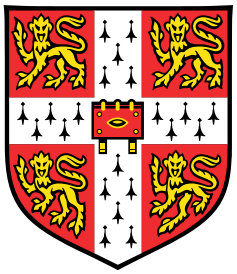
Inferring tumour evolution from single-cell and multi-sample data



Edith M. Ross

Cancer Research UK Cambridge Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Clare College March 2018

Summary

Inferring tumour evolution from single-cell and multi-sample data

Edith M. Ross

Tumour development has long been recognised as an evolutionary process during which cells accumulate mutations and evolve into a mix of genetically distinct cell subpopulations. The resulting genetic intra-tumour heterogeneity poses a major challenge to cancer therapy, as it increases the chance of drug resistance. To study tumour evolution in more detail, reliable approaches to infer the life histories of tumours are needed. This dissertation focuses on computational methods for inferring trees of tumour evolution from single-cell and multi-sample sequencing data.

Recent advances in single-cell sequencing technologies have promised to reveal tumour heterogeneity at a much higher resolution, but single-cell sequencing data is inherently noisy, making it unsuitable for analysis with classic phylogenetic methods. The first part of the dissertation describes OncoNEM, a novel probabilistic method to infer clonal lineage trees from noisy single nucleotide variants of single cells. Simulation studies are used to validate the method and to compare its performance to that of other methods. Finally, OncoNEM is applied in two case studies.

In the second part of the dissertation, a comprehensive collection of existing multi-sample approaches is used to infer the phylogenies of metastatic breast cancers from ten patients. In particular, shallow whole-genome, whole exome and targeted deep sequencing data are analysed. The inference methods comprise copy number and point mutation based approaches, as well as a method that utilises a combination of the two. To improve the copy number based inference, a novel allele-specific multi-sample segmentation algorithm is presented. The results are compared across methods and data types to assess the reliability of the different methods.

In summary, this thesis presents substantial methodological advances to understand tumour evolution from genomic profiles of single cells or related bulk samples.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text or declared below:

1. Oscar M. Rueda and Stephen-John Sammut performed the QDNAseq analysis and the variant calling in Chapter 4.

2. Stephen-John Sammut performed the PyClone analysis in Chapter 4.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution.

It does not exceed the prescribed word limit for the Degree Committee of the Faculties of Clinical Medicine and Veterinary Medicine.

Edith M. Ross
March 2018

This dissertation contains material from the following manuscripts:

- E.M. Ross and F. Markowetz
  OncoNEM: Inferring tumour evolution from single-cell sequencing data.
  Genome Biology 2016; 17:69

- E.M. Ross, K. Haase, P. Van Loo and F. Markowetz
  Allele-specific multi-sample copy number segmentation
  bioRxiv:166017 | doi:10.1101/166017

- L. De Mattos Arruda†, S.-J. Sammut†, E.M. Ross, et al.
  The integrated genomic and immune landscapes of lethal metastatic breast cancer
  in preparation

- E.M. Ross and F. Markowetz
  What are Dirichlet process mixture models?
  in preparation

Other manuscripts published or prepared during the course of the PhD:

- V.B. Kuchler†, E.S. Polson†, C. Abbosh‡, E.M. Ross‡, R.K. Mathew, H.A. Beard, E. Chuntharpursat, J. Williams, A. Patel, A. Droop, D.J. Beech, P. Chumas, S.C. Short, M. Lorger, R.S. Bon, S.J. Allison, S. Zhu, F. Markowetz, and H. Wurdak
  The small molecule KHS101 exhausts energy metabolism in glioblastoma cells
  submitted

- M.A.A. Castro, I. de Santiago, T.M. Campbell, C. Vaughn, T.E. Hickey, E. Ross, W.D. Tilley, F. Markowetz, B.A.J. Ponder and K.B. Meyer
  Regulators of genetic risk of breast cancer identified by integrative network analysis
  Nature Genetics 2016; 48(1):12-21

† and ‡ denote equal contribution.

Acknowledgements

I would like to thank my supervisor Florian Markowetz for giving me the opportunity to conduct my research in his group, for proposing the topic of this thesis, for his guidance, support and trust over the past years.

Thank you also to past and present members of the Markowetz lab for their friendship and for stimulating discussions. In particular, I thank Geoff Macintyre, Ines de Santiago and Andrew Holding for their advice throughout the years. Special thanks go to Anne Trinh, Leon Chlon and Ruben Drews for being the best PhD buddies I could have wished for.

I am grateful to Heiko Wurdak, Leticia De Mattos Arruda, Stephen-John Sammut and Carlos Caldas for wonderful collaborations.

I would also like to acknowledge the Cancer Research UK Cambridge Institute for funding my research and for providing a fantastic work environment. Furthermore, I would like to thank Ann Kaminski for supporting all of the PhD students at CRUK CI in the background.

Thank you to all members of Clare Boat Club for incredible memories.

Finally, I thank Helge and my family for their love and support. I could not have done this without you.

Table of contents

Chapter 1

Introduction

This thesis contributes to several aspects of the inference of tumour evolution. My main contribution is a method to infer tumour evolution from single-cell sequencing data. Furthermore, I led the evolutionary analysis of a study of ten lethal metastatic breast cancers and adapted existing methods to infer tumour evolution from multi-sample bulk sequencing data.
This first chapter gives an overview of tumour evolution and the challenges it poses for treating the disease (Section 1.1). It explains different sequencing approaches to analyse the genomic markers of tumour evolution and describes the main challenges for inferring tumour phylogenies (Section 1.2).

1.1 Tumour evolution

The somatic evolution of cells within a multicellular organism explains both how cancer arises and why curing it is so difficult [104]. Advances in genome sequencing have enabled researchers to study the mutations of a tumour at an unprecedented resolution. However, sequencing on its own reveals little about the development and the genetic heterogeneity of a tumour. Computational methods are needed to infer tumour phylogenies, which describe the evolutionary history of a tumour while also providing information about the genetic heterogeneity within the tumour at the time of sampling. Inference of tumour phylogenies is the main topic of this thesis.

Cancer development Cancer develops through an evolutionary process, which is based on two fundamental biological processes (i) the accumulation of mutations and (ii) natural selection acting on the phenotypic differences caused by the mutations. Cells acquire

mutations through various mechanisms including exposure to mutagens as well as cell-intrinsic processes such as the replication of deoxyribonucleic acid (DNA) [158]. Most mutations do not have an effect on the phenotype of the cell and are called *neutral* or *passenger mutations* [166]. Alternatively, they can be disadvantageous, in which case they may slow down cell division or cause cell death. Some mutations, however, can change the phenotype of a cell in a way that confers a growth advantage, i.e. they increase the ratio of cell birth to death. These mutations are called *drivers* [166]. If a cell acquires such a mutation, this can lead to an increased proliferation of the cell compared to its neighbours. This is called a *clonal expansion* [115]. Clonal expansions of cells are common in the human body. Most of these clones have a limited growth potential and lead to benign growths like skin moles [158]. The accumulation of mutations, however, does not have to stop there. Cells within an expanding clone accumulate further mutations. A descendent clone carries all mutations of its parent clone in addition to its newly acquired mutations. This leads to a subset pattern of mutations, where early clones contain a subset of the mutations of the later clones. Through this multi-step process of accumulation of advantageous mutations and subsequent clonal expansions, a single founder cell can grow into a heterogeneous tumour mass.

Eventually, cells can become cancerous, which means that in addition to the uncontrolled division of tumour cells, they acquire the ability to invade nearby tissues and spread to distant organs through the bloodstream to form metastases [56]. Metastatic cancer is largely incurable, and more than 90% of cancer deaths are due to metastases [162], highlighting the importance of studying this process. Emerging evidence indicates that the spreading of cancer cells can occur through multiple routes and in different directions [161, 93]. Metastases can be seeded by clusters of circulating tumour cells [1]. Cells that have been generated through the ongoing evolution within a metastasis can invade another already existing metastasis in a process called *cross-metastasis seeding* [161]. Furthermore, cells from a metastasis can re-infiltrate the primary tumour (*self-seeding*). In combination, these processes can lead to complex mixtures of cell populations within the primary and metastatic sites.

The mutational landscape of tumours During tumour evolution cells acquire different types of mutations. A very common type are *single-nucleotide variants* (SNVs) [3]. These are single base-pair substitutions. Mutations that affect more than one base-pair are called *structural variants* which include indels [3], copy number aberrations [175] as well as inversions and translocations [106]. *Indels* are mutations that lead to a small insertion or deletion of one or a few nucleotides. *Copy number aberrations* (CNAs) in contrast lead to gains or losses of large parts of the genome, ranging in size from a few kilobases to whole

chromosomes. In addition to genetic changes, epigenetic changes can also play a role in tumour evolution [77, 61].

The time frame of cancer evolution is extremely variable, ranging from years to several decades [8, 52]. This also applies to the rate at which mutations occur. While mutations may accumulate slowly at the beginning of tumour development, the mutation rate can increase once genes that, under normal circumstances, are responsible for maintaining a stable genome have been mutated themselves [87, 11, 88]. Consequently, the number of mutations found in tumours varies widely. It is particularly small in paediatric tumours, which develop over short time frames and harbour an average of 9.6 non-synonymous mutations per tumour. *Non-synonymous mutations* are mutations that change the encoded amino acid sequence of a protein [171]. Melanomas and lung tumours show a high number of mutations containing approximately 200 non-synonymous mutations per tumour, reflecting the role of mutagens in the aetiology of the disease [166]. Tumours that have deficiencies in DNA damage repair pathways can harbour thousands of non-synonymous mutations [166]. For breast cancers, which are the subject of study in Chapter 4, the median number of non-synonymous mutations is estimated to be around 33 [166].

Selective forces shape tumour evolution The proliferative success of a tumour cell depends on its genetic background and the resulting consequences on the phenotype of the cell. Whether or not a mutation acts as a driver or passenger can change over time as it depends on the selective pressures acting on the cell. Furthermore, it can depend on the genetic background of the cell in which it occurs [154]. Nevertheless, researchers have defined a list of *driver genes* which are considered to be relevant for cancer development if mutated in certain ways or aberrantly expressed [166]. Usually, the vast majority of mutations in a tumour are passengers, with the number of driver gene mutations in a typical tumour ranging from two to eight [166].

Examples of selective pressures acting on the phenotypic differences caused by mutations are competition for space and resources such as oxygen. Other constraints are imposed by the tissue microenvironment, for example through the immune system, which can select against cells containing neoantigen-causing mutations [171]. Cancer treatments can be an extreme form of selective pressure, ideally eradicating all of the tumour cells but none of the normal ones. Emerging evidence suggests that in some cases tumours might not be subjected to any strong selective pressures, leading to neutral evolution, in which all the different clones in the tumour expand at the same rate [157].

Treating cancer in the light of evolution Tumour evolution has important implications for cancer treatment and outcome. Since the accumulation of mutations is a stochastic process, every cancer is different, a phenomenon called *inter-tumour heterogeneity*. Consequently, response to treatment is variable and there is a need for personalised treatments that are tailored towards the molecular features of the disease. In response to this, molecular subtypes have been defined for many cancer types, such as breast cancer [121, 156, 22], glioblastoma [165] and pancreatic cancer [7]. Subtypes have been shown to be prognostic of patient outcome and cancers of some subtypes show distinct responses to certain treatments [128]. Often subtype definitions are based on gene expression patterns only [121, 156], but some also integrate genomic features in the classification [22, 7] or were subsequently used to identify genomic features that are associated with subtypes [165].

In addition to inter-tumour heterogeneity, tumour evolution also leads to *intra-tumour heterogeneity*. This means that cells within a tumour differ from each other. This heterogeneity manifests itself spatially and temporally [150]. Accordingly, biopsies from a single region are unlikely to reveal the full heterogeneity of a tumour [50]. Nevertheless, single biopsies are commonly used to inform treatment decisions [50], because of the difficulties and risks associated with obtaining tissue biopsies [21]. This can thwart therapeutic success, as to avoid relapse, all cell populations within a tumour need to be eradicated by the treatment. Often, tumours shrink initially upon treatment administration but later regrow due to a resistant cell population. This can be caused by resistant cells that were already present in the tumour before the treatment [64, 29] or, given that tumours keep evolving for as long as they exist, by new mutations that were acquired during treatment [5]. In fact, emerging evidence suggests that therapeutic intervention can accelerate the growth of resistant populations by inducing new mutations and removing competing populations that are sensitive to the treatment [52, 70, 74].

To detect signs of regrowth earlier, methods that to monitor the treatment response non-invasively through circulating tumour cells [119, 89] and circulating tumour DNA (ctDNA) [30, 42] from liquid biopsies are being developed.

Why study tumour evolution? Given that every cancer is different, studying the evolution of a tumour long after a patient has died may seem futile. However, cancer development can be seen as a repeated evolutionary experiment [154]. Unlike the evolution of species, which is a single experiment that started billions of years ago and is still ongoing, cancer starts over and over again with every patient. Even though the exact genetic makeup of each cancer is different, the conditions in which cancers evolve are similar. This imposes constraints that may limit the range of possible trajectories for cancer evolution [154]. A

consequence of these mechanistic commonalities are the recurrently mutated cancer driver genes mentioned above.

Reliable methods for the inference of tumour life histories are important for several reasons. Inferring trees of evolution across many patients may help elucidate the unifying patterns of tumour evolution at a higher resolution. For example, instead of only analysing if a gene is mutated in a tumour, inferring trees enables us to test whether certain combinations of mutations occur in the same subpopulations. It may also provide the opportunity to obtain a clearer picture of the evolutionary dynamics of tumours, which could be a major stepping-stone towards enhanced treatment strategies for cancer. In the future, phylogenetic trees could also be of use in the clinic, as they enable the identification of clonal mutations which are likely to be better targets for therapy than subclonal mutations [117, 120]. Ultimately, studying tumour evolution could enable researchers to predict the next evolutionary step of a tumour [62] and could help to turn cancer into a manageable chronic disease, similar to HIV where treatment strategies are guided by evolutionary principles to slow down the emergence of resistance and the progression of the disease [37].

1.2 Inferring trees of tumour evolution

The evolution of cancer is not a new concept. In fact, Theodor Boveri already hypothesised more than 100 years ago that "every tumour has its origin in a single cell", which contains a "faulty assembly of chromosomes as a consequence of an abnormal event" and that the transformation from benign to malignant tumours is due to further changes in the chromosomes [14]. His theories were developed through cytological observations. Peter Nowell consolidated the theoretical framework for clonal evolution in his seminal paper from 1976 [115]. His conclusions were, however, still mainly drawn from observations at the chromosome level.

Rapid advances in sequencing technologies in recent years have dramatically changed our ability to analyse genomic changes. Genomes can now be studied at base-pair resolution at relatively low cost. Nevertheless, inferring the evolutionary history of a tumour remains challenging. The following section gives a brief overview of the basic approaches to inferring tumour phylogenies. It mainly focuses on SNV based methods as these are the most common ones. A more detailed description, including copy number based methods, can be found in Chapter 2.

Bulk versus single-cell sequencing Most methods developed for the inference of tumour evolution use data derived from *bulk sequencing* of tumour samples. Bulk sequencing means

that the genetic material of all cells in the sample is pooled and sequenced. The resulting sequencing reads cannot be traced back to the individual cells and the genotypes inferred are therefore a combination of the different cell populations in the sample. There are two main types of approaches for the inference of trees of tumour evolution from bulk samples.

The first kind of approach builds *sample trees* and can only be used if multiple, phylogenetically related samples are available [151, 81, 132]. It assumes that every bulk sample contains cells of a single population. If this assumption is violated, the inferred sample tree still reflects similarities between samples, but these can differ from the true underlying phylogenetic relationships [4].

The second type are *deconvolution* based approaches [139, 105, 27]. Unlike the sample tree approaches, these methods do not assume that all mutations of a sample occurred in a single population. Most deconvolution based methods consist of two inference steps. First, they deconvolve the mixed signal of the bulk sample to infer clusters of mutation which occurred together during tumour evolution [139]. In the second step, mutation cluster frequencies are used to infer evolutionary relationships between the clusters [126]. Together, the mutation clusters and their relationships define the genotypes of the subpopulations in the sample. Deconvolution-based approaches can be applied to both single and multiple samples [139, 105]. However, using multiple samples helps to infer the population structure at a higher resolution [68]. Furthermore, mutation cluster frequencies of a single sample are often compatible with multiple tree structures. In certain cases, multiple samples can resolve these ambiguities.

Recently, *single-cell sequencing* techniques have been developed. Unlike the mixed signal of bulk-sequencing, single-cell sequencing provides information about the *co-occurrence* of mutations. Multiple single-cell sequences are needed to infer a tree, and ideally, these cells should be a representative sample of the different subpopulations in the tumour. If genotype inference using single-cell sequencing was perfect, inferring the phylogeny of the cells could be done using classic phylogenetic methods such as maximum parsimony or neighbour-joining. However, due to technical limitations, variants inferred from single-cell sequencing are very noisy [169, 58]. Consequently, standard methods often fail to identify subpopulations within the sequenced cells, turning even a seemingly simple task, such as mapping cells to clones, into a challenge. This thesis contributes to the challenge of inferring trees from SNVs of single cells. It proposes a probabilistic model that exploits the subset relationships of mutation patterns between related cells and accounts for genotyping errors and missing data.

Data pre-processing Most point mutation based approaches require binary present/absent calls, read depths of the reference and mutant alleles or variant allele frequencies. These can be obtained using standard variant callers. *Germline* mutations are usually filtered out using matched normal samples, as only *somatic* mutations are informative about the structure of tumour phylogenies. In the following, we refer to these mutations as SNVs. In general, however, point mutation based methods apply to all mutations that can be considered as independent from each other, including indels.

Data pre-processing for copy number based approaches is often more difficult. There are two main challenges, which are related to the two main steps of inferring copy number profiles. First, the genome has to be *segmented* into regions of constant copy number [114]. Second, relative copy number values need to be converted into absolute integer values by estimating the *ploidy* and *purity* of the sample [164]. To date, the segmentation of allele-specific copy number data has been done on a per sample basis, even when multiple related samples are analysed. Noise in the data can cause misalignment of segment boundaries that are shared between related samples, which in turn leads to problems during tree inference. This thesis presents an allele-specific multi-sample segmentation algorithm to address this problem.

Benchmarking One of the biggest limitations for method development in the area of tumour phylogenies is the lack of a gold standard data set for method benchmarking. For any tumour, the ground truth tree is unknown, and new methods are often only validated using simulated data, which always contains biases.

At the same time, many of the application studies that infer trees only use a single method. This is partially because studies often only generate a single type of data, which in turn limits the range of methods that can be used. It is, however, difficult to estimate the uncertainty in the tree based on a single method, as many tools only generate a single solution that may seem definitive to the user. While resampling approaches like bootstrapping could be used to assess the stability when only one method is used, this is rarely done. Furthermore, resampling might not address method specific artefacts in the tree reconstruction.

The danger of ignoring uncertainty in the tree inference is that the biological conclusions drawn could be a consequence of artefacts in the tree reconstruction. Artefacts could include the inferred relationships between samples, the branch lengths as well as instances of spurious convergent evolution [4].

In an attempt to provide a more realistic benchmarking scenario for methods inferring tumour evolution, ICGC and TCGA jointly initiated a DREAM challenge (ICGC-TCGA-DREAM Somatic Mutation Calling Challenge –Tumor Heterogeneity and Evolution) [143].

The challenge consisted of several tasks, including the inference of tumour purity, the number of mutation clusters, their cellular prevalence and the phylogenetic relationships between subpopulations. The challenge was, however, limited to the benchmarking of single-sample bulk methods. A comprehensive comparison of multi-sample methods is still lacking.

To infer trees as part of an autopsy study, we decided to use a comprehensive set of multi-sample methods. Some of these methods infer sample trees while others perform deconvolution of subpopulation as part of the tree inference. Furthermore, the methods vary in the input data they require and the type of variants they analyse, with some being SNV based and others being copy number based or a combination of the two. This is possible due to the large variety of sequencing data available for the same tumour in this study, which included shallow whole genome sequencing, whole exome sequencing and targeted deep sequencing.

1.3 Thesis organisation

In summary, to infer tumour phylogenies, challenges need to be addressed at every level of the data analysis process (i) data pre-processing, in particular for copy number segmentation, (ii) tree inference, in particular in the area of single-cell phylogenies, and (iii) benchmarking of tree inference algorithms in realistic settings. This thesis contributes to all of these three components. The organisation of the thesis is outlined in the following.

Computational approaches to inferring tumour evolution Chapter 2 provides an overview of current approaches to inferring tumour phylogenies. It describes the basic principles and assumptions behind inference methods for tumour evolution. It focuses on methods that have been used to infer phylogenies from single-cell data and bulk sample-based approaches that are applicable to multiple related samples.

Inferring tumour evolution from single-cell sequencing data In Chapter 3 we develop a method for inferring clonal lineage trees from single-cell sequencing data. The main assumption of this method is that genotypes of the different cells are connected through subset relationships, which reflect the accumulation of mutations during tumour evolution. Unlike previous methods, we probabilistically account for genotyping errors to model the uncertainty associated with variant calls from single-cell sequencing. Furthermore, we develop a search algorithm, to efficiently search the space of possible trees. We validate our approach in simulation studies and apply it in two case studies.

Inferring tumour evolution using multi-sample bulk sequencing data Chapter 4 focuses on the inference of tumour phylogenies from bulk sequencing data of multiple related samples. We first present a new algorithm for allele-specific multi-sample segmentation to reduce noise in multi-sample copy number data. We then use this approach in combination with a representative set of multi-sample phylogeny methods to study the evolutionary history of 10 cases of lethal metastatic breast cancer. Depending on the method requirements we use shallow whole genome, whole exome or targeted sequencing data of the same samples. The chapter concludes with a comprehensive comparison of the inferred phylogenies across methods and data types to assess the robustness of the results.

Chapter 2

Computational methods for inferring tumour evolution

In this chapter I describe computational models for inferring tumour evolution. I start with defining the aim of tumour phylogenetics (Section 2.1) and describe the common evolutionary assumptions inference models are based on (Section 2.2). Section 2.3 reviews bulk-sequencing methods, while Section 2.4 discusses single-cell based approaches. This chapter contains text from Ross and Markowetz [136] and a figure from Ross and Markowetz [137].

2.1 Aim of tumour phylogenetics

Tumour phylogenetics aims to infer a tree that describes the evolutionary history of a tumour. Each *node* of the tree corresponds to a subpopulation of the tumour that is characterised by a distinct genotype. *Leaf nodes* correspond to current subpopulations that have been observed in the tumour sample. In contrast, *internal nodes* correspond to ancestral populations. Different from the evolution of species, ancestral populations do not have to be driven to extinction by tumour evolution. Therefore, internal nodes can describe *observed* or *unobserved* ancestral populations. Populations can be unobserved due to extinction, or because they have not been sampled. The *edge* connecting two nodes corresponds to the mutations that distinguish the respective subpopulations. The number of mutations assigned to an edge is often encoded in its *length*. To reflect the process of evolution edges are *directed*.

In case of single-cell sequencing, each sample corresponds to a single subpopulation and can, therefore, be assigned to a single node of the tree. In case of bulk-sequencing, a sample can contain cells from multiple subpopulations. To describe the composition of a

bulk sample, edges are drawn from its constituent subpopulations to the sample, yielding a *directed acyclic graph*.

2.2 Assumptions

To reconstruct tumour phylogenies we have to make assumptions about how tumours evolve. Most methods make the following two assumptions. Firstly, every site is only mutated once. This is often referred to as the *infinite sites assumption* [91]. Secondly, mutations do not revert or disappear. Combined, these assumptions describe a *perfect* and *persistent phylogeny*, where mutations accumulate over time and are never lost [13]. These assumptions are not always justified. For example, deletions leading to loss of heterozygosity can remove previously acquired mutations.

The main alternative to these assumptions are *maximum parsimony* or *minimum evolution* methods, which are based on the principle of Occam's razor. These approaches allow for multiple changes including back mutations at the same site but assume that the true underlying tree is the one that requires the lowest number of changes to explain the data [38]. Noisy data can, however, limit the applicability of these approaches.

2.3 Bulk-sequencing based methods

Bulk samples can contain cells from multiple subpopulations. Most bulk-sequencing based methods aim to identify the different subpopulations in one or multiple related samples as well as their phylogenetic relationships. This section gives an introduction to the most commonly used approaches in this field. An overview of bulk-sequencing methods is shown in Table 2.1 at the end of this section.

2.3.1 Data

Currently, the most commonly used sequencing technologies produce short reads. Illumina's HiSeq 4000 platform, for example, generates reads with a maximum length of 2×150 base-pairs [63]. As point mutations are distributed sparsely across the genome, a single read rarely covers multiple mutations [174]. At the same time, bulk sequencing reads do not contain information about their cellular origin. Therefore, we cannot tell if two reads stem from the same or different cells. In summary, this means that bulk sequencing does not provide direct information about which mutations occur together in a cell. In the future, this might improve due to longer sequencing reads or synthetic long reads [122, 79]. Given the

current setting, however, computational methods are needed to infer the co-occurrence of mutations.

2.3.2 Basic idea

Most bulk methods identify groups of co-occurring mutations by clustering them according to their frequency. The motivation behind this approach is based on the perfect and persistent phylogeny assumption described above. Under this assumption, two mutations, A and B, that occurred at the same time in the same cell are always inherited together. This means that all cells that carry A also carry B and vice versa. Therefore, the *cellular prevalences* of these mutations are equal [139], where the cellular prevalence of a mutation is defined as the fraction of cancer cells that carry the mutation. It is also referred to as *cancer cell fraction* (CCF) [84].

Having the same cellular prevalence is a requirement for co-occurrence, but is not sufficient. Mutations that have the same frequency do not have to appear in the same cell. They could, for example, belong to different subpopulations that happen to have a similar cellular prevalence in the sample. Nevertheless, many deconvolution-based methods infer the co-occurrence of mutations purely based on mutation frequencies.

Once mutations have been clustered, cluster frequencies are used to infer phylogenetic relationships. In the following, we explain the clustering and tree reconstruction approaches in more detail. First, however, we describe how mutation frequencies are estimated from the sequencing data.

2.3.3 Estimating mutation frequencies

Bulk genome sequencing enables us to identify mutations and provides estimates of their *variant allele frequencies* (VAFs). The VAF of a mutation is the relative frequency of the variant allele in the population. It can be estimated directly from the sequencing data as the fraction of reads carrying the mutant allele. Naturally, the accuracy of VAF estimates increases with sequencing depth. To infer mutation clusters, VAFs need to be transformed into cellular prevalences. The cellular prevalence of a mutation differs from its VAF for three main reasons.

- **Germline ploidy** For diploid genomes, the fraction of cells carrying the mutation is twice as high as its VAF because every cell that carries the mutation contains one mutant and one wild-type allele.

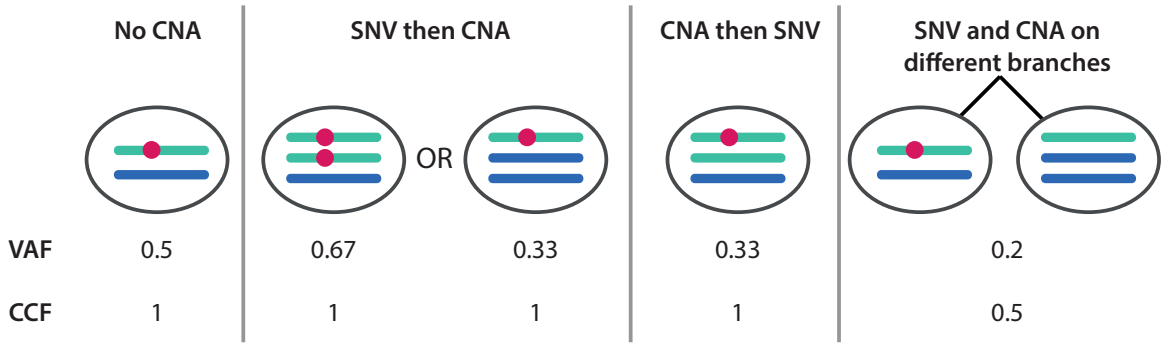


Fig. 2.1 The effect of a copy number change on the frequency of a variant allele (red dot) in the population depends on the phylogenetic relationship between the two mutation events. Various VAFs can correspond to the same cellular prevalence of the variant allele (CCF). The last case, where SNV and CNA lie on different branches, assumes that the two populations are present in equal proportions.

- **Purity** Tumour samples usually contain a mixture of tumour and normal cells. The *purity*, also referred to as *cellularity*, measures the fraction of tumour cells in the sample. Most methods require this as an input parameter. It is often estimated using copy number inference tools such as ASCAT [164] and ABSOLUTE [16]. Alternatively, it can be estimated from histopathology images of the tumour [139]. A low purity leads to small VAFs.

- **Copy number changes** The VAF of a mutation depends on the number of copies of a locus in a cell. In a normal diploid cell, this is two. Due to copy number changes the number of copies of different loci can vary widely.

Accounting for purity and germline ploidy is straightforward, as they affect the relationship between VAF and cellular prevalence in the same way for all mutations. Local copy number changes, in contrast, can affect VAFs in different ways. Apart from the location of SNV and CNA, the effect depends on the phylogenetic relationship between these mutations. This makes accounting for copy number changes much more difficult. Examples of how CNAs can impact VAFs are shown in Figure 2.1. To complicate things even further, copy number changes can overlap. This means that the VAF of a single point mutation can be affected by several copy number changes with independent phylogenetic relationships.

Most methods do not contain built-in functions to adjust VAFs for copy number changes. Some ignore copy number changes entirely, which can lead to spurious mutation clusters (see Figure 2.2). Others recommend excluding mutations that fall into regions of copy number changes [105, 176]. Copy number changes can, however, affect large parts of the genome. Therefore, restricting the analysis to regions without copy number changes may discard much

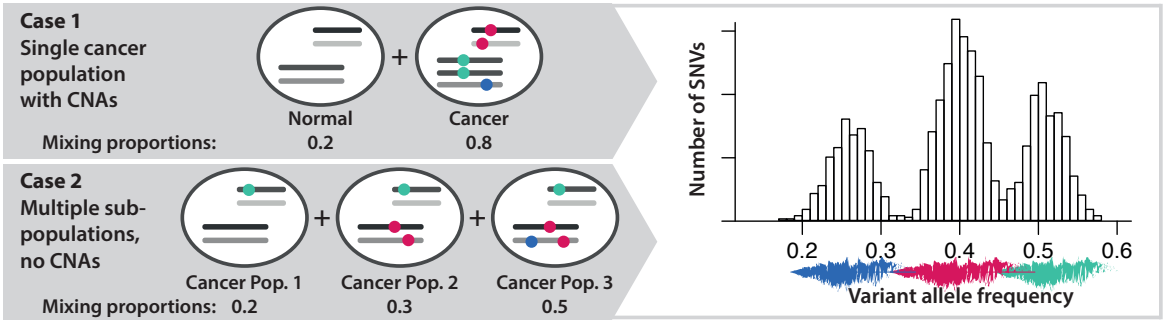


Fig. 2.2 Copy number changes (Case 1) and intra-tumour heterogeneity (Case 2) can both lead to a multi-modal distribution of variant allele frequencies. SNVs marked by dots are representatives of a much larger number of variants with the same phylogenetic history. Variant allele frequencies of each cluster show dispersion due to the noise introduced in the data generation process.

of the phylogenetic information, including disease-relevant driver mutations, limiting the biological interpretation of the inferred phylogenies. Furthermore, the clustering approaches that are used in the subsequent analysis require sufficiently large data sets to yield good results. For these reasons, robust estimation of cellular prevalences is an important part of the phylogenetic analysis.

PyClone [139] is one of the methods that model the impact of copy number changes on VAFs. It does this by integrating over the different possible relationships between CNA and SNV. However, it does not take into account the case where CNAs and SNVs occur in independent populations and makes further simplifying assumptions about the co-occurrence of CNA and SNV events that disregard some of the possible genotype states [139, 27, 84]. CHAT [84], PhyloWGS [27] and SuperFreq [145] estimate the phylogenetic relationship and take all of the possible phylogenetic combinations between an SNV and a single CNA into account. CHAT assigns phylogenetic relationships and explicitly calculates cellular prevalences using a probabilistic model that excludes ambiguous cases. SuperFreq, in turn, assigns phylogenetic relationships where possible and increases the uncertainty of the cellular prevalence estimates in ambiguous cases. PhyloWGS does not explicitly calculate cellular prevalences for each variant but instead incorporates the effect of copy number changes in its probabilistic clustering model. Rather than picking one phylogenetic relationship or marginalising over them, EXPANDS [6] calculates distributions of possible cellular prevalence values and clusters these distributions to infer mutation clusters. However, like PyClone it does not take into account all of the possible combinations of CNAs and SNVs.

Alternatively, where methods do not account for copy number changes, cellular prevalence estimates of third-party tools can be used as input instead of VAFs. One such tool,

which purely focuses on the estimation of cellular prevalences, is OncoPhase [17]. It uses the VAFs of phased germline variants as additional information for inferring the order between CNAs and somatic SNVs.

All of these approaches make important advances towards a more accurate estimation of cellular prevalences. Nevertheless, limitations remain due to the complexity of the problem. The common copy number inference tools infer copy number states of a single or at most two aberrant cell populations [27]. Heterogeneous samples can, however, contain many more subpopulations. In these cases, the estimated copy numbers are a mixture of the different subpopulations. This can lead to inaccuracies in the cellular prevalence estimates. Furthermore, the infinite sites assumption does not hold for CNAs because they involve large parts of the genome. Therefore, SNVs can be affected by multiple copy number changes with different phylogenetic relationships. This is generally not taken into account.

2.3.4 Clustering approaches

Assuming that cellular prevalences have been estimated, the next step is to infer mutation clusters. The most commonly used method for clustering mutation frequencies are mixture models. Mixture models in different variations are used by SciClone [105], PyClone, PhyloSub [68], PhyloWGS, CHAT, LICHeE [126] and CTPsingle [31]. The following section gives a brief introduction to mixture models.

2.3.4.1 Model-based clustering

Mixture models describe the data distribution $f(x)$ as a mixture of component distributions [43]

$$f(x) = \sum_{k=1}^{K} \pi_k F(x|\theta_k), \tag{2.1}$$

where K is the number of mixture components, π_k are the mixing proportions of the different components with $\sum_{k=1}^{K} \pi_k = 1$ and $F(x|\theta_k)$ is the density function of the mixture components with component specific parameters θ_k. The component density depends on the type of data being analysed. LICHeE and CHAT use Gaussian distributions to model mutation frequencies. PhyloSub, PhyloWGS, Clomial [176], CloneHD [39], Canopy [67] and TITAN [53] model counts of normal and variant alleles using binomial distributions, while PyClone, CTPsingle and Cloe [97] use beta-binomial distributions. Copy number read depths are modelled through multinomial distributions by THetA [116] and TITAN.

Inference Direct inference of the parameters that meet the maximum likelihood criterion is difficult. They can, however, be estimated using the Expectation Maximisation (EM) algorithm, as done by LICHeE. The EM is a general iterative optimisation algorithm. In the setting of mutation cluster inference, the algorithm alternates between assigning component responsibilities to each mutation (expectation step) and updating component parameters (maximisation step) [57].

Model selection Model-based clustering and other techniques used for tumour phylogeny inference require the number of components as input parameter. Since the number of subpopulations is not known *a priori* and to avoid overfitting, model selection is needed. To do this, models with varying numbers of clusters are inferred, and a selection criterion is used to choose the preferred model. Many different selection criteria exist. Methods including THetA, Clomial, CloneHD and Canopy use the Bayesian Information Criterion (BIC), which penalises for the number of parameters in the model. The BIC is defined as

$$\mathrm{BIC} = -2 \cdot \hat{L} + \log(N) \cdot d, \tag{2.2}$$

where $\hat{L}$ is the log-likelihood of a given model, N is the number of data points, and d is the number of free parameters of the model [57]. TITAN uses a measure called *S_Dbw* validity index [55], which minimises the variance within each cluster and maximises the density-based separation between clusters.

2.3.4.2 A Bayesian approach to mixture models

Instead of using model selection, mixture models can be extended to fully probabilistic non-parametric models that treat the selection of the number of components as part of the inference problem. The core of these approaches is a probabilistic model that is derived using a Bayesian framework, and that describes how the observed data is generated.

Finite mixture models Assuming the number of components K is known, defining the generative model is straightforward [51]: As in the simple mixture model case, the data stems from a set of component-wise distributions $F(x|\theta_k)$. However, the component parameters θ_k and component weights π_k are unknown *a priori*. For this reason, we treat them as variables, which follow certain probability distributions, so called *priors*. A convenient choice for the component weight prior is the *Dirichlet distribution*, which models the probability of K competing events. The prior over the component parameters depends on the data that is being analysed. In the following we denote it as $G_0(\theta_k)$. This defines a joint distribution over

the data points $\boldsymbol{x} = \{x_1, \ldots, x_N\}$, their cluster assignments $\boldsymbol{c} = \{c_1, \ldots, c_N\}$ and the cluster parameters $\theta = \{\theta_1, \ldots, \theta_K\}$

$$P(\boldsymbol{x}, \boldsymbol{c}, \theta) = \prod_{k=1}^{K} G_0(\theta_k) \prod_{n=1}^{N} F(x_n|\theta_{c_n}) P(c_n), \tag{2.3}$$

where $P(c_n)$ is the component weight prior for data point n with $c_n \in \{1, \ldots, k\}$.

Infinite mixture models As mentioned above, K is usually unknown and needs to be treated as a variable, too. This can be achieved using a probabilistic model called *Dirichlet process*. Like the sample from a Dirichlet distribution, a sample from the Dirichlet process can be interpreted as a sample of component weights. Instead of producing a fixed number of components K, however, the Dirichlet process generates an infinite number, where the component weights still sum up to 1. This is achieved by an iterative sampling procedure, where the weight of component k is sampled as

$$\pi_k \sim \text{Beta}(1, \alpha) \left(1 - \sum_{j=1}^{k-1} \pi_j\right). \tag{2.4}$$

Due to the infinite number of components these models are also called *infinite mixture models*. Alternatively, they are referred to as *Dirichlet process mixture models* [51]. Despite the name, the number of components that are populated by data points are finite in practice, because the number of data points is finite itself. Consequently, the clustering solution will consist of a finite number of clusters, too.

An intuitive way to understand the Dirichlet process is the *stick breaking* construction [2]. Given a stick of unit length, a piece of length π_k is broken off the stick, where π_k is determined by a draw from a Beta distribution $\text{Beta}(1, \alpha)$. The remaining stick is used for the next stick breaking step. This procedure is repeated an infinite number of times to generate the component weights. As the remaining stick becomes smaller and smaller, the sampled component weights become smaller too. The sampling procedure of *Dirichlet process mixture models* is summarised in Figure 2.3.

Inference The aim is to infer the posterior probability of cluster assignments given the data $P(\boldsymbol{c}|\boldsymbol{x})$. This is usually achieved using Bayes theorem. However in case of both finite and infinite mixture models, the posterior is intractable because it requires summing over all possible partitionings of the data with K components [51]. Nevertheless, the generative model enables us to approximate the posterior distribution. The most common procedure to

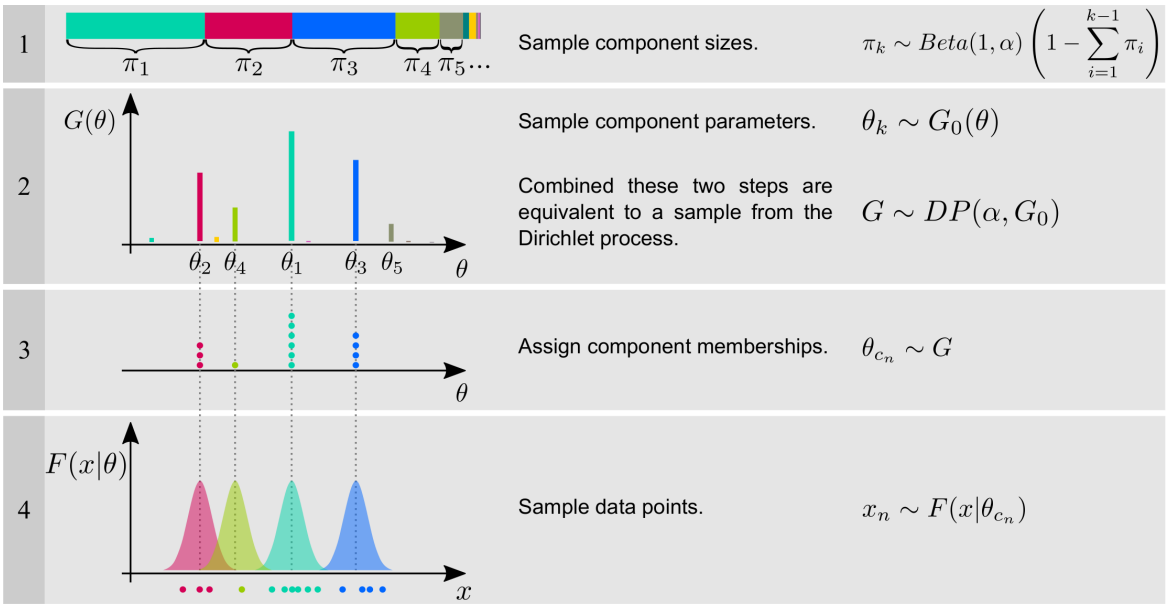


Fig. 2.3 Visual summary of the Dirichlet process mixture model sampling procedure.

do this is *Markov Chain Monte Carlo* (MCMC) sampling, which generates a sample from the posterior distribution. It is used by CHAT, PyClone, CTPsingle, PhyloSub, PhyloWGS and Canopy. The disadvantage of this approach is that a large number of steps may be required until the Markov Chain reaches its equilibrium distribution and it is difficult to assess convergence [12]. An alternative approach, *variational Bayes*, is used by SciClone. Instead of sampling, variational Bayes approximates the posterior distribution with a simpler family of distributions and minimises the Kullback-Leibler divergence between posterior and the approximative distribution [12]. An advantage of variational Bayes over MCMC is that it is usually faster, and convergence can be easily assessed, but it may yield biased results [51].

2.3.4.3 Clustering multiple samples

As explained above, frequency-based clustering may erroneously group mutations of different subpopulations, if they have the same cellular prevalence. Additional samples can increase the resolution of the clustering solution if they contain different fractions of the subpopulations [68] (see Figure 2.4). Most mixture model-based methods can cluster cellular prevalences of multiple related samples simultaneously by using multi-dimensional distributions.

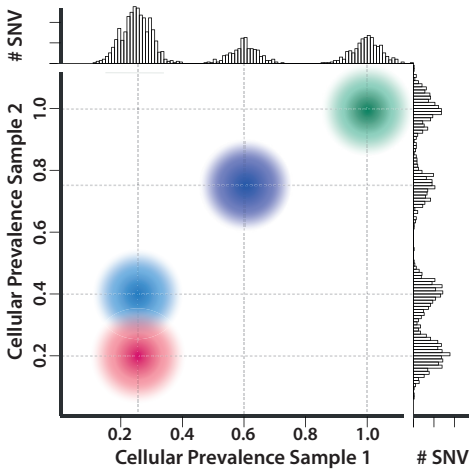


Fig. 2.4 Toy example of mutation clustering based on two samples. In sample 1 the two low frequency clusters (red and light blue) have the same cellular prevalence and can therefore not be distinguished. In sample 2, the two cluster occur at different frequencies and can therefore be resolved.

2.3.5 From mutation clusters to trees

In the following, we assume that mutations have successfully been clustered into groups. Many methods stop at this point and do not infer phylogenies. It is, however, important to note that mutation clusters do not directly correspond to the genotypes of the subpopulations. Each mutation cluster is thought to be the result of a clonal expansion. A subpopulation may, however, be the result of several successive rounds of clonal expansions. Therefore, the genotype of a subpopulation may be a combination of several mutation clusters.

To obtain the full SNV genotypes of the tumour subpopulations the phylogenetic relationships between mutation clusters need to be inferred. In this step, methods rely again on the perfect and persistent phylogeny assumptions described in Section 2.2. They can be used to derive two rules constraining the possible topologies of tumour phylogenies, called the *sum rule* and the *crossing rule* [68].

Sum rule The sum rule states that in a branching phylogeny where A is ancestral to both B and C, the sum of population frequencies of B and C must be lower or equal to the population frequency of A. If this is not the case, the phylogeny must be linear. This rule is also called the *pigeonhole principle* [112]. Consider the following example of three mutation clusters A, B and C, with $\mathrm{CCF}(A) = 0.8$, $\mathrm{CCF}(B) = 0.6$ and $\mathrm{CCF}(C) = 0.4$. A and B cannot belong to different branches of the phylogeny because the sum of their population frequencies is larger than 1. A sum larger than 1 could only be explained with a violation of the infinite sites

assumption. The same applies to A and C. Therefore, A must be an ancestor of both B and C. B and C in turn cannot belong to different branches, because $\text{CCF}(B) + \text{CCF}(C) > \text{CCF}(A)$. The sum rule can be applied to single as well as multiple samples.

Crossing rule The crossing rule states that two clusters A and B must belong to different branches of the phylogeny if $\text{CCF}(A) > \text{CCF}(B)$ in one of the samples and $\text{CCF}(B) > \text{CCF}(A)$ in another sample.

Since the crossing rule relies on multiple samples, the mutation clusters of a single sample are not sufficient to rule out a linear phylogeny, and even with multiple samples, these constraints do not always define a unique phylogeny. Likewise, it is possible that the frequencies of mutation clusters across multiple samples define incompatible phylogenies. This is because the methods described above infer mutation clusters without taking into account phylogenetic constraints.

Sum and crossing rules are used by all methods that infer phylogenies from mutation clusters, but methods differ in details of how they apply them. LICHeE, for example, requires an error range for cluster CCFs as input parameter and takes this into account when inferring the ordering between clusters. SCHISM [113] performs a hypothesis test for each pair of mutations (i, j) to infer if mutation i can be an ancestor of mutation j. In a second step, it aggregates the votes of each mutation pair across clusters to infer if cluster I can be an ancestor of cluster J. The resulting matrix is then used to infer phylogenies. Methods also differ in how they report ambiguities in the tree topologies. LICHeE infers multiple tree topologies, whereas PhyloWGS and PhyloSub report partial order graphs if unique solutions cannot be inferred. SuperFreq in turn always assumes a linear relationship, unless it can be ruled out via the crossing rule.

2.3.6 Other approaches

2.3.6.1 Joint clustering and tree building

Some methods attempt to infer clusters and tree structure jointly, to avoid the problem of phylogenetically incompatible mutation clusters.

PhyloSub and PhyloWGS use a *tree structured stick breaking* process [2]. This mixture model approach uses a tree-structured prior instead of a flat Dirichlet process prior to simultaneously infer clonal genotypes and their phylogenetic relationship. Like most mixture model-based methods, PhyloSub and PhyloWGS use MCMC sampling for inference.

However, the large space of possible phylogenetic trees can be problematic for sampling methods [35].

Other methods aim to decompose the observed data matrix of mutations across samples into a matrix of clonal genotypes ($|\text{mutations}| \times |\text{clones}|$) and a matrix of clone frequencies ($|\text{clones}| \times |\text{samples}|$). Varieties of this approach are used by CITUP [95] and AncesTree [34]. TrAp [159] aims to infer clonal genotypes from a single sample in a similar fashion but has to make strong additional assumptions such as sparsity constraints to address identifiability issues.

Clomial also uses matrix decomposition but does not impose phylogenetic constraints on the genotype matrix. The result may therefore not be compatible with a tree structure. Most of these methods infer solutions using integer linear programming (ILP) optimisation.

2.3.6.2 Relaxed phylogenetic assumptions

El-Kebir et al. [35] developed SPRUCE, which infers a *multi-state perfect phylogeny*. In this setting, a locus can change its state multiple times but may only change to the same state once. This is called the *infinite alleles assumption*. Marass et al. [97] developed a latent feature model, called Cloe, which allows for loss of mutations and convergent evolution.

2.3.6.3 Sample-based methods

In contrast to deconvolution-based approaches, other methods build trees directly on the mutation profiles of multiple samples of the same tumour. These methods either assume that mixing of subpopulations in a sample is rare or that each sample consists of a single subpopulation.

Limited mixing of subpopulations Recently, Reiter et al. [132] developed a point mutation based method called *Treeomics*, which assumes that samples rarely contain a mix of subclones from independent branches of the phylogenetic tree. While this cannot be assumed for primary tumours, previous research suggests that it may be a suitable assumption for samples from metastases, as metastasis formation is a bottleneck that reduces genetic diversity [103, 80, 161].

As input Treeomics uses two matrices of mutation sites versus samples, containing the total number of reads and the number of mutant reads. Based on the number of reference and mutant alleles, Treeomics calculates the posterior probability of a mutation being present at a given site in a given sample. It then extracts binary mutation patterns from the data, where each mutation pattern is defined by the set of samples which contain the variant.

For every unique mutation pattern, a reliability score is calculated based on the posterior mutation probabilities. Finally, a mixed integer linear program (MILP) solver is used to find the set of compatible mutation patterns which maximises the total reliability score. Mutation patterns are considered compatible if they do not violate the perfect and persistent phylogeny assumption. The compatible mutation patterns define the inferred phylogeny. In a post-processing step, incompatible mutation patterns with a high reliability score are used to identify mixing of subpopulations within samples. An advantage of this approach is that it does not rely on mutation frequencies to identify co-occurrence of mutations.

No mixing of subpopulations If samples consist of a single subpopulation, they can be used directly to build trees. Trees, where samples are directly assigned to nodes, are called *sample trees* [4]. Even though the assumption of a single subpopulation per sample may not hold, sample tree methods can be useful if the resolution of the data is not high enough to reveal the clonal composition of each sample. If the assumption is violated the tree will still represent similarities between samples, but it might differ from the true underlying phylogenetic trajectory [4]. Sample tree methods are mainly used in the context of copy number based phylogenies because the ability to deconvolute copy number profiles is still limited.

In principle, classic phylogenetic methods like neighbour joining on a Euclidean distance matrix or maximum parsimony could be used to infer sample trees. However, these methods consider each genomic location individually and do not take into account horizontal dependencies in the data, which are caused by the fact that copy number changes affect a group of neighbouring positions at a time. Consequently, a single copy number change would yield different distances depending on whether it is small or large. For this reason, distances that are calculated using individual genomic locations do not reflect the real number of copy number events and are unsuitable for inferring phylogenies from this data [148].

To account for horizontal dependencies, copy number specific approaches have been developed. One of these is TuMult [81]. It uses the shared chromosome breakpoints between samples to infer phylogenies. However, TuMult is restricted to a small number of samples and uses the total copy number, discarding information from the two parental alleles [151].

An alternative approach is MEDICC [151]. In contrast to most deconvolution-based methods, MEDICC does not rely on the infinite sites assumption, which is often violated in copy number data due to overlapping events. Instead, it takes a minimum evolution approach. Like many classic phylogenetic methods, it is a distance-based approach. However, its distance measure is based on copy number events instead of single-character changes. MEDICC takes allele-specific copy number data and infers a phylogeny that minimises the

Table 2.1 Overview of phylogenetic methods for bulk sequencing data. Methods are grouped by the task they perform: clustering, tree inference, a combination of the two or sample tree inference. Abbreviations: CG - Clonal genotype, DPMM - Dirichlet process mixture model, EM - Expectation maximisation, HMM - Hidden Markov model, (M)ILP - (Mixed) integer linear programming, QIP - Quadratic integer programming. *SubcloneSeeker can be applied to a maximum of two samples simultaneously.

Method	Infers clusters	Infers phylogeny	Single- or multi-sample	Adjusts VAFs for CNAs	Phylogenetic markers	Clustering (C) and tree inference (T) method	Year	Source
THetA	y	n	s	-	CNAs	C: matrix decomposition + custom optimisation	2013	[116]
TITAN	y	n	s	-	CNAs	C: HMM + EM	2014	[53]
CHAT	y	n	s	y	SNVs + CNAs	C: DPMM + MCMC	2014	[84]
SciClone	y	n	s/m	n	SNVs	C: Beta mixture model + variational Bayes	2014	[105]
PyClone	y	n	s/m	limited	SNVs	C: DPMM + MCMC	2014	[139]
EXPANDS	y	n	s	limited	SNVs	C: hierarchical clustering of probability distributions	2014	[6]
Clomial	y	n (CG)	m	n	SNVs	C: Matrix decomposition + EM	2014	[176]
CloneHD	y	n (CG)	m	y	SNVs + CNAs	C: HMM + local optimisation	2014	[39]
BayClone	y	n (CG)	m	n	SNVs	C: Latent feature model + MCMC	2015	[152]
RecBTP	n	y	s	-	SNVs	T: optimisation (local search)	2014	[54]
SubcloneSeeker	n	y	s/2*	-	SNVs	T: optimisation (exhaustive enumeration)	2014	[129]
SCHISM	n	y	m	-	SNVs	T: Likelihood ratio test + optimisation (heuristic)	2015	[113]
TrAp	y	y	s	n	SNVs	C+T: Matrix decomposition with sparsity constraints + optimisation (exhaustive enumeration)	2013	[159]
PhyloSub	y	y	s/m	limited	SNVs	C+T: Tree-structured stick breaking + MCMC	2014	[68]
LICHeE	optional	y	m	n	SNVs	C: heuristic multi-step clustering. T: builds evolutionary constraint network to find best scoring spanning trees	2015	[126]
CITUP	y	y	m	n	SNVs	C+T: Matrix decomposition + optimisation (QIP)	2015	[95]
AncesTree	y	y	m	n	SNVs	C+T: Matrix decomposition + optimisation (ILP)	2015	[34]
PhyloWGS	y	y	s/m	y	SNVs + CNAs	C+T: Tree-structured stick breaking + MCMC	2015	[27]
CTPsingle	y	y	s	n	SNVs	C: DPMM + MCMC. T: optimisation (MILP)	2016	[31]
SuperFreq	y	y	m	y	SNVs + CNAs	C: recursive clustering using custom heuristic, T: optimisation, assumes linear phylogeny if it cannot be ruled out.	2016	[145]
Cloe	y	y	m	n	SNVs	C+T: Latent feature model + MCMC	2016	[97]
SPRUCE	y	y	m	y	SNVs + CNAs	C: custom method, T: combinatorial optimisation	2016	[35]
Canopy	y	y	s/m	y	SNVs + CNAs	C+T: custom likelihood model + MCMC	2016	[67]
TuMult	-	y	m	-	large CNAs	assumes homogeneous sample, T: max. parsimony + optimisation	2010	[81]
MEDICC	-	y	m	-	CNAs	assumes homogeneous sample, T: min. evolution + optimisation	2014	[151]
Treeomics	-	y	m	-	SNVs	assumes mostly homogeneous sample. T: optimisation (ILP)	2016	[132]

event distance using finite state transducers. A limitation of both TuMult and MEDICC is that they do not account for noise in the copy number profiles.

2.3.7 Summary

In summary, many bulk sequencing approaches covering different data types, and sample numbers exist. However, they come with important limitations.

Deconvolution-based methods rely on clustering mutations by their cellular prevalences, but estimating cellular prevalences remains challenging due to copy number changes. Furthermore, the cellular prevalences of mutation clusters often overlap, this leads to uncertainty in the assignment of mutations to clusters and, in the worst case, can lead to clustering of independent groups of mutations. Frequency-based clustering is especially problematic for low-frequency mutations because their VAF estimates are noisier and because the chance of overlapping low-frequency clusters is higher. Additionally, cluster prevalences may not define a unique tree structure or may be phylogenetically incompatible.

Other methods aim to infer clonal genotypes directly. To overcome identifiability issues, some of these methods rely on strong evolutionary assumptions, which go far beyond the perfect and persistent phylogeny assumption and whose biological justification is unclear. Other methods reduce the solution space by limiting the maximum number of inferred subpopulations to the number of samples.

Treeomics and sample tree methods assume that mixing of subpopulations is rare or non-existent, respectively and should, therefore, be applied with caution. Furthermore, they require a relatively large number of samples.

2.4 Single-cell approaches

Recent advances in single-cell sequencing technologies have promised to reveal tumour heterogeneity at a much higher resolution [107, 169, 58]. The advantage of single-cell sequencing for phylogeny reconstruction is that it provides direct information about the co-occurrence of mutations. However, single-cell sequencing comes with its own challenges. The following section gives an introduction to single-cell sequencing and the types of errors and biases that can be introduced in the process.

2.4.1 Single-cell sequencing technologies

Single cell isolation In the first step, single cells need to be isolated. Methods to do this are micro-manipulation [179, 58, 169, 85], fluorescence activated cell sorting (FACS)

[23], laser-capture microdissection [10, 45, 170] and microfluidics [36, 168]. Some of these methods can introduce biases in the cell selection, which need to be taken into account in the experimental design [48]. If cells are not separated correctly, two cells, called *doublets*, may be isolated together. Genotypes inferred from doublets are a combination of the two cells, making tree inference more difficult. Doublet rates depend on the platform used for single-cell isolation and other experimental conditions such as cell concentration. For Drop-Seq, which uses microfluidics for cell isolation, estimates of doublet frequencies range from 0.36% to 11.3% [94]. Fluidigm assessed doublet rates for the C1 integrated fluidic circuit platform across different conditions and cell types and estimated a doublet rate of 30% (standard deviation 10%) [41]. Microscopy images can help to exclude doublets from further analysis. However, in a study by Macosko et al. [94] only a third of the doublets could be identified using this technique, as the majority of doublet cells are stacked on top of each other [41].

Single cell whole genome amplification A typical cancer cell contains between 6 and 12 picograms of DNA [109]. In comparison, high-throughput sequencing requires micrograms of DNA [141]. For this reason, the single-cell DNA needs to be amplified before it can be sequenced. DNA amplification it is a major source of noise in single-cell sequencing data. Noise caused by whole genome amplification (WGA) of single-cells includes

- false-positive mutations caused by amplification errors,
- false-negative mutations caused by allelic dropout events (ADO) in which one allele at a heterozygous locus is not amplified,
- missing genotypes caused by regions without amplification and
- uneven coverage caused by amplification biases [110].

Three major WGA techniques have been used for single-cell sequencing [59] and each of them has different advantages and disadvantages regarding technical errors.

The first technique is called *degenerate oligonucleotide-primed polymerase chain reaction* (DOP-PCR) [160] and was used as part of the first single-cell DNA sequencing method, single-nucleus sequencing [107]. DOP-PCR primers consist of random sequences at the 3' end and a fixed sequence at the 5' end. In the first amplification step, the primers bind to the DNA via their random part and are extended. In the second step, the DNA fragments are amplified by PCR using the fixed 5' sequences. In every round of PCR, the DNA fragments are doubled. A drawback of this exponential amplification is that small differences in the amplification levels of different genomic regions in the early stages get exponentially enlarged

later on, yielding over- and under-amplified regions and a single-cell genome coverage of only about 10% [109]. Due to the low coverage, DOP-PCR is unsuitable for detecting SNVs from single cells. However, it can be used to identify copy number aberrations in the range of megabases and larger [59].

The second technique is called *multiple displacement amplification* (MDA) [26]. It is the basis of two single-cell DNA sequencing methods, nuc-seq [58, 169] and single nucleus exome sequencing (SNES) [83]. MDA uses random primers and amplification is performed under isothermal conditions which enable the use of Φ29 DNA polymerase. This polymerase is characterised by a strong strand displacement activity and a high replication fidelity. Thereby, MDA can achieve a single-cell genome coverage of more than 90%[109] and yields datasets with relatively low false positive rates [59]. However, like DOP-PCR, MDA is non-linear leading to uneven coverage. In case of MDA, the amplification bias is variable from cell to cell and the read depth can therefore not be normalised based on a set of reference cells. For these reasons, MDA is used primarily for SNV detection from single-cells [58, 169, 85].

The third technique, *multiple annealing and looping-based amplification cycles* (MALBAC) [179], is an amplification technique specifically developed for single-cell DNA sequencing. Like DOP-PCR, it uses primers with random 3' and fixed 5' sequences for pre-amplification. The primers are designed such that the 3' and 5' ends of the full amplicons are complementary and can hybridise to each other. This prevents the amplicons from being amplified further and ensures that the first stage of the amplification is quasilinear, leading to a lower amplification bias. The pre-amplification step is followed by PCR. Advantages of MALBAC compared to MDA are a more uniform coverage as well as a lower ADO rate. The false positive rate, however, is higher due to the use of a different DNA polymerase [59]. Even though MALBAC is not free from amplification bias, the bias is similar between replicates of cells, which makes normalisation of the sequencing depth feasible. This makes MALBAC the preferred technique for the analysis of single-cell copy-number aberrations.

Huang et al. [59] performed a comprehensive comparison of the different amplification methods using commercially available WGA kits on diploid human cell lines. They report ADO rates of 0.78 for DOP-PCR, 0.33 to 0.38 for MDA and 0.21 to 0.28 for MALBAC. The corresponding false positive rates were 9.6×10^{-4} for DOP-PCR, 8.2×10^{-5} to 1.3×10^{-4} for MDA and 2.4×10^{-4} to 3.8×10^{-4} for MALBAC. Others have reported ADO rates between 0.16 to 0.43 and false positive rates between 2.67×10^{-5} to 6.7×10^{-5} [58, 169, 85] for MDA. In the genotype matrices produced by Li et al. [85] and Hou et al. [58] 55.2% and 57.7% of values are missing, respectively.

In addition to the technical errors mentioned above, amplification can produce chimeric DNA fragments, which are amplification artefacts that contain sequences from different parts of the genome [48]. For this reason, inference of structural variants like small insertions and deletions from single cells remains challenging.

Whole genome, exome and targeted sequencing Once amplified, the single-cell DNA can be subjected to standard sequencing protocols used for bulk DNA including whole genome, exome or targeted sequencing. Whole genome sequencing offers to yield most information about the genome. However, it is also the most expensive. In the context of single-cell sequencing it is mostly used for the analysis of copy number alterations and shallow sequencing is used to reduce the cost [107].

Alternatively, one can restrict sequencing to selected regions of the genome using exome or targeted sequencing. Targeted sequencing can, for example, be based on a particular set of genes. Leung et al. [82] used 200 known cancer-associated genes in their protocol for targeted DNA sequencing from single nuclei. The main advantage of targeted sequencing is that it drastically reduces the cost per cell, which means that more cells can be analysed. However, this needs to be weighed up against the size of the sequenced regions, which limits the number of mutations that can be detected [48].

Data preprocessing Even though false positive rates (FPR) in the order of 10^{-4} might seem small at first sight, the fact that the genome is extremely large means that false positives can easily outnumber true somatic variants [89]. For the human exome alone, which contains approximately 30 megabases [111], an FPR of 10^{-4} would yield around 3000 false positives.

The number of false positives is usually reduced by census-based variant calling, which only selects variants that are observed in multiple cells. Hou et al. [58] and Xu et al. [169] for example select the threshold t, the minimum number of cells in which a mutation needs to occur in order to be called, such that the expected number of false positive mutation sites is smaller than 1, i.e. they choose the smallest t that satisfies

$$\sum_{i=t}^{n} \binom{n}{i} \mathrm{FPR}^i (1-\mathrm{FPR})^{n-i} \cdot S_{\mathrm{exome}} < 1, \tag{2.5}$$

where n is the number of single cells, and S_{exome} is the size of the exome. However, consensus-based variant filtering cannot remove sites of recurrent sequencing errors [108] or errors caused by doublets. Furthermore, it implies that private mutations of single-cells cannot be identified.

2.4.2 Inferring trees from single-cell sequencing data

Various methods have been used to infer trees from single-cell datasets of SNVs. These include classic phylogenetic methods as well as new heuristic and probabilistic approaches. A summary of probabilistic single-cell phylogeny methods is shown in Table 2.2.

- Yu et al. [173] used *UPGMA* (Unweighted Pair Group Method with Arithmetic Mean), whereas Xu et al. [169] applied the *neighbour joining* algorithm. Both are agglomerative hierarchical clustering methods. Hughes et al. [60] used neighbour joining trees as input for a likelihood optimisation method, which is based on a general time-reversible substitution model. Another classic phylogenetic approach is Bayesian phylogenetic inference as used by Eirew et al. [33].

- Potter et al. [127] defined subpopulations by grouping cells with identical genotypes into clones and then applied a *maximum parsimony* approach. Their data sets were derived by single-cell qPCR of few genetic markers. Their definition of subpopulations is not suitable for noisy single-cell data with hundreds of genetic markers, where the observed genotypes of any two cells differ. Furthermore, maximum parsimony does not account for noise.

- Some methods first attempt to cluster cells into subpopulations and then infer minimum spanning trees. Gawad et al. [47] do this using model-based clustering, whereas Yuan et al. [174] use k-means and hierarchical clustering.

None of the methods mentioned so far accounts for the noise of single-cell datasets. However, probabilistic models have started to emerge.

- Yuan et al. [174] developed a probabilistic model called *BitPhylogeny*. It uses a tree-structured mixture model in combination with MCMC to cluster cells and to infer their phylogenetic relationships. While mixture models are widely used and valuable, they require large data sets in order to converge to an accurate representation of the underlying distributions. Current single-cell data sets in contrast are small, containing usually fewer than 100 cells [107, 169, 58, 85, 173, 167, 60, 89].

- Roth et al. [140] use a hierarchical Bayesian mixture model to infer clones and genotypes. In contrast to Bitphylogeny, their method *Single Cell Genotyper* accounts for doublet errors. Furthermore, they use variational Bayes for the inference. However, their approach has only been tested on targeted sequencing data, which has different error profiles than the more commonly used WGA-based sequencing [110].

Table 2.2 Overview of probabilistic phylogenetic methods for single-cell sequencing data. All of these methods use point-mutations as phylogenetic markers. *ddClone requires both single-cell and bulk sequencing data.

Method	Model	Inference	Phylogenetic representation	Year	Source
BitPhylogeny	Tree-structured stick breaking	MCMC	clone tree	2015	[174]
OncoNEM	custom likelihood model	custom search algorithm	clone tree	2016	Chapter 3
SCITE	custom likelihood model	MCMC	mutation tree	2016	[65]
Single Cell Genotyper	hierarchical Bayesian model	variational Bayes	clonal genotypes	2016	[140]
ddClone*	Dirichlet process mixture model	MCMC	mutation clusters	2017	[144]

- Recently, Salehi et al. [144] developed *ddClone*. This method uses a combination of single-cell and bulk sequencing data. It infers mutation clusters from bulk data using a Dirichlet process mixture model, where the single-cell sequencing data serves as prior for the mutation clusters and their prevalences.

- Kim and Simon [75] proposed a heuristic method for inferring *mutation trees*. These are trees in which each node corresponds to a mutation instead of a clone. Recently, Jahn et al. [65] developed a probabilistic version of this approach called *SCITE*. Since the model behind SCITE is related to the method presented in the next chapter, it will be discussed in more detail at the end of Chapter 3.

As with bulk-sequencing, attempts to infer phylogenies from copy number data are less common.

- Navin et al. [107] and Wang et al. [167] used neighbour joining on Euclidean distances to infer trees from single-cell copy-number profiles obtained by whole-genome sequencing. As explained in the previous section, the Euclidean distance is an unsuitable distance measure for copy number profiles.

- Chowdhury et al. [19, 20] used *Steiner trees* to infer phylogenies from single-cell copy number profiles obtained from fluorescent in situ hybridisation. Their algorithms, however, only infer trees from low-dimensional genotype spaces.

- Garvin et al. [46] developed *Ginkgo*, which clusters samples using the Pearson correlation as dissimilarity measure.

All in all, phylogeny inference from single cells is less advanced than inference from bulk data and suitable methods to infer single-cell phylogenies are still lacking. Current methods do not account for the large proportion of noise in the data or they require data sets that are much larger than the ones currently available.

Chapter 3

Inference of single-cell phylogenies

The last chapter described current approaches to inferring tumour phylogenies from single-cell sequencing data and their limitations. This chapter presents a new approach, OncoNEM, to overcome these limitations (Section 3.1). The main contribution is a probabilistic model, which accounts for the noise in single-cell data, and a heuristic search algorithm to explore the search space. I assess the robustness of OncoNEM in simulations (Section 3.2) and compare its performance with that of competing methods (Section 3.3), which were chosen to be a representative selection of the approaches described in Chapter 2. Section 3.4 presents the results of applying OncoNEM in two case studies: a data set containing 44 single tumour cells from a muscle-invasive bladder transitional cell carcinoma and a data set containing 58 single tumour cells from an essential thrombocythemia. Finally, I describe an extension of OncoNEMs to infer mutation trees (Section 3.5) and discuss OncoNEM's limitations (Section 3.6). This chapter contains text and figures from Ross and Markowetz [136].

3.1 A probabilistic model for single-cell phylogenies

3.1.1 Modelling noisy subset relationships

As described in Chapter 1 and 2, tumour evolution produces cell populations containing different sets of mutations and, under the infinite sites assumption, the subset relationships of the mutations describe the phylogenetic relationships between the different populations.

A probabilistic method that uses noisy subset relationships to infer graphs are Nested Effects Models (NEMs) [99, 100]. More specifically, NEMs have been developed to infer signalling hierarchies from the noisy measurements of gene expression changes from gene perturbation screens.

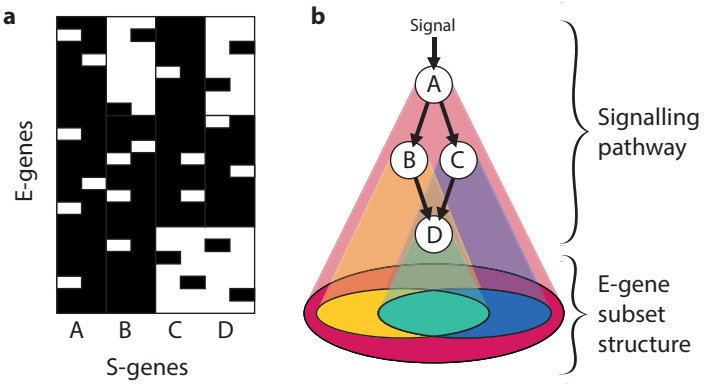


Fig. 3.1 Nested Effects Models in a nutshell. a – The input to NEMs is a noisy data matrix from a gene perturbation screen. b – The signalling pathway is inferred based on the subset structure of perturbation effects.

The input to NEMs is a matrix of a set of signalling genes (S-genes) versus a set of effect genes (E-genes). S-genes are genes that are thought to be related through an unknown signalling pathway. E-genes are genes whose expression levels change if one of the signalling genes is perturbed, for example using RNAi. The matrix describes for each pair of S-genes and E-genes if perturbing S affected the gene expression of E (Figure 3.1a). The main assumption behind NEMs is that the genes that show an effect upon perturbation of a particular S-gene S_i will be a subset of the genes that show an effect upon perturbation of a signalling gene that is upstream of S_i (Figure 3.1b). NEMs offer a probabilistic scoring function which uses this assumption to evaluate the fit between a given pathway model and the observed data.

3.1.2 Inferring signalling pathways versus tumour phylogenies using NEMs

Similar to NEMs, we aim to infer a graph from noisy subset relationships. For this reason, we used NEMs as the basis for developing OncoNEM (Oncogenetic Nested Effects Model). OncoNEM is an automated method for reconstructing clonal lineage trees from somatic single nucleotide variants (SSNVs) of multiple single tumour cells that exploits the nested structure of mutation patterns of related cells. In the following sections, we will describe OncoNEM in detail. First, however, we discuss some of the key differences between the inference of signalling pathways from gene perturbation screens and the inference of tumour phylogenies from single-cell single nucleotide variants.

Inversion of subset relationships E-genes affected by downstream signalling genes are subsets of the E-genes affected by upstream signalling genes. Those up- and downstream subset relationships are inverted in the setting of phylogenies. As mutations accumulate over time, populations that occurred early during tumour development will carry a subset of the mutations that are present in their descendants.

Targeted perturbations versus random mutations In the setting of NEMs, it is known which node in the network is perturbed due to the targeted nature of the intervention. However, the effect of the perturbation on the pathway components cannot be observed directly. Instead, we observe effects on downstream reporters, but the connection between the pathway components and the reporters is unknown *a priori*.

In case of OncoNEMs, the perturbations are spontaneously occurring mutations, which means that the origin of the perturbation signal is unknown *a priori*. However, the effect of a perturbation on the graph components can be observed directly in the form of the single-cell genotypes.

Graph size and structure For OncoNEMs it is not only unknown in which cell a mutation originated, but it is also possible that a mutation originated from an ancestral tumour cell population that is not even represented in our data set. For this reason, the size of the oncoNEM graph is unknown *a priori*. The size of the inferred signalling graph, in contrast, is defined by the number of perturbed genes. In addition to the graph sizes, the structures also differ: Signalling pathways can have complex structures. To reflect this, NEMs infer directed acyclic graphs. Tumour evolution is thought of as a branching process. Therefore, OncoNEMs infer directed, arbitrary trees.

Data NEMs have been developed to handle complete data, OncoNEMs, in turn, need to handle large amounts of missing values to be useful in practice.

3.1.3 Likelihood model of OncoNEMs

To model the accumulation of mutations, we assume that each locus gets mutated only once (infinite sites assumption [76]) and that mutations are never lost. Under these assumptions, direct relationships between clones imply that the mutations of the ancestral clone are a subset of the descendants' mutations. To define the likelihood of a tree given the observed genotypes, OncoNEM predicts the expected mutation patterns based on the tree and then scores the fit between predicted and observed mutations patterns while probabilistically accounting for genotyping errors. A schematic illustration of the OncoNEM scoring model

is shown in Figure 3.2. The derivation of the scoring function is described in the following. It is adapted from the derivation of NEMs [99] and takes the differences mentioned above into account.

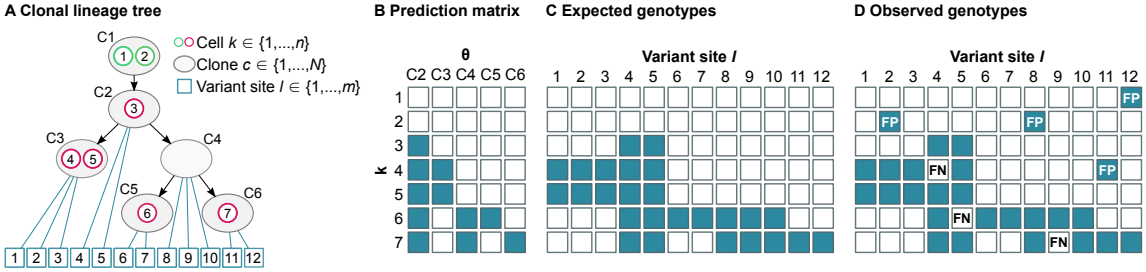


Fig. 3.2 Toy example of the OncoNEM scoring model. A – Hypothesis of a clonal lineage tree that describes the subpopulations of a tumour (grey circles) and their relationships (black arrows). B – This tree can be represented as a prediction matrix that predicts the mutation pattern we expect to see across all k cells for a mutation that occurred in a certain clone θ. C – Assuming that we know the originating clone of every mutation (blue lines in clonal lineage tree), we can extend the prediction matrix to a full matrix of expected genotypes. D – To score the tree, expected genotypes are compared to observed genotypes. The more mismatches (FP: false positive, FN: false negative) there are, the lower the likelihood of the tree given the data. Since the origin of a mutation is unknown *a priori*, the full likelihood of the lineage tree is calculated by marginalising over all possible origins for every mutation.

Data We assume that the variants of the single cells have already been called and filtered so that the data set only contains the somatic variant sites. Let $D = (d_{kl})$ be the matrix of observed genotypes where $k \in \{1,\ldots,n\}$ is the label of a single cell and $l \in \{1,\ldots,m\}$ is the index of a mutation site. Let $d_{kl} \in \{0,1,\texttt{NA}\}$ denote the mutation status of cell k at site l, where 0, 1 and `NA` encode an unmutated, mutated or unknown site, respectively.

Clonal lineage trees We assume that a clonal lineage tree is a directed, not necessarily binary tree $\mathscr{T}$ whose root is the unmutated normal. Every node of this tree represents a clone $c \in \{1,\ldots,N\}$ that contains zero, one or multiple cells of the data set. Let $c(k)$ denote the clone that contains cell k. In the following, we assume without loss of generality that the root has index 1.

OncoNEM An OncoNEM has two parts: the clonal lineage tree $\mathscr{T}$ and the occurrence parameter $\Theta = \{\theta_l\}_{l=1}^{m}$, where θ_l takes the value c of the clone where mutation l originated.

The core of our method is a function that defines the probability of the OncoNEM given a data set D and is derived in the following. Using a Bayesian approach, the posterior

probability of $\mathscr{T}$ and Θ given D can be written as

$$P(\mathscr{T},\Theta|D) = \frac{P(D|\mathscr{T},\Theta)\,P(\Theta|\mathscr{T})\,P(\mathscr{T})}{P(D)}. \tag{3.1}$$

The model prior $P(\mathscr{T})$ can be used to incorporate prior biological knowledge. We assume it to be uniform over the search space. The normalising factor $P(D)$ is the same for all models and not necessary to compute when comparing them. Therefore,

$$P(\mathscr{T},\Theta|D) \propto P(D|\mathscr{T},\Theta)\,P(\Theta|\mathscr{T}). \tag{3.2}$$

Likelihood for known Θ Let us assume that we know for each locus l in which clone the mutation occurred and that no mutations occur in the normal. This is equivalent to restricting the parameter space of θ_l to $\{2,\ldots,N\}$ and is justified by stringent variant filtering of the input data.

Given $\mathscr{T}$ and Θ, we can predict the genotype of every cell: if c is the clone in which a mutation occurred, the mutation is present in c and all descendants of c and absent in all other clones, i.e. given $\theta_l = c$, the tree determines the predicted genotype δ_{kl}.

Finally, to calculate the likelihood of $(\mathscr{T},\Theta)$, we compare the expected genotypes with the observed ones. We model the genotyping procedure as draws of binary random variables ω_{kl} from the sample space $\Omega = \{0,1\}$ and assume that, given $\mathscr{T}$ and Θ, the random variables are independent and identically distributed according to the probability distribution

$$P(\omega_{kl}|\delta_{kl}) = \begin{pmatrix} P(0|0) & P(1|0) \\ P(0|1) & P(1|1) \end{pmatrix} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}, \tag{3.3}$$

where α and β are global probabilities of false positive and false negative draws, respectively.

We interpret the observed genotypes d_{kl} as events from the event space $\mathscr{P}(\Omega) = \{\emptyset,\{0\},\{1\},\{0,1\}\}$, where a missing value corresponds to the event $\{0,1\}$. Then, the probability of the observed genotypes D given $\mathscr{T}$ and Θ is

$$P(D|\mathscr{T},\Theta) = \prod_{l=1}^{m} P(D_l|\mathscr{T},\theta_l) = \prod_{l=1}^{m}\prod_{k=1}^{n} P(\omega_{kl} \in d_{kl}|\delta_{kl}), \tag{3.4}$$

where

$$P(\omega_{kl} \in d_{kl} | \delta_{kl}) = \begin{cases} 1-\alpha & \text{if } d_{kl} = \{0\} \text{ and } \delta_{kl} = 0 \\ \alpha & \text{if } d_{kl} = \{1\} \text{ and } \delta_{kl} = 0 \\ \beta & \text{if } d_{kl} = \{0\} \text{ and } \delta_{kl} = 1 \\ 1-\beta & \text{if } d_{kl} = \{1\} \text{ and } \delta_{kl} = 1 \\ 1 & \text{if } d_{kl} = \{0,1\} \end{cases} \tag{3.5}$$

is the probability of a single observation given the predicted genotype.

Likelihood for unknown Θ So far we assumed Θ to be known, but this is generally not the case. To derive the likelihood of the entire data matrix, we treat Θ as a nuisance parameter and marginalise over it. Furthermore, we make two assumptions: First, the occurrence of one mutation is independent of the occurrence of all other mutations, i.e.

$$P(\Theta|\mathscr{T}) = \prod_{l=1}^{m} P(\theta_l|\mathscr{T}), \tag{3.6}$$

and second, the prior probability of a mutation to occur in a clone is

$$P(\theta_l = c|\mathscr{T}) = \begin{cases} 0 & \text{if } c \text{ is the normal } (c=1), \\ \frac{1}{N-1} & \text{otherwise.} \end{cases} \tag{3.7}$$

Then the marginal likelihood is

$$\begin{aligned} P(D|\mathscr{T}) &= \int P(D|\mathscr{T},\Theta) P(\Theta|\mathscr{T}) \mathrm{d}\Theta \\ &= \prod_{l=1}^{m} \int P(D_l|\mathscr{T},\theta_l) P(\theta_l|\mathscr{T}) \mathrm{d}\theta_l \\ &= \frac{1}{(N-1)^m} \prod_{l=1}^{m} \sum_{c=2}^{N} \prod_{k=1}^{n} P(\omega_{kl} \in d_{kl}|\mathscr{T}, \theta_l = c) \\ &= \frac{1}{(N-1)^m} \prod_{l=1}^{m} \sum_{c=2}^{N} \prod_{k=1}^{n} P(\omega_{kl} \in d_{kl}|\delta_{kl}) . \end{aligned} \tag{3.8}$$

3.1.4 Algorithms for OncoNEMs inference

The search space of cell lineage trees with n nodes contains n^{n-2} models [155], making exhaustive enumeration infeasible for trees with more than nine nodes. This is further complicated by the fact that the size of the tree is not known *a priori* due to unobserved subpopulations.

To address these challenges and to identify high-scoring models in the space of possible tree structures, we developed a sequence of three inference algorithms.

We start with an initial search, where we restrict the model space to cell lineage trees, i.e. trees where each node corresponds to a single cell. This yields a first estimate of the tree. The second step tests whether adding unobserved clones to the tree substantially increases the likelihood. This is to identify ancestral subpopulations that are not represented in the single cell sample. The third step yields the final model of the clonal lineage tree by clustering cells within the previously derived tree into clones. An overview of the inference steps is shown in Figure 3.3 and details of the algorithms are described in the following sections.

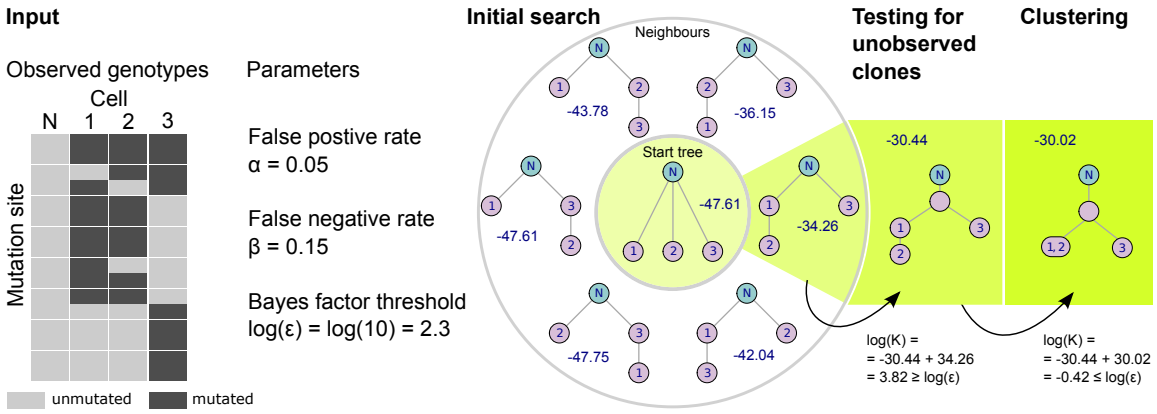


Fig. 3.3 Toy example of OncoNEM inference steps. Given the observed genotypes and the input parameters α and β, the log-likelihood of the start tree, which is by default a star-shaped tree, is -47.61. In the first step of the initial search, all neighbours of the star tree are scored. The highest scoring tree obtained in this step has a log-likelihood of -34.26. In this toy example, the highest scoring tree of the first step is also the best cell lineage tree, overall. Therefore, the initial search terminates with this tree as solution. In the first refinement step, we find that inserting an unobserved node into the branch point of our current tree increases the log-likelihood by 3.82. Since this improvement is larger than the Bayes factor threshold of 2.3, the solution with the unobserved clone is accepted. In the final refinement step, cells are clustered along edges. In the toy example, only one clustering step does not decrease the log-likelihood by more than $\log(\varepsilon)$.

Step 1) Initial search: Building a cell tree

The initial search algorithm performs a local heuristic search and is based on two lists containing trees:

a) The list of scored trees. This list contains all trees which have been scored already.

b) The priority queue. This list contains all trees which have been scored at some point during the search but whose neighbours have not yet been scored explicitly.

Both lists are ordered by the likelihood of the trees as defined in Section 3.1.3, with the tree with the highest likelihood coming first.

By default, the search starts with a tree that has a star topology, i.e. a tree where all cells are attached to the normal root. We score the start tree and add it to both the priority queue and the list of scored trees. We pick the highest scoring tree from the priority queue list and generate all of its neighbours. Each neighbour is scored and added to both lists, unless it is already on the list of scored trees, in which case we do nothing. This step is repeated until the highest scoring tree has not changed for more than δ generations, where δ is a parameter set by the user. For a detailed description of the algorithm see Algorithm 1.

Unlike hill climbing [86], which only considers the neighbourhood of its current best solution and can therefore easily get stuck in local optima, our algorithm uses the highest scoring tree in the priority list as a starting point for the next search step. This tree does not have to be a neighbour of the current best solution and also does not have to have a higher score than the current best solution. To prevent the algorithm from reanalysing solutions, each tree can only be scored and added to the priority queue once.

Step 2) Refinement: Testing for unobserved clones

The number of sequenced single cells is usually small compared to the tumour size. Consequently, some clones of the tumour may not be represented in the single cell sample. This problem is similar to the 'unknown unknowns' problem in reconstructing biological pathways [142], where latent variables which cause additional patterns in the observed data set can be inferred. In the OncoNEM setting, unobserved clones with at least two child clones create additional mutation patterns and can therefore potentially be inferred (Figure 3.4). OncoNEM accounts for this possibility by testing if there is a lineage tree with additional, unobserved branch nodes that can better explain the observed data (see Algorithm 2). Unobserved clones that connect observed clones linearly cannot be inferred, but also do not change the shape of the tree.

To test for unobserved clones, the algorithm identifies all branch points of the tree inferred in step 1. For each branch point, it then generates a new tree by inserting an unobserved

Algorithm 1: Heuristic search algorithm. D is the genotype Matrix, α and β are the error rates and startTrees is the list of trees the heuristic search is started from. The algorithm terminates if the highest scoring solution has not changed for more than δ iterations.

```
1  Function heuristicSearch(D, α, β, startTrees, δ)
2      initialize scoredTrees ← empty;
3      initialize priorityQueue ← empty;
4      initialize counter ← 0;
5      for tree in startTrees do
6          score tree;
7          add tree to scoredTrees;
8          add tree to priorityQueue;
9      end
10     bestTree ← scoredTrees[1];
11     while counter ≤ δ do
12         currentTree ← priorityQueue[1];
13         delete currentTree from priorityQueue;
14         for every neighbour of currentTree do
15             if neighbour ∉ scoredTrees then
16                 score neighbour;
17                 add neighbour to scoredTrees;
18                 add neighbour to priorityQueue;
19             end
20         end
21         if bestTree ≠ scoredTrees[1] then
               /* Highest scoring solution changed                    */
22             counter ← 0;
23             bestTree ← scoredTrees[1];
24         else
25             counter ← counter +1;
26         end
27     end
28     return scoredTrees[1]
29 end
```

node into this point. The collection of these trees are used as start trees in a new search that optimises the position of the unobserved node in the tree. A larger model is accepted if the Bayes factor of the larger versus the smaller model is greater than a threshold ε (see below). If the larger model passes the threshold, these expansion steps are repeated. Otherwise, the algorithm terminates with the smaller solution.

Algorithm 2: Expansion algorithm – tests for unobserved clones. $\mathscr{T}_n$ represents the cell lineage tree with n nodes inferred by the initial `heuristicSearch()`, ε is the Bayes factor threshold. As before, D is the genotype Matrix, α and β are the error rates and δ defines the termination criterion for the heuristic search.

```
Function expandTree(D, α, β, T_n, δ, ε)
    initialize i ← 0;
    repeat
        i ← i + 1;
        /* Generate start trees                                    */
        startTrees ← star tree with n + i nodes;
        branchPoints ← vector of nodes in T_{n+i-1} that have at least two children;
        for j in branchPoints do
            Generate a new tree by inserting an unobserved node into j;
            Add tree to startTrees;
        end
        scoredTrees ← heuristicSearch(D, α, β, startTrees, δ);
        T_{n+i} ← highest scoring tree in scoredTrees in which every unobserved node
          has at least two children;
        /* Calculate Bayes factor                                  */
        K ← P(D|T_{n+i}) / P(D|T_{n+i-1})
    until K < ε;
    return T_{n+i-1}
end
```

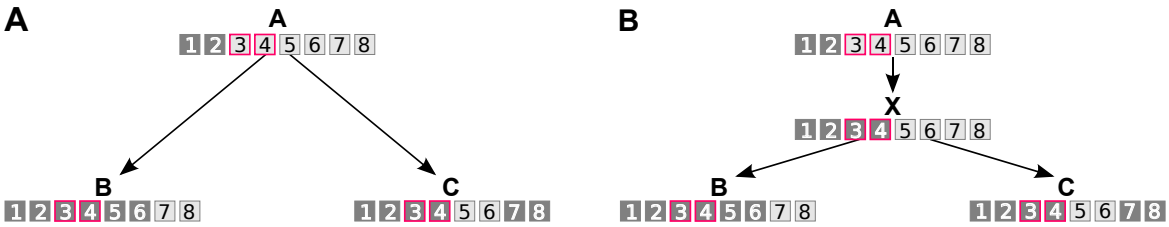


Fig. 3.4 Unobserved clones lead to differences between observed and predicted mutation patterns. To explain the observed mutation patterns the model shown in (A) treats positions 3 and 4 as false negatives in clone A or false positives in clone B and C. Alternatively, we can explain the data without observational errors by adding an intermediate unobserved clone X as shown in (B).

Step 3) Refinement: Clustering cells into clones

The clustering procedure tests if the data can be explained better or equally well by a clonal lineage tree in which multiple cells correspond to the same node (see Algorithm 3). Nodes are clustered iteratively along branches until merging cells into clones decreases the likelihood by more than a factor of $1/\varepsilon$ compared to the best clustering solution found so far. Cells may

be clustered into clones because they are genetically very similar or because of the limited information content of the data, which can be due to genotyping errors, missing values or a restricted number of SSNVs in the sequenced regions of the genome.

Algorithm 3: Clustering algorithm. $\mathscr{T}_{start}$ represents the cell lineage tree inferred by `expandTree()`. As before, D is the genotype Matrix, α and β are the error rates and ε is the Bayes factor threshold.

Function `clusterTree`(D, α, β, $\mathscr{T}_{start}$, ε)
  initialize $\mathscr{T} \leftarrow \mathscr{T}_{\text{start}}$; `/* Current tree */`
  initialize $\mathscr{T}^* \leftarrow \mathscr{T}_{\text{start}}$; `/* Best tree scored so far */`
  repeat
    for *every edge* e_i **do**
      Generate clustered tree $\mathscr{T}_{e_i}$ from $\mathscr{T}$ by merging the clones connected by e_i;
    end
    $\mathscr{T}_{e_i^*} \leftarrow \arg\max_{\mathscr{T}_{e_i}} P(D|\mathscr{T}_{e_i})$;
    $K \leftarrow P(D|\mathscr{T}^*)/P(D|\mathscr{T}_{e_i^*})$;
    if $K \leq \varepsilon$ **then**
      `/* Accept clustering solution */`
      $\mathscr{T} \leftarrow \mathscr{T}_{e_i^*}$;
      if $P(D|\mathscr{T}^*) < P(D|\mathscr{T}_{e_i^*})$ **then**
        `/* Save clustering solution as new best tree */`
        $\mathscr{T}^* \leftarrow \mathscr{T}_{e_i^*}$;
      end
    end
  until $K > \varepsilon$;
  return $\mathscr{T}$
end

Choosing the Bayes factor threshold ε

Choosing the parameter ε is a trade-off between inferring distinct clones with little support from the data and overly strict clustering. In this setting, choosing $\varepsilon > 1$ means that we prefer the smaller model unless the strength of evidence for the larger model compared to the smaller one exceeds a certain threshold. Jeffreys' [66] or Kass and Raftery's [73] scale for the interpretation of the Bayes factor can be used as guidance. We used a value of $\varepsilon = 10$, which denotes strong evidence according to Jeffreys' scale.

Estimating the occurrence parameter Θ

Given a lineage tree, we can estimate which clones acquired which mutations during tumour development. To do this, we calculate the posterior probability of a mutation having occurred in clone c. Using a uniform prior for the occurrence parameter $\theta_l \in \{2,\ldots,N\}$, we obtain

$$P(\theta_l = c \mid \mathscr{T}, D) = \frac{1}{Z}\prod_{k=1}^{n} P(\omega_{kl} \in d_{kl} \mid \mathscr{T}, \theta_l = c), \tag{3.9}$$

with normalising constant

$$Z = \sum_{c=2}^{N}\prod_{k=1}^{n} P(\omega_{kl} \in d_{kl} | \mathscr{T}, \theta_l = c). \tag{3.10}$$

The branch lengths L of the tree can be estimated as the expected number of mutations that separate a clone c from its parent $\mathrm{pa}(c)$,

$$L_{\mathrm{pa}(c),c} = \sum_{l=1}^{m} P(\theta_l = c | \mathscr{T}, D). \tag{3.11}$$

Estimating model parameters α and β

Previous studies have estimated FPRs and ADO rates from the sequencing data [169, 58]. These error rates are, however, not equivalent to the false positive and false negative parameters α and β used by OncoNEM. This is due to two reasons (*i*) the pre-processing steps that are applied to the sequencing data to generate the final genotype matrix and (*ii*) doublets.

Effect of pre-processing on error parameters The FPR estimated in sequencing studies is usually the number of false variants per sequenced base pair

$$\mathrm{FPR} = \frac{\sum_k \sum_{l\in L} \mathbb{1}_{d_{kl}=1 \wedge \delta_{kl}=0}}{n|L|}, \tag{3.12}$$

where d_{kl} and δ_{kl} are the observed and true genotype at site l and in cell k, respectively, and L is the set of all sequenced sites. Let $L_i \subset L$ be the set of sites for which mutations have been observed in i cells. As $L_0, L_1, \ldots, L_n$ form a partition of L, the FPR can be expressed as

$$\begin{aligned}\mathrm{FPR} &= \frac{\sum_k \left(\sum_{l\in L_0} \mathbb{1}_{d_{kl}=1\wedge\delta_{kl}=0} + \sum_{l\in L_1} \mathbb{1}_{d_{kl}=1\wedge\delta_{kl}=0} + \ldots + \sum_{l\in L_k} \mathbb{1}_{d_{kl}=1\wedge\delta_{kl}=0}\right)}{n\left(|L_0| + |L_1| + \ldots + |L_n|\right)} \\ &= \frac{\sum_k \left(\sum_{l\in L_1} \mathbb{1}_{d_{kl}=1\wedge\delta_{kl}=0} + \ldots + \sum_{l\in L_k} \mathbb{1}_{d_{kl}=1\wedge\delta_{kl}=0}\right)}{n\left(|L_0| + |L_1| + \ldots + |L_n|\right)}.\end{aligned} \tag{3.13}$$

To obtain the genotype matrix which is used as an input to OncoNEM, uninformative sites that show no mutations across all cells are removed, i.e. only sites $l \in L_1 \cup \ldots \cup L_k$ are selected. In this case the false positive parameter α is

$$\alpha = \frac{\sum_k \left(\sum_{l \in L_1} \mathbb{1}_{d_{kl}=1 \wedge \delta_{kl}=0} + \ldots + \sum_{l \in L_k} \mathbb{1}_{d_{kl}=1 \wedge \delta_{kl}=0} \right)}{n \left(|L_1| + \ldots + |L_k| \right)}. \tag{3.14}$$

Given that most of the sites sequenced will not show a mutation, $|L_0|$ will be large, which means that the estimated FPR will be much lower than α. If a consensus-based variant filtering threshold t is applied, only mutations that are observed in at least t cells are selected and α is changed further to

$$\alpha = \frac{\sum_k \left(\sum_{l \in L_t} \mathbb{1}_{d_{kl}=1 \wedge \delta_{kl}=0} + \ldots + \sum_{l \in L_k} \mathbb{1}_{d_{kl}=1 \wedge \delta_{kl}=0} \right)}{n \left(|L_t| + \ldots + |L_k| \right)}. \tag{3.15}$$

While α will be lower for a consensus filtered data set than α in equation 3.14, the exact effect of this step is difficult to estimate.

The impact of filtering is not limited to α but affects the false negative parameter β, as filtering removes sites that have a high ADO rate preferentially. Furthermore, the data matrix is binarised as part of the pre-processing. In this step, all homozygous mutant sites are interpreted as heterozygous normal/mutant. This step reduces the false negative rate (FNR) by approximately half and further explains why β is expected to differ from the genome-wide estimate of the ADO rate.

Effect of doublets on error parameters Sequencing doublets can create mutation patterns that are incompatible with a tree structure. These mutation patterns further increase the number of false positives and false negatives that are needed to explain the data. At the same time, these mutation patterns are not caused by sequencing errors in the strict sense, as they are present in the original DNA.

While all of the points discussed above are expected to change the error rates of the final dataset, the exact impact on the parameters is difficult to estimate. Therefore, we chose to estimate error rates for our model directly from the data.

We treat the selection of model parameters as part of the learning problem and estimate them using a maximum likelihood approach, similar to Zeller et al. [177]. We create a grid of parameter combinations α and β and optimise $\mathscr{T}$ given these parameters using the heuristic search algorithm. Then, we choose the parameter combination that yields the highest scoring tree and infer a clonal lineage tree as described above.

This parameter estimation process is computationally expensive compared to the tree inference. However, it can easily be parallelised, and the grid of parameter combinations can be coarse as OncoNEM is robust to changes in the model parameters around the optimum (see simulation results). Furthermore, the range of tested parameter combinations can be reduced in the presence of prior knowledge.

3.1.5 Implementation

OncoNEM is freely available as an R package on Bitbucket [134] under a GPL3 license. The user functions and the tree refinement algorithms have been implemented in R [130] whereas the computationally expensive search algorithm has been implemented in C++ and integrated into R using the package Rcpp [32].

3.2 Validation through simulations

We performed comprehensive simulations to assess the robustness of OncoNEM to errors in the parameter estimates and compared its performance to six baseline methods. For these simulation studies, datasets were created in a two-step procedure that consists of (*i*) generating a tree structure and (*ii*) simulating the corresponding genotypes.

Simulating clonal lineage trees. To simulate a tree with c clones, we select clone one to be the root and the parent of the second clone. Then, the remaining clones are added iteratively by choosing a non-root node that is already part of the tree with uniform probability as parent.

In the case of simulating trees with unobserved clones, we count how many nodes in the simulated tree have at least two children. If this number is greater than or equal to the desired number of unobserved clones c_u, we randomly choose c_u of these nodes as unobserved clones. Otherwise, a new tree is simulated. Next, we assign one cell to every observed clone. For the remaining cells, clones are chosen iteratively with a probability proportional to the current clone size, to generate clones of different sizes.

Simulating genotype observations. For every mutation site, we choose the occurrence parameter θ_l with uniform probability from all non-root nodes. Given Θ and the tree structure, the full matrix of true genotypes is obtained by setting an entry to 1 if the mutation occurred in a clone that is ancestral to the cell's clone or if the mutation occurred in the clone containing the cell itself, and 0 otherwise.

Observed genotypes are derived from true genotypes by (*i*) setting a fraction p_{missing} of randomly chosen values to `NA`, (*ii*) setting a fraction α of unmutated, non-missing entries to 1 and (*iii*) setting a fraction β of mutated, non-missing entries to 0. If this yields sites without any observed mutations, we add, for each of these sites, a false positive to a randomly chosen cell. Finally, we randomise the order of cells in the matrix of observed genotypes, to avoid bias in the method testing.

OncoNEM is robust to changes in error parameters α and β To test if our method can infer the main model parameters, α and β, and to evaluate the robustness of our method to errors in those estimates, we simulated a tree containing 10 clones, 2 of which were unobserved, with a total number of 20 cells. A corresponding genotype matrix with 200 SNVs was simulated using an α of 0.2, a β of 0.1 and 20% of missing values. Then, we inferred clonal lineage trees as described above, using various combinations of false positive and false negative rates, and compared the inferred trees to the ground truth. As Figure 3.5A shows, a large range of parameter combinations yield solutions that are close to the original tree in terms of pairwise cell shortest-path distance and V-measure with both the inferred and the ground truth parameters lying in the middle of this range. Similar results were obtained on a second data set that was simulated using a much lower α of 10^{-5} (see Figure A.1). These results demonstrate that OncoNEM is robust to changes in the model parameters.

OncoNEM estimates model parameters accurately In the second simulation study, we further assessed the parameter estimation accuracy of OncoNEM. To generate different test data sets, we varied simulation parameters such as noise levels, number of cells, number of mutation sites, number of clones, fraction of missing values and the number of unobserved clones.

For the case of unknown error rates, we compared the estimated FPR and FNR to the ground truth parameters. As shown in Figure 3.5B, the estimated parameters are close to the ground truth parameters for all but the single clone case. This demonstrates that OncoNEM estimates model parameters accurately over a wide range of simulation settings.

OncoNEM is robust to changes in ε Next, we assessed the sensitivity of OncoNEM to changes in the Bayes factor threshold ε. We applied OncoNEM to each simulated data set described in the previous section, using various values for ε and recoded the inferred number of clones (see Figure 3.6). In all simulation scenarios the number of clones is largely independent of ε, unless this parameter is set to very low values ($\varepsilon < 5$). Throughout all

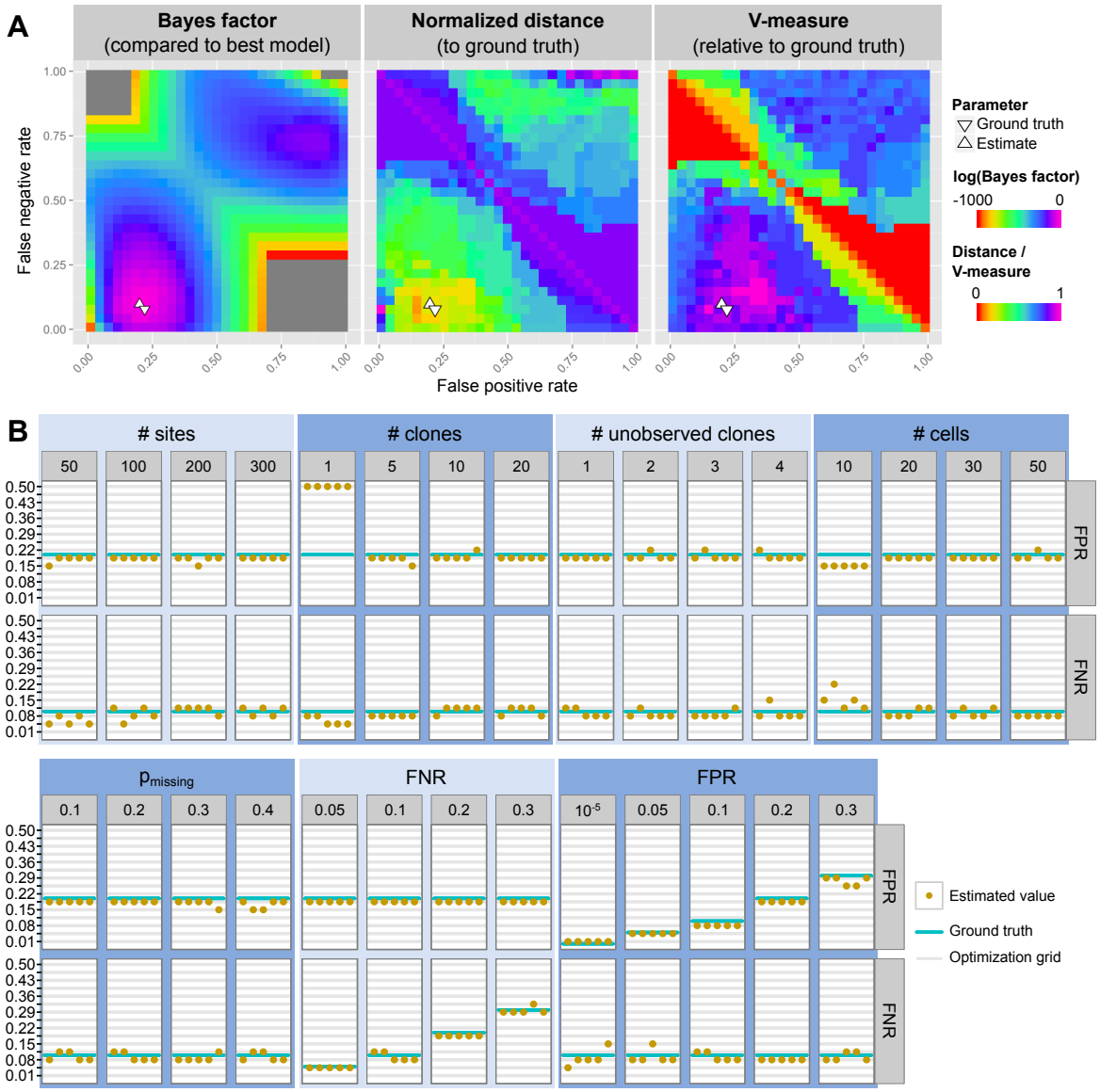


Fig. 3.5 Parameter estimation. A – Dependence of OncoNEM results on inference parameters. Log Bayes factor of highest scoring model inferred with given parameter combination relative to highest scoring model overall. The inferred parameters ($\hat{\alpha} = 0.22$, $\hat{\beta} = 0.08$) are close to the ground truth ($\alpha = 0.2$, $\beta = 0.1$). A large range of parameter combinations around the ground truth parameters yield solutions close to the ground truth tree in terms of pairwise cell shortest-path distance and V-measure. The distance was normalised to the largest distance observed between any inferred tree and the ground truth. B – Parameter estimation accuracy. FPRs and FNRs estimated by OncoNEM for various simulation settings with five replicates each. The blue lines mark the ground truth parameters. The grey lines mark the grid values over which FPR and FNR were optimised.

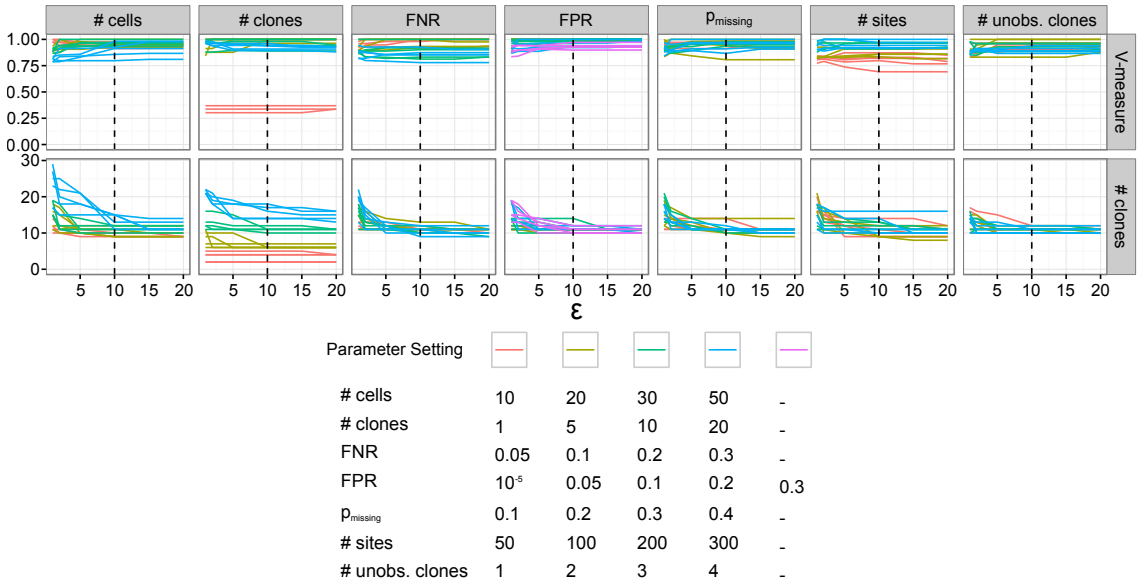


Fig. 3.6 Dependence of OncoNEM's clustering solution on Bayes factor threshold ε. This figure shows the V-measure and the number of clones of the OncoNEM solution as a function of epsilon for various simulation scenarios. Every line corresponds to one data set of the method comparison study. Lines are colour coded by parameter setting for the varied simulation parameter. The following settings were used as default parameters: 20 cells, 10 clones, FNR 0.1, FPR 0.2, 20% missing values, 200 mutation sites and 2 unobserved clones. In all simulation scenarios the number of clones is largely independent of ε unless it is set to be unreasonably small ($\varepsilon < 5$). The threshold ε used throughout the simulation and case studies is 10 (dashed line), and thus well within the stable range.

further simulation and case studies ε was kept constant at 10, which is well within the stable range.

3.3 Method comparison

3.3.1 Competing methods

Finally, using the same simulated data as above, we compared the performance of OncoNEM in case of known and unknown inference parameters to the performance of the six baseline methods. As representatives of classic phylogenetic methods we used likelihood optimisation of neighbour joining trees, as applied by Hughes et al. [60], and Bayesian phylogenetic inference, as used by Eirew et al. [33]. Both methods yield solutions where each cell corresponds to a different leaf in the tree. This type of tree is not directly comparable to the simulated one. To at least be able to evaluate the clustering solutions of the two methods,

we identified subpopulations of cells within these trees by hierarchical clustering of the trees' distance matrices with Silhouette score-based model selection. As representatives of hierarchical clustering based methods and the approaches used by Gawad et al. [47] and Yuan et al. [174], we used hierarchical and k-centroids clustering with Silhouette score-based model selection and subsequent minimum spanning tree construction. Furthermore, we compared our method to BitPhylogeny [174] and a method for inferring mutation trees by Kim and Simon [75].

For all but Kim and Simon's method, clustering performance was assessed using the V-measure, whereas the overall tree reconstruction accuracy was measured using the pairwise cell shortest-path distance. Both measures are described in the next section. Since Kim and Simon's method neither infers the position of the sequenced cells within the tree nor performs any clustering, V-measure and single cell shortest-path distance cannot be used to assess its performance. Instead, we calculated the accuracy of the inferred mutation orders.

3.3.2 Tree comparison measures

Measuring the accuracy of mutation orders

We define the *mutation order accuracy* of a tree $\mathscr{T}_1$ given the ground truth tree $\mathscr{T}_2$ as the average of

- the fraction of correctly inferred pairwise mutation orders, i.e. the probability that mutation a is upstream of mutation b in $\mathscr{T}_1$ given that a is upstream of b in $\mathscr{T}_2$, and
- the fraction of correctly inferred mutually exclusive mutations, i.e. the probability that two mutations a and b lie on separate branches in $\mathscr{T}_1$ given that a and b lie on separate branches in $\mathscr{T}_2$

for all mutations that belong to different clusters in $\mathscr{T}_2$.

Measuring clustering performance

Clustering performance was assessed using the V-measure [133], an entropy-based cluster evaluation measure that is calculated as the harmonic mean of the completeness and homogeneity scores of the clustering solution, where the completeness score measures the success of including all cells of a certain subpopulation in a cluster and the homogeneity score measures the success of including only cells from the same subpopulation in a cluster. The V-measure takes values from 0 to 1, with higher values indicating better performance.

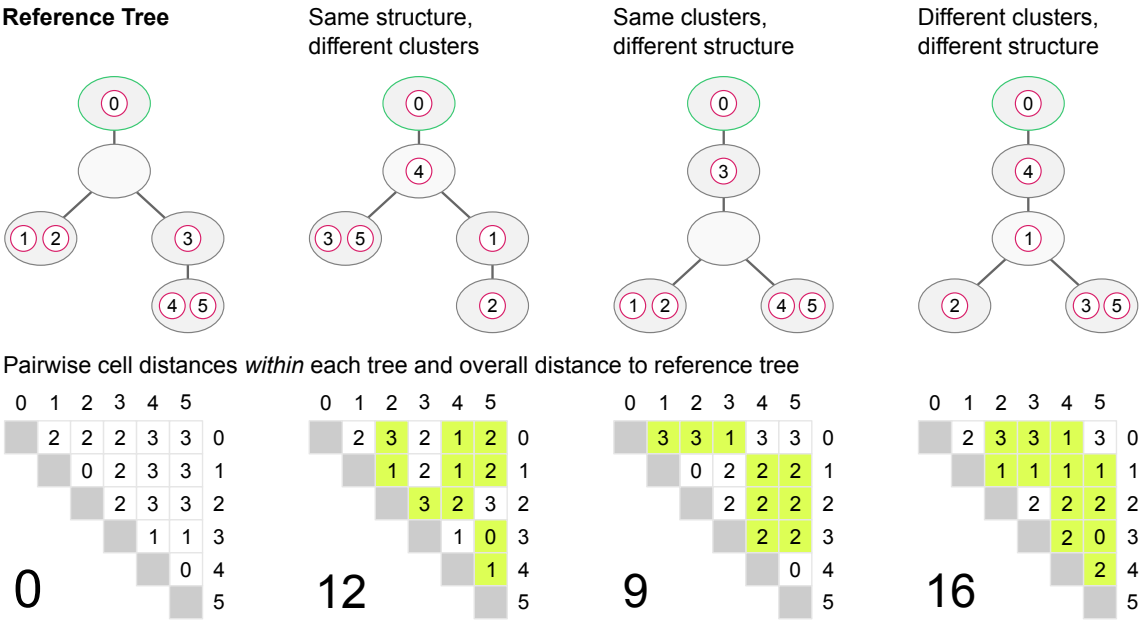


Fig. 3.7 Comparing Clonal Trees with the pairwise cell shortest-path distance. The yellow entries in the pairwise distance matrices indicate differences to the reference tree.

Measuring tree distances

Yuan et al. [174] developed the *consensus node-based shortest-path distance* to compare trees of clonal evolution, but this distance only takes nodes into account that appear in all trees that are being compared.

In order to take all of the observed nodes into account, we developed a distance measure called *pairwise cell shortest-path distance* (see Figure 3.7). This measure calculates the distance between two trees which are built on the same set of cells but may contain a differing number of nodes (clones). Briefly, for every pair of cells i and j, we compute the shortest-path $d_{ij}(\cdot)$ between the two cells in each tree. If the two cells belong to the same clone, their shortest-path distance is 0. Otherwise, the shortest-path distance equals the number of edges (regardless of direction) that separate the clones of the two cells. Finally, we sum up the absolute differences between the shortest-path distances of all unordered pairs of cells in the two trees to obtain the overall pairwise cell shortest-path distance. A formal definition of this distance and a proof that it defines a metric is provided in the following.

Definition of pairwise cell shortest-path distance Let $\mathscr{T}_1$ and $\mathscr{T}_2$ be two trees on the same set of cells $1,\ldots,n$ in which all leaf nodes contain at least one cell. In order to mark the root and to ensure that every node of the tree is taken into account in the distance measure, we add an extra cell with index 0 to the root. For two cells i and j in $0,\ldots,n$, let $d_{ij}(\mathscr{T})$ be

the number of edges separating the clones $c(i)$ and $c(j)$ in tree $\mathscr{T}$. With this we define the *pairwise cell shortest-path distance* between two trees $\mathscr{T}_1$ and $\mathscr{T}_2$ as

$$d(\mathscr{T}_1, \mathscr{T}_2) = \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} |d_{ij}(\mathscr{T}_1) - d_{ij}(\mathscr{T}_2)|.$$

In order to show that this distance is a metric, we first prove the following lemma.

Lemma If $d_{ij}(\mathscr{T}_1) = d_{ij}(\mathscr{T}_2)$ for $\forall\, i, j \in \{0, \ldots, n\}$, then $\mathscr{T}_1 = \mathscr{T}_2$.

Proof We prove this by showing that, if $d_{ij}(\mathscr{T}_1) = d_{ij}(\mathscr{T}_2)$ for all i, j in $0, \ldots, n$, we can define a bijection $\phi : V(\mathscr{T}_1) \mapsto V(\mathscr{T}_2)$ that maps the vertices of $\mathscr{T}_1$ to $\mathscr{T}_2$ so that

(I) cell i is in clone v_j in tree $\mathscr{T}_1 \Leftrightarrow$ cell i is in clone $\phi(v_j)$ in tree $\mathscr{T}_2$ and

(II) $v_i = \mathrm{pa}(v_j) \Leftrightarrow \phi(v_i) = \phi\left(\mathrm{pa}(v_j)\right)$, where $\mathrm{pa}(v_j)$ is the parent of node v_j.

We construct the bijection in several steps.

(I) Assignment of observed clones.
The distance matrix uniquely defines the observed clones of a tree: Two cells i and j belong to the same clone if and only if $d_{ij}(\mathscr{T}) = 0$. From this it follows that two trees with the same cell distances contain the same observed clones, i.e. we can define a bijection ϕ between the observed clones of $\mathscr{T}_1$ and $\mathscr{T}_2$.

(II) Parent-child relationships between observed clones.
Let i and j be two cells in separate clones of $\mathscr{T}_1$, i.e. $c(i) \neq c(j)$. Then, $c(i) = \mathrm{pa}(c(j))$ if and only if

(i) $d_{ij}(\mathscr{T}) = 1$ and

(ii) $d_{0j}(\mathscr{T}) = d_{0i}(\mathscr{T}) + 1$.

Since the pairwise cell distances are equal in both trees, ϕ preserves the parent-child relationships between observed clones.

(III) Assignment of unobserved clones.
Let v_i be an unobserved clone of $\mathscr{T}_1$. Then v_i has an observed descendant $v_j \in \mathscr{T}_1$, since all leaf nodes of $\mathscr{T}_1$ contain at least one cell by the definition of the pairwise cell shortest-path distance. We call v_i the n-th ancestor of v_j. Since the distances from the

root to v_j in $\mathscr{T}_1$ and to $\phi(v_j)$ in $\mathscr{T}_2$ are the same, respectively, we can assign $\phi(v_i)$ as the n-th ancestor of $\phi(v_j)$.

This is a well-defined function as the assignment does not depend on the specific choice of the descendant v_j. To see this, choose another observed descendant $\bar{v}_j$. Then v_i is the $\bar{n}$-th ancestor of $\bar{v}_j$. The distances $d_{0v_j}(\mathscr{T}_1)$, $d_{0\bar{v}_j}(\mathscr{T}_1)$ and $d_{v_iv_j}(\mathscr{T}_1)$ uniquely determine the length of the branches connecting the three nodes via v_i. Since those distances are the same in $\mathscr{T}_2$ it follows that the n-th ancestor of $\phi(v_j)$ is the same node as the $\bar{n}$-th ancestor of $\phi(\bar{v}_j)$.

This function defines a bijection between the vertices since the same construction from $\mathscr{T}_2$ to $\mathscr{T}_1$ defines an inverse.

(IV) General parent-child relationships.
Assuming that $v_i = \mathrm{pa}(v_j)$ we consider the following cases to show that $\phi(v_i) = \mathrm{pa}(\phi(v_j))$:

(i) v_i and v_j observed – see step (II).

(ii) $v_i = \mathrm{pa}(v_j)$ and v_i unobserved
Assign $\phi(v_i)$ following step (III) as first ancestor of $\phi(v_j)$. Then $\phi(v_i) = \mathrm{pa}(\phi(v_j))$.

(iii) $v_i = \mathrm{pa}(v_j)$, v_i observed and v_j unobserved
Choose v_k so that it is an observed descendant of v_j. Assign $\phi(v_j)$ following step (III). Then $\phi(v_j)$ is the n-th ancestor of $\phi(v_k)$ and $\phi(v_i)$ is the $(n+1)$-th ancestor of $\phi(v_k)$, so that $\phi(v_i) = \mathrm{pa}(v_j)$.

(iv) $v_i = \mathrm{pa}(v_j)$ and both unobserved
Choose v_k so that it is an observed descendant of v_j. Assign $\phi(v_i)$ and $\phi(v_j)$ following step (III). Then $\phi(v_j)$ is the n-th ancestor of $\phi(v_k)$ and $\phi(v_i)$ is the $(n+1)$-th ancestor of $\phi(v_k)$, so that $\phi(v_i) = \mathrm{pa}(v_j)$.

The same construction from $\mathscr{T}_2$ to $\mathscr{T}_1$ defines the inverse of ϕ. Therefore, $v_i = \mathrm{pa}(v_j) \Leftrightarrow \phi(v_i) = \phi\left(\mathrm{pa}(v_j)\right)$. □

With this, we can show that $d(\mathscr{T}_1, \mathscr{T}_2)$ satisfies all the requirements of a metric.

Proof that pairwise cell shortest-path distance is a metric

(I) Non-negativity: $d(\mathscr{T}_1, \mathscr{T}_2) \geq 0$
Non-negativity follows from defining the distance as a sum of absolute values.

(II) Coincidence axiom: $d(\mathscr{T}_1, \mathscr{T}_2) = 0$ if and only if $\mathscr{T}_1 = \mathscr{T}_2$

(i) If $\mathscr{T}_1 = \mathscr{T}_2$, then $d(\mathscr{T}_1, \mathscr{T}_2) = d(\mathscr{T}_1, \mathscr{T}_1) = 0$.

(ii) If $d(\mathscr{T}_1, \mathscr{T}_2) = 0$, then $\mathscr{T}_1 = \mathscr{T}_2$ as shown in the lemma.

(III) Symmetry: $d(\mathscr{T}_1, \mathscr{T}_2) = d(\mathscr{T}_2, \mathscr{T}_1)$
Symmetry follows from $|d_{ij}(\mathscr{T}_1) - d_{ij}(\mathscr{T}_2)| = |d_{ij}(\mathscr{T}_2) - d_{ij}(\mathscr{T}_1)|$.

(IV) Triangle inequality: $d(\mathscr{T}_1, \mathscr{T}_3) \leq d(\mathscr{T}_1, \mathscr{T}_2) + d(\mathscr{T}_2, \mathscr{T}_3)$
The usual triangle inequality for real numbers implies that

$$|d_{ij}(\mathscr{T}_1) - d_{ij}(\mathscr{T}_3)| \leq |d_{ij}(\mathscr{T}_1) - d_{ij}(\mathscr{T}_2)| + |d_{ij}(\mathscr{T}_2) - d_{ij}(\mathscr{T}_3)|.$$

Hence for the pairwise cell shortest-path distance we find

$$\begin{aligned} d(\mathscr{T}_1, \mathscr{T}_3) &= \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} |d_{ij}(\mathscr{T}_1) - d_{ij}(\mathscr{T}_3)| \\ &\leq \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} |d_{ij}(\mathscr{T}_1) - d_{ij}(\mathscr{T}_2)| + \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} |d_{ij}(\mathscr{T}_2) - d_{ij}(\mathscr{T}_3)| \\ &= d(\mathscr{T}_1, \mathscr{T}_2) + d(\mathscr{T}_2, \mathscr{T}_3). \end{aligned}$$

□

3.3.3 Results

The results of the method comparison are shown in Figure 3.8. OncoNEM substantially outperforms the other methods for all simulation scenarios but the single clone case. It consistently yields results that have a smaller distance to the ground truth and a higher V-measure than the baseline methods or, in the case of oncogenetic trees, infers the order of mutation with a much higher accuracy. Overall, OncoNEM's performance in the case of unknown model parameters is comparable to its performance with given parameters.

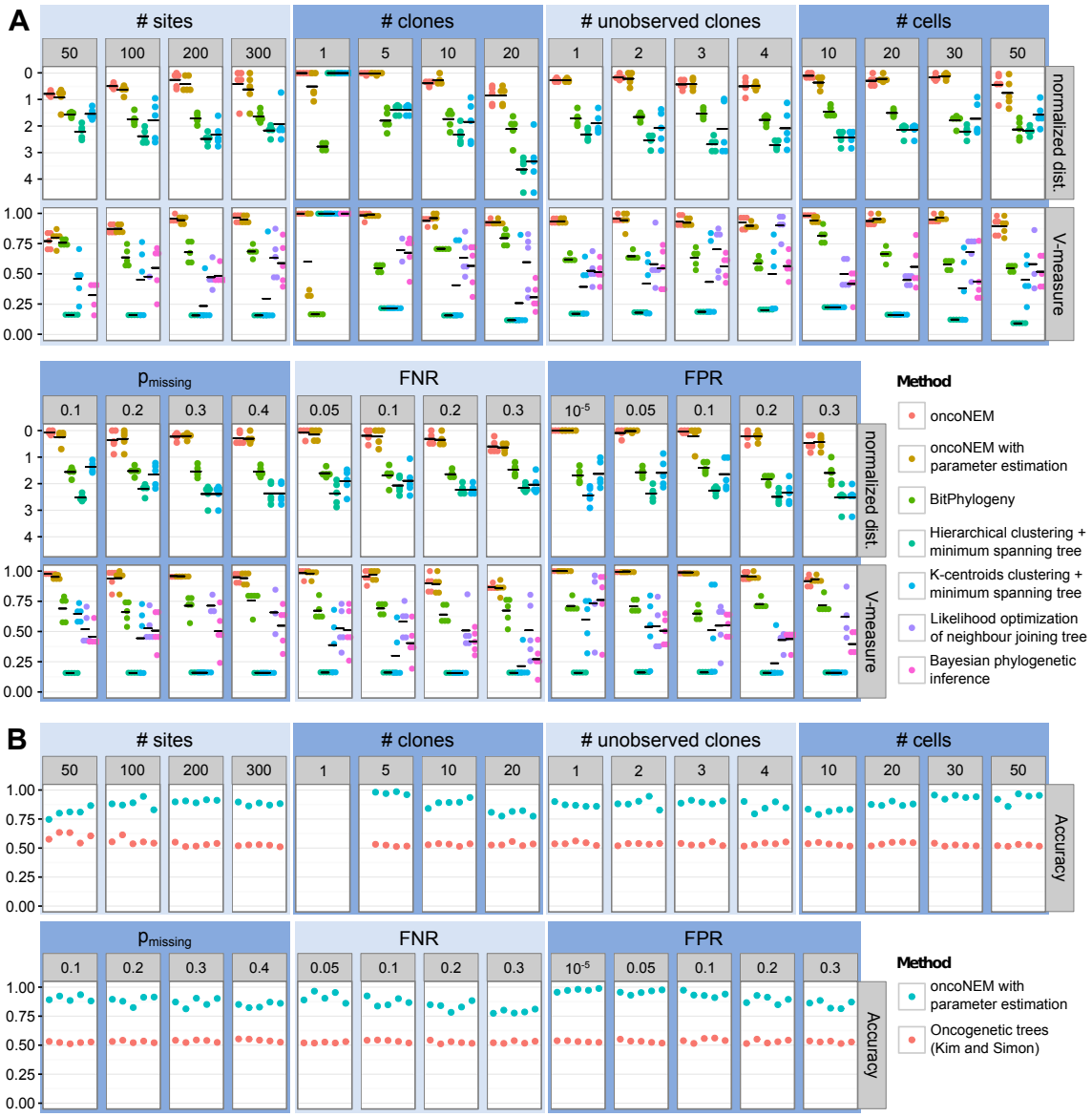


Fig. 3.8 OncoNEM performance assessment. A – Performance comparison of OncoNEM and five baseline methods. Shown are the distance and V-measure of inferred trees to ground truth. Results of single simulations are marked by dots and coloured by method, while black horizontal bars indicate the mean over five simulations for each method. The distances shown were normalised for the number of cells n in the trees and were obtained by dividing the pairwise cell shortest-path distances by $n(n-1)/2$. Distances could only be calculated for three of the baseline methods. Values of the varied parameters are shown in the panels at the top. As default parameters, we used an FNR of 0.1, an FPR of 0.2, 200 sites, 10 clones, no unobserved clones, 20 cells and 20% of missing values. B – Performance comparison of OncoNEM and Kim and Simon's oncogenetic tree method. Shown is the mutation order accuracy of the inferred trees for each of the simulated datasets. This measure is undefined for datasets without mutually exclusive mutations. Therefore, no values are shown for the single clone case and the first replicate of the 5 clone scenario, for which the simulated tree is linear.

3.4 Case studies

3.4.1 Analysis of a muscle-invasive bladder transitional cell carcinoma

We used OncoNEM to infer the evolutionary history of a muscle-invasive bladder transitional cell carcinoma previously analysed by Li et al. [85], who performed single-cell exome sequencing of 44 tumour cells, as well as exome sequencing of normal and tumour tissue. Li *et al* estimated the average ADO rate to be 0.4 and the FPR to be 6.7×10^{-5}. Using a census-filtering threshold of 3, they identified 443 SSNVs across the 44 cells. In their final genotype matrix, 55.2% of the values were missing.

We binarised the genotype matrix by setting homozygous normal sites to 0 and hetero- or homozygous mutant sites to 1 and applied OncoNEM as described above. The resulting tree is shown in Figure 3.9b. The single linear branch from the normal suggests that all cells in the data set are descendants of a single founder cell. The tree contains three major subpopulations. The least mutated of these subpopulations carries about a quarter of the detected mutations. These trunk mutations are shared by almost all of the analysed cells. This early clone gave rise to multiple divergent subpopulations two of which are large and again diversified into smaller subclones.

These results agree with the results of Li et al. who inferred three main subpopulations (A, B, C) with B and C having evolved from A. Mapping the clone labels of Li et al. onto the OncoNEM tree shows, however, that the assignment of cells to clones differs between the two approaches (see Figure 3.10). Li et al. also inferred the origins of 8 mutations in 7 genes that are commonly altered in muscle-invasive bladder transitional cell carcinomas. A comparison of their results with the posterior probability of θ inferred by OncoNEM is shown in Table 3.1. The assignment of mutations to clones agrees in 7 out of 8 cases.

OncoNEM estimated the FPR to be 0.185 (see Figure 3.9a). This error rate is higher than the expected value under the binomial model used for consensus filtering by Li et al., which suggests that there might be recurrent sequencing errors in the dataset. The FNR was estimated to be 0.08. This estimated value lies within the expected range of less than half the estimated ADO rate. See the parameter estimation section within the Methods for an explanation of the conceptual differences between the original error rates estimated by Li *et al* and the OncoNEM parameters.

To test the robustness of our results, we inferred trees using model parameters that are slightly different from the estimated ones (see Figure A.2). The structure and the overall features of the resulting trees are close to the original estimate, which further supports our results.

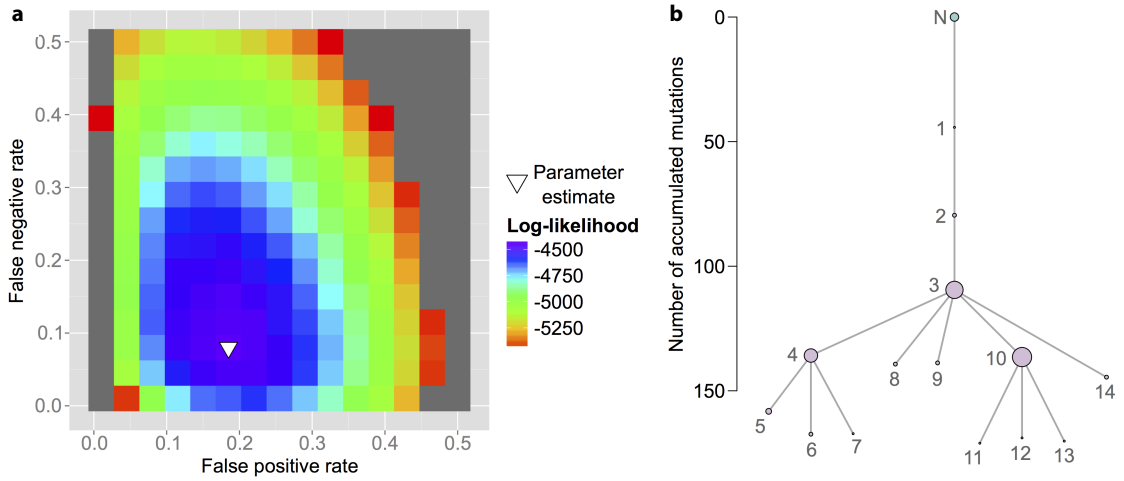


Fig. 3.9 Results inferred by OncoNEM from bladder cancer dataset. a – The estimated error rates are $\alpha = 0.185$ and $\beta = 0.08$. b – The inferred tree suggests a branching evolution with three major subpopulations.

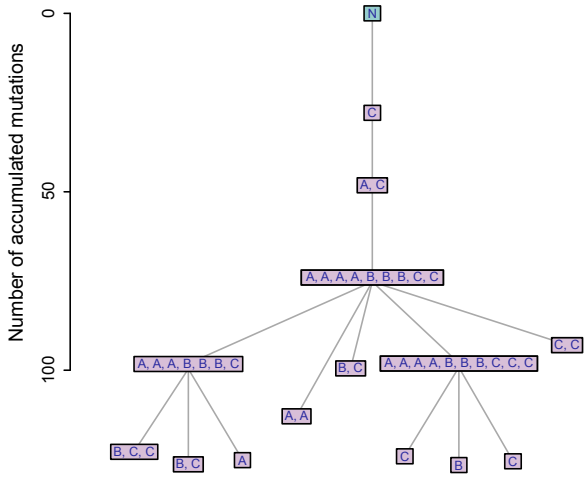


Fig. 3.10 Mapping of clone labels assigned by Li et al. [85] onto cells in oncoNEM tree. Comparison of clone labels between the two trees shows that the assignment of cells to clones differs between oncoNEM and the results by Li et al.

Impact of loss of heterozygosity on inference results The OncoNEM model assumes that mutations are never lost. Deletions that lead to loss of heterozygosity (LOH) are however common in various types of cancer.

We expect that our algorithm is able to infer good solutions despite LOH events, as long as the fraction of mutations affected by LOH is relatively small. In this case, LOH-affected sites will simply contribute to the error rates of false positives and false negatives, depending on whether the deletion occurred early or late after the original occurrence of the SNV.

Table 3.1 Comparison of origin of mutations inferred by OncoNEM with origins inferred by Li *et al*. Posterior probabilities of θ inferred by OncoNEM for the eight recurrently mutated genes analyzed by Li *et al*. A, B and C denote the clones inferred by Li *et al*, 1 to 14 denote the clones inferred by OncoNEM. Visual comparison of the OncoNEM tree with the phylogeny inferred by Li *et al* suggests that clone A corresponds to clones $1-3$, clone B corresponds to clones $4-7$ and clone C corresponds to clones $10-13$. Overall, both methods assign mutations to the same clones. *KIAA1958*[1] denotes mutation at chromosome 9, position 114376732. *KIAA1958*[2] denotes mutation at chromosome 9, position 114376902.

		1	2	3	4	5&7	6	10	11&12	13	8,9&14
A	*NIPBL*	***0.33***	***0.33***	***0.33***	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	CFTR	**0.45**	**0.45**	**0.09**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	DHX57	**0.45**	**0.45**	**0.09**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ASTN1	**0.25**	**0.25**	**0.51**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B	*ATM*	0.00	0.00	0.00	**1.00**	**0.00**	**0.00**	0.00	0.00	0.00	0.00
C	*COL6A3*	0.07	0.07	0.07	0.00	0.00	0.00	**0.76**	**0.01**	**0.00**	0.00
	KIAA1958[1]	0.00	0.00	0.00	0.00	0.00	0.00	**0.98**	**0.00**	**0.00**	0.00
	KIAA1958[2]	0.19	0.19	0.19	0.02	0.00	0.02	**0.38**	**0.00**	**0.00**	0.00

To support this claim, we identified the LOH-affected regions of the bladder cancer from a bulk-sequencing analysis by Li et al. (see Table A.1) and removed all mutations within these regions from the mutation dataset (6.3% of all variant sites). We then applied OncoNEM to this reduced data set and compared the solution to the one obtained from the full data set. Figure A.3 shows that the inferred tree is largely stable and the overall tree structure remains the same.

3.4.2 Analysis of a case of essential thrombocythemia

In the second case study, we applied OncoNEM to a data set derived by single-cell exome sequencing of 58 single cells from an essential thrombocythemia [58]. Hou et al. estimated the average ADO rate to be 0.42 and the FPR to be 6.4×10^{-5}. Using a census-filtering threshold of 5, they identified 712 SSNVs. Their final genotype matrix contained 57.7% of missing values.

The genotypes were binarised, and OncoNEM was applied as in the previous case study. The inferred tree is shown in Figure 3.11b. Again, the tree suggests that all tumour cells are descendants of a single founder cell. The majority of cells belong to subpopulations that are related through a linear trajectory. All detected branching events have occurred late during tumour development, i.e. after the tumour had already acquired more than 60% of its mutations.

These results agree with the somatic mutant allele frequency spectrum analysis of Hou et al. that suggests that the neoplasm is of monoclonal origin [58], while Kim et al. inferred a mutation tree with a complex hierarchy [75]. Using BitPhylogeny, Yuan et al. [174] inferred a polyclonal origin. However, with 58 cells, the data set might be too small for their method to converge.

OncoNEM estimated the FPR and FNR to be 0.255 and 0.185, respectively (see Figure 3.11 a). The FPR estimate is again higher than expected under the binomial model, whereas the FNR lies within the expected range. As in the previous case study, running OncoNEM with similar parameters yields similar trees (see Figure A.4).

Given the error rates inferred by OncoNEM, the log-likelihood of the BitPhylogeny tree computed under the OncoNEM model is -11584, whereas the OncoNEM tree has a log-likelihood of -9964. The fact that the OncoNEM solution has a much higher likelihood than the BitPhylogeny tree shows that the differences are not due to the heuristic nature of OncoNEM's search algorithm, but instead suggest that BitPhylogeny did not converge to the optimal solution.

In summary, these two case studies show how OncoNEM can extend and improve on previous analyses of this data set.

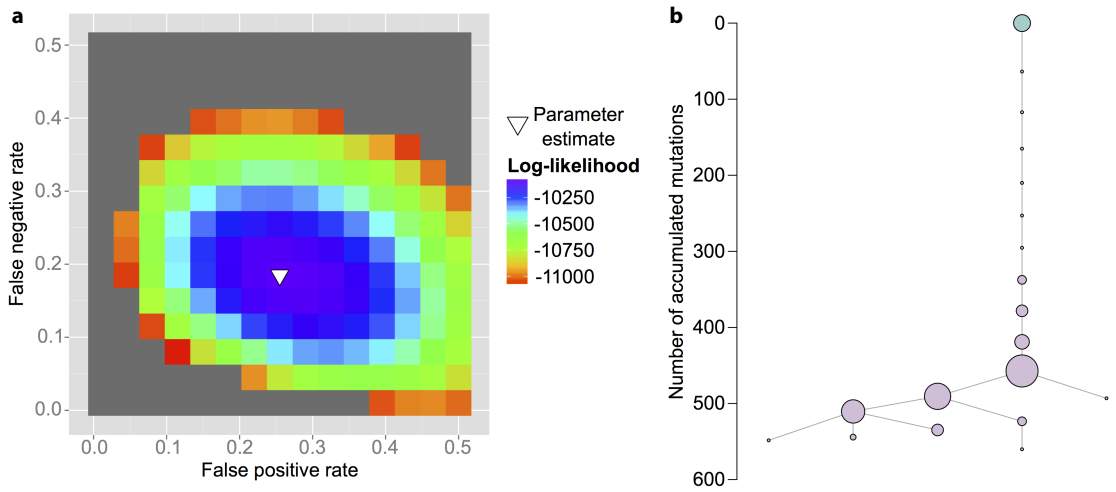


Fig. 3.11 Results inferred by OncoNEM from essential thrombocythemia dataset. a – The estimated error rates are $\alpha = 0.255$ and $\beta = 0.185$. b –The inferred tree suggests a largely linear evolution with some small subpopulations branching off late during tumour evolution.

3.5 From cell trees to mutation trees

So far we have focused on the inference of cell or clonal lineage trees. With some small modifications, the OncoNEM model can be adapted to infer mutation trees. These are trees that describe the order in which mutations were acquired during tumour evolution. The main difference between mutation trees and clonal lineage trees is that the nodes of mutation trees are mutations, whereas the nodes of cell lineage trees are cells. Adapting NEMs to infer mutation trees yields the likelihood model used by SCITE (single-cell inference of tumour evolution) [65], which was published shortly after OncoNEM. The option to infer mutation trees is also implemented in the OncoNEM R package.

Even though the concepts of clonal lineage trees and mutation trees are closely related and, essentially, both trees can describe the same information, one inference model can be preferable over the other, depending on the input data. If the number of cells n is larger than the number of mutations m, inferring mutation trees can be advantageous, as this reduces the size of the inferred model in comparison to cell lineage trees.

To infer cell lineage trees, we marginalise over the attachment of mutations, whereas to infer mutation trees, we marginalise over the attachment of cells. For this reason, the inference model for the mutation tree case is even closer related to traditional NEMs than the model for cell lineage trees (see Section 3.1.2 for a comparison of the cell lineage tree model and traditional NEMs).

No inversion of subset relationships As with NEMs, the patterns of upstream nodes are a superset of the patterns of downstream nodes, or in other words, mutations that occurred late during tumour evolution are present in a subset of the cells that carry an upstream mutation.

Targeted perturbations In case of both cell lineage and mutation trees, mutations correspond to perturbations. In case of mutation trees, however, the node of mutation l is the origin of the perturbation. This is directly analogous to NEMs, where the signalling genes are the origins of the perturbations. As with NEMs, the effects of the perturbations are observed via downstream reporters and the connection between graph components and reporters is unknown *a priori*.

Graph size and structure As with NEMs, the graph size is defined by the number of perturbations. There are no unobserved nodes. The only difference in comparison to NEMs is that the search space is the space of arbitrary trees instead of the space of directed acyclic graphs.

Comparison with SCITE

Theoretical comparison

Despite the similarity in the likelihood models, the tools OncoNEM and SCITE [65] show important differences. Some of these are conceptual while others are technical.

Clonal lineage tree versus mutation tree OncoNEM was primarily developed to infer clonal lineage trees, whereas the main focus of SCITE is to infer mutation trees. The software packages of both methods offer options to infer the other type of tree. SCITE, however, does not infer unobserved nodes, which are needed to account for unobserved subpopulations in the inference of cell trees.

Inference of error parameters OncoNEM estimates both false positive and false negative rates. The current implementation of SCITE only infers the false negative rate. Jahn et al. [65] simply use the FPR estimated directly from the sequencing data. However, as shown earlier, the FPR estimate from the sequencing data and the false positive rate of the likelihood model are different conditional probabilities.

Input data OncoNEM uses binary input data, SCITE has two different options (*i*) a binary mode and (*ii*) a mode that distinguishes between heterozygous mutant/normal sites and homozygous mutant sites.

Inference algorithms OncoNEM uses a heuristic neighbourhood search, whereas SCITE uses Markov-chain Monte-Carlo sampling (MCMC). A disadvantage of the neighbourhood search is that for large trees the neighbourhood can become very big and the search becomes too computationally expensive. The purpose of MCMC is to generate a sample from the posterior distribution. Thereby MCMC does not generate just one solution tree but a sample from the distribution of solutions, which allows assessing the uncertainty. In case of OncoNEM, a list of highest scoring trees can be used to assess stability of the inferred solution.

Comparison in a simulation study

Salehi et al. [144] compared the performance of OncoNEM and SCITE in a simulation study.

Data simulation Salehi et al. simulated data sets for 10 different tree topologies. In each case, 10 clones and 48 mutation sites were simulated. Their simulation procedure consists of three steps.

First, they simulate genotypes for a given tree using a Generalised Dollo Model. Under this model, a point mutation can only be gained once during the evolution of a tumour, but it can also be lost again once or several times on independent branches. This is to take into account the effects of deletions. Second, they simulate the single-cell data. To sample the number of cells corresponding to each clone, they sample from a Dirichlet-multinomial distribution

$$(n_1, \ldots, n_M) \sim \text{Dirichlet-multinomial}(\lambda\Phi),$$

where Φ is a vector containing the ground truth cellular prevalences of the clones and λ is a distortion parameter. The higher λ, the closer the simulated cellular prevalences are to the ground truth prevalences. Datasets containing between 50 and 1000 cells were simulated. However, OncoNEMs performance for data sets with 200 cells or more could not be assessed because of computational constraints. Third, they introduce ADO errors and doublets. Apart from errors caused by doublets, no false positives are simulated.

In total Salehi generate data for three scenarios (*i*) perfect data without ADO and doublets (denoted as $\lambda = \infty$), (*ii*) data with moderate levels of sampling distortion ($\lambda = 10$) in combination with an ADO rate of 30% as well as 30% doublet cells and (*iii*) data with a high level of sampling distortion ($\lambda = 1.12$), which is reflective of real data according to the authors of the study, and again 30% of ADO and doublet cells.

Analysis The data was analysed using OncoNEM, SCITE and ddClone, the method developed by the authors of this study. As described in Chapter 2, ddClone uses bulk-sequencing data in addition to the single-cell data for the inference. As in our own simulation studies, the V-measure was used to evaluate the clustering performance. In addition, the accuracy of cellular prevalence estimates was evaluated. The accuracy of the inferred tree topology was not assessed.

Results The results of the simulation study are summarised in Table 3.2. Unsurprisingly, ddClone outperforms OncoNEM and SCITE in terms of cellular prevalence estimation error in case of moderate or high sampling distortion, as it obtains additional information about

Table 3.2 Average cellular prevalence estimation error (top) and V-measure (bottom) with standard deviation. Data from Salehi et al. [144]. Simulation parametes: (i) $\lambda = \infty$, 0% ADO, 0% doublets; (ii) $\lambda = 10$, 30% ADO, 30% doublets; (ii) $\lambda = 1.12$, 30% ADO, 30% doublets;

$\mathbf{e}_\Phi$	(i)	(ii)	(iii)
OncoNEM	0.04 ± 0.01	0.13 ± 0.03	0.17 ± 0.03
SCITE	0.06 ± 0.01	0.18 ± 0.05	0.18 ± 0.05
ddClone	0.06 ± 0.01	0.07 ± 0.02	0.09 ± 0.03

V-measure	(i)	(ii)	(iii)
OncoNEM	0.90 ± 0.03	0.81 ± 0.03	0.74 ± 0.06
SCITE	0.87 ± 0.09	0.75 ± 0.05	0.71 ± 0.08
ddClone	0.86 ± 0.04	0.79 ± 0.09	0.77 ± 0.06

cellular prevalences from the bulk-sequencing data. OncoNEM and SCITE show a very similar performance in all cases. In terms of V-measure ddClone does not show a significant performance advantage in comparison to OncoNEM. OncoNEM and SCITE also show a similar performance, with OncoNEM performing marginally better on average.

3.6 Limitations

OncoNEM has limitations, some of which were mentioned in a research highlight article by Davis and Navin [24].

Limitations of model assumptions

The OncoNEM likelihood model is based on the infinite sites assumption, i.e. it assumes that during tumour evolution every mutation site is only mutated once. Deletions can violate this assumption as they can lead to the loss of somatic SNVs. If a deletion only affects a single SNV and occurred in a clone that arose late during tumour evolution, the deletion affects few of the entries in the genotype matrix. OncoNEM would account for the resulting discrepancies between observed and expected genotype matrix by assuming additional false positives or negatives. As the number of additional errors needed to explain the observed data is small, the likelihood scores will differ only marginally from the scores in the case without deletion. Therefore, the deletion should not have a significant influence on the inference result.

However, copy number changes can occur early during tumour evolution. Likewise, they often affect larger regions of the genome. This means deletions can lead to systematically

missing mutations. In principle, OncoNEM, which does not account for deletions, could still explain the observed data by assuming genotyping errors. In these cases, however, the assumption that observed genotypes are independent and identically distributed is violated.

For these reasons, our general recommendation is to blacklist LOH-affected regions before OncoNEM inference if copy-number informative data is available. If the evolution of a tumour is known to be copy number driven and LOH affects large parts of the genome, we recommend using a copy number based method for inferring tumour evolution.

Similar to deletions, sequencing of doublets can also yield mutation patterns that are systematically incompatible with a tree structure under the infinite sites assumption. To account for this, one could extend the OncoNEM search space from directed trees to directed acyclic graphs. However, this would make the search even more computationally expensive due to an even larger search space. In addition, it is unclear how robustly OncoNEM would be able to distinguish whether incompatible mutation patterns are caused by sequencing errors or doublets. Nevertheless, the simulations performed by Salehi et al. [144] demonstrate that OncoNEM in its current form can yield good clustering solutions even in the presence of doublets.

Limitations of global error parameters

OncoNEM uses error parameters that are constant across all sites and cells. Ideally, the OncoNEM error probabilities would depend on the confidence of the mutation call in a cell- and site-specific manner. While this is easy to implement, an approach to estimate these error rates is lacking. One possibility could be to estimate local error parameters by adjusting global parameters on the basis of local sequencing parameters such as quality scores. Our simulation studies showed, however, that OncoNEM is robust to changes in the error parameters. This suggests that local adjustments of these parameters may not be necessary to obtain robust results and it would be interesting to test this in future simulations.

Restrictions on data set size

OncoNEM's search algorithms are suitable for handling single-cell datasets of the current size. In the future data sets are likely to contain many more cells and the inference of cell trees with the current algorithms will become too computationally expensive.

One of the limiting factors is the heuristic neighbourhood search. As the number of nodes in the tree grows, the number of trees in the neighbourhood that are scored in each step becomes larger and larger. To make the search algorithm more efficient, one could optimise the heuristic search algorithm, for example by limiting the number of neighbouring

trees generated in each step. In addition, the parameter estimation could be accelerated by implementing a local parameter optimisation instead of a brute force grid search.

Even with these adaptions, OncoNEM will not be able to infer cell trees for data sets with a very large number of cells. In this case, we recommend inferring mutation trees instead of cell trees, which means that the computational cost is mainly determined by the number of mutations instead of the number of cells. Of course, the number of mutations in the data set may also be too large. In most cases, it should, however, be possible to reduce the number of mutations to a number that OncoNEM can handle by limiting the data to mutations that are of special interest, such as non-synonymous mutations or driver mutations.

Dependence on data quality

As with bulk-sequencing inference methods, single-cell methods depend on the quality of the input data. On the one hand, this concerns the quality of the sequencing itself which in turn depends on parameters such as allele drop-out and doublet rates. On the other hand, the collected cells need to be a representative sample of the tumour to yield a sensible solution. Sampling biases or under-sampling lead to an underestimation of tumour heterogeneity. Increasing the number of cells can improve all of these issues, but this needs to be balanced against sequencing costs and computational constraints. More work is needed to estimate the number of cells that are required to infer reliable phylogenies.

Chapter 4

Inference of multi-sample phylogenies

So far the thesis has focused on the inference of phylogenetic trees from single-cell sequencing data. This chapter describes the phylogenetic analysis of ten lethal metastatic breast cancers using multi-sample bulk sequencing data. In the first part, I present a new allele-specific multi-sample copy number segmentation algorithm that improves the copy number based phylogenetic inference (Section 4.1). The second part describes the phylogenetic analysis. First, I give an overview of the data (Section 4.2.1) and describe the steps of the phylogenetic analysis (Section 4.2.2), which include the application of the new segmentation algorithm. Then, I present the inferred trees and compare results across methods and data types (Section 4.2.3). This chapter contains text from Ross et al. [135] and figures from De Mattos Arruda et al. [25].

4.1 Allele-specific multi-sample segmentation

As explained in Chapter 2, the reliability of sample tree inference from copy number data depends on the accurate estimation of the number of mutation events that separate the copy number profiles of different samples. Two methods that aim to do this are MEDICC [151] and TuMult [81]. Both of these methods use breakpoint based input data, where breakpoints are defined as genomic locations that show a change in copy number status [81]. Given the smallest difference between the breakpoint locations of two samples, these methods assume that the breakpoints were caused by different events. Therefore, breakpoints that are shared between samples need to be perfectly aligned with each other to produce phylogenies of high quality.

Current approaches to inferring copy number profiles The main two steps of inferring allele-specific copy number profiles are (i) the segmentation of chromosomes into

regions of constant copy number and (ii) the estimation of purity and ploidy values to convert the segmented relative copy number values into integer copy number profiles. The most frequently used methods to perform the second step are ASCAT (Allele-Specific Copy number Analysis of Tumors) by Van Loo et al. [164] and ABSOLUTE by Carter et al. [16]. For the segmentation step, a wide variety of methods exist.

Segmentation algorithms can be distinguished by the mathematical model they use, by the type of data they segment, total or allele-specific copy number data, and depending on whether they perform single- or multi-sample segmentation.

- Most segmentation algorithms use circular binary segmentation (CBS) [118], hidden Markov models (HMM) [44, 98] or regression-based approaches such as piecewise constant fitting (PCF) [123] and regression trees [18].

- Some methods only segment *total* copy number data, while others perform *allele-specific segmentation*. Allele-specific segmentation can detect copy number neutral changes, for example when one allele is deleted, and the other one is duplicated. Allele-specific data is therefore preferable for phylogenetic inference and is used as input data by MEDICC.

- Some segmentation algorithms can only be applied to a single sample at a time. *Multi-sample segmentation* algorithms segment multiple related samples jointly and infer both private and shared segment boundaries.

Allele-specific versions of some of the methods mentioned above exist. Patchwork [102], for example, uses CBS and aspcf [114] uses PCF. However, current allele-specific methods segment each sample on its own and, due to the noise in the data, the inferred locations of breakpoints that are shared between samples may differ.

To address this problem, the authors of the copy number-based tumour phylogeny inference algorithm MEDICC performed extensive experimental breakpoint validation of their data sets [150]. This is an expensive approach which has often been omitted by similar papers. Mangiola et al. [96] and Gerlinger et al. [50], for example, used size-based heuristic filters for CNAs instead. However, the fact that a segment is small does not automatically imply that it is an artefact. Size-based filters do not assess the evidence for a given segment in the raw segmentation data and therefore cannot distinguish between segments that correspond to small highly amplified regions and segments that are the result of misaligned segment boundaries.

While experimental breakpoint validation was not feasible in our study, we wanted to use a more rigorous approach than a heuristic size filter. We, therefore, addressed the problem of

multi-sample breakpoint detection by developing *asmultipcf* (allele-specific multi-sample piecewise constant fitting), an algorithm performing joint allele-specific segmentation of multiple samples to infer private and shared segment boundaries of phylogenetically related samples.

Asmultipcf is based on two copy number segmentation algorithms developed by Nilsen et al. [114], which use penalised least square principles to fit piecewise constant segments to the data. The first algorithm, *aspcf*, performs allele-specific segmentation and is the segmentation algorithm used in ASCAT Van Loo et al. [164]. The second one, *multipcf*, is a multi-sample segmentation algorithm, which is however not allele-specific. Combining these two algorithms provides a straightforward approach to address the multi-sample breakpoint inference problem for our phylogenetic analyses. As an extension to the existing algorithms, asmultipcf handles missing values, making extensive data filtering unnecessary. The algorithm is described in detail in the following.

4.1.1 Data for allele-specific copy number segmentation

Allele-specific copy number segmentation makes use of two data types *log ratios* (logR) and *B-allele frequecies* (BAF) of germline heterozygous sites. They can be obtained from sequencing data or SNP (single nucleotide polymorphism) arrays [9].

Log ratio The logR is a relative measure of the total copy number of the cancer cell population at a given genomic locus. An increase in relation to the baseline logR indicates an amplification of the genome, whereas a decrease indicates a deletion. LogR values alone can be used to estimate absolute non-allele-specific copy number profiles. In the following, we assume that logR values have been adjusted for GC biases before segmentation, for example using ASCAT's *GCcorrect* function.

B-allele frequency The BAF measures the relative abundance of two alternative alleles A and B at a genomic locus [9],

$$\mathrm{BAF} = \frac{n_i^B}{n_i^A + n_i^B}. \tag{4.1}$$

In noise-free data, a homozygous locus has a BAF of 0 or 1, whereas a heterozygous site of copy number 2 has a BAF of 0.5.

4.1.2 Copy number segmentation by piecewise constant fitting

In their piecewise constant fitting algorithms, Nilsen et al. [114] use a penalised least squares approach to evaluate the fit of a segmentation solution to the data. This section describes the basic principles of their single-sample, multi-sample and allele-specific segmentation algorithms.

Single sample segmentation (pcf) The simplest case is single-sample non-allele-specific segmentation. Assume we are given a sequence of logR values $\mathbf{y} = (y_1, \ldots, y_p)$. Then, the aim is to find a segmentation solution $S = \{I_1, \ldots, I_M\}$ that minimises the cost function

$$L(S|\mathbf{y}) = \sum_{I \in S} \sum_{j \in I} (y_j - \bar{y}_I)^2 + \gamma |S|, \tag{4.2}$$

where $\bar{y}_I$ is the average of the values y_j in segment I and where γ is a penalty parameter that controls the number of segments $|S|$.

Multi-sample segmentation (multipcf) In the multi sample case, we have logR measurements $\mathbf{y}_i$ for each of the samples $i = 1, \ldots, n$. The aim is to find a single segmentation that minimises the sum of the costs across all samples

$$L(S|\mathbf{y}_1, \ldots, \mathbf{y}_n) = \sum_{i=1}^{n} L(S|\mathbf{y}_i). \tag{4.3}$$

Allele-specific segmentation (aspcf) The allele-specific case, is essentially a multi-sample segmentation problem. Instead of segmenting logR tracks of different samples, we are given a logR, and a BAF track $\mathbf{y}_1$ and $\mathbf{y}_2$ of a single sample and the aim is to segment them jointly. Before this can be done, some pre-processing steps are necessary. Most importantly BAFs in $\mathbf{y}_2$ are mirrored by replacing y_j with $1 - y_j$ if $y_j > 0.5$ in order to obtain a single track in regions of allelic imbalance. Furthermore, we assume that logR and BAF values are paired, i.e. each BAF has a corresponding logR value. Then the aim is to find a segmentation that minimises the sum of the costs for the logR and the BAF data

$$L(S|\mathbf{y}_1, \mathbf{y}_2) = L(S|\mathbf{y}_1) + L(S|\mathbf{y}_2). \tag{4.4}$$

4.1.3 Extension to allele-specific multi-sample segmentation

As noted by Nilsen et al. [114], allele-specific segmentation and multi-sample segmentation are closely related. Therefore, to perform allele-specific multi-sample segmentation, we use BAF and logR tracks of all samples in one large multi-sample segmentation.

Asmultipcf largely uses the same pre-processing steps as the allele-specific single sample algorithm proposed by Nilsen et al. [114]. There is, however, one important difference. The aspcf algorithm removes sites with missing BAF values before the segmentation. In the multi-sample setting, this is problematic, as removing sites with incomplete data in just a single samples could lead to loss of a significant amount of data. To allow for missing values in the data matrix we extend the penalised least squares approach of Nilsen et al. [114] by using a weighted least squares function that models missing values in the data matrix.

Given the input data of n samples across p germline variant sites, the pre-processing yields a single matrix $\mathbf{Y} = (y_{ij}) \in \mathbb{R}^{2n \times p}$ that contains both logR and BAF values.

A cost function for weighted segmentation A weight matrix $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{2n \times p}$ is derived by assigning w_{ij} a weight of 0 if y_{ij} is missing and 1 otherwise. Then all missing values in $\mathbf{Y}$ are assigned an arbitrary numeric value. Our aim is to find a segmentation $S = \{I_1, \ldots, I_M\}$ that minimizes the least squares cost function

$$L(S|\mathbf{Y}, \mathbf{W}, \gamma) = \sum_{i=1}^{2n} L(S|\mathbf{y}_{i.}, \mathbf{w}_{i.}, \gamma) \tag{4.5}$$

$$= \sum_{i=1}^{2n} \sum_{I \in S} \sum_{j \in I} w_{ij}(y_{ij} - \bar{y}_{i,I})^2 + \gamma|S|, \tag{4.6}$$

where the best fit on a given segment I is the weighted average of the observations on that segment

$$\bar{y}_{i,I} = \frac{\sum_{j \in I} w_{ij} y_{ij}}{\sum_{j \in I} w_{ij}}.$$

Expanding the square in (4.6) we find

$$L(S|\mathbf{Y},\mathbf{W},\gamma) = \sum_{i=1}^{2n}\sum_{I\in S}\sum_{j\in I} w_{ij}\left(y_{ij}^2 - 2y_{ij}\bar{y}_{i,I} + \bar{y}_{i,I}{}^2\right) + \gamma|S| \tag{4.7}$$

$$= \sum_{i=1}^{2n}\sum_{I\in S}\sum_{j\in I} w_{ij}\left(y_{ij}^2 - 2y_{ij}\frac{\sum_{j\in I} w_{ij}y_{ij}}{\sum_{j\in I} w_{ij}} + \frac{\left(\sum_{j\in I} w_{ij}y_{ij}\right)^2}{\left(\sum_{j\in I} w_{ij}\right)^2}\right) + \gamma|S| \tag{4.8}$$

$$= \sum_{i=1}^{2n}\sum_{I\in S}\left((\sum_{j\in I} w_{ij}y_{ij}^2) - 2\frac{\left(\sum_{j\in I} w_{ij}y_{ij}\right)^2}{\sum_{j\in I} w_{ij}} + \frac{\left(\sum_{j\in I} w_{ij}y_{ij}\right)^2}{\sum_{j\in I} w_{ij}}\right) + \gamma|S| \tag{4.9}$$

$$= \sum_{i=1}^{2n}\sum_{I\in S}\left((\sum_{j\in I} w_{ij}y_{ij}^2) - \frac{\left(\sum_{j\in I} w_{ij}y_{ij}\right)^2}{\sum_{j\in I} w_{ij}}\right) + \gamma|S|. \tag{4.10}$$

To find an optimal segmentation, we can omit the term that is independent of the segmentation and instead minimise the cost function

$$L'(S|\mathbf{Y},\mathbf{W},\gamma) = -\sum_{i=1}^{2n}\sum_{I\in S}\frac{\left(\sum_{j\in I} w_{ij}y_{ij}\right)^2}{\sum_{j\in I} w_{ij}} + \gamma|S|. \tag{4.11}$$

An exact algorithm for weighted segmentation For a large number of loci p, naive optimisation is not feasible. Nilsen et al. [114], however, noted that the global segmentation problem can be broken down into smaller subproblems and used this to develop a dynamic programming algorithm. The idea behind their algorithm is that, given a breakpoint, the optimal segmentation solutions on either side of it are independent of each other. In the following, we adapt their dynamic programming algorithm to our weighted problem.

Algorithm 1: asmultipcf

Input: Matrix $\mathbf{Y}$ of log-transformed copy numbers and B allele frequencies (BAF); weight matrix $\mathbf{W}$; penalty $\gamma > 0$;
Output: Segment start indices and matrix of segment averages $\overline{\mathbf{Y}}$;
Notation: In the following, $\mathbf{A}_k$ and $\mathbf{C}_k$ denote matrices of size $2n \times k$ that are used to store the sum in the numerator and denominator of the cost function (Equation 4.11), respectively, for k different segments I, where column i corresponds to the segment of length $k-i+1$ ranging from position $k-(k-i+1)$ to k. $\mathbf{d}_k$ is a vector of length k and stores the first term of the cost function for the k segments. $\mathbf{e}_k$ is a vector of length $k+1$ and stores the overall cost of different segmentation solutions. Vectors that are obtain by selecting column k of a matrix $\mathbf{M}$ are denoted as $\mathbf{m}_{\cdot k}$.

1. Initialise $\mathbf{A}_k$ and $\mathbf{C}_k$ as empty matrices ($\mathbf{A}_0 = [\,]$, $\mathbf{C}_0 = [\,]$). The initial value for the overall cost is zero ($\mathbf{e}_0 = 0$).

2. Iterate for $k = 1, \ldots, p$

   - $\mathbf{A}_k = [\mathbf{A}_{k-1}\ 0] + \mathbf{w}_{.k}\mathbf{y}_{.k}$
   - $\mathbf{C}_k = [\mathbf{C}_{k-1}\ 0] + \mathbf{w}_{.k}$
   - $\mathbf{d}_k = -\mathbf{1}^T \left(\mathbf{A}_k \circ \mathbf{A}_k \circ \mathbf{C}_k^{\circ -1}\right)$ where $\circ$ denotes an element-wise matrix product and $\mathbf{C}_k^{\circ -1}$ the element-wise inverse
   - $\mathbf{e}_k = [\mathbf{e}_{k-1}\ \min(\mathbf{d}_k + \mathbf{e}_{k-1} + \gamma)]$

   storing also the index $t_k \in 1, \ldots, k$ at which the minimum in the last step is achieved.

3. Find segment start indices from right to left as $s_1 = t_p$, $s_2 = t_{s_1 - 1}$, ..., $s_M = 1$, where $M \geq 1$.

4. Find segment averages

$$\overline{\mathbf{y}}_{.m} = \frac{\left(\mathbf{w}_{.s_m}\mathbf{y}_{.s_m} + \cdots + \mathbf{w}_{.s_{m-1}-1}\mathbf{y}_{.s_{m-1}-1}\right)}{\left(\mathbf{w}_{.s_m} + \cdots + \mathbf{w}_{.s_{m-1}-1}\right)}$$

   for $m = 1, \ldots, M$, where $s_0 = p + 1$.

A heuristic algorithm for large data sets Algorithm 1 is of order $O(np^2)$, where p is the number of input loci and n is the number of samples. This means that the segmentation becomes computationally expensive for long sequences. However, instead of allowing breakpoints at any of the p positions, Nilsen et al. [114] noted that we can pre-select potential breakpoints and thereby reduce the runtime to $O(nq^2)$ where q is the number of potential breakpoints. To identify potential breakpoints, different heuristics can be used. Here, we apply Algorithm 1 to overlapping subsequences, combine all of the inferred breakpoints and use them as input for the subsequent global segmentation. As in the implementation by Nilsen et al. [114] we use subsequences of length 5000 with an overlap of 1000. Algorithm 2 describes the fast heuristic version of asmultipcf.

Algorithm 2: Fast asmultipcf

Input: Matrix $\mathbf{Y}$ of log-transformed copy numbers and B allele frequencies; Weight matrix $\mathbf{W}$; penalty $\gamma > 0$;
Output: Segment start indices and segment averages

1. Split data set into overlapping subsequences and apply steps 1 and 2 of Algorithm 1 to each of them in order to find potential breakpoints $r_0, r_1, \ldots, r_q$ where $r_0 = 1$ and $r_q = p + 1$.
2. Aggregate sequences between breakpoints by setting $x_{ik} = \sum_{j=r_{k-1}}^{r_k-1} w_{ij} y_{ij}$ and $v_{ik} = \sum_{j=r_{k-1}}^{r_k-1} w_{ij}$.
3. Calculate segmentation solution by using the aggregated matrices $\mathbf{X}$ and $\mathbf{V} \in \mathbb{R}^{2n \times q}$ as input to Algorithm 1 instead of $\mathbf{Y}$ and $\mathbf{W}$, respectively.

Post-processing Both algorithms yield a single segmentation solution S for all samples. However, we expect that only some of the segments will be shared between all samples while others will be private. While ASCAT can be run directly on the global segmentation solution, removing unnecessary breakpoints on a per sample base can reduce noise in the segment average estimates by generating larger segments. To refine breakpoints individually for each sample, we simply use the breakpoints inferred from the multi-sample segmentation and rerun steps 2 and 3 of Algorithm 2 on each sample individually based on these potential breakpoints.

Implementation Asmultipcf is part of the ASCAT R package from version 2.5 onwards [163]. The asmultipcf function contains a parameter to select whether the exact or the fast algorithm should be run, as well as an option to include the per-sample breakpoint refinement. Furthermore, samples can be weight adjusted to account for quality differences in the data.

4.1.4 Segmentation case studies

This section compares tree inference results obtained from single- versus multi-sample segmentation data using two case studies.

Case 298 This case is relatively simple with few copy number changes and only eight metastasis samples. Using the default values for multi-sample segmentation, asmultipcf inferred an average of 44 segments per sample. For single-sample segmentation, the penalty parameter was set to 80 to achieve a similar segmentation resolution (average of 44.1 segments per sample). The subsequent steps including the ASCAT and MEDICC analysis were performed as described in section 4.2.2. The resulting trees are shown in Figure 4.1. The branches leading to the leaf nodes are much longer in the single-sample segmentation

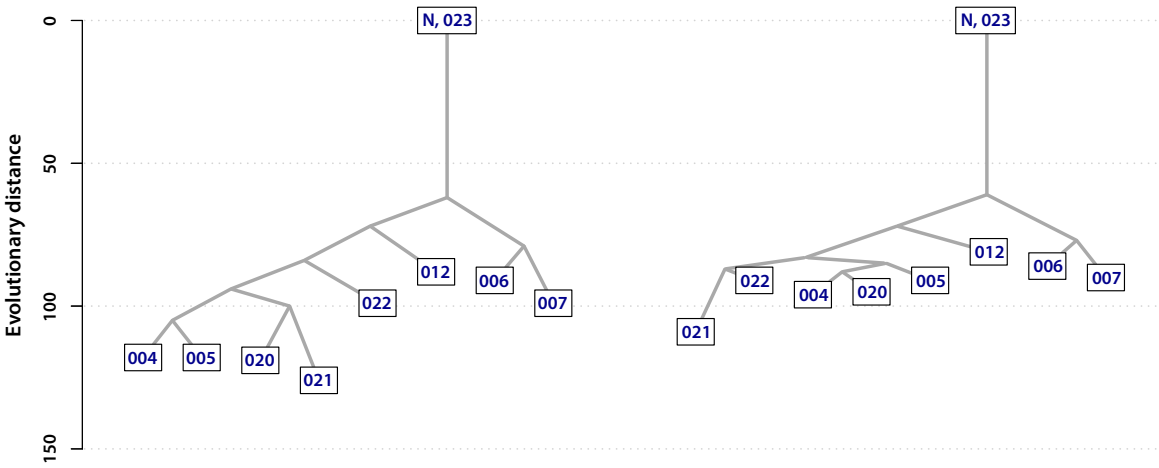


Fig. 4.1 Comparison of MEDICC trees inferred for case 298 using single-sample segmentation (left) and multi-sample segmentation (right).

tree than the multi-sample segmentation tree. There are two alternative explanations for the difference in branch lengths: (i) under-segmentation of the multi-sample data, which could make samples look more similar to each other than they actually are, leading to shorter branches or (ii) misalignment of segment boundaries in the single-sample data, which forces MEDICC to infer more copy number changes than there actually are, leading to longer branches. The fact that the number of segments in the input data for both trees is comparable suggests that the longer branches are not due to copy number changes that are only detected by the single-sample segmentation.

In the tree inferred from individually segmented samples, all branches connecting internal nodes have a similar evolutionary distance. In comparison, the tree inferred from jointly segmented samples shows a clearer divide between sample 012 and the group of closely related samples 004, 005, 020, 021, 022. This divide is also inferred by other methods such as OncoNEM (see Section 4.2.3.2), which shows that multi-sample segmentation may help to infer a more accurate tree structure. It also suggests that the long leaf branches in the single-sample segmentation tree are due to reason (ii), misaligned segment boundaries. However, this cannot be proven definitively without knowledge of the ground truth.

Case 308 The previous case shows that given few samples and a small number of copy number changes, the overall tree structure may be recovered even with single-sample segmentation, but the inference of finer details of the tree like branch lengths is less reliable. However, as the number of samples increases the number of misaligned segment boundaries increases, too. An example for this is case 308, for which 19 samples were available. Multi-sample

segmentation was run with default parameters. For the single-sample segmentation, the penalty parameter was set to the maximum of 800 to reduce the number of noisy fragments.

Figure 4.2 shows the segmentation of chromosome 19 for this case. Instead of the 3 breakpoints inferred by multi-sample segmentation, single-sample segmentation of all samples yields a total of 11 breakpoints. Overall single-sample segmentation infers five clearly distinct segments, whereas multi-sample segmentation infers four, omitting a separate segment at the start of chromosome 19. While this could indicate a lower sensitivity of the multi-sample segmentation, the set of samples for which this segment was inferred does not match with the tree structure inferred by the other methods (see Figure A.5). This suggests that the extra segment inferred by single-sample segmentation is an artefact. It could, however, also be a clonal mutation that was missed in a large number of samples.

Running MEDICC on the single-sample segmentation data failed due to computational constraints, as a result of the large number of segments. To reduce the size of the dataset, we removed all segments spanning less than 500 000 bp, as done by Mangiola et al. [96]. This yields an average of 38.5 segments per sample, while multi-sample segmentation produced an average of 37 segments per sample. Despite the size filtering, the total number of segments of the single-sample data was much higher than that of the multi-sample data (104 versus 53). MEDICC inference on a 320GB node took 51.5 hours for the single-sample data versus less than 2 hours for the multi-sample data.

The resulting trees are shown in Figure 4.3. They differ in the placement of samples (018 and 022), which, according to LICHeE, contain a mix of the two major subpopulations (see Figure A.5). As with case 298, the evolutionary distances between samples that are inferred to be closely related by alternative methods are smaller in the multi-sample segmentation tree than in the single-sample segmentation tree. Furthermore, finer details of the relationship structures are more similar between the multi-sample segmentation tree and the trees inferred by alternative methods. For example, in the multi-sample segmentation tree, samples 005 and 014 are placed close to the cluster containing 006, 007, 008, 009 and 010, as inferred by OncoNEM, Treeomics and LICHeE (see Figure A.5).

4.1.5 Discussion

Our case studies suggest that the independent segmentation of related samples can artificially inflate estimates of tumour heterogeneity. We developed an algorithm for the joint allele-specific segmentation of multiple samples, which addresses this problem. A limitation of multi-sample segmentation is that it can potentially underestimate tumour heterogeneity because CNAs that are shared by many samples are more likely to be detected than CNAs that are shared by fewer samples or private. Nevertheless, even though the total number of

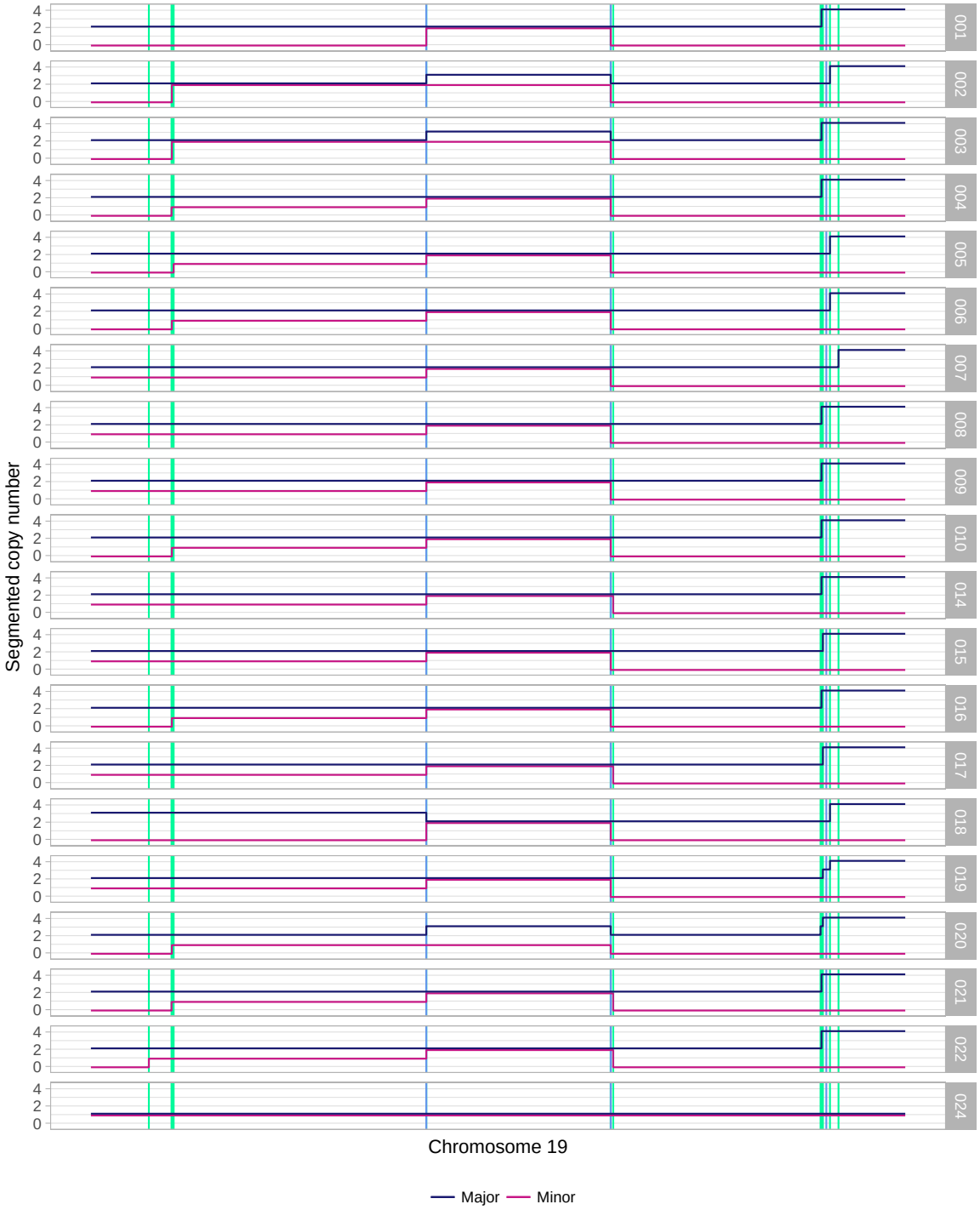


Fig. 4.2 Copy number inference of chromosome 19 for case 308. The blue and magenta lines show the major and minor integer copy number values inferred with ASCAT by single-sample segmentation. These lines are plotted with a slight vertical offset to show both lines in regions of balanced allele frequencies. The vertical green and blue lines indicate the breakpoints inferred with single-sample and multi-sample segmentation, respectively.

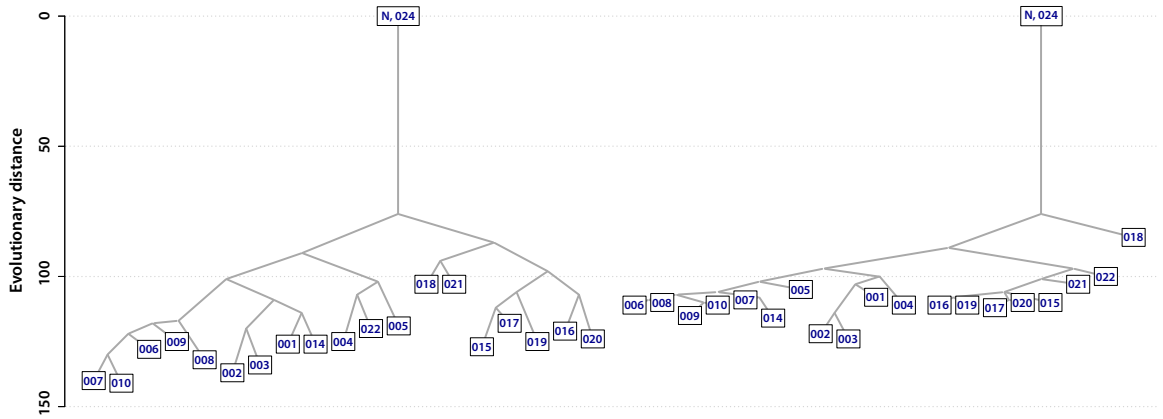


Fig. 4.3 Comparison of MEDICC trees inferred for case 308 using single-sample segmentation (left) and multi-sample segmentation (right).

segments inferred by multi-sample segmentation was smaller in our test cases, the resulting trees were more similar to trees inferred by other methods.

To avoid under-segmentation of the multi-sample method, the penalty parameter γ that controls the number of inferred segments could be further optimised. As always, there is a trade-off between sensitivity and specificity. If the segmented data is to be analysed with downstream tools such as MEDICC, computational constraints of these methods also need to be taken into account. Multi-sample segmentation can substantially reduce the computation time of MEDICC.

Overall, our case studies show that allele-specific multi-sample segmentation can improve the copy number-based phylogenetic analysis of multiple related samples.

4.2 Phylogenetic analysis of metastatic breast cancers

So far, this chapter focused on a new method for multi-sample segmentation. The motivation for developing this algorithm was a phylogenetic study, which analysed a data set of metastatic breast cancers using various multi-sample inference algorithms and which is described in the remainder of this chapter.

4.2.1 Data

Samples The following study contains samples from ten breast cancer patients who died with metastatic disease that had become resistant to multiple lines of therapy. The samples analysed comprise metastatic samples of each patient (average: 10.6 per patient, range: 3 to

19), which were collected during autopsies, as well as primary tumour samples from surgery for six out of the ten patients. For some of the patients, cell-free DNA from plasma samples and other body fluids that had been collected at varying time points was also available (average: 4.7 per patient, range: 1 to 9). For all patients, blood samples were used as matched normal.

Sequencing Shallow whole genome sequencing (sWGS) data was available for the primary tumour samples, the metastasis samples and most of the fluid samples. Furthermore, a subset of metastases (average: 7.6 per patient, range: 1 to 19) had been analysed using whole exome sequencing (WES). All samples had been subjected to deep targeted sequencing (TS) using a total of 499 amplicons, which had been designed based on somatic mutations discovered in the WES data (mean 49.9 amplicons/patient, range 13-189 amplicons, mean sequencing depth of 7570 to 29891x). Since not all data types are available for all samples, the samples used for tree inference vary depending on the input required by each method.

Variant calling Data pre-processing was performed by Oscar Rueda and Stephen-John Sammut (Cancer Research Cambridge Institute). In particular, germline SNVs were called from WES data of matched blood samples. Somatic SNVs were called from WES data of all other samples and were rigorously filtered to exclude germline variants and other false positives.

4.2.2 Methods

In the following, we describe the phylogenetic analysis using a set of five different phylogenetic methods (see Chapter 2 for a more detailed description of the methods).

Method selection We decided to use three sample tree methods, MEDICC, Treeomics and OncoNEM, and two deconvolution methods, PyClone in combination with LICHeE as well as SuperFreq. MEDICC is the only method that performs a comprehensive phylogenetic analysis of copy number profiles. Treeomics is a point mutation based method that has specifically been developed for the analysis of metastases and performs limited subclone detection. While OncoNEM has been developed for the analysis of point mutations from single cells, it can also be applied to multi-sample data, assuming that mixing of subpopulations is limited. In contrast to Treeomics, it does not detect subpopulations but can be applied to large datasets.

PyClone is a standard method for clustering mutation frequencies but does not reconstruct phylogenies. Therefore, we use LICHeE to infer trees from the PyClone clusters. SuperFreq

Table 4.1 Overview of phylogenetic methods used for the analysis of metastatic breast cancers including the sequencing data that was used to infer the tree, the type of variants used as phylogenetic markers, the input data required, the type of tree inferred and whether the method performs clonal deconvolution.

Method	Data	Variants	Input	Tree type	Clonal deconvolution
MEDICC	sWGS + WES	CNA	Allele-specific copy number profiles	sample tree	no
OncoNEM	WES / TSeq	SNV	Binary mutation profiles	sample tree	no
Treeomics	WES / TSeq	SNV	Total read depth + read depth of variant allele at every SNV location across all samples	sample tree	limited
SuperFreq	WES	SNV and CNA	Aligned reads (bam files) + preliminary variant calls (vcf files)	clone tree	yes
PyClone + LICHeE	WES / TSeq + sWGS	SNV	Variant allele frequencies, copy number profiles + purity estimates	clone tree	yes

is one of the few methods that use both point mutations and copy number changes as phylogenetic markers. It performs clustering and tree inference.

All methods were applied to the WES data, while OncoNEM, Treeomics and PyClone/LICHeE were also applied to the TS data. An overview of the methods used is shown in table 4.1.

MEDICC A combination of shallow whole-genome sequencing and exome sequencing data was used to infer the allele-specific copy number profiles used for the MEDICC analysis. The reason for this is that the higher coverage of WES data is needed to estimate BAFs, while sWGS is expected to provide better estimates of logRs than the exome sequencing data due to smaller biases in the sequencing depth.

The use of sWGS and WES is not ideal, as they both provide data of limited resolution. The read counts of sWGS data need to be binned to estimate log ratios, while the BAFs can only be estimated for a small part of the genome. For this reason, the inferred allele-specific copy number profiles are expected to be of relatively low resolution. LogR values were estimated from sWGS using QDNAseq 1.2.4 [147, 146] by Oscar Rueda (Cancer Research UK Cambridge Institute). The QDNAseq estimates are based on bins of size 100kb. In addition to estimating binned logR values, QDNAseq was also used to adjust estimates for sequence mappability and GC content of the DNA to account for biases in the read depth. BAFs were calculated from WES data at sites of previously inferred germline SNVs using alleleCount 3.1.1 [15].

To make logR and BAF data compatible for segmentation, each BAF was assigned a binned logR based on the location of the corresponding germline SNV using a custom R script. Log ratios and B-allele frequencies were segmented on a per case basis using the allele-

specific multi-sample segmentation algorithm described in Section 4.1 with default penalty parameter. Using the segmented BAF and logR values, allele-specific copy number profiles were inferred with ASCAT 2.5 [163]. The raw ASCAT copy number profiles were visually compared across samples for each case, and ASCAT was rerun with manually adjusted ploidy and purity estimates where necessary, to obtain the final discrete copy number profiles. The reason for this is that diploid and tetraploid solutions often fit almost equally well and copy number callers like ASCAT and ABSOLUTE struggle to infer the true underlying ploidy. An example of this is shown in Figure 4.4. To date, this is an unsolved problem in the inference of copy numbers and to obtain high-quality copy number profiles manual curation is needed. We selected a single baseline ploidy, diploid or tetraploid, for each case by comparing samples within a case and while aiming to achieve a fit where the raw, unrounded copy number values lie close to integer values.

MEDICC assumes that a region that has been affected by a homozygous deletion cannot show an increase in copy number later in the phylogeny. This makes sense from a biological perspective. However, given that MEDICC does not account for noise in the data, spurious homozygous deletions can have a big impact on the inferred phylogeny. For this reason, regions that show a homozygous deletion in any sample were removed from all samples of that case. Furthermore, regions with constant integer copy number values across all samples were compressed to single integer values to reduce the size of the input data. To comply with MEDICC requirements, a maximum copy number cut-off of nine was applied to both major and minor copy number profiles, replacing any values exceeding this threshold. Finally, MEDICC (devel branch, commit da7ed4a) [149] with ancestor reconstruction switched on was used to infer the phylogenies. Trees were plotted using a custom R script.

Treeomics Treeomics was used to infer phylogenies from both WES and targeted sequencing data. For each case, Treeomics was used to calculate the posterior probabilities of a variant being present based on total read depth and the number of reads covering the alternative allele. All sites where the posterior probabilities were lower than 0.5 in all samples were removed, as these are likely to be false positives.

Treeomics offers a variety of inference options. In particular, it has an option to test for subclones as part of the tree inference as well as different options that allow reducing the memory requirements for the inference. A disadvantage of the low memory options is that the detection of the optimal solution is no longer guaranteed and it is not recommended to use this in combination with subclone detection (Johannes Reiter, personal communication). For our analysis, subclone detection was a priority. Therefore, we decided to reduce the number of samples for some of the cases, to make the Treeomics analysis with subclone

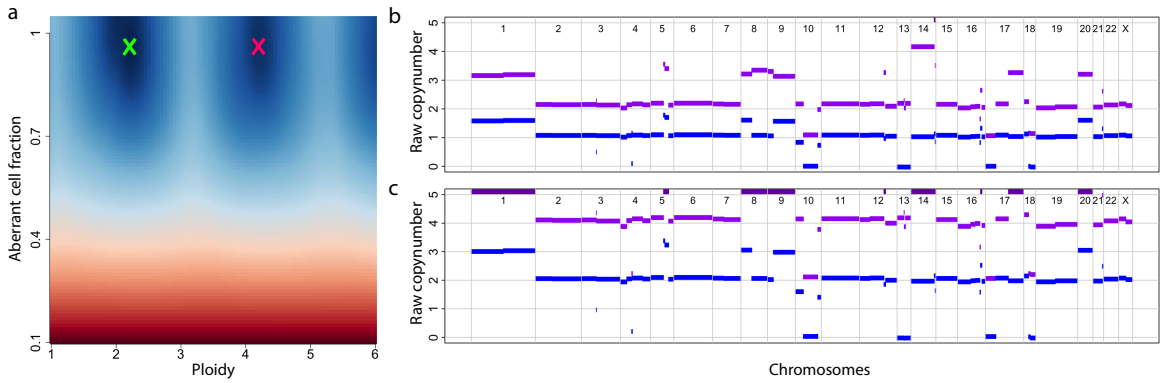


Fig. 4.4 Copy number estimation for sample 004 of case 298. a – ASCAT sunrise plot. This plot shows the goodness of fit for varying ploidy and purity parameters. The green cross marks the parameters estimated by ASCAT, the red cross marks the manually selected parameter values. The corresponding raw copy number profiles are shown in panels b and c, respectively. b – Despite the high goodness of fit for the ploidy parameter of 2.2 estimated by ASCAT, the raw copy number values of chromosome 1, 9 as well as large regions of chromosome 8 are far away from integer values. c – Given a ploidy parameter of 4.2, the raw values of most regions of the genome are close to integer values. For 4 out of the 8 samples of case 298, ASCAT automatically estimated a ploidy of around 4 (data not shown).

detection computationally feasible on a 320GB cluster node. To keep the mutation profiles as diverse as possible and to maintain a good representation of the different tumour populations, samples that had a mutation profile similar to one of the remaining samples were excluded preferentially. Finally, Treeomics 1.7.3 [131] was applied to each case with subclone detection switched on and all other parameters set to default. Trees were plotted using a custom R script.

OncoNEM Like Treeomics, OncoNEM was used to infer phylogenies from both WES and targeted sequencing data. Binary mutation profiles were obtained from the filtered Treeomics posterior probability matrices by setting all entries with a mutation probability smaller than 0.5 to 0 and to 1 otherwise. Alternatively, standard variant callers could have been used to infer binary mutation profiles. However, using the output from Treeomics helps ensure that differences in the results between the OncoNEM and Treeomics are due to methodological reasons instead of differences in the input data. The OncoNEM analysis was performed using error parameter optimisation over a parameter range from 0.0001 to 0.1. The Bayes factor threshold epsilon was set to 2 in order to avoid overly strict clustering.

SuperFreq SuperFreq was used to infer clone trees and the corresponding sample compositions from WES data. As input, SuperFreq requires preliminary liberal variant calls which are then filtered as part of the SuperFreq pipeline. Pileup files of the WES data were generated using samtools 1.3.1 mpileup [49] using a maximum depth threshold of 10000, a minimum mapping quality of 1 and a minimum base quality of 15. Variant calling was performed using VarScan 2.4.3 mpileup2cns [78] with a p-value filter of 0.01, no strand-bias filter and the variant flag set to only obtain variant sites as recommended in the SuperFreq manual. Finally, SuperFreq 0.9.17 [40] was run with default parameters using the normal of all cases apart from DET52 as reference normals for each case. SuperFreq does not generate figures or data files of the inferred tree structure, but only visualises the subset relationships and sample compositions in a fish plot. Tree structures were therefore extracted manually and plotted using custom scripts.

PyClone and LICHeE PyClone and LICHeE were used to infer clone trees and the corresponding sample compositions from both WES and targeted sequencing data.

LICHeE offers different options for the tree inference. Either it can infer trees directly from VAFs of SNVs, in which case it performs clustering of VAFs before inferring the clone tree, or it can be applied to previously inferred mutation clusters in which case it only performs the tree inference step. We decided to use the second option because, in contrast to other deconvolution methods, LICHeE does not offer the possibility to adjust variant allele frequencies for copy number aberration related biases.

Therefore, the first step of the analysis was to infer mutation clusters and their cellular prevalences for each sample using PyClone 0.13.0 [138]. PyClone was run using a beta-binomial density. A minimum cluster size of 3 was selected for WES data, while no minimum cluster size was set for TS data, due to the smaller number of mutations. The input parental copy number and tumour content (purity) estimates were inferred with ASCAT 2.4 using allele-specific single-sample segmentation and default parameters. The start and end points of copy number segments are defined by the location of heterozygous germline sites. Variants that fall between segments have an unknown copy number status and were excluded from the analysis. PyClone was run for 40000 iterations using a burn-in sample of 20000. This analysis was performed by Stephen-John Sammut (Cancer Research UK Cambridge Institute).

To remove spurious clusters from the PyClone output, two filtering steps were performed. First, low prevalence clusters that did not exceed a cellular prevalence of 0.1 in the WES data in any of the samples were removed, as they tend to attract noisy mutations of low frequency. This step was skipped for the targeted sequencing data, where VAF estimates are more precise. Second, if multiple clusters were present in all samples, all but the cluster

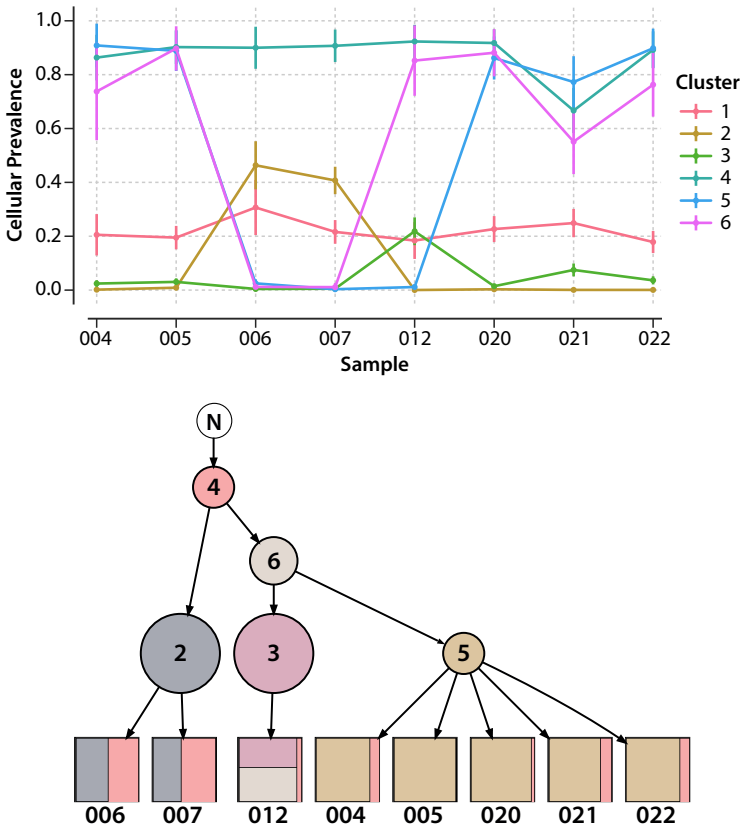


Fig. 4.5 Example of PyClone and LICHeE results for case 298. Top – PyClone clustering result. There are two clusters which are present in all samples, cluster 1 and cluster 4. The sum of the cellular prevalences of these two clusters exceeds 1 for most samples and the cellular prevalence of cluster 1 is smaller than the cellular prevalence of cluster 4. According to the sum-rule (see Chapter 2) this means that these two clusters describe a linear phylogeny where the subpopulation carrying the mutations of cluster 1 is a descendant of the subpopulation carrying the mutations of cluster 4. The same applies to the clusters 4 and 6. However, according to the crossing rule (see Chapter 2), cluster 1 and cluster 6 describe a branching phylogeny, because the cellular prevalence of cluster 1 is sometimes smaller and sometimes larger than the cellular prevalence of cluster 6. At the same time, the sum of the cellular prevalences of cluster 1 and 6 is higher than the cellular prevalence of cluster 4. This shows that the three clusters 1, 4 and 6 are not compatible. For this reason, cluster 1 was excluded from the analysis. Clusters 4 and 5 are also incompatible: Sample 020 suggests that they form a linear phylogeny because the sum of the cellular prevalence is larger than 1, but in sample 021 the cellular prevalence estimate of cluster 5 is higher than that of cluster 4. However, the error bars indicate a large uncertainty in the cellular prevalence estimates. LICHeE takes cellular prevalence errors into account through a global parameter that needs to be specified by the user. Bottom – Given a cellular prevalence estimate error of 0.3, LICHeE infers a phylogeny where a subpopulation carrying cluster 4 occurred first. It gave rise to a subpopulation carrying cluster 6 which in turn gave rise to a subpopulation carrying cluster 5. Furthermore, samples 012 shows a private mutation cluster (3) and samples 006 and 007 belong to a subpopulation defined by cluster 2, which branched off from the clonal cluster 4.

with the highest cellular prevalence were removed. In many cases, these ubiquitous medium prevalence clusters are incompatible with most of the other clusters, which suggests that they are spurious clusters (see Figure 4.5). A possible explanation for these clusters are errors in the copy number based adjustment of variant allele frequencies. While LICHeE can handle incompatible clusters by using a user-specified error range for the cellular prevalence estimates, keeping too many incompatible clusters in the input data yields a solution that ignores a large portion of the inferred clusters.

In addition to cellular prevalences for each mutation cluster and sample, LICHeE requires a binary profile for each cluster that indicates if the cluster is classified as present in a given sample. We decided to infer these binary profiles based on the VAFs of the mutation set rather than the cellular prevalences, as the copy number adjustment needed to obtain cellular prevalences is likely to add noise to the data. In particular, the cellular prevalence estimates of low-frequency clusters are expected to be less accurate than those of high-frequency clusters, because the copy number profiles used for the adjustment are dominated by the major population in the sample.

For the WES data, a mutation was classified as present in a given sample if its VAF was larger than 0.01. For the TS data, we used the VAFs of 8 control samples that had been sequenced for each case to select the variant calling threshold in a site-specific manner. More precisely, a mutation was called present in a sample if its VAF was at least three standard deviations higher than the mean VAF of the control samples at the corresponding site. For both WES and TS data, a mutation cluster was called present if at least 40% of its mutation were called present in a sample.

Finally, LICHeE (commit 238770c) [125] was used to infer clone trees and sample compositions, assuming a cellular prevalence estimate error of 0.3.

4.2.3 Results

In this section, we discuss the tree inference results. Comparing the trees inferred by the different methods in a quantitative manner is challenging for several reasons.

Trees vary in their structure. Some methods infer sample trees, whereas others infer clone trees and the corresponding sample compositions. In case of clone trees and assuming that mixing of subpopulations is detected, the resulting graph is not a tree, but a directed acyclic graph, where nodes that correspond to samples can have multiple parents.

Building blocks vary between trees. The sample composition of each tree depends on various factors. First, the required data type needs to be available for a given method.

Second, the data of a sample has to be of sufficient quality to be analysed, which can also be method specific. Third, some methods have computational constraints regarding the number of samples they can analyse, requiring the exclusion of samples.

Evolutionary markers vary between trees. MEDICC uses copy number changes, whereas SuperFreq uses a combination of CNAs and SNVs. Treeomics, OncoNEM and LICHeE analyse only SNVs. Even if methods analyse the same type of variants, the composition of the variants analysed can vary. For MEDICC we used copy number profiles inferred with ASCAT from sWGS and WES data, whereas SuperFreq uses its own pipeline to call copy number aberrations purely from WES data, yielding different sets of CNAs. Furthermore, SuperFreq filters out most of the SNVs that are in regions affected by copy number changes and also removes mutations that are only detected in a single sample. For this reason, the SuperFreq inference is based only on a fraction of the SNVs used by other SNV-based methods. Similarly, many SNVs are removed during the PyClone and LICHeE analysis.

For these reasons, robust quantitative comparison of the different methods is challenging. In the following, we restrict ourselves to a mainly qualitative comparison. We start by evaluating distances between trees to assess the similarity of trees inferred by different methods across cases. Then we discuss two of the ten cases in detail and finally summarise the main observed advantages and disadvantages of the different methods.

4.2.3.1 Comparison of tree similarities across cases

To assess the similarity of the inferred trees across all cases, we calculated pairwise distances between the WES trees of all methods. To make the results of different methods comparable, we converted the directed acyclic graphs inferred by LICHeE and SuperFreq into trees by assigning each sample containing a mix of subpopulations to the branch of its major subpopulation. We then removed samples that were not analysed by all methods from the tree and normalised the total branch length of the tree to one. Next, we calculated the pairwise-sample shortest path distance for each pair of methods and each case as described in Chapter 3. To account for the fact that the number of samples varies across cases, we normalised the distances by the number of sample pairs. Finally, we averaged the distances across cases. The resulting distance matrix is shown in Figure 4.6.

Trees inferred by the sample tree methods Treeomics, OncoNEM and MEDICC are more similar to each other than to the results of deconvolution methods. This is not surprising because the deconvolution methods often group several samples into a single subpopulation,

which yields a pairwise-sample distance of zero, while the sample tree methods assign most samples to separate branches. More surprisingly, the agreement between the two deconvolution methods SuperFreq and LICHeE is even lower than between either of them and the sample tree methods.

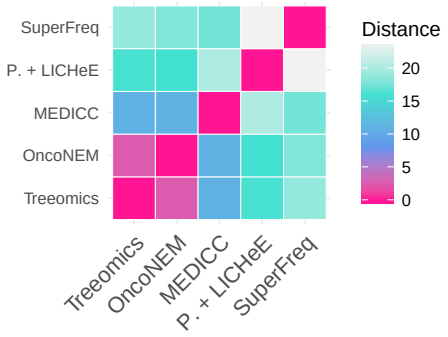


Fig. 4.6 Heatmap of normalised pairwise-sample shortest path distances averaged across all cases.

4.2.3.2 Case 298

Trees inferred from WES and sWGS data are shown in Figure 4.7, whereas trees inferred from TS data are shown in Figure 4.9. In the figures for MEDICC, OncoNEM and Treeomics, distances in the direction of the y-axis correspond to the evolutionary distance or the number of accumulated mutations. In the figures for SuperFreq and LICHeE, the numbers in the nodes of the trees correspond to the number of mutations in the clusters. Samples in these figures are coloured to reflect the subpopulations they contain. The size of the coloured areas roughly corresponds to the cellular prevalence estimates of the subpopulations.

Tree structure Based on WES and sWGS data, all methods infer a clear branching evolution, with the meninges metastases (006, 007) on one main branch and the lymph node (004, 005), liver (020, 021, 022) and lung (012) metastases on the other main branch (see Figure 4.7). Within the second main branch, OncoNEM, Treeomics, MEDICC and LICHeE infer an early divergence that gave rise to the lung metastasis, whereas SuperFreq does not identify any private mutations for sample 012. The inferred relationships between the lymph node and liver metastases differ between methods. OncoNEM, Treeomics and MEDICC

all infer different albeit close relationships, whereas LICHeE concludes that they all belong to the same subpopulation. SuperFreq infers that 004, 005, 020, 021 belong to a single subpopulation whereas 022 belongs to a descendant of that subpopulation.

The TS based inference confirms these results (see Figure 4.9). However, using this data, LICHeE cannot identify mutations private to 012 due to variant filtering (see Figure 4.10). OncoNEM and Treeomics, yield identical results and infer that 020 and 021 belong to an ancestral population of 004, 005 and 022. The WES based tree that agrees best with this result is the one inferred by Treeomics.

Mixing of subpopulation Out of the tested methods LICHeE, SuperFreq and Treeomics are the only ones designed to detect mixing of subpopulations within samples. SuperFreq infers that a small fraction of the cells in sample 012 belong to the same subpopulation as samples 006 and 007. While LICHeE does not detect mixing in the WES data, its TS data findings support the SuperFreq result. Treeomics does not detect mixed subpopulations in either of the datasets.

Two different scenarios can explain the mixing of subpopulations in 012. First, it could be due to cross-seeding between metastases. Alternatively, 012 may have been seeded by an ancestral stem population, which evolved into two independent subpopulations within the metastasis and then gave rise to the other metastases. As the time at which the different metastases occurred is unknown, none of the two possibilities can be selected with certainty.

Both SuperFreq and LICHeE infer that multiple samples contain cells from the stem population. However, as described in the methods section, the inference of cellular prevalences can be unreliable. This is also indicated by the fact that cellular prevalence estimates vary strongly between data types. For samples 006 and 007, for example, LICHeE infers a cellular prevalence of the stem population of 0.44 and 0.5 based on the WES data versus 0 and 0.05 based on the TS data.

Mutation assignment To assess the branch length estimation, we measured the branch lengths of the core tree, i.e. branches connecting the root of the tree with the most recent common ancestor of each of the three main subpopulations. We then calculated pairwise Pearson correlation coefficients between those branch lengths of all WES based trees (see Figure 4.11). The highest correlation was found between the OncoNEM and Treeomics estimates (0.98). This is expected because these methods are the most similar ones. Even though MEDICC analyses copy number changes, the branch length estimates showed a high correlation with OncoNEM and Treeomics (0.85 and 0.87). SuperFreq's branch length estimates were weakly correlated with those of OncoNEM (0.46), Treeomics (0.40) and

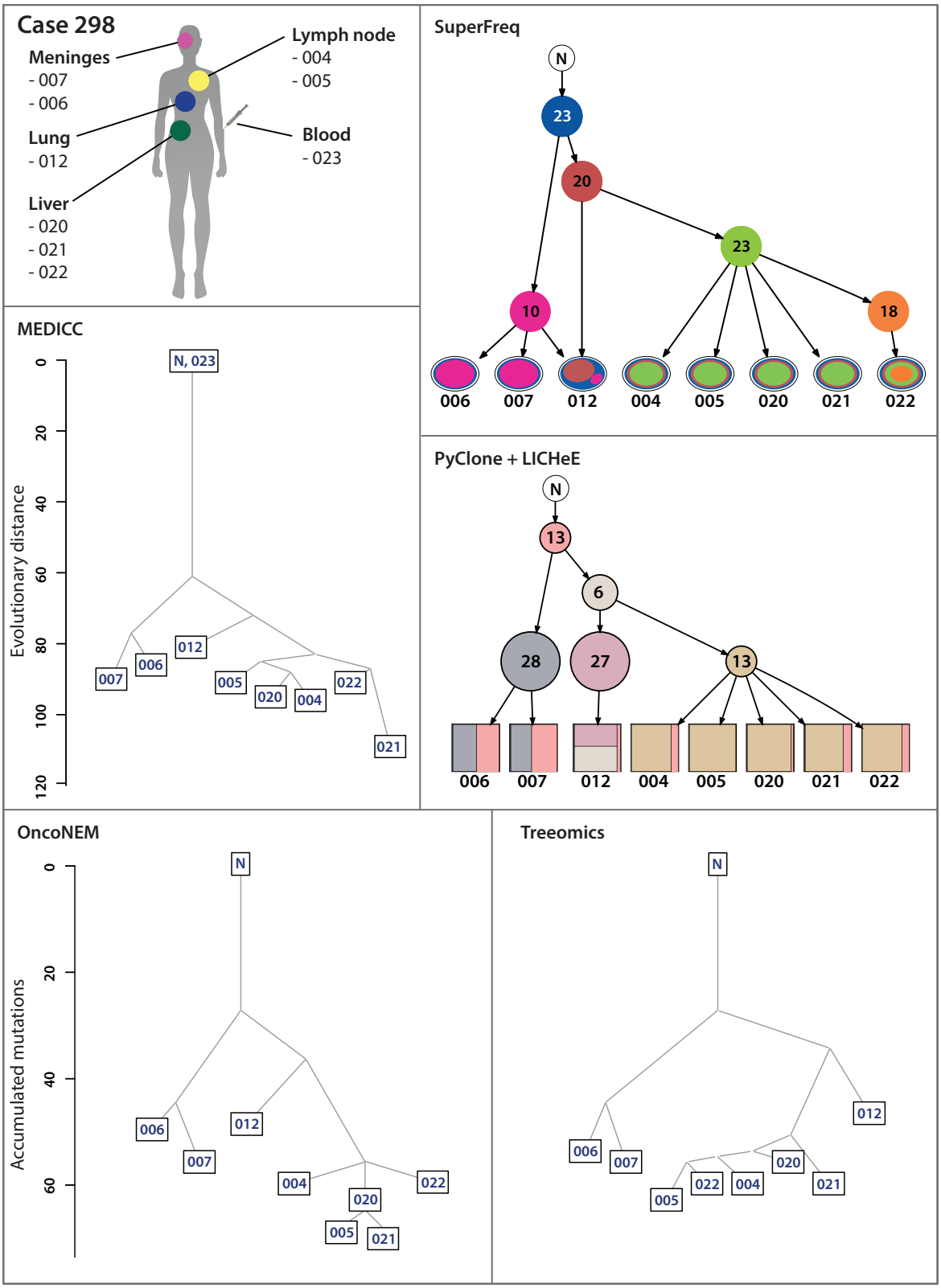


Fig. 4.7 Tree inference from WES and sWGS data for case 298.

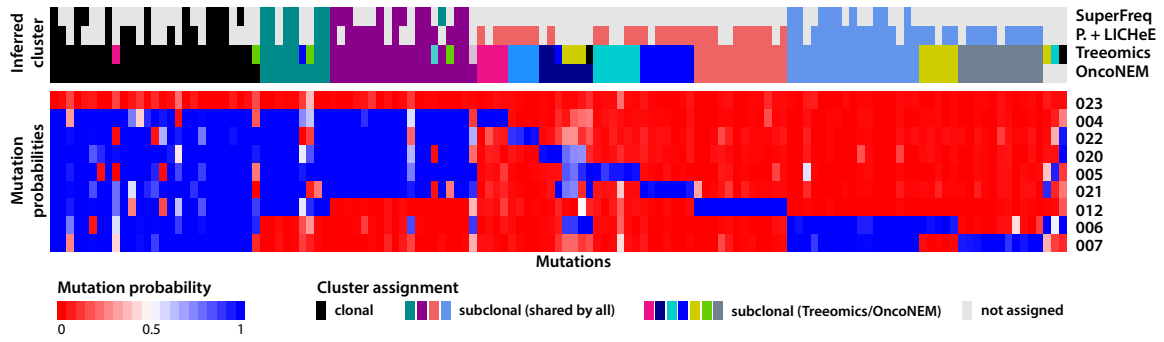


Fig. 4.8 WES derived point mutation matrix of case 298 (bottom) and cluster assignments of mutations for all SNV based methods (top). Clusters that were assigned to equivalent branches by different methods are marked in the same colour. Mutations that were not assigned to any cluster are shown in grey.

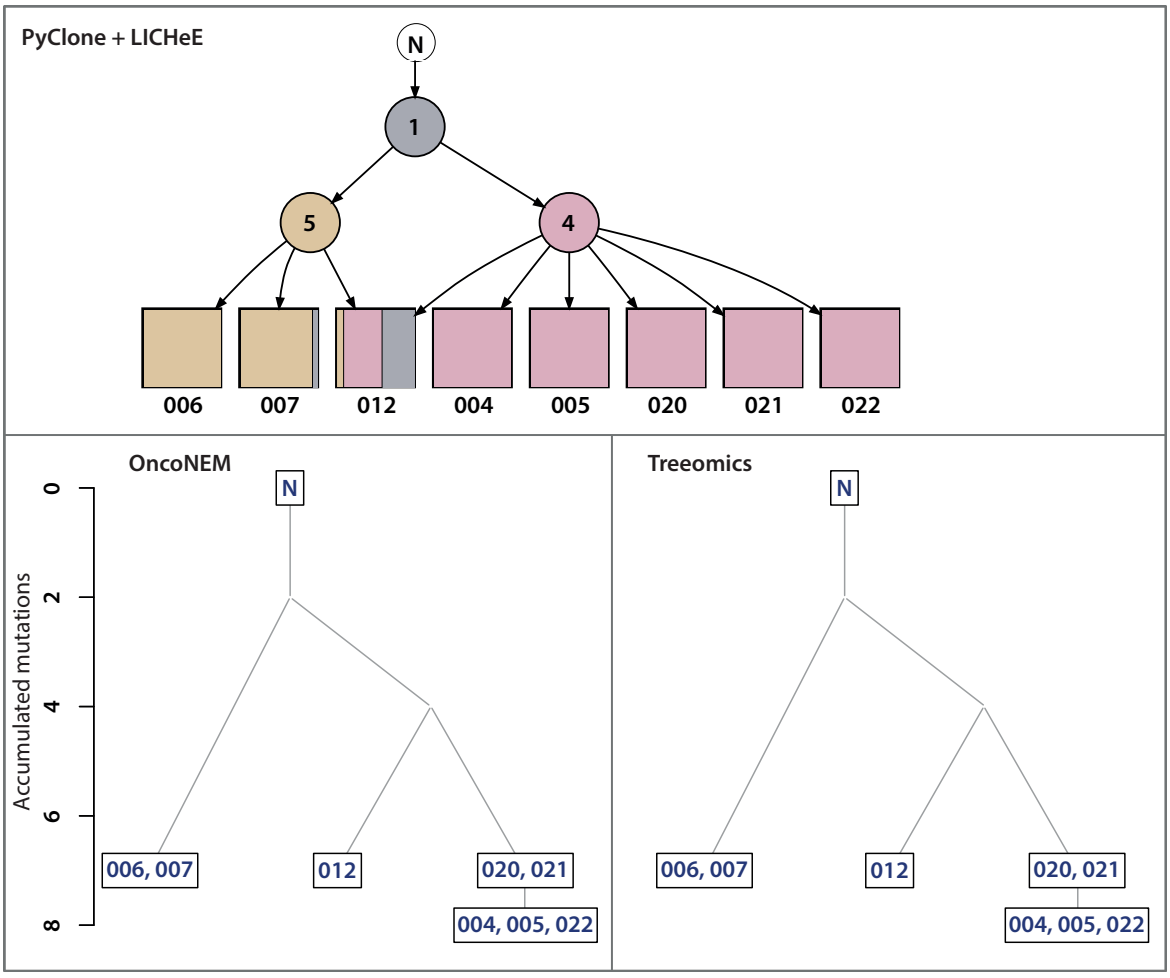


Fig. 4.9 Tree inference from targeted sequencing data for case 298.

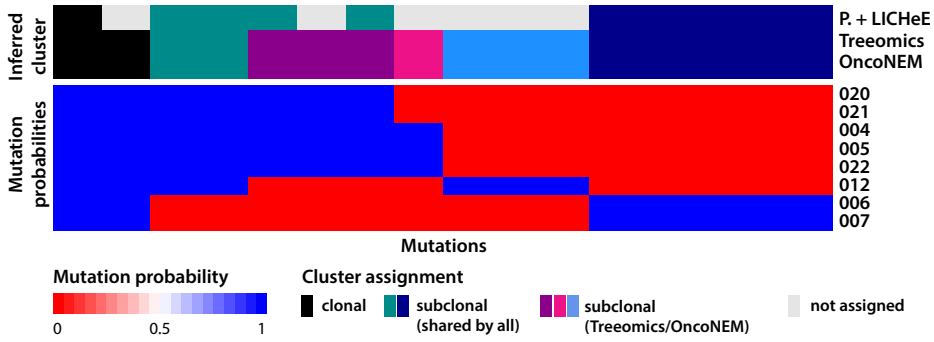


Fig. 4.10 TS derived point mutation matrix of case 298 (bottom) and cluster assignments of mutations (top). Clusters that were assigned to equivalent branches by different methods are marked in the same colour.

MEDICC (0.44), whereas LICHeE showed a negative correlation in relation to most of the methods.

Next, we evaluated how many point mutations of the WES data passed the analysis pipelines. Treeomics did not filter out any of the mutations, while OncoNEM's edge assignment was ambiguous for 2.3% of the mutations. PyClone/LICHeE and SuperFreq analysis removed 33.6% and 72.5% of mutations, respectively.

Furthermore, we tested if methods group the same mutations into clusters. A visual comparison of mutation clusters is shown in Figures 4.8 (WES) and 4.10 (TS). Considering only those mutations that are assigned by all methods (31 out of 131), the methods achieve perfect agreement in the WES case. In the TS case, LICHeE differs from Treeomics and OncoNEM in the assignment of 2 out of 16 mutations. Overall, Treeomics and OncoNEM infer very similar mutation clusters, achieving a V-measure cluster similarity score of 0.91 for WES data, even though they assign almost all mutations. LICHeE performs well at assigning mutations that are present in many of the samples. However, it misclassifies a large fraction of the mutations that are present in a single sample. For example, using the WES data, it clusters all the private mutations of samples 004, 005, 012, 020, 021 and 022 together and assigns them to a single mutation cluster that is private to sample 012. Similarly, it does not distinguish between the private mutations of sample 006 and 007 and clusters them together with the mutations shared by both samples. For this reason, its V-measure in relation to Treeomics and OncoNEM is relatively low (0.77 and 0.78 respectively for WES data). SuperFreq only assigns the mutations most methods agree on, which suggests that it restricts itself to high confidence assignments.

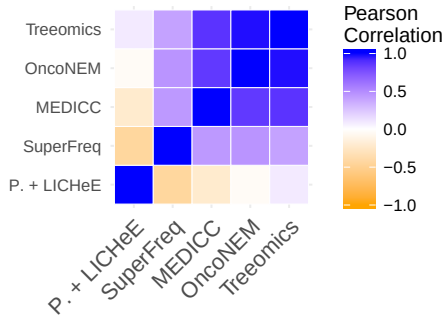


Fig. 4.11 Pairwise Pearson correlation of branch lengths in the core tree of WES based trees of case 298.

4.2.3.3 Case 290

Tree structure Overall, the phylogenetic analysis of 290 reveals three main subpopulations (Figure 4.12). Population A is represented by a single sample (024) from a stomach metastasis and was identified by all methods but SuperFreq. Population B comprises the brain metastases (005, 007, 008) as well as a subset of the liver metastases (018, 019). Population C includes the remaining liver metastases (015, 016-A, 016-B, 021).

The sample tree methods differ in their assignments of two liver metastases, 014 and 017. Treeomics, OncoNEM and MEDICC either assign them to population B or C. The uncertainty in placing these samples is not surprising, given that LICHeE infers that both contain a mix of cells from B and C. SuperFreq finds mixing of independent populations only in sample 014.

While the methods mostly agree on the assignment of samples to subpopulations, the tree structure connecting them is less clear. From the WES data, OncoNEM, Treeomics and MEDICC infer that population A branches off from an early ancestral population of B and C. LICHeE infers that all three developed from one shared most recent common ancestor. SuperFreq fails to identify a stem subpopulation. However, it infers one mutation cluster that is present in all samples (blue node in Figure 4.12). Most of the mutations in this cluster are classified as clonal by the other methods. This is shown in Figure 4.13 where the ubiquitous SuperFreq cluster is marked in yellow, and the clonal mutations of the other methods are marked in black. This finding suggests that the ubiquitous mutation cluster does not belong to an independent population but defines the common ancestor of the other two branches inferred by SuperFreq. Based on the TS data, in contrast, Treeomics and OncoNEM infer that

population B branched off from an early ancestral population of A and C, whereas LICHeE infers only two diverging branches, B and C.

Subpopulations Overall, mixed subpopulations were detected much more frequently than in the previous case. Out of the 14 samples, a mix of the two main subpopulations B and C was detected in eight samples by at least one method and in four using at least two different methods or data types (see Table 4.2). LICHeE inferred more mixing than any other method and identified more events in the TS than in the WES data. Treeomics was unable to detect any mixing of subpopulations in the WES but detected some mixing in the TS data. SuperFreq could only be applied to WES data and identified mixing in a single sample.

Given that the ground truth is unknown, it is impossible to decide whether LICHeE detects more mixing events than the other methods because it is more sensitive or because it infers more false positives. Its sensitivity can, however, be adapted by changing the filtering thresholds for the PyClone output. Out of the four mixing events that were identified using different methods or data types and which are therefore less likely to be artefacts, LICHeE identifies 100% of the tested cases from TS data and 75% from WES data. Treeomics identifies 75% from TS and 0% from WES data and SuperFreq detects 25% from WES data. The fact that Treeomics did not identify mixing in the WES data while all of the other methods detected at least one event suggests that it has low sensitivity on this type of data.

Even though OncoNEM cannot identify mixing of subpopulations, the mutation matrices support some of the mixing events, as they reveal repeatedly occurring mutation patterns that are incompatible with a tree structure. The WES mutation matrix supports three of the four high confidence mixing events (Figure 4.13), while the TS analysis supports two (Figure 4.15). Regions of incompatible mutations that agree with the mixing events detected by other methods are marked with white horizontal bars in these figures.

All in all, it seems likely that there are multiple metastases with a combination of the subpopulations B and C. Given that there are multiple mixed samples, this has to be at least partially due to multi-seeding or cross-seeding events.

Mutation assignment As in the previous case, we evaluated the correlation of branch lengths in the stable core tree between different methods (see Figure 4.16). For this analysis, the ubiquitous SuperFreq cluster was interpreted as a clonal cluster. Again, the branch lengths of the core tree estimated by OncoNEM and Treeomics show a strong correlation (0.97). The correlation of MEDICC with OncoNEM and Treeomics is 0.85 and 0.89, respectively. Unlike in the previous case, OncoNEM and Treeomics also show a good correlation with LICHeE (0.83 and 0.76) and SuperFreq (0.81 and 0.66).

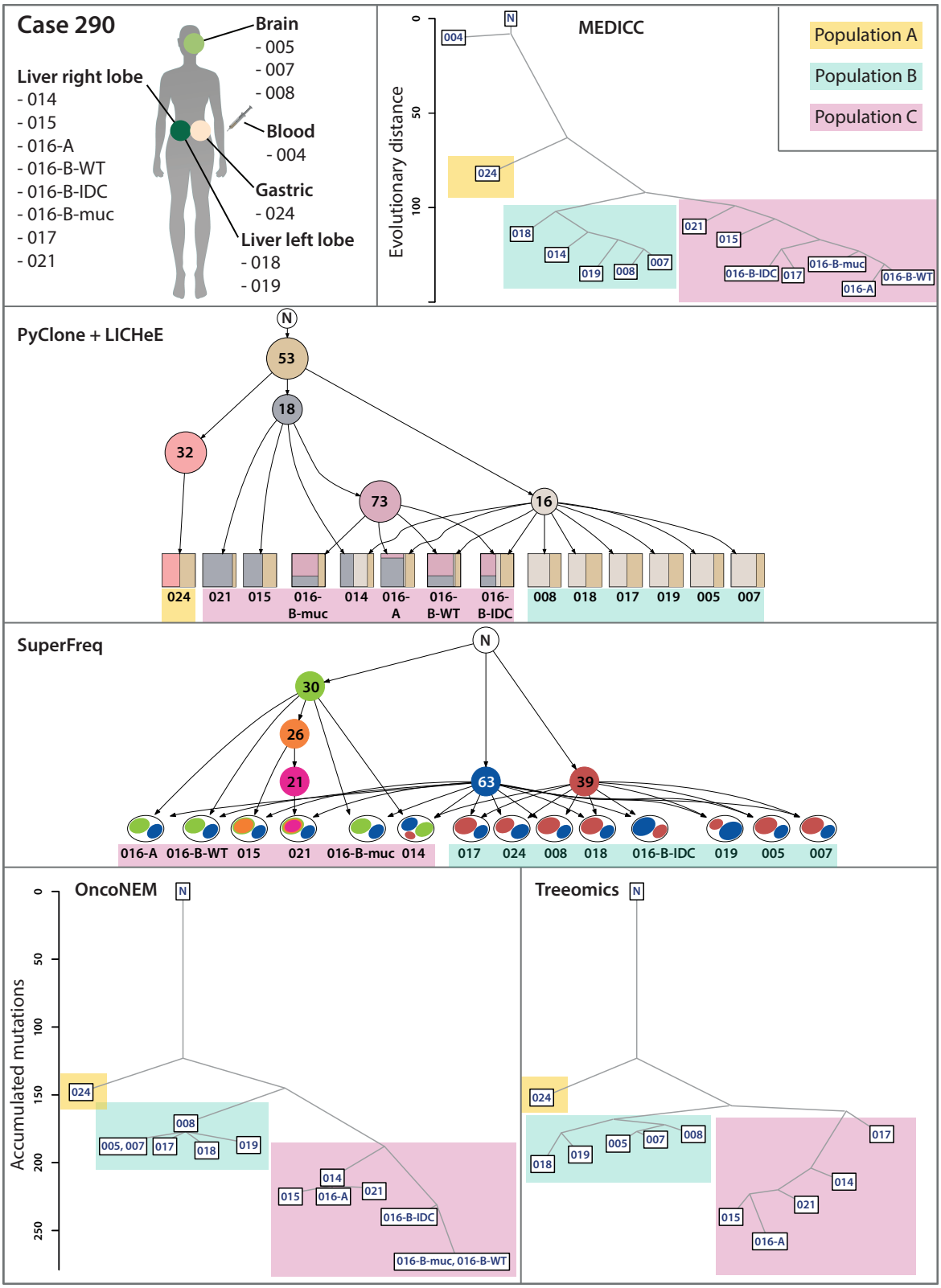


Fig. 4.12 Tree inference from WES data for case 290. Samples are coloured by their main subpopulation.

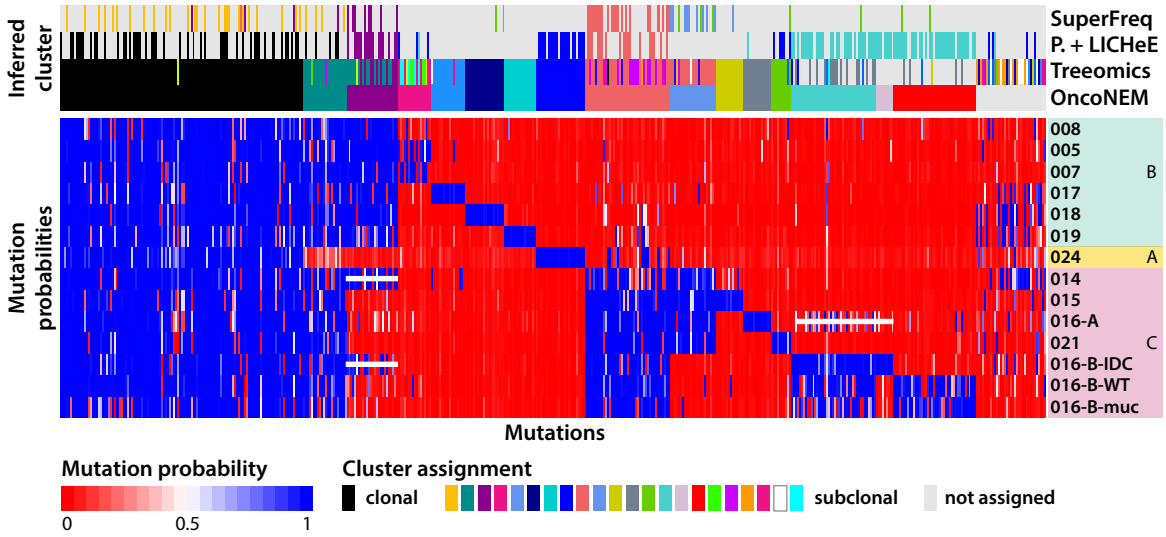


Fig. 4.13 WES derived point mutation matrix of case 290 (bottom) and cluster assignments of mutations (top). Clusters that were assigned to equivalent branches by different methods are marked in the same colour. SuperFreq failed to identify a clonal cluster and instead assigned the mutations in this cluster (yellow) to an independent branch in the tree. The mutation matrix reveals mutation patterns which are incompatible with a tree structure and therefore indicate mixing of independent subpopulations. Regions of incompatible mutation patterns are marked with white horizontal bars.

However, many of the variants get lost during the PyClone and LICHeE inference pipeline. Out of the initial 499 variants identified by WES, the 12 clusters inferred by PyClone contain 336. Only 8 clusters pass the quality filtering that removes spurious clusters, reducing the number of variants to 219. Out of the remaining 8 clusters, LICHeE ignores 3. The final tree, therefore, contains only 192 mutations, which corresponds to 38% of initial mutations. In case of SuperFreq, 17% of point mutations passed variant filtering, whereas OncoNEM assigns 93% of mutations to a cluster. Treeomics uses all of the mutations provided. The unassigned mutations in Figure 4.13 are due to the smaller number of samples analysed by Treeomics.

As before, we tested if methods assign mutations to equivalent branches in the trees they infer (see Figures 4.13 and 4.15). In this case, methods do not entirely agree on the clustering of the mutations assigned by all methods. Figure 4.16 shows the pairwise V-measure cluster similarity scores, which were calculated based on mutations assigned by all methods only. The V-measure is highest between SuperFreq and LICHeE as well as OncoNEM and LICHeE but is again impacted by the large fraction of unassigned mutations. Taking into account the other mutations as well, PyClone again clusters low-frequency mutations together even if

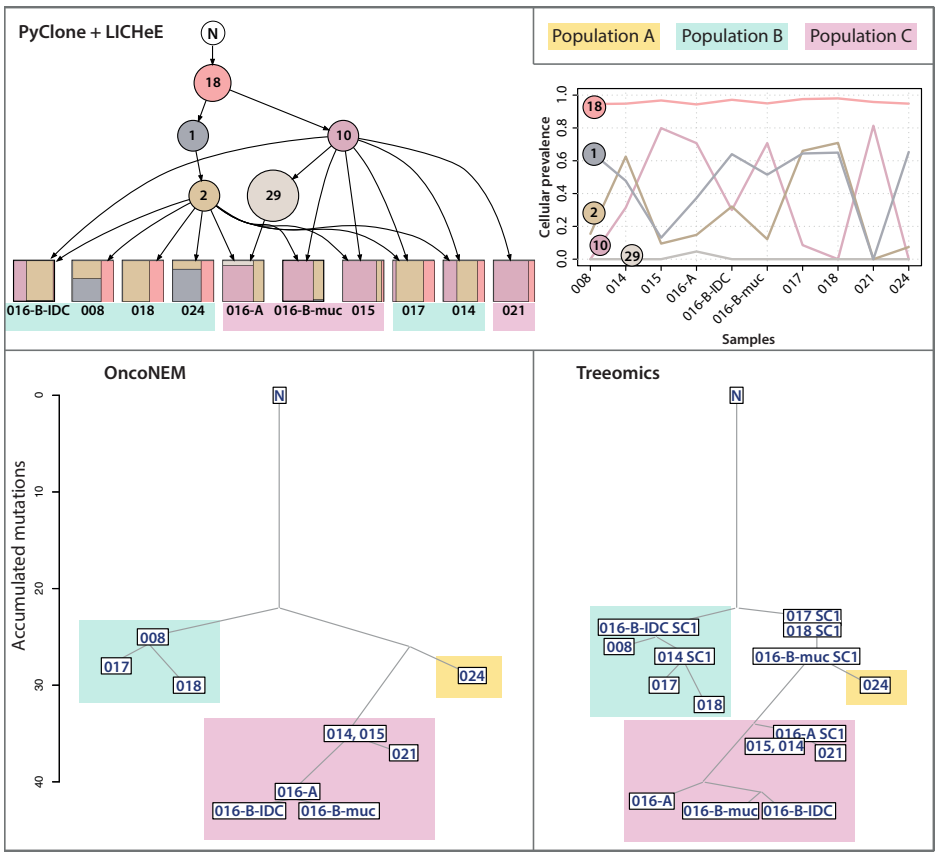


Fig. 4.14 Tree inference from targeted sequencing data for case 290. The line graph in the top right corner shows the cellular prevalence spectrum inferred by PyClone. The cellular prevalence of the cluster containing 29 mutations is compatible with multiple tree topologies. LICHeE, however only infers the structure shown on the left.

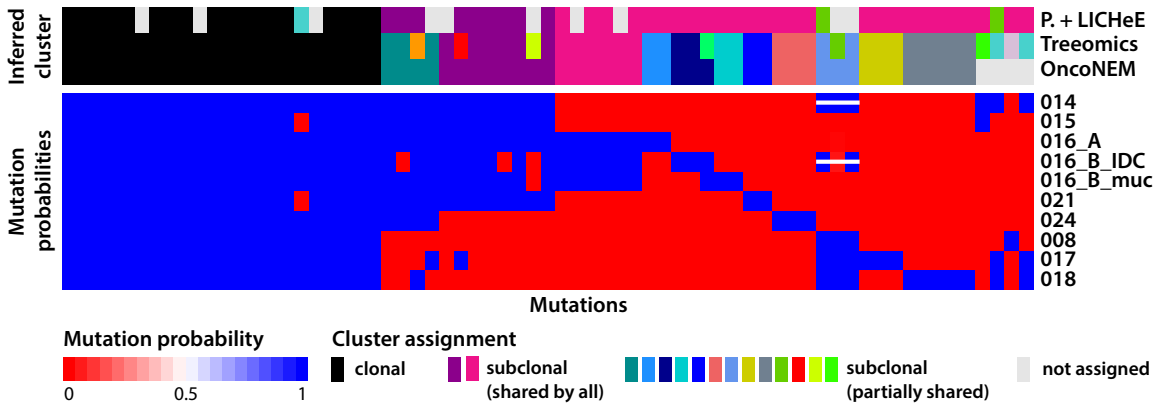


Fig. 4.15 TS derived point mutation matrix of case 290 (bottom) and cluster assignments of mutations (top). The white bars indicate mutation patterns that could be due to mixing of subpopulations.

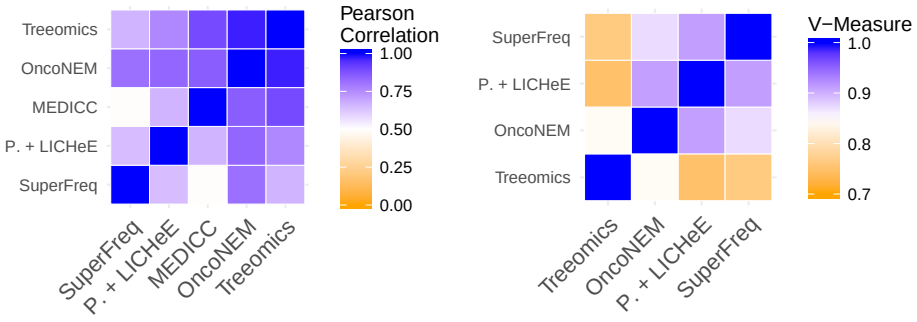


Fig. 4.16 Pairwise Pearson correlation of branch lengths in the core tree (left) and V-measure cluster similarity of mutations assigned by all methods (right) of WES based trees of case 290.

Table 4.2 Inference of subpopulations by different methods for case 290. Shown are all samples for which at least one method inferred mixing of subpopulations B and C in at least one data type. A tick mark indicates that mixing was detected, whereas a cross symbolises the opposite. Samples that had to be excluded from an analysis are marked with NA. Note that Treeomics identifies mixing of subpopulations within population C in some samples. These events are not included in this table.

Method	Data	014	015	016-A	016-B-WT	016-B-muc	016-B-IDC	017	018
P. + LICHeE	WES	✓	×	✓	✓	×	✓	×	×
	TS	✓	✓	✓	NA	✓	✓	✓	×
Treeomics	WES	×	×	×	NA	NA	NA	×	×
	TS	✓	×	×	NA	×	✓	✓	✓
SuperFreq	WES	✓	×	×	×	×	×	×	×

they occur in different samples (turquoise cluster in Figure 4.13 and pink cluster in 4.15), whereas SuperFreq excludes most of them.

In addition to the clustering challenges posed by low-frequency mutations, inferring the relationships of the resulting low prevalence clusters is also problematic as the evolutionary constraints defined by sum- and crossing-rule are unlikely to define a unique phylogeny. An example for this is the low-frequency cluster inferred by PyClone in the TS data, which contains 29 mutations (pink cluster in Figure 4.15). Based on the cellular prevalence estimates shown in Figure 4.14, several tree topologies are possible. In fact, the cluster could be downstream of any of the other clusters. Even though LICHeE can in principle infer multiple compatible trees, LICHeE only generated a single tree in this case and concluded that the low-frequency cluster lies downstream of the cluster containing ten mutations. For

some of the mutations in this cluster, this is a clear contradiction of the observed mutation patterns.

Together, the challenges of clustering and phylogeny inference of low-frequency mutations indicate that a higher filtering threshold for PyClone clusters should be used during the analysis to infer reliable phylogenies. This would further reduce the number of assigned variants. Treeomics and OncoNEM can infer phylogenies with higher resolution in this case, but their ability to detect subclone mixing is limited or non-existent. Furthermore, their performance may be worse for samples with more mixed subpopulations, as explained in Chapter 2.

4.2.3.4 Conclusions of method comparison

Assignment of mutations One important motivation for inferring trees is to assess which mutations are shared by different populations in the tumour. Throughout the LICHeE analysis pipeline, a large proportion of mutations is lost, limiting the ability to characterise the subpopulations in the tumour. In our study, this problem is even more pronounced for SuperFreq, which excludes private mutations and many of the mutations in regions of copy number aberrations. On average, the fraction of the genome affected by CNAs is larger in breast cancer than in many other cancers [124]. Therefore, SuperFreq may be better suited to analyse SNVs in other types of cancer. Treeomics and OncoNEM assign most of the variants to branches in the tree.

Mutation clusters An advantage of OncoNEM and Treeomics, is that they infer the tree structure and mutation clusters simultaneously. Mutation clusters inferred by these methods seem to be more reliable than those inferred by LICHeE and SuperFreq. This particularly concerns low prevalence and private mutations. Stronger filtering thresholds could reduce the number of spurious clusters but also reduce resolution. Reliability of mutation clusters inferred by MEDICC could not be assessed.

Resolution In general, OncoNEM and Treeomics infer very similar trees, however, the smaller branches within the major subpopulations are often not reproducible. A disadvantage of Treeomics is its limited ability to analyse data sets containing many samples. Even though MEDICC infers trees from copy number data, the results agree well with those of OncoNEM and Treeomics. At the same time, MEDICC is very sensitive to changes in the copy number fits, which shows that MEDICC does not handle noise well. LICHeE and SuperFreq infer trees of lower resolution, often clustering samples together that can be reliably distinguished

by the sample tree methods. Overall, LICHeE results seem to agree better with the other methods than those of SuperFreq.

As expected, the reliability of sample tree inference suffers if samples are a mixture of populations from different parts of the tree. Given that metastasis formation is an evolutionary bottleneck, metastases are likely to be more homogeneous than primary tumours. Performance of sample tree methods is likely to be worse for multi-sample data of primary tumours or for tumours with more cross-seeding between metastases.

Mixing of subpopulations An advantage of Treeomics over OncoNEM is the inference of mixed subpopulations. In practice, however, Treeomics did not identify mixtures of subpopulations in any of the cases from WES data. Using TS data, it was able to identify mixed populations in some of the cases. SuperFreq which can only be applied to WES data identified mixed populations in few cases. Overall, LICHeE is the most sensitive method, detecting mixing events in both WES and TS data. However, the specificity of the subpopulation detection could not be assessed due to the lack of a ground truth.

Usability The MEDICC analysis is the most labour intensive because ploidy estimates need to be curated manually. This makes MEDICC difficult to apply for inexperienced users. Furthermore, manual ploidy selection can introduce biases in the analysis.

PyClone also requires copy number profiles as input. However, it appears to be less sensitive to errors in the ploidy estimates than MEDICC and yields good results even if ploidy values have not been manually adjusted. In our study, LICHeE required extensive filtering of the PyClone input clusters. Without this, LICHeE ignored many of input clusters due to incompatibilities between clusters. Even though LICHeE offers the possibility to cluster mutations itself, it cannot adjust for copy number changes. We were unable to produce trees that resembled those of any of the other methods by clustering VAFs using LICHeE (data not shown). This could be different for samples with fewer CNAs.

SuperFreq is easy to use because almost all of the data pre-processing steps are implemented within its pipeline. At the same time, it offers no options to optimise the pipeline for a given data set and does not provide options for manual interventions. Even though SuperFreq infers subset relationships between mutation clusters, it only produces fish plots, whereas trees need to be generated by the user.

OncoNEM and Treeomics are both easy to use, working well with default parameters.

4.2.4 Summary

In summary, all methods have different advantages and disadvantages. The choice of method depends on the data available, the type of cancer that is being analysed and the biological question. Using multiple methods helps to identify stable regions of trees, to robustly assign mutations and to increase confidence in inferred mixing events.

Chapter 5

Summary and outlook

Over the past decade, cancer genome studies have generated large quantities of sequencing data. Novel computational tools are essential to analyse this data. In particular, the aim of inferring life histories of tumours poses novel challenges to computational biology. This dissertation presents methods and analyses that address multiple challenges in the field of tumour phylogenetics:

1. Single nucleotide variants obtained through single-cell DNA sequencing are noisy and classic phylogenetic methods are not suitable to analyse this kind of data. In Chapter 3, I develop a Bayesian scoring function to evaluate the fit between a tree hypothesis and observed mutation data. I present an efficient heuristic search algorithm and demonstrate the applicability of our approach in two case studies.

2. While statistical methods can be used to infer phylogenies from noisy data, the quality of phylogenetic analysis also depends on the quality of the input data. In the first part of Chapter 4, I describe a novel approach for the joint segmentation of copy number profiles of multiple related samples and demonstrate its advantages in two case studies.

3. A plethora of methods exist for the inference of tumour phylogenies from bulk sequencing data of multiple samples. An extensive comparison of these methods is lacking, and no best practice guide exists. In the second part of Chapter 4, I demonstrate how multiple, conceptually different methods can be used in a comprehensive phylogenetic analysis to assess the reliability of the inferred results.

In summary, this thesis contributes important methodology to the field of tumour phylogenetics and demonstrates the usefulness of multi-sample bulk sequencing approaches developed by others. It thereby highlights significant advances that have been made over

the past few years. Nevertheless, challenges and opportunities in different parts of the field remain to be addressed.

Phylogenetic methods for single-cell copy number data Copy number alterations can affect overlapping regions of the genome, which means that copy number data usually contains horizontal dependencies [151]. Therefore, OncoNEM and similar methods are unsuitable to infer phylogenies from this type of data. MEDICC, in turn, is not computationally efficient enough to analyse large numbers of cells and does not take into account noise in the data. Future single-cell inference approaches could use copy number information on its own or in combination with SNVs to analyse the phylogenies of copy number driven tumours.

Combining phylogenies and gene expression phenotypes Recent advances have enabled researchers to sequence both the genome and the transcriptome of a single cell [28, 92]. Combining single-cell phylogenies with single-cell transcriptomics will allow researchers to gain insights into how gene expression changes as tumours evolve.

Tumour evolution in the tissue context Cancers develop in a complex microenvironment, and it has been shown that interactions between tumour and stromal cells can promote the growth of primary tumours as well as the dissemination to other organs [90, 72]. Despite the evidence that the microenvironment plays a role in tumour evolution, most phylogenetic studies ignore the tissue context of samples. The most commonly used methods for single-cell isolation require tissue disaggregation, leading to a loss of information about the spatial origin of cells. Single-cell isolation techniques such as laser-capture microdissection, in contrast, offer the opportunity to analyse the genome as well as the spatial tissue context of a cell [153]. In future, this could enable researchers to study the connection between tumour evolution and the tumour microenvironment at a higher resolution.

Best practice guide for tumour phylogenetics More and more cancer studies that perform genome sequencing incorporate phylogenies in their analyses. This thesis demonstrates that a wide variety of phylogenetic methods have been developed for this purpose. However, few researchers beyond the method development community make use of these methods. Instead, many studies resort to unsuitable classic phylogenetic methods or manual inference. Zhao et al. [178], for example, used multiple classic phylogenetic methods for their analysis of SNVs but did not choose any methods developed for the inference of cancer evolution. Similarly, Jiménez-Sánchez et al. [69] used a parsimony method to infer phylogenies from SNVs. Joung et al. [71] also use a parsimony method but applied it to copy number data. Like

many bulk sequencing-based methods, Yates et al. [172] used a Dirichlet process mixture model to infer mutation clusters, but trees were inferred manually instead of using one of the automated approaches available. All of these studies were published within the last two years. Explanations as to why methods are not being used more widely could be a lack of visibility of methods beyond the method development community or a lack of trust in the methods. Alternatively, users might be overwhelmed by the number of options available and differences between methods may not be obvious to users outside the field. A best practice user guide could help to alleviate this problem. It could contain analyses like the ones described in Chapter 4 but should comprise a more comprehensive set of methods.

Going beyond tree inference As tumour phylogenetics is coming of age, tumour life histories can be inferred with reasonable confidence. While there is still room for improvement on the side of inference algorithms, the biggest challenge ahead lies in the interpretation of the inferred phylogenies. So far, inference algorithms are used to observe individual evolutionary trajectories that happened in the past. While this may be interesting from a basic science point of view, individual phylogenies are of limited use because they do not allow us to distinguish features of the tree that are important for the development and progression of the tumour from those that are not. Therefore, an important question for the future of the field will be if and how inferred phylogenies from many patients can be combined to further our understanding of cancer biology and to improve treatment strategies for future patients.

The holy grail would be to use tumour phylogenies of previous cases to predict the next evolutionary steps of new tumours [62]. This goal, however, is still far out of reach and many intermediary questions need to be answered first. For example, one question regarding the practicability of this approach is how one can aggregate the phylogenetic information across cases.

Matsui et al. [101] made a step in this direction by developing PhyC, an algorithm that clusters phylogenetic trees according to their shape. While the shape of the tree might be predictive of survival, as Matsui et al. suggest, other approaches are likely to be more informative for predicting tumour progression pathways. An alternative approach could, therefore, be to analyse phylogenies on a more functional level, for example, by aggregating information from the phylogenetic trees on the basis of mutated genes or pathways. This poses a challenging project in itself, as it is currently unclear if and how this could be done.

Even if we achieve to summarise the information gained from phylogenetic trees, many important questions remain unanswered: How do parameters such as the number of samples and the spatial distribution of the samples taken influence the resolution of the tree, its shape and ultimately the aggregated results? How many cases are needed and how detailed do the

phylogenetic trees need to be in order to achieve robust results? Most importantly, is the aggregated result informative? What can we learn from it? Can we discover reoccurring patterns of mutated pathways? How can we use all of this to make predictions?

All these questions show that there is a long way to go before these approaches can have an impact on real lives. Instead of focusing on incremental improvements by further optimising phylogenetic inference algorithms, it is crucial to approach the next steps that go beyond tree inference in order to ensure that tumour phylogenetics will lead to more than singular observations. This does not imply that further technical improvements of phylogenetic algorithms will be unnecessary. However, further optimisations of phylogenetic algorithms should be guided by the requirements of the aggregation methods that are yet to be developed. In any case, without progress which brings us closer to summarising phylogenetic information across cases to generate robust biological conclusions, the information gained through tumour phylogenetics cannot be used to its full potential.

References

[1] Aceto, N., Bardia, A., Miyamoto, D. T., Donaldson, M. C., Wittner, B. S., Spencer, J. A., Yu, M., Pely, A., Engstrom, A., Zhu, H., Brannigan, B. W., Kapur, R., Stott, S. L., Shioda, T., Ramaswamy, S., Ting, D. T., Lin, C. P., Toner, M., Haber, D. A., and Maheswaran, S. (2014). Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell*, 158(5):1110–1122.

[2] Adams, R. P., Ghahramani, Z., and Jordan, M. I. (2010). Tree-Structured Stick Breaking for Hierarchical Data. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 19–27. Curran Associates, Inc.

[3] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Borresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjord, J. E., Foekens, J. A., Greaves, M., et al. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421.

[4] Alves, J. M., Prieto, T., and Posada, D. (2017). Multiregional Tumor Trees Are Not Phylogenies. *Trends in Cancer*, 10(0):e1003703.

[5] Amirouchene-Angelozzi, N., Swanton, C., and Bardelli, A. (2017). Tumor Evolution as a Therapeutic Target. *Cancer Discovery*.

[6] Andor, N., Harness, J. V., Müller, S., Mewes, H. W., and Petritsch, C. (2014). EXPANDS: Expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics*, 30(1):50–60.

[7] Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A.-M., Gingras, M.-C., Miller, D. K., Christ, A. N., Bruxner, T. J. C., Quinn, M. C., Nourse, C., Murtaugh, L. C., Harliwong, I., Idrisoglu, S., Manning, S., Nourbakhsh, E., Wani, S., Fink, L., Holmes, O., Chin, V., et al. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, 531(7592):47–52.

[8] Beerenwinkel, N., Antal, T., Dingli, D., Traulsen, A., Kinzler, K. W., Velculescu, V. E., Vogelstein, B., and Nowak, M. A. (2007). Genetic progression and the waiting time to cancer. *PLoS Computational Biology*, 3(11):2239–2246.

[9] Beerenwinkel, N., Schwarz, R. F., Gerstung, M., and Markowetz, F. (2015). Cancer evolution: Mathematical models and computational inference. *Systematic Biology*, 64(1):e1–e25.

[10] Bhattacherjee, V., Mukhopadhyay, P., Singh, S., Roberts, E. A., Hackmiller, R. C., Greene, R. M., and Pisano, M. M. (2004). Laser capture microdissection of fluorescently labeled embryonic cranial neural crest cells. *Genesis*, 39:58–64.

[11] Bielas, J. H., Loeb, K. R., Rubin, B. P., True, L. D., and Loeb, L. A. (2006). Human cancers express a mutator phenotype. *PNAS*, 48(103):18238–18242.

[12] Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1 A):121–144.

[13] Bonizzoni, P., Braghin, C., Dondi, R., and Trucco, G. (2012). The binary perfect phylogeny with persistent characters. *Theoretical Computer Science*, 454:51–63.

[14] Boveri, T. (2007). Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris. *Journal of Cell Science*, 121(Supplement 1):1–84.

[15] CancerIT (Genome Research Ltd.) (2017). alleleCount 3.1.1. https://github.com/cancerit/alleleCount.

[16] Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M., and Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421.

[17] Chedom-Fotso, D., Ahmed, A. A., and Yau, C. (2016). OncoPhase: Quantification of somatic mutation cellular prevalence using phase information. *bioRxiv*, doi:10.1101/046631.

[18] Chen, X., Gupta, P., Wang, J., Nakitandwe, J., Roberts, K., Dalton, J. D., Parker, M., Patel, S., Holmfeldt, L., Payne, D., Easton, J., Ma, J., Rusch, M., Wu, G., Patel, A., Baker, S. J., Dyer, M. A., Shurtleff, S., Espy, S., Pounds, S., Downing, J. R., Ellison, D. W., Mullighan, C. G., and Zhang, J. (2015). CONSERTING : integrating copy-number analysis with structural-variation detection. *Nature Methods*, 12(July 2014):527–530.

[19] Chowdhury, S. A., Shackney, S. E., Heselmeyer-Haddad, K., Ried, T., Schäffer, A. A., and Schwartz, R. (2013). Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations. *Bioinformatics*, 29.

[20] Chowdhury, S. A., Shackney, S. E., Heselmeyer-Haddad, K., Ried, T., Schäffer, A. A., and Schwartz, R. (2014). Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics. *PLoS Computational Biology*, 10(7):e1003740.

[21] Crowley, E., Di Nicolantonio, F., Loupakis, F., and Bardelli, A. (2013). Liquid biopsy: monitoring cancer-genetics in the blood. *Nature Reviews Clinical Oncology*, 10(8):472–484.

[22] Curtis, C., Shah, S. P., Chin, S.-f., Turashvili, G., Rueda, O. M., Dunning, M. J., Ha, G., Haffari, G., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., and Purushotham, A. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486:346–352.

[23] Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P. S., Rothenberg, M. E., Leyrat, A. A., Sim, S., Okamoto, J., Johnston, D. M., Qian, D., Zabala, M., Bueno, J., Neff, N. F., Wang, J., Shelton, A. A., Visser, B., Hisamori, S., Shimono, Y., van de Wetering, M., Clevers, H., Clarke, M. F., and Quake, S. R. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology*, 29(12):1120–1127.

[24] Davis, A. and Navin, N. E. (2016). Computing tumor trees from single cells. *Genome Biology*, 17:113.

[25] De Mattos Arruda, L., Sammut, S.-J., Ross, E. M., et al. (2017). The integrated genomic and immune landscapes of lethal metastatic breast cancer. *In preparation*.

[26] Dean, F. B., Nelson, J. R., Giesler, T. L., and Lasken, R. S. (2001). Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Research*, 11(6):1095–1099.

[27] Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L., and Morris, Q. (2015). PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35.

[28] Dey, S. S., Kester, L., Spanjaard, B., Bienko, M., and van Oudenaarden, A. (2015). Integrated genome and transcriptome sequencing of the same cell. *Nature Biotechnology*, 33(3):285–289.

[29] Diaz Jr, L. a., Williams, R. T., Wu, J., Kinde, I., Hecht, J. R., Berlin, J., Allen, B., Bozic, I., Reiter, J. G., Nowak, M. a., Kinzler, K. W., Oliner, K. S., and Vogelstein, B. (2012). The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature*, 486(7404):4–7.

[30] Diehl, F., Schmidt, K., Choti, M. a., Romans, K., Goodman, S., Li, M., Thornton, K., Agrawal, N., Sokoll, L., Szabo, S. a., Kinzler, K. W., Vogelstein, B., and Diaz, L. a. (2008). Circulating mutant DNA to assess tumor dynamics. *Nature Medicine*, 14(9):985–990.

[31] Donmez, N., Malikic, S., Wyatt, A. W., Gleave, M. E., Collins, C. C., and Sahinalp, S. C. (2016). Clonality Inference from Single Tumor Samples Using Low Coverage Sequence Data. In Singh, M., editor, *Research in Computational Molecular Biology. RECOMB 2016*, pages 83–94, Cham. Springer.

[32] Eddelbuettel, D. (2013). Seamless R and C++ Integration with Rcpp.

[33] Eirew, P., Steif, A., Khattra, J., Ha, G., Yap, D., Farahani, H., Gelmon, K., Chia, S., Mar, C., Wan, A., Laks, E., Biele, J., Shumansky, K., Rosner, J., McPherson, A., Nielsen, C., Roth, A. J. L., Lefebvre, C., Bashashati, A., de Souza, C., Siu, C., Aniba, R., Brimhall, J., Oloumi, A., Osako, T., Bruna, A., Sandoval, J. L., Algara, T., Greenwood, W., Leung, K., Cheng, H., Xue, H., Wang, Y., Lin, D., Mungall, A. J., Moore, R., Zhao, Y., Lorette, J., Nguyen, L., Huntsman, D., Eaves, C. J., Hansen, C., Marra, M. A., Caldas, C., Shah, S. P., and Aparicio, S. (2015). Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539):422–426.

[34] El-Kebir, M., Oesper, L., Acheson-Field, H., and Raphael, B. J. (2015). Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70.

[35] El-Kebir, M., Satas, G., Oesper, L., Raphael, B. J., Jung, J., Maire, C., Ligon, K., Meyerson, M., Love, J., Mermel, C., Al., E., and Al., E. (2016). Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. *Cell Systems*, 3(1):43–53.

[36] Fan, H. C., Wang, J., Potanina, A., and Quake, S. R. (2011). Whole-genome molecular haplotyping of single cells. *Nature Biotechnology*, 29(1):51–57.

[37] Feder, A. F., Rhee, S. Y., Holmes, S. P., Shafer, R. W., Petrov, D. A., and Pennings, P. S. (2016). More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. *eLife*, 5:e10670.

[38] Felsenstein, J. (1988). Phylogenies from Molecular Sequences: Inference and Reliability. *Annual Review of Genetics*, 22(1):521–565.

[39] Fischer, A., Vázquez-García, I., Illingworth, C. J. R., and Mustonen, V. (2014). High-definition reconstruction of clonal composition in cancer. *Cell Reports*, 7(5):1740–1752.

[40] Flensburg, C. (2017). SuperFreq 0.9.17. https://github.com/ChristofferFlensburg/superFreq.

[41] Fluidigm Inc. (2016). Doublet Rate and Detection on the C1 IFCs. Technical report, Fluidigm Inc.

[42] Forshew, T., Murtaza, M., Parkinson, C., Gale, D., Tsui, D. W. Y., Kaper, F., Dawson, S.-J., Piskorz, A. M., Jimenez-Linan, M., Bentley, D., Hadfield, J., May, A. P., Caldas, C., Brenton, J. D., and Rosenfeld, N. (2012). Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Science Translational Medicine*, 4(136):136ra68.

[43] Fraley, C. (2002). Model-based clustering, discriminant analysis, and density estimation. J. *Journal of the American Statistical Association*, 97(458):611–631.

[44] Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., and Jain, A. N. (2004). Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, 90:132–153.

[45] Frumkin, D., Wasserstrom, A., Itzkovitz, S., Harmelin, A., Rechavi, G., and Shapiro, E. (2008). Amplification of multiple genomic loci from single cells isolated by laser micro-dissection of tissues. *BMC Biotechnology*, 8:17.

[46] Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G. S., Hicks, J., Wigler, M., and Schatz, M. C. (2015). Interactive analysis and assessment of single-cell copy-number variations. *Nature Methods*, 12(11):1058–1060.

[47] Gawad, C., Koh, W., and Quake, S. R. (2014). Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *PNAS*, 111.

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

[49] Genome Research Ltd. (2017). Samtools 1.3.1. http://www.htslib.org.

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

[51] Gershman, S. J. and Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56:1–12.

[52] Greaves, M. and Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, 481(7381):306–313.

[53] Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Gilks, C. B., Huntsman, D. G., McAlpine, J. N., Aparicio, S., and Shah, S. P. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumor whole genome sequence data. *Genome Research*, 24:1881–1893.

[54] Hajirasouliha, I., Mahmoody, A., and Raphael, B. J. (2014). A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, 30(12):78–86.

[55] Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002). Clustering Validity Checking Methods : Part II. *SIGMOD Record*, 31(3):19–27.

[56] Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70.

[57] Hastie, T., Tibshirani, R., and Friedman, J. H. J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer.

[58] Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., Wu, H., Ye, X., Ye, C., Wu, R., Jian, M., Chen, Y., Xie, W., Zhang, R., Chen, L., Liu, X., Yao, X., Zheng, H., Yu, C., Li, Q., Gong, Z., Mao, M., Yang, X., Yang, L., Li, J., Wang, W., Lu, Z., Gu, N., Laurie, G., Bolund, L., Kristiansen, K., Wang, J., Yang, H., Li, Y., Zhang, X., and Wang, J. (2012). Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell*, 148(5):873–885.

[59] Huang, L., Ma, F., Chapman, A., Lu, S., and Xie, X. S. (2015). Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. *Annual Review of Genomics and Human Genetics*, 16:79–102.

[60] Hughes, A. E. O., Magrini, V., Demeter, R., Miller, C. A., Fulton, R., Fulton, L. L., Eades, W. C., Elliott, K., Heath, S., Westervelt, P., Ding, L., Conrad, D. F., White, B. S., Shao, J., Link, D. C., DiPersio, J. F., Mardis, E. R., Wilson, R. K., Ley, T. J., Walter, M. J., and Graubert, T. A. (2014). Clonal Architecture of Secondary Acute Myeloid Leukemia Defined by Single-Cell Sequencing. *PLoS Genetics*, 10(7):e1004462.

[61] Hugo, W., Shi, H., Sun, L., Piva, M., Song, C., Kong, X., Moriceau, G., Hong, A., Dahlman, K., Johnson, D., Sosman, J., Ribas, A., and Lo, R. (2015). Non-genomic and Immune Evolution of Melanoma Acquiring MAPKi Resistance. *Cell*, 162(6):1271–1285.

[62] Hutchinson, L. (2014). Predicting cancer's next move. *Nature Reviews Clinical Oncology*, 11(2):61–62.

[63] Inc., I. (2017). HiSeq 3000/HiSeq 4000 System quality and performance. https://www.illumina.com/systems/sequencing-platforms/hiseq-3000-4000/specifications.html.

[64] Inukai, M., Toyooka, S., Ito, S., Asano, H., Ichihara, S., Soh, J., Suehisa, H., Ouchida, M., Aoe, K., Aoe, M., Kiura, K., Shimizu, N., and Date, H. (2006). Presence of epidermal growth factor receptor gene T790M mutation as a minor clone in non-small cell lung cancer. *Cancer Research*, 66(16):7854–7858.

[65] Jahn, K., Kuipers, J., and Beerenwinkel, N. (2016). Tree inference for single-cell data. *Genome Biology*, 17:86.

[66] Jeffreys, H. (1998). *Theory of Probability*. Oxford University Press, Oxford.

[67] Jiang, Y., Qiu, Y., Minn, A. J., and Zhang, N. R. (2016). Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *PNAS*, 113(37):E5528–37.

[68] Jiao, W., Vembu, S., Deshwar, A. G., Stein, L., and Morris, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, 15:35.

[69] Jiménez-Sánchez, A., Memon, D., Pourpe, S., Veeraraghavan, H., Li, Y., Vargas, H. A., Gill, M. B., Park, K. J., Zivanovic, O., Konner, J., Ricca, J., Zamarin, D., Walther, T., Aghajanian, C., Wolchok, J. D., Sala, E., Merghoub, T., Snyder, A., and Miller, M. L. (2017). Heterogeneous Tumor-Immune Microenvironments among Differentially Growing Metastases in an Ovarian Cancer Patient. *Cell*, 170(5):927–938.e20.

[70] Johnson, B. E., Mazor, T., Hong, C., Barnes, M., Aihara, K., McLean, C. Y., Fouse, S. D., Yamamoto, S., Ueda, H., Tatsuno, K., Asthana, S., Jalbert, L. E., Nelson, S. J., Bollen, A. W., Gustafson, W. C., Charron, E., Weiss, W. A., Smirnov, I. V., Song, J. S., Olshen, A. B., Cha, S., Zhao, Y., Moore, R. A., Mungall, A. J., Jones, S. J. M., Hirst, M., Marra, M. A., Saito, N., Aburatani, H., Mukasa, A., Berger, M. S., Chang, S. M., Taylor, B. S., and Costello, J. F. (2014). Mutational Analysis Reveals the Origin and Therapy-Driven Evolution of Recurrent Glioma. *Science*, 343(6167):189–194.

[71] Joung, J.-G., Ha, S. Y., Bae, J. S., Nam, J.-Y., Gwak, G.-Y., Lee, H.-O., Son, D.-S., Park, C.-K., Park, W.-Y., Joung, J.-G., Ha, S. Y., Bae, J. S., Nam, J.-Y., Gwak, G.-Y., Lee, H.-O., Son, D.-S., Park, C.-K., and Park, W.-Y. (2017). Nonlinear tumor evolution from dysplastic nodules to hepatocellular carcinoma. *Oncotarget*, 8(2):2076–2082.

[72] Joyce, J. a. and Pollard, J. W. (2009). Microenvironmental regulation of metastasis. *Nature Reviews Cancer*, 9(4):239–52.

[73] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

[74] Kim, H., Zheng, S., Amini, S. S., Virk, S. M., Mikkelsen, T., Brat, D. J., Grimsby, J., Sougnez, C., Muller, F., Hu, J., Sloan, A. E., Cohen, M. L., Van Meir, E. G., Scarpace, L., Laird, P. W., Weinstein, J. N., Lander, E. S., Gabriel, S., Getz, G., Meyerson, M., Chin, L., Barnholtz-Sloan, J. S., and Verhaak, R. G. W. (2015). Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Research*, 25(3):316–327.

[75] Kim, K. I. and Simon, R. (2014). Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics*, 15:27.

[76] Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903.

[77] Knoechel, B., Roderick, J. E., Williamson, K. E., Zhu, J., Lohr, J. G., Cotton, M. J., Gillespie, S. M., Fernandez, D., Ku, M., Wang, H., Piccioni, F., Silver, S. J., Jain, M., Pearson, D., Kluk, M. J., Ott, C. J., Shultz, L. D., Brehm, M. A., Greiner, D. L., Gutierrez, A., Stegmaier, K., Kung, A. L., Root, D. E., Bradner, J. E., Aster, J. C., Kelliher, M. A., and Bernstein, B. E. (2014). An epigenetic mechanism of resistance to targeted therapy in T cell acute lymphoblastic leukemia. *Nature Genetics*, 46(4):364–370.

[78] Koboldt, D. (2017). VarScan 2.4.3. http://dkoboldt.github.io/varscan/.

[79] Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., Kertesz, M., and Snyder, M. (2014). Whole-genome haplotyping using long reads and statistical methods. *Nature Biotechnology*, 32(3):261–266.

[80] Kumar, A., Coleman, I., Morrissey, C., Zhang, X., True, L. D., Gulati, R., Etzioni, R., Bolouri, H., Montgomery, B., White, T., Lucas, J. M., Brown, L. G., Dumpit, R. F., DeSarkar, N., Higano, C., Yu, E. Y., Coleman, R., Schultz, N., Fang, M., Lange, P. H., Shendure, J., Vessella, R. L., and Nelson, P. S. (2016). Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nature Medicine*, 22(4):369–378.

[81] Letouzé, E., Allory, Y., Bollet, M. a., Radvanyi, F., and Guyon, F. (2010). Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biology*, 11(7):R76.

[82] Leung, M. L., Wang, Y., Kim, C., Gao, R., Jiang, J., Sei, E., and Navin, N. E. (2016). Highly multiplexed targeted DNA sequencing from single nuclei. *Nature Protocols*, 11(2):214–235.

[83] Leung, M. L., Wang, Y., Waters, J., and Navin, N. E. (2015). SNES: single nucleus exome sequencing. *Genome Biology*, 16(1):55.

[84] Li, B. and Li, J. Z. (2014). A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biology*, 15(9):473.

[85] Li, Y., Xu, X., Song, L., Hou, Y., Li, Z., Tsang, S., Li, F., Im, K. M., Wu, K., Wu, H., Ye, X., Li, G., Wang, L., Zhang, B., Liang, J., Xie, W., Wu, R., Jiang, H., Liu, X., Yu, C., Zheng, H., Jian, M., Nie, L., Wan, L., Shi, M., Sun, X., Tang, A., Guo, G., Gui, Y., Cai, Z., Li, J., Wang, W., Lu, Z., Zhang, X., Bolund, L., Kristiansen, K., Wang, J., Yang, H.,

Dean, M., and Wang, J. (2012). Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *GigaScience*, 1:12.

[86] Lin, S. and Kernighan, B. W. (1973). An Effective Heuristic Algorithm for the Traveling-Salesman Problem. *Operations Research*, 21(2):498–516.

[87] Loeb, L. A. (1991). Mutator Phenotype May Be Required for Multistage Carcinogenesis. *Cancer Research*, 51:3075–3079.

[88] Loeb, L. A., Bielas, J. H., and Beckman, R. A. (2008). Cancers exhibit a mutator phenotype: Clinical implications. *Cancer Research*, 68(10):3551–3557.

[89] Lohr, J. G., Adalsteinsson, V. A., Cibulskis, K., Choudhury, A. D., Rosenberg, M., Cruz-Gordillo, P., Francis, J. M., Zhang, C.-Z., Shalek, A. K., Satija, R., Trombetta, J. J., Lu, D., Tallapragada, N., Tahirova, N., Kim, S., Blumenstiel, B., Sougnez, C., Lowe, A., Wong, B., Auclair, D., Van Allen, E. M., Nakabayashi, M., Lis, R. T., Lee, G.-S. M., Li, T., Chabot, M. S., Ly, A., Taplin, M.-E., Clancy, T. E., Loda, M., Regev, A., Meyerson, M., Hahn, W. C., Kantoff, P. W., Golub, T. R., Getz, G., Boehm, J. S., and Love, J. C. (2014). Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nature Biotechnology*, 32(5):479–484.

[90] Lorusso, G. and Rüegg, C. (2008). The tumor microenvironment and its contribution to tumor evolution toward metastasis. *Histochemistry and Cell Biology*, 130(6):1091–1103.

[91] Ma, J., Ratan, A., Raney, B. J., Suh, B. B., Miller, W., and Haussler, D. (2008). The infinite sites model of genome evolution. *PNAS*, 105(38):14254–61.

[92] Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., Van der Aa, N., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., Ponting, C. P., and Voet, T. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12(6):519–522.

[93] Macintyre, G., Van Loo, P., Corcoran, N. M., Wedge, D. C., Markowetz, F., and Hovens, C. M. (2017). How subclonal modeling is changing the metastatic paradigm. *Clinical Cancer Research*, 23(3):630–635.

[94] Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

[95] Malikic, S., McPherson, A. W., Donmez, N., and Sahinalp, C. S. (2015). Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–1356.

[96] Mangiola, S., Hong, M. K. H., Cmero, M., Kurganovs, N., Ryan, A., Costello, A. J., Corcoran, N. M., Macintyre, G., and Hovens, C. M. (2016). Comparing nodal versus bony metastatic spread using tumour phylogenies. *Scientific Reports*, 6:33918.

[97] Marass, F., Mouliere, F., Yuan, K., Rosenfeld, N., and Markowetz, F. (2016). A phylogenetic latent feature model for clonal deconvolution. *Annals of Applied Statistics*, 10(4):2377–2404.

[98] Marioni, J. C., Thorne, N. P., and Tavaré, S. (2006). BioHMM: A heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, 22(9):1144–1146.

[99] Markowetz, F., Bloch, J., and Spang, R. (2005). Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics*, 21(21):4026–4032.

[100] Markowetz, F., Kostka, D., Troyanskaya, O. G., and Spang, R. (2007). Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, 23(13):305–312.

[101] Matsui, Y., Niida, A., Uchi, R., Mimori, K., and Shimamura, T. (2017). phyC : Clustering cancer evolutionary trees. *PLoS Computational Biology*, 13(5):e1005509.

[102] Mayrhofer, M., DiLorenzo, S., and Isaksson, A. (2013). Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. *Genome Biology*, 14(3):R24.

[103] McPherson, A., Roth, A., Laks, E., Masud, T., Bashashati, A., Zhang, A. W., Ha, G., Biele, J., Yap, D., Wan, A., Prentice, L. M., Khattra, J., Smith, M. A., Nielsen, C. B., Mullaly, S. C., Kalloger, S., Karnezis, A., Shumansky, K., Siu, C., Rosner, J., Chan, H. L., Ho, J., Melnyk, N., Senz, J., Yang, W., Moore, R., Mungall, A. J., Marra, M. A., Bouchard-Côté, A., Gilks, C. B., Huntsman, D. G., McAlpine, J. N., Aparicio, S., and Shah, S. P. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*, 48(7):758–767.

[104] Merlo, L. M., Pepper, J. W., Reid, B. J., and Maley, C. C. (2006). Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*, 6(12):924–935.

[105] Miller, C. A., White, B. S., Dees, N. D., Griffith, M., Welch, J. S., Griffith, O. L., Vij, R., Tomasson, M. H., Graubert, T. A., Walter, M. J., Ellis, M. J., Schierding, W., DiPersio, J. F., Ley, T. J., Mardis, E. R., Wilson, R. K., and Ding, L. (2014). SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Computational Biology*, 10(8):e1003665.

[106] Mitelman, F., Johansson, B., and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*, 7(4):233–245.

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

[108] Navin, N. E. (2014). Cancer genomics: one cell at a time. *Genome Biology*, 15(8):452.

[109] Navin, N. E. (2015). The first five years of single-cell cancer genomics and beyond.

[110] Navin, N. E. and Chen, K. (2016). Genotyping tumor clones from single-cell data. *Nature Methods*, 13(7):555–556.

[111] Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., Bamshad, M., Nickerson, D. A., and Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276.

[112] Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., and Lau, K. W. (2012). The life history of 21 breast cancers. *Cell*, 149:994–1007.

[113] Niknafs, N., Beleva-Guthrie, V., Naiman, D. Q., and Karchin, R. (2015). SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing. *PLoS Computational Biology*, 11(10):1–26.

[114] Nilsen, G., Liestøl, K., Van Loo, P., Moen Vollan, H. K., Eide, M. B., Rueda, O. M., Chin, S.-F., Russell, R., Baumbusch, L. O., Caldas, C., Børresen-Dale, A.-L., and Lingjaerde, O. C. (2012). Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics*, 13:591.

[115] Nowell, P. C. (1976). The Clonal Evolution of Tumor Cell Populations. *Science*, 194:23–28.

[116] Oesper, L., Mahmoody, A., and Raphael, B. J. (2013). THetA : Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology*, 14(7):R80.

[117] Oliveira, M., Dienstmann, R., Bellet, M., Perez-Garcia, J. M., Gómez-Pardo, P., Muñoz-Couselo, E., Vidal, M., Ortega, V., Soberino, J., Zamora, E., Hierro, C., Ruiz, F., Nuciforo, P., Vivancos, A., Cortes, J., and Saura, C. (2016). Clonality of PIK3CA mutations (mut) and efficacy of PI3K/AKT/mTOR inhibitors (PAMi) in patients (pts) with metastatic breast cancer (MBC). *Journal of Clinical Oncology*, 34(15_suppl):528.

[118] Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572.

[119] Pachmann, K., Camara, O., Kavallaris, A., Krauspe, S., Malarski, N., Gajda, M., Kroll, T., Jörke, C., Hammer, U., Altendorf-Hofmann, A., Rabenstein, C., Pachmann, U., Runnebaum, I., and Höffken, K. (2008). Monitoring the response of circulating epithelial tumor cells to adjuvant chemotherapy in breast cancer allows detection of patients at risk of early relapse. *Journal of Clinical Oncology*, 26(8):1208–1215.

[120] Pearson, A., Smyth, E., Babina, I. S., Herrera-Abreu, M. T., Tarazona, N., Peckitt, C., Kilgour, E., Smith, N. R., Geh, C., Rooney, C., Cutts, R., Campbell, J., Ning, J., Fenwick, K., Swain, A., Brown, G., Chua, S., Thomas, A., Johnston, S. R. D., Ajaz, M., Sumpter, K., Gillbanks, A., Watkins, D., Chau, I., Popat, S., Cunningham, D., and Turner, N. C. (2016). High-level clonal FGFR amplification and response to FGFR inhibition in a translational clinical trial. *Cancer Discovery*, 6(8):838–851.

[121] Perou, C. M., Sørlie, T., Eisen, M. B., Rijn, M. V. D., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A.-L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406:747–752.

[122] Peters, B. A., Kermani, B. G., Sparks, A. B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y. T., Haas, J., Robasky, K., Zaranek, A. W., Lee, J.-H., Ball, M. P., Peterson, J. E., Perazich, H., Yeung, G., Liu, J., Chen, L., Kennemer, M. I., Pothuraju, K., Konvicka, K., Tsoupko-Sitnikov, M., Pant, K. P., Ebert, J. C., Nilsen, G. B., Baccash, J., Halpern, A. L., Church, G. M., and Drmanac, R. (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, 487(7406):190–5.

[123] Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J.-J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6:27.

[124] Pikor, L., Thu, K., Vucic, E., and Lam, W. (2013). The detection and implication of genome instability in cancer. *Cancer and Metastasis Reviews*, 32(3-4):341–352.

[125] Popic, V. (2016). LICHeE. https://github.com/viq854/lichee.

[126] Popic, V., Salari, R., Hajirasouliha, I., Kashef-Haghighi, D., West, R. B., and Batzoglou, S. (2015). Fast and scalable inference of multi-sample cancer lineages. *Genome Biology*, 16:91.

[127] Potter, N. E., Ermini, L., Papaemmanuil, E., Cazzaniga, G., Vijayaraghavan, G., and Titley, I. (2013). Single-cell mutational profiling and clonal phylogeny in cancer. *Genome Research*, 23.

[128] Prat, A., Bianchini, G., Thomas, M., Belousov, A., Cheang, M. C. U., Koehler, A., Gómez, P., Semiglazov, V., Eiermann, W., Tjulandin, S., Byakhow, M., Bermejo, B., Zambetti, M., Vazquez, F., Gianni, L., and Baselga, J. (2014). Research-Based PAM50 subtype predictor identifies higher responses and improved survival outcomes in HER2-Positive breast cancer in the NOAH Study. *Clinical Cancer Research*, 20(2):511–521.

[129] Qiao, Y., Quinlan, A. R., Jazaeri, A. A., Verhaak, R. G., Wheeler, D. A., and Marth, G. T. (2014). SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biology*, 15(8):443.

[130] R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[131] Reiter, J. G. (2017). Treeomics 1.7.3. https://github.com/johannesreiter/treeomics.

[132] Reiter, J. G., Makohon-Moore, A. P., Gerold, J. M., Bozic, I., Chatterjee, K., Iacobuzio-Donahue, C. A., Vogelstein, B., and Nowak, M. A. (2017). Reconstructing metastatic seeding patterns of human cancers. *Nature Communications*, 8:14114.

[133] Rosenberg, A. and Hirschberg, J. (2007). V-measure: a conditional entropy-based external cluster evaluation measure.

[134] Ross, E. M. (2016). OncoNEM 1.0. https://bitbucket.org/edith_ross/onconem.

[135] Ross, E. M., Haase, K., Loo, P. V., and Markowetz, F. (2017). Allele-specific multi-sample copy number segmentation. *bioRxiv*, doi:10.1101/166017.

[136] Ross, E. M. and Markowetz, F. (2016). OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biology*, 17:69.

[137] Ross, E. M. and Markowetz, F. (2017). What are Dirichlet process mixture models? *In preparation*.

[138] Roth, A. (2016). PyClone 0.13.0. https://bitbucket.org/aroth85/pyclone.

[139] Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., and Biele, J. (2014). PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398.

[140] Roth, A., McPherson, A., Laks, E., Biele, J., Yap, D., Wan, A., Smith, M. A., Nielsen, C. B., McAlpine, J. N., Aparicio, S., Bouchard-Côté, A., and Shah, S. P. (2016). Clonal genotype and population structure inference from single-cell tumor sequencing. *Nature Methods*, 13(7):573–576.

[141] Rusk, N. (2013). Single cells go fully genomic. *Nature Methods*, 10(3):190–191.

[142] Sadeh, M. J., Moffa, G., and Spang, R. (2013). Considering unknown unknowns: reconstruction of nonconfoundable causal relations in biological networks. *Journal of Computational Biology*, 20(11):920–932.

[143] Sage Bionetworks (2015). ICGC-TCGA-DREAM Somatic Mutation Calling Challenge–Tumor Heterogeneity and Evolution. https://www.synapse.org/smchet.

[144] Salehi, S., Steif, A., Roth, A., Aparicio, S., Bouchard-Côté, A., and Shah, S. P. (2017). ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome Biology*, 18:44.

[145] Savas, P., Teo, Z. L., Lefevre, C., Flensburg, C., Caramia, F., Alsop, K., Mansour, M., Francis, P. A., Thorne, H. A., Silva, M. J., Kanu, N., Dietzen, M., Rowan, A., Kschischo, M., Fox, S., Bowtell, D. D., Dawson, S. J., Speed, T. P., Swanton, C., and Loi, S. (2016). The Subclonal Architecture of Metastatic Breast Cancer: Results from a Prospective Community-Based Rapid Autopsy Program "CASCADE". *PLoS Medicine*, 13(12):e1002204.

[146] Scheinin, I., Sie, D., and Bengtsson, H. (2017). QDNAseq 1.2.4. https://bioconductor.org/packages/release/bioc/html/QDNAseq.html.

[147] Scheinin, I., Sie, D., Bengtsson, H., van de Wiel, M. A., Olshen, A. B., van Thuijl, H. F., van Essen, H. F., Eijk, P. P., Rustenburg, F., Meijer, G. A., Reijneveld, J. C., Wesseling, P., Pinkel, D., Albertson, D. G., and Ylstra, B. (2014). DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Research*, 24(12):2022–32.

[148] Schwartz, R. and Schäffer, A. A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229.

[149] Schwarz, R. F. (2016). MEDICC. https://bitbucket.org/rfs/medicc.

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

[151] Schwarz, R. F., Trinh, A., Sipos, B., Brenton, J. D., Goldman, N., and Markowetz, F. (2014). Phylogenetic Quantification of Intra-tumour Heterogeneity. *PLoS Computational Biology*, 10(4):e1003535.

[152] Sengupta, S., Wang, J., Lee, J., Müller, P., Gulukota, K., Banerjee, A., and Ji, Y. (2015). BayClone: Bayesian nonparametric inference of tumor subclones using ngs data. In *Proceedings of The Pacific Symposium on Biocomputing (PSB)*, pages 467–478.

[153] Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630.

[154] Sidow, A. and Spies, N. (2015). Concepts in solid tumor evolution. *Trends in Genetics*, 31(4):208–214.

[155] Sloane, N. J. A. (2010). The Online Encyclopedia of Integer Sequences. Published electronically at http://oeis.org. Sequence A000169.

[156] Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., Børresen-Dale, A.-L., and Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *PNAS*, 100(14):8418–23.

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

[158] Stratton, M. R., Campbell, P. J., and Andrew F, P. (2009). The cancer genome. *Nature*, 458(7239):719–724.

[159] Strino, F., Parisi, F., Micsinai, M., and Kluger, Y. (2013). TrAp: A tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Research*, 41(17):e165.

[160] Telenius, H., Carter, N. P., Bebb, C. E., Nordenskjöld, M., Ponder, B. A. J., and Tunnacliffe, A. (1992). Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics*, 13(3):718–725.

[161] Turajlic, S. and Swanton, C. (2016). Metastasis as an evolutionary process. *Science*, 352(6282):169–175.

[162] Valastyan, S. and Weinberg, R. A. (2011). Tumor metastasis: Molecular insights and evolving paradigms. *Cell*, 147(2):275–292.

[163] Van Loo, P. et al. (2017). ASCAT 2.5. https://github.com/Crick-CancerGenomics/ascat.

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

[165] Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H. S., Hodgson, J. G., James, C. D., Sarkaria, J. N., Brennan, C., Kahn, A., Spellman, P. T., Wilson, R. K., Speed, T. P., Gray, J. W., Meyerson, M., Getz, G., Perou, C. M., and Hayes, D. N. (2010). Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1):98–110.

[166] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr., L. A., and Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, 339(6127):1546–1558.

[167] Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., Multani, A., Zhang, H., Zhao, R., Michor, F., Meric-Bernstam, F., and Navin, N. E. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160.

[168] White, A. K., VanInsberghe, M., Petriv, O. I., Hamidi, M., Sikorski, D., Marra, M. A., Piret, J., Aparicio, S., and Hansen, C. L. (2011). High-throughput microfluidic single-cell RT-qPCR. *PNAS*, 108(34):13999–14004.

[169] Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., He, W., Zeng, L., Xing, M., Wu, R., Jiang, H., Liu, X., Cao, D., Guo, G., Hu, X., Gui, Y., Li, Z., Xie, W., Sun, X., Shi, M., Cai, Z., Wang, B., Zhong, M., Li, J., Lu, Z., Gu, N., Zhang, X., Goodman, L., Bolund, L., Wang, J., Yang, H., Kristiansen, K., Dean, M., Li, Y., and Wang, J. (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, 148(5):886–895.

[170] Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R. H., Eshleman, J. R., Nowak, M. A., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., and Iacobuzio-Donahue, C. A. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, 467(7319):1114–1117.

[171] Yarchoan, M., Johnson, B. A., Lutz, E. R., Laheru, D. A., and Jaffee, E. M. (2017). Targeting neoantigens to augment antitumour immunity. *Nature Reviews Cancer*, 17(4):209–222.

[172] Yates, L. R., Knappskog, S., Wedge, D., Farmery, J. H., Gonzalez, S., Martincorena, I., Alexandrov, L. B., Van Loo, P., Haugland, H. K., Lilleng, P. K., Gundem, G., Gerstung, M., Pappaemmanuil, E., Gazinska, P., Bhosle, S. G., Jones, D., Raine, K., Mudie, L., Latimer, C., Sawyer, E., Desmedt, C., Sotiriou, C., Stratton, M. R., Sieuwerts, A. M., Lynch, A. G., Martens, J. W., Richardson, A. L., Tutt, A., Lønning, P. E., and Campbell, P. J. (2017). Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell*, 32(2):169–184.e7.

[173] Yu, C., Yu, J., Yao, X., Wu, W. K., Lu, Y., Tang, S., Li, X., Bao, L., Li, X., Hou, Y., Wu, R., Jian, M., Chen, R., Zhang, F., Xu, L., Fan, F., He, J., Liang, Q., Wang, H., Hu, X., He, M., Zhang, X., Zheng, H., Li, Q., Wu, H., Chen, Y., Yang, X., Zhu, S., Xu, X., Yang, H., Wang, J., Zhang, X., Sung, J. J., Li, Y., and Wang, J. (2014). Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. *Cell Research*, 24(6):701–712.

[174] Yuan, K., Sakoparnig, T., Markowetz, F., and Beerenwinkel, N. (2015). BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biology*, 16:36.

[175] Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., Lawrence, M. S., Zhang, C.-Z., Wala, J., Mermel, C. H., Sougnez, C., Gabriel, S. B., Hernandez, B., Shen, H., Laird, P. W., Getz, G., Meyerson, M., and Beroukhim, R. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, 45(10):1134–1140.

[176] Zare, H., Wang, J., Hu, A., Weber, K., Smith, J., Nickerson, D., Song, C. Z., Witten, D., Blau, C. A., and Noble, W. S. (2014). Inferring Clonal Composition from Multiple Sections of a Breast Cancer. *PLoS Computational Biology*, 10(7):e1003703.

[177] Zeller, C., Frohlich, H., and Tresch, A. (2009). A Bayesian network view on nested effects models. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009(1):195272.

[178] Zhao, Z.-M., Zhao, B., Bai, Y., Iamarino, A., Gaffney, S. G., Schlessinger, J., Lifton, R. P., Rimm, D. L., and Townsend, J. P. (2016). Early and multiple origins of metastatic lineages within primary tumors. *PNAS*, 113(8):2140–2145.

[179] Zong, C., Lu, S., Chapman, A. R., and Xie, X. S. (2012). Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. *Science*, 338(6114):1622–1626.

Abbreviations

ADO ..... allelic dropout

BAF ..... B-allele frequency

CCF ..... cancer cell fraction

CNA ..... copy number aberration

DNA ..... deoxyribonucleic acid

DOP-PCR .... degenerate oligonucleotide-primed polymerase chain reaction

FNR ..... false negative rate

FPR ..... false positive rate

logR ..... log ratio, relative measure of sequencing depth

LOH ..... loss of heterozygosity

MDA ..... multiple displacement ampli cation

MALBAC ..... multiple annealing and looping-based amplification cycles

PCR ..... polymerase chain reaction

SNV ..... single nucleotide variant

sWGS ..... shallow whole genome sequencing

TS ..... targeted sequencing

VAF ..... variant allele frequency

WES ..... whole exome sequencing

WGA ..... whole genome amplification

Appendix A

Supplementary figures and tables

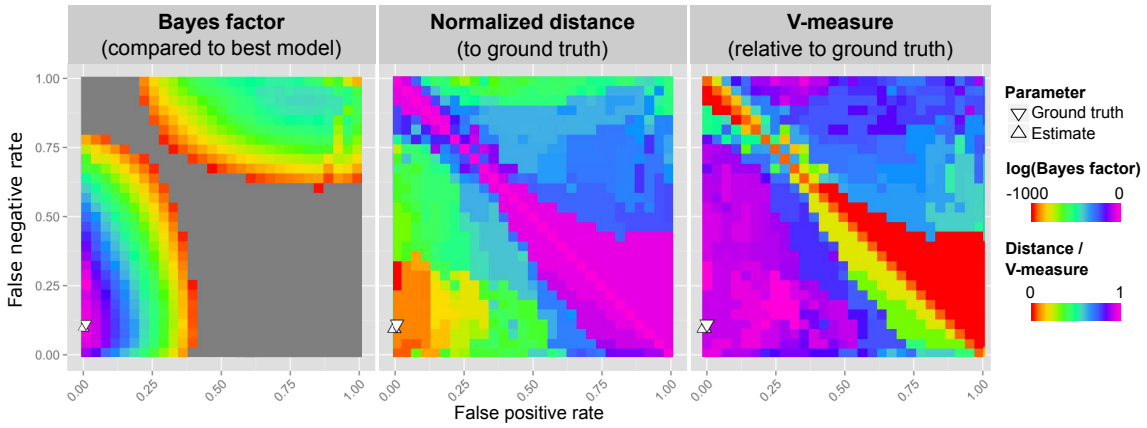


Fig. A.1 Dependence of oncoNEM on inference parameters. This second example uses a much lower FPR of 10^{-5} and shows again that (i) oncoNEM estimates error parameters that are close to the ground truth parameters and (ii) oncoNEM is robust to changes in those parameters. The left panel shows the log Bayes factor of the highest scoring model inferred with the respective parameter combination relative to highest scoring model overall. The second and the third panels show that a large range of parameter combinations around the ground truth parameters yield solutions close to the ground truth tree in terms of pairwise cell shortest-path distance and V-measure. The distance was normalised to the largest distance observed between any inferred tree and the ground truth.

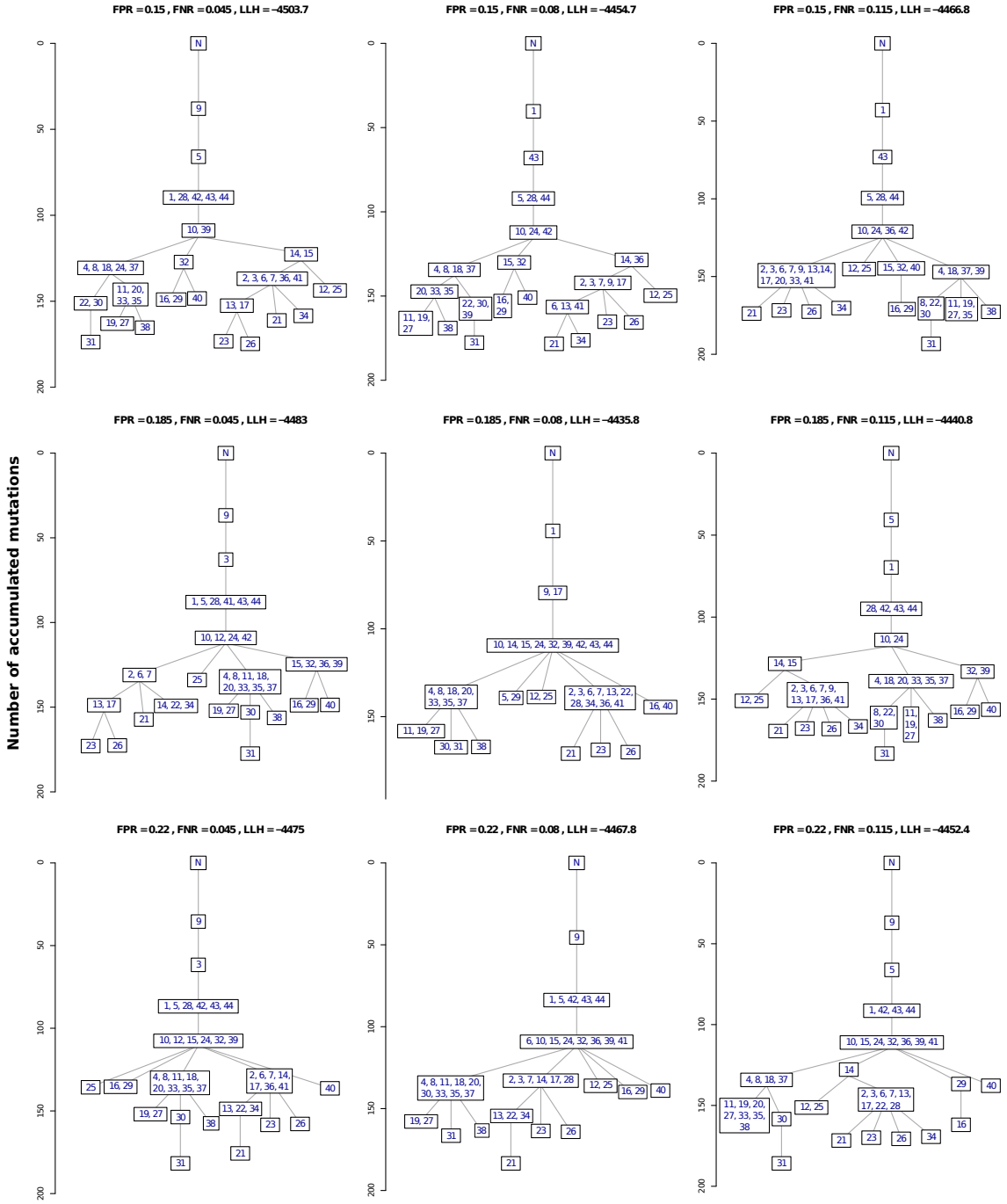


Fig. A.2 Trees inferred with estimated parameters (middle tree) and parameters that are similar to the estimated ones (outer trees) for data set by Li et al. [85]. Even if the inference parameters are varied, the overall structure and features of the oncoNEM tree are preserved.

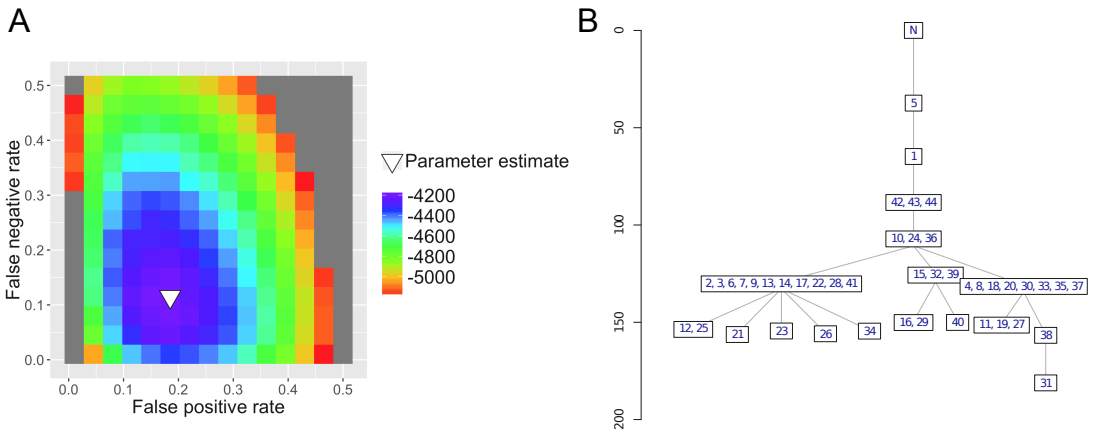


Fig. A.3 OncoNEM solution based on a subset of the mutations in the bladder cancer data set. For this analysis all mutations within regions affected by loss of heterozygosity were excluded from the data set. Genomic regions of the bladder cancer affected by loss of heterozygosity are shown in Table S1. Panel A shows the likelihood landscape with the inferred error parameters (FPR = 0.185, FNR = 0.115), which is close to the parameters estimated for the full data set. Panel B shows the inferred tree. As for the complete data set, the result suggests that initially the tumour underwent a linear evolution and then branched into two major subpopulations and some smaller ones.

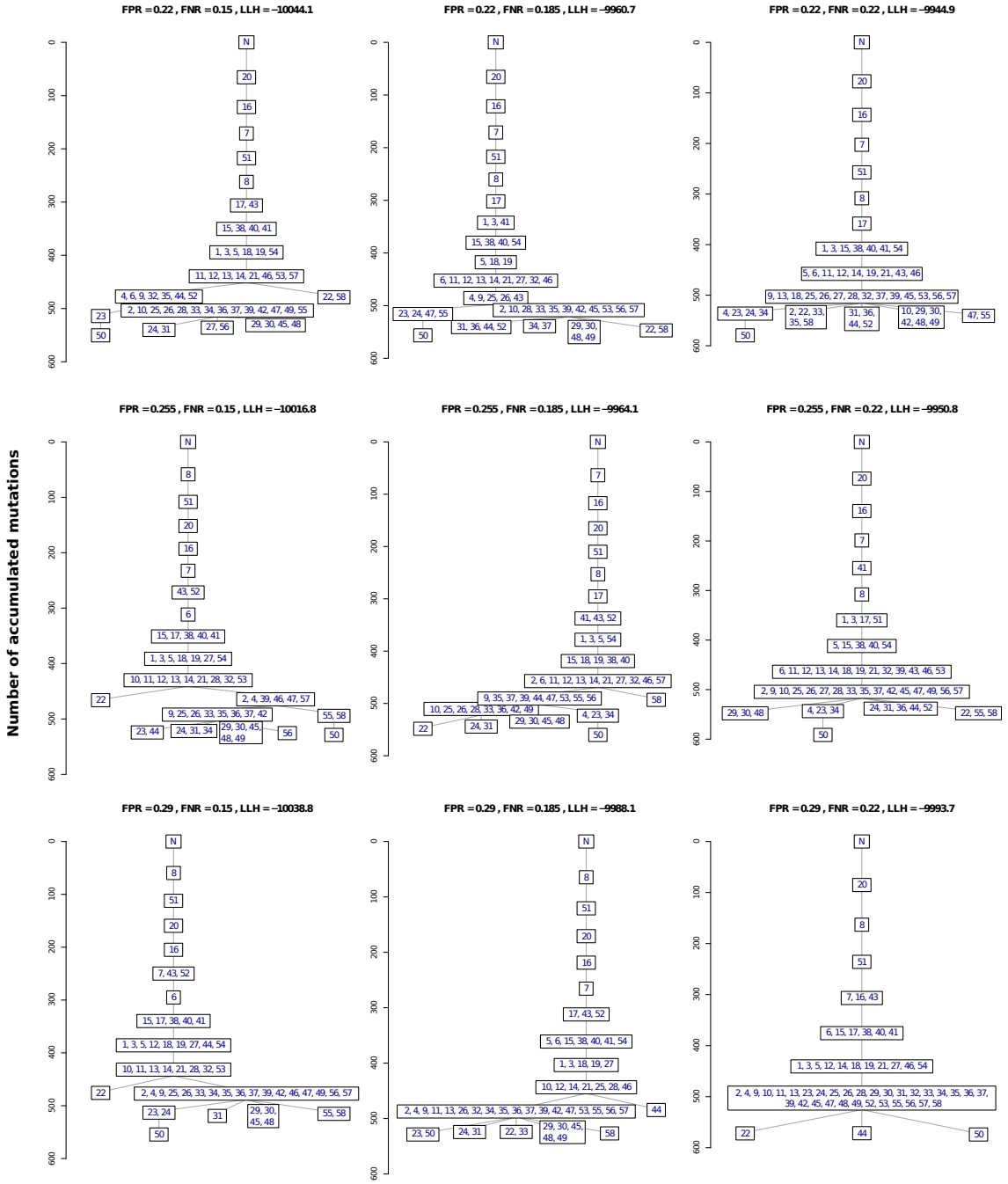


Fig. A.4 Trees inferred with estimated parameters (middle tree) and parameters that are similar to the estimated ones (outer trees) for data set by Hou et al. [58]. Even if the inference parameters are varied, the overall structure and features of the oncoNEM tree are preserved.

Table A.1 This table summarises all large genomic regions of the bladder cancer that are affected by loss of heterozygosity as shown in Figure S5 of Li et al. [85]. To assess the effect of loss of heterozygosity on the oncoNEM result, the onoNEM inference was repeated on a subset of the original mutation data containing only SNVs that lie outside these regions.

Chr	Region	Position (bp)
2	q33.3 - q37.3	>205 600 000
9	entire chromosome	
10	q25.3 - q26.3	>114 900 000
11	p	<52 900 000
22	q	>11 800 000

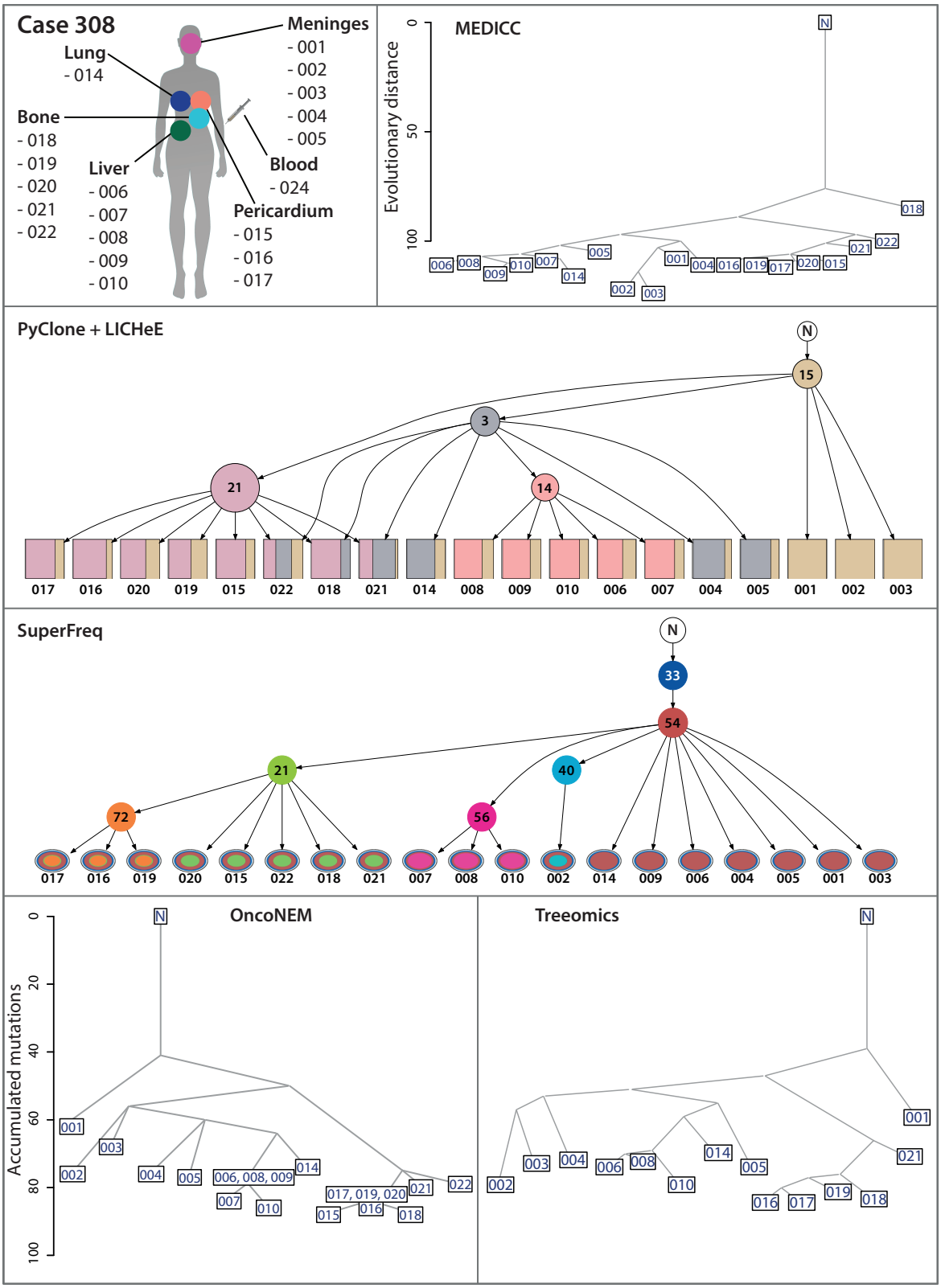


Fig. A.5 Tree inference from WES and sWGS data for case 308.