Somatic mutation analysis for the study of clonal evolution in cancer

by

Fatemeh Dorri

MSc, University of Waterloo, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Computer Science)

The University of British Columbia
(Vancouver)

April 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Somatic mutation analysis for the study of clonal evolution in cancer

submitted by **Fatemeh Dorri** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Computer Science**.

Examining Committee:

Dr. Sohrab Shah, Department of Pathology and Laboratory Medicine, Department of Molecular Oncology
Co-Supervisor

Dr. Anne Condon, Department of Computer Science
Co-Supervisor

Dr. Alexandre Bouchard-Côté, Department of Statistics
Supervisory Committee Member

Dr. Gabriela V. Cohen Freue, Department of Statistics
University Examiner

Dr. Sara Mostafavi, Department of Statistics
University Examiner

Abstract

Next generation sequencing (NGS) technology provides researchers an opportunity to study cancer genomes at different resolutions. In particular, detection and interpretation of the smallest somatic changes of the genome (single nucleotide variants) are now tractable at scale. However, significant challenges in the analysis of both bulk tumour and single cell sequencing methods remain to fully exploit the advance in technology development. Two emerging areas of applying sequencing technology to better ascertain properties of cancer evolution are (i) sequencing multiple tumour biopsies from the same patient, and (ii) single cell genome sequencing. Both of these advances represent computational challenges that I address through development of novel methods in this thesis. The first proposed method (Chapter 2) incorporates prior clonal information to improve the accuracy of detecting SNVs across the genome of multiple bulk tumour samples. The second proposed method (Chapter 3) is a statistical model that exploits the underlying phylogeny of individually sequenced cells to detect SNVs in every individual cell. The latter method identifies clone specific SNVs without the requirement of deconvolving the results from bulk sequencing data. The resultant accurate detection of SNVs (Chapter 4) helps enhance insight on the evolutionary process of tumours and genetic pathways. Together, the methods provide a toolbox for comprehensive profiling of SNVs for the study of tumour dynamics.

Lay Summary

The accumulation of genetic mutations disrupts the regular activity of cells and leads to development of tumours. As cancerous cells divide, their acquired mutations are passed down to their descendants. This thesis proposes methods that exploit this information for detecting mutations with a resolution down to one cell. We expect the methods to be applied for the study of tumours clonal dynamics. Deciphering the clonal dynamics of tumours can potentially lead to enhanced cancer diagnosis and treatment.

Preface

Chapter 2 is a modified version of the material published in Dorri et al. [Dorri, F., Jewell, S., Bouchard-Côté, A., Sohrab P. Shah. Somatic mutation detection and classification through probabilistic integration of clonal population information. Nature Communications Biology 2, 44 (2019)]. The MuClone model was jointly conceived by myself, Dr. Shah, and Dr. Bouchard-Côté. I developed the MuClone software with Sean Jewell. I performed all computational analyses and I am the original creator and copyright holder of all the figures presented in this chapter. I co-wrote the text with Sean Jewell, Dr. Shah and Dr. Bouchard-Côté.

Chapter 3 contains unpublished material that is being prepared for a submission to a peer review journal. The model was jointly conceived by myself, Dr. Shah and Dr. Bouchard-Côté. I developed the CellMutScope software and performed all computational analyses described. The software was implemented within the Corrupt package. I wrote the text and I am the original creator and copyright holder of all the figures presented in this chapter.

Chapter 4 contains unpublished material that is being prepared for submission to a peer reviewed journal. I performed all computational analyses described in chapter 4. I wrote the text. The figures are generated from a modified version of Tyler Funnel's code for similar figures. I am the original creator and copyright holder of all figures presented in this chapter. The data presented in this chapter from a manuscript in preparation with Sohrab Salehi. The xenografting data was collected with informed patient consent according to procedures approved by the Ethics Committee at the University of British Columbia, under protocols H06-00289 BCCA-TTR-BREAST and H11-01887 Neoadjuvant Xenograft Study.

Table of Contents

List of Tables

List of Figures

Glossary

CNV Copy number variants

DLP Direct library preparation

DNA Deoxyribonucleic acid

FFPE Formalin-Fixed Paraffin-Embedded

NGS Next generation sequencing

PCR Polymerase chain reaction

PT Parallel tempering

ROC Receiver operating characteristic

SCS Single cell sequencing

SNP Single nucleotide polymorphism

SNV Single nucleotide variants

SV Structural variants

TNBC Triple-negative breast cancer

VAF Variant allelic fraction

WES Whole exome sequencing

WGA Whole genome amplification

WGS Whole genome sequencing

Acknowledgements

First and foremost, I wish to thank multitude of people who helped me throughout my journey as a PhD student. I would like to express my sincere gratitude to my advisor, Prof. Sohrab Shah, for giving me an invaluable opportunity to work on challenging and interesting projects over the past 6 years. Also, I would like to thank Prof. Anne Condon for all her support during these years. I would like to extend my appreciation for Prof. Alexandre Bouchard-Côté for serving on my dissertation committee. It has been a pleasure to work with and learn from him during these years.

I believe that graduate school was one of the most interesting pages of my life. I would like to thank all my friends and my colleagues at BC Cancer research Center, Dr. Razieh Annabestani, Dr. Sayeh Rajabi, Dr. Nilgoon Zarei, and many others for filling this page of my life with the most wonderful memories. Thank you all!

I would like to thank all who helped with their spiritual support. I feel a deep sense of gratitude for my mother and father, Zahra and Behrouz, who formed part of my vision, and taught me things that really matter in life. Their infallible love and support has always been my strength. Their patience and sacrifice will remain my inspiration throughout my life. I like to thank my twin sister, Faezeh and my brother-in-law, Hamid Mahini for their sincere help and support. They are always thanked in my heart for their enormous love. I would like to thank my lovely daughter, Elina, for all the joy and energy she gave me after her birth. Last but not least, I am deeply grateful to the love of my life and my husband, Mohsen Keshavarz Akhlaghi, for all support and sacrifice he made during every single step of this Journey.

It is impossible to remember all, and I apologize to those I have inadvertently left out. Lastly, thank you all and thank God!

To my lovely Elina

Chapter 1

Introduction

> *"It is paradoxical, yet true, to say, that the more we know, the more ignorant we become in the absolute sense, for it is only through enlightenment that we become conscious of our limitations. Precisely one of the most gratifying results of intellectual evolution is the continuous opening up of new and greater prospects."*
>
> - Nikola Tesla

Cancer is an evolutionary process driven by the accumulation of genetic variants. Intra-tumour heterogeneity is an inevitable consequence of this evolutionary process (Section 1.1). Identification of the underlying heterogeneous populations (clones) is an important step toward understanding the constituent parts of a tumour. Genetic variants can occur at single nucleotides, or there may be larger variants in the genome (Section 1.2). Because single nucleotide variants (SNVs) can be used as a marker in deciphering clonal dynamics, and because they are important in genetic pathways, accurate SNV detection is critical. Advanced sequencing technologies provide raw data with SNV information covered (1.3). In spite of numerous methods developed in the field (1.4), it is still desired to use this data to detect SNVs with higher accuracy. Incorporating the data obtained from different sequencing technologies, we proposed methods to detect SNVs with high accuracy, particularly SNVs that are only present in a relatively small population of cells (Section (1.5))

1.1 Clonal evolution in cancer and intra-tumour heterogeneity

During the cell division process, the DNA content of a cell has to be accurately transmitted in order to maintain the genetic integrity. However, the limited fidelity of this process results in genetic variants. Genetic variants arise in DNA of cells through different mechanisms including DNA division process or exposure to physical or chemical agents. DNA damage is induced by both endogenous (e.g. metabolic by-products such as reactive oxygen species) and exogenous factors (e.g. UV radiation and viruses) [38]. It is instructive to note the normal haploid human genome consists of about 3×10^9 base pairs. So, with the approximate 10^{-9} base pair mutation (single nucleotide variant) rate, it is likely that daughter cells inherit a new mutation during every division process [73, 130]. Unrepaired mutations are passed to the daughter cells and all subsequent descendants. Cells also possess a DNA repair mechanism. The rate of mutation accumulation is therefore a combined function of improper DNA replication inducing mutations and defective DNA repair mechanism [15]. Defects in DNA replication process or repair mechanism, or exposure to exogenous factors can increase the overall rate of mutations. For example, the mutation rates respectively reach as high as 12.6, 135.1 and 173.9 mutations per megabase of genome in high-grade serous ovarian cancer, breast cancer, and non-small cell lung cancer [20].

Most mutations (neutral or passenger mutations) do not impact a cell's phenotype. A vital phynotype, cell birth to death ratio, is affected by either (i) acquiring disadvantageous mutations that slow down the division process (this causes higher cell death rate), or (ii) driver mutations that cause an increase in cell birth rate. Acquiring a driver mutation can lead to a higher proliferation rate and development of cancerous cells [14, 52]. Besides genetic variants, epigenetic mechanisms like methylation and heterogeneous tumour micro-environments may also affect the proliferation and survival of tumour cells [142].

A group of cells with similar genetic variants and similar phenotypic properties are called clones [120]. Clonal expansion is the uncontrolled growth of a group of tumour cells with higher survival rate [48, 52]. A growing tumour cell population includes different (heterogeneous) clones with distinct genetic profiles [88, 92].

Intra-tumour heterogeneity describes the existence of heterogeneous clones with distinct phenotypic properties. For example, a clone with selective advantage survives better under some treatment [49, 55], or it can spread to other sites in the body by metastasis.

Intra-tumour heterogeneity is observed in various cancers including lung [27, 143], breast [36, 89, 90, 108, 137], ovarian [8, 106, 123], pancreatic [16, 136], kidney [42, 43], colorectal [118, 122], brain [117, 121], acute lymphoblastic leukemia [4], and prostate [18, 26, 51]. Intra-tumour heterogeneity complicates the accurate diagnosis of tumour, the prediction of disease progresses, and treatments [68, 77]. For example, in non–small cell lung cancer, T790M mutation is found to be resistant to EGFR inhibitors [29]. KRAS mutations in colon cancers is a marker of resistance to EGFR-I therapy [112]. BRCA1 reversion mutations in breast and ovarian cancers are identified as conferring resistance to PARP inhibitors and/or cisplatin [91]. ESR1 mutations are found to cause resistance in ER+ breast cancers treated with tamoxifen [61].

Identification of clones and measuring their dynamics (variations in clone's fitness and growth) is critical in deciphering the evolutionary process of cancer. The process of cells acquiring new mutations and passing them on to the daughter cells leaves a record of ancestral relationship between cells (see Figure 1.1 panel a). Basically, each tumour cell is a copy of its parents with a few probable mutations. Therefore, tumour evolution can be studied by inferring the genetic content of each cell. In order to determine the genetic profile of the existing clones, tissue samples dissected from parts of the body, are sequenced and aligned to the reference genome. This provides the raw data for inferring the genetic profile and for the subsequent studies.

Single nucleotide variants are the most common type of genetic variants [66]. In spite of the progress in developing SNV detection algorithms in recent years, the problem remains challenging particularly for detecting SNVs that exist only in a relatively small subset of cells (rare clones). This thesis is focused on detection of SNVs for the study of clonal populations and their dynamics. We showed that injecting clonal information improves the performance of detecting SNVs.

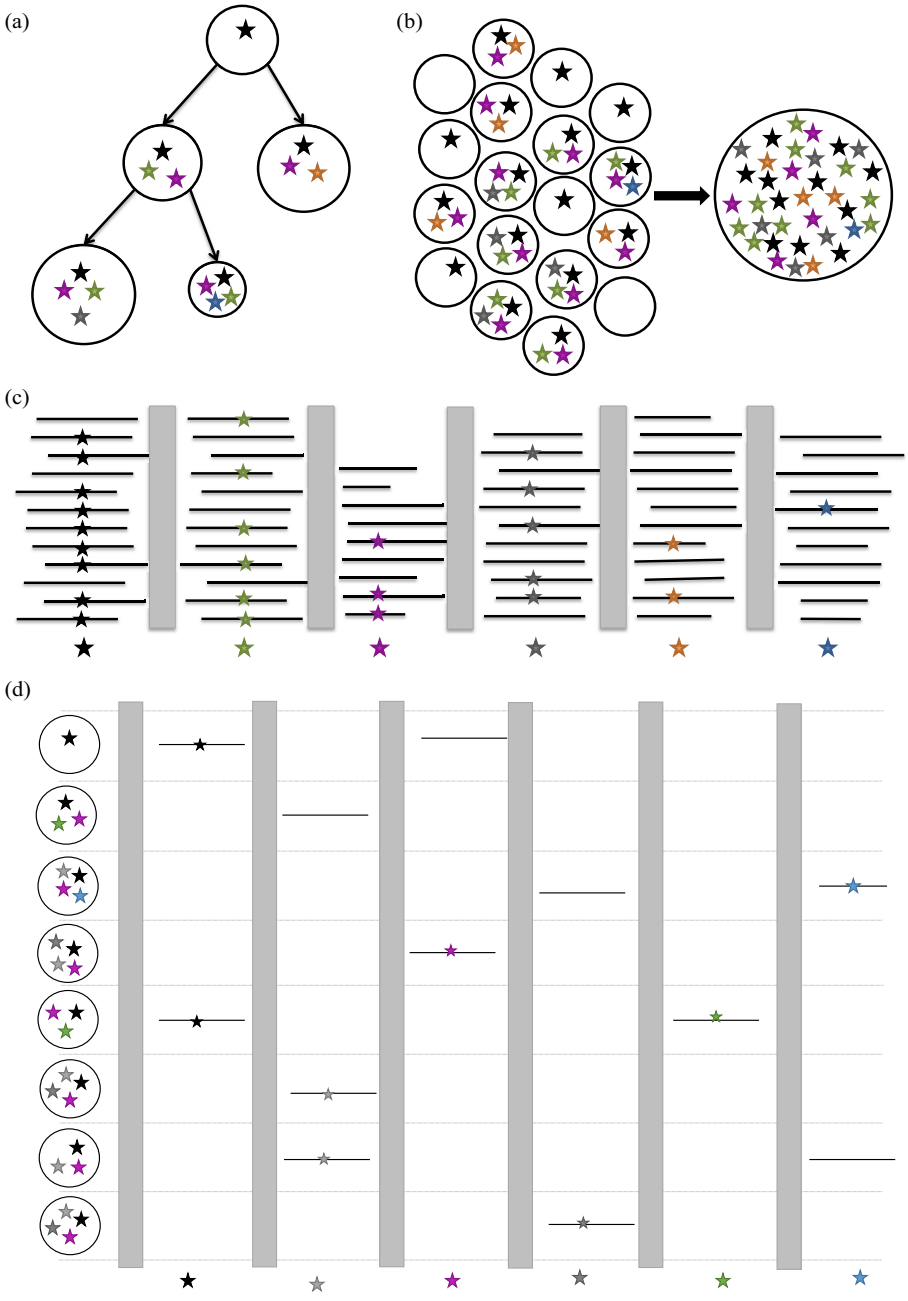

Figure 1.1: (Caption next page.)

Figure 1.1: (a) A hypothetical phylogenetic tree shows the ancestral relationship between clones (circles). Each clone is identified by mutations (stars). Unlike traditional phylogenetic trees, the internal clones may contribute to the observed data. The diameter of the circles (clones) is proportional to the clone's abundance. The black star represents an ancestral mutation, and other stars are clone specific mutations. (b) A sample of bulk sequencing data includes multiple cells from different clones. Their DNA content are extracted and are mixed together. The association of cells' genotype are lost in this step. (c) Horizontal black lines represent reads from the bulk sample that are aligned to the reference genome. Stars on the reads represents a variant allele compared with the reference genome. The blue star is a low prevalence mutation with a weak signal. (d) Single cells are separated, individually sequenced, and separately aligned to the reference genome. Data has the cells' genotype information.

1.2 The landscape of genetic variants in tumours

Over the evolutionary process of cancer, tumour genome may acquire various genetic variants including single nucleotide variants (SNV) and structural variants (variants that are across a larger number of nucleotide sites).

1.2.1 Single nucleotide variants

SNVs are point-like mutations that are a result of a single nucleotide substitution in the cells genome. Somatic SNVs refer to variants that are present in tumour cells. Germline SNVs refer to variants that are present in both normal and tumour cells. Despite being a relatively simple genomic aberration, detection of somatic SNVs in a small sub-population of cells (rare clones) is challenging. For example, Figure 1.1 panel c illustrates a high prevalence SNV and a low prevalence one by black and blue stars, respectively. It is obvious that while detection of high prevalence SNVs is rather trivial, detection of low prevalence ones is challenging due to weak signal to noise ratio. The weak signal is mainly because of (i) the presence of normal cells in the sample, (ii) copy number changes, and (iii) existence of SNVs only in a small population of cells.

1.2.2 Structural variants

In contrast with SNVs that are point-like, structural variants (SV) are changes in a large region of the genome. SVs are classified into different types depending on the nature of the variants. Deletion or multiplication of a region of the genome is called copy number variant (CNV). New genome arrangements resulting from genome breakage at different positions and subsequent attachment of the broken ends is referred as structural rearrangement. For example, inversion type rearrangement is when a segment of the genome is reversed end-to-end. Or translocation type rearrangement is when at least a segment of the genome moves to a new position on the same or on another chromosome.

1.2.3 SNVs and SVs in the study of tumour dynamics

Both SNVs and SVs contain valuable information for the study of the evolutionary dynamics of tumour clones. The selective advantage and the abundance of SNVs across the genome (numbering in the thousands) form a statistically robust marker. The signal of this marker is affected by CNVs. For example, the autosomal chromosomes of a diploid cell have two copies of each genome segment. If a SNV exists only in one of the copies, the number of variant nucleotides (mutated alleles) is half of the total number of alleles. However, if there is a CNV overlapping with the SNV locus, the variant allele ratio at that locus can be different. Therefore, consideration of CNV data can be instrumental in detection of SNVs, particularly the low prevalence ones. In both of the suggested models in this thesis (Chapter 2 and Chapter 3) CNV information is incorporated to improve the accuracy of detecting SNVs.

1.3 Next generation sequencing technologies

Next generation sequencing (NGS) significantly decreases the sequencing cost through massive parallel sequencing of millions of DNA fragments. Whole human genome was first sequenced using NGS [25, 78].

There are different NGS platforms, however their workflow can be summarized as follows (see Figure 1.2):

Library Preparation - DNA molecules of cells from a dissected sample are col-

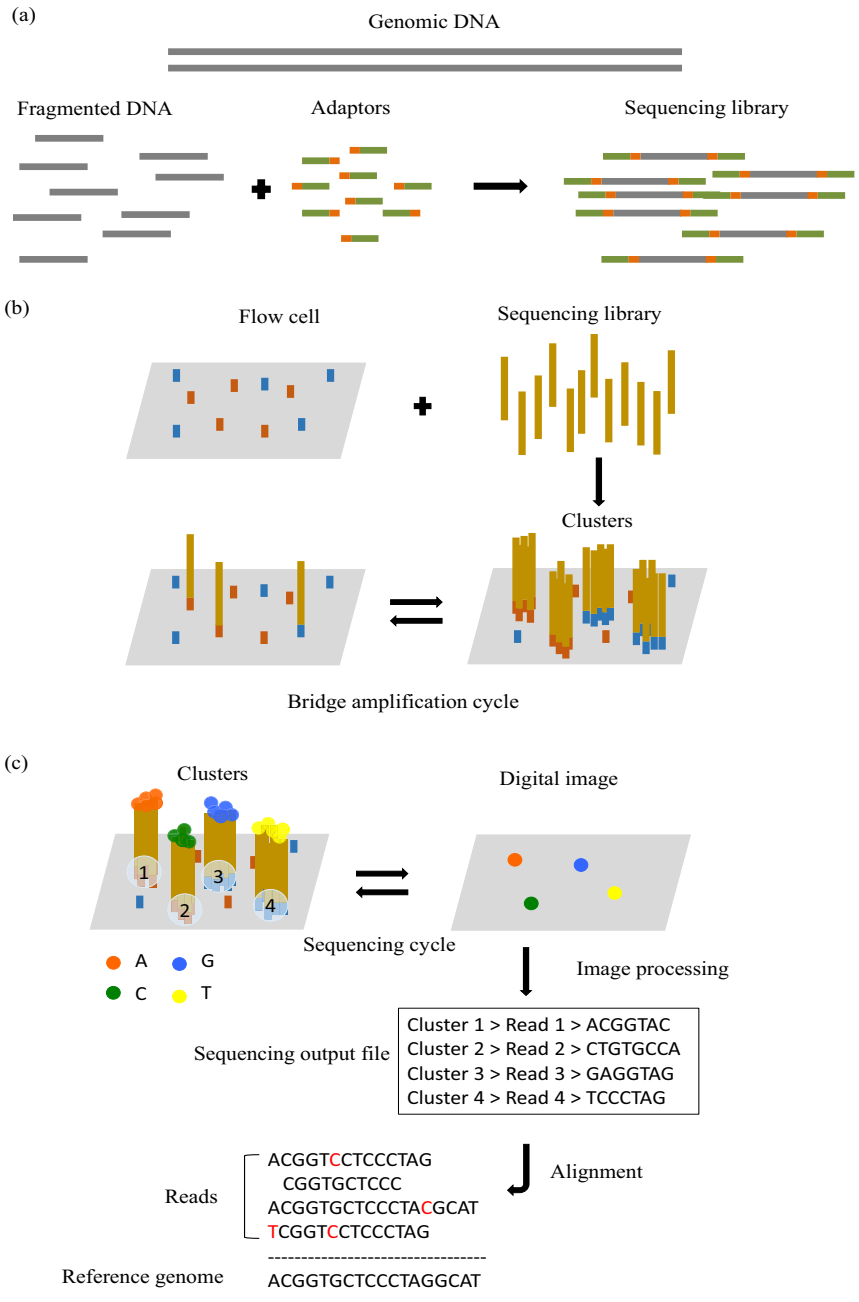

Figure 1.2: (Caption next page.)

Figure 1.2: Workflow of next generation sequencing technology (a) NGS library is prepared by fragmenting DNA molecules and ligating specialized adaptors to both fragment ends. (b) The library is loaded into flow cell, and the fragmented DNAs are attached to the surface. Each fragment is amplified into a clonal cluster through a polymerase chain reaction. (c) The flow cell is imaged, and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the bases. This cycle is repeated n times to create a read of length n. Reads are aligned to a reference genome using bioinformatics software. The aligned reads are used as raw data to identify genetic variants (modified figure from [59]).

lected. DNA molecules are fragmented, tagmented and ligated by adaptor sequences. In this step, the association of fragments and cells is lost.

Cluster Generation - The library is loaded into a flow cell where fragmented DNAs are amplified in a process called polymerase chain reaction (PCR). In this step each fragment is amplified into distinct clusters through a bridge amplification process (see Figure 1.2 panel a).

Sequencing - The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated n times to create a read of length n. This image-processing step is prone to errors. Different parameters characterize the accuracy of this step. These parameters are utilized as a measure of the errors in the subsequent analysis (see Figure 1.2 panel b).

Alignment and Data Analysis - The reads are aligned to a reference genome using bioinformatics software. Different post processing steps may be required for correcting sequencing and alignment errors. After alignment, the genetic variants can be inferred from the reads by comparing them with a reference genome (see Figure 1.2 panel c). The changes in genetic variants across time or across different anatomical locations demonstrate the clonal dynamics of the tumour.

In NGS data, each locus is covered by multiple overlapping reads that each represents a DNA fragment from the original sample [6]. This data format provides a major benefit for the analysis of variants in the genome. The reason is that the digital counts of the nucleotides (Adenine, Cytosine, Guanine, or Thymine) in each

locus quantify the nucleotides statistics. In the context of cancer, the allele counts are proportional to the population prevalence of an allele. This fact can be used to infer the prevalence of the clonal populations [100].

Several sequencing technologies were developed to study cancer genomes. We review two categories that are used to generate data for the models described in this thesis.

1.3.1 Bulk genome sequencing

Bulk sequencing technologies sequence multiple cells of a sample tissue altogether. Therefore, the output data is representative of the genome of a pool of tissue cells. The cells most likely belong to different populations (see Figure 1.1 panel a and b). The data must be deconvolved in order to identify the underlying genomic profiles of each cell or clone.

The sequencing can be done across whole genome (whole genome sequencing) or only across the exome regions (exome sequencing). The sequencing can also be done across a pre-specified segments of the genome (targeted sequencing).

Whole genome and exome sequencing

Both whole genome and whole exome sequencing start with a large pool of cells. The cells are lysed, and their DNA contents are extracted. The cell genotype association is lost. In whole genome sequencing (WGS) the entire genome of the admixed DNA is sequenced, while in whole exome sequencing (WES) only the exome segments of the genome are amplified and sequenced. In present whole genome sequencing, samples from DNA fragments are sequenced with the median coverage of 106x [32, 94, 96, 107, 108]. It means the average number of reads covering a position of genome after alignment is about 106. The median coverage is expected to increase with advances in the sequencing technologies. This number is larger in WES since it does not include all segments of the genome. One issue with both WGS and WES is that the input DNA comes from a pool of cells. It means normal cells are sequenced together with cancerous (tumour) cells. Therefore, the resultant reads can represent DNA fragments from both the tumour cells and uninteresting normal cells. In addition, the population of tumour cells is heterogeneous. All such

factors complicate detecting genetic variants (SNV or CNV) unique to rare clones.

Stochastic sampling of DNA fragments and low coverage of WGS or WES make DNA fragments of rare clones not represented or weakly represented by the reads (see the blue star mutation in Figure 1.1 panel (c). Copy number variants can also affect the observed mutation signal. For example, in Figure 1.1 panel a, the green and purple mutations have similar prevalence. However, due to copy number variants the allelic prevalence of purple mutation is lower than the prevalence of the green one (see Figure 1.1 panel c). The CNV can make the SNV signal weak, and that causes SNV detection tool to misinterpret it as an artifact. Therefore in bulk sequencing data, the allelic prevalence does not directly and properly represent the clone populations. The issue of the weak signal of SNVs can be partly addressed by increasing the sequencing coverage rate.

Deep-targeted sequencing

To overcome the issue with low coverage in WGS and WES, deep-targeted sequencing can be used [32, 107, 108]. In deep-targeted sequencing only selected regions of interest from the genome are amplified. The regions are usually the aberrant positions that are initially identified from WGS. Their polymerase chain reaction (PCR) primers for the sequencing are then designed accordingly. The advantage of this approach is that only a small portion of the genome is sequenced. This results in a data with a higher coverage of about 1000x-10000x. The higher coverage is beneficial through providing a stronger signal for accurate detection of SNVs (particularly the ones that are only present in a small sub-population of cells). Accurate detection of SNVs is critical for inferring the prevalence of underlying sub-populations and their dynamics.

1.3.2 Single cell sequencing

Single Cell Sequencing (SCS) technology sequences the genome of individual cells. This addresses the problem of DNA fragments being disassociated from the cells. In SCS, the first step is to isolate a cell from the primary sample to produce a viable single cell without a bias to any specific sub-population. The process is based on recent developments in flow sorting or microfluidics [10, 89, 128, 141]. A normal

single cell contains only about 7 pg of DNA which is not enough material for current sequencing platforms. So, the cell needs to go through either whole genome amplification (WGA), targeted polymerase chain reaction based protocols, or tagmentation approaches to produce enough material for sequencing [76, 82, 141]. SCS technology is expensive; it is in its early stages; and it is prone to different technical errors. The errors include allele drop-out, uneven coverage across the genome, and sequencing multiple cells (if two cells are sequenced together, the measurement will be doublet). Figure 1.1 panel d, synthetically depicts the reads from single cell sequencing aligned to the reference genome. As shown, the data is sparse with uneven coverage across the genome.

1.4 Literature review on SNV calling models

Accurate detection of somatic single nucleotide variants is important in defining clonal composition of the human cancers. The genomic data from tumour tissue is accessible either through bulk sequencing of multiple cells or single cell sequencing. However, each data type has its own challenges. The data from bulk sequencing is generated from admixed DNA material of a large pool of cells. The pooled data complicates the detection of SNVs that are only present in a subset of cells (small clone). In single cell sequencing, the cell genotype association is maintained. However, detection of SNVs remains challenging because the data is very sparse with uneven coverage across the genome.

1.4.1 Literature review on SNV calling from bulk sequencing data

Theoretically, SNVs can be identified given data with enough read depth coverage regardless of their variant allelic ratio. However, detecting SNVs with high confidence is challenging due to the artifacts and the limited read depth coverage. Most of the artifacts are present in low frequency and they interfere with the detection of low prevalence SNVs [135]. SNV calling has been an active research field for years. The state-of-the-art tools are designed to detect SNVs from matched tumour-normal samples, a single tumour sample, or multiple normal and tumour samples. Matched normal and tumour samples are required to detect somatic SNVs. If the matched normal sample is not available or calling germline is desired, a single sample SNV

caller will be needed. In addition, there has been an increase in deciphering clonal population dynamics over clonally related multiple samples (same patient samples taken at different points in time or space). This has resulted in abundance of WGS or WES data from multiple samples. With the access to this data, methods incorporating multiple clonally related samples are proposed to detect SNVs with higher accuracy.

Matched tumour-normal SNV calling

There is a wide range of methods proposed for SNV calling using matched normal tumour samples. Most of the approaches can be categorized into four groups: (i) Bayesian approaches, (ii) statistical test approaches, (iii) heuristic approaches, and (iv) machine learning approaches.

Bayesian approach SomaticSniper, JointSNVMix2, Samtools, Virmid, FaSD-somatic and SNVSniffer are examples of the models taking Bayesian approach for detecting SNVs [24, 67, 79, 81, 99, 127]. These methods evaluate the posterior probability of the joint genotype of normal and tumour samples (assuming they are diploid). The methods may apply some filtering as the post processing step to improve the performance. The posterior probability of the joint genotype is calculated by

$$P(G_T, G_N | D_T, D_N) = \frac{P(D_T, D_N | G_T, G_N) P(G_T, G_N)}{\sum_{g_T, g_N \in G} P(D_T, D_N | g_T, g_N) P(g_T, g_N)}. \tag{1.1}$$

G_T and G_N, tumour and normal genotypes, take value from $G = \{AA, AC, AG, AT, CC, CG, CT, GG, GT, TT\}$, where A, C, G, T represents DNA nucleotides. D_T and D_N are tumour and normal data (aligned reads), respectively. The prior genotype probability, $P(G_T, G_N)$ depends on the prior knowledge of mutation rate. The joint genotype likelihood follows a Binomial probability distribution with the mean parameter that depends on the genotype and sequencing error rate [135]. SomaticSniper and FaSD-somatic detect a site as a SNV if the probability of having different normal and tumour genotypes is high [79, 127]. SAMtools follows the same idea but uses log-likelihood ratio instead of posterior probability [24]. JointSNVMix2, Virmid and SNVsniffer simplify the 10 genotype state of G into

AA, *AB* and *BB* [67, 81, 99]. *A* denotes the allele matching the reference and *B* is the one that does not match. Therefore, the probability of marking a site as a SNV is calculated by

$$P(Somatic) = P(AA,AB) + P(AA,BB), \tag{1.2}$$

where the normal genotype alleles match the reference (AA or homozygous reference). The tumour genotype is heterozygous in the first term (AB), and a non-reference homozygous genotype in the second term (BB) [135]. Virmid considers the tumour as a mixture of normal and tumour samples with somatic mutations [67]. The model infers the tumour content and joint genotype. JointSNVMix2 applies a hierarchical Bayesian model to estimate the joint genotype probabilities [99]. SNVsniffer initially calls high confidence SNVs using a set of heuristic thresholds. Then, the resultant low confidence calls are further tested using the joint genotype model [81].

The assumption that the normal and tumour samples are diploid, is relaxed in Strelka, MuTect, deepSNV, and EBCall [22, 44, 105, 111]. These models usually incorporate the joint variant allele frequencies (VAF), denoted by (f_T, f_N), instead of the joint genotypes (G_T, G_N). Strelka infers the VAFs of normal and tumour samples. Assuming the normal sample genotype as a homozygous reference, the model calculates the probability of difference in VAFs of normal and tumour samples. MuTect considers two models (i) wildtype model, and (ii) mutation model. The wildtype model assumes that the non-reference reads come from artifacts, whereas mutation model assumes that the variant alleles come from an unknown frequency. MuTect then computes which of the aforementioned models are more likely [22].

In local realignment, the reads are assembled together, and the read supports are then counted. Mutect2, Platypus, FreeBayes incorporate the local realignment step [3, 39, 98]. This strategy improves the performance particularly in regions prone to errors. It also gives information about the coexistence of mutations.

Statistical test approach In this approach the model's null hypothesis is wildtype and the alternative hypothesis is variant [19, 44, 111, 131]. LoLopicker and LoFreq use a sequencing quality score to infer the error rates [19, 131]. DeepSNV estimates the error rates directly from the input normal and tumour samples, while EBcall uses other independent control samples [44, 111]. Note LoLoPicker and

deepSNV are particularly designed for deep targeted sequencing data to detect low prevalence SNVs [19, 44].

Heuristic approach The methods with heuristic approach make use of different thresholds to identify mutation candidates, and then apply ad hoc rules to determine somatic SNVs [53, 69, 74, 135]. For instance, VarScan2 requires the variant allelic ratio of more than 8% with at least 2 reads supporting the variant [69]. The thresholds are mainly set above the expected artifact rate. For the loci passed the thresholds, VarDict, VarScan2 and Shimmer apply Fisher's exact test in a two-by-two table of read counts [53, 69, 74]. Small p-value denotes somatic SNVs that have normal and tumour variant reads significantly different.

Machine learning approach Another group of methods for calling SNVs are based on machine learning. The methods classify mutations into either (i) wild-type/germline or (ii) somatic. MutationSeq and SNooper use supervised classifiers like random forest, while BAYSIC applies an unsupervised latent class model to detect somatic SNVs [17, 31, 119].

Single sample SNV calling

Single sample SNV calling is required when the matched normal sample is not available (e.g Formalin-Fixed Paraffin-Embedded (FFPE) samples), or when SNV calling is required for a tumour deep targeted sequencing data. Outlyzer, ISOWN, and SomVarIUS, Pisces, SPLINTER, SNVer, SiNVICT are examples of the models particularly designed for deep targeted sequencing data [35, 64, 70, 86, 116, 124, 129]. Outlyzer, ISOWN, and SomVarIUS use FFPE data in their analysis [64, 86, 116], and SNVMix2 use single tumour WGS/WES data [47].

OutLyzer estimates the background noise and calls the variants that are above the noise level [86]. SPLINTER generates an error model of the pooled samples to detect low prevalence variants. SNVer calls variants if the VAF exceeds a set of thresholds based on Binomial distribution [129]. SNVMix2 model is based on the inference of the posterior probability of different genotypes [47].

ISOWN, SiNVICT and SomVArIUS are other single sample SNV calling methods that are capable of distinguishing somatic and germline SNVs [64, 70, 116]. ISOWN uses a supervised machine learning approach to classify the variants. The

classification uses different features including the membership of databases (like COSMIC for somatic mutations and dpSNP for germlines) [64]. SomVarIUS considers the VAF distribution of heterozygous germline SNPs as a measure to detect SNVs. The loci that are placed on the left tail of the distribution, are identified as somatic SNVs [116]. SiNVICT is designed for detecting low prevalence SNVs from circulating tumour DNA (ctDNA). The loci that have VAF significantly lower than 0.5, are called somatic SNVs [70].

Multiple sample SNV calling

Intra-tumour heterogeneity complicates the detection of SNVs. In particular, the clone specific mutations that are only present in a small fraction of cells may be missed. Increasing interest in deciphering clonal population dynamics from multiple samples collected over time or over anatomical space, results in availability of multiple clonally related sequencing data. With the access to this data, methods like MultiSNV, SNV-PPILP and MultiGems are proposed [63, 87, 126]. MultiSNV jointly analyses all available samples under a Bayesian framework to improve the performance of calling SNVs that are shared [63]. SNV calls from GATK are refined and corrected by using phylogeny information across multiple samples [83]. MultiGeMS calls SNVs using a statistical model selection procedure and accounts for enzymatic substitution sequencing errors [87].

1.4.2 Literature review on SNV detection from single cell sequencing data

Single cell sequencing data brings an opportunity for detecting SNVs at each cell. However, the data is not perfect and introduces challenges including: low breadth and depth coverage, biased allelic counts, and possibility of sequencing multiple cells (most probably two cells). Monovar addressed the low and uneven data coverage by pooling the sequencing data across cells without assuming any dependency across sites [140]. SCcaller identifies SNVs while accounting for local allelic amplification biases. However, it cannot recover mutations from drop-out events or from loss of heterozygosity [33]. Conbase and LiRA utilize read phasing to correct for errors and allelic drop-out [12, 54]. Conbase unlike LiRA performs joint vari-

ant calling across the population of cells. Although Conbase and LiRA use read phasing in order to decrease false discovery rate (FDR) and increase specificity, the analysis of bases is only possible for regions in proximity to single nucleotide polymorphisms (SNPs). This requirement refrains to call SNVs across any bases that are covered by reads. SCIΦ detects SNVs while inferring the phylogenies with mutations propagated along tree branches [115]. Although SCIΦ claims to detect mutation in single cells with very low or even no coverage, it requires at least two cells to show an alternative nucleotide count of at least three [115]. Genotyper is a method to infer clonal genotypes. It is based on a statistical model coupled with a mean-field variational inference approach [101].

The advent of DLP protocol resulted in data that consists of thousands of cells with very low coverage (about 0.07x). The existing methods are only scalable to hundreds of cells and require more coverage for calling SNVs. Therefore, a SNV caller that considers the characteristics and scale of the data is needed.

1.5 Motivations and research contribution

We discussed the evolutionary process of cancer where mutations show a record of ancestral relationships. Therefore, the accuracy of mutation calling is critical in inferring the underlying phylogeny and the clonal dynamics. We reviewed the existing sequencing technologies and the methods for detecting SNVs using (i) bulk sequencing data, and (ii) single cell sequencing data. Conventionally, mutations are the input for inferring the underlying clones' phylogeny. In this thesis, we show that injecting underlying clonal structure into the SNV detection model improves the performance of detecting SNVs. Our method improves the accuracy of inferring the phylogeny and detecting mutations through multiple iterations.

The first contribution is to use the underlying clonal information (can be inferred from deep targeted sequencing data) in a model to accurately detect SNVs. This is done using WGS or WES data from multiple clonally related bulk tumour samples. The samples can be from different anatomical locations or multiple samples from different time points. We encode detecting SNVs in a generative probabilistic framework, and we call it MuClone [34]. It performs joint statistical inference on multiple observations (from multiple samples) of the allele counts of a

locus of interest. We evaluated the performance of MuClone on both synthetic and real data in Chapter 2.

The second contribution is to use the underlying phylogeny (can be inferred from CNV data) in a model to detect SNVs in every individual cell. The model utilizes single cell sequencing data from one or multiple clonally related tissue samples. We call the model CellMutScope, and it incorporates a Bayesian statistical framework. It performs joint statistical inference on the allelic counts of multiple cells at a locus of interest, and outputs the posterior probability of being a mutation at the loci for every individual cell. Also, we frame incorporation of SNV data in the Corrupt model (Corrupt infers the underlying phylogeny of cells using CNV data). The extended tree with both SNV and CNV discloses the underlying tree phylogeny in more detail. The performance of CellMutScope is benchmarked is Chapter 3, and its performance on real data is presented in Chapter 4. Chapter 5 concludes by a summary of contributions and a discussion of future works.

Chapter 2

Somatic mutation detection and classification through probabilistic integration of clonal population structure

"A very great deal more truth can become known than can be proven."

- Richard Feynman

This chapter introduces MuClone, a new model for detection of SNVs across multiple clonally related whole genome sequencing data or whole exome sequencing data. In addition to detection of SNVs, MuClone classifies SNVs into biologically meaningful groups to allow study of clonal dynamics. The problem and the existing approaches are reviewed in Sections 2.1 and 2.2, respectively. Detection of SNVs using WGS or WES of multiple tumour samples by MuClone is presented in Section 2.3. We demonstrated that incorporating clonal information into joint analysis of multiple samples improves SNV detection. This has a particular advantage for low prevalence SNVs. The performance of MuClone on both synthetic and real data is reported in Section 2.4. Section 2.5 concludes and discusses the next steps.

2.1 Introduction

Genomic accumulation of somatic SNVs can disrupt the regular activity of cells and can result in cancer initiation and progression. Collectively, the complete repertoire of SNVs across a cancer genome (numbering in the thousands) form a statistically robust marker for inferring clonal populations to study tumour evolution. As such, accurate detection of all somatic SNVs, including those with low prevalence, is vital as they can define clones with phenotypic properties of interest. Mechanistic association of specific clones with properties such as treatment resistance, metastatic potential, and fitness under therapeutic selective pressures remains a key objective of biomedical investigators studying tumour progression.

Phylogenetic analysis can encode the evolutionary lineage of tumour cells across time and anatomic space [46, 60, 71, 84, 90, 107, 136]. Sequencing of multiple samples of a cancer to reconstruct evolutionary patterns and drug response profiles is increasingly common. For example, in rapid autopsy programs, at the time of a patient's death, tens to hundreds of metastatic samples are collected for future study [113, 134]. Recent multi-sample sequencing studies in renal, lung, ovary, breast, colorectal and other cancers have revealed striking evolutionary and clinically important properties of cancers [36, 42, 60, 84]. However, the analytical methods to detect mutations from such experimental designs are still immature, and few studies have leveraged shared statistical strength across samples to detect mutations with greater sensitivity.

In the limiting case, all cells likely harbour unique genomes; however due to the nature of branched evolutionary processes, clones can be coarsely modelled as major clades in the cell lineage phylogeny of a cancer. These clades share the majority of mutations, and therefore define first approximations to the genotypes of clones. Clonal genotypes and their relative abundances in the cancer cell population can be approximated by clustering mutations measured in bulk tissues and estimating their cellular prevalences (the variant tumour cell fraction) [100, 139].

Phylogenetic algorithms mostly use mutations (represented as binary genetic markers), as inputs to infer the branched evolutionary lineages of tumour cells [30, 95]. Thus, mutation detection accuracy will ultimately impact the performance of phylogenetic inference algorithms.

Detection of low prevalence mutations is a major challenge due to typically small signal to noise ratios, owing to: (i) contamination by normal cells; (ii) genome copy number alteration; and (iii) the presence of mutations in only a small fraction of tumour cells (intra-tumour heterogeneity). In this work, we illustrate that knowledge of the clonal population structure improves detection of mutations defining low prevalence clonal genotypes.

2.2 Literature review

Although SNV calling algorithms are ubiquitous in the literature, it remains challenging to detect low prevalence mutations. Algorithms have been developed for calling mutations from a single sample [47, 69], paired (matched normal and tumour) samples [22, 31, 67, 99, 105], or multiple samples [63, 126]. We list several popular algorithms. Mutationseq uses a feature based classifier for calling mutations [31], where the features are constructed from matched paired normal and tumour samples. Strelka is a method for somatic SNV and small indel detection from sequencing data of matched normal and tumour samples [105]. It is based on a Bayesian approach that uses normal and tumour samples' allele frequencies with normal expected genotype structure. MuTect uses a Bayesian classifier that employs various filters to ensure high specificity to detect mutations from matched normal and tumour samples [22]. FreeBayes uses short read alignments to call the most likely genotypes for the population at each position. It can be run in single mode using only one tumour sample or in multiple mode utilizing multiple tumour samples from the same patient [39]. FreeBayes can detect somatic mutations if germlines are manually removed. MultiSNV jointly analyses all available samples under a Bayesian framework to improve the performance of calling shared mutations [63]. SNV calls from GATK [83] are refined and corrected by using phylogeny information across multiple samples [103, 126].

2.3 Method

MuClone uses previously known cellular prevalence information to improve mutation detection and classification. For each sample, MuClone detects mutations from joint analysis of multiple samples. We encode this process in a generative

probabilistic framework to perform joint statistical inference of multiple observations (from multiple samples) of the variant allele counts of a mutation of interest. The inputs to the model are: the number of variant reads, and the depth for a set of sequenced loci from multiple samples derived from the same patient; a measure of allele specific copy number at each locus, in each sample, with tumour content; and the cellular prevalence and the abundance of underlying mutation clusters. MuClone outputs (i) a probability for each locus, at each sample, of being a mutation, and (ii) its cluster.

The probabilistic graphical model of MuClone is depicted in Figure 2.1.

2.3.1 Problem formulation

In our proposal, MuClone, we exploit prior knowledge of tumour content, tumour cellular prevalence, and copy number information across multiple samples to improve detection of somatic single nucleotide variants, and in particular, low prevalence ones. Our model uses mutation clusters and copy number information obtained from standard approaches [50, 125]. In the first step, a set of stringent SNV calls or validated SNVs (using targeted sequencing data) is used to infer mutation cluster information. Then, MuClone uses the inferred cluster information to more accurately call mutations across genome (whole genome or exome sequencing data). In addition to calling mutations, MuClone also classifies mutations into clusters based on cellular prevalence. This provides the user with the opportunity to profile mutation changes across time and space, and adds a rich layer of interpretation into the detection process.

2.3.2 Model description

We first define $g_{m,n}$, the genotype of a given locus n in sample m. Samples are indexed by $1 \dots M$ and loci are indexed by $1 \dots N$. Given the copy number $c_{m,n}$, the possible genotype states are $\mathscr{G} = \{A \dots A, AA \dots B, A \dots BB, \dots, B \dots B\}$, where each element has a length equal to $c_{m,n}$. For example, the genotype ABB refers to the genotype with one reference allele A and two variant alleles B. For simplicity, we assume the number of reads containing the variant alleles, $b_{m,n}$ at a given locus with genotype $g_{m,n}$ and read depth $d_{m,n}$ follows a Binomial distribution with

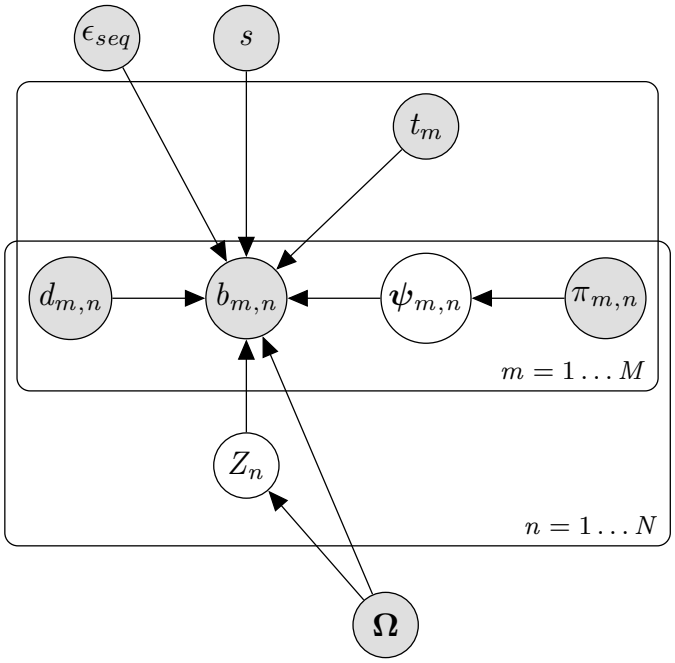

Figure 2.1: Probabilistic graphical model for MuClone: white nodes are unobserved variables; grey shaded nodes are observed variables. The variables $m \in \{1 \ldots M\}$ and $n \in \{1 \ldots N\}$ index the samples and the loci, respectively. In sample m, $d_{m,n}$ is the total number of reads aligned at locus n and $b_{m,n}$ is the number of aligned reads with B alleles. The genotype state is $\psi_{m,n}$ and $\pi_{m,n}$ is the prior over the genotype states. The tumour content of sample m is t_m and the error rate is ε_{seq}. The parameter s stands for the precision parameter. Tumour clusters prior and their cellular prevalence information are encoded in $\mathbf{\Omega}$ and the variable Z_n denotes the mutation cluster.

genotype specific variant probability $p(g_{m,n})$

$$b_{m,n}|d_{m,n}, p(g_{m,n}) \sim \text{Binomial}\left(d_{m,n}, p(g_{m,n})\right). \tag{2.1}$$

For $g_{m,n} \in \mathscr{G}$, the variant probability $p(g_{m,n}) : \mathscr{G} \rightarrow [0,1]$ is defined as

$$p(g_{m,n}) = \begin{cases} \frac{\mathscr{B}(g_{m,n})}{\mathscr{C}(g_{m,n})} & \mathscr{B}(g_{m,n}) \neq 0, , \mathscr{B}(g_{m,n}) \neq \mathscr{C}(g_{m,n}) \\ \varepsilon_{seq} & \mathscr{B}(g_{m,n}) = 0, \\ 1 - \varepsilon_{seq} & \mathscr{B}(g_{m,n}) \neq 0, \mathscr{B}(g_{m,n}) = \mathscr{C}(g_{m,n}), \end{cases} \tag{2.2}$$

where $\mathscr{B}(g_{m,n}) : \mathscr{G} \rightarrow \mathbb{N}$ and $\mathscr{C}(g_{m,n}) : \mathscr{G} \rightarrow \mathbb{N}$ return the number of the variant alleles of genotype $g_{m,n}$ and its copy number, respectively. For example $\mathscr{B}(ABB) = 2$ and $\mathscr{C}(ABB) = 3$. The variable $\varepsilon_{seq} > 0$ is a small positive constant that accounts for sequencing error. It allows for non-zero variant reads, due to sequencing error, when there are no variant alleles in genotype $g_{m,n}$.

However, since the sequenced reads are independently sampled from an infinite pool of DNA fragments, at a given locus, each read may belong to the normal, reference, or variant population. The normal population stands for normal cells; the reference population are tumour cells which do not have the mutation at the given locus; and the variant population are the ones carrying the mutation. Therefore, using a single genotype state, $g_{m,n}$, introduces error into our analysis. To account for this fact, we consider using the full genotype state, $\psi_{m,n} = (g_{N_{m,n}}, g_{R_{m,n}}, g_{V_{m,n}})$, at a given locus n and sample m, to model the number of variant reads. Normal population fraction is $1 - t_m$ where t_m is the tumour content (the proportion of tumour cells in the sample) of sample m and the cellular prevalence of the mutation is ϕ_m^z which is the fraction of tumour cells carrying the mutation. According to our prior knowledge, we assume mutations are clustered into K clusters. For a given locus n, $Z_n = z \in \{1, \ldots, K\}$ defines which cluster the mutation belongs to. If a position is not a mutation then it belongs to wildtype cluster identified by $Z_n = 0$.

Therefore, for a mutation with cellular prevalence ϕ_m^z and tumour content t_m, the variant allele probability is denoted by $\xi(\psi_{m,n}, \phi_m^z, t_m)$. It is proportional to the sum of the (properly scaled) variant probabilities from each population:

$$\begin{aligned} \xi(\psi_{m,n}, \phi_m^z, t_m) \propto\ & (1 - t_m)\mathscr{C}(g_{N_{m,n}})p(g_{N_{m,n}}) + \\ & t_m(1 - \phi_m^z)\mathscr{C}(g_{R_{m,n}})p(g_{R_{m,n}}) + \\ & t_m\phi_m^z\mathscr{C}(g_{V_{m,n}})p(g_{V_{m,n}}), \end{aligned} \tag{2.3}$$

where the first term $(1-t_m)\mathscr{C}(g_{N_{m,n}})p(g_{N_{m,n}})$ is proportional to the probability of sampling a read containing variant allele from the normal population, and the second and third terms, $t_m(1-\phi_m^z)\mathscr{C}(g_{R_{m,n}})p(g_{R_{m,n}})$ and $t_m\phi_m^z\mathscr{C}(g_{V_{m,n}})p(g_{V_{m,n}})$, are proportional to the probabilities of sampling a read containing variant alleles from the reference and variant populations, respectively.

Considering the full genotype state, the number of reads containing the variant alleles at a given locus n that belongs to cluster Z_n follows a Binomial distribution with probability

$$\mu(Z_n) \quad = \quad \begin{cases} \varepsilon_{seq} & \text{if } Z_n = 0 \\ \xi(\psi_{m,n}, \phi_m^z, t_m) & \text{if } Z_n = z \text{ and } z \in \{1, \ldots, K\}, \end{cases} \tag{2.4}$$

where ε_{seq} accounts for sequencing error in wildtype cluster and $\xi(\psi_{m,n}, \phi_m^z, t_m)$ is the variant alleles probability for nth locus, mth sample from zth cluster. According to Equation (2.3), tumour content and cellular prevalence information are incorporated to estimate $\xi(\psi_{m,n}, \phi_m^z, t_m)$.

Since empirical evidence shows that variant read data is overdispersed, we replace the Binomial model (2.1) with a BetaBinomial model

$$b_{m,n}|d_{m,n}, \mu(Z_n), s \sim \text{BetaBinomial}\left(b_{m,n}|d_{m,n}, \mu(Z_n), s\right), \tag{2.5}$$

where $\mu(Z_n)$ is the expected variant alleles probability and the hyperparameter s is the precision parameter of the BetaBinomial distribution. The BetaBinomial distribution in Equation (2.5) assigns a small chance for mutation when the locus is wildtype, otherwise it is governed by the prior clonal information.

To fully express our model, for each locus, we assume the genotype state follows a categorical distribution with probability vector $\pi_{m,n} \in [0,1]^{|\mathscr{G}|}$ whose ith element is the probability of the ith genotype state,

$$\psi_{m,n}|\pi_{m,n} \sim \text{Categorical}\left(\pi_{m,n}\right). \tag{2.6}$$

The number of possible genotype states, denoted by $|\mathscr{G}|$, is finite given the copy number information. For simplicity, we assume every element of $\pi_{m,n}$ is equal to $\frac{1}{|\mathscr{G}|}$.

In addition, we also assume that the clonal assignment of a locus, denoted by Z_n, follows categorical distribution with probability vector $\boldsymbol{\tau}$:

$$Z_n|\boldsymbol{\tau} \sim \text{Categorical}\,(\boldsymbol{\tau})\,. \tag{2.7}$$

Our probabilistic framework can be succinctly written as

$$\begin{aligned} b_{m,n}|d_{m,n},\mu(Z_n),s &\sim \text{BetaBinomial}\,(b_{m,n}|d_{m,n},\mu(Z_n),s)\,,\\ \psi_{m,n}|\pi_{m,n} &\sim \text{Categorical}\,(\pi_{m,n})\,,\\ Z_n|\boldsymbol{\tau} &\sim \text{Categorical}\,(\boldsymbol{\tau})\,. \end{aligned} \tag{2.8}$$

2.3.3 Inference

Based on the generative model introduced in (2.8) mutations are inferred via the posterior probability distribution of a locus n belonging to cluster z:

$$P(Z_n = z|b_{m,n},d_{m,n},s) \propto \tau_z \prod_{m=1}^{M} \sum_{i\in I} \pi_{m,n_i} \mathscr{L}(Z_n = z|b_{m,n},d_{m,n},s), \tag{2.9}$$

where the variable i indexes $\pi_{m,n}$ over the genotype states, $\boldsymbol{I} = \{1\ldots|\mathscr{G}|\}$. The posterior probability of locus n belongs to cluster z is proportional to the likelihood of observing $b_{m,n}$ number of nucleotides matching the variant alleles times the prior over tumour cluster z. The tumour cluster prior τ_z is the fraction of mutations belonging to cluster z and it has been tuned according to wildtype prior; the tumour cluster prior and the cellular prevalence information are encoded in $\boldsymbol{\Omega}$. Wildtype prior is our prior information if a locus is a mutation. If we don't have any information we can set it to 0.5. The likelihood function, $\mathscr{L}(Z_n = z|b_{m,n},d_{m,n},s)$, is the BetaBinomial distribution defined in 2.5.

Based on basic decision theory, a decision can be extracted from a posterior distribution given a loss function. Under the loss function $\ell(z,z') = \mathbf{1}[\mathbf{1}[z=0] \neq \mathbf{1}[z'=0]]$, the decision is simply the maximum a posteriori (MAP). That is, if the probability η of belonging to any of the tumour clusters is greater than 0.5, we conclude that the locus is mutated in at least one of the M samples. The value of η

is

$$\eta = \sum_{z=1}^{K} P(Z_n = z | b_{m,n}, d_{m,n}, s).$$

If locus n is mutated in at least one of the M samples, then the probability of mutation, in each sample, is calculated separately as

$$P_{m,n}(\text{mutant}) = \sum_{j \in \boldsymbol{J}_m^*} P(Z_n = j | b_{m,n}, d_{m,n}, s), \tag{2.10}$$

where $\boldsymbol{J}_m^*$ is the set of clusters of sample m whose cellular prevalences are greater than a fixed positive threshold called Φ_T,

$$\boldsymbol{J}_m^* = \{j \mid \phi_m^j > \Phi_T\}. \tag{2.11}$$

The threshold Φ_T distinguishes the clusters of sample m in which their non-zero cellular prevalence are due to actual variant alleles. The default value of Φ_T is zero. However, depending on the method used for estimating cellular prevalences, it can be set to another positive value, if some non-zero input cellular prevalences indicate wildtype clusters.

In addition, MuClone assigns the locus to cluster z^* that maximizes

$$z^* = \underset{z \in \{1,\ldots,K\}}{\text{argmax}}\ P(Z_n = z | b_{m,n}, d_{m,n}, s). \tag{2.12}$$

This classifies mutations to one of the previously known clusters. The classification of mutations helps in biological interpretation and phylogenetic analysis of the data.

2.4 Experimental results

2.4.1 Synthetic data

In this section, we examine the performance of MuClone on simulated data. In what follows, we generate N loci from M samples with K underlying tumour mutation clusters with sequencing error rate ε_{seq} and tumour content t_m.

We first randomly generate an evolutionary relationship between clusters, viewed

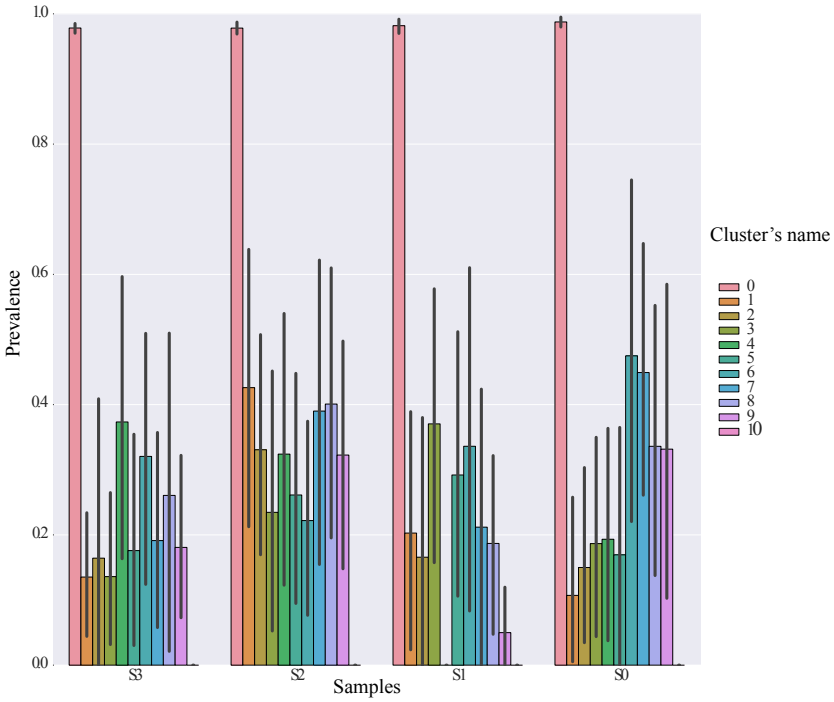

Figure 2.2: Distribution of the cellular prevalence of clusters across different samples in multiple runs.

as a binary phylogenetic tree. Each node in the tree represents a mutation cluster. The root node represents the ancestral cluster. The cellular prevalences of the first descendant, $\phi_{1^{st}}$, is sampled from a Uniform distribution over $[0, \phi_{\text{parent}}]$, where ϕ_{parent} is the cellular prevalence of the parent node (cluster). The cellular prevalences of the second descendant, $\phi_{2^{nd}}$, is sampled from a Uniform distribution over $[0, \phi_{\text{parent}} - \phi_{1^{st}}]$, defined so that the sum of the children's prevalences do not exceed their parent's cellular prevalence. The absence or presence of each cluster, in each sample, is sampled from a Bernoulli distribution that assigns equal probability to both outcomes. Distribution of the cellular prevalence of 10 clusters across 4 samples in multiple runs is depicted in Figure 2.2. Figure 2.3 is an example of cellular prevalences of clusters across different samples in one random run.

If a cluster is not present in a sample, the corresponding cellular prevalence will be 0. See Figure 2.4 for an example of this process. Loci are assigned to a

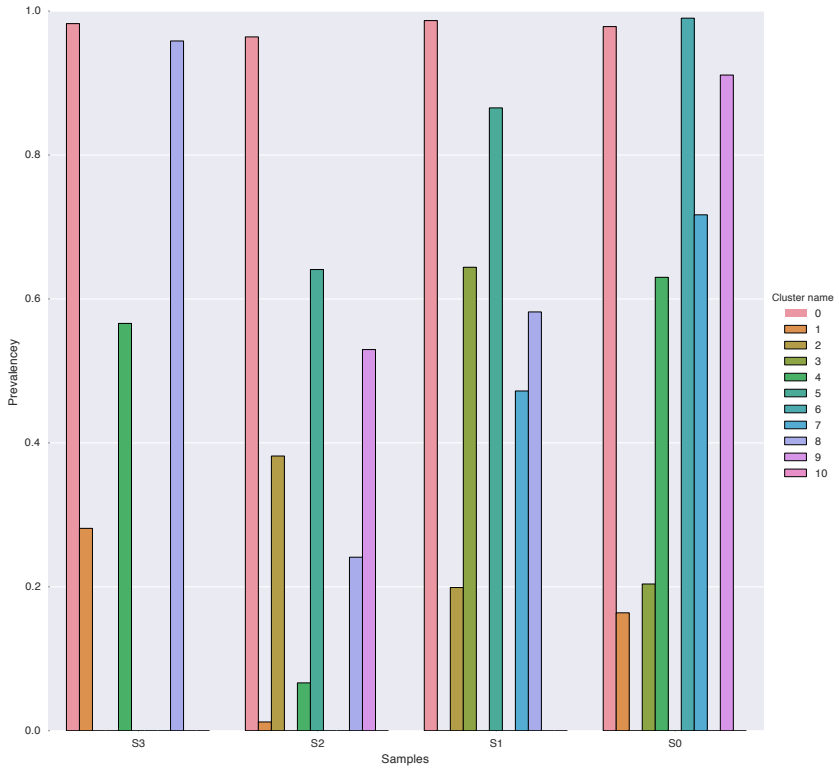

Figure 2.3: Distribution of the cellular prevalence of clusters across different samples in one random run.

cluster uniformly at random from $\{0,\ldots,K\}$, where cluster 0 represents the wild-type cluster and $\{1,\ldots,K\}$ are mutation clusters. For each locus, in each sample, the number of reads overlapping the locus (depth) is sampled from a Poisson distribution with mean d_m. Wildtype copy number is deterministically set to 2, and a copy number profile (major and minor copy number) is generated according to the following steps: The total copy number, c^t, is sampled uniformly at random from $\{1,\ldots,c_{\max}\}$. An integer number, c^b, is randomly (following a discrete Uniform distribution) picked from 1 to c^t, and c^a is defined as $c^a = c^t - c^b$. Lastly, the major copy number is set to the maximum of c^b and c^a; the minor copy number is set to the minimum of those two values. Then, corresponding to each cluster, the number of variant reads are sampled from the Beta-Binomial distribution described in

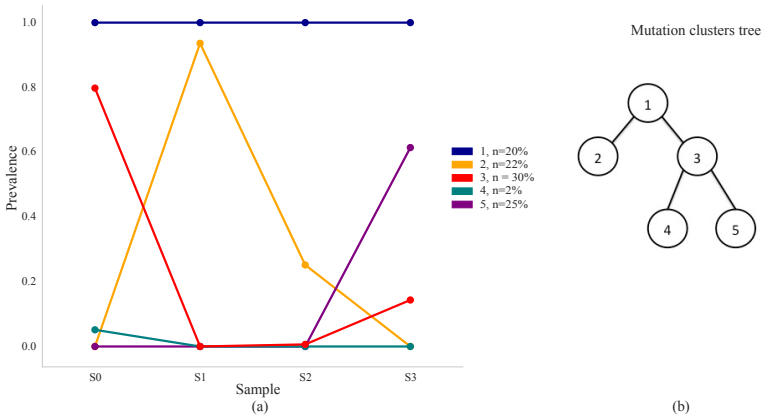

Figure 2.4: A clonal information example. (a) Cellular prevalences of underlying mutation clusters across multiple samples. (b) The relationship between mutation clusters is represented as a tree.

Equation (2.5) with precision parameter equal to 1000.

Synthetic data evaluation We simulated 10 synthetic data for 20000 loci from 4 samples of a patient, with 5 underlying clusters, including an ancestral cluster. The maximum copy number was 5, and error rate was 0.01. The average sequencing depth was assumed to be 100 for all samples.

To assess the performance and robustness of MuClone, we systematically shield MuClone from clonal information (Figure 2.5). In particular, the cellular prevalence information was perturbed by (i) adding noise to its value, or (ii) removing the cellular prevalence information of the clusters. The noise was generated from a normal distribution with mean zero and standard deviations: 0, 0.01, 0.1, and 0.2. The noise value, ν, was added to the cellular prevalence of the cluster, while bounding the resulting value between 0 and 1, that is,

$$\phi^{*z}_{m} = \min(\max(\phi^{z}_{m} + \nu, 0), 1),$$

where ϕ^{*z}_{m} and ϕ^{z}_{m} are the perturbed and original cellular prevalence of cluster z and sample m, respectively. The clusters which their clonal information was removed, were randomly chosen with equal probabilities. As expected, both sensitivity and

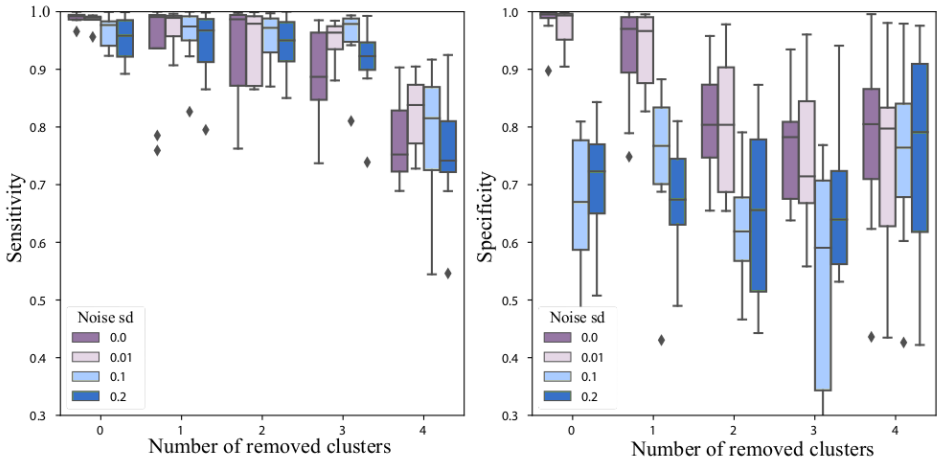

Figure 2.5: MuClone's performance with inaccurate clonal information: 10 synthetic datasets generated for 20000 loci, from 4 samples of a hypothetical patient, with 5 underlying clusters. The maximum copy number is 5, error rate is set to 0.01, and average sequencing depth is approximately 100. To assess performance, we add noise from a mean zero normal with standard deviation equal to 0 (dark purple), 0.01 (light purple), 0.1 (light blue), and 0.2 (dark blue) to the cellular prevalence and also remove the clonal information of different number of clusters. (a) Sensitivity and (b) Specificity of MuClone with parameters: wildtype prior $= 0.5$, $\Phi_T = 0.02$, error rate $= 0.01$, tumour content $= 0.75$, and precision parameter $= 1000$.

specificity were highest with complete and accurate clonal information; see Figure 2.5. This suggests that incorporating clonal information can improve mutation detection accuracy and gives evidence to support MuClone's approach. Furthermore, since the sensitivities were only marginally impacted by adding noise to the clonal information, MuClone should be able to cope with modest misspecification of the prior. However, specificity can decrease if the cellular prevalence is reduced to levels associated with the wildtype cluster and sensitivity can improve if adding noise increases the cellular prevalence to levels associated with a removed mutation cluster.

Naturally, accuracy was most severely impacted with reduced/corrupted clonal information; see Figure 2.5. For modest level of noise (noise standard deviation 0

and 0.01), the sensitivity and specificity of removing various numbers of clusters were compared through a Kruskal-Wallis test ($4e{-}5 \leq$ p-values $\leq 1e{-}4$) which shows that the change in performance due to clonal information is significant. In noiseless settings, the confidence interval for the difference (of zero and four removed clusters) in mean sensitivity and specificity are $[0.16, 0.26]$ and $[0.08, 0.32]$, respectively. When the noise standard deviation is equal to 0.01, these intervals are $[0.11, 0.21]$ and $[0.09, 0.36]$.

We also explored how sensitivity and specificity changes as a function of the wildtype prior and the threshold Φ_T used to distinguish the cellular prevalence cutoff of a mutation cluster. In Figure 2.6, we tested MuClone with wildtype prior values 0.5, 0.75, and 0.99, and with Φ_T values 0.001, 0.01, 0.02, 0.03, and 0.04. In the case that the wildtype prior equals 0.5, we assumed that a locus is equally likely to be a mutation or not (when no other information is provided). MuClone's sensitivity and specificity were near 1 for $\Phi_T = 0.02$ and wildtype prior equal to 0.5. As expected, with small values of Φ_T, the sensitivity and specificity decreased since it is difficult to distinguish between wildtypes and mutations with small cellular prevalences. The sensitivity also decreased for large values of Φ_T because mutations were miscalled as wildtypes. When the error rate was 0.01, and wildtype prior was 0.5, the optimal Φ_T was about 0.02. We used these values for the following experiments.

The performance of MuClone was tested with various tumour content (from 0.1 to 0.99) and different error rates (0.01 and 0.001); see Figure 2.7. For samples with tumour content greater than 0.5, the sensitivity and specificity remained close to 1. Sensitivity and specificity decreased to only about 0.9 when the tumour content in the sample was as low as 0.1. These results establish promising performance over different ranges of tumour content with different error rates (likely scenarios in real data). In addition, we also explored the performance of MuClone for samples with different coverage (mean depth): 30, 60 and 100; see Figure 2.8. Intuitively, the performance was higher when we had more coverage. Since MuClone leverages cellular prevalence information to improve the performance of mutation detection, the performance gain was noticeable when the variant allelic ratio resolution supports the given cellular prevalence resolution (in our analysis the cellular prevalence of mutations were greater than 0.02).

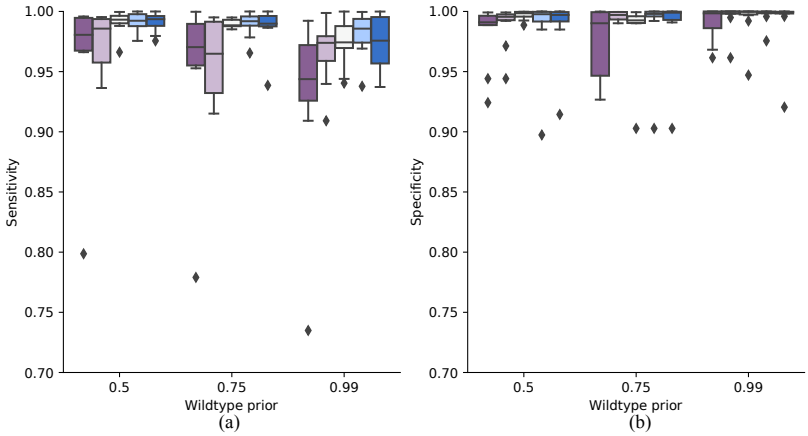

Figure 2.6: MuClone's performance as a function of wildtype prior: 10 synthetic datasets generated for 20000 loci, from 4 samples of a hypothetical patient, with 5 underlying clusters. The maximum copy number is 5 and error rate is set to 0.01, and average sequencing depth is approximately 100. We asses the performance for Φ_T equal to 0.001 (dark purple), 0.01 (light purple), 0.02 (white smoke), 0.03 (light blue), 0.04 (dark blue). (a) Sensitivity and (b) Specificity of MuClone with parameters: error rate = 0.01, tumour content = 0.75, and precision parameter = 1000.

Figure 2.9 demonstrates how well mutations were classified by MuClone. The input clonal information was perturbed by adding noise from zero mean normal distribution with with standard deviation 0.01 to simulate a more realistic scenario. In Figure 2.9(a), each bin (i,j) shows the fraction of mutations in cluster i that were classified into cluster j by MuClone. Figure 2.9(a) shows that 85% of mutations were classified into the correct cluster.

In order to show that the classification errors occurred between clusters with small phylogenetic distance, we define a misclassification index to quantify phylogenetic distance; calculated as

$$\text{Misclassification index} = \frac{\sum_{i \neq j} q_{(i,j)} \times \frac{dist_{(i,j)} - dist_i^{\min}}{dist_i^{\max} - dist_i^{\min}}}{\sum_{i \neq j} q_{(i,j)}}, \tag{2.13}$$

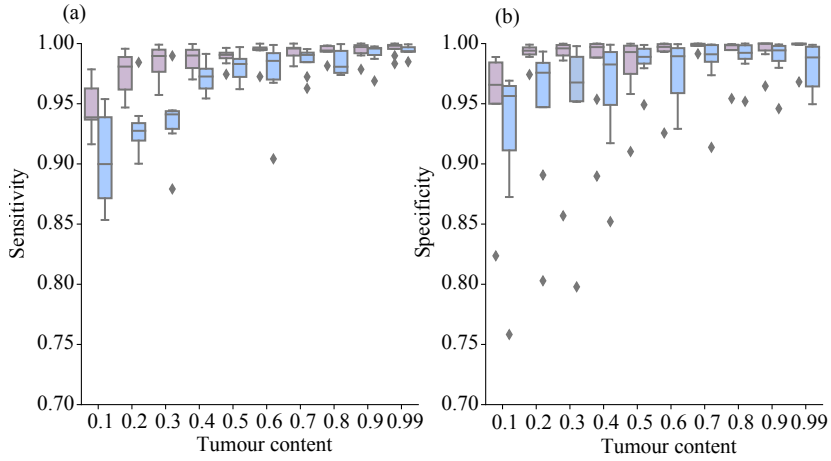

Figure 2.7: MuClone's performance as a function of tumour content: 10 synthetic datasets generated for 20000 loci, from 4 samples of a hypothetical patient, with 5 underlying clusters. The maximum copy number is 5, and average sequencing depth is approximately 100. We asses the performance for error rate equal to 0.001 (light purple) and 0.01 (light blue), and tumour content from 0.1 to 0.99. (a) Sensitivity and (b) Specificity of MuClone with parameters: wildtype prior = 0.5, ϕ_T = 0.02, and precision parameter = 1000.

where $q_{(i,j)}$ is the number of mutations in cluster i that have been classified into cluster j, and the Euclidean distance between the cellular prevalences of cluster i and j is denoted by $dist_{(i,j)}$. The distance of the closest and farthest cluster to cluster i is denoted by $dist_i^{\min}$ and $dist_i^{\max}$, respectively. In Figure 2.9(b), small misclassification indices demonstrate that misclassified mutations occur between close clusters. This can be interpreted as phylogenetically recently separated clusters.

2.4.2 Real data

Two real data sets with multiple samples for each patient were used to evaluate the performance of MuClone. The first data set was multiple whole genome sequencing data from 7 patients with high grade serous ovarian cancer. The second data set was multiple whole exome sequencing data from 8 patients with non-small-cell lung cancer (NSCLC).

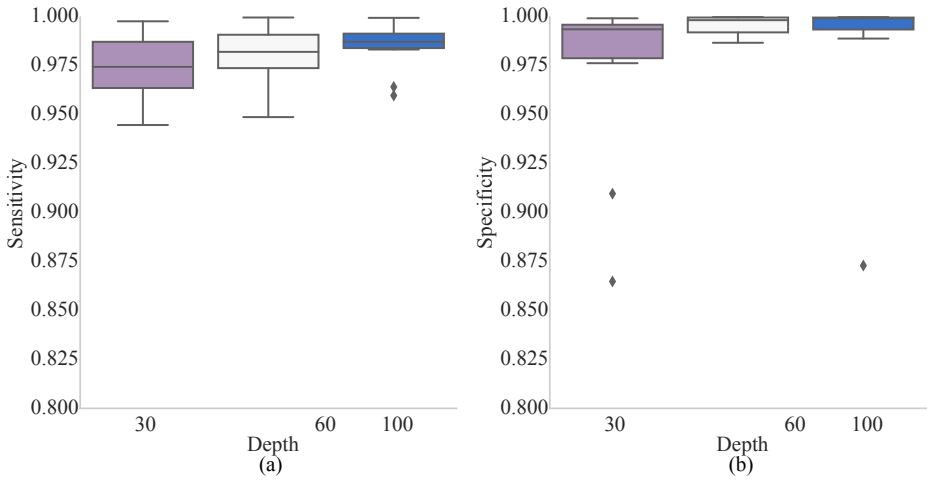

Figure 2.8: MuClone's performance as a function of different depth: 10 synthetic datasets generated for 20000 loci, from 4 samples of a hypothetical patient, with 5 underlying clusters. The maximum copy number is 5, error rate is set to 0.01 and average sequencing depth is approximately 100. (a) Sensitivity and (b) Specificity of MuClone with parameters: wildtype prior $= 0.5$, $\phi_T = 0.02$, error rate $= 0.01$, tumour content $= 0.75$, and precision parameter $= 1000$.

High grade serous ovarian cancer We tested MuClone's performance on whole genome sequencing data (with depth 30x) from multiple tumour samples surgically resected from high grade serous ovarian cancer patients [84]. The samples were obtained from different spatially distributed metastatic sites. Brief details about the number of samples for each patient, sample sites and the number of validated loci for each patient are shown in Table 2.1. Germline mutations were excluded from the list.

The copy number, tumour purity, and mutation cluster information for experimentally re-validated mutation status were taken from the phylogenetic study of high-grade serous ovarian cancer (see the supplementary note of the paper [84]). Mutation clusters were estimated with PyClone [100] on the deep targeted sequencing data (>1000x coverage) from the same samples and in three patients with accompanying single cell sequencing data (see Table S16 in the phylogenetic study of high-grade serous ovarian cancer paper [84]). Copy number and tumour

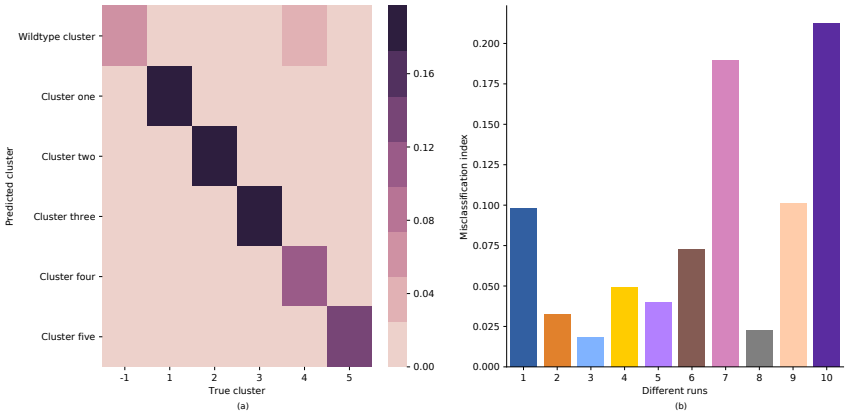

Figure 2.9: MuClone's classification performance: 10 synthetic datasets generated for 20000 loci, from 4 samples of a hypothetical patient, with 5 underlying clusters. The maximum copy number is 5, error rate is set to 0.01. (a) Bin (i,j) shows the fraction of mutations in cluster i that were classified into cluster j by MuClone. 85% of the mutations are classified correctly. (b) Misclassification index for 10 independent samples. MuClone with parameters: wildtype prior $= 0.5$, $\Phi_T = 0.02$, error rate $= 0.01$, tumour content $= 0.75$, and precision parameter $= 1000$.

purity estimates were calculated with the TITAN software [50]. In order to eliminate germlines, loci with variant nucleotides in the corresponding normal sample were removed from the dataset. Then, the performance of MuClone was benchmarked against Strelka [105] (v2.0.15), MutationSeq [31] (v4.3.7), MuTect [3], FreeBayes [39] (v1.2.0-2), MultiSNV [63] and naive MuClone. Naive MuClone is a version of MuClone where no clonal information is provided (that is, all mutations are from an ancestral cluster).

In Figure 2.10, the performance of MuClone is compared with other methods executed with default settings. For each patient, p, we assessed performance by averaging Youden's index, sensitivity, and specificity across different samples. For

Patient	Samples	#Validated positions	Anatomic samples
1	6	153	Right Ovary Site 1-4; Omentum Site 1; Small Bowel Site l
2	4	46	Omentum Site 1,2;Right Ovary Site 1,2
3	4	99	Right Ovary Site 1,2;Omentum Site 1; Left Ovary Site 2
4	5	69	Right Ovary Site 1-4;Right Pelvic Side Wall
7	3	59	Left Ovary Site 1; Brain Metastasis; Right Pelvic Mass
9	5	72	Right Ovary Site 1; Left Ovary Site 1; Omentum Site 1,2
10	4	136	Right Ovary Site 1-4

Table 2.1: Summary of high grade serous ovarian cancer data set [84].

patient p, with n_p samples, these are calculated as

$$\begin{aligned} \textit{Sensitivity}_p &= \frac{1}{n_p}\sum_{i=1}^{n_p} \textit{Sensitivity}_p^i, \\ \textit{Specificity}_p &= \frac{1}{n_p}\sum_{i=1}^{n_p} \textit{Specificity}_p^i, \\ \textit{Youden's index}_p &= \frac{1}{n_p}\sum_{i=1}^{n_p} \left(\textit{Sensitivity}_p^i + \textit{Specificity}_p^i - 1\right), \end{aligned} \tag{2.14}$$

where $\textit{Sensitivity}_p^i$, $\textit{Specificity}_p^i$ and $\textit{Youden's index}_p^i$ are the sensitivity, specificity and Youden's index of sample i and patient p, respectively. In aggregate, MuClone outperforms other methods by improving sensitivity without compromising specificity; see Figure 2.10. The performance (sensitivity, specificity and Youden's index) and the Receiver Operating Characteristic (ROC) curves for each patient are separately depicted in Figures 2.11 to 2.24. False negatives arise mainly because the WGS data is under-represented (the average depth of the WGS data is about 30x) and lacks variant alleles that are present in the targeted sequencing data. False positives arise due to erroneous signal from sequencing technical artefacts.

In Figure 2.10, Strelka, MutationSeq, MuTect and Naive MuClone have lower performance as they do not incorporate information across multiple samples. FreeBayes was run on multiple samples and germlines were removed manually, but since the method only considers tumour samples, it had the lower performance versus other methods.

To assess the performance of MuClone and MultiSNV, we conducted a two

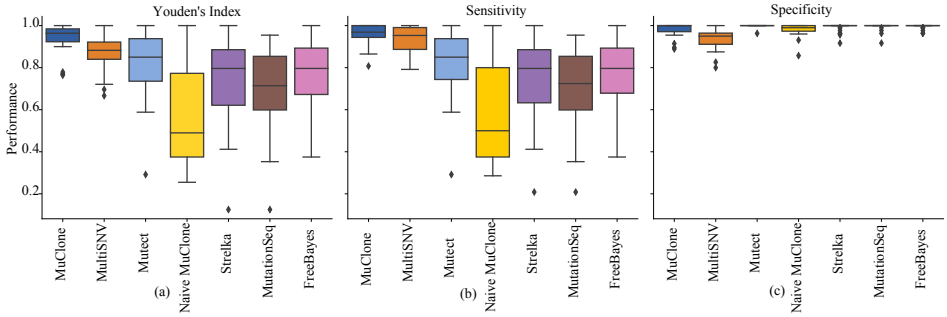

Figure 2.10: Performance comparison of different methods on whole genome sequencing data from patients with high grade serous ovarian cancer: (a) Youden's index, (b) Sensitivity, and (c) Specificity across different mutation detection methods (from left to right: MuClone (dark blue), MultiSNV (orange), MuTect (light blue), Naive MuClone (yellow), Strelka (purple), MutationSeq (brown), and FreeBayes (pink)). MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $= 1000$.

sided t-test for the difference in the mean of Youden's index evaluated on mutation calls from MuClone and MultiSNV. The 95% confidence interval is $[0.03, 0.1]$, with p-value equal to 0.0006; this shows that the difference is statistically significant. Importantly, MuClone improves sensitivity, enabling the detection of more mutations across the whole genome.

Figure 2.25 depicts the classification of mutations into clusters relative to the ground truth, as defined by running PyClone on the data (omitting singleton clusters [84]). Each bin (i,j) of Figure 2.25 shows the fraction of mutations in cluster i that were classified into cluster j by MuClone; 93% are correctly classified by MuClone. Moreover, we notice that misclassified mutations were classified into phylogenetically similar clusters (the misclassification index for patient 1 was 0.015).

Non-small-cell lung cancer We tested MuClone's performance on early-stage NSCLC samples from the TRACERx data set [60]. (See Table 2.2). To help obtain the clonal and subclonal census, multiple tumour regions for each patient were sequenced by Illumina HiSeq. We used the copy number, purity estimate, and the mutation cluster information available in the Supplementary Material of the paper [60]. In the TRACERX study, the cellular prevalence was calculated from

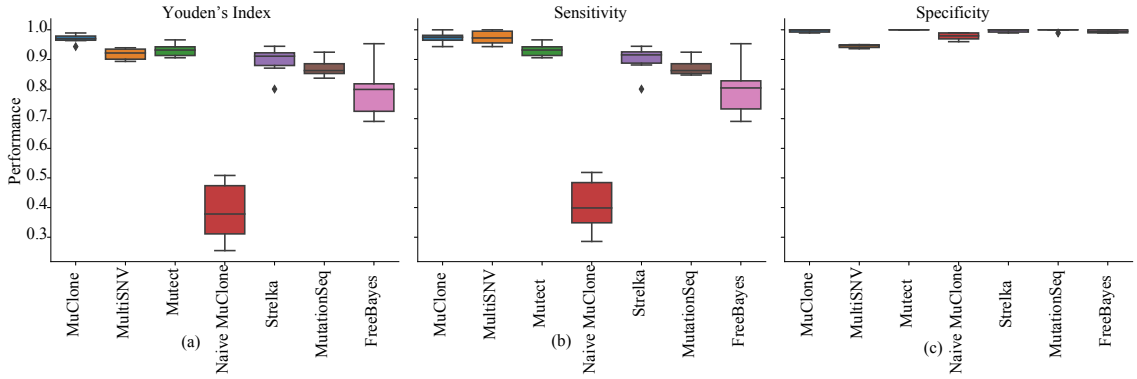

Figure 2.11: Performance comparison of different methods on whole genome sequencing data for patient 1 with high grade serous ovarian cancer: (a) Youden's index, (b) Sensitivity, and (c) Specificity across different mutation detection methods (from left to right: MuClone (dark blue), MultiSNV (orange), MuTect (light blue), Naive MuClone (yellow), Strelka (purple), MutationSeq (brown), and FreeBayes (pink)). MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $= 1000$.

the whole exome sequencing data on a set of stringent mutations that were selected from MuTect and VarScan2 results with post-processing. In addition, the TRACERx study added a few mutations to reduce missed subclonal mutations; see Supplementary Appendix of TRACERx study [60].

To compare the performance of MuClone with Strelka, MultiSNV, and MuTect, we randomly selected 8 patients with subclonal mutations from the TRACERx data set (see Table S2 Supplementary Appendix 1 of the paper [60]). The TRACERx study generated a re-validated and curated list of mutations for their analysis; see Supplementary Appendix 2 of TRACERx study [60]. The mutations with full copy number information across all 8 patients were used as ground truth to evaluate performance.

We evaluated the false negative rates of mutation calling across several methods; see Table 2.3. Altogether, out of 7238 mutations, MuClone missed 475 mutations while Strelka, MultiSNV and MuTect missed 7205, 5720, and 1086 mutations, respectively. Hence, borrowing statistical strength, as done in MuClone, across samples likely increases sensitivity to real mutations.

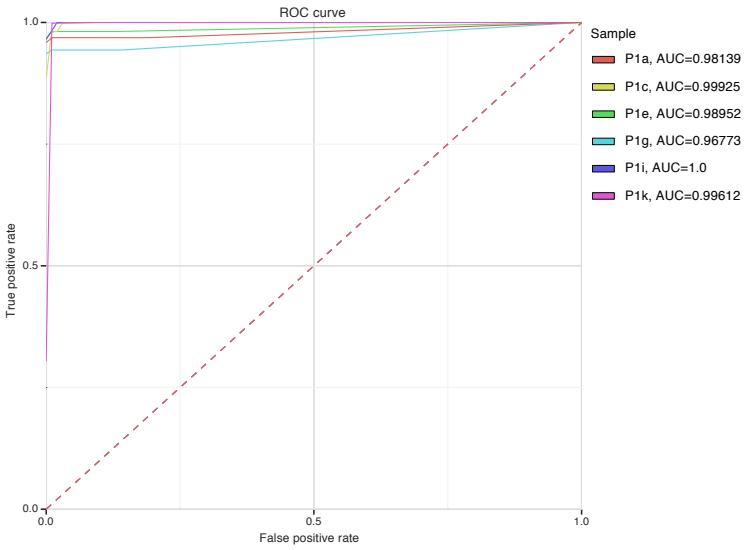

Figure 2.12: MuClone's Roc curves and the area under the curve (AUC) for patient 1. MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $=$ 1000.

We next ran MuClone, MultiSNV and MuTect on the whole exome data from multiple samples of 8 patients to ascertain specificity. We note that MuClone removes reads with mapping quality less than 5 and for positions that have (i) a variant nucleotide in a normal sample, (ii) more than 40% filtered basecalls (A basecall is filtered if more than 3 mismatches occur between the read and the reference within a window of 20 bases on each side of the site.); or (iii) more than 75% of the reads that cross the site have deletions in any of the samples [105]. For exome sequencing data, mutations were called if the corresponding MuClone mutation probability is greater than 0.9. The other methods were executed with default settings. The total number of calls and the number of common calls between different methods (restricted to positions with copy number information) at the patient level is depicted in Figure 2.26. A high degree of variation across callers was observed.

Altogether, MuClone called 13556 mutations while MultiSNV and MuTect called 31374 and 11915, respectively. MultiSNV output the largest number of

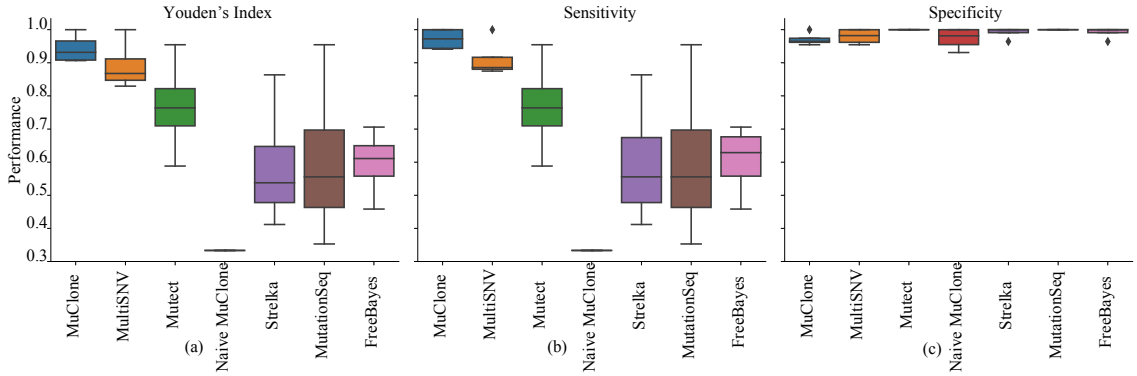

Figure 2.13: Performance comparison of different methods on whole genome sequencing data for patient 2 with high grade serous ovarian cancer: (a) Youden's index, (b) Sensitivity, and (c) Specificity across different mutation detection methods (from left to right: MuClone (dark blue), MultiSNV (orange), MuTect (light blue), Naive MuClone (yellow), Strelka (purple), MutationSeq (brown), and FreeBayes (pink)). MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $= 1000$.

Patient	Samples	Cancer type
CRUK0003	4	Adenocarcinoma
CRUK0004	4	Adenocarcinoma
CRUK0005	4	Adenocarcinoma
CRUK0013	5	Adenocarcinoma
CRUK0062	7	Squamous-Cell Carcinoma
CRUK0063	5	Squamous-Cell Carcinoma
CRUK0065	6	Squamous-Cell Carcinoma
CRUK0094	4	Other

Table 2.2: Summary of the NSCLC data set [60].

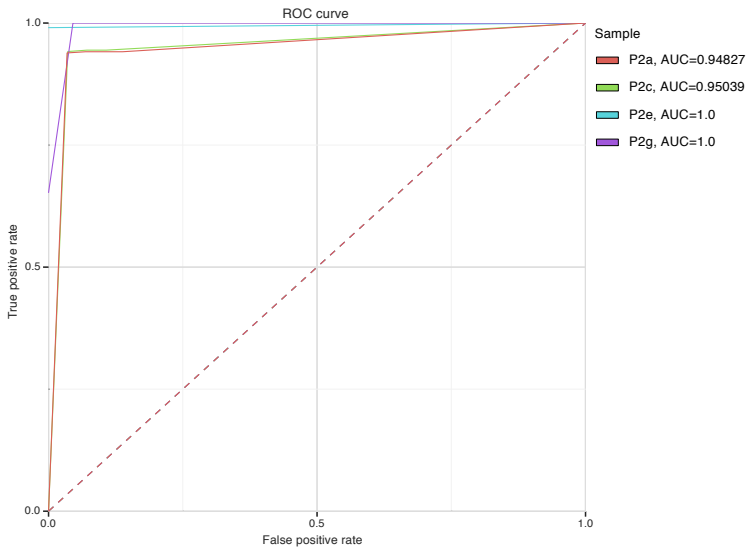

Figure 2.14: MuClone's Roc curves and the area under the curve (AUC) for patient 2. MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $=$ 1000.

Patient	MuClone	MultiSNV	MuTect	Strelka
CRUK0003	52	350	60	430
CRUK0004	16	188	36	240
CRUK0005	212	1736	236	2040
CRUK0013	26	490	270	540
CRUK0062	30	469	42	609
CRUK0063	75	445	40	510
CRUK0065	60	1902	342	2640
CRUK0094	4	140	60	196

Table 2.3: Total number of false negative calls across multiple samples of non-small cell lung cancer patients for different algorithms.

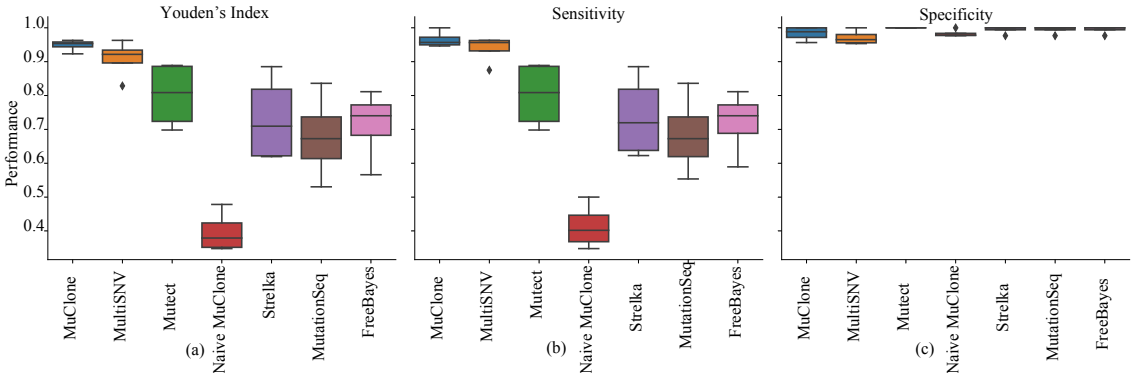

Figure 2.15: Performance comparison of different methods on whole genome sequencing data for patient 3 with high grade serous ovarian cancer: (a) Youden's index, (b) Sensitivity, and (c) Specificity across different mutation detection methods (from left to right: MuClone (dark blue), MultiSNV (orange), MuTect (light blue), Naive MuClone (yellow), Strelka (purple), MutationSeq (brown), and FreeBayes (pink)). MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $= 1000$.

calls in all of the samples, while MuTect and MuClone output similar number of calls. Figure 2.26 also demonstrates the mutations used in the TRACERx study and their overlap with the mutation calls in different methods. The set of mutations overlapping between MuClone and TRACERx is most similar; this suggests that the increase in sensitivity conferred by MuClone does not come at the expense of specificity.

We also explored the performance of MuClone when clonal information differs in the number of input clusters or the value of the cellular prevalence; see Figure 2.27. We perturbed the value of the cellular prevalences (estimated by PyClone) by adding noise from a mean zero normal distribution with different standard deviations: 0, 0.01, 0.1, and 0.2. We see that MuClone is robust to slight changes of cellular prevalence values. We also shielded MuClone from different fractions of the clonal information and that decreased performance more than adding noise. In general, this result shows that more accurate clonal information provides better mutation detection.

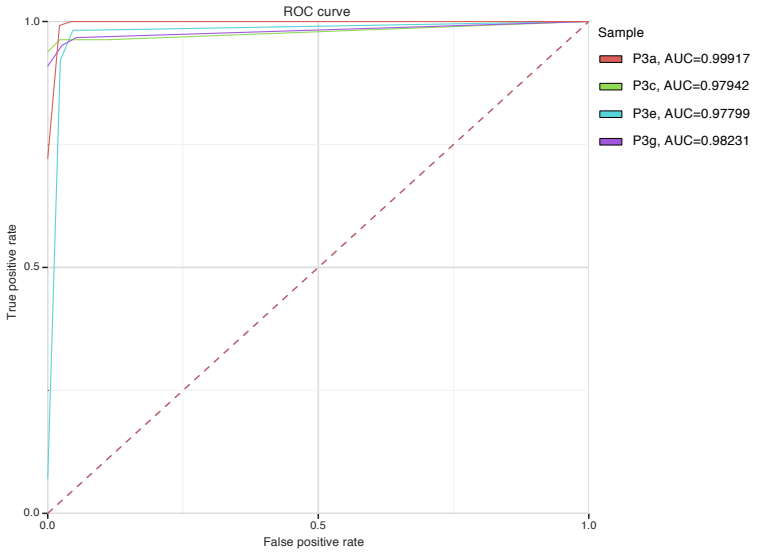

Figure 2.16: MuClone's Roc curves and the area under the curve (AUC) for patient 3. MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $=$ 1000.

2.5 Conclusion

We studied the use of clonal information for the purpose of somatic mutation detection and classification in multi-sample whole genome sequencing data. The proposed statistical framework uses the clusters cellular prevalences and copy number information for detection and classification of low prevalence mutations. Our proposal, MuClone, outperformed other popular mutation detection tools, while providing the added benefit of classifying whole genome sequencing mutations into biologically relevant groups. Both synthetic and real data results showed that using the cellular prevalences of tumour clusters can improve mutation detection sensitivity. Importantly, our results suggest improvement in sensitivity can be achieved without compromising specificity.

Since the accuracy of detecting mutations can affect the performance of phylogenetic analysis, we suggest improvement from using MuClone will impact the field of multi-region sequencing for cancer evolution studies. As the field matures,

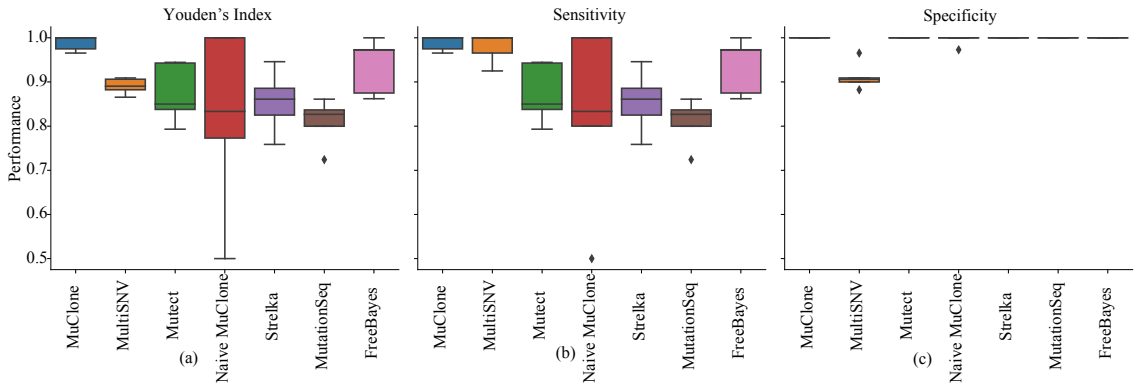

Figure 2.17: Performance comparison of different methods on whole genome sequencing data for patient 4 with high grade serous ovarian cancer: (a) Youden's index, (b) Sensitivity, and (c) Specificity across different mutation detection methods (from left to right: MuClone (dark blue), MultiSNV (orange), MuTect (light blue), Naive MuClone (yellow), Strelka (purple), MutationSeq (brown), and FreeBayes (pink)). MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $= 1000$.

we expect that the method presented here will be incorporated into more analytically comprehensive modelling of whole genome sequencing data when multiple samples are used to infer properties of clonal dynamics. Next steps are in developing a unified iterative algorithm that alternates between identifying the phylogenetic structure of the constituent clones comprising each tumour sample, and detection of mutations leveraging the new phylogenetic structure.

As sequencing costs continue to decrease (e.g. with Illumina's NovoSeq platform), multi-sample whole genome sequencing of tumours will continue to proliferate (e.g. rapid autopsy program) as a viable experimental design. Thus, MuClone will be an asset in the arsenal of analytical methods deployed to interpret evolutionary properties of cancer and to gain insights into clonal dynamics in time and space.

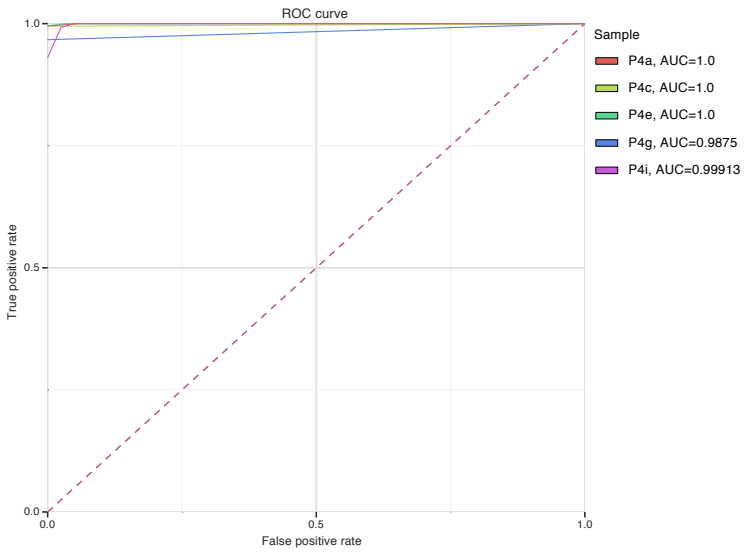

Figure 2.18: MuClone's Roc curves and the area under the curve (AUC) for patient 4. MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $=$ 1000.

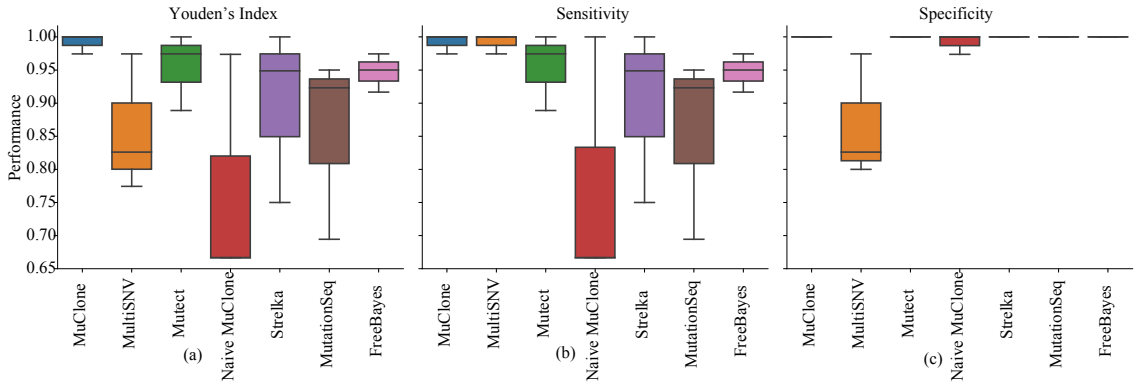

Figure 2.19: Performance comparison of different methods on whole genome sequencing data for patient 7 with high grade serous ovarian cancer: (a) Youden's index, (b) Sensitivity, and (c) Specificity across different mutation detection methods (from left to right: MuClone (dark blue), MultiSNV (orange), MuTect (light blue), Naive MuClone (yellow), Strelka (purple), MutationSeq (brown), and FreeBayes (pink)). MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $= 1000$.

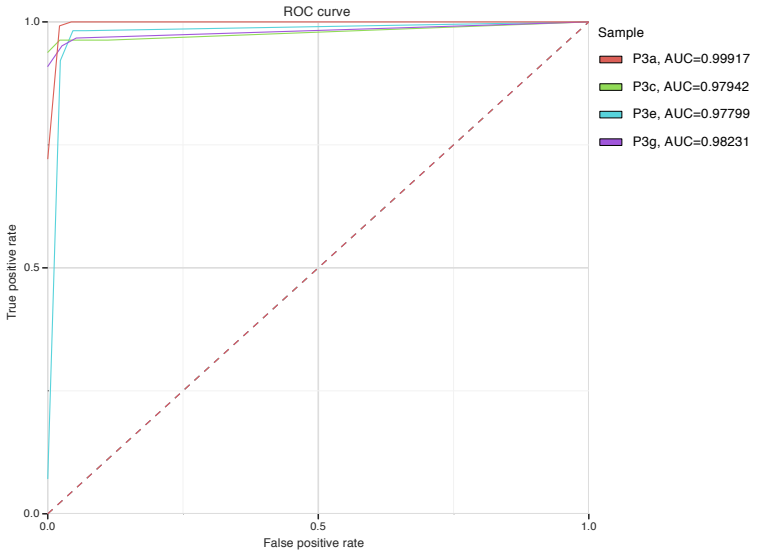

Figure 2.20: MuClone's Roc curves and the area under the curve (AUC) for patient 7. MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $=$ 1000.

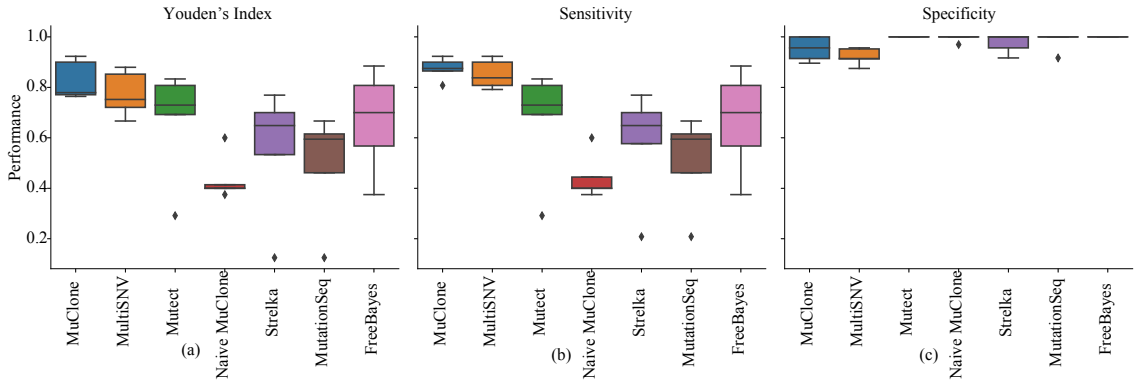

Figure 2.21: Performance comparison of different methods on whole genome sequencing data for patient 9 with high grade serous ovarian cancer: (a) Youden's index, (b) Sensitivity, and (c) Specificity across different mutation detection methods (from left to right: MuClone (dark blue), MultiSNV (orange), MuTect (light blue), Naive MuClone (yellow), Strelka (purple), MutationSeq (brown), and FreeBayes (pink)). MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $= 1000$.

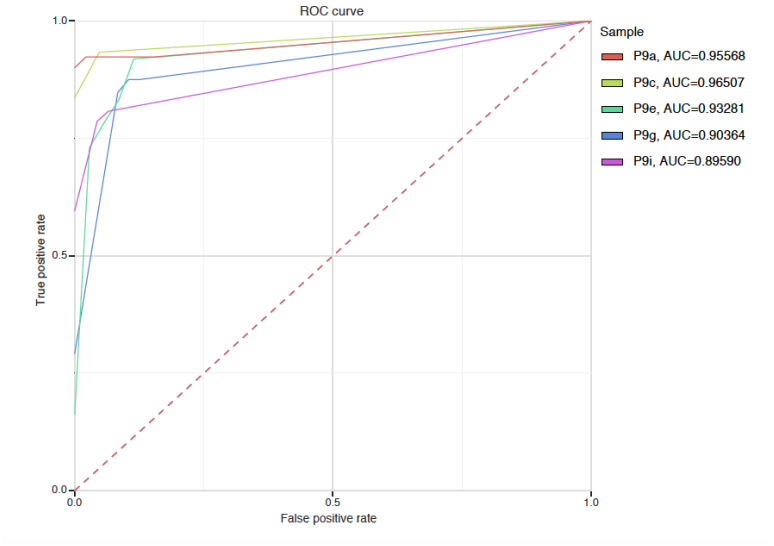

Figure 2.22: MuClone's Roc curves and the area under the curve (AUC) for patient 9. MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $=$ 1000.

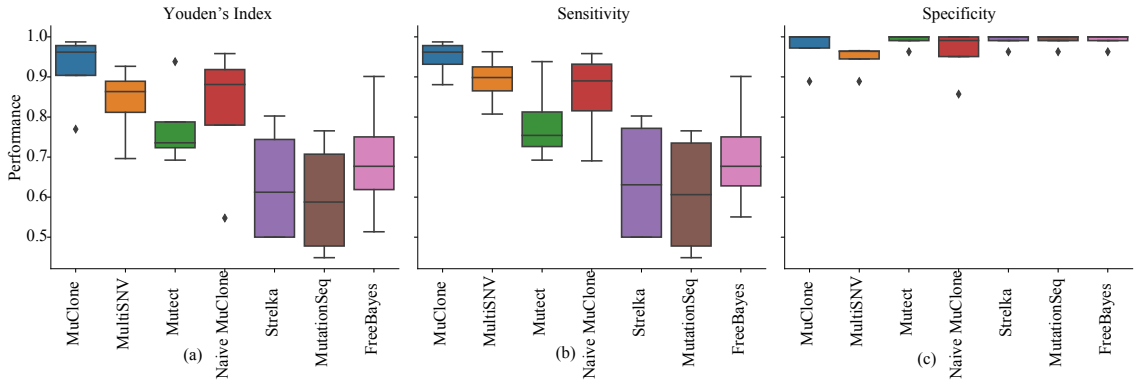

Figure 2.23: Performance comparison of different methods on whole genome sequencing data for patient 10 with high grade serous ovarian cancer: (a) Youden's index, (b) Sensitivity, and (c) Specificity across different mutation detection methods (from left to right: MuClone (dark blue), MultiSNV (orange), MuTect (light blue), Naive MuClone (yellow), Strelka (purple), MutationSeq (brown), and FreeBayes (pink)). MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $= 1000$.

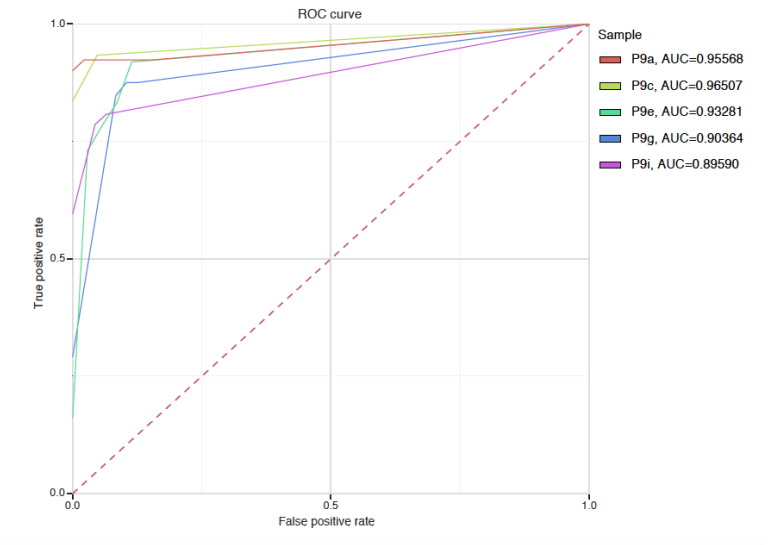

Figure 2.24: MuClone's Roc curves and the area under the curve (AUC) for patient 10. MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, tumour content $= 0.75$, error rate $= 0.01$, and precision parameter $=$ 1000.

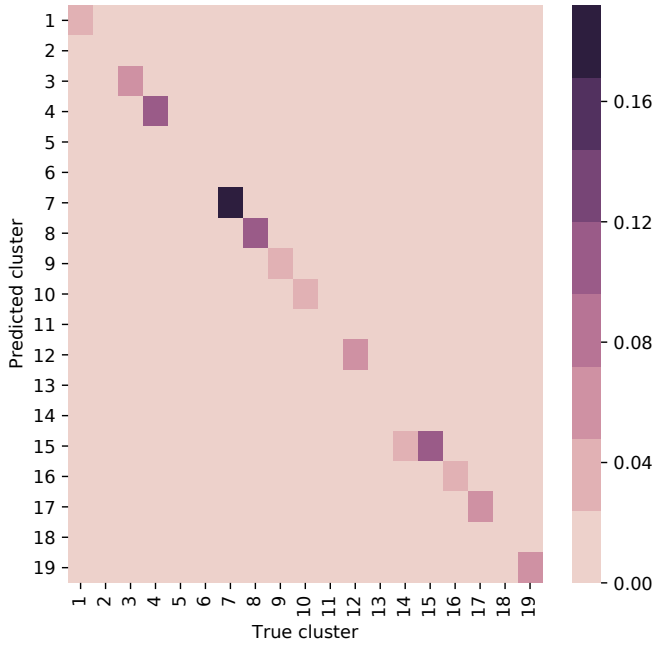

Figure 2.25: Classification of 153 mutations of patient 1 with high grade serous ovarian cancer across 6 samples. Bin (i,j) shows the fraction of mutations in cluster i that were classified into cluster j by MuClone. MuClone parameters are: wildtype prior $= 0.5$, $\Phi_T = 0.02$, error rate $= 0.01$, tumour content $= 0.75$, and precision parameter $= 1000$. 93% of the elements are diagonal.

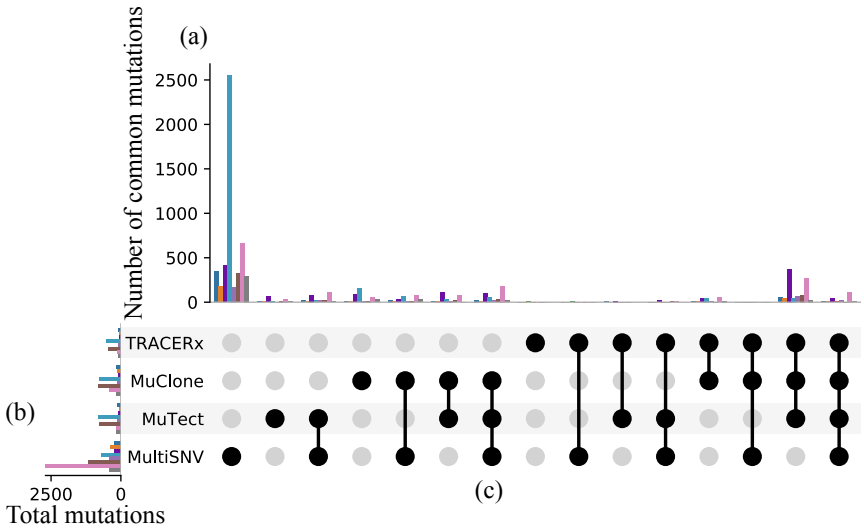

Figure 2.26: Comparison of detected mutations from MultiSNV, MuTect, MuClone, and TRACERx on whole exome sequencing data from non-small cell lung cancer patients: CRUK0003 (dark blue), CRUK0004 (orange), CRUK0005 (dark purple), CRUK0013 (light blue), CRUK0062 (light purple), CRUK0063 (brown), CRUK0065 (pink), and CRUK0094 (grey). To illustrate overlap in detected mutations from all combinations of different methods, in panel (a) the number of mutations called in all selected methods, but not in any other method, are plotted for each patient. Dark circles in the columns of panel (c) indicate selected methods. For example, the first column in panel (c) indicates mutations that are only called by MultiSNV (all circles are grey except the one corresponding to MultiSNV). Panel (b) displays the total number of mutations called for each method. MuClone removes reads with mapping quality less than 5 and positions which have (i) a variant nucleotide in a normal sample, (ii) more than 40% filtered basecalls; or (iii) more than 75% of the reads that cross the site have deletions in any of the samples. Other methods executed with default settings. See UpSet tool for additional visualization details [80].

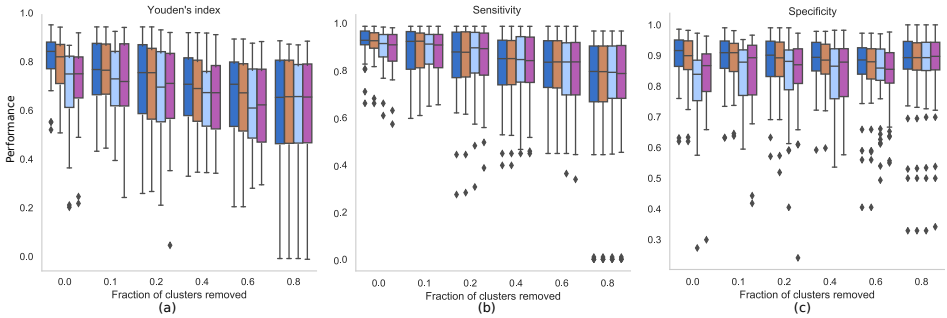

Figure 2.27: MuClone's performance with inaccurate clonal information: The mutations with full copy number information across all 8 non-small cell lung cancer patients are used as ground truth. To asses the performance, we add noise from a zero mean normal with standard deviation equal to 0 (dark blue), 0.01 (orange), 0.1 (light blue), and 0.2 (purple) and also remove the clonal information of different fractions of clusters. (a) Youden's index, (b) Sensitivity, and (c) Specificity of MuClone with different fraction of clusters information removed.

Chapter 3

Single cell somatic mutation detection through incorporation of the underlying phylogeny

"Nothing truly valuable can be achieved except by the unselfish cooperation of many individuals."

- Albert Einstein

This chapter introduces a new model, CellMutScope, to detect mutations (SNVs) in each cell based on the joint analysis of a large number (1000 - 10,000 or more) of whole genome single cell sequencing data. The model detects SNVs incorporating the underlying CNV phylogenetic tree. Moreover, we expand the utility of Corrupt model that infers the underlying phylogeny of a group of cells based on CNV data [1]. We frame how to incorporate SNV data in the Corrupt model (i) to infer the underlying phylogeny using SNVs with or without CNV data, or (ii) to extend the input underlying phylogenetic tree with the SNVs placed on the tree. The extended tree discloses the cell-to-cell genomic variability information and shows similar cells that are grouped together. Sections 3.1 and 3.2 define the problem and review the existing approaches, respectively. Section 3.3 reviews the fast scalable Corrupt model that infers the underlying CNV tree. Section 3.4 frames how to

incorporate SNV data into the Corrupt model and introduces CellMutScope. Section 3.5 synthetically evaluates the results in terms of SNV calls and the phylogeny when SNV data is also used.

3.1 Introduction

Single cell genomics is critical for deciphering the evolutionary process (clonal dynamics) of cancer during cancer initiation, progression, and during ongoing evolution as a patient undergoes treatment. Bulk sequencing protocols sequence a mixture of cells; computational methods are then applied to deconvolve the data and to infer the underlying clones. However, it is challenging to resolve minor clones from bulk data. The challenge is because of the presence of sequencing errors, and also because of sampling issues related to intra-tumour heterogeneity in solid tumours [141]. Single cell sequencing data promises to reveal the genomic details of single cells and the underlying tumour clones. Genomic variations (single nucleotide mutations, copy number variation, and structural variants) contain a record of ancestral relationships between the cells, and the relationship is reflected in a phylogenetic tree. Although single cell sequencing technologies provide a direct data to infer the phylogeny (to study the clones and their dynamics), experimental challenges in capturing nuclei of individual cells and difficulties in sequencing them with even coverage have delayed the large scale analysis of single cell genomes [41, 75].

The low quality of single cell sequencing data (SCS data) is rooted in the fact that a single cell only contains about 6-7 pg of genomic DNA while the typical amount of DNA needed to construct a library in next generation sequencing (NGS) platforms is on the order of 1000 pg [21]. This means that the genome or parts of the genome needs to be amplified before sequencing [102, 109]. However, amplification of DNA is a stochastic process correlated with genome accessibility and GC content [11]. In order to generate a sufficient quantity of DNA for sequencing, several protocols based on whole genome amplification (WGA) have been suggested [9, 40, 56, 75, 89, 128, 145]. While WGA permits sequencing to higher coverage depth and breadth [37, 128, 141], it introduces coverage biases. In order to overcome the coverage bias issue, Zahn et al. have proposed a scalable

micro-fluidics based protocol, called Direct Library Preparation (DLP) [141]. This protocol requires no pre-amplification and has lower biases compared to WGA protocols. In the DLP protocol DNA of individual cells are first fragmented into short sequencing inserts, and then through a few cycles of PCR, the sequencing adaptors and index barcodes are added. This process produces exact duplicates that can be removed computationally. The breadth and depth of DLP single cell sequencing data coverage is lower than the bulk whole genome sequencing data. However, the resultant 'bulk-equivalent' data generated by aggregating cells has uniform coverage similar to the bulk data [141]. The uniform coverage accommodates CNV calling, but the sparsity of the data introduces difficulties for SNV calling. In addition, the isolation of cells may fail and multiple cells (mostly double cells) may be sequenced together.

The DLP data which is composed of thousands of cells, is sequenced shallowly (< 1x) across the genome. Following the uniform coverage of the DLP data, Zahn et al. proposed a tool for CNV calling [141]. Bouchard et al. then utilized the CNV data to infer the underlying phylogeny in the Corrupt [1].

In this chapter, we frame how to utilize SNV data in the Corrupt model (i) to infer the underlying phylogeny, or (ii) to extend the underlying CNV tree by adding the SNVs on the tree. The extended tree (that includes both CNVs and SNVs) shows the genomic variability between individual cells and reveals cells that are grouped together. Incorporation of SNV data leads to a more distinguished identification of minor clone structures, and this helps resolve the phylogeny in more detail. The resultant phylogenetic tree contains a record of the ancestral relationships between cells.

The computational complexity of the Corrupt model tree inference at each iteration is linear with the number of cells and traits. Missing data in Bayesian models requires a higher number of iterations before convergence [72]. This is also the case for the above framework. We observe an undesirable increase in computational time with missing rate in the SNV data. Therefore, it is reasonable to use only the CNV data in tree inference step and add the SNVs afterwards.

We propose a model, CellMutScope, to detect SNVs in every single cell. We posit that incorporating the CNV data and the underlying phylogenetic tree across a large number of whole genome single cell sequencing data help overcome the

sparsity of single cell sequencing data. In contrast to MuClone's output which is a vector of length N (number of loci), CellMutScope output is a matrix of size $M \times N$ (M is the number of cells[1]. MuClone calls a mutation to be present if it is present in any of the cells in the sample, while CellMutScope calls the mutation status of each cell individually. The overall consolidated framework that unites SNVs and CNVs over a phylogeny will improve the understanding of the tumour clonal dynamics. It also helps understanding of how the combination of different genetic variations (CNVs and SNVs) evolve to the observed results in the current population.

3.2 Literature review

Most of the existing methods utilize SCS data obtained from WGA protocols. While these protocols are efficient in preparing enough material for the sequencing, they have high rate of allelic drop-out (about 10% -20%). The high allelic drop-out rate is the result of random amplification of only one allele at a heterozygous genotype site. This complicates the SNV detection and results in a coverage bias. Moreover, existing methods are only focused on the analysis of SNVs across sites with enough coverage (roughly more than 5x).

Monovar addressed the issue of low and uneven data coverage by pooling the sequencing data across cells with an assumption that the sites are independent [140]. Monovar uses data with coverage between 6x to 24x, and it skips the ones with less coverage. It has the asymptotic complexity of $O(M^3)$ for genotyping M single cells. SCcaller identifies SNVs for each cell while accounting for local allelic amplification biases [33]. The model requires the data coverage about 5x, and it requires the single nucleotide polymorphisms (SNPs) data. This data is not always available. Furthermore, it can recover mutations neither from drop-out events nor from loss of heterozygosity [33]. Conbase and LiRA utilize read-phasing to correct for errors and allelic drop-outs [12, 54]. Conbase unlike LiRA performs joint variant calling across the population of cells. The use of read phasing decreases false discovery rate and increases specificity. However, such analysis is only possible for regions in proximity to SNPs. This prevents calling SNVs across all of the bases covered by the reads. SCIΦ simultaneously detects

[1]M represents the number of samples in Chapter 2. A sample in this chapter represents a cell.

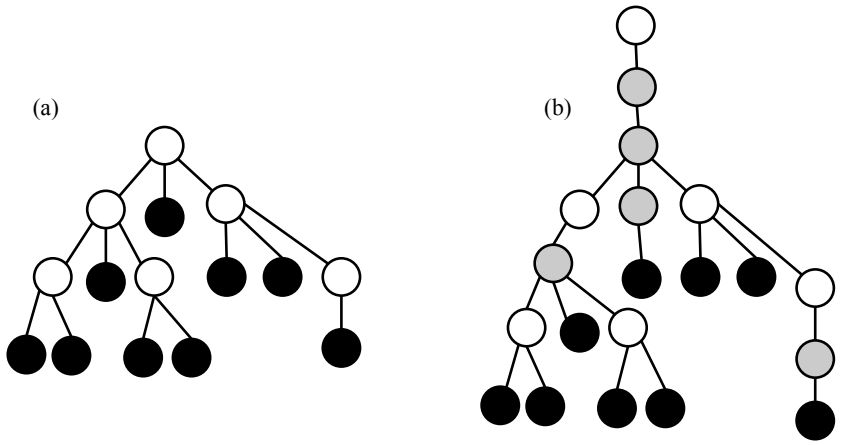

Figure 3.1: (a) A schematic view of CNV tree where the black nodes represent cells, and the blank nodes represent CNV traits. (b) A schematic view of the tree after adding SNV loci. The grey nodes represent SNV traits.

SNVs and infers the underlying phylogeny. Although SCIΦ claims to detect mutations in single cells with very low or even no coverage, it requires at least two cells to show an alternative nucleotide count of at least three. The runtime complexity of SCIΦ is $O(X \times max(MN, C_u))$, where C_u is the number of unique coverage values of the experiment for M cells and N loci, and X is the number of iterations [115].

We propose a method, CellMutScope, that calls SNVs given the underlying phylogeny. Its computational complexity is $O(M+N)$. With this scalable framework each SNV locus is quickly processed. Inferring the tree from CNV data itself has a computational complexity of $O(M+N)$ per iteration [1]. Therefore, the unified framework provides a comprehensive genomic data analysis tool for large scale low-coverage genome-wide single cell data.

Figure 3.1(a) shows a schematic view of a CNV tree in which the black nodes represent cells, and white nodes represent CNV traits. Figure 3.1(b) shows the tree after running CellMutScope and adding SNV loci (The grey nodes represent SNV traits).

3.3 Corrupt model

Most of the copy number analysis methods start with dividing the genome into non-overlapping segments. Each segment is assumed to be homogeneous in copy number, and the number of overlapping reads is proportional to the DNA copy number at the corresponding segment. The issue of coverage bias in SCS data obtained from WGA, hinders development of methods to utilize CNV data. Navin et al. and Wang et al. infer the copy number of single cells from WGA based SCS data [89, 128]. The DLP data is more uniform and has lower coverage bias compared to WGA protocols. Therefore, the data is more suitable for detecting copy number alterations. Zahn et al. developed a method that exploits DLP data to infer the copy numbers [141]. Following this work, Bouchard et al. have proposed Corrupt model that incorporates CNV data and infers the underlying phylogeny with $O(M+N)$ per iteration [1].

Corrupt is a Bayesian framework for inferring phylogenetic trees using copy number information from large scale low-coverage genome-wide single cell data. A schematic view of the tree is depicted in Figure 3.1(a). In order to simplify the site dependencies, Corrupt encodes CNV data into binary change points. The copy number data is indexed by genomic bins, and each bin has an integer copy number value. In the binary change point representation of data, the data is indexed by the 'space' between two adjacent bins; and it is a change point if there is a copy number change between two bins. Let us define n as a locus between two adjacent bins, and m as an index for the cells. The copy number value of the bins before and after locus n in cell m is denoted by $\mathfrak{c}^{-}_{m,n}$ and $\mathfrak{c}^{+}_{m,n}$, respectively. Let M and N denote the disjoint set of cells and loci, respectively. The trait value for cell $m \in M$ and locus $n \in N$ is defined as:

$$x_{m,n} = \mathbf{1}[\mathfrak{c}^{-}_{m,n} \neq \mathfrak{c}^{+}_{m,n}] = \begin{cases} 1, & \text{if } \mathfrak{c}^{-}_{m,n} \neq \mathfrak{c}^{+}_{m,n}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

The matrix $x = (x_{m,n})$ is not observed directly, since $\mathfrak{c}^{-}_{m,n}$, $\mathfrak{c}^{+}_{m,n}$ are unobserved. Therefore, the proposed observation probability model is $p(y|x, \theta^{CNV})$, where y is the observed data and θ^{CNV} encodes the error rates in copy number calls. Let us define $\theta^{CNV}_{n} = (r^{FP}_{m,n}(\theta), r^{FN}_{m,n}(\theta))$ to represent the locus specific false positive

and false negative rates of cell m. The error rates are $r^{FP}_{m,n}(\theta) \in (0,1)$ and $r^{FN}_{m,n}(\theta) \in (0,1)$. For simplicity, the likelihood probability distribution denoted by $p(y^{CNV}_{m,n}|x^{CNV}_{m,n}, \theta^{CNV})$ is written as:

$$p(y_{m,n}|x_{m,n},\theta) = p(y_{m,n}|x_{m,n}, r^{FP}_{m,n}(\theta), r^{FN}_{m,n}(\theta)) = e^{r^{FP}_{n}(\theta), r^{FN}_{n}(\theta)}_{x_{m,n}, y_{m,n}}, \qquad (3.2)$$

where the error matrix is

$$e^{r^{FP}_{m,n}(\theta), r^{FN}_{m,n}(\theta)} = \begin{bmatrix} 1 - r^{FP}_{m,n}(\theta) & r^{FP}_{m,n}(\theta) \\ r^{FN}_{m,n}(\theta) & 1 - r^{FN}_{m,n}(\theta) \end{bmatrix}.$$

In the simplest case, these parameters are globally constant across all cells and loci, where

$$\begin{aligned} r^{FP}_{m,n}(\theta) &= r^{FP} \\ r^{FN}_{m,n}(\theta) &= r^{FN}. \end{aligned}$$

Therefore,

$$p(y|x,\theta) = \prod_{m\in M}\prod_{n\in N} p(y_{m,n}|x_{m,n}, r^{FP}(\theta), r^{FN}(\theta)), \qquad (3.3)$$

$$= (r^{FP})^{n^{FP}} (r^{FN})^{n^{FN}} (1 - r^{FP})^{n^{N} - n^{FN}} (1 - r^{FN})^{n^{P} - n^{FP}}, \qquad (3.4)$$

where n^{FP} and n^{FN} denotes the number of false positive and false negative instances, respectively. They are computed as:

$$\begin{aligned} n^{FP} &= \sum_{m\in M}\sum_{n\in N} \mathbb{I}[x_{m,n} = 0, y_{m,n} = 1], \\ n^{FN} &= \sum_{m\in M}\sum_{n\in N} \mathbb{I}[x_{m,n} = 1, y_{m,n} = 0], \end{aligned}$$

where, n^{P} and n^{N} are the number of positive and negative instances, with

$$n^{p} = \sum_{m\in M}\sum_{n\in N} \mathbb{I}[y_{m,n} = 1],$$

and

$$n^N = |M||N| - n^P.$$

Initially, the parameters are estimated using the whole matrix, then at each iteration, the parameters are updated by comparing the row or the column that is modified compared to the previous iteration (each iteration or move modifies only one row or column of the matrix x). In the Corrupt inference model, the error rate parameters can be set to local or global.

3.3.1 Probability model

An underlying assumption in the Corrupt is the validity of perfect phylogeny in the latent data ($x_{m,n}$), and not in the observed data ($y_{m,n}$). The model can be described by a two-step generative process: (i) sampling a mutation tree, and (ii) sampling cell assignments.

Sampling a mutation tree Let $\mathscr{T}$ denote a set of t^z trees spanning node set $\mathscr{V}$, where $\mathscr{V}$ includes root node v^* plus a node for each of the N loci. The root node v^* puts an implicit direction on the edges of the tree. There is a directed path from node/locus n to n' in t^z, if and only if the trait indexed by n is emerged in a cell prior to the trait indexed by n'.

Sampling cell assignments Assign each cell to a node in t^z. If cell m is assigned to locus n, then the cell has all the traits on the shortest path from locus n to root v^*. If a cell is assigned to node v^*, then it doesn't have any of the traits on the tree.

3.3.2 Inference

The goal is to explore the space of trees. Corrupt model proposes a move with computational complexity of $O(N+M)$ to explore the exponentially large number of neighbours. The move consists of four steps:

1. Remove a locus (trait) node $n \in N$ from the tree.
2. Select a node v that is not a cell node on the tree from which node n is removed.
3. Pick a (possibly empty) subset from the children of v.

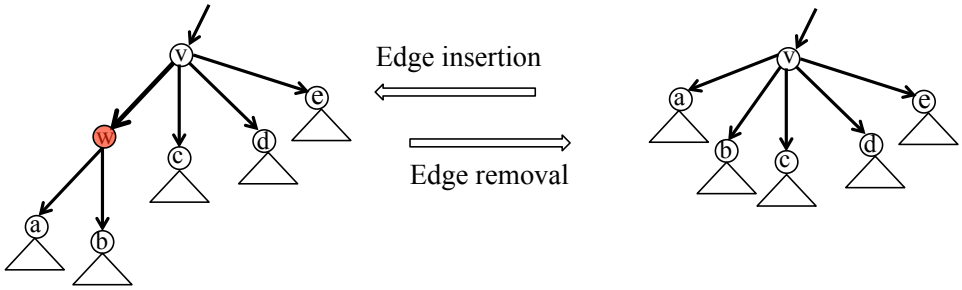

Figure 3.2: A schematic of removing(adding) a node from(to) the tree. The bold edge and node w are removed, and the nodes connected to w is connected to the parent node v. The triangles represent a subtree rooted at an specific internal node. [modified figure from [1]]

4. Add a node w to represent locus n by adding an edge from v to w. Move the subset of selected nodes from previous step to hang from w (See Figure 3.2.).

The above steps are followed when sampling or maximizing. In order to select a node, the following probability is efficiently calculated:

$$\bar{\rho}_v = \frac{\rho_v}{\sum_{\tilde{v}\in R}\rho_{\tilde{v}}}, \tag{3.5}$$

where $\tilde{v} \in R$, and $R = \{v^*\} \bigcup N \backslash \{n\}$. The value of ρ_v is

$$\rho_v = \sum_{t\in\mathscr{N}_v^n\left(t_{\backslash n}\right)} p(t)p(y|x(t),\theta), \tag{3.6}$$

where $\mathscr{N}_v^n\left(t_{\backslash n}\right)$ denotes the set of neighbour trees of the tree t with locus node n removed. After further simplification, $\bar{\rho}_v$ is estimated as

$$\begin{aligned}\bar{\rho}_v &= \frac{\rho_v}{\sum_{\tilde{v}\in R}\rho_{\tilde{v}}}, \\ &= \frac{\left(\frac{\prod_{v_i\in Children(v)}(p^0_{v_i}+p^1_{v_i})}{p^0_v}\right)}{\sum_{\tilde{v}\in R}\left(\frac{\prod_{\tilde{v}_i\in Children(\tilde{v})}(p^0_{\tilde{v}_i}+p^1_{\tilde{v}_i})}{p^0_{\tilde{v}}}\right)},\end{aligned} \tag{3.7}$$

where for all other $v \in R$ and $s \in \{0,1\}$,

$$p_v^s = \prod_{v'' \in Children(v)} p_{v''}^s. \tag{3.8}$$

For all nodes m corresponding to a cell and $s \in \{0,1\}$, the likelihood probability function is defined as:

$$p_{m,n}^s = p(y_{m,n}|x_{m,n}, s, r_n^{FP}(\theta), r_n^{FN}(\theta)). \tag{3.9}$$

After the attachment node v is selected, for each child v_i of v, a Bernouli distribution with success parameter $\frac{p_{vi}^1}{p_{vi}^1+p_{vi}^0}$ is considered. The success corresponds to moving v_i into a child of the newly re-introduced node w. The realization of the Bernoulli variables can be sampled from for approximating the posterior distribution. Also, it can be maximized for adding a node to the tree.

3.3.3 Approximation of the posterior distribution

The posterior distribution is

$$\pi(t,\theta) \propto p(t)p(\theta)p(y|x(t),\theta).$$

Here the deterministic function is denoted by $x(t)$. Given $t^z \in \mathscr{T}$, the matrix X is a deterministic function obtained by setting $x_{m,n} = 1$ if node m is a descendant of node n in t^z, and zero otherwise. Different methods are used to approximate the posterior function:

1. A direct Markov chain Monte Carlo (MCMC) scheme.
2. A non-reversible Parallel Tempering (PT) algorithm [133].
3. The sequential change of measure (SCM) scheme from [28] with the adaptive scheme from [144].

3.3.4 Summary via minimum Bayes estimator

Minimum Bayes risk estimator is used to summarize the posterior distribution:

$$\underset{t\in\mathscr{T}}{\operatorname{argmin}} \sum_{t'\in\mathscr{T}} \int L(t,t')\pi(t,\mathrm{d}\theta), \tag{3.10}$$

where $t' \in \mathscr{T}$, and

$$L(t,t') = \sum_{n\in N}\sum_{m\in M} |x_{m,n}(t) - x_{m,n}(t')|.$$

This minimization problem itself can be solved through a greedy procedure. More details can be found in [1].

3.4 Method

The Corrupt model utilizes CNV data for inferring the underlying phylogeny. In Section 3.4.1 we frame how to use SNV data (allelic read counts) with or without CNV data (i) to infer the underlying phylogeny, or (ii) to add SNVs on the underlying CNV tree. With this framework, we can infer the underlying phylogeny tree using both CNV and SNV data. This latter approach (adding the SNV loci to the tree) helps skip the computationally expensive inference step (when both CNVs and SNVs are used in tree inference) and infers the extended tree (with both CNVs and SNVs) in a time efficient manner. Next, we propose a new model, CellMutScope, for detecting SNVs in every single cell through adding the SNV loci to the underlying tree. Incorporating the underlying phylogeny and CNV data help overcome the high missing rate of the data and detects SNVs in every individual cell. The resultant extended tree and the SNV calls discloses further detail about the clonal structure of the cells that were not distinguishable based on only CNV data.

3.4.1 Incorporating single nucleotide data in the Corrupt model

The Corrupt model takes binary representation of copy number data. This data depends on the unobserved copy number variables to infer the phylogeny. The proposed observation probability model is defined in Equation 3.8. In this section, we posit an observation probability model for SNV data to extend the utility of the

Corrupt model, where the observed data is $y_{m,n}^{SNV} = (b_{m,n}, d_{m,n}, c_{m,n})$. For locus n in cell m, the values of $d_{m,n}$, $b_{m,n}$ and $c_{m,n}$ respectively represent: the total number of reads covering the locus, the number of reads with a variant allele (compared to the reference genome), and copy number. The observed data, in particular $b_{m,n}$ depends on the mutation status of cell m at locus n, denoted by $x_{m,n}^{SNV}$. The error model is $\theta^{SNV} = (\varepsilon_{FP}, \varepsilon_{FN})$, where ε_{FP} and ε_{FN} are false positive rate and false negative rate, respectively. Therefore

$$q_{m,n}^{s} = p(y_{m,n}^{SNV} | x_{m,n}^{SNV}, \theta^{SNV}) = p(b_{m,n} | d_{m,n}, c_{m,n}, x_{m,n}^{SNV} = s, \theta^{SNV}), \tag{3.11}$$

where $d_{m,n}$ and $c_{m,n}$ are given. The likelihood probability of cell node m is denoted by $q_{m,n}^{s}$, where $s \in \{0,1\}$. For $s = 1$, $q_{m,n}^{s}$ reflects the likelihood of cell m being mutated at locus n; and for $s = 0$, $q_{m,n}^{s}$ reflects the likelihood of cell m not being mutated at locus n.

The probability $q_{m,n}^{s}$ follows a mixture of binomial distributions depending on all possible genotype states of locus n at cell m. Given the copy number $c_{m,n}$, the possible genotype states are $\mathscr{G} = \{A\ldots A, AA\ldots B, A\ldots BB, \ldots, B\ldots B\}$, where each element has a length equal to $c_{m,n}$. For example, the genotype AAB refers to a genotype with one variant allele B and 2 reference allele A. For each genotype state g_i, where i indexes the elements of $\mathscr{G}$, the mean parameter of the corresponding binomial distribution is denoted by $\xi_{m,n}^{i}$:

$$\xi_{m,n}^{i} = \begin{cases} \frac{\mathscr{B}(g_i)}{c_{m,n}}, & 1 \leq \mathscr{B}(g_i) < c_{m,n}, \\ 1 - \varepsilon_{FP}, & \mathscr{B}(g_i) = c_{m,n}, \\ \varepsilon_{FP}, & \text{otherwise}, \end{cases} \tag{3.12}$$

where $\mathscr{B}(g_i)$ represents the number of variant allele of genotype g_i. Therefore, for $s = 1$,

$$q_{m,n}^{1} = p(b_{m,n} | d_{m,n}, c_{m,n}, x_{m,n}^{SNV} = 1, \theta^{SNV}) \tag{3.13}$$

$$= \sum_{i=1}^{c_{m,n}} p(g_i)[\xi_{m,n}^{b_{m,n}} (1 - \xi_{m,n})^{d_{m,n} - b_{m,n}}] \tag{3.14}$$

$$+ \varepsilon_{FN}[\varepsilon_{FP}^{b_{m,n}} (1 - \varepsilon_{FP})^{d_{m,n} - b_{m,n}}].$$

The value of $p(g_i)$ equals $\frac{1-\varepsilon_{FN}}{c_{m,n}}$, and ε_{FN} represents the error due to mutation loss or tree errors.

If the mutation status of cell m at locus n is wildtype (i.e mutation is not present), then the possible genotype states should not have any variant allele. The only possible genotype state is $\{A \ldots A\}$. The mean parameter of the binomial distribution equals ε_{FP} (false positive rate). Therefore,

$$q^0_{m,n} = p(b_{m,n}|d_{m,n}, c_{m,n}, x^{SNV}_{m,n} = 0, \varepsilon_{FP}). \tag{3.15}$$

Algorithm 1 summarizes the assumed model to generate the data where $g_{m,n}$ is the true genotype of locus n of cell m.

With the proposed probability model for SNVs, we can incorporate both SNV data and CNV data to infer the underlying tree phylogeny in the Corrupt model. Therefore, Equation 3.4 is updated as:

$$p(y|x,\theta) = \prod_{m\in M}\prod_{n\in N_{CNV}} p(y^{CNV}_{m,n}|x^{CNV}_{m,n}, \theta^{CNV}) \prod_{n\in N_{SNV}} p(y^{SNV}_{m,n}|x^{SNV}_{m,n}, \theta^{SNV}), \tag{3.16}$$

where M and N is the disjoint set of cells and loci, respectively. N includes both CNV and SNV traits, and $N = N_{CNV} + N_{SNV}$. Therefore, Equation 3.7 (when locus n is removed from the tree) can be rewritten in the following form:

$$\bar{\rho}_v = \frac{\left(\frac{\prod_{v_i\in Children(v)}(\gamma^0_{v_i}+\gamma^1_{v_i})}{\gamma^0_v}\right)}{\sum_{\bar{v}\in R}\left(\frac{\prod_{\bar{v}_i\in Children(\bar{v})}(\gamma^0_{\bar{v}_i}+\gamma^1_{\bar{v}_i})}{\gamma^0_{\bar{v}}}\right)}, \tag{3.17}$$

where γ^s_v, for $s \in \{0,1\}$ is:

$$\gamma^s_v = \begin{cases} p^s_v, & \text{If } n \text{ represents a CNV loci,} \\ q^s_v, & \text{If } n \text{ represents a SNV loci.} \end{cases}$$

Algorithm 1: Data's underlying generative model

Input: $d_{m,n}, c_{m,n}, \theta^{SNV}, s$;
Output: $b_{m,n}$;
if $s = 1$ **then**
  draw $l \sim \text{Bernoulli}(\varepsilon_{FN})$;
  if $l = 0$ **then**
    draw $\mathscr{B}(g_{m,n}) \sim \text{Uniform}([1, c_{m,n}])$;
    if $\mathscr{B}(g_{m,n}) = c_{m,n}$ **then**
      draw $b_{m,n} \sim \text{Binomial}(d_{m,n}, 1-\varepsilon_{FP})$;
    else
      $\xi_{m,n} = \frac{g_{m,n}}{c_{m,n}}$;
      draw $b_{m,n} \sim \text{Binomial}(d_{m,n}, \xi_{m,n})$;
    end
  else
    draw $b_{m,n} \sim \text{Binomial}(d_{m,n}, \varepsilon_{FP})$;
  end
else
  draw $b_{m,n} \sim \text{Binomial}(d_{m,n}, \varepsilon_{FP})$;
end
return $b_{m,n}$;

For $v \in R = \{v^*\} \bigcup N \backslash \{n\}$, and $s \in \{0,1\}$, the value of q_v^s is

$$q_v^s = \prod_{v'' \in Children(v)} q_{v''}^s. \tag{3.18}$$

For the cell nodes that are the leaves of the tree $q_v^s = q_{m,n}^s$.

3.4.2 Detection of SNVs at every individual cell

Given the underlying tree (denoted by t) and the read counts data (denoted by y), here the goal is to calculate the posterior probability of $x_{m,n}$, where $x_{m,n}$ denotes the mutation status of locus n at cell m. According to the underlying tree, cells are attached to the trait nodes, and they are leaves of the tree (See Figure 3.1 (b)). The

joint probability distribution of $x_{m,n}$, y and t can be written as:

$$p(x_{m,n},y,t) = \sum_{v\in R}\sum_{t'\in\mathscr{N}_v^n(t\backslash n)} p(x_{m,n},t',y) \tag{3.19}$$

$$= \sum_{v\in R}\sum_{t'\in\mathscr{N}_v^n(t\backslash n)} p(x_{m,n}|t')p(y|t')p(t'), \tag{3.20}$$

where R is the set of all loci nodes in the tree (including the root) excluding locus n. The joint probability distribution is calculated as

$$p(x_{m,n}=1,y,t) = \sum_{v\in\mathscr{P}(m,t)}\sum_{t'\in\mathscr{N}_v^n(t\backslash n)} p(y|t')p(t'), \tag{3.21}$$

where $x_{m,n}$ is a deterministic function of t' which belongs to $\mathscr{N}_v^n(t\backslash n)$. The value of $x_{m,n}$ is 1 if locus n is an ancestor of cell m, otherwise it is 0. The set $\mathscr{P}(m,t)$ denotes all the nodes on the path from cell m to the root of the tree (including the root and excluding the cell m node). An example of the path on an imaginary tree is depicted in Figure 3.3. The nodes coloured in green belong to $\mathscr{P}(m,t)$. Therefore, the posterior probability distribution of $x_{m,n}=1$ yields

$$p(x_{m,n}=1|y,t) = \frac{p(x_{m,n}=1,y,t)}{p(y,t)} = \frac{\sum_{v\in\mathscr{P}(m,t)}\sum_{t'\in\mathscr{N}_v^n(t\backslash n)} p(y|t')p(t')}{p(y,t)}. \tag{3.22}$$

Rewriting Equation 3.22 assuming uniform probability distribution for $p(t')$ yields:

$$\begin{aligned}
p(x_{m,n}=1|y,t) &\propto \sum_{v\in\mathscr{P}(m,t)}\sum_{t'\in\mathscr{N}_v^n(t\backslash n)} p(y|t'),\\
&= \sum_{v\in\mathscr{P}(m,t)}\sum_{t'\in\mathscr{N}_v^n(t\backslash n)}\prod_{n'\in N}\prod_{m'\in M} p(y_{m',n'}|t'),\\
&= \sum_{v\in\mathscr{P}(m,t)}\sum_{t'\in\mathscr{N}_v^n(t\backslash n)}\prod_{\substack{n'\in N\\ n'\neq n}}\prod_{m'\in M} p(y_{m',n'}|t')\prod_{m'\in M} p(y_{m',n}|t'),\\
&= K_1\sum_{v\in\mathscr{P}(m,t)}\sum_{t'\in\mathscr{N}_v^n(t\backslash n)}\prod_{m'\in M} p(y_{m',n}|t'),\\
&= K_1\sum_{v\in\mathscr{P}(m,t)}\sum_{t'\in\mathscr{N}_v^n(t\backslash n)}\prod_{m'\in M_{\backslash v}} p(y_{m',n}|t')\prod_{m'\in M_v} p(y_{m',n}|t'),
\end{aligned}$$

where N denotes the set of all trait nodes, M denotes the set of all cell nodes, M_v denotes the cells that are a descendant of node v, and $M_{\setminus v}$ denotes the cells that are a not descendant of node v. The product of the likelihood contributions for non-descendant nodes can be calculated by taking the product of q^0_m for all cells, divided by the ones that are descendant of v:

$$\prod_{m' \in M_{\setminus v}} q^0_{m'} = \frac{q^0_{v^*}}{q^0_v}.$$

Therefore:

$$p(x_{m,n} = 1|y,t) \propto K_1 \sum_{v \in \mathscr{P}(m,t)} \frac{q^0_{v^*}}{q^0_v} \sum_{t' \in \mathscr{N}_v^n(t \setminus n)} \prod_{m' \in M_v} p(y_{m',n}|t'). \tag{3.23}$$

The likelihood contribution of descendant cells can be re-indexed by a binary vector of $\mathbf{s} = (s_1, s_2, \ldots, s_k)$, where $s_i \in \{0,1\}$, and $s_i = 1$ if the child v is to be moved into a child of the node n. The value of k denotes the number of children of v. The i^*th child of v which is on the path from node v to cell m is called v^*_i. This implies $s_{i^*} = 1$ (See Figure 3.3). Therefore:

$$\sum_{t' \in \mathscr{N}_v^n(t \setminus n)} \prod_{m' \in M_v} p(y_{m',n}|t') = q^1_{v^*_m} \sum_{s_1=0}^{1} \sum_{s_2=0}^{1} \cdots \sum_{s_{i-1}=0}^{1} \sum_{s_{i+1}=0}^{1} \cdots \sum_{s_k=0}^{1} \prod_{\substack{i=1 \\ i \neq i^*}}^{k} q^{s_i}_{v_i}. \tag{3.24}$$

Rewriting Equation 3.23 using Equation 3.24 yields:

$$\begin{aligned}
p(x_{m,n} = 1|y,t) &\propto K_1 \sum_{v \in \mathscr{P}(m,t)} \frac{q^0_{v^*}}{q^0_v} q^1_{v^*_m} \sum_{s_1=0}^{1} \sum_{s_2=0}^{1} \cdots \sum_{s_{i-1}=0}^{1} \sum_{s_{i+1}=0}^{1} \cdots \sum_{s_k=0}^{1} \prod_{\substack{i=1 \\ i \neq i^*}}^{k} q^{s_i}_{v_i}, \\
&= K_1 \sum_{v \in \mathscr{P}(m,t)} \frac{q^0_{v^*}}{q^0_v} q^1_{v^*_m} \prod_{\substack{i=1 \\ i \neq i^*}}^{k} (q^0_{v_i} + q^1_{v_i}), \\
&= K_1 \sum_{v \in \mathscr{P}(m,t)} \frac{q^0_{v^*}}{q^0_v} \frac{\prod_{i=1}^{k} (q^0_{v_i} + q^1_{v_i})}{(q^0_{v_{i^*}} + q^1_{v_{i^*}})} q^1_{v_{i^*}}, \\
&= K_1 q^0_{v^*} \sum_{v \in \mathscr{P}(m,t)} \frac{q^1_{v_{i^*}}}{q^0_v (q^0_{v_{i^*}} + q^1_{v_{i^*}})} \prod_{i=1}^{k} (q^0_{v_i} + q^1_{v_i}).
\end{aligned} \tag{3.25}$$

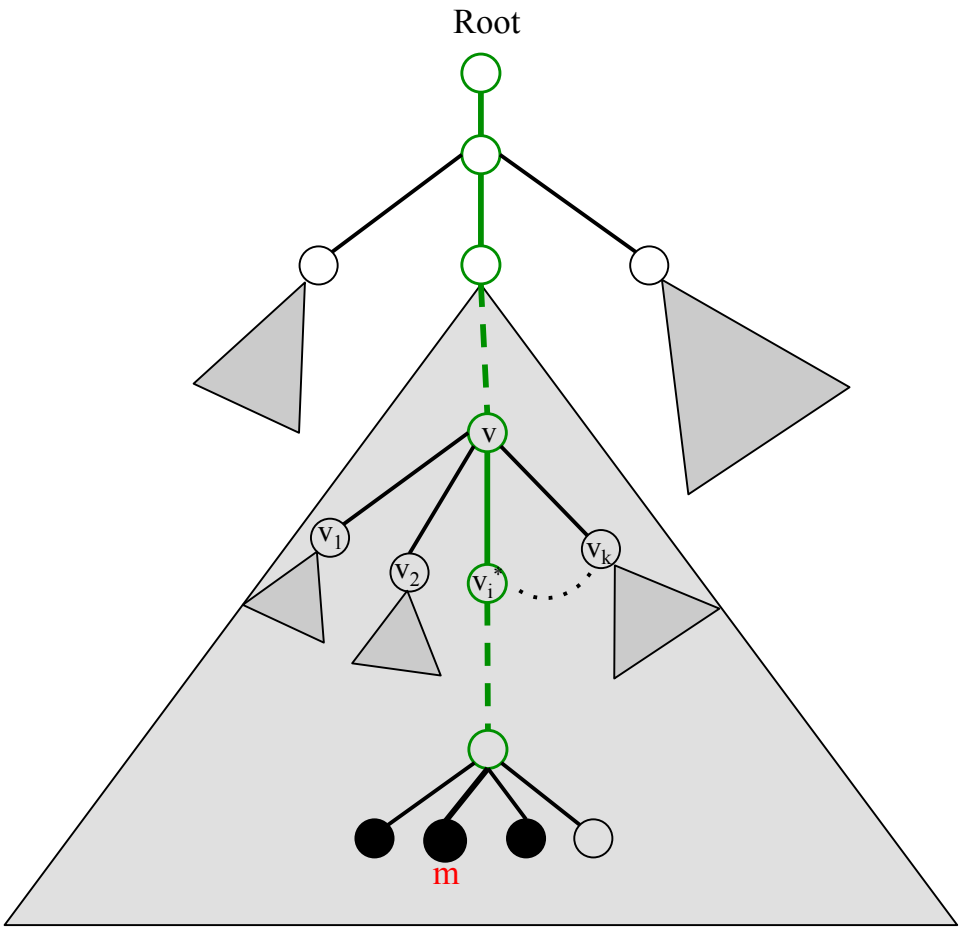

Figure 3.3: A schematic view of the underlying tree inferred from CNV and SNV loci across multiple cells. Black and white nodes represent cells and loci, respectively. The grey triangle represents a subtree rooted at a node. It includes all of the nodes and edges in the subtree.

With this approach, the posterior probability of $x_{m,n}$ is calculated based on placing locus n at different places on the tree. This considers different subset of cells (sub-trees of the underlying tree) to calculate the posterior probability of $x_{m,n}$.

3.4.3 Model's asymptotic complexity

The computational complexity of 3.25 is $O(MN)$ with M number of cells and N number of loci. In order to reduce the complexity of calculating $p(x_{m,n} = 1|y,t)$ for each locus per cell, $\mathscr{P}'(m,t)$ is defined to denote the nodes sitting on the path from root to cell m, excluding the root node and including the cell m node. Then,

$$q^*{}_v = \prod_{i=1}^{k} (q^0_{v_i} + q^1_{v_i}). \tag{3.26}$$

Therefore,

$$K_1 q^0_{v^*} \sum_{v \in \mathscr{P}(m,t)} \frac{q^1_{v_{i^*}}}{q^0_v (q^0_{v_{i^*}} + q^1_{v_{i^*}})} \prod_{i=1}^{k} (q^0_{v_i} + q^1_{v_i}) = K_1 q^0_{v^*} \sum_{v \in \mathscr{P}'(m,t)} \frac{q^1_v}{(q^0_v + q^1_v)} \frac{q^*{}_{parent(v)}}{q^0_{parent(v)}}.$$

Calculating $p(x_{m,n} = 1|y,t)$ with a recursive approach reduces the complexity from $O(MN)$ to $O(M+N)$, where $N = n_{SNV} + n_{CNV}$.

3.5 Benchmarking experiments

3.5.1 Synthetic data generation

To generate the synthetic data, a tree is uniformly sampled from the set of undirected trees with $N = N_{SNV} + N_{CNV}$ number of nodes. Then cells are assigned to the trait nodes such that the cell nodes are the leaves of the the tree. The internal nodes are randomly assigned to either a SNV locus or to a CNV locus. For each SNV locus, the copy number and the number of reads covering the locus (depth) are independently generated. The copy number is generated from a discrete uniform distribution from 0 to $c_{\max}$ ($c_{\max}$ equals 5 in the analysis); the depth is generated from a Poisson distribution given the coverage. Given the depth and the copy number, a random genotype is uniformly sampled from all possible genotype

states ($\mathscr{G}$). The number of variant reads is sampled from a binomial distribution given depth, copy number, false positive and false negative error rate parameters (See Algorithm 1).

3.5.2 Synthetic data evaluation

In this section we evaluate both the topology of the inferred tree, and the accuracy of SNV calls. We define a distance D to evaluate the tree topology performance. Consider X to be a matrix derived from the phylogeny, where $x(i,j) = 1$ if locus i is an ancestor of cell j. If X_1 is derived based on a reference tree and X_2 is derived from an inferred tree, then the distance between X_1 and X_2 is estimated by

$$D = \sum_i \sum_j |(x_1(i,j) - x_2(i,j))|/z, \tag{3.27}$$

where z is the total number of entries of matrix X_1 or X_2 (both are the same size).

The performance of SNV calls are estimated by sensitivity, specificity and Youden's index. They are estimated given both the binary SNV calls, and the ground truth calls (derived from the reference tree). The SNV calls are binarized by setting a threshold equal to 0.5 (probability greater than 0.5 is 1 and smaller than 0.5 is 0). The topology of the reference tree provides the ground truth of SNV calls in which locus i is mutated in the entire descendant cells of a sub-tree rooted at i.

In Figure 3.4 and 3.5, we demonstrate the performance of SNV calls for data with different coverages. The data is generated for 200 SNV loci, 100 CNV loci across 1000 cells. We randomly picked 90% of SNV loci and utilized them with the CNV loci for inferring the tree. The remaining 10% SNV loci are added afterwards. The nodes are assigned to 10 random tree topologies. We assign a copy number to each SNV locus. The copy number is uniformly sampled from $\{1 \dots 5\}$. The genotype is also uniformly sampled from all possible genotype states given the copy number. The number of reads and the number of variant alleles are generated following the binomial probability distribution defined in 3.15 and 3.13 given the genotype and different coverage means of 0.04, 0.07, 0.1, 0.5 or 2. The inference is done for 10 different seeds, and non-reversible Parallel Tempering (PT) algorithm [133] with 1000 and 5000 number of iterations with one chain [1]. The

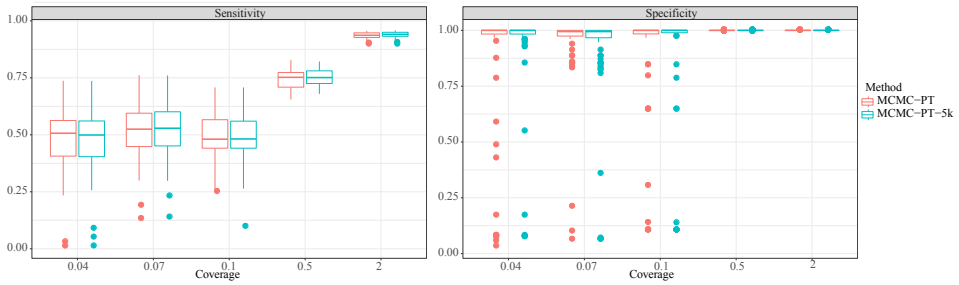

Figure 3.4: The performance (Sensitivity and Specificity) of SNV calls across data coverages of $0.04, 0.07, 0.1, 0.5$, and 2. The data is generated for 200 SNV loci, 100 CNV loci across 1000 cells for 10 different topologies. Only 90% of SNV loci are used for the inference, and the remaining loci are added to the tree afterwards. Copy number is uniformly sampled from $\{1 \dots 5\}$. MCMC-PT and MCMC-PT-5k denote the inference algorithm using PT with 1000 and 5000 iterations, respectively.

values of ε_{FP} and ε_{FN} are equal to 0.01. The inference parameters are constant through all the following analysis. As expected, the higher coverage the better performance. As depicted in Figure 3.4 and 3.5, the performance is not very high with 0.07x coverage. However, the performance reaches almost 1 with 2x coverage. Note that the results are evaluated per cell, and these coverages are all lower than the coverage of the data used in all other existing methods. Figure 3.6 depicts the performance of the model (in terms of tree topology) improves with more coverage.

Next, we investigate the performance of the CellMutScope when different portion of SNV loci are used for inferring the tree and the remaining loci are added to the inferred tree. Using a subset of SNVs is helpful for reducing the run time. The data is generated for 200 SNV loci, 100 CNV loci across 1000 cells. The data coverage is 0.07 and $c_{\max}$ is equal to 5. The percentage of loci used for the inference are 20%, 50% and 90%. In Figure 3.7 and 3.8, we explore how the performance changes with different proportion of SNVs used for tree inference. The performance of detecting SNV is higher when 90% percent of SNV loci are provided. However, the difference is not significant. Figure 3.9, suggests that incorporating SNVs improves resolving the phylogeny. However, for larger number of the nodes, we may need to increase the number of iterations to reach a higher accuracy.

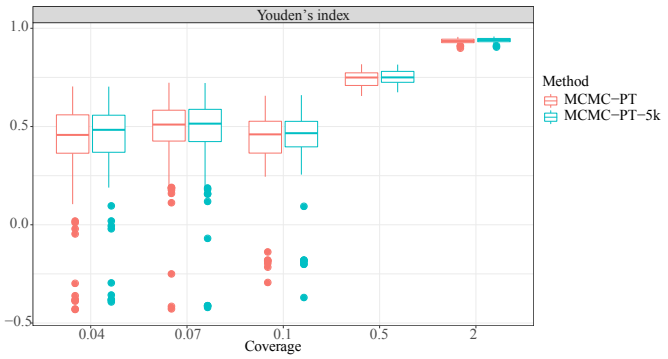

Figure 3.5: The performance (Youden's index) of SNV calls across data coverages of $0.04, 0.07, 0.1, 0.5$, and 2. The data is generated for 200 SNV loci, 100 CNV loci across 1000 cells for 10 different topology. Only 90% of SNV loci are used for the inference, and the remaining are added to the tree afterwards. Copy number is uniformly sampled from $\{1 \dots 5\}$. MCMC-PT and MCMC-PT-5k denote the inference algorithm using PT with 1000 and 5000 iterations, respectively.

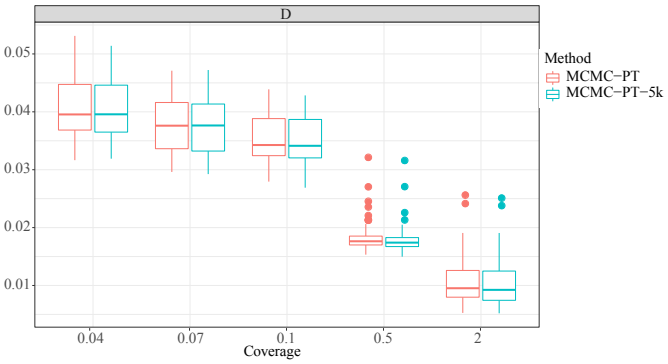

Figure 3.6: The performance (D) of tree topology across data coverages of $0.04, 0.07, 0.1, 0.5$, and 2. The data is generated for 200 SNV loci, 100 CNV loci across 1000 cells for 10 different topology. Only 90% of SNV loci are used for the inference, and the remaining are added to the tree afterwards. Copy number is uniformly sampled from $\{1 \dots 5\}$. MCMC-PT and MCMC-PT-5k denote the inference algorithm using PT with 1000 and 5000 iterations, respectively.

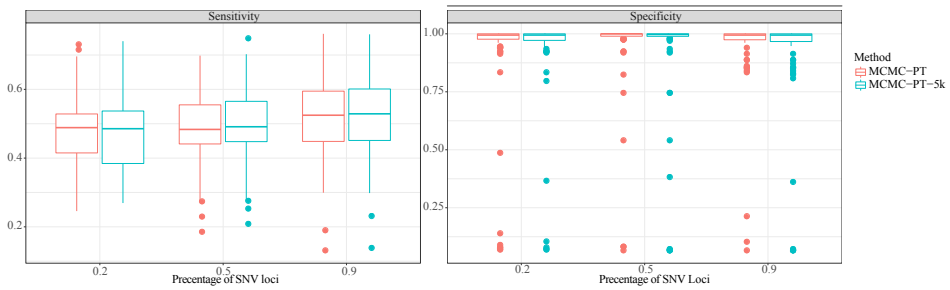

Figure 3.7: The performance (Sensitivity and Specificity) of SNV calls if various percentage of SNV loci are used for the inference, and the remaining are added afterwards. The data is generated for 200 SNV loci, 100 CNV loci across 1000 cells for 10 different topology. Copy number is uniformly sampled from $\{1 \dots 5\}$ and the coverage mean is 0.07. MCMC-PT and MCMC-PT-5k denote the inference algorithms, and use PT with 1000 and 5000 iterations, respectively.

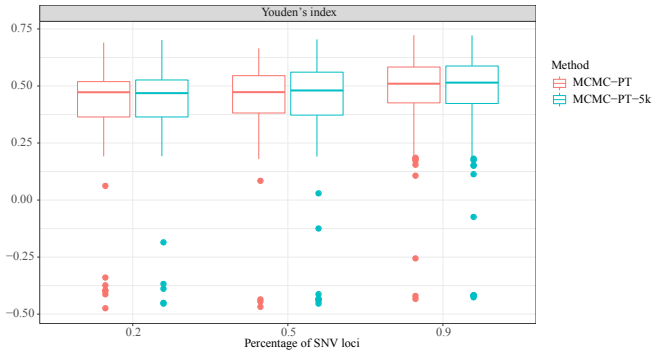

Figure 3.8: The performance (Youden's index) of SNV calls if various percentage of SNV loci are used for the inference, and the remaining are added afterwards. The data is generated for 200 SNV loci, 100 CNV loci across 1000 cells for 10 different topology. Copy number is uniformly sampled from $\{1 \dots 5\}$ and the coverage mean is 0.07. MCMC-PT and MCMC-PT-5k denote the inference algorithm, and use PT with 1000 and 5000 iterations, respectively.

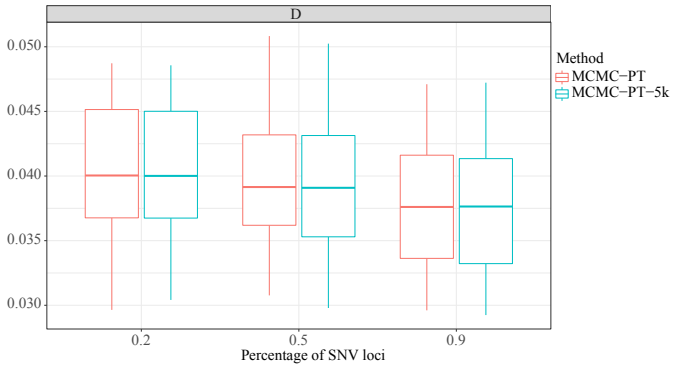

Figure 3.9: The performance (D) of tree topology if various percentage of SNV loci are used for the inference, and the remaining are added afterwards. The data is generated for 200 SNV loci, 100 CNV loci across 1000 cells for 10 different topology. Copy number is uniformly sampled from $\{1 \dots 5\}$ and the coverage mean is 0.07. MCMC-PT and MCMC-PT-5k denote the inference algorithms, and use PT with 1000 and 5000 iterations, respectively.

In Figures 3.10 and 3.11, we explore how the number of cells affects the performance of SNV calls. The data is generated for 200 SNV loci, 100 CNV loci across 100, 200, 500, 1000, 5000 and 10000 cells. The coverage mean is 0.07 and only 20% of SNV loci are used for the tree inference. The other parameters stay similar to the previous analysis. For this number of loci, there is a slight increase in the performance by increasing the number of cells from 100 to 1000. However, further increase of the number of cells slightly drops the performance. This can be explained by the need to increase the number of iterations for tree inference with larger number of cells. Since the SNV data has a lot of missing values, this negatively affects the inference [72].

3.6 Conclusion

We studied the use of tree phylogeny for the purpose of SNV detection across a large number of clonally related cells. The proposed statistical framework uses the copy number profile of each cell and the underlying phylogeny inferred based on the copy number data. Our proposed method, CellMutScope, is fast and scalable

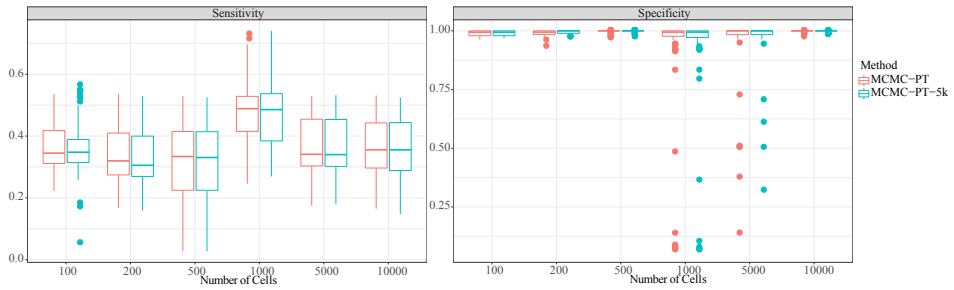

Figure 3.10: The performance (Sensitivity and Specificity) of SNV calls across different number of cells. The data is generated for 200 SNV loci, 100 CNV loci across 100, 200, 500, 1000, 5000, and 10000 cells for 10 different topology. Copy number is uniformly sampled from $\{1 \ldots 5\}$, and the coverage mean is 0.07. Only 20% of SNV loci are used for tree inference. MCMC-PT and MCMC-PT-5k denote the inference algorithms, and use PT with 1000 and 5000 iterations respectively.

to the large number of cells. With the benefit of placing SNVs (when SNVs are incorporated into the Corrupt model) on the underlying tree beside CNV traits, it enables us to detect minor clones that were not distinguishable with the CNV data. The model uses DLP data which has lower depth coverage compared to the data used in other existing methods. In order to increase the coverage for the data, computationally similar cells can be grouped together such that the pseduo-cells have higher coverage (cells grouped together called as a pseduo-cell). Another approach would be an experimental increase in the number of lanes at the cost of a decrease in the number of cells (that keeps experimental cost fixed.).

Detection of SNVs at single cell level facilitates identification of clone specific driver mutations using transcriptomic single cell sequencing data. As the field matures, we expect that the method presented here to be incorporated into analytically comprehensive modelling of single cell whole genome sequencing data (specially when thousands of cells are used to infer the tumour dynamics). For future work we suggest incorporating other genomic aberrations like indels and structural variants in reconstructing the underlying tree phylogeny.

As sequencing cost continues to decrease and the quality of single cell sequencing data continues to increase, single cell sequencing of multiple clonally related

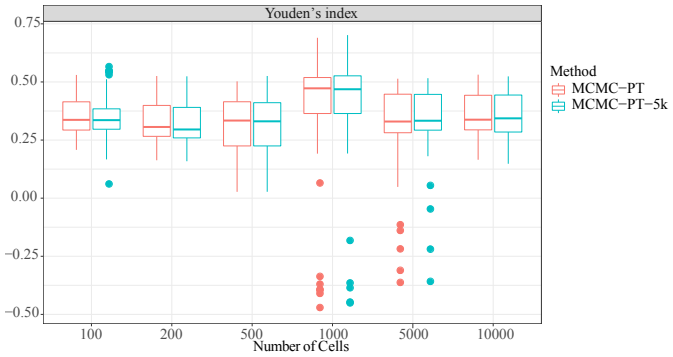

Figure 3.11: The performance (Youden's index) of SNV calls across different number of cells. The data is generated for 200 SNV loci, 100 CNV loci across 100, 200, 500, 1000, 5000, and 10000 cells for 10 different topology. Copy number is uniformly sampled from $\{1 \dots 5\}$, and the coverage mean is 0.07. Only 20% of SNV loci are used for tree inference. MCMC-PT and MCMC-PT-5k denote the inference algorithms, and use PT with 1000 and 5000 iterations, respectively.

cells will proliferate as a valuable experimental design. Thus CellMutScope will be an asset in the toolbox of analytical methods exploited to interpret tumour dynamics in time and space and to enhance the insight into tumours evolutionary process.

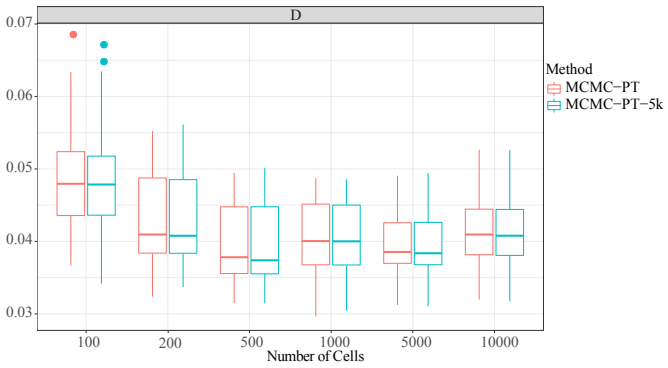

Figure 3.12: The performance (D) of tree topology across different number of cells. The data is generated for 200 SNV loci, 100 CNV loci across 100, 200, 500, 1000, 5000, and 10000 cells for 10 different topology. Copy number is uniformly sampled from $\{1 \ldots 5\}$, and the coverage mean is 0.07. Only 20% of SNV loci are used for tree inference. MCMC-PT and MCMC-PT-5k denote the inference algorithms, and use PT with 1000 and 5000 iterations, respectively.

Chapter 4

Real data application

> *"Each piece, or part, of the whole of nature is always merely an approximation to the complete truth, or the complete truth so far as we know it. In fact, everything we know is only some kind of approximation because we know that we do not know all the laws as yet."*
>
> - Richard P. Feynman

This chapter presents the results of analysis base on SNV data using single cancerous cell sequencing data. Section 4.1 discusses the impact of accurate SNV detection on deciphering the clonal dynamics of tumours. Section 4.2 briefly describes the real data analysed in this chapter. Section 4.3 explains the initial data cleaning and main analysis. It also suggests a quality control step to improve the quality of the results. Finally, Section 4.4 shows the results and the genomic interpretations.

4.1 Introduction

Cancer is driven through the accumulation of genetic mutations at cellular level. Mutations can disrupt the regular activities of proteins. They can lead to a higher proliferation rate, if they confer selective advantage. A specific mutation may differently affect various tumours and patients. The difference in their impact leads to intra- and inter- heterogeneity of tumour cells, and that can explain the difference in therapy responses and survival rates.

Sequencing technologies provide data toward the goal of revealing the associations between mutations and clinical outcomes. Insight on tumour clonal dynamics offers a new perspective to improve the diagnosis, prognosis and treatment decisions beyond conventional approaches [93].

Mutations that occur in high impact genes or non-synonymous coding genes can be the underlying cause of cells phenotypes. Several genes such as TP53, BRCA1, BRCA2, PTEN, pRb, etc. are well known tumour suppressor genes which inhibit cell proliferation and tumour development [5, 97]. Another group of genes, reported as housekeeping genes, like ACTB, GAPDH, AHSP, B2M, etc. govern or prevent cell growth [97, 114]. Mutations that affect the regular activity of these genes promote unconditional cell growth. For example, recent studies reveal the driving role of genes PIK3CA, PTEN, AKT1, PIK3CA, ERBB2, FGFR1, MYC, MAP3K1, TP53, CCND1 and RB1 in breast cancer [5, 65, 132], CDKN2A, KRAS, PTEN, RB1, CCNE1, EVI2A and LCP2 in ovarian cancer [45] and SPOP, FOXA1 and MED12 in prostate cancer [7].

In this chapter, we focus on the analysis of three real data sets, each compromising thousands of clonally related cancerous cells. The model imputes the missing values of the data and reveals that clone specific mutations occurred in high impact genes. We evaluate the performance and report the high impact genes identified in these data sets.

4.2 Data

The first sample is from a patient with a high-grade serous ovarian cancer. The samples are across different points in time and in anatomical spaces. The other two data sets are from patients with breast cancer tissues xenografted in mice and transplanted across multiple time points.

4.2.1 Ovarian cancer

High-grade serous ovarian cancer that arises from the serous epithelial layer in the abdominopelvic cavity is the most aggressive subtype of ovarian cancer [13]. Tissue samples from female patients with high-grade serous ovarian cancer are used to sequence cell lines in more than 1966 libraries [76]. The cell lines are from

the 3 clonally related high grade serous samples, sourced from one primary tumour ($SA1090$) and two relapse specimens after therapy ($SA921$ and $SA922$). $SA1090$ (or called $OV2295$) and $SA922$ (or called $OV2295(R)$) are from ascites, and $SA921$ (or called $TOV2295$) is from the solid tumour. For simplicity, these three samples together are called $OV2295$.

4.2.2 Breast cancer xenograft samples

Breast cancer is the most common cause of death in women diagnosed with cancer [110]. Medical practitioners increasingly use hormone receptors to categorize breast cancer. Hormone receptors of healthy breast cells receive messages and provide instructions to cells for their growth and functionality [2]. A cancer is called estrogen-receptor-positive (or ER+), if it has receptors for estrogen. It is called progesterone-receptor-positive (PR+), if it has progesterone receptors. A smaller percentage of breast cancers, about 20%, may have an excess amount of HER2 protein (HER2+). This means that the cancer cells may receive signals from ER, PR or HER2 to promote their growth. Hormonal therapies can disrupt the activity of these hormones to slow down the growth rate or even stop the proliferation. Triple-negative breast cancer (TNBC) is the one that gives negative test result for estrogen receptors, progesterone receptors, and excess HER2 protein. Because of the lack of these hormones, it is likely that the patients with TNBC do not respond to hormone therapies.

Tumour samples from a HER2+ patient, and a Triple-negative breast cancer patient were taken and transplanted in mouse to generate patient-deriven xenografts (PDX) across multiple generations (passages). Sample $SA532$ is from a HER2+ patient, and it was serially passaged for up to 10 generations, $X1$, $X2$, ..., $X10$. The sequenced data compromises 8381 cells. Sample $SA609$ is from a patient with Triple-negative breast cancer, and it was serially passaged for up to 10 generations, $X1$, $X2$, ..., $X10$. The sequenced data compromises the total number of 10553 cells. We have used the cells from the passages that is studied by Salehi et al.

4.3 Data analysis

4.3.1 Data cleaning

The initial cleaning steps remove cells with the number of read counts less than $500,000$ or the quality below 0.75 [76]. Also the s-phase cells are removed at this step as described by Salehi et al. [104]. The total number of 1345 cells are retained from *OV*2295. However, we only used 891 cells from the study reported by Laks et al. in [76]. As a first approximation to clones, Laks et al. used dimensionality reduction and clustering to identify 9 cell subsets with shared copy number profiles. Then 891 cells are selected from subsets with more than 50 cells. The number of cells retained after the initial cleaning for *SA*532 and *SA*609 is 2400 and 3243, respectively.

4.3.2 Main analysis

The first step in the analysis is to prepare the copy number data and the underlying tree using the aligned reads. A schematic view of input data is depicted in Figure 4.1(a). The copy number of the cells is inferred using the approach suggested by Zahn et. al in [141]. As a result, the whole genome of each cell is segmented by bins, and each bin is associated with a copy number and a quality measure. A schematic view of the CNV data is depicted in Figure 4.1(b). The bins with quality less than 0.99 are ignored in the Corrupt analysis. Corrupt utilizes the CNV call bins to infer the underlying tree (See Figure 4.1 (c)) [1]. Corrupt is then run with $10,000$ number of iterations across 10 chains. The false positive and false negative error rates are set globally, and the maximum rate for false positive and false negative are 0.1 and 0.5, respectively [1].

Figures 4.5, 4.6 and 4.7 depict the copy number profile and the inferred tree using Corrupt. The tree is cut to identify subset of cells sharing similar copy number profiles by using the approach of Salehi et al. in [104]. For *SA*609, we used the super tree inferred by Salehi et al., and removed the cells that are not in the set of *SA*609 cells. The distribution of the number of cells from different samples in each clone for *OV*2295, *SA*532, and *SA*609 is depicted in Figures 4.2, 4.3 and 4.4, respectively.

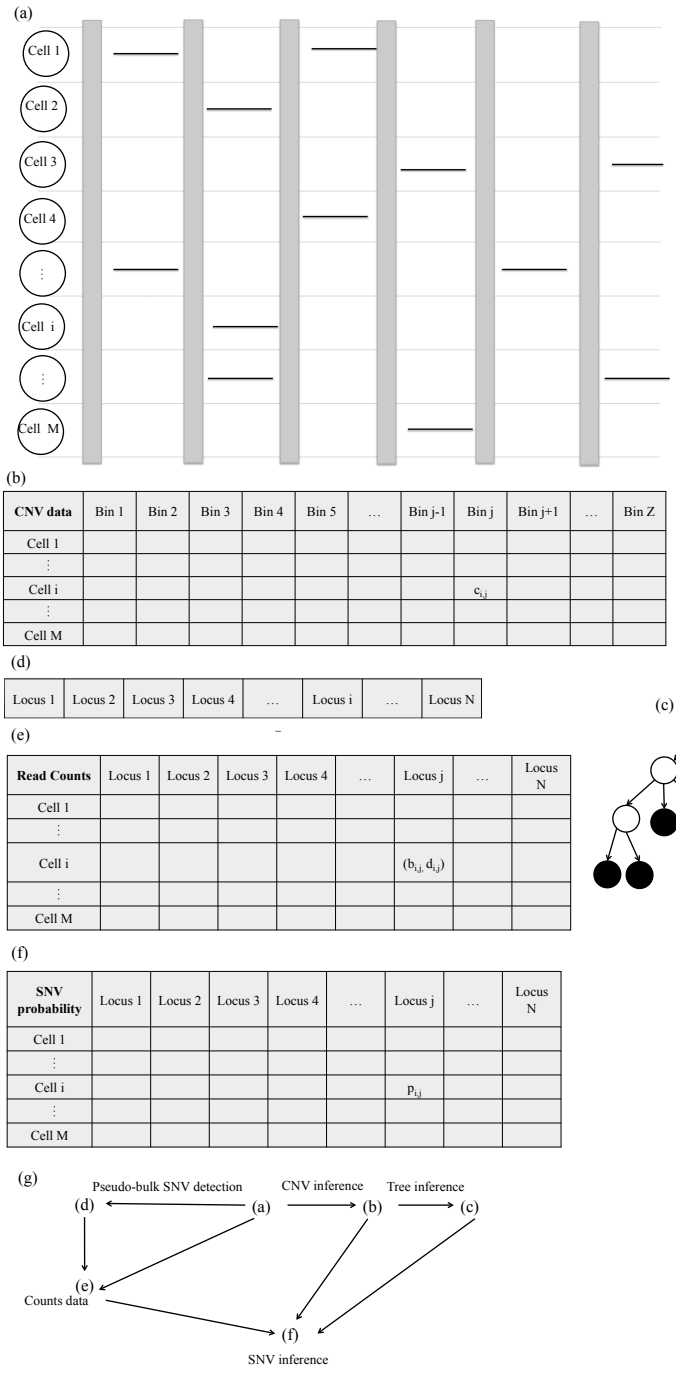

Figure 4.1: (Caption next page.)

Figure 4.1: (a) A schematic view of single cell sequencing data. The black short horizontal lines represent reads that are associated with each cell's identity. The vertical grey lines separate different segments of the genome. (b) The CNV data inferred by using the single cell sequencing data depicted in (a). The genome is segmented into continuous bins. The CNV calling algorithm assigns copy number $c_{i,j}$ to cell i and bin j. (c) A hypothetical phylogenetic tree inferred by Corrupt. The black circles represent cells, and the blank circles represent CNV alteration. (d) The list of loci with high SNV call probability from the pseduo-bulk analysis. (e) The raw data, $(b_{i,j}, d_{i,j})$, denote the number of variant and matched reads at loci j and cell i. (f) SNV call probability inferred by CellMutScope, for loci j and cell i, is denoted by $p_{i,j}$. (g) The dependencies between different parts of the analysis.

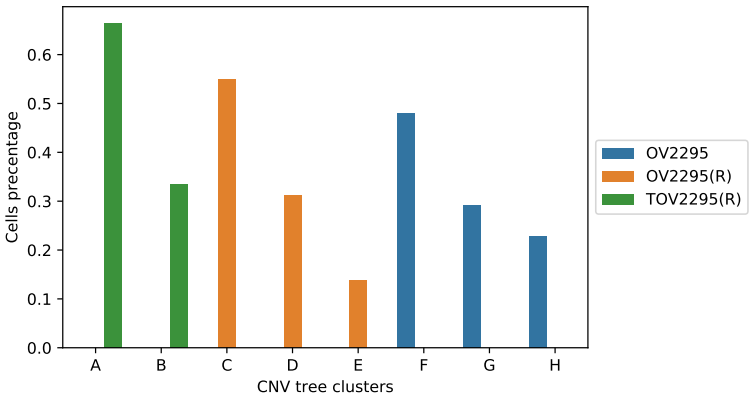

Figure 4.2: The distribution of the number of cells from different samples in each clone for *OV*2295.

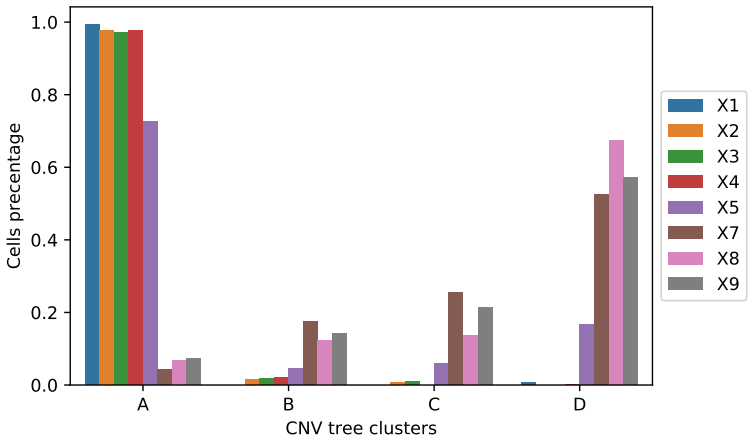

Figure 4.3: The distribution of the number of cells from different samples in each clone for *SA*532.

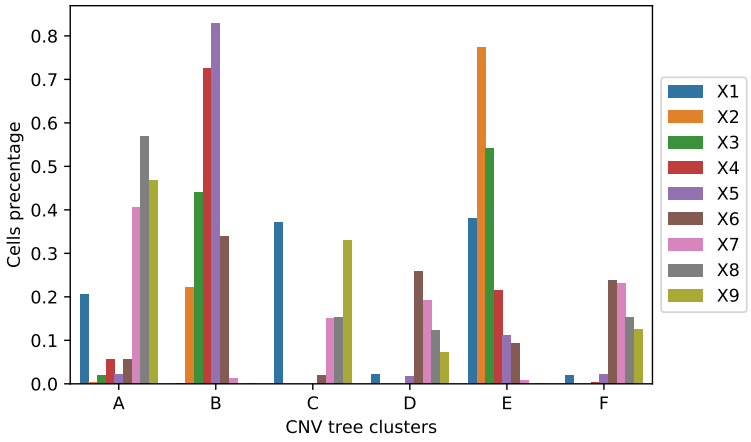

Figure 4.4: The distribution of the number of cells from different samples in each clone for *SA*609.

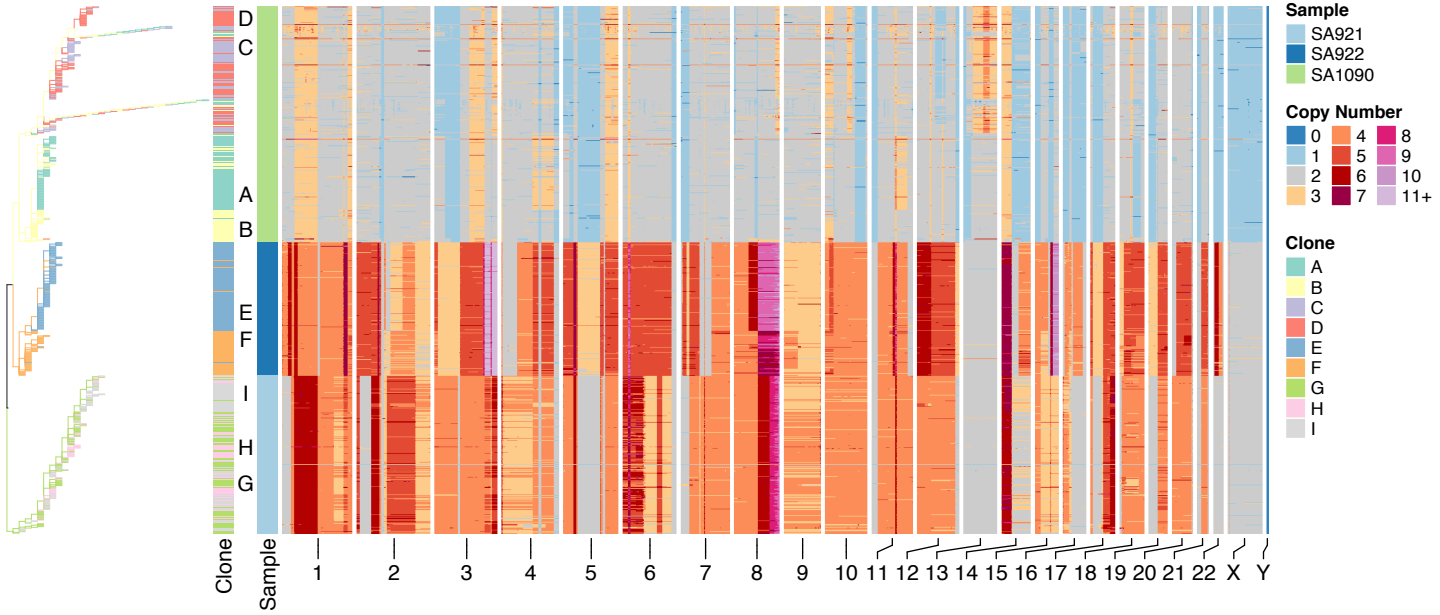

Figure 4.5: The copy number heatmap for $OV2295$ data. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The cells are from three samples $OV2295$ ($SA1090$), $OV2295(R)$ ($SA921$) and $TOV2295(R)$ ($SA922$). The tree includes 9 clones identified in Laks et al. study [76].

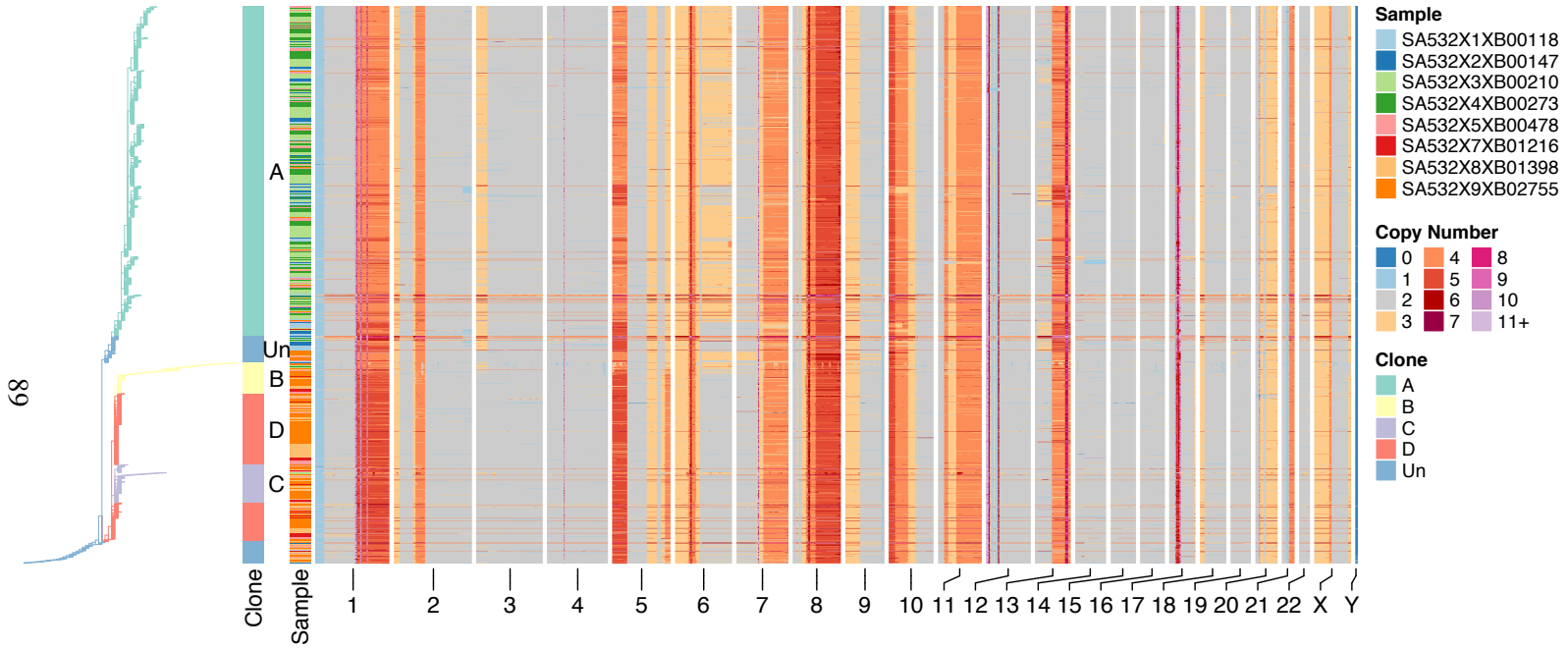

Figure 4.6: The copy number heatmap for *SA*532 data. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

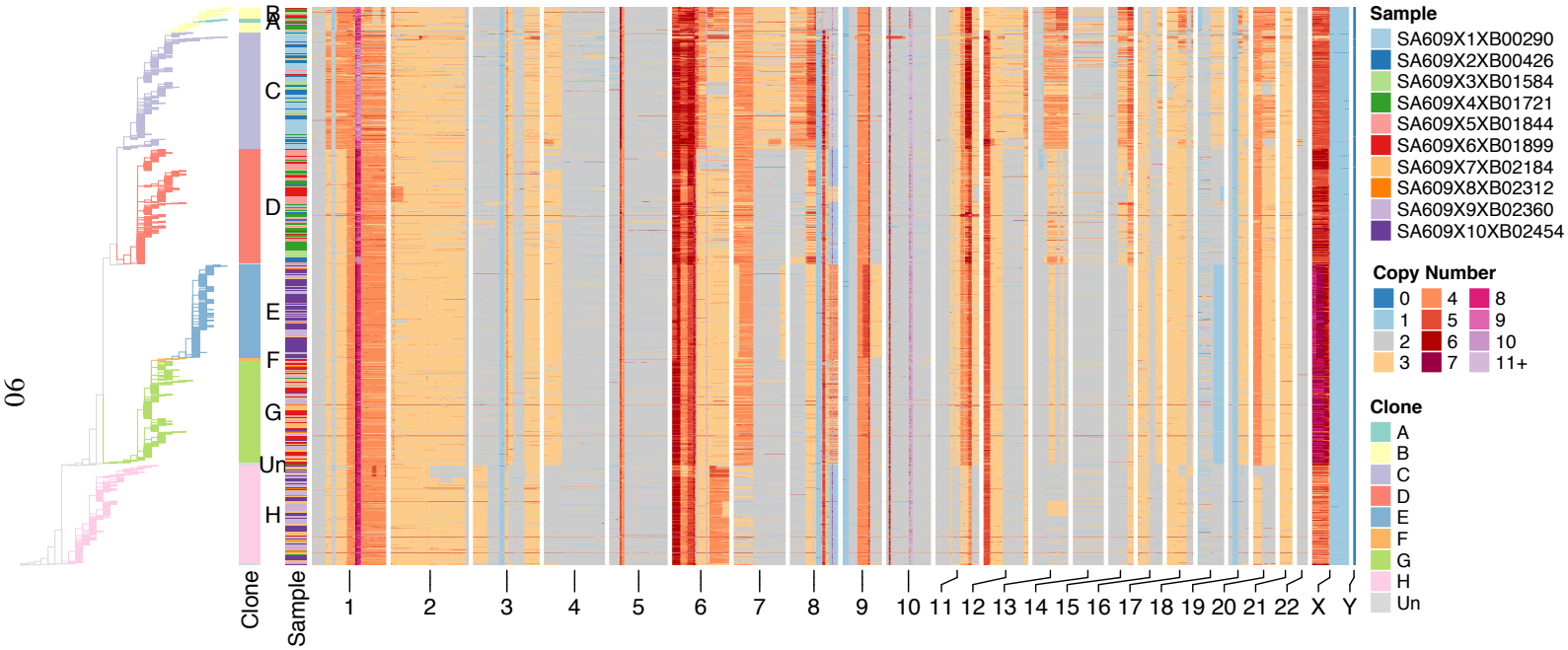

Figure 4.7: The copy number heatmap for *SA*609 data. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

Next, a group of SNV loci candidates is selected from the aggregated data in which all cells were collapsed into a 'pseudo-bulk' genome (See Figure 4.1 (d)). MutationSeq and Strelka are run for selecting SNV loci candidates from the pseudo-bulk sample [31, 105]. The loci with MutationSeq probability greater than 0.9 and Strelka's Phred quality score of 20 or more are identified as SNV candidates. In addition, each locus should demonstrate at least two variant reads across the cells. The total number of 14020, 11416 and 15446 SNV loci are selected for *OV*2295, *SA*532 and *SA*609, respectively. The cells with no variant across the loci are removed. This condition reduces the number of cells to 2400 for *SA*532, 3243 for *SA*609, and 731 for *OV*2295. Table 4.1 summarizes some statistics about the cells used in the following analysis. As specified in this table, *OV*2295 has a larger number of reads, and a higher breadth, and a higher depth coverage in comparison to *SA*532 and *SA*609.

CellMutScope described in Chapter 3, incorporates the reads data (See Figure 4.1 (e)), the copy number data (See Figure 4.1 (b)), and the underlying tree (See Figure 4.1 (c)) to detect SNVs per cell. Output of CellMutScope is schematically depicted in Figure 4.1 (g). The value of ε_{FN} is approximated as

$$\varepsilon_{FN} = \frac{n_{deletion}}{n_{total}} \times \frac{1}{2}, \tag{4.1}$$

where $n_{deletion}$ is the number of bins with copy number less than 2, and n_{total} is the total number of bins across the genome. The value $\frac{1}{2}$ is the probability that the mutation has occurred on the missing strand. The value of ε_{FP} is 0.001. The overall picture of the pipeline and dependencies is depicted in Figure 4.1 (g).

4.4 Experimental results

CellMutScope is applied on *OV*2295, *SA*532, and *SA*609. Since the model is based on placing SNV loci on the tree, the distribution of SNV call probabilities is bimodal. It depends on the position of the node (if it is an ancestor of a cell or not). Figures 4.8, 4.9 and 4.10 depict the SNV call probability distributions for *OV*2295, *SA*532 and *SA*609, respectively. Figures 4.8 (a), 4.9 (a) and 4.10 (a) confirm that the distribution has two peaks at 0 and 1. Zoomed in at the distribution of SNV call

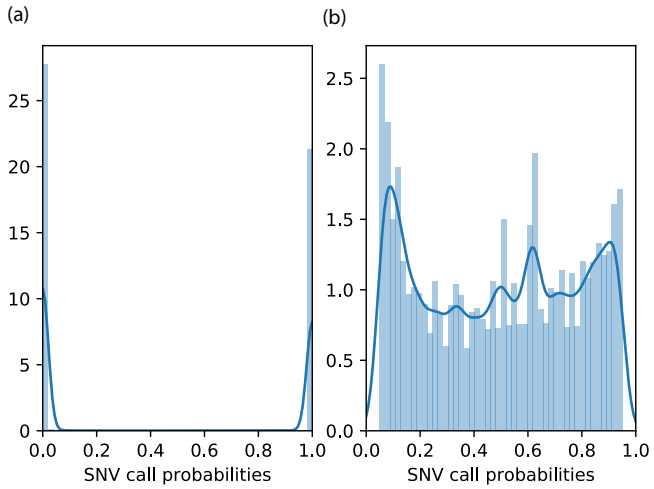

Figure 4.8: (a) SNV call probability distribution for *OV*2295. (b) SNV call probability distribution for *OV*2295 for the probabilities between 0.05 and 0.95.

probabilities between 0.05 and 0.95 are depicted in Figures 4.8 (b), 4.9 (b) and 4.10 (b). There is also a peak at 0.5 that is smaller than the peaks at 0 and 1. The 0.5 peak represents for the loci that the model was not confident in SNV calling. This can be partly explained by missing data, low coverage, or inconsistency of SNV data with the tree. The coverage data supports this assumption as well.

Figure 4.11 depicts the SNV call probabilities for *OV*2295. For the sake of visualization, a list of 200 SNV loci candidates is randomly sampled, from the set of all loci with equal probabilities. Moreover, the loci are sorted using *R*'s 'hclust' function with 'average' metric and 'manhattan' distance. Figure 4.12 depicts the number of variant reads in each cell across the same set of loci. The loci in SNV call probability plot (Figure 4.11) and the plot that depicts the number of variant reads (Figure 4.12) follow a exactly similar order. Considering both Figures 4.12 and 4.11, the number of variant reads in the area (black area in Figure 4.12) that CellMutScope has low probability (blue area in Figure 4.11) is small.

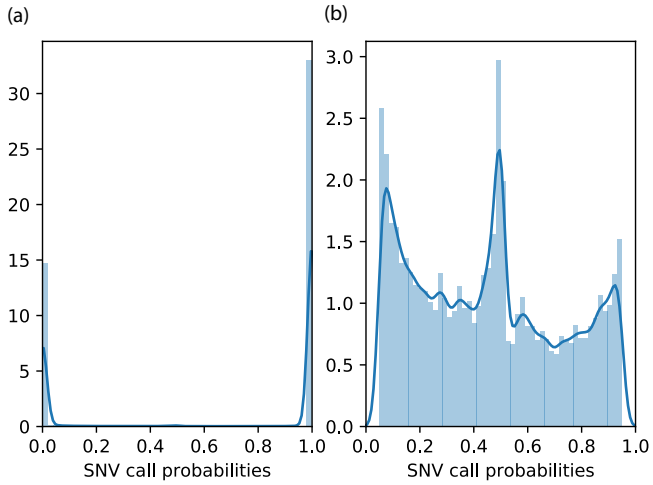

Figure 4.9: (a) SNV call probability distribution for *SA*532. (b) SNV call probability distribution for *SA*532 for the probabilities between 0.05 and 0.95.

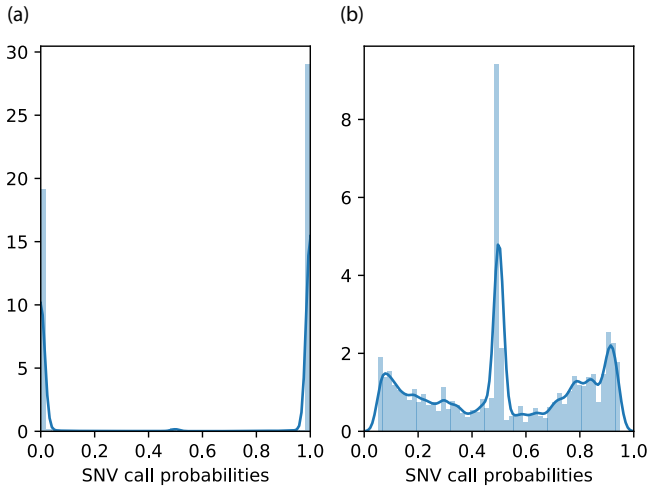

Figure 4.10: (a) SNV call probability distribution for *SA*609. (b) SNV call probability distribution for *SA*609 for the probabilities between 0.05 and 0.95.

Metric	OV2295	SA609	SA532
Number of cells	731	3243	2400
Number of SNV loci candidates	14020	15446	11416
Total reads	8.558316e+06	1.446370e+06	2.635680e+06
Coverage breadth	1.563200e-01	3.452980e-02	5.356769e-02
Coverage depth	1.776126e-01	3.615207e-02	5.871972e-02
Total mapped reads	7.913032e+06	1.329028e+06	2.376801e+06
Unmapped reads	6.452840e+05	1.173421e+05	2.588786e+05

Table 4.1: The table shows the total number of cells, the total number of SNV loci candidates, the average number of reads, the average breadth coverage, the average depth coverage, the average total number of mapped reads and the average total number of unmapped reads across the cells in each sample.

Figure 4.11: The heatmap depicts *OV*2295 SNV call probability of 200 random loci across the cells. The blue colour depicts the low probabilities near 0, the white colour depicts the probabilities around 0.5, and the orange colour depicts the probabilities around 1. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The cells are from three samples *OV*2295 (*SA*1090), *OV*2295(*R*) (*SA*921) and *TOV*2295(*R*) (*SA*922). The tree includes 9 clones identified in Laks et al. study [76].

Figure 4.12: The heatmap depicts the number of variant reads observed in each cell of $OV2295$ for 200 random loci. No variant read is depicted in black, and reads with variant alleles are depicted in bright colours. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The cells are from three samples $OV2295$ ($SA1090$), $OV2295(R)$ ($SA921$) and $TOV2295(R)$ ($SA922$). The tree includes 9 clones identified in Laks et al. study [76].

For *SA*532, Figure 4.13 depicts the SNV call probabilities, and Figure 4.14 depicts the number of variant reads for 200 selected loci across the cells. A similar approach is used for sorting SNV loci across the x-axis. As depicted in Figure 4.13, more detailed clonal structure appears in clone *A* that was not apparent with only the CNV data and smaller group of cells sharing similar SNV profile. Compared with the results from *OV*2295, we see more variant reads in the area (black area in Figure 4.14) that CellMutScope has low probability (blue area in Figure 4.13). For *SA*609, Figures 4.15 and 4.16 show SNV call probabilities and the number of variant read across 200 random loci. As shown, subgroups of cells in clone *C* and *H* sharing similar SNVs can represent sub-clones in clone *C* and *H*. Similarly, for *SA*609 (see Figures 4.15 and 4.16), there are a few cells having variant reads while they do not have a high SNV call probability. This suggests the need for a quality control step to remove the inconsistent loci (i.e. the loci having SNV data inconsistent with the tree structure). Comparing the results from *OV*2295, *SA*532 and *SA*609, we can see *SA*532 and *SA*609 have a higher rate of cells with variant reads while they do not have a high SNV call probability. This can be explained by the difference between the coverage depth of the samples. *OV*2295 has higher coverage depth. The coverage depth of *OV*2295, *SA*532 and *SA*609 is 0.17, 0.03 and 0.05, respectively (See Table 4.1).

Quality control step A SNV is called present, if the value of SNV call probability from CellMutScope is greater or equals to 0.8. For clones that SNV is not present, we count the number of cells with variant reads normalized to the total number of cells in the clone. The average number of variant reads is expected to be greater than the average coverage times the number of cells in each clone divided by the copy number (assuming only one of the copies has mutation). This implies the normalized number of variant reads should be greater than the average coverage divided by copy number. With the average coverage of 0.05 and 0.01 and copy number 10, the threshold is 0.005 and 0.001, respectively.

Figure 4.13: The heatmap depicts *SA*532 SNV call probability for 200 random loci across the cells. The blue colour depicts the low probabilities near 0, the white colour depicts the probabilities around 0.5, and the orange colour depicts the probabilities around 1. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

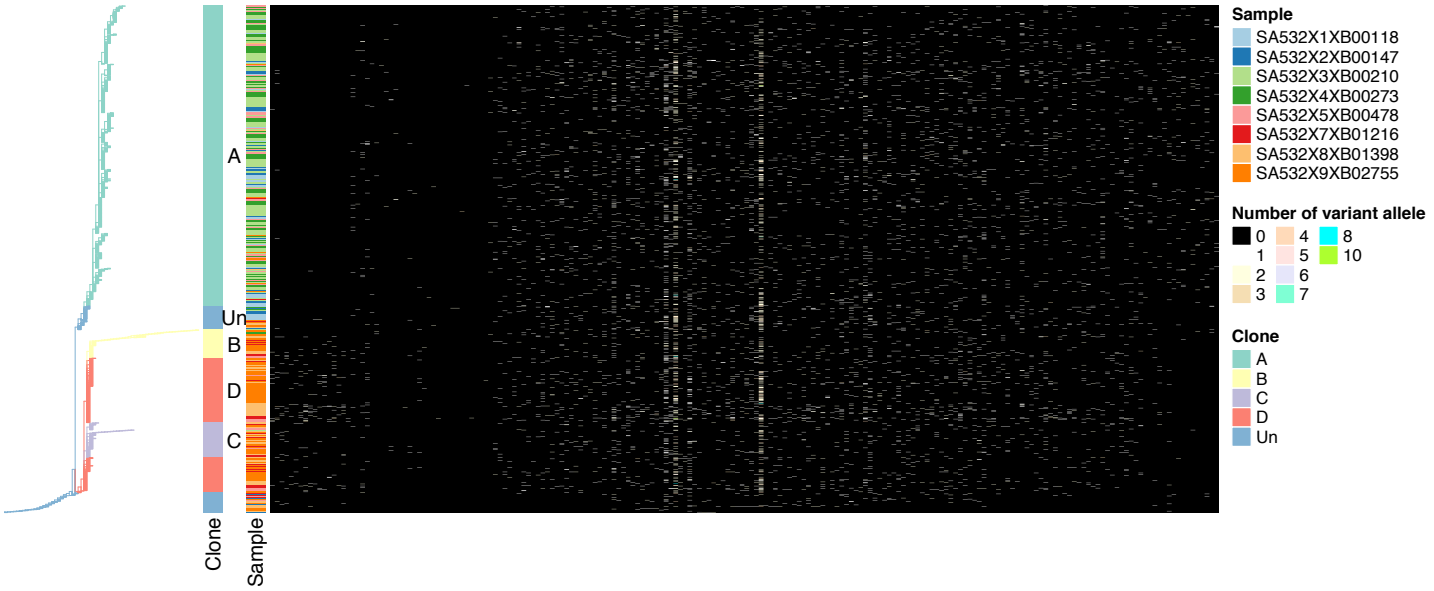

Figure 4.14: The heatmap depicts the number of variant reads observed in each cell of *SA*532 for 200 random loci. No variant read is depicted in black, and reads with variant alleles are depicted in bright colours. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

Figure 4.15: The heatmap depicts *SA*609 SNV call probability for 200 random loci across the cells. The blue colour depicts the low probabilities near 0, the white colour depicts the probabilities around 0.5, and the orange colour depicts the probabilities around 1. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

Figure 4.16: The heatmap depicts the number of variant reads observed in each cell of *SA*609 for 200 random loci. No variant read is depicted in black, and reads with variant alleles are depicted in bright colours. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

Figure 4.17: The heatmap depicts *OV*2295 SNV call probability for 200 random loci passed the quality control step with threshold equals 0.001 across the cells. The blue colour depicts the low probabilities near 0, the white colour depicts the probabilities around 0.5, and the orange colour depicts the probabilities around 1. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The cells are from three samples *OV*2295 (*SA*1090), *OV*2295(R) (*SA*921) and *TOV*2295(R) (*SA*922). The tree includes 9 clones identified in Laks et al. study [76].

Figure 4.18: The heatmap depicts the number of variant reads observed in each cell of $OV2295$ for 200 random loci passed the quality control step with threshold equals 0.001 across the cells. No variant read is depicted in black, and reads with variant alleles are depicted in bright colours. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The cells are from three samples $OV2295$ ($SA1090$), $OV2295(R)$ ($SA921$) and $TOV2295(R)$ ($SA922$). The tree includes 9 clones identified in Laks et al. study [76].

Figure 4.19: The heatmap depicts the SNV call probability of 200 random loci passed the quality control step with threshold equals 0.005 across the cells. The blue colour depicts the low probabilities near 0, the white colour depicts the probabilities around 0.5, and the orange colour depicts the probabilities around 1. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The cells are from three samples $OV2295$ ($SA1090$), $OV2295(R)$ ($SA921$) and $TOV2295(R)$ ($SA922$). The tree includes 9 clones identified in Laks et al. study [76].

Figure 4.20: The heatmap depicts the number of variant reads observed in each cell for 200 random loci passed the quality control step with threshold equals 0.005 across the cells. No variant read is depicted in black, and reads with variant alleles are depicted in bright colours. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The cells are from three samples $OV2295$ ($SA1090$), $OV2295(R)$ ($SA921$) and $TOV2295(R)$ ($SA922$). The tree includes 9 clones identified in Laks et al. study [76].

Sample	Original	TH=0.005	TH=0.001
OV2295	14020	12574	12574
SA609	15446	13052	9962
SA532	11416	9648	8403

Table 4.2: The table shows the total number of cells before quality control step (original) and the total number of cells after quality control with threshold equal 0.005 and 0.001.

For *OV*2295, the total number of SNVs passing quality control step with both thresholds 0.005 and 0.001 is 12574 (see Table 4.2). Figures 4.17 and 4.18 depict SNV call probability for 200 loci randomly sampled from the set of loci that pass the quality check with a threshold of 0.001 for *OV*2295. Figures 4.19 and 4.20 depict results for the same data with a threshold of 0.005. As shown in Table 4.2, for *SA*532, the total number of SNVs passing quality control step with thresholds 0.005 and 0.001 are 9648 and 8403, respectively. For *SA*609, the numbers are 13052 and 9962 for thresholds of 0.005 and 0.001, respectively. The results of similar analysis for *SA*532 and *SA*609 with thresholds of 0.001 and 0.005 are depicted in Figures 4.21 to 4.28. See Figures 4.21 and 4.22 for *SA*532 results with a threshold of 0.001. See Figures 4.23 and 4.24 for *SA*532 results with a threshold of 0.005. See Figures 4.25 and 4.26 for *SA*609 results with a threshold of 0.001. See Figures 4.27 and 4.28 for *SA*609 results with a threshold of 0.005. As shown in the results after the quality control step, we observe lower rate of cells having variant reads while they do not have a high SNV call probability.

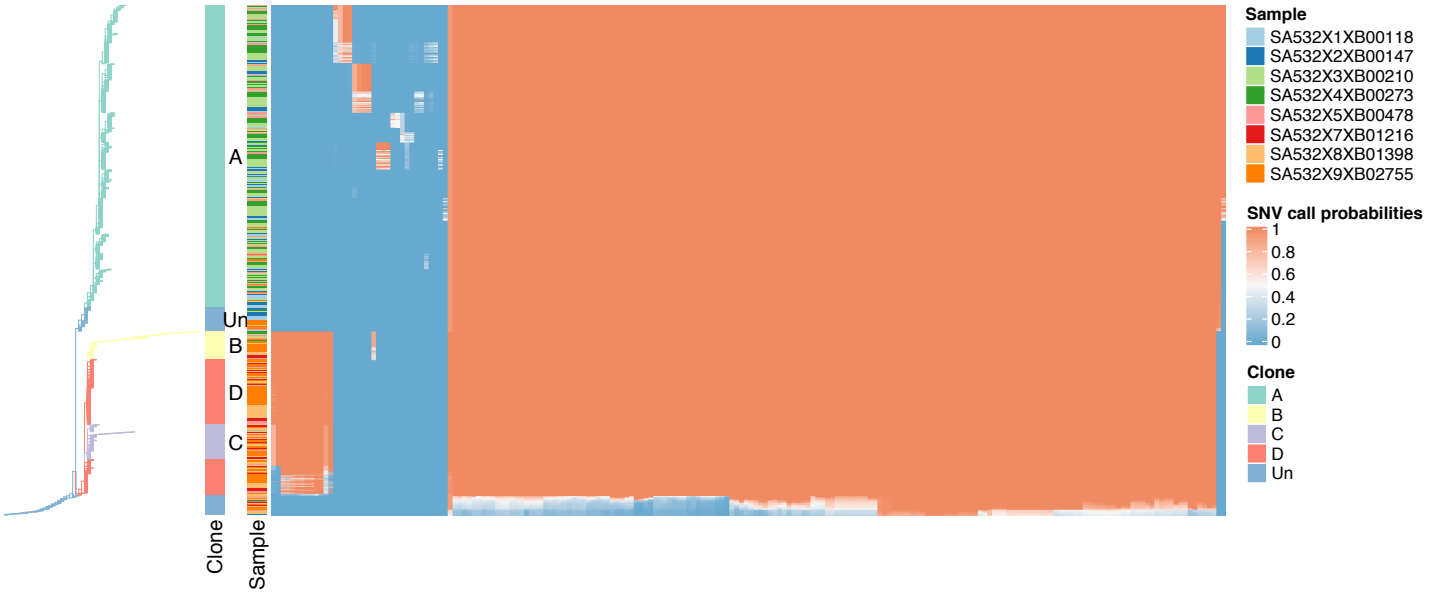

Figure 4.21: The heatmap depicts *SA*532 SNV call probability of 200 random loci passed the quality control step with threshold equals 0.001. The blue colour depicts the low probabilities near 0, the white colour depicts the probabilities around 0.5, and the orange colour depicts the probabilities around 1. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

Figure 4.22: The heatmap depicts the number of variant reads observed in each cell of *SA*532 for 200 random loci passed the quality control step with threshold equals 0.001. No variant read is depicted in black, and reads with variant alleles are depicted in bright colours. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

Figure 4.23: The heatmap depicts *SA*532 SNV call probability of 200 random loci passed the quality control step with threshold equals 0.005. The blue colour depicts the low probabilities near 0, the white colour depicts the probabilities around 0.5, and the orange colour depicts the probabilities around 1. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et. al. study [104].

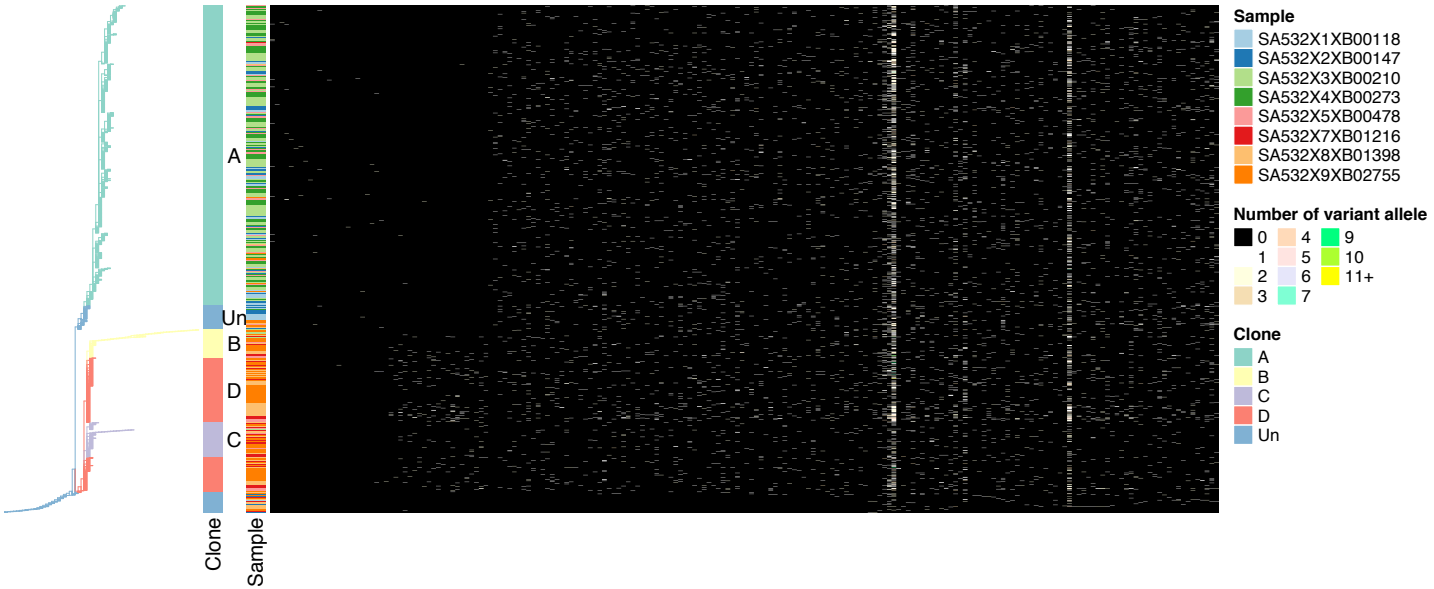

Figure 4.24: The heatmap depicts the number of variant reads observed in each cell of *SA*532 for 200 random loci passed the quality control step with threshold equals 0.005. No variant read is depicted in black, and reads with variant alleles are depicted in bright colours. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

Figure 4.25: The heatmap depicts *SA*609 SNV call probability of 200 random loci passed the quality control step with threshold equals 0.001. The blue colour depicts the low probabilities near 0, the white colour depicts the probabilities around 0.5, and the orange colour depicts the probabilities around 1. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

Figure 4.26: The heatmap depicts the number of variant reads observed in each cell of *SA*609 for 200 random loci passed the quality control step with threshold equals 0.001. No variant read is depicted in black, and reads with variant alleles are depicted in bright colours. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. studies [104].

Figure 4.27: The heatmap depicts *SA*609 SNV call probability of 200 random loci passed the quality control step with threshold equals 0.005. The blue colour depicts the low probabilities near 0, the white colour depicts the probabilities around 0.5, and the orange colour depicts the probabilities around 1. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

Figure 4.28: The heatmap depicts the number of variant reads observed in each cell of *SA*609 for 200 random loci passed the quality control step with threshold equals 0.005. No variant read is depicted in black, and reads with variant alleles are depicted in bright colours. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

To confirm that the structure of the tree is an important piece of information for detecting the SNVs, particularly the ones that are only present in a subset of cells (clone specific), we introduce a perturbed tree by permuting the cells on the tree while the tree structure is fixed. We repeated this process 20 times. Figures 4.29, 4.31 and 4.33 show the mean probability of SNV calls across the perturbed trees. The results show that the perturbation disables the model from detecting the clone specific mutations, and only the ancestral mutations survived. Figures 4.30, 4.32 and 4.34 show the SNV variant reads corresponding to the permutation analysis of Figures 4.29, 4.31 and 4.33, respectively. This proves the tree structure is a crucial element of the model to enable capturing the clone specific mutations and addressing the sparsity of the data.

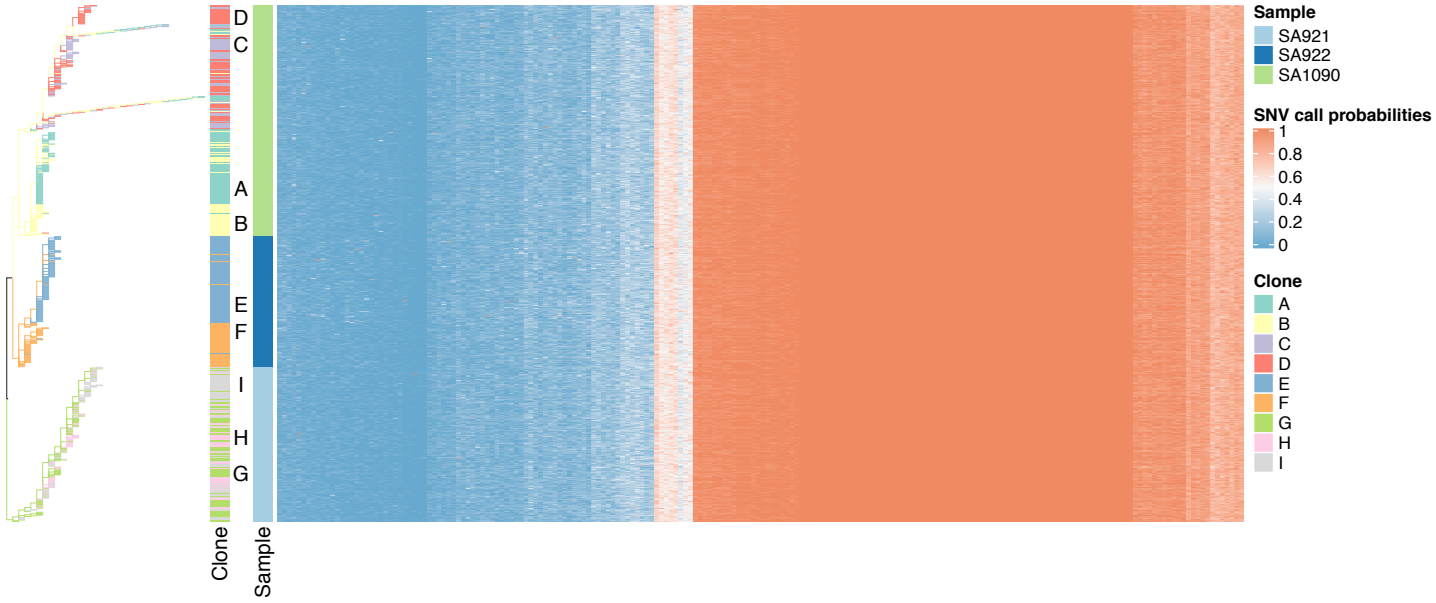

Figure 4.29: The heatmap depicts the mean SNV call probability of 200 random loci across the cells of $OV2295$ over 20 perturbed tree (quality control step with threshold equals 0.001 is applied). The blue colour depicts the low probabilities near 0, the white colour depicts the probabilities around 0.5, and the orange colour depicts the probabilities around 1. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The cells are from three samples $OV2295$ ($SA1090$), $OV2295(R)$ ($SA921$) and $TOV2295(R)$ ($SA922$). The tree includes 9 clones identified in Laks et al. study [76].

Figure 4.30: The heatmap depicts the number of variant reads observed in each cell of $OV2295$ for 200 random loci passed the quality control step with threshold equals 0.001. No variant read is depicted in black, and reads with variant alleles are depicted in bright colours. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The cells are from three samples $OV2295$ ($SA1090$), $OV2295(R)$ ($SA921$) and $TOV2295(R)$ ($SA922$). The tree includes 9 clones identified in Laks et al. study [76].

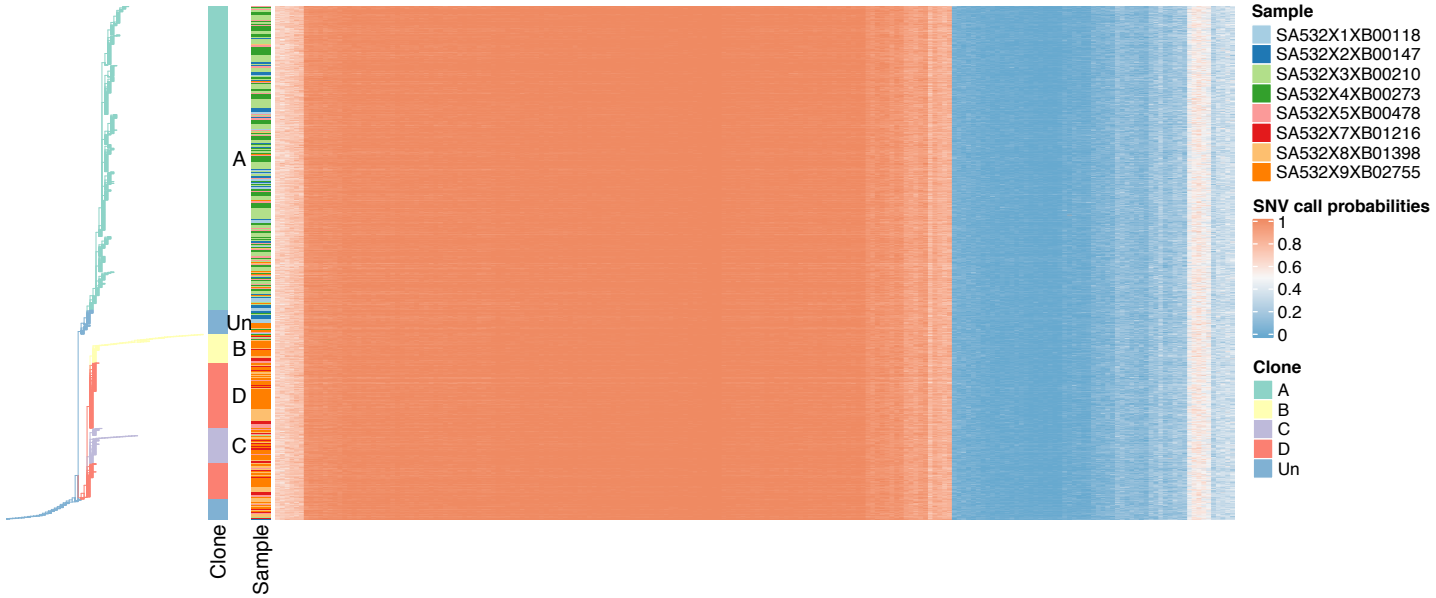

Figure 4.31: The heatmap depicts the mean SNV call probability of 200 random loci across the cells of *SA*532 over 20 perturbed tree (quality control step with threshold equals 0.001 is applied). The blue colour depicts the low probabilities near 0, the white colour depicts the probabilities around 0.5, and the orange colour depicts the probabilities around 1. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

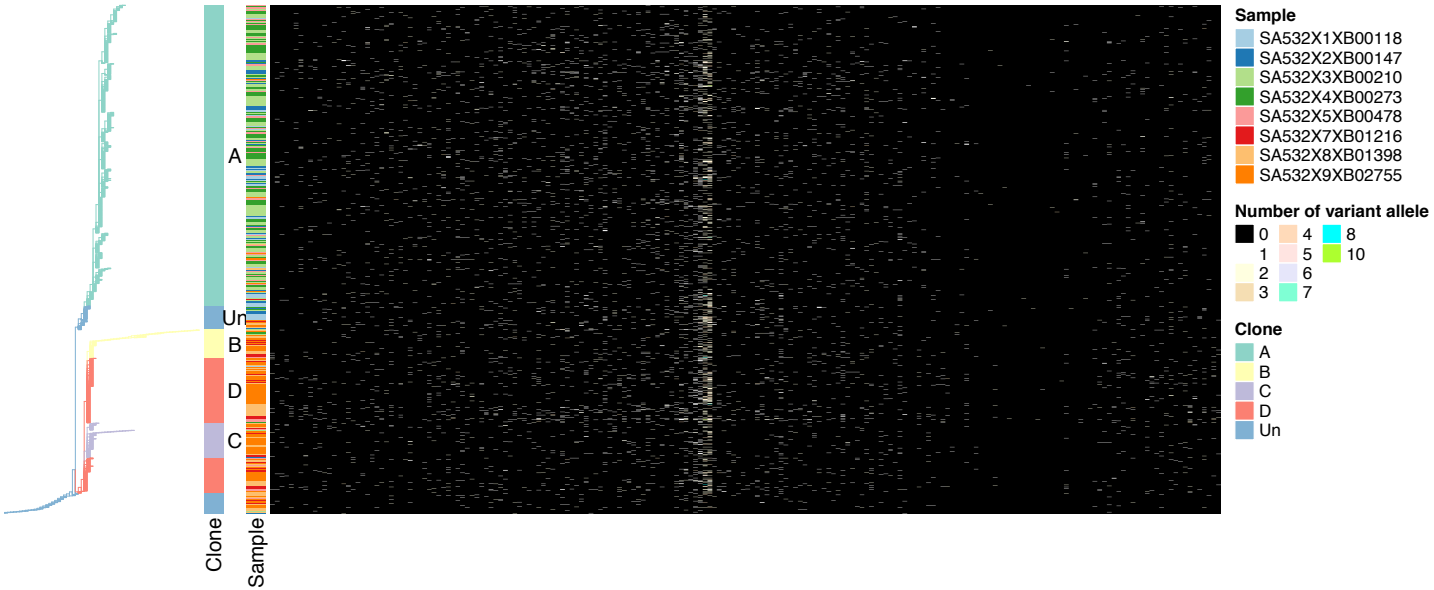

Figure 4.32: The heatmap depicts the number of variant reads observed in each cell of *SA*532 for 200 random loci passed the quality control step with threshold equals 0.001. No variant read is depicted in black, and reads with variant alleles are depicted in bright colours. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. studies [104].

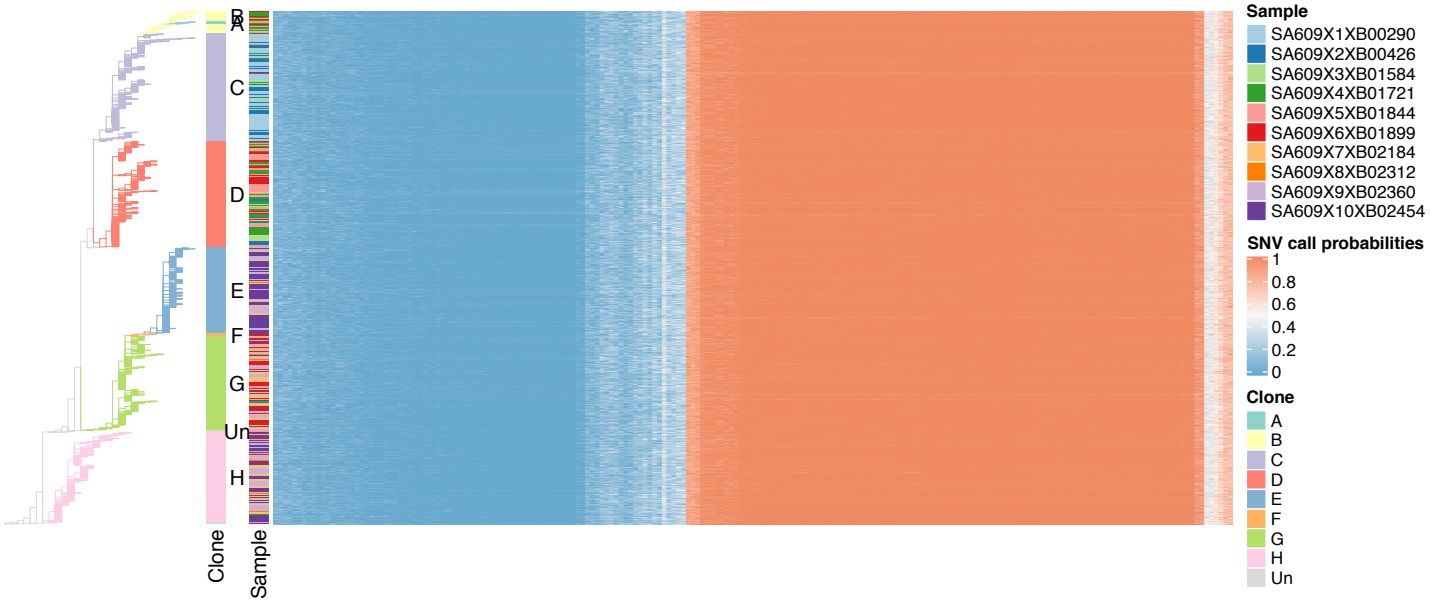

Figure 4.33: The heatmap depicts the mean SNV call probability of 200 random loci across the cells of *SA*609 over 20 perturbed tree (quality control step with threshold equals 0.001 is applied). The blue colour depicts the low probabilities near 0, the white colour depicts the probabilities around 0.5, and the orange colour depicts the probabilities around 1. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

Figure 4.34: The heatmap depicts the number of variant reads observed in each cell of *SA*609 for 200 random loci passed the quality control step with threshold equals 0.001. No variant read is depicted in black, and reads with variant alleles are depicted in bright colours. The inferred underlying phylogeny tree based on CNV data is on the left side of the heatmap. The samples and the clones are colour coded. The clones are identified using the approach by Salehi et al. study [104].

4.5 Mutations in high impact genes

We first annotate (using SnpEff [23]) the SNV loci that pass the quality control step. The loci that are annotated as a high impact gene (or a non-synonymous coding genes) and are only present in at most two clones are identified. The clone specific genes for *OV*2295, *SA*532 and *SA*609 are listed in Tables 4.3, 4.4 and 4.5, respectively. For *OV*2295, there are SNVs in GLG1 that are specific to clone *C*, and there are SNVs in DNHD1 that are specific to clone *H*. For *SA*532, there are SNVs in BCCIP, PTDSS2, PLEKHH1, UBALD1, NEURL4, MGAT5B, DOCK6, CBLC, TRIM28, PARS2, TP53RK, FKBP7, SPEG, TRANK1, SEL1L3, GRIA1, HIST1H2AM, PSMB9 and DHRS13 genes that are only present in clone *A*. Some SNVs that are only present in clone *B* occur on SMARCA5; and some of the ones that are only present in clone *C* occur on NOC3L and SPEG. Moreover, SNVs on XIAP gene are only present in clone *D*. For *SA*609, there are SNVs that are only present in clone *D* and occur in TLX1, OTUD7A, ZNF112 and MROH2A; and there are SNVs only present in clone *E* and occur in ARHGAP21, SLC12A5 and ZNF622. Some of the clone *H* specific mutations occur in DGKA, CARHSP1, PANK4, TTN, LRRC61, BAI1, MAGEC3 and NHS.

We explore if any of these genes are in cancer genes census. We find TLX1 in clone D of SA609 and CBLC in clone A of SA532 in cancer gene census. Also, we find SMARCA5 in clone B of SA532, and SLC34A2 in clone B and C of SA609. Both of the genes are reported as translocation partner genes. Reviewing the literature, it is interesting that we find studies reporting the above genes effective in different cancers. Hideshima et al. report TP53RK confers poor prognosis in multiple myeloma tumours. Huang et al. report XIAP possesses a critical role in promotion of cell survival and maintenance of cellular homeostasis for breast and colon carcinoma [57]. BCCIP is reported to have a role in the maintenance of genomic integrity [58, 85]. Jin et al. report overexpression of SMARCA5 correlates with cell proliferation and migration in breast cancer [62]. Yu et al. report SLC12A5 as an oncogene in clone cancer [138]. Further gene expression analysis can be done to confirm if any of the identified mutations has a significant impact in the dynamics of studied tumours.

Clone	Gene list
C	GLG1
H	DNHD1

Table 4.3: High impact and non-synonymous coding genes identified having SNVs that are only present in at most 2 clones of *OV*2295.

Clone	Gene list
A	BCCIP, PTDSS2, PLEKHH1, UBALD1, NEURL4, GPX4 GRIA1, HIST1H2AM, PSMB9, DHRS13, HSPB6 MGAT5B, DOCK6, CBLC, TRIM28, PARS2, TP53RK, FKBP7, SPEG, TRANK1, SEL1L3,
B	SMARCA5
C	NOC3L, SPEG
D	XIAP
B, D	PYCRL , CCKBR, SFI1, ANK2

Table 4.4: High impact and non-synonymous coding genes identified having SNVs that are only present in at most 2 clones of *SA*532.

4.6 Conclusion

We applied CellMutScope to three real datasets each compromising thousands of clonally related cells. The model detects SNVs per cell from the input data with very low breadth and depth coverage (See Table 4.2). Although the results show that better performance is achieved with a higher coverage, the applicable coverage is still lower than the ones achieved in other existing methods. In order to have a DLP data with higher coverage we can either increase the number of lanes or synthetically group similar cells as a pseudo-cell with higher coverage. The latter approach can introduce some errors as cells with different genomic profiles maybe mistakenly grouped. The results demonstrates the CellMutScope identifies clone specific mutations that are typically considered hardly detectable. In some of the existing clones, minor clonal structures (that were not distinguished with only the CNV data) are identified. A few of clone specific SNVs occur in high impact and non-synonymous coding genes. This may suggest the underling genetic pathways. The next step is to validate if these mutations have real impacts on the expansion

Clone	Gene list
D	TLX1, OTUD7A, ZNF112, MROH2A
E	ARHGAP21, SLC12A5, ZNF622
C, B	SHISA2, NOX5, TNFAIP8L1, OTUD7B, MIEF1, CLNK, SLC34A2
E, G	OTUD1, GCGR, MON1A, RASSF1, CWH43, DNAJC25, CTNNBIP1

Table 4.5: High impact and non-synonymous coding genes identified having SNVs that are only present in at most 2 clones of *SA*609.

or shrinkage of their associated clones. Single cell transcriptome sequencing data should be exploited to investigate if any of these mutations has any impact on the expression of the genes.

Chapter 5

Conclusion

"Learn from yesterday, live for today, hope for tomorrow. The important thing is not to stop questioning."

- Albert Einstein

5.1 Summary of contributions

This dissertation outlines two computational methods and their application on real world cancer datasets. The methods can be used to identify the portfolio of SNVs through incorporation of clonal information. The accurate identification of SNVs benefits the inference of tumour dynamics.

Mutation detection and classification through probabilistic integration of clonal population structure

We developed a statistical method, MuClone, to detect and classify mutations from multiple clonally related tumour whole genome or exome sequencing samples. Our method models SNV genotype, normal contamination, and clonal prevalence, all of which confound the detection of low prevalence SNVs. We showed that incorporating clonal information improves the detection of SNVs, particularly the low prevalence ones. In addition, we confirmed multi sample bulk sequencing data is a viable experimental design (e.g rapid autopsy program that at the time of a pa-

tient's death, tens to hundreds of metastatic samples are collected for future study) which can be exploited to interpret evolutionary properties of cancer.

Single cell somatic mutation detection through incorporation of the underlying phylogeny

We developed a statistical model, CellMutScope, to detect mutations across thousands of clonally related sequenced single cells. Incorporating the underlying phylogeny and copy number profile of cells, enabled us to overcome the low breadth and depth coverage of the data. By sharing statistical strength, we showed that the model imputes the missing data and detects SNVs in each cell. We also frame how we can incorporate SNV data into the Corrupt model (i) to infer the underlying tree phylogeny using SNV data with or without CNV data, or (ii) to extend the underlying CNV data with SNV loci placed on the tree. Therefore, we confirmed that with single cell sequencing data, it is possible to reveal the genomic profile of the underlying tumour clones in more detail.

5.2 Future work and discussion

This thesis focused on identification and interpretation of SNVs across clonally related tumour samples while exploiting clonal information. The first proposed model in Chapter 2 is restricted to the use of information from a flat clustering of clones. We believe exploiting all information about the underlying phylogeny of clones will improve the performance of detecting mutations. Although tumour evolution is the result of accumulation of SNVs with other types of genomic aberrations, the proposed method is only focused on the detection of SNVs. An important subject for future research is to integrate analysis of different aberrations in tumours with the underlying phylogeny to increase accuracy of detecting genomic aberrations, and to decipher tumour evolutionary process. We expect the joint analysis of genomic data will improve the accuracy of detecting genomic aberrations. Exploiting bulk genome sequencing data combined with other type of data (e.g. single cell sequencing data, or transcriptome sequencing data) can provide a more comprehensive insight.

With access to single cell sequencing data, we proposed a model in Chapter 3

to detect SNVs in every individual cell by incorporating the underlying phylogenetic tree. The model helps distinguish clone specific mutations and it discloses minor clone structure in the data. Identification of driver SNVs and their fitness can be extended to include all genomic aberrations. This should enhance the capacity of interpretations in tracking various aberrations. In addition, measuring gene expression of tumour clones links functional consequences to somatic aberrations. Therefore, genomic and transcriptomic single cell sequencing data can complement each other to give us insights on changes of gene expressions as a tumour evolve. This enables us to reconstruct tumours life histories.

It is observed that cells evolve through accumulation of genomic aberrations, epigenetic modifications and translation. This results in the heterogeneity of cell populations. Single cell sequencing data together with other biological information, open an avenue for probing the tumour dynamics in various patients by identifying distinct patterns. The difference in tumour dynamic patterns can lead towards a novel classification of tumour subtypes. This can help clinical decisions and treatments. The emerging field is called personalized medicine in which the treatment is personalized based on an individual's genomic profile.

Our understanding about cancer as a very complex disease can be improved by the study of the evolutionary process of cells under different conditions. Extensions of the experimental techniques like cell lines, patient derived xenografted samples, and genetically engineered mouse models can provide data to systematically study the evolutionary dynamics of cells. For example, we can study the evolutionary consequences of introducing mutations into the model or investigate the response of xenograft populations to therapeutic intervention (like gene knockouts). The models presented here can be applied at scale to comprehensively analyse SNVs in different experimental designs. The results can help reveal the mechanisms of tumourigenesis, treatment resistance, and metastatic progression in different conditions.

In spite of advances in sequencing technologies and different experiments, little is known about how to predict the evolutionary dynamics of cell populations, and how this information should be used in practice. Little findings from different experiments will add up, and will eventually enhance our understanding of tumour dynamics. Therefore, another milestone lies in interrelationship of clinical deci-

sions and computational results. Here, there are critical questions that are must be answered: What can we learn from inferred tumour dynamics? How should we aggregate computational results for clinically valid conclusions? How many computational cases are sufficient for clinical conclusions? Are the discovered tumour dynamic patterns reoccurring? If they are, how long the inferred data will remain valid in an evolving population? How frequent should we expect to find a de novo tumour dynamic pattern? How much do the parameters used in the analysis (e.g. number of samples or distribution of longitudinal/anatomical samples) affect the results? All these questions reflect the ambiguities and uncertainties laid under this field; and they should set our expectations about the impact of the field in practice.

I hope the advances in computational methods, in particular the ones we described here help improve our understanding about cancer and eventually help save more lives.

Bibliography

[1] Corrupt. URL https://github.com/UBC-Stat-ML/nowellpack. → pages 55, 57, 59, 60, 63, 65, 73, 84

[2] URL https://www.breastcancer.org. → page 83

[3] Mutect2. URL https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_cancer_m2_MuTect2.php. → pages 13, 35

[4] K. Anderson, C. Lutz, F. W. van Delft, C. M. Bateman, Y. Guo, S. M. Colman, H. Kempski, A. V. Moorman, I. Titley, J. Swansbury, L. Kearney, T. Enver, and M. Greaves. Genetic variegation of clonal architecture and propagating cells in leukaemia. Nature, 469(7330):356–61, 2011. → page 3

[5] L. Angus, M. Smid, S. M. Wilting, J. van Riet, A. Van Hoeck, L. Nguyen, S. Nik-Zainal, T. G. Steenbruggen, V. C. G. Tjan-Heijnen, M. Labots, J. M. G. H. van Riel, H. J. Bloemendal, N. Steeghs, M. P. Lolkema, E. E. Voest, H. J. G. van de Werken, A. Jager, E. Cuppen, S. Sleijfer, and J. W. M. Martens. The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. Nature Genetics, 51(10): 1450–1458, 2019. → page 82

[6] S. A. Aparicio and D. G. Huntsman. Does massively parallel dna resequencing signify the end of histopathology as we know it? The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland, 220(2):307–315, 2010. → page 8

[7] C. E. Barbieri, S. C. Baca, M. S. Lawrence, F. Demichelis, M. Blattner, J.-P. Theurillat, T. A. White, P. Stojanov, E. Van Allen, N. Stransky, E. Nickerson, S.-S. Chae, G. Boysen, D. Auclair, R. C. Onofrio, K. Park,

N. Kitabayashi, T. Y. MacDonald, K. Sheikh, T. Vuong, C. Guiducci, K. Cibulskis, A. Sivachenko, S. L. Carter, G. Saksena, D. Voet, W. M. Hussain, A. H. Ramos, W. Winckler, M. C. Redman, K. Ardlie, A. K. Tewari, J. M. Mosquera, N. Rupp, P. J. Wild, H. Moch, C. Morrissey, P. S. Nelson, P. W. Kantoff, S. B. Gabriel, T. R. Golub, M. Meyerson, E. S. Lander, G. Getz, M. A. Rubin, and L. A. Garraway. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nature Genetics, 44(6):685–689, 2012. → page 82

[8] A. Bashashati, G. Ha, A. Tone, J. Ding, L. M. Prentice, A. Roth, J. Rosner, K. Shumansky, S. Kalloger, J. Senz, W. Yang, M. McConechy, N. Melnyk, M. Anglesio, M. T. Y. Luk, K. Tse, T. Zeng, R. Moore, Y. Zhao, M. A. Marra, B. Gilks, S. Yip, D. G. Huntsman, J. N. McAlpine, and S. P. Shah. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. Journal of Pathology, 231(1):21–34, 2013. → page 3

[9] T. Baslan, J. Kendall, L. Rodgers, H. Cox, M. Riggs, A. Stepansky, J. Troge, K. Ravi, D. Esposito, B. Lakshmi, M. Wigler, N. Navin, and J. Hicks. Genome-wide copy number analysis of single cells. Nature Protocols, 7(6):1024–1041, 2012. → page 56

[10] T. Baslan, J. Kendall, B. Ward, H. Cox, A. Leotta, L. Rodgers, M. Riggs, S. D'Italia, G. Sun, M. Yong, K. Miskimen, H. Gilmore, M. Saborowski, N. Dimitrova, A. Krasnitz, L. Harris, M. Wigler, and J. Hicks. Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. Genome Research, 125(5):714–724, 2015. → page 10

[11] Y. Benjamini and T. P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic acids research, 40(10): e72–e72, may 2012. → page 56

[12] C. L. Bohrson, A. R. Barton, M. A. Lodato, R. E. Rodin, L. J. Luquette, V. V. Viswanadham, D. C. Gulhan, I. Cortés-Ciriano, M. A. Sherman, M. Kwon, M. E. Coulter, A. Galor, C. A. Walsh, and P. J. Park. Linked-read analysis identifies mutations in single-cell DNA-sequencing data. Nature Genetics, 51(4):749–754, 2019. → pages 15, 58

[13] D. D. Bowtell, S. Böhm, A. A. Ahmed, P.-J. Aspuria, R. C. Bast, V. Beral, J. S. Berek, M. J. Birrer, S. Blagden, M. A. Bookman, J. D. Brenton, K. B. Chiappinelli, F. C. Martins, G. Coukos, R. Drapkin, R. Edmondson, C. Fotopoulou, H. Gabra, J. Galon, C. Gourley, V. Heong, D. G. Huntsman,

M. Iwanicki, B. Y. Karlan, A. Kaye, E. Lengyel, D. A. Levine, K. H. Lu, I. A. McNeish, U. Menon, S. A. Narod, B. H. Nelson, K. P. Nephew, P. Pharoah, D. J. Powell, P. Ramos, I. L. Romero, C. L. Scott, A. K. Sood, E. A. Stronach, and F. R. Balkwill. Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. Nature Reviews Cancer, 15(11):668–679, 2015. → page 82

[14] I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. W. Kinzler, B. Vogelstein, and M. A. Nowak. Accumulation of driver and passenger mutations during tumor progression. Proceedings of the National Academy of Sciences, 107(43):18545–18550, 2010. → page 2

[15] B. B. Campbell, N. Light, D. Fabrizio, M. Zatzman, F. Fuligni, R. de Borja, S. Davidson, M. Edwards, J. A. Elvin, K. P. Hodel, W. Zahurancik, Z. Suo, T. Lipman, K. Wimmer, C. P. Kratz, D. C. Bowers, T. W. Laetsch, G. P. Dunn, T. M. Johanns, M. R. Grimmer, I. V. Smirnov, V. Larouche, D. Samuel, A. Bronsema, M. Osborn, D. Stearns, P. Raman, K. A. Cole, P. B. Storm, M. Yalon, E. Opocher, G. Mason, G. A. Thomas, M. Sabel, B. George, D. S. Ziegler, S. Lindhorst, V. M. Issai, S. Constantini, H. Toledano, R. Elhasid, R. Farah, R. Dvir, P. Dirks, A. Huang, M. A. Galati, J. Chung, V. Ramaswamy, M. S. Irwin, M. Aronson, C. Durno, M. D. Taylor, G. Rechavi, J. M. Maris, E. Bouffet, C. Hawkins, J. F. Costello, M. S. Meyn, Z. F. Pursell, D. Malkin, U. Tabori, and A. Shlien. Comprehensive Analysis of Hypermutation in Human Cancer. Cell, 171: 1042–1056.e10, 2017. → page 2

[16] P. J. Campbell, S. Yachida, L. J. Mudie, P. J. Stephens, E. D. Pleasance, L. A. Stebbings, L. A. Morsberger, C. Latimer, S. McLaren, M.-L. Lin, D. J. McBride, I. Varela, S. A. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, C. A. Griffin, J. Burton, H. Swerdlow, M. A. Quail, M. R. Stratton, C. Iacobuzio-Donahue, and P. A. Futreal. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature, 467(7319):1109–13, 2010. → page 3

[17] B. L. Cantarel, D. Weaver, N. McNeill, J. Zhang, A. J. Mackey, and J. Reese. Baysic: a bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC bioinformatics, 15(1):104, 2014. → page 14

[18] S. Carreira, A. Romanel, J. Goodall, E. Grist, R. Ferraldeschi, S. Miranda, D. Prandi, D. Lorente, J.-S. Frenel, C. Pezaro, A. Omlin, D. N. Rodrigues, P. Flohr, N. Tunariu, J. S de Bono, F. Demichelis, and G. Attard. Tumor

clone dynamics in lethal prostate cancer. Science translational medicine, 6 (254):254ra125, 2014. → page 3

[19] J. Carrot-Zhang and J. Majewski. Lolopicker: detecting low allelic-fraction variants from low-quality cancer samples. Oncotarget, 8(23):37032, 2017. → pages 13, 14

[20] Z. R. Chalmers, C. F. Connelly, D. Fabrizio, L. Gay, S. M. Ali, R. Ennis, A. Schrock, B. Campbell, A. Shlien, J. Chmielecki, F. Huang, Y. He, J. Sun, U. Tabori, M. Kennedy, D. S. Lieber, S. Roels, J. White, G. A. Otto, J. S. Ross, L. Garraway, V. A. Miller, P. J. Stephens, and G. M. Frampton. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Medicine, 9(1):34, 2017. → page 2

[21] D. Chen, H. Zhen, Y. Qiu, P. Liu, P. Zeng, J. Xia, Q. Shi, L. Xie, Z. Zhu, Y. Gao, G. Huang, J. Wang, H. Yang, and F. Chen. Comparison of single cell sequencing data between two whole genome amplification methods on two sequencing platforms. Scientific Reports, 8(1):4963, 2018. → page 56

[22] K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, and G. Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology, 31(3):213–219, 2013. → pages 13, 20

[23] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly, 6(2): 80–92, 2012. → page 122

[24] . G. P. Consortium. A map of human genome variation from population-scale sequencing. Nature, 467(7319):1061–1073, 2010. → page 12

[25] I. H. G. S. Consortium et al. Finishing the euchromatic sequence of the human genome. Nature, 431(7011):931, 2004. → page 6

[26] C. S. Cooper, R. Eeles, D. C. Wedge, P. Van Loo, G. Gundem, L. B. Alexandrov, B. Kremeyer, A. Butler, A. G. Lynch, N. Camacho, C. E. Massie, J. Kay, H. J. Luxton, S. Edwards, Z. Kote-Jarai, N. Dennis, S. Merson, D. Leongamornlert, J. Zamora, C. Corbishley, S. Thomas, S. Nik-Zainal, S. O'Meara, L. Matthews, J. Clark, R. Hurst, R. Mithen, R. G. Bristow, P. C. Boutros, M. Fraser, S. Cooke, K. Raine, D. Jones,

A. Menzies, L. Stebbings, J. Hinton, J. Teague, S. McLaren, L. Mudie, C. Hardy, E. Anderson, O. Joseph, V. Goody, B. Robinson, M. Maddison, S. Gamble, C. Greenman, D. Berney, S. Hazell, N. Livni, C. Fisher, C. Ogden, P. Kumar, A. Thompson, C. Woodhouse, D. Nicol, E. Mayer, T. Dudderidge, N. C. Shah, V. Gnanapragasam, T. Voet, P. Campbell, A. Futreal, D. Easton, A. Y. Warren, C. S. Foster, M. R. Stratton, H. C. Whitaker, U. McDermott, D. S. Brewer, and D. E. Neal. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. Nature genetics, 47(4):367–72, 2015. → page 3

[27] E. C. de Bruin, N. McGranahan, R. Mitter, M. Salm, D. C. Wedge, L. Yates, M. Jamal-Hanjani, S. Shafi, N. Murugaesu, A. J. Rowan, E. Grönroos, M. A. Muhammad, S. Horswell, M. Gerlinger, I. Varela, D. Jones, J. Marshall, T. Voet, P. Van Loo, D. M. Rassl, R. C. Rintoul, S. M. Janes, S.-M. Lee, M. Forster, A. Tanya, D. Lawrence, M. Falzon, A. Capitanio, T. T. Harkins, C. C. Lee, W. Tom, E. Teefee, S.-C. Chen, S. Bengum, A. Rabinowitz, B. Phillimore, B. Spencer-Dene, G. Stamp, Z. Szallasi, N. Matthews, A. Stewart, P. Campbell, and C. Swanton. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science, 346(6206):251–256., 2014. → page 3

[28] P. Del Moral, A. Doucet, and A. Jasra. Sequential monte carlo samplers. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(3):411–436, 2006. → page 64

[29] M. G. Denis, A. Vallée, and S. Théoleyre. EGFR T790M resistance mutation in non small-cell lung carcinoma. Clinica Chimica Acta, 444: 81–85, 2015. → page 3

[30] A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris. Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome biology, 16(1):35, 2015. → page 19

[31] J. Ding, A. Bashashati, A. Roth, A. Oloumi, K. Tse, T. Zeng, G. Haffari, M. Hirst, M. A. Marra, A. Condon, S. Aparicio, and S. P. Shah. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. Bioinformatics, 28(2):167–175, 2012. → pages 14, 20, 35, 91

[32] L. Ding, M. J. Ellis, S. Li, D. E. Larson, K. Chen, J. W. Wallis, C. C. Harris, M. D. McLellan, R. S. Fulton, L. L. Fulton, et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature, 464(7291):999, 2010. → pages 9, 10

[33] X. Dong, L. Zhang, B. Milholland, M. Lee, A. Y. Maslov, T. Wang, and J. Vijg. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. Nature methods, 14(5):491—493, May 2017. → pages 15, 58

[34] F. Dorri, S. Jewell, A. Bouchard-Côté, and S. P. Shah. Somatic mutation detection and classification through probabilistic integration of clonal population information. Communications Biology, 2(1):44, 2019. → page 16

[35] T. Dunn, G. Berry, D. Emig-Agius, Y. Jiang, S. Lei, A. Iyer, N. Udar, H.-Y. Chuang, J. Hegarty, M. Dickover, B. Klotzle, J. Robbins, M. Bibikova, M. Peeters, and M. Strömberg. Pisces: an accurate and versatile variant caller for somatic and germline next-generation sequencing data. Bioinformatics, 35(9):1579–1581, 10 2018. → page 14

[36] P. Eirew, A. Steif, J. Khattra, G. Ha, D. Yap, H. Farahani, K. Gelmon, S. Chia, C. Mar, A. Wan, et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. Nature, 518(7539):422, 2015. → pages 3, 19

[37] E. Falconer, M. Hills, U. Naumann, S. S. S. Poon, E. A. Chavez, A. D. Sanders, Y. Zhao, M. Hirst, and P. M. Lansdorp. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nature methods, 9(11):1107–1112, nov 2012. → page 56

[38] E. C. Friedberg, L. D. McDaniel, and R. A. Schultz. The role of endogenous and exogenous DNA damage and mutagenesis. Current Opinion in Genetics & Development, 14(1):5–10, 2004. → page 2

[39] E. Garrison and G. Marth. Haplotype-based variant detection from short-read sequencing. Preprint at arXiv:1207.3907v2, 2012. → pages 13, 20, 35

[40] C. Gawad, W. Koh, and S. R. Quake. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. Proceedings of the National Academy of Sciences of the United States of America, 111(50):17947–52, 2014. → page 56

[41] C. Gawad, W. Koh, and S. R. Quake. Single-cell genome sequencing: current state of the science. Nature Review Genetics, 17(3):175–188, 2016. → page 56

[42] M. Gerlinger, A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N. Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. R. Santos, M. Nohadani, A. C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P. A. Futreal, and C. Swanton. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. 366(10):883–892, 2012. → pages 3, 19

[43] M. Gerlinger, S. Horswell, J. Larkin, A. J. Rowan, M. P. Salm, I. Varela, R. Fisher, N. McGranahan, N. Matthews, C. R. Santos, P. Martinez, B. Phillimore, S. Begum, A. Rabinowitz, B. Spencer-Dene, S. Gulati, P. A. Bates, G. Stamp, L. Pickering, M. Gore, D. L. Nicol, S. Hazell, P. A. Futreal, A. Stewart, and C. Swanton. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nature genetics, 46(3):225–33, 2014. → page 3

[44] M. Gerstung, C. Beisel, M. Rechsteiner, P. Wild, P. Schraml, H. Moch, and N. Beerenwinkel. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nature communications, 3:811, 2012. → pages 13, 14

[45] O. Gevaert, V. Villalobos, B. I. Sikic, and S. K. Plevritis. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. Interface focus, 3(4):20130013, aug 2013. → page 82

[46] R. Govindan, L. Ding, M. Griffith, J. Subramanian, N. D. Dees, K. L. Kanchi, C. A. Maher, R. Fulton, L. Fulton, J. Wallis, K. Chen, J. Walker, S. Mcdonald, R. Bos, D. Ornitz, D. Xiong, M. You, D. J. Dooling, and M. Watson. GENOMIC LANDSCAPE OF NON-SMALL CELL LUNG CANCER IN SMOKERS AND NEVER SMOKERS. Cell, 150(6): 1121–1134, 2012. → page 19

[47] R. Goya, M. G. F. Sun, R. D. Morin, G. Leung, G. Ha, K. C. Wiegand, J. Senz, A. Crisan, M. A. Marra, M. Hirst, D. Huntsman, K. P. Murphy, S. Aparicio, and S. P. Shah. SNVMix: Predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics, 26(6): 730–736, 2010. → pages 14, 20

[48] M. Greaves and C. C. Maley. Clonal evolution in cancer. Nature, 481 (7381):306, 2012. → page 2

[49] W. W. R. H. J. G. Bloom. Histological grading and prognosis of breast cancer. British Journal of Cancer, 11(3):359–377, 1957. → page 3

[50] G. Ha, A. Roth, J. Khattra, J. Ho, D. Yap, L. M. Prentice, N. Melnyk, A. McPherson, A. Bashashati, E. Laks, et al. Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome research, 24(11):1881–1893, 2014. → pages 21, 35

[51] M. C. Haffner, T. Mosbruger, D. M. Esopi, H. Fedor, C. M. Heaphy, D. A. Walker, N. Adejola, M. Gurel, J. Hicks, A. K. Meeker, M. K. Halushka, J. W. Simons, W. B. Isaacs, A. M. De Marzo, W. G. Nelson, and S. Yegnasubramanian. Tracking the clonal origin of lethal prostate cancer. Journal of Clinical Investigation, 123(11):4918–4922, 2013. → page 3

[52] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. cell, 100(1): 57–70, 2000. → page 2

[53] N. F. Hansen, J. J. Gartner, L. Mei, Y. Samuels, and J. C. Mullikin. Shimmer: detection of genetic alterations in tumors using next-generation sequence data. Bioinformatics, 29(12):1498–1503, 2013. → page 14

[54] J. Hård, E. Al Hakim, M. Kindblom, Å. K. Björklund, B. Sennblad, I. Demirci, M. Paterlini, P. Reu, E. Borgström, P. L. Ståhl, J. Michaelsson, J. E. Mold, and J. Frisén. Conbase: a software for unsupervised discovery of clonal somatic mutations in single cells through read phasing. 20(1):68, 2019. → pages 15, 58

[55] G. H. Heppner and B. E. Miller. Tumor heterogeneity: biological implications and therapeutic consequences. Cancer and Metastasis Review, 2(1):5–23, 1983. → page 3

[56] Y. Hou, L. Song, P. Zhu, B. Zhang, Y. Tao, X. Xu, F. Li, K. Wu, J. Liang, D. Shao, H. Wu, X. Ye, C. Ye, R. Wu, M. Jian, Y. Chen, W. Xie, R. Zhang, L. Chen, X. Liu, X. Yao, H. Zheng, C. Yu, Q. Li, Z. Gong, M. Mao, X. Yang, L. Yang, J. Li, W. Wang, Z. Lu, N. Gu, G. Laurie, L. Bolund, K. Kristiansen, J. Wang, H. Yang, Y. Li, X. Zhang, and J. Wang. Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm. Cell, 148(5):873–885, 2012. → page 56

[57] X. Huang, X.-n. Wang, X.-d. Yuan, W.-y. Wu, P. E. Lobie, and Z. Wu. XIAP facilitates breast and colon carcinoma growth via promotion of p62 depletion through ubiquitination-dependent proteasomal degradation. Oncogene, 38(9):1448–1460, 2019. → page 122

[58] S. C. Huhn, J. Liu, C. Ye, H. Lu, X. Jiang, X. Feng, S. Ganesan, E. White, and Z. Shen. Regulation of spindle integrity and mitotic fidelity by BCCIP. Oncogene, 36(33):4750–4766, 2017. → page 122

[59] Illumina. An introduction to next generation sequencing technology, January 2011. URL http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf. → page 8

[60] M. Jamal-Hanjani, G. A. Wilson, N. McGranahan, N. J. Birkbak, T. B. Watkins, S. Veeriah, S. Shafi, D. H. Johnson, R. Mitter, R. Rosenthal, et al. Tracking the evolution of non–small-cell lung cancer. New England Journal of Medicine, 376(22):2109–2121, 2017. → pages 19, 37, 38, 40

[61] R. Jeselsohn, G. Buchwalter, C. De Angelis, M. Brown, and R. Schiff. ESR1 mutations—a mechanism for acquired endocrine resistance in breast cancer. Nature reviews. Clinical oncology, 12(10):573–583, oct 2015. → page 3

[62] Q. Jin, X. Mao, B. Li, S. Guan, F. Yao, and F. Jin. Overexpression of SMARCA5 correlates with cell proliferation and migration in breast cancer. Tumor Biology, 36(3):1895–1902, 2015. → page 122

[63] M. Josephidou, A. G. Lynch, and S. Tavaré. multiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples. Nucleic acids research, 43(9):e61, 2015. → pages 15, 20, 35

[64] I. Kalatskaya, Q. M. Trinh, M. Spears, J. D. McPherson, J. M. Bartlett, and L. Stein. Isown: accurate somatic mutation identification in the absence of normal tissue controls. Genome medicine, 9(1):59, 2017. → pages 14, 15

[65] Z. Kan, B. S. Jaiswal, J. Stinson, V. Janakiraman, D. Bhatt, H. M. Stern, P. Yue, P. M. Haverty, R. Bourgon, J. Zheng, M. Moorhead, S. Chaudhuri, L. P. Tomsho, B. A. Peters, K. Pujara, S. Cordes, D. P. Davis, V. E. H. Carlton, W. Yuan, L. Li, W. Wang, C. Eigenbrot, J. S. Kaminker, D. A. Eberhard, P. Waring, S. C. Schuster, Z. Modrusan, Z. Zhang, D. Stokoe, F. J. de Sauvage, M. Faham, and S. Seshagiri. Diverse somatic mutation

patterns and pathway alterations in human cancers. Nature, 466(7308): 869–873, 2010. → page 82

[66] C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, M. D. M. Leiserson, C. A. Miller, J. S. Welch, M. J. Walter, M. C. Wendl, T. J. Ley, R. K. Wilson, B. J. Raphael, and L. Ding. Mutational landscape and significance across 12 major cancer types. Nature, 502(7471):333–339, 2013. → page 3

[67] S. Kim, K. Jeong, K. Bhutani, J. H. Lee, A. Patel, E. Scott, H. Nam, H. Lee, J. G. Gleeson, and V. Bafna. Virmid: accurate detection of somatic mutations with sample impurity inference. Genome biology, 14(8):R90, 2013. → pages 12, 13, 20

[68] M. Kleppe and R. L. Levine. Tumor heterogeneity confounds and illuminates: assessing the implications. Nature medicine, 20(4):342–4, 2014. → page 3

[69] D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, and L. Ding. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics, 25(17):2283–2285, 2009. → pages 14, 20

[70] C. Kockan, F. Hach, I. Sarrafi, R. H. Bell, B. McConeghy, K. Beja, A. Haegert, A. W. Wyatt, S. V. Volik, K. N. Chi, et al. Sinvict: ultra-sensitive detection of single nucleotide variants and indels in circulating tumour dna. Bioinformatics, 33(1):26–34, 2016. → pages 14, 15

[71] R. Kridel, F. C. Chan, A. Mottok, M. Boyle, P. Farinha, K. Tan, B. Meissner, A. Bashashati, A. McPherson, A. Roth, et al. Histological transformation and progression in follicular lymphoma: a clonal evolution study. PLoS medicine, 13(12):e1002197, 2016. → page 19

[72] J. Kuipers, P. Suter, and G. Moffa. Efficient structure learning and sampling of bayesian networks. 2018. → pages 57, 77

[73] S. Kumar and S. Subramanian. Mutation rates in mammalian genomes. Proceedings of the National Academy of Sciences, 99(2):803–808, 2002. → page 2

[74] Z. Lai, A. Markovets, M. Ahdesmaki, B. Chapman, O. Hofmann, R. McEwen, J. Johnson, B. Dougherty, J. C. Barrett, and J. R. Dry. Vardict:

a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic acids research, 44(11):e108–e108, 2016. → page 14

[75] E. Laks, H. Zahn, D. Lai, A. McPherson, A. Steif, J. Brimhall, J. Biele, B. Wang, T. Masud, D. Grewal, C. Nielsen, S. Leung, V. Bojilova, M. Smith, O. Golovko, S. Poon, P. Eirew, F. Kabeer, T. R. de Algara, S. R. Lee, M. J. Taghiyar, C. Huebner, J. Ngo, T. Chan, S. Vatrt-Watts, P. Walters, N. Abrar, S. Chan, M. Wiens, L. Martin, R. W. Scott, M. T. Underhill, E. Chavez, C. Steidl, D. D. Costa, Y. Ma, R. J. N. Coope, R. Corbett, S. Pleasance, R. Moore, A. J. Mungall, M. A. Marra, C. Hansen, S. P. Shah, and S. Aparicio. Resource: Scalable whole genome sequencing of 40,000 single cells identifies stochastic aneuploidies, genome replication states and clonal repertoires. bioRxiv, 2018. → page 56

[76] E. Laks, A. McPherson, H. Zahn, D. Lai, A. Steif, J. Brimhall, J. Biele, B. Wang, T. Masud, J. Ting, D. Grewal, C. Nielsen, S. Leung, V. Bojilova, M. Smith, O. Golovko, S. Poon, P. Eirew, F. Kabeer, T. Ruiz de Algara, S. R. Lee, M. J. Taghiyar, C. Huebner, J. Ngo, T. Chan, S. Vatrt-Watts, P. Walters, N. Abrar, S. Chan, M. Wiens, L. Martin, R. W. Scott, T. M. Underhill, E. Chavez, C. Steidl, D. Da Costa, Y. Ma, R. J. N. Coope, R. Corbett, S. Pleasance, R. Moore, A. J. Mungall, C. Mar, F. Cafferty, K. Gelmon, S. Chia, G. J. Hannon, G. Battistoni, D. Bressan, I. Cannell, H. Casbolt, C. Jauset, T. Kovačević, C. Mulvey, F. Nugent, M. P. Ribes, I. Pearsall, F. Qosaj, K. Sawicka, S. Wild, E. Williams, S. Aparicio, E. Laks, Y. Li, C. O'Flanagan, A. Smith, T. Ruiz, S. Balasubramanian, M. Lee, B. Bodenmiller, M. Burger, L. Kuett, S. Tietscher, J. Windager, E. Boyden, S. Alon, Y. Cui, A. Emenari, D. Goodwin, E. Karagiannis, A. Sinha, A. T. Wassie, C. Caldas, A. Bruna, M. Callari, W. Greenwood, G. Lerda, Y. Lubling, A. Marti, O. Rueda, A. Shea, O. Harris, R. Becker, F. Grimaldi, S. Harris, S. Vogl, J. A. Joyce, J. Hausser, S. Watson, S. Shah, A. McPherson, I. Vázquez-García, S. Tavaré, K. Dinh, E. Fisher, R. Kunes, N. A. Walton, M. Al Sa'd, N. Chornay, A. Dariush, E. G. Solares, C. Gonzalez-Fernandez, A. K. Yoldas, N. Millar, X. Zhuang, J. Fan, H. Lee, L. S. Duran, C. Xia, P. Zheng, M. A. Marra, C. Hansen, S. P. Shah, and S. Aparicio. Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. Cell, 179(5):1207–1221.e22, 2019. → pages 11, 82, 84, 88, 95, 96, 102, 103, 104, 105, 116, 117

[77] D. A. Landau, S. L. Carter, P. Stojanov, A. McKenna, K. Stevenson, M. S. Lawrence, C. Sougnez, C. Stewart, A. Sivachenko, L. Wang, Y. Wan, W. Zhang, S. A. Shukla, A. Vartanov, S. M. Fernandes, G. Saksena,

K. Cibulskis, B. Tesar, S. Gabriel, N. Hacohen, M. Meyerson, E. S. Lander, D. Neuberg, J. R. Brown, G. Getz, and C. J. Wu. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell, 152(4): 714–726, 2013. → page 3

[78] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. Fitzhugh, et al. Erratum: Initial sequencing and analysis of the human genome: International human genome sequencing consortium (nature (2001) 409 (860-921)). Nature, 412(6846):565–566, 2001. → page 6

[79] D. Larson, C. Harris, K. Chen, D. Koboldt, T. Abbott, D. Dooling, T. Ley, E. Mardis, R. Wilson, and L. Ding. Somaticsniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics (Oxford, England), 28:311–7, 12 2011. → page 12

[80] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. Upset: visualization of intersecting sets. IEEE transactions on visualization and computer graphics, 20(12):1983–1992, 2014. → page 53

[81] Y. Liu, M. Loewer, S. Aluru, and B. Schmidt. Snvsniffer: an integrated caller for germline and somatic single-nucleotide and indel mutations. BMC systems biology, 10(2):47, 2016. → pages 12, 13

[82] I. C. Macaulay and T. Voet. Single Cell Genomics: Advances and Future Perspectives. PLoS Genetics, 10(1):1–9, 2014. → page 11

[83] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. Genome research, 20(9):1297–1303, 2010. → pages 15, 20

[84] A. McPherson, A. Roth, E. Laks, T. Masud, A. Bashashati, A. W. Zhang, G. Ha, J. Biele, D. Yap, A. Wan, L. M. Prentice, J. Khattra, M. A. Smith, C. B. Nielsen, S. C. Mullaly, S. Kalloger, A. Karnezis, K. Shumansky, C. Siu, J. Rosner, H. L. Chan, J. Ho, N. Melnyk, J. Senz, W. Yang, R. Moore, A. J. Mungall, M. A. Marra, A. Bouchard-Côté, C. B. Gilks, D. G. Huntsman, J. N. McAlpine, S. Aparicio, and S. P. Shah. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. Nature Genetics, 48(October 2015):758–767, 2016. → pages 19, 34, 36, 37

[85] X. Meng, J. Fan, and Z. Shen. Roles of BCCIP in chromosome stability and cytokinesis. Oncogene, 26(43):6253–6260, sep 2007. → page 122

[86] E. Muller, N. Goardon, B. Brault, A. Rousselin, G. Paimparay, A. Legros, R. Fouillet, O. Bruet, A. Tranchant, F. Domin, et al. Outlyzer: software for extracting low-allele-frequency tumor mutations from sequencing background noise in clinical practice. Oncotarget, 7(48):79485, 2016. → page 14

[87] G. H. Murillo, N. You, X. Su, W. Cui, M. P. Reilly, M. Li, K. Ning, and X. Cui. MultiGeMS: detection of SNVs from multiple samples using model selection on high-throughput sequencing data. Bioinformatics, 32 (10):1486–1492, 01 2016. → page 15

[88] N. Navin, A. Krasnitz, L. Rodgers, K. Cook, J. Meth, J. Kendall, M. Riggs, Y. Eberling, J. Troge, V. Grubor, D. Levy, P. Lundin, S. Maner, A. Zetterberg, J. Hicks, and M. Wigler. Inferring tumor progression from genomic heterogeneity. Genome Research, 20(1):68–80, 2010. → page 2

[89] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks, and M. Wigler. Tumour evolution inferred by single-cell sequencing. Nature, 472(7341):90–94, 2011. → pages 3, 10, 56, 60

[90] S. Nik-Zainal, P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna, A. Shlien, S. L. Cooke, J. Hinton, A. Menzies, L. A. Stebbings, C. Leroy, M. Jia, R. Rance, L. J. Mudie, S. J. Gamble, P. J. Stephens, S. McLaren, P. S. Tarpey, E. Papaemmanuil, H. R. Davies, I. Varela, D. J. McBride, G. R. Bignell, K. Leung, A. P. Butler, J. W. Teague, S. Martin, G. Jonsson, O. Mariani, S. Boyault, P. Miron, A. Fatima, A. Langerod, S. A. J. R. Aparicio, A. Tutt, A. M. Sieuwerts, k. Borg, G. Thomas, A. V. Salomon, A. L. Richardson, A. L. Borresen-Dale, P. A. Futreal, M. R. Stratton, and P. J. Campbell. The life history of 21 breast cancers. Cell, 149(5): 994–1007, 2012. → pages 3, 19

[91] S. M. Noordermeer and H. van Attikum. PARP Inhibitor Resistance: A Tug-of-War in BRCA-Mutated Cells. Trends in Cell Biology, 29(10): 820–834, oct 2019. → page 3

[92] P. C. Nowell. The clonal evolution of tumor cell populations. Science, 194 (4260):23–28, 1988. → page 2

[93] S. E. Park, K. Park, E. Lee, J.-Y. Kim, J. S. Ahn, Y.-H. Im, C. Lee, H. Jung, S. Y. Cho, W.-Y. Park, R. Cristescu, and Y. H. Park. Clinical implication of tumor mutational burden in patients with HER2-positive refractory metastatic breast cancer. OncoImmunology, 7(8):e1466768, aug 2018. → page 82

[94] E. D. Pleasance, P. J. Stephens, S. O'Meara, D. J. McBride, A. Meynert, D. Jones, M.-L. Lin, D. Beare, K. W. Lau, C. Greenman, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature, 463(7278):184, 2010. → page 9

[95] V. Popic, R. Salari, I. Hajirasouliha, D. Kashef-Haghighi, R. B. West, and S. Batzoglou. Fast and scalable inference of multi-sample cancer lineages. Genome biology, 16(1):91, 2015. → page 19

[96] P. Priestley, J. Baber, M. P. Lolkema, N. Steeghs, E. de Bruijn, C. Shale, K. Duyvesteyn, S. Haidari, A. van Hoeck, W. Onstenk, P. Roepman, M. Voda, H. J. Bloemendal, V. C. Tjan-Heijnen, C. M. van Herpen, M. Labots, P. O. Witteveen, E. F. Smit, S. Sleijfer, E. E. Voest, and E. Cuppen. Pan-cancer whole genome analyses of metastatic solid tumors. bioRxiv, 2019. → page 9

[97] B. K. Rajendran and C.-X. Deng. Characterization of potential driver mutations involved in human breast cancer by computational approaches. Oncotarget, 8(30):50252–50272, July 2017. → page 82

[98] A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg, A. O. M. Wilkie, G. McVean, and G. Lunter. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature genetics, 46(8):912–918, 2014. → page 13

[99] A. Roth, J. Ding, R. Morin, A. Crisan, G. Ha, R. Giuliany, A. Bashashati, M. Hirst, G. Turashvili, A. Oloumi, M. A. Marra, S. Aparicio, and S. P. Shah. JointSNVMix: A probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics, 28(7):907–913, 2012. → pages 12, 13, 20

[100] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. PyClone: statistical inference of clonal population structure in cancer. Nature Methods, 11(4):396–398, 2014. → pages 9, 19, 34

[101] A. Roth, A. McPherson, E. Laks, J. Biele, D. Yap, A. Wan, M. A. Smith, C. B. Nielsen, J. N. McAlpine, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. Clonal genotype and population structure inference from single-cell tumor sequencing. Nature Methods, 2016. → page 16

[102] A. J. L. Roth. Probabilistic models for the identification and interpretation of somatic single nucleotide variants in cancer genomes. PhD thesis, University of British Columbia, 2015. → page 56

[103] R. Salari, S. S. Saleh, D. Kashef-Haghighi, D. Khavari, D. E. Newburger, R. B. West, A. Sidow, and S. Batzoglou. Inference of tumor phylogenies with improved somatic mutation discovery. Journal of Computational Biology, 20(11):933–944, 2013. → page 20

[104] S. Salehi, F. Kabir, A. Bouchard-Côté, S. P. Shah, et al. Quantitative genotype and phenotype fitness modeling in cancerusing single cell timeseries population dynamics. → pages 84, 89, 90, 98, 99, 100, 101, 107, 108, 109, 110, 111, 112, 113, 114, 118, 119, 120, 121

[105] C. T. Saunders, W. S. W. Wong, S. Swamy, J. Becq, L. J. Murray, and R. K. Cheetham. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics, 28(14):1811–1817, 2012. → pages 13, 20, 35, 39, 91

[106] R. F. Schwarz, C. K. Y. Ng, S. L. Cooke, S. Newman, J. Temple, A. M. Piskorz, D. Gale, K. Sayal, M. Murtaza, P. J. Baldwin, N. Rosenfeld, H. M. Earl, E. Sala, M. Jimenez-Linan, C. A. Parkinson, F. Markowetz, and J. D. Brenton. Spatial and Temporal Heterogeneity in High-Grade Serous Ovarian Cancer: A Phylogenetic Analysis. PLoS Medicine, 12(2):1–20, 2015. → page 3

[107] S. P. Shah, R. D. Morin, J. Khattra, L. Prentice, T. Pugh, A. Burleigh, A. Delaney, K. Gelmon, R. Guliany, J. Senz, C. Steidl, R. A. Holt, S. Jones, M. Sun, G. Leung, R. Moore, T. Severson, G. A. Taylor, A. E. Teschendorff, K. Tse, G. Turashvili, R. Varhol, R. L. Warren, P. Watson, Y. Zhao, C. Caldas, D. Huntsman, M. Hirst, M. A. Marra, and S. Aparicio. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature, 461(7265):809–813, 2009. → pages 9, 10, 19

[108] S. P. Shah, A. Roth, R. Goya, A. Oloumi, G. Ha, Y. Zhao, G. Turashvili, J. Ding, K. Tse, G. Haffari, A. Bashashati, L. M. Prentice, J. Khattra, A. Burleigh, D. Yap, V. Bernard, A. McPherson, K. Shumansky, A. Crisan,

R. Giuliany, A. Heravi-Moussavi, J. Rosner, D. Lai, I. Birol, R. Varhol, A. Tam, N. Dhalla, T. Zeng, K. Ma, S. K. Chan, M. Griffith, A. Moradian, S.-W. G. Cheng, G. B. Morin, P. Watson, K. Gelmon, S. Chia, S.-F. Chin, C. Curtis, O. M. Rueda, P. D. Pharoah, S. Damaraju, J. Mackey, K. Hoon, T. Harkins, V. Tadigotla, M. Sigaroudinia, P. Gascard, T. Tlsty, J. F. Costello, I. M. Meyer, C. J. Eaves, W. W. Wasserman, S. Jones, D. Huntsman, M. Hirst, C. Caldas, M. A. Marra, and S. Aparicio. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature, 486(7403):395–9, 2012. → pages 3, 9, 10

[109] E. Shapiro, T. Biezuner, and S. Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics, 14:618, jul 2013. → page 56

[110] G. N. Sharma, R. Dave, J. Sanadya, P. Sharma, and K. K. Sharma. Various types and management of breast cancer: an overview. Journal of advanced pharmaceutical technology & research, 1(2):109–126, apr 2010. → page 83

[111] Y. Shiraishi, Y. Sato, K. Chiba, Y. Okuno, Y. Nagata, K. Yoshida, N. Shiba, Y. Hayashi, H. Kume, Y. Homma, M. Sanada, S. Ogawa, and S. Miyano. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. Nucleic Acids Research, 41(7):e89–e89, 03 2013. → page 13

[112] A. D. Siddiqui and B. Piperdi. KRAS Mutation in Colon Cancer: A Marker of Resistance to EGFR-I Therapy. Annals of Surgical Oncology, 17(4): 1168–1176, 2010. → page 3

[113] M. B. Siegel, X. He, K. A. Hoadley, A. Hoyle, J. B. Pearce, A. L. Garrett, S. Kumar, V. J. Moylan, C. M. Brady, A. E. Van Swearingen, et al. Integrated rna and dna sequencing reveals early drivers of metastatic breast cancer. The Journal of clinical investigation, 128(4), 2018. → page 19

[114] N. Silver, S. Best, J. Jiang, and S. L. Thein. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. BMC molecular biology, 7:33, oct 2006. → page 82

[115] J. Singer, J. Kuipers, K. Jahn, and N. Beerenwinkel. Single-cell mutation identification via phylogenetic inference. Nature Communications, 9(1): 5144, 2018. → pages 16, 59

[116] K. S. Smith, V. K. Yadav, S. Pei, D. A. Pollyea, C. T. Jordan, and S. De. Somvarius: somatic variant identification from unpaired tissue samples. Bioinformatics, 32(6):808–813, 2015. → pages 14, 15

[117] A. Sottoriva, I. Spiteri, S. G. M. Piccirillo, A. Touloumis, V. P. Collins, J. C. Marioni, C. Curtis, C. Watts, and S. Tavaré. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proceedings of the National Academy of Sciences of the United States of America, 110 (10):4009–14, 2013. → page 3

[118] A. Sottoriva, H. Kang, Z. Ma, T. A. Graham, M. P. Salomon, J. Zhao, P. Marjoram, K. Siegmund, M. F. Press, D. Shibata, and C. Curtis. A Big Bang model of human colorectal tumor growth. Nature Genetics, 47(3): 209–216, 2015. → page 3

[119] J.-F. Spinella, P. Mehanna, R. Vidal, V. Saillour, P. Cassart, C. Richer, M. Ouimet, J. Healy, and D. Sinnett. Snooper: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. BMC genomics, 17(1):912, 2016. → page 14

[120] M. R. Stratton. Exploring the genomes of cancer cells: progress and promise. science, 331(6024):1553–1558, 2011. → page 2

[121] H. Suzuki, K. Aoki, K. Chiba, Y. Sato, Y. Shiozawa, Y. Shiraishi, T. Shimamura, A. Niida, K. Motomura, F. Ohka, T. Yamamoto, K. Tanahashi, M. Ranjit, T. Wakabayashi, T. Yoshizato, K. Kataoka, K. Yoshida, Y. Nagata, A. Sato-Otsubo, H. Tanaka, M. Sanada, Y. Kondo, H. Nakamura, M. Mizoguchi, T. Abe, Y. Muragaki, R. Watanabe, I. Ito, S. Miyano, A. Natsume, and S. Ogawa. Mutational landscape and clonal architecture in grade II and III gliomas. Nat Genet, 47(5):458–468, 2015. → page 3

[122] C. Thirlwell, O. C. C. Will, E. Domingo, T. A. Graham, S. A. C. McDonald, D. Oukrif, R. Jeffrey, M. Gorman, M. Rodriguez-Justo, J. Chin-Aleong, S. K. Clark, M. R. Novelli, J. A. Jankowski, N. A. Wright, I. P. M. Tomlinson, and S. J. Leedham. Clonality Assessment and Clonal Ordering of Individual Neoplastic Crypts Shows Polyclonality of Colorectal Adenomas. Gastroenterology, 12(11):1058–1060, 2010. → page 3

[123] A. a. Tone, M. K. McConechy, W. Yang, J. Ding, S. Yip, E. Kong, K.-K. Wong, D. M. Gershenson, H. Mackay, S. Shah, B. Gilks, A. V. Tinker,

B. Clarke, J. N. McAlpine, and D. Huntsman. Intratumoral heterogeneity in a minority of ovarian low-grade serous carcinomas. BMC cancer, 14(1): 982, 2014. → page 3

[124] F. Vallania, E. Ramos, S. Cresci, R. D. Mitra, and T. E. Druley. Detection of rare genomic variants from pooled sequencing using splinter. JoVE (Journal of Visualized Experiments), (64):e3943, 2012. → page 14

[125] P. Van Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C. M. Perou, A.-L. Børresen-Dale, and V. N. Kristensen. Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences, 107 (39):16910–16915, 2010. → page 21

[126] K. E. Van Rens, V. Mäkinen, and A. I. Tomescu. SNV-PPILP: Refined SNV calling for tumor data using perfect phylogenies and ILP. Bioinformatics, 31(7):1133–1135, 2015. → pages 15, 20

[127] W. Wang, P. Wang, F. Xu, R. Luo, M. P. Wong, T.-W. Lam, and J. Wang. Fasd-somatic: a fast and accurate somatic snv detection algorithm for cancer genome sequencing data. Bioinformatics, 30(17):2498–2500, 2014. → page 12

[128] Y. Wang, J. Waters, M. L. Leung, A. Unruh, W. Roh, X. Shi, K. Chen, P. Scheet, S. Vattathil, H. Liang, A. Multani, H. Zhang, R. Zhao, F. Michor, F. Meric-Bernstam, and N. E. Navin. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature, 512(7513):1–15, 2014. → pages 10, 56, 60

[129] Z. Wei, W. Wang, P. Hu, G. J. Lyon, and H. Hakonarson. Snver: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic acids research, 39(19): e132–e132, 2011. → page 14

[130] M. J. Williams, A. Sottoriva, and T. A. Graham. Measuring clonal evolution in cancer with genomics. Annual Review of Genomics and Human Genetics, 20(1):309–329, 2019. → page 2

[131] A. Wilm, P. P. K. Aw, D. Bertrand, G. H. T. Yeo, S. H. Ong, C. H. Wong, C. C. Khor, R. Petric, M. L. Hibberd, and N. Nagarajan. Lofreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic acids research, 40(22):11189–11201, 2012. → page 13

[132] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjöblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, N. Silliman, S. Szabo, Z. Dezso, V. Ustyanksky, T. Nikolskaya, Y. Nikolsky, R. Karchin, P. A. Wilson, J. S. Kaminker, Z. Zhang, R. Croshaw, J. Willis, D. Dawson, M. Shipitsin, J. K. V. Willson, S. Sukumar, K. Polyak, B. H. Park, C. L. Pethiyagoda, P. V. K. Pant, D. G. Ballinger, A. B. Sparks, J. Hartigan, D. R. Smith, E. Suh, N. Papadopoulos, P. Buckhaults, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, V. E. Velculescu, and B. Vogelstein. The genomic landscapes of human breast and colorectal cancers. Science, 318(5853): 1108–1113, 2007. → page 82

[133] F. Wu. Irreversible Parallel Tempering and an Application to a Bayesian Nonparametric. PhD thesis, Oxford University, 2017. → pages 64, 73

[134] T. Xie, M. Musteanu, P. P. Lopez-Casas, D. J. Shields, P. Olson, P. A. Rejto, and M. Hidalgo. Whole exome sequencing of rapid autopsy tumors and xenograft models reveals possible driver mutations underlying tumor progression. PloS one, 10(11):e0142631, 2015. → page 19

[135] C. Xu. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Computational and Structural Biotechnology Journal, 16:15–24, 2018. → pages 11, 12, 13, 14

[136] S. Yachida, S. Jones, I. Bozic, T. Antal, R. Leary, B. Fu, M. Kamiyama, R. H. Hruban, J. R. Eshleman, M. A. Nowak, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, and C. A. Iacobuzio-Donahue. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature, 467 (7319):1114–7, 2010. → pages 3, 19

[137] L. R. Yates, M. Gerstung, S. Knappskog, C. Desmedt, G. Gundem, P. Van Loo, T. Aas, L. B. Alexandrov, D. Larsimont, H. Davies, Y. Li, Y. S. Ju, M. Ramakrishna, H. K. Haugland, P. K. Lilleng, S. Nik-Zainal, S. McLaren, A. Butler, S. Martin, D. Glodzik, A. Menzies, K. Raine, J. Hinton, D. Jones, L. J. Mudie, B. Jiang, D. Vincent, A. Greene-Colozzi, P.-Y. Adnet, A. Fatima, M. Maetens, M. Ignatiadis, M. R. Stratton, C. Sotiriou, A. L. Richardson, P. E. Lønning, D. C. Wedge, and P. J. Campbell. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nature medicine, 21(7):751–759, 2015. → page 3

[138] C. Yu, J. Yu, X. Yao, W. K. K. Wu, Y. Lu, S. Tang, X. Li, L. Bao, X. Li, Y. Hou, R. Wu, M. Jian, R. Chen, F. Zhang, L. Xu, F. Fan, J. He, Q. Liang, H. Wang, X. Hu, M. He, X. Zhang, H. Zheng, Q. Li, H. Wu, Y. Chen,

X. Yang, S. Zhu, X. Xu, H. Yang, J. Wang, X. Zhang, J. J. Y. Sung, Y. Li, and J. Wang. Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. Cell Research, 24(6):701–712, 2014. → page 122

[139] K. Yuan, T. Sakoparnig, F. Markowetz, and N. Beerenwinkel. Bitphylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. Genome biology, 16(1):36, 2015. → page 19

[140] H. Zafar, Y. Wang, L. Nakhleh, N. Navin, and K. Chen. Monovar: single-nucleotide variant detection in single cells. Nature Methods, 13:505, apr 2016. → pages 15, 58

[141] H. Zahn, A. Steif, E. Laks, P. Eirew, M. VanInsberghe, S. P. Shah, S. Aparicio, and C. L. Hansen. Scalable whole-genome single-cell library preparation without preamplification. Nature Methods, 14(2):167–173, Feb. 2017. → pages 10, 11, 56, 57, 60, 84

[142] A. W. Zhang, A. McPherson, K. Milne, D. R. Kroeger, P. T. Hamilton, A. Miranda, T. Funnell, N. Little, C. P. de Souza, S. Laan, et al. Interfaces of malignant and immunologic clonal dynamics in ovarian cancer. Cell, 173(7):1755–1769, 2018. → page 2

[143] J. Zhang, J. Fujimoto, J. Zhang, D. C. Wedge, X. Song, J. Zhang, S. Seth, C.-W. Chow, Y. Cao, C. Gumbs, K. A. Gold, N. Kalhor, L. Little, H. Mahadeshwar, C. Moran, A. Protopopov, H. Sun, J. Tang, X. Wu, Y. Ye, W. N. William, J. J. Lee, J. V. Heymach, W. K. Hong, S. Swisher, I. I. Wistuba, and P. A. Futreal. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science (New York, N.Y.), 346(6206):256–9, 2014. → page 3

[144] Y. Zhou, A. M. Johansen, and J. A. Aston. Toward automatic model comparison: An adaptive sequential monte carlo approach. Journal of Computational and Graphical Statistics, 25(3):701–726, 2016. → page 64

[145] C. Zong, S. Lu, A. R. Chapman, and X. S. Xie. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science, 338(6114):1622–1626, 2012. → page 56