well of a 96-well plate for

barcoding using the Nextera XT DNA library prep kit (Illumina, USA). Single cell barcoded products were pooled together in one lane for sequencing on HiSeq 2500. Raw sequencing data were processed through the following pipeline to determine gene expression patterns of every single cell of SUM149 and K562 cell lines.

4.2.3 Simulated sequencing data

All simulated sequencing data in this work were generated using RSEM (coupled with STAR aligner) [13, 14]. STAR and RSEM aligned and estimated the transcript abundance of a UHRR sample (SRR950078) and a HBRR sample (SRR950079) against ENSEMBL 90 human reference cDNA sequences. RSEM simulator then took the quantification results and generated 48 samples based on the UHRR sample and 48 based on the HBRR sample.

4.2.4 Benchmarks

The UHRR, HBRR, simulated, and Fluidigm Polaris samples were mapped against ENSEMBL 90 human reference cDNA sequences [15]. The mouse embryonic stem cells with spike-ins were mapped against ENSEMBL 90 mouse reference cDNA sequences. The quantification and imputation results of simulated 96-cell dataset, the mouse embryonic stem cell dataset, and the Fluidigm Polaris dataset were also analyzed using Seurat [16].

4.3 Results

Isoform quantification methods are able to quantify transcripts accurately with abundant reads from bulk RNA sequencing. When compared against qRT-PCR assays, (logarithm-transformed) estimated transcript abundance from methods such as STAR-RSEM can achieve correlation coefficients as high as 0.7 (Figure 4.1). Yet their performance on single cell data, each cell of which has much fewer reads, is not well studied. We performed a simulation experiment to benchmark their performance on samples with much fewer reads. Literature reports the numbers of reads per cell generated by commonly used single-cell sequence techniques range from a few tens of thousands to a few million. Therefore, we used RSEM to simulate samples with 10,000 to 1,000,000 reads based on UHRR and HBRR samples. As shown in the downsampling simulation experiment (Figure 4.2), the performance of these methods decreased to 0.2~0.5 when the number of reads dropped below a million reads. In the experiment, the simulator takes the estimated expression profile and generates reads accordingly. While the source expression profile of such simulation remains the same, the estimated transcript abundance can be drastically different, and they are also different from the source profile as well. The experiment on simulated samples with different number of reads shows that the fewer reads are present in the sample, the less accurate these methods estimate (Figure 4.1).

To improve the estimation accuracy for samples with limited number reads, quantification methods need more reads. Fortunately, in single cell sequencing, cells of same types often have similar gene-level expression profile and form clusters. Assuming cells with similar gene-level expression have similar transcript-level expression, reads from cells with similar gene expression

profiles may benefit transcript abundance estimation.

The algorithm design of Seekmer is based on the above read pooling idea. Seekmer consists of two part. The first part is an alignment-free mapper that resembles transcript quantification tools like Kallisto (Figure 4.3) [17]. To brief the algorithm, the mapper first collects all possible K-mers from reference transcriptomic sequences. Each K-mer is associated with a distinct set of transcripts to which the K-mer can be mapped. To facilitate the mapping process, the algorithm groups together K-mers that are next to each other on transcript sequences and share same sets of transcripts together. By convention, these groups formed by contiguous K-mers are called contigs. When mapping a read, the mapper searches for a K-mer that can be mapped to transcripts. Starting from the first matched K-mer, the mapper keep extending the match by iteratively jumping over the terminal K-mers of the contigs and mapping K-mers. The mapper then assigns the read to the intersection of transcript sets associated with all mapped contigs. Reads with no mappable K-mers are discarded. After mapping all the reads in each cell, the mapper gathers a long list of transcript sets and how many reads can be mapped to these sets. These mapping data are then used in the later imputation algorithm to guild cell pooling. In sum, this part is similar to common transcript quantification algorithms but calculates mappable read counts instead of the abundance.

The second part of Seekmer is the read pooling estimator (Figure 4.4). Seekmer takes the read mapping data and calculates the number of reads that can be uniquely mapped to each gene. Entries of different isoforms from same genes are merged, and ambiguous reads are discarded. The gene read counting results is similar to those from HTseq annotation. Seekmer then calculates

the Pearson correlation coefficients of gene read counts for all pairs of cells, which gives a correlation matrix. Seekmer also applies K-mean clustering ($K = 2$) on the coefficients in the matrix. The clustering differentiates the higher coefficients among cells with similar expression profiles against those lower ones among cells with different profiles. The lower coefficients in the correlation matrix are zero out. Seekmer then raise all elements in the processed correlation matrix to higher power to get the weight matrix for all cells. The weight matrix reduces the contribution of cells with less similar profiles, so the final abundance inference for a cell will be driven by the reads from the cell itself and those cells with almost identical expression profiles. After getting the weight matrix, Seekmer imputes the abundance of transcripts by pooling reads from other cells. The statistical model for the imputation is a similar model to traditional transcript quantification tools like Cufflinks, RSEM, and Kallisto, but with reads from other cells. Given a target cell to impute, Seekmer first normalizes read counts from other cells to the same number of the target cell. Then the read counts are multiplied by the precalculated weights. By optimizing the scoring function of the model, Seekmer estimates the transcript-level abundance of the cells.

We first tested Seekmer on a dataset simulated by RSEM. Given the expression profiles of the UHRR and HBRR samples, the RSEM simulator generated 48 samples similar to UHRR and 48 to HBRR. Each of these samples have varying number of reads from 10,000 to 1,000,000, the amount of which is similar to the numbers of reads per cell obtained from commonly used single-cell sequencing techniques. Therefore, the simulated dataset resembles a testing 96-cell sequencing data from Fluidigm full-length sequencing platform on two cell types. Figure 4.5 shows the performance of Seekmer on this simulated single-cell dataset. The imputation improved the cor-

relation of the transcript abundance against the qRT-PCR results from 0.54 to 0.72, much closer to the bulk results of 0.8. The performance is not sensitive to the change of the power parameter (Figure 4.6). The improvement signifcantly reduces the variation caused by the downsampling (Figure 4.7). Yet the difference between the clusters is preserved (Figure 4.8). Imputation also alleviated the artifact in the clusters introduced by imbalanced read counts (Figure 4.9).

One concern against such weight pooling approach is that the model may mistakenly pool in cells that are actually dissimilar. This may happen in cases for cells with number of reads that are too low. The resulting gene count data of these cells can be moderately similar to arbitrary cells in the sample. The situation might get worse when these poorly sequenced cells belong to a small cluster in comparison to other clusters present in the dataset. Yet as shown in Figures 4.10 and 4.11, Seekmer performed well in imputing undersampled cells with low read counts. In this test, the numbers of cells in each cell type vary from 6 to 48. While the performance of Seekmer indeed dropped slightly when the number of cells decreased (from 0.7 to 0.62), the resulting performance is still much better than quantification on individual cells (0.3). Specifically on those poorly sequenced cells whose numbers of reads are lower than 100,000, Seekmer was able to achieve correlation coefficients as high as 0.6 as well (Figure 4.12 and 4.13).

To further assess how well Seekmer differentiate different isoforms from same genes, we further tested Seekmer on real-world SMRT-seq data of mouse embryonic stem cells with spike-ins. The dataset contains SIRV spike-ins, sequences of which resembles alternative splicing isoforms from eukaryotic organisms. The SIRV spike-in data can be used to evaluate the isoform-level imputation results. An interesting point of the SIRV spike-ins is that the reference annotations

contain entries that are not present in the spike-in sample. This resembles the real-world transcript quantification scenarios. Eukaryotic genes may generate tens of different splicing isoforms, but they may not be present in a single cell at the same time. These decoy entries allow testing of whether Seekmer's imputation can reduce the false positives.

Figure 4.14 shows that Seekmer performed well on this real-world benchmark dataset. Seekmer uniformly improved the quantification by read-pooling imputation. The average performance was improved to 0.6. Unlike above simulated cases, Seekmer did not collapse the clusters into almost a single dot in the PCA plot but preserved the intra-cluster variation (Figure 4.15). Yet the cluster are well separated in this case. The similar pattern can be seen in the t-distributed stochastic neighbour embedding (t-SNE) analysis as well (Figure 4.16). The cluster pattern is quite similar to what is observed in direct quantification results, but the separation is much more obvious. The clustering is free of the artifact from read counts (Figure 4.17).

In addition to the spike-in validation, we also applied Seekmer to a Fluidigm Polaris sequencing data on two cell lines. One is the leukemia cell line K562, and the other is the triple-negative breast cancer cell line SUM149. Both cell lines are well studied and have many splicing isoforms have been characterized. Again, Seekmer imputation results show similar clustering patterns to that of the direct quantification profiles (Figure 4.18). In this case, a group of poorly sequenced cells forms a small cluster, mixing both types of cells (Figure 4.19). The similar cluster pattern is also preserved in the t-SNE analysis as well (Figure 4.20). Other than the poorly sequenced cell cluster, the other clusters well separated the cells of two cell lines. Both clusters of cell lines are well characterized by marker genes. For example, both GYPA and GYPB genes are

highly expressed in K562 cells (Figure 4.21) And EGFR is highly expressed in SUM149 cells (Figure 4.22). Another interesting example is CD44, which is known to have alternative splicing in SUM149 cells. Figure 4.23 shows that not only the canonical transcript of CD44 is expressed in the SUM149 cells, the v6 transcript (CD44-206) is also detected.

4.4 Discussion

Seekmer aims at providing transcript-level abundance information for single-cell studies. Most single-cell studies focus on gene-level transcriptomic characterization, but not much at the transcript level [2]. This is mainly due to the limitation of the current single-cell sequencing techniques. Commonly used single-cell sequencing methods need to deal with the trade-off between covering large number of cells and sampling more transcripts in each cell [18]. Yet even methods optimized for in-depth analysis for each cell yield much less reads than bulk RNA sequencing [6]. Methods like SMART-seq 2 generate a few million reads per cell [19]. Without enough reads, the performance of transcript inference drop significantly, both reported and shown in our simulation test [20–22]. On top of that, the most commonly used single-cell sequencing platforms such as Drop-seq and 10x sequencing rely on barcodes to track the identity of the transcript molecules [23, 24]. These barcodes are often attached at the terminal of the transcript molecules and occupy one end of the pair-ended sequencing. This results in single-end sequencing of the actual transcript sequences and heavily terminal-biased sequencing [25]. These are not desirable in splicing isoform inference. Because short-read sequencing cannot capture all splicing junctions

in one single-reads, isoform inference algorithms needs reads of (ideally) even coverage over the transcript molecules. This is critical for the statistical model in isoform inference algorithms. Unfortunately, due to the aforementioned reasons, it is difficult to obtain transcript-level information from single-cell RNA-seq data [26]. In sum, even coverage and adequate reads per cell are the two major challenges of inferring single-cell transcript abundance.

While single-cell sequencing platforms such as SMRT-seq provides less biased coverage, the requirement on the numbers of reads per cell is difficult to accomodated through engineering methods at the moment. From the algorithm aspect, inference algorithms can be much more accurate if they can borrow reads from cells with similar expression profiles [9]. This is the key to imputation. However, calculating the similarities among the transcriptomic profiles of cells needs the profiles themselves, thus forming a dilemma. To break the cycle, Seekmer uses the gene-level expression profiles to calculate the similarities. It assumes that similar gene-level expression implies similar isoforms-level expression. The assumption seems quite strict, yet justifiable. Popular single-cell sequencing analysis still rely on clustering of cells over their gene-level expression profiles to determine the expression profiles of the cell types [16, 27, 28]. We also show that the observed variations within the cell clusters are partially attributed to the poorly sampled sequences. In our simulation test, such pooling approach not only recovered the expression profiles of individual cells, also resisted the effect of dissimilar cells even when imputing cells from a much smaller cluster. The precedent algorithms and our experiments well justified the algorithm choice.

There are a few limitations in Seekmer right now. One limitation is that Seekmer calculates

the similarity between cells based on read counts of all genes. Considering most platforms can detect about a few thousands genes in each cell, more than half of the elements in gene expression vectors of cells are zeros. Also, for studies involving cells from same tissues or organs, they might have many genes with similar expression levels. This may lead to high correlations between cells and cause Seekmer to be insensitive to minor differences between cells. This issue is not present in the experiments shown in this paper. A potential reason for the absence might be that the 2-way clustering and raising the correlation coefficients to higher power alleviate the issue. In case the issue occurs, one solution is to use most variable genes instead of all genes to calculate the correlation matrix. By excluding genes that have low variation in their expression levels across all cells, the algorithm will be more sensitive to the differences between cells and thus more selective in pooling reads. Another limitation is that the current implementation of Seekmer is not scalable to handle cases like multiple thousand cells. Inference for more than a few thousands of cells might take days. This is currently acceptable as major full-length single-cell sequencing platforms accept at most a few hundred cells per batch. At the same time, this can be easily improved in algorithm implementation and future hardware upgrade. For now, we do not apply pre-mature optimizaiton.

4.5 Conclusion

This work presents Seekmer, an imputation method that estimates single-cell transcript abundance through read pooling. The weighted pooling approach enables estimating the abundance of the

transcripts per cell much more accurately. The algorithm is able to differentiate splicing isoforms from same genes in the imputation process and performs well even for limited numbers of cells. In sum, Seekmer provides more accurate transcript profiling analysis and empowers researchers in single-cell splicing studies.

4.6 Figures and Tables

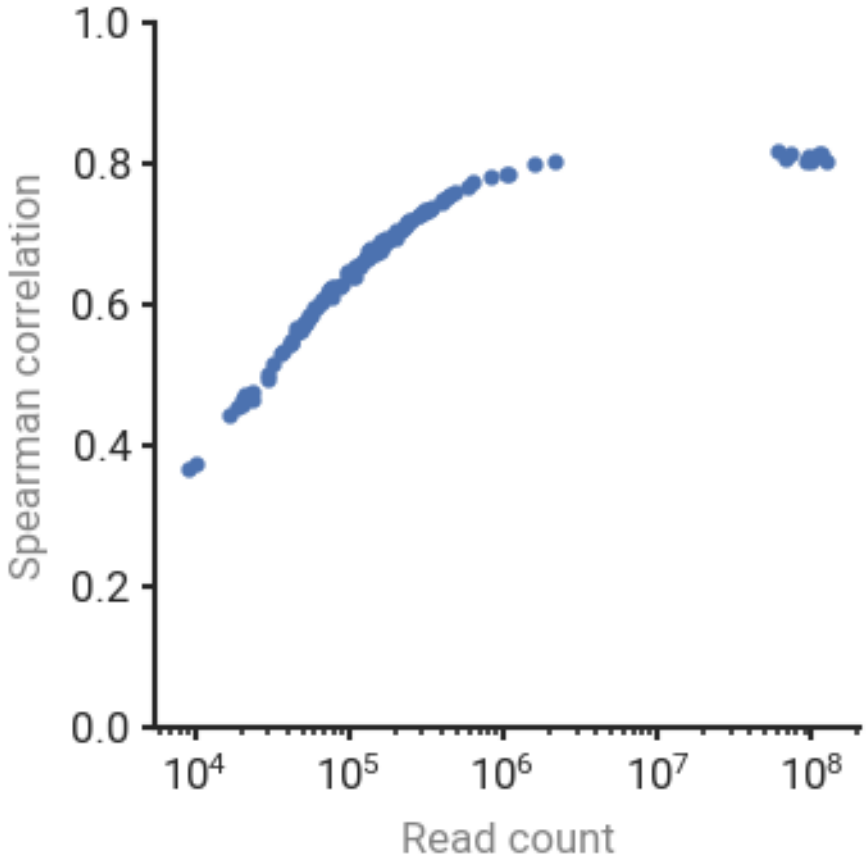


Figure 4.1: Low read counts compromises the accuracy of transcript abundance estimation.
A series of downsampling experiments based on the expression profiles of UHRR and HBRR reveals that the accuracy of the transcript abundance estimation drastically decreases when the numbers of reads drop below one million.

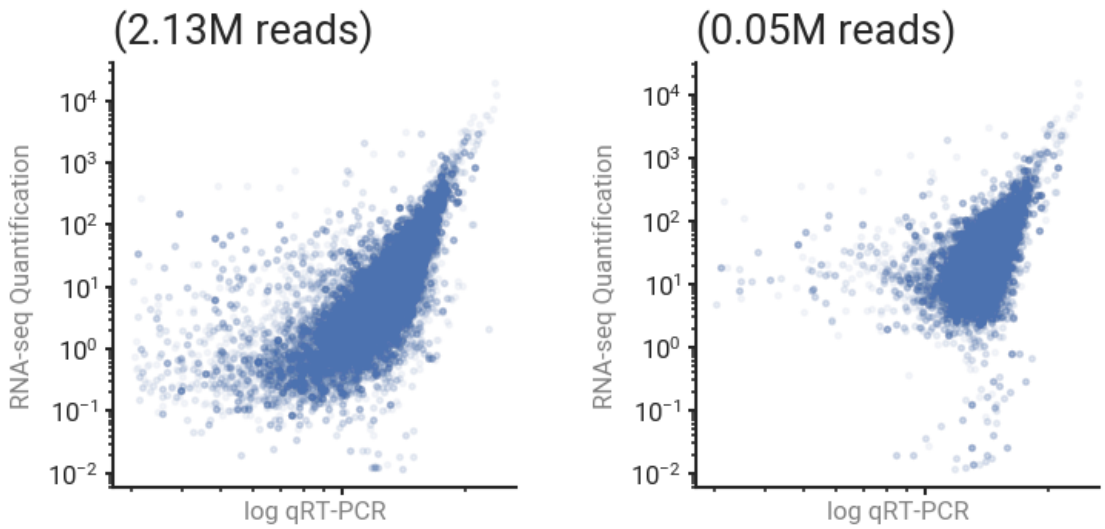


Figure 4.2: Low read counts hurts the estimation of transcripts of intermediate expression. Comparison of estimated abundance of transcripts from RNA sequencing against the qRT-PCR results on two cells of different read counts shows the inaccurate estimation of transcripts of intermediate expression levels. With less than 10 million reads, transcripts with expression levels less than 10 TPM are either not observed or wrongly quantified.

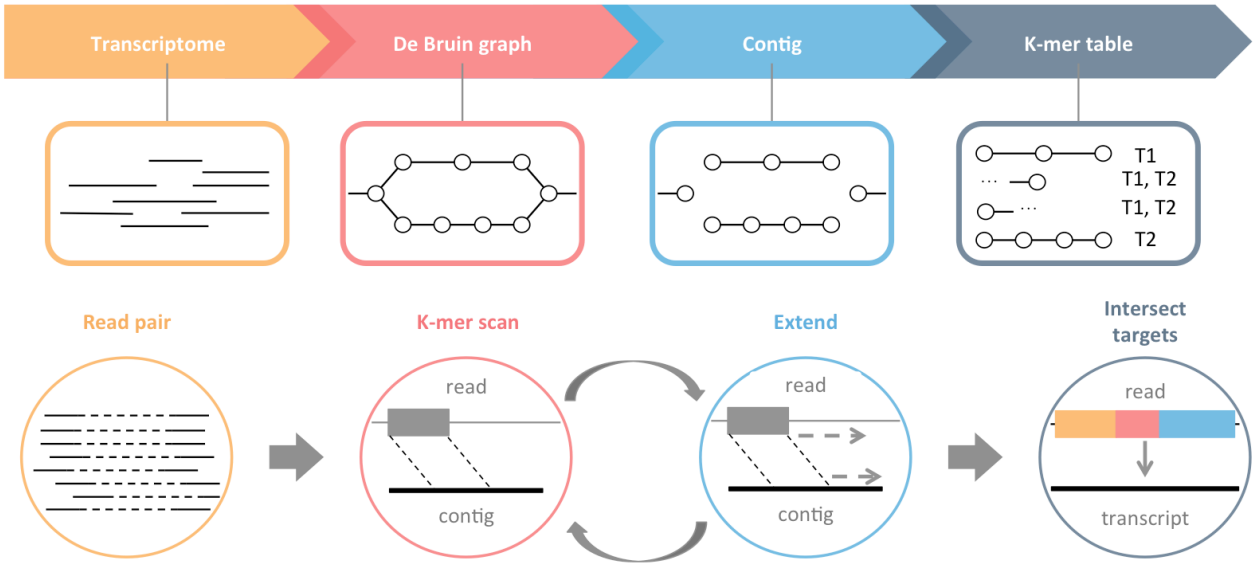


Figure 4.3: Seekmer transcript quantification approach.
Seekmer first collects all possible K-mers from the reference transcriptome sequences and associates them to their target transcripts. It connects adjacent K-mers together to form de Bruijn graph and dissects the graph into contigs of K-mers that share same sets of target transcripts. The contigs of K-mers with their targets are used as the index of the transcriptomic sequences. Given RNA sequencing reads, Seekmer looks for a K-mer that matches the index. Seekmer then extends along the reference contig from the K-mer and maps as many contigs as possible. If there is no exact match when extending K-mer search, Seekmer performs a heuristic local alignment based on SIFT4 algorithm to correct potential sequencing errors or SNPs. Seekmer then assign the reads to the intersection of all transcript sets obtained from mapped contigs. Finally, Seekmer performs an expectation-maximization to quantify the expression of transcripts based on all mapped reads.

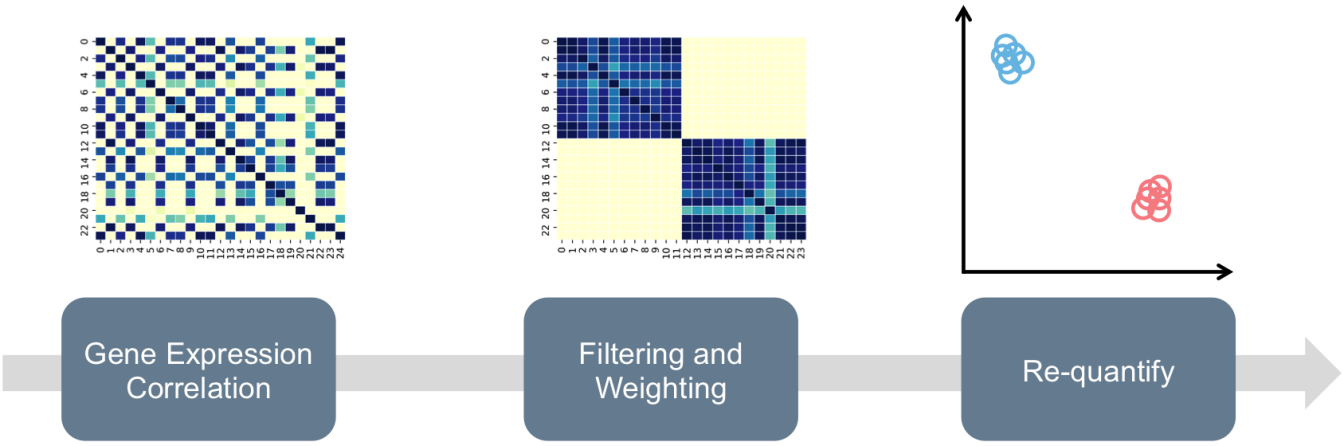


Figure 4.4: Seekmer read pooling approach.
Seekmer calculates an initial gene expression profile by direct mapping and quantification over sequencing data from individual cells. Then for all pairs of cells, Seekmer calculates the correlation of their gene expression levels. By clustering the correlation scores, Seekmer identifies cells that are highly correlated. Then Seekmer pools reads from correlated cells with weights that are simply powered correlation scores. From pooled reads, Seekmer re-estimates the transcript expression levels.

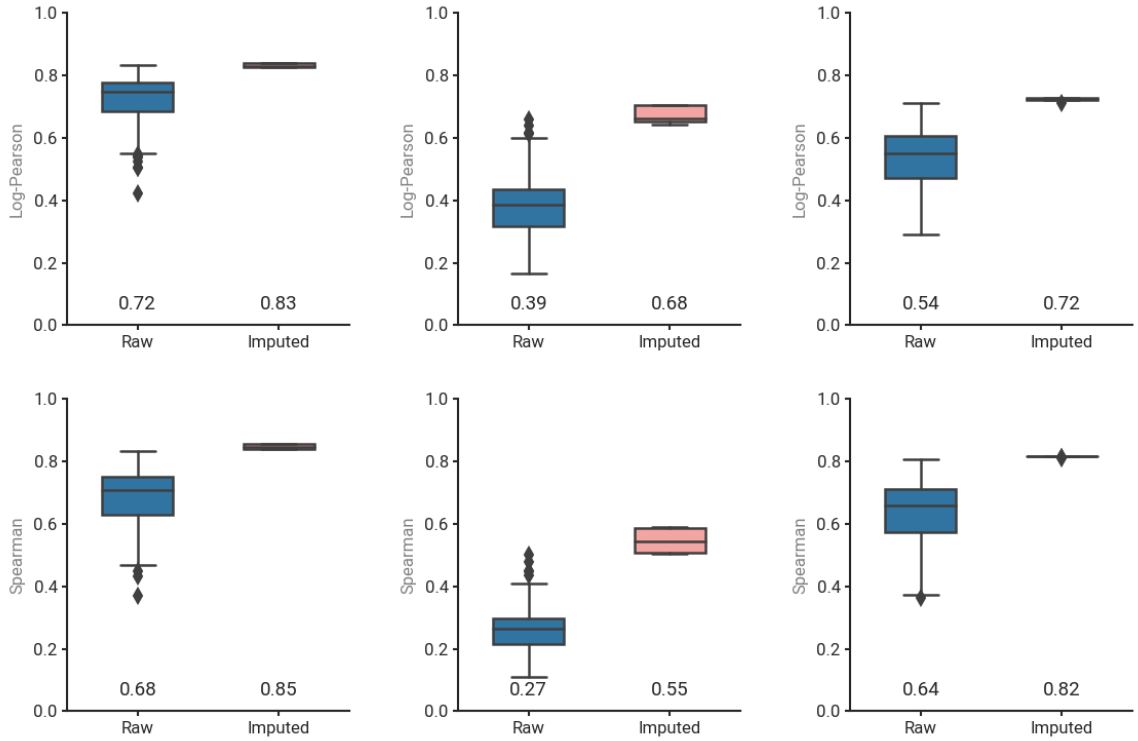


Figure 4.5: Seekmer performs well on simulated single cell RNA sequencing data.
Seekmer boosted the accuracy of transcript abundance inference by read-pooling imputation over a series of simulated single-cell RNA sequencing data. The performance is evaluated against the gene (left) and transcript (middle) expression levels used in the simulation and the qRT-PCR quantification results (right). The direct quantification and imputation results are evaluated in terms of Log-Pearson correlation coefficient (in the upper panel) and Spearman correlation coefficient (in the lower panel).

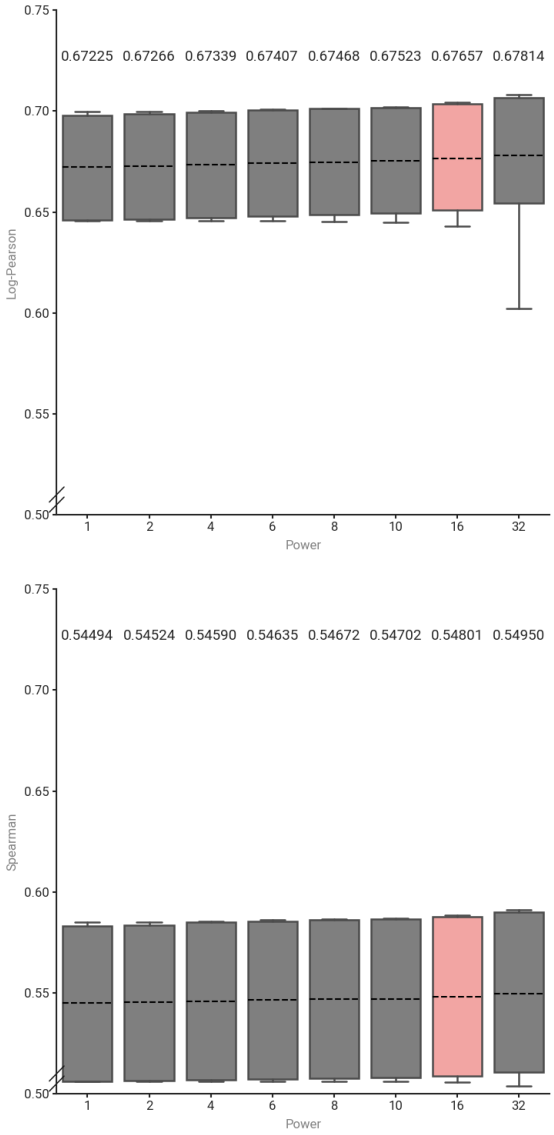


Figure 4.6: Seekmer is not sensitive to the weighting power parameter.
Seekmer's imputation accuracy is not very sensitive to the weighting power parameter. The default power of 16 gives the best performance among all tested parameter values in terms of Log-Pearson correlation coefficient (in the upper panel) and Spearman correlation coefficient (in the lower panel). The difference, however, is marginal.

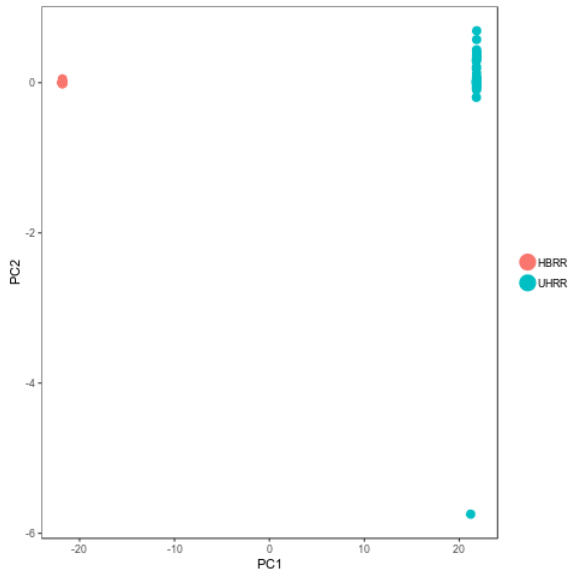


Figure 4.7: Seekmer greatly reduces the expression variation due to down sampling.
Seurat clustering of the simulated single-cell data shows a clear clustering of two types of cells, each simulated from a bulk RNA sequencing sample. The principal component analysis over the most variable genes shows clear clustering of cells using the direct abundance estimation. The same analysis using the imputed expression levels shows much tighter clustering.

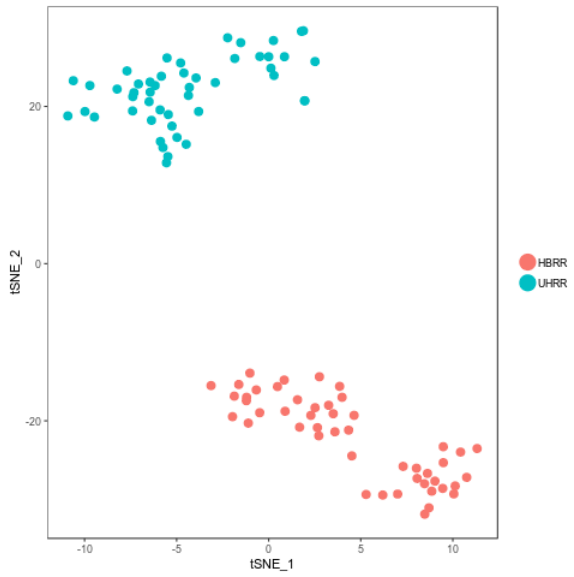


Figure 4.8: Seekmer preserves the cluster structure in the simulated data.
Seurat clustering of the simulated single-cell data shows a clear clustering of two types of cells, each simulated from a bulk RNA sequencing sample. The t-SNE analysis over the most variable genes shows clear clustering of cells using both the direct abundance estimation and the imputed expression levels, with similar cluster shapes.

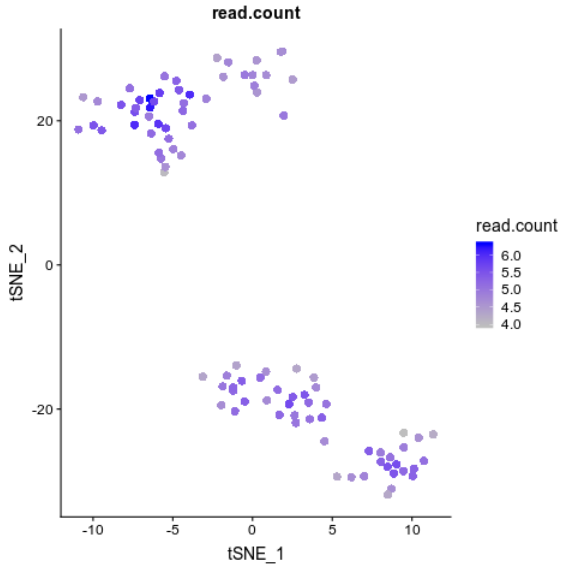
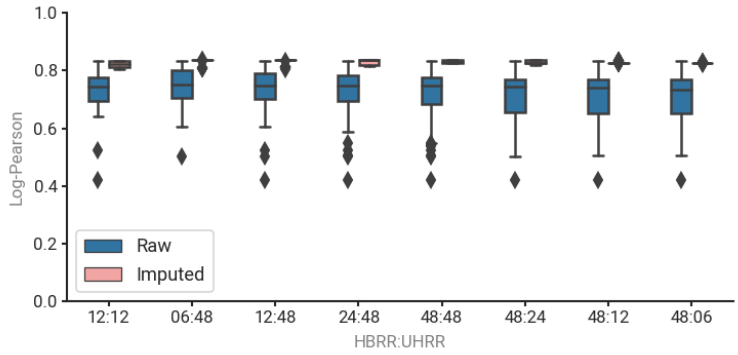
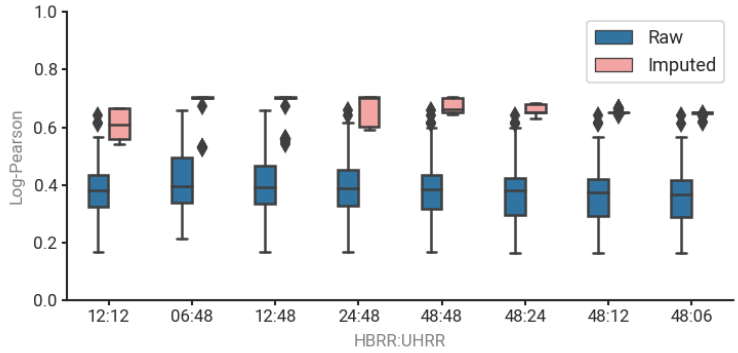


Figure 4.9: Seekmer removes biases of read counts from clustering.
Seurat clustering over the direct abundance estimation shows a strong artifact introduced by read count imbalance. Poorly sequenced cells often occupy the boundaries of the clusters. Such an effect is absent in the clustering over the imputed expression levels.

Gene-level



Transcript-level



qRT-PCR

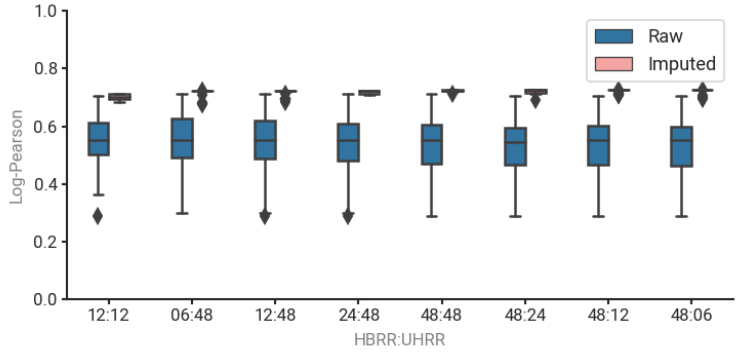
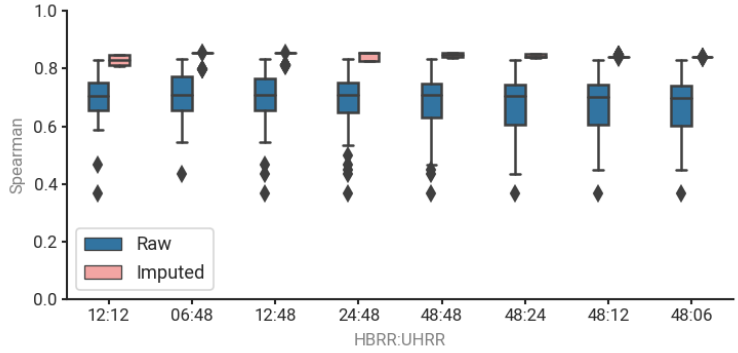
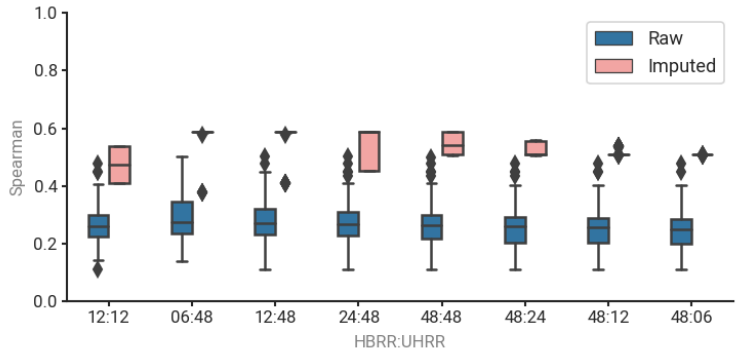


Figure 4.10: Seekmer is not sensitive to cell counts (Log-Pearson scores).
The accuracy of Seekmer's imputation does not fluctuate over varying numbers of cells in the clusters. The performance is evaluated against the gene (top) and transcript (middle) expression levels used in the simulation and the qRT-PCR quantification results (bottom) in terms of Log-Pearson correlation coefficient.

Gene-level



Transcript-level



qRT-PCR

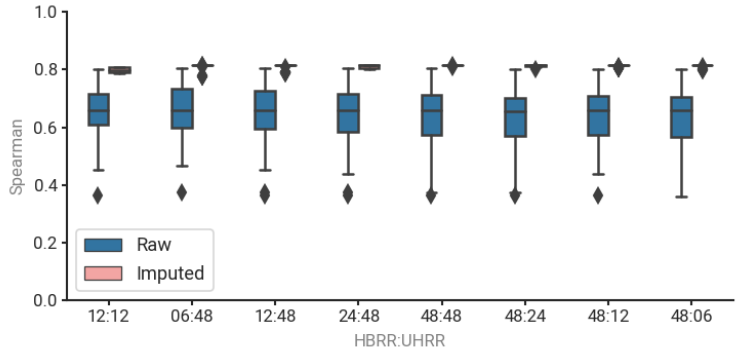
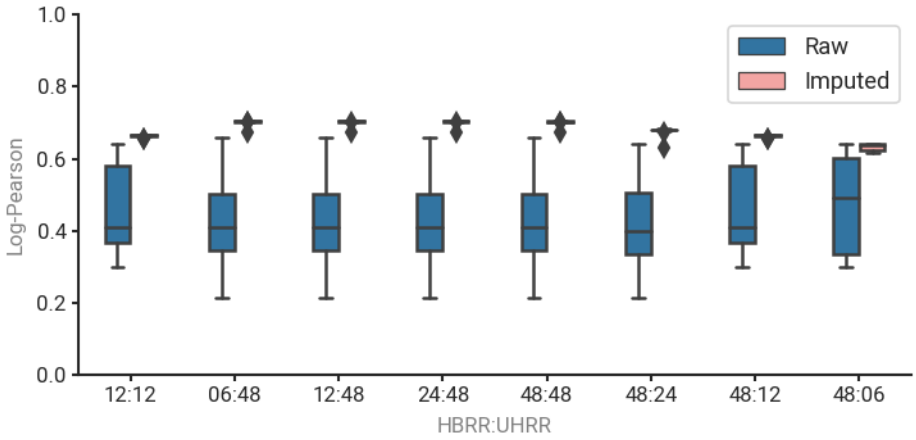


Figure 4.11: Seekmer is not sensitive to cell counts (Spearman scores).
The accuracy of Seekmer's imputation does not fluctuate over varying numbers of cells in the clusters. The performance is evaluated against the gene (top) and transcript (middle) expression levels used in the simulation and the qRT-PCR quantification results (bottom) in terms of Spearman correlation coefficient.

Log-pearson



Spearman

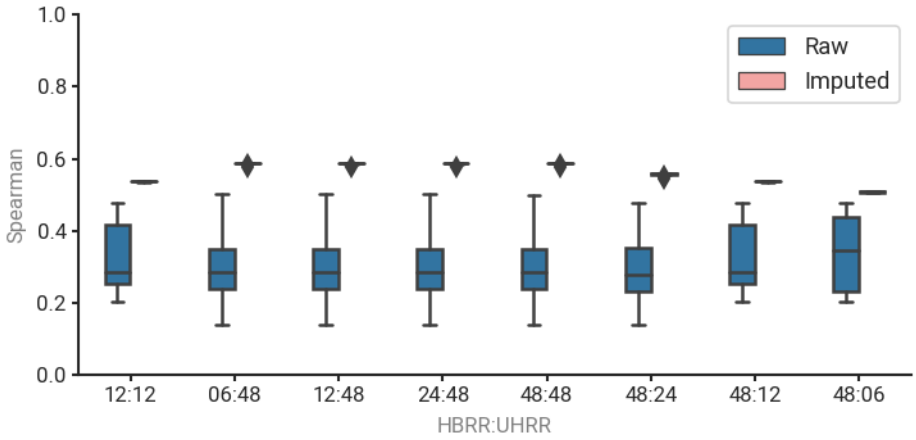
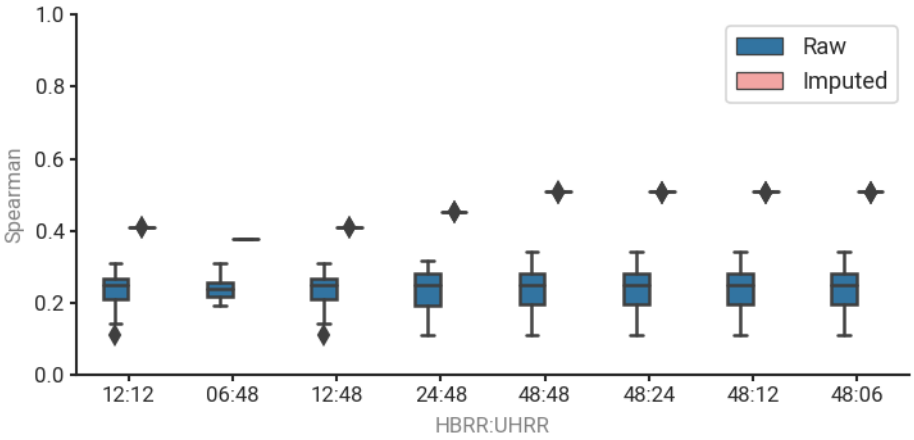


Figure 4.12: Seekmer can impute small cell clusters accurately (UHRR cells).
The X-axis is the number of cells in all groups. Even for cell clusters as small as 6 cells, Seekmer can still achieve Log-pearson correlation coefficients as high as 0.5.

Log-pearson



Spearman

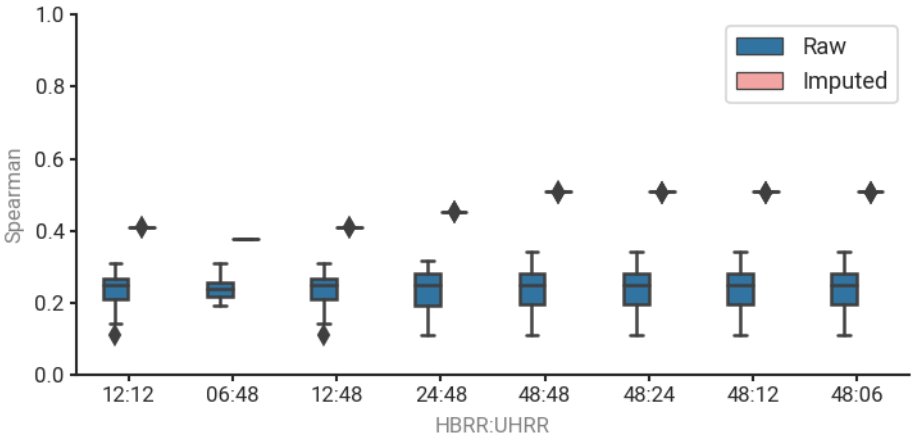


Figure 4.13: Seekmer can impute small cell clusters accurately (HBRR cells).
The X-axis is the number of cells in all groups. Even for cell clusters as small as 6 cells, Seekmer can still achieve Log-pearson correlation coefficients as high as 0.4.

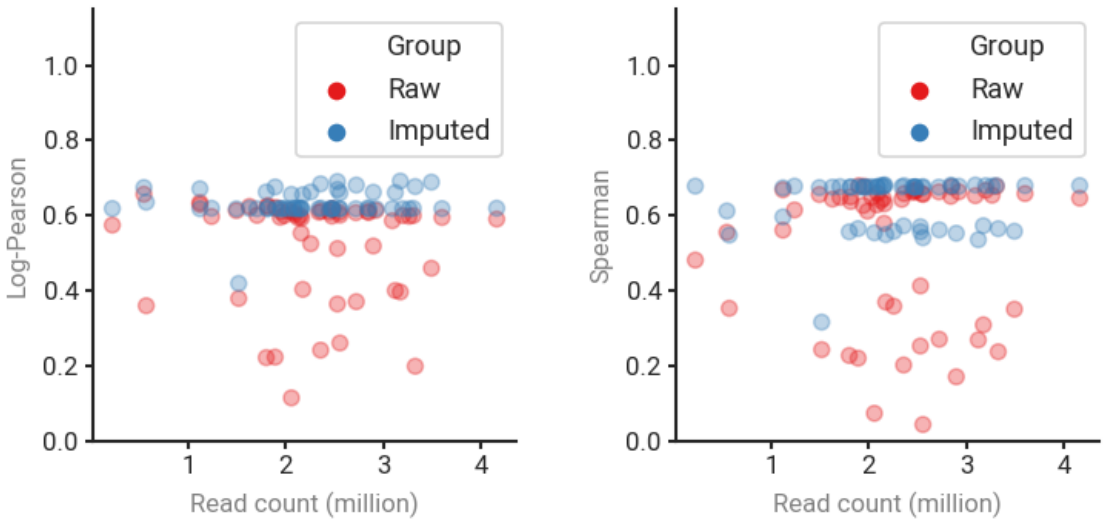


Figure 4.14: Seekmer imputation improves the quantification of SIRV spike-in sequences.
The performance is measured in terms of Log-Pearson (left) and Spearman (right) correlation coefficients.

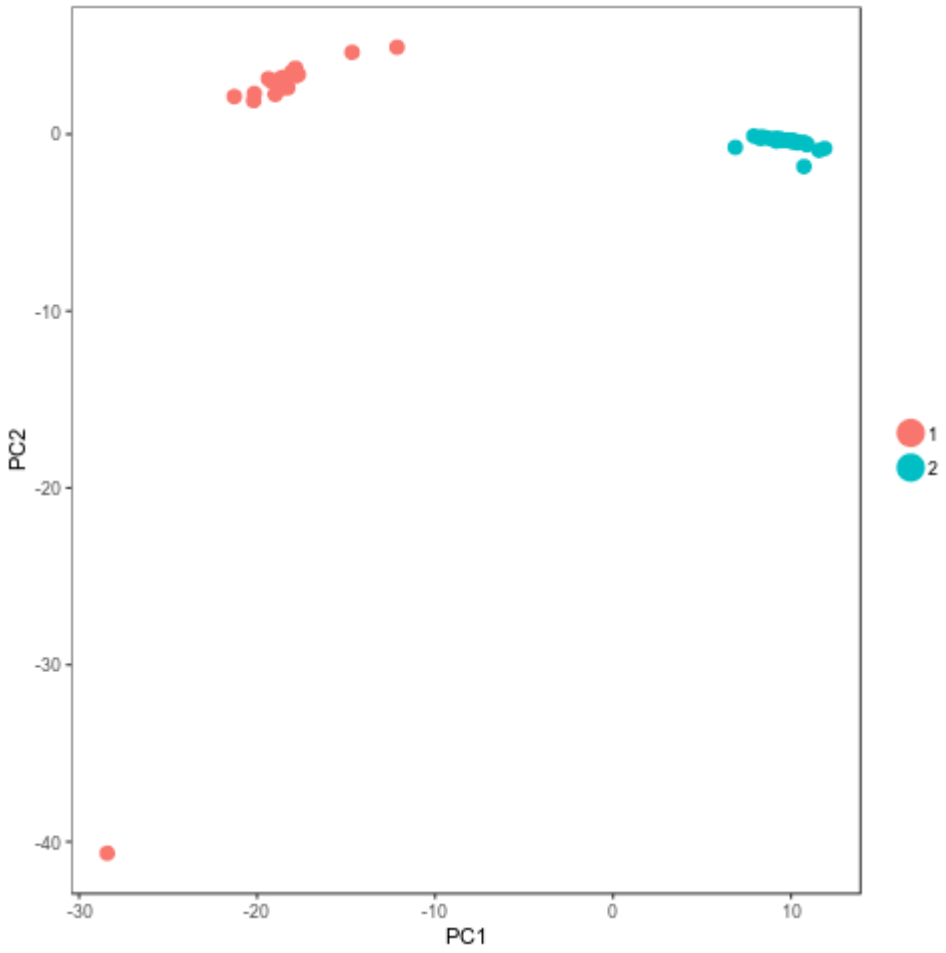


Figure 4.15: Seekmer preserved intra-cluster variation.
Seekmer still preserved the intra-cluster variation in the real-world spike-in data, unlike what was seen in simulated data.

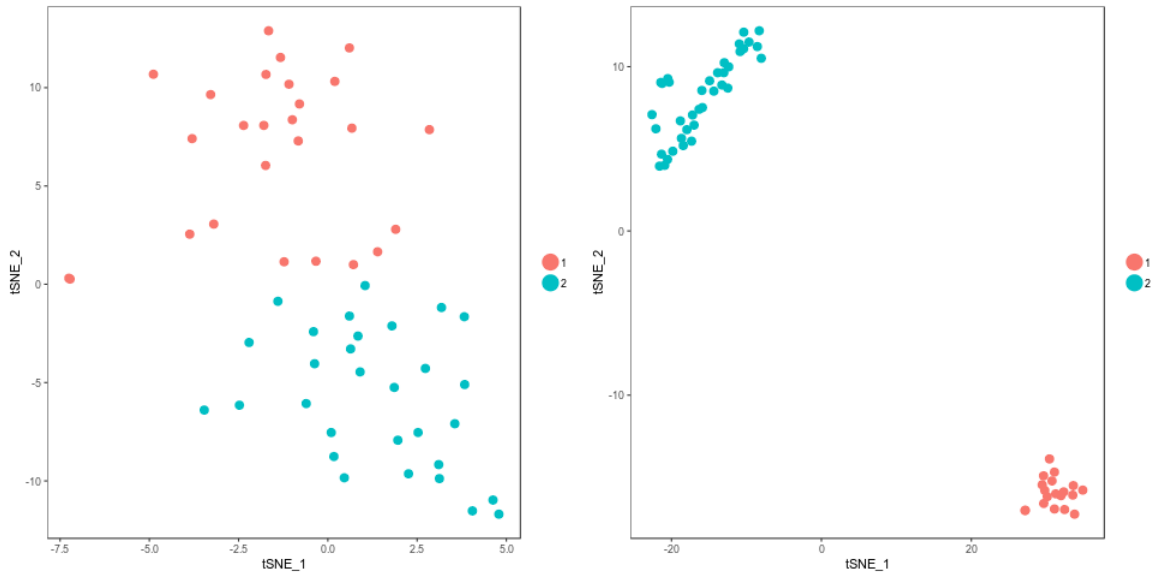


Figure 4.16: Seekmer preserved cluster structure.
Seekmer still preserved the cluster structure seen in the direct quantification data (left). The t-SNE of the imputed data (right) does show larger separation between the cluster.

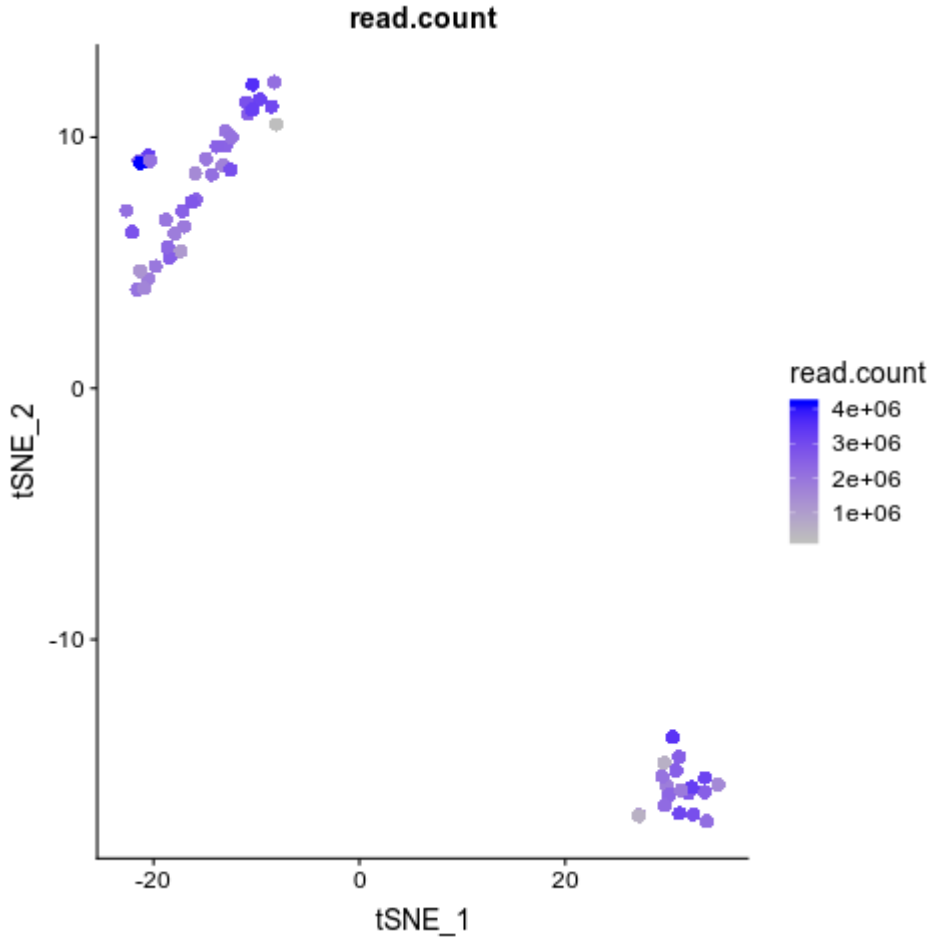


Figure 4.17: Seekmer clustering is not affect by imbalanced read counts.
The t-SNE clustering of Seekmer imputation data over the spike-in cells does not show artifacts caused by imbalanced read counts.

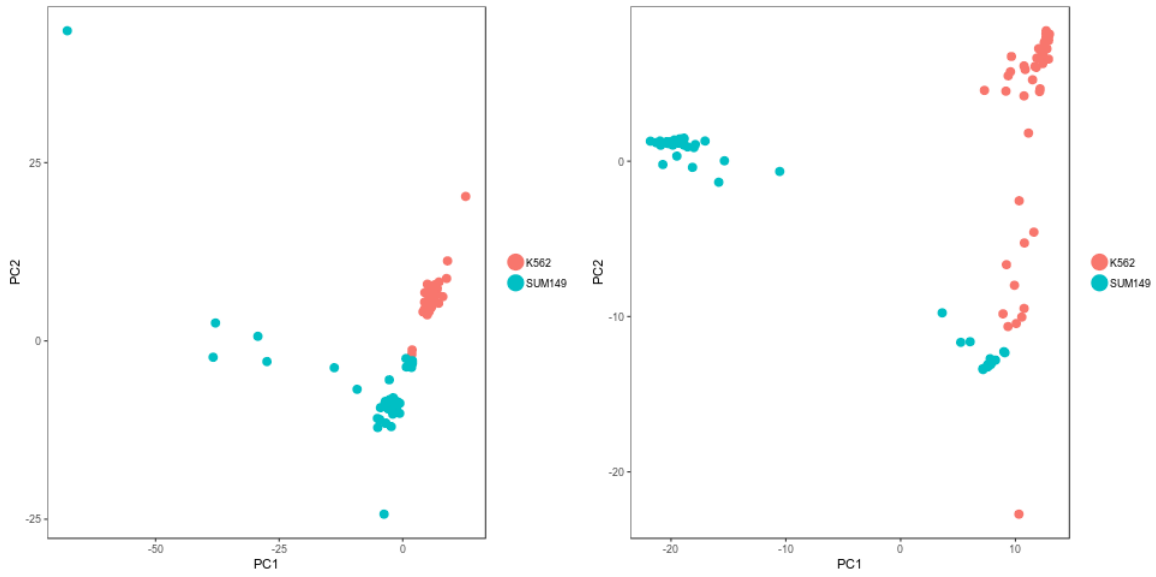


Figure 4.18: Clusters of Fluidigm Polaris data in both raw quantification (left) and imputed results (right) are similar.

Similar to the spike-in data, Seekmer preserved the clustering patterns in its imputation results.

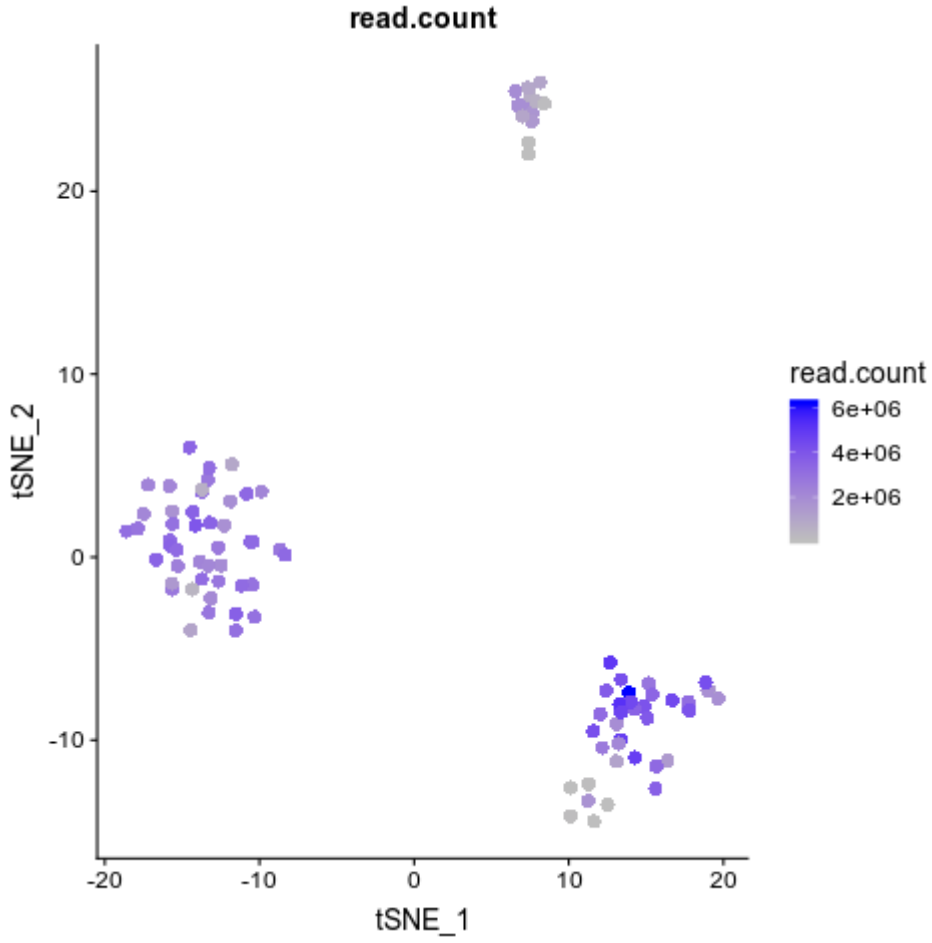


Figure 4.19: Pooly sequenced cells in Fluidigm Polaris data forms a separate cluster.

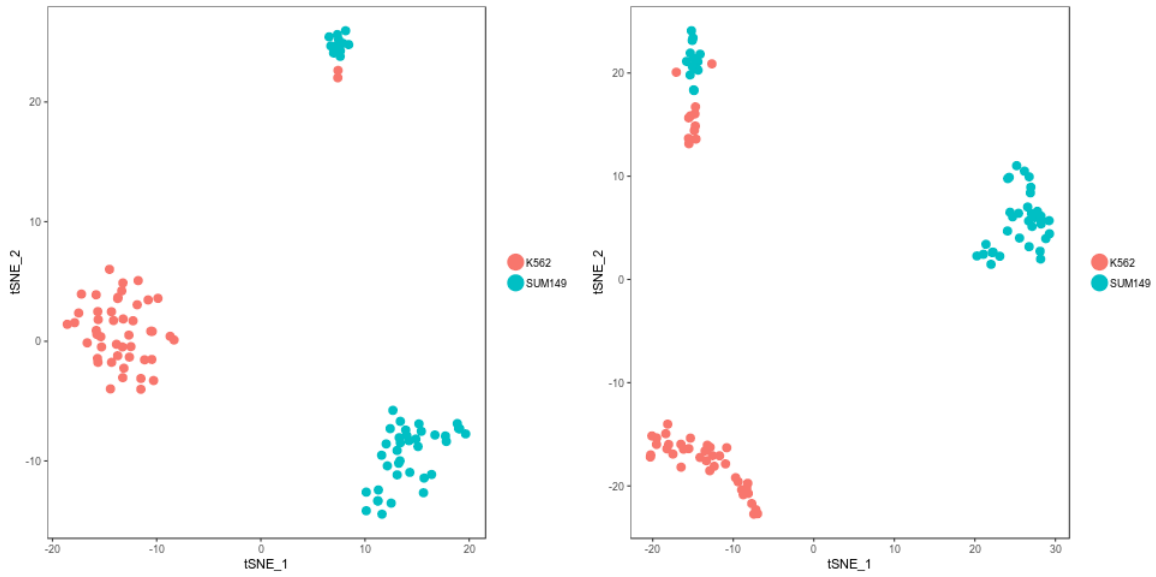


Figure 4.20: The t-SNE analysis of the Seekmer imputation over the Fluidigm Polaris dataset. Similar to the spike-in data, Seekmer preserved the clustering patterns in its imputation results. The left panel shows the direct quantification data, while the right shows the imputed results.

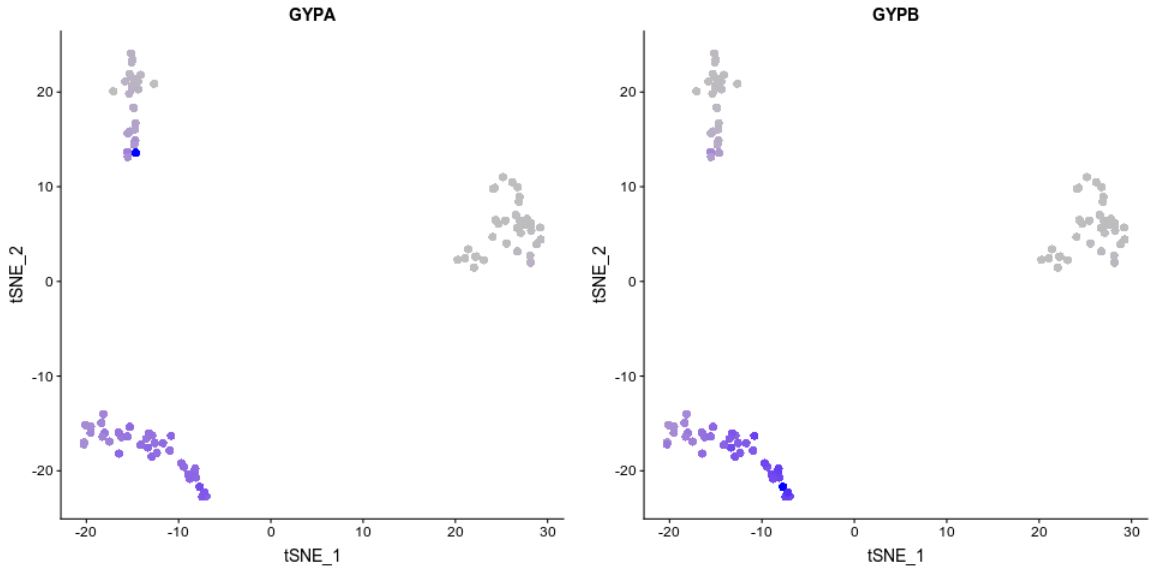


Figure 4.21: GYPA and GYPB genes are well expressed in K562 cells.

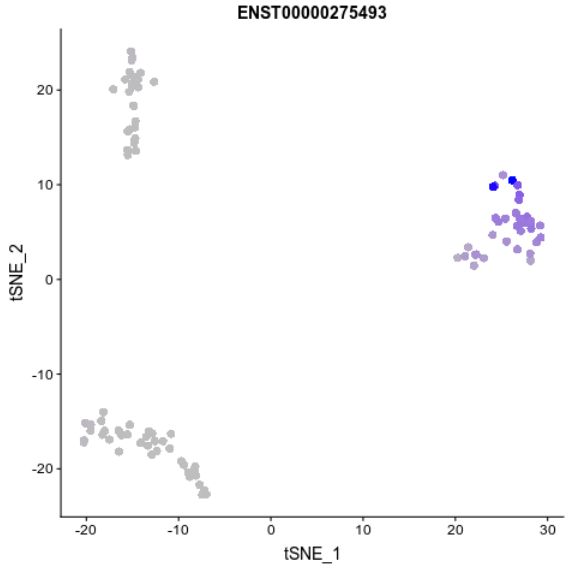


Figure 4.22: The SUM149 cells express the canonical transcript of EGFR (EGFR-201).

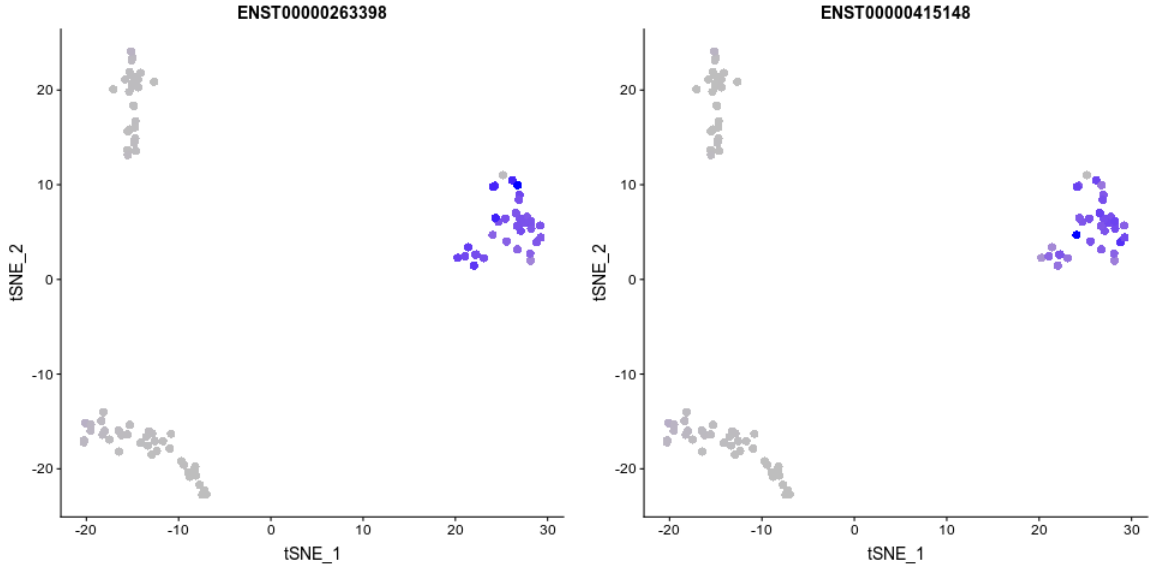


Figure 4.23: The SUM149 cells express both the canonical transcript of CD44 and the transcript with v6 exon.

The left is CD44-201, and the right is CD44-206.

4.7 Reference

1. Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M. & Stamm, S. Function of alternative splicing. *Gene* **514,** 1–30. ISSN: 0378-1119 (Feb. 2013).

2. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. En. *Experimental & Molecular Medicine* **50,** 96. ISSN: 2092-6413 (Aug. 2018).

3. Wang, Y. & Navin, N. E. Advances and Applications of Single Cell Sequencing Technologies. *Molecular cell* **58,** 598–609. ISSN: 1097-2765 (May 2015).

4. Rizzetto, S., Eltahla, A. A., Lin, P., Bull, R., Lloyd, A. R., Ho, J. W. K., Venturi, V. & Luciani, F. Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. *Scientific Reports* **7.** ISSN: 2045-2322. doi:10.1038/s41598-017-12989-x. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5630586/> (2019) (Oct. 2017).

5. Zhang, C., Zhang, B., Lin, L.-L. & Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18,** 583. ISSN: 1471-2164 (Aug. 2017).

6. Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A. B., Sloan, S. A., Luo, W., Fedrigo, O., Ross, M. E. & Tilgner, H. U. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. en. *Nature Biotechnology* **36,** 1197–1202. ISSN: 1546-1696 (Dec. 2018).

7. AlJanahi, A. A., Danielsen, M. & Dunbar, C. E. An Introduction to the Analysis of Single-Cell RNA-Sequencing Data. *Molecular Therapy. Methods & Clinical Development* **10,** 189–196. ISSN: 2329-0501 (Aug. 2018).

8. Qiu, P. Embracing the dropouts in single-cell RNA-seq data. en. *bioRxiv,* 468025 (Nov. 2018).

9. Zhang, L. & Zhang, S. Comparison of computational methods for imputing single-cell RNA-sequencing data. eng. *IEEE/ACM transactions on computational biology and bioinformatics.* ISSN: 1557-9964. doi:10.1109/TCBB.2018.2848633 (June 2018).

10. Moussa, M. & Măndoiu, I. I. Locality Sensitive Imputation for Single Cell RNA-Seq Data. eng. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology.* ISSN: 1557-8666. doi:10.1089/cmb.2018.0236 (Feb. 2019).

11. Mongia, A., Sengupta, D. & Majumdar, A. McImpute: Matrix Completion Based Imputation for Single Cell RNA-seq Data. English. *Frontiers in Genetics* **10.** ISSN: 1664-8021. doi:10.3389/fgene.2019.00009. <https://www.frontiersin.org/articles/10.3389/fgene.2019.00009/full> (2019) (2019).

12. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. En. *Nature Communications* **9,** 997. ISSN: 2041-1723 (Mar. 2018).

13. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. eng. *BMC bioinformatics* **12,** 323. ISSN: 1471-2105 (Aug. 2011).

14. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. en. *Bioinformatics* **29,** 15–21. ISSN: 1367-4803 (Jan. 2013).

15. Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A. & Flicek, P. Ensembl 2018. eng. *Nucleic Acids Research* **46,** D754–D761. ISSN: 1362-4962 (Jan. 2018).

16. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. eng. *Nature Biotechnology* **33,** 495–502. ISSN: 1546-1696 (May 2015).

17. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. en. *Nature Biotechnology* **34,** 525–527. ISSN: 1546-1696 (May 2016).

18. Chen, W., Li, Y., Easton, J., Finkelstein, D., Wu, G. & Chen, X. UMI-count modeling and differential expression analysis for single-cell RNA sequencing. eng. *Genome Biology* **19,** 70. ISSN: 1474-760X (2018).

19. Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S. & Sandberg, R. Full-length RNA-seq from single cells using Smart-seq2. en. *Nature Protocols* **9,** 171–181. ISSN: 1750-2799 (Jan. 2014).

20. Stupnikov, A., O'Reilly, P. G., McInerney, C. E., Roddy, A. C., Dunne, P. D., Gilmore, A., Ellis, H. P., Flannery, T., Healy, E., McIntosh, S. A., Savage, K., Kurian, K. M., Emmert-Streib, F., Prise, K. M., Salto-Tellez, M. & McArt, D. G. Impact of Variable RNA-Sequencing Depth on Gene Expression Signatures and Target Compound Robustness: Case Study Examining Brain Tumor (Glioma) Disease Progression. *JCO precision oncology* **2.** ISSN: 2473-4284. doi:10.1200/PO.18.00014. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6186166/> (2019) (Sept. 2018).

21. Williams, A. G., Thomas, S., Wyman, S. K. & Holloway, A. K. RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]* **83,** 11.13.1–11.13.20. ISSN: 1934-8266 (Oct. 2014).

22. Haas, B. J., Chin, M., Nusbaum, C., Birren, B. W. & Livny, J. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics* **13,** 734. ISSN: 1471-2164 (Dec. 2012).

23. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A. & McCarroll, S. A. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. English. *Cell* **161,** 1202–1214. ISSN: 0092-8674, 1097-4172 (May 2015).

24. Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J. & Bielas, J. H. Massively parallel digital transcriptional profiling of single cells. en. *Nature Communications* **8,** 14049. ISSN: 2041-1723 (Jan. 2017).

25. Arzalluz-Luque, Á. & Conesa, A. Single-cell RNAseq for the study of isoforms—how is that possible? *Genome Biology* **19.** ISSN: 1474-7596. doi:10.1186/s13059-018-1496-z. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6085759/> (2019) (Aug. 2018).

26. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. & Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. en. *Nature Biotechnology* **28,** 511–515. ISSN: 1546-1696 (May 2010).

27. Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H. & van Oudenaarden, A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. en. *Nature* **525,** 251–255. ISSN: 1476-4687 (Sept. 2015).

28. Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R. & Hemberg, M. SC3: consensus clustering of single-cell RNA-seq data. eng. *Nature Methods* **14,** 483–486. ISSN: 1548-7105 (May 2017).

Chapter 5

Conclusion

The advances of experimental techniques have exposed us to more diverse and heterogeneous biological phenomena. To understand the mechanisms, researchers have deployed numerous computational techniques, from parsimony to likelihood models, from statistical learning methods to deep neural networks, to summarize the phenomena of interest. Each time a category of wrongly predicted phenomena is determined, a more complex model is proposed to replace its predecessor. However, while the capability of the models to capture more intricate relationships increases, the interpretability of the models often declines. Having powerful computational models to help study the mechanisms behind the biological heterogeneity is an everlasting challenge.

In this work, we present the development of three computational models to study biological heterogeneity of different scales. Each of the model well demonstrated their capability in capturing important factors that are often missed in the traditional analysis.

Chapter 2 described a deep neural network approach to model heterogeneity in molecular evo-

lution. The purpose of using a deep neural network model is to allow incorporation of more terms for branch-wise and site-wise heterogeneity without incurring computational burden. The model provides a powerful fitting function that approximates the conditional probability distribution of the quartet tree topologies given the observed sequences. We showed that the model has a great performance in some of the tough cases such as long-branch attractions.

The model, however, is not easily extensible to model more than four species. The formulation of this network has a fixed dimension for the input data. This limitation is shared by many early deep neural network [1]. In our model, the dimension restriction over the sequence length is alleviated by the adaptive average pooling design. Yet the number of species is still limited. There is no meaningful ordering of the species in the model for more than four input species. The output of the model is not extensible as well. The number of possible topologies of phylogenetic trees grows super-exponentially over the number of leaf taxons. The current output layer does not work.

The same requirement over the variable input data dimension is also seen in tasks such as protein contact prediction [2]. In such tasks, the feature dimention of individual amino acids becomes the channel. The other dimension becomes variable. Such design can be borrowed similarly to the phylogenetic inference. At the same time, traditional formulation of structured prediction is often described as a classification problem, in which the predictor discriminate the prediction value associated with the correct structure against that of the highest-scored wrong structure [3]. In this case, the tree can be expressed as a distance matrix which can be reliably converted to a unrooted tree topology. The classification problem can then be formulated to

discriminate the matrices.

The training data simulator determines what goes into the deep neural network model. With much less computational restriction of the inference process, there are many new parameters can be further incorporated into the model. Since it is easy to express gaps and insertions in the input encoded sequences, it is now possible to incorporate models for insertions and deletions [4, 5]. Since the convolutional layers can examine neighbouring sites simultaneously, incorporating site-dependencies may also improve the power of the model [6]. Incorporating terms to more realistically describe the evolution process and choosing proper parameter values in the simulator would allow the model to infer real-world phylogeny more accurately.

Chapter 3 described a heuristic inference methods for resolving tumour heterogeneity. The hinting strategy of the model allows fast clustering of the mutations without sacrificing accuracy. In contrast to methods that over-specify the clusters, the conservative inference strategy allows more accurate clustering of the mutations. Specifically, FastClone preserves the co-clustering relationship of the mutations. The program won the DREAM SMC-Het Challenge.

However, there is a limit for heterogeneous inference with only somatic mutations and copy number profiles. The mutation calling is prone to error [7]. The issues with mutation calling programs may even affect the clinical interpretation of the subclonal inference [8]. Also, due to the nature of the clustering, low-frequency somatic mutations might always clustered together. These somatic mutations have been repeatedly reported to be clinically relavent, especially in their contributions to prognosis [9, 10]. Placing these mutations into the right place would be difficult. To further improve the subclonal inference, integration of data from orthogonal techniques

might be necessary.

Chapter 4 described an imputation method for estimating single-cell transcript abundance. The read-pooling approach to impute the single-cell transcripts not only recovers the expression levels of genes and transcripts that are not sampled, but also helps further refine the transcript deconvolution results. The model preserves the clustering structures of the samples and reduces the variation introduced by the sampling process. The simulation test and the real-world data analysis have demonstrated the accuracy of the imputation results.

The imputation results, however, lacks statistical meaning in certain downstream analysis. Common analysis in single-cell transcriptomic studies involves differential expression analysis and marker identification [11, 12]. These analysis tools often share similar traits with their bulk analysis counterparts, such as read count modelling [13–15]. They are not designed to work with imputed data directly, because the imputed read counts and expression levels do not carry the same statistical meaning as the direct observed one. They should not be used in such analysis. Differential expression analysis over imputed expression data at the transcript level needs additional modifications to the existing models.

All three tasks involves resolving biological signals from a noisy data which can be attributed to both biological and irrelavent factors. All three methods involve elements that are not statistically sound, such as the deep neural network itself, the density-hinted optimization, and the weighting strategy in single-cell imputation. It is a well known problem in the field of optimization algorithms that rarely conditions for theoretical guarantees are met in real-world problems [16]. Even so, some methods are way too time-consuming to achieve the marginal benefit of

the optimal solution [17]. Also, the gaps between the real biological scenario and the statistical models often render the mathematical optimal solutions suboptimal or even undesirable for biological interpretaion of the real-world data [18, 19]. Therefore, from the aspect of algorithm design, choosing effective approximation or heuristics with proper designed validation is more pragmatic and relavent in biological data mining.

The advances of biological assays have provided voluminous data for more complicated computational modelling techniques. Methods such as deep neural network can accurately approximate the nonlinear impact of factors on the phenomena of interest. Yet the current approach to interpret the mechanisms through direct examination of the model formulation have become less relavant in the era of the new modelling regime. Therefore, improvements in both interpretation and interpretable modelling are emergent.

5.1 Reference

1. He, K., Zhang, X., Ren, S. & Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. en. *arXiv:1406.4729 [cs]* **8691.** arXiv: 1406.4729, 346–361 (2014).

2. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. en. *PLOS Computational Biology* **13,** e1005324. ISSN: 1553-7358 (Jan. 2017).

3. Tsochantaridis, I., Joachims, T., Hofmann, T. & Altun, Y. Large Margin Methods for Structured and Interdependent Output Variables. *Journal of Machine Learning Research* **6,** 1453–1484. ISSN: ISSN 1533-7928 (2005).

4. Rivas, E. Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC Bioinformatics* **6,** 63. ISSN: 1471-2105 (Mar. 2005).

5. Warnow, T. Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. *PLoS Currents* **4.** ISSN: 2157-3999. doi:10.1371/currents.RRN1308. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3299439/> (2019) (Mar. 2012).

6. Fernandes, A. D. & Atchley, W. R. Site-specific evolutionary rates in proteins are better modeled as non-independent and strictly relative. eng. *Bioinformatics (Oxford, England)* **24,** 2177–2183. ISSN: 1367-4811 (Oct. 2008).

7. Ewing, A. D., Houlahan, K. E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T. N., Bare, J. C., P'ng, C., Waggott, D., Sabelnykova, V. Y., ICGC-TCGA DREAM Somatic Mutation Calling Challenge participants, Kellen, M. R., Norman, T. C., Haussler, D., Friend, S. H., Stolovitzky, G., Margolin, A. A., Stuart, J. M. & Boutros, P. C. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. eng. *Nature Methods* **12,** 623–630. ISSN: 1548-7105 (July 2015).

8. Noorbakhsh, J., Kim, H., Namburi, S. & Chuang, J. H. Distribution-based measures of tumor heterogeneity are sensitive to mutation calling and lack strong clinical predictive power. En. *Scientific Reports* **8,** 11445. ISSN: 2045-2322 (July 2018).

9. Zhao, X., Little, P., Hoyle, A. P., Pegna, G. J., Hayward, M. C., Ivanova, A., Parker, J. S., Marron, D. L., Soloway, M. G., Jo, H., Salazar, A. H., Papakonstantinou, M. P., Bouchard, D. M., Jefferys, S. R., Hoadley, K. A., Ollila, D. W., Frank, J. S., Thomas, N. E., Googe, P. B., Ezzell, A. J., Collichio, F. A., Lee, C. B., Earp, H. S., Sharpless, N. E., Hugo, W., Wilmott, J. S., Quek, C., Waddell, N., Johansson, P. A., Thompson, J. F., Hayward, N. K., Mann, G. J., Lo, R. S., Johnson, D. B., Scolyer, R. A., Hayes, D. N. & Moschos, S. J. The Prognostic Significance of Low-Frequency Somatic Mutations in Metastatic Cutaneous Melanoma. eng. *Frontiers in Oncology* **8,** 584. ISSN: 2234-943X (2018).

10. Klebanov, N., Artomov, M., Goggins, W. B., Daly, E., Daly, M. J. & Tsao, H. Burden of unique and low prevalence somatic mutations correlates with cancer survival. En. *Scientific Reports* **9,** 4848. ISSN: 2045-2322 (Mar. 2019).

11. Dal Molin, A., Baruzzo, G. & Di Camillo, B. Single-Cell RNA-Sequencing: Assessment of Differential Expression Analysis Methods. English. *Frontiers in Genetics* **8.** ISSN: 1664-8021. doi:10.3389/fgene.2017.00062. <https://www.frontiersin.org/articles/10.3389/fgene.2017.00062/full> (2019) (2017).

12. Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* **20,** 40. ISSN: 1471-2105 (Jan. 2019).

13. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. en. *Nature Methods* **11,** 740–742. ISSN: 1548-7105 (July 2014).

14. Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., Kaper, F., Fan, J.-B., Zhang, K., Chun, J. & Kharchenko, P. V. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. en. *Nature Methods* **13,** 241–244. ISSN: 1548-7105 (Mar. 2016).

15. Ntranos, V., Yi, L., Melsted, P. & Pachter, L. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. En. *Nature Methods* **16,** 163. ISSN: 1548-7105 (Feb. 2019).

16. Le, Q. V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B. & Ng, A. Y. *On Optimization Methods for Deep Learning* in *Proceedings of the 28th International Conference on International Conference on Machine Learning* event-place: Bellevue, Washington, USA (Omnipress, USA, 2011), 265–272. ISBN: 978-1-4503-0619-5. <`http://dl.acm.org/citation.cfm?id=3104482.3104516`> (2019).

17. Shapiro, A. & Homem-de-Mello, T. On the Rate of Convergence of Optimal Solutions of Monte Carlo Approximations of Stochastic Programs. *SIAM Journal on Optimization* **11,** 70–86. ISSN: 1052-6234 (Jan. 2000).

18. Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y. & DePristo, M. A. A universal SNP and small-indel variant caller using deep neural networks. en. *Nature Biotechnology* **36,** 983–987. ISSN: 1546-1696 (Oct. 2018).

19. Du Plessis, L., Leventhal, G. E. & Bonhoeffer, S. How Good Are Statistical Models at Approximating Complex Fitness Landscapes? *Molecular Biology and Evolution* **33,** 2454–2468. ISSN: 0737-4038 (Sept. 2016).