Abstract of "Computational Characterization of Heterogeneity and Rearrangements in Cancer" by Layla Oesper, Ph.D., Brown University, May 2015.

Cancer is a disease resulting from genomic mutations that occur during an individual's lifetime and cause the uncontrolled growth of a collection of cells into a tumor. These mutations range from single nucleotide variants to larger rearrangements such as copy number aberrations or structural variants that duplicate, delete or rearrange entire segments of DNA. As we enter the era of personalized medicine, where a patient's treatment may be tailored to their specific genomic architecture, accurate identification and interpretation of the set of mutations within each patient's genome is increasingly important.

Despite numerous recent advances in DNA sequencing technologies, many challenges still exist for measuring genomic mutations – especially for cancer genomes. For example, tumors often exhibit *intra-tumor heterogeneity* where individual cells in a single tumor contain different complements of mutations. Additionally, cancer genomes are highly rearranged and often contain *complex rearrangement patterns* that amplify, delete or interweave distant regions of the reference genome. These challenges confound the identification and interpretation of the complete set of mutations in a tumor. Therefore, we have developed a suite of algorithms that directly address these challenges.

With respect to intra-tumor heterogeneity, we describe several algorithms that infer the composition of heterogeneous tumors from both single and multi-sample datasets. In the case of multi-sample data, our approach is also able to infer the evolutionary history of the tumor sample. With respect to complex rearrangements, we present algorithms to find the most likely set of rearrangements present in a cancer genome, and to determine whether such a set of rearrangements occurred simultaneously or as part of a sequence of individual events. We demonstrate the advantages of our methods over competing algorithms using both simulated and real DNA sequence data. This collection of algorithms forms an initial step towards enabling improved characterization of genomic mutations in tumor samples.

Computational Characterization of Heterogeneity and Rearrangements in Cancer

by

Layla Oesper

B.A., Pomona College, 2005

Certificate, University of Wisconsin at Madison, 2010

Sc. M., Brown University, 2012

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Computer Science at Brown University

Providence, Rhode Island

May 2015

This dissertation by Layla Oesper is accepted in its present form by
the Department of Computer Science as satisfying the dissertation requirement
for the degree of Doctor of Philosophy.

Date _____ _____
Benjamin J. Raphael, Director

Recommended to the Graduate Council

Date _____ _____
Amy Greenwald, Reader

Date _____ _____
Michael L. Littman, Reader

Date _____ _____
Teresa Przytycka, Reader
(NCBI, NLM, NIH)

Approved by the Graduate Council

Date _____ _____
Peter M. Weber
Dean of the Graduate School

Vita

Layla Oesper was born in Colorado Springs, Colorado on June 11, 1983. She attended Pomona College in Claremont, California where she received a B.A in Mathematics in 2005. She then worked for Epic Systems Corporation (a company that makes electronic medical records) in Madison, Wisconsin for three years before deciding to return to school. In 2010 she received a certificate in Computer Science from the University of Wisconsin at Madison. She then joined the Computer Science department at Brown University in Providence, Rhode Island in 2010. She earned her Sc.M. in Computer Science from Brown University in 2012 while studying for her Ph.D under the advisement of Ben Raphael. During her time at Brown University she was the recipient of various awards including a National Science Foundation Graduate Research Fellowship (2011-2014) and an Anita Borg Scholarship (2014). She has also been heavily involved in a number of initiatives related to diversity in computing including organizing the Graduate Women in Computer Science (GWiCS) group and coordinating a new scholarship program that sends students to conferences related to diversity in computing. In the fall of 2015 she will start a tenure-track position as an Assistant Professor of Computer Science at Carleton College in Northfield, Minnesota.

Acknowledgements

"We have done the impossible, and that makes us mighty."

Captain Malcolm Reynolds – *Firefly*

This dissertation is only possible through the effort, collaboration and support of many people. I am deeply indebted to my co-authors on the work presented here: Ben Raphael, Gryte Satas, Ahmad Mahmoody, Anna Ritz, Sarah Aerni, Ryan Drebin, Mohammed El-Kebir, Hannah Acheson-Field and Caleb Weinreb. The many discussions, ideas and enthusiasm that each brought to the table were essential to driving this work forward.

In particular, I can not thank my adviser, Ben Raphael, enough. His drive, ingenuity and seemingly unending source of ideas still amaze me. I have learned so much from him about not only about how to do research, but also the crucial task of how to effectively present research. He has taught me to strive for excellence and to aim for more than I think is possible, because I will never know for sure what is possible until I try. The successes that I have had over the past 5 years have in no small way been due to his influence. I can only hope to be as influential to my students in future years.

I also want to acknowledge the rest of my thesis committee – Amy Greenwald, Michael L. Littman and Teresa Przytycka. They have been extremely valuable colleagues to have as I have been finishing up this thesis.

I am extremely indebted to the former and current members of the Raphael Lab: Anna Ritz, Crystal Valentine, Suzanne Sindi, Fabio Vandin, Hsin-Ta Wu, Max Leiserson, Ahmad Mahmoody, Gryte Satas, Mohammed El-Kebir, Matt Reyna and Iman Hajirasouliha. We certainly know how to

both work hard and have a good time. I will indeed miss being a part of this group. In particular, I want to acknowledge Anna Ritz for helping to guide me (and keep me sane) during my first year and more recently while I was on the job market. I also want to thank Max Leiserson for always being willing to follow through on crazy adventures like the CS Conference Chicken and the Watson Cup.

There have been many friends here at Brown who have helped to make my time here wonderful. Mike Hughes and Ryan Cabeen – we've been in this together since day one. I have really appreciated the support and friendship you have provided during the past five years. Mike – We still have to collaborate on something, someday. Betsy Hilliard – through all the sushi dinners and coffee breaks you have been with me for more adventures (and initiatives) than I can count. I look forward to seeing all that you accomplish in the coming years. To the numerous others not named here – people really are what makes a place special. You all have made my time at Brown truly special.

I would never have discovered the field of Computer Science if it wasn't for Cole Rottweiler who convinced me to take my first ever Computer Science class back at Pomona College. Who would have thought back then that I would have ever ended up here? Certainly not me. But, I guess Cole knew something that I didn't.

Both A-Staff and T-Staff in the Computer Science department here at Brown University are absolutely amazing. I cannot even imagine what this process would have been like without the help of people like Lauren Clarke, Kathy Kirman, Max Salvas, Angel Murakami and many others. I would also be remiss if I didn't thank Nathaniel Gill. Not only do you seamlessly administer everything for CCMB, you are just a joy to hang out with.

My family – Mom, Dad and Erin – you have been so supportive of me throughout this process. I am extremely grateful for your energy, thoughtfulness and willingness to listen if things were not going well. Lastly, I must say a big thank you to Y. You literally found me in a forest when I needed you, and the world has been a better place since then.

Okay, I'm ready now. Bring on the rubber chicken!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

DNA is the chemical compound that contains the instructions for the development of nearly all living organisms. For simplicity, a DNA sequence is often described as a string on a 4 character alphabet, $\{A, C, G, T\}$, representing different nucleotides/bases that form a strand of DNA. The human genome is a code consisting of $\sim$3 billion bases and is organized into 23 pairs of *chromosomes*, or subsequences. While the exact sequence of the human genome is nearly identical ($>$ 99.5% similarity) for all individuals [160], there are important difference that make each individual's genome unique to them. However, it is important to note that an individual's genome is not static, which is how cancer develops.

Cancer is a disease resulting from genomic alterations called *somatic mutations* – those that occur during the individual's lifetime – and cause the uncontrolled growth of a collection of cells into a tumor. These mutations range from *single nucleotide variants* (SNVs) where a single base pair is changed, to larger rearrangements called *structural variants* that affect whole segments of DNA (Figure 1.1). As we enter the era of personalized medicine, where a patient's treatment may be tailored to their specific genomic architecture, accurate identification of the set of mutations within each patient's genome and how they arose is increasingly important.

One of the most common ways to measure genomic mutations is through high-throughput DNA sequencing experiments. Despite numerous recent advances in such technologies, many challenges still exist for measuring genomic mutations. Many of these challenges arise from the fact that no technology currently exists that can read off the complete ordered sequence of billions of nucleotides in the human genome. Instead, most current technologies produces many millions or billions of short

reads of length ∼100-300 nucleotides - each representing a short sequence that may exist somewhere in the genome. Furthermore, these short reads contain errors and are subject to other artifacts from the sequencing process.

Further challenges arise in measuring cancer genomes specifically from high-throughput DNA sequence data. For instance, tumors often exhibit *intra-tumor heterogeneity* where different cells in a single tumor contain a different complement of mutations. Most DNA sequencing technologies sequence the genomes from a collection of cells, rather than a single genome and therefore the resulting genomic measurements reflect a mixed population where the signal to detect individual mutations is often diluted. Furthermore, many cancer genomes contain complex genomic rearranged, and often contain extensive duplicated or deleted sequences compared to the normal/healthy genome from which there were derived. Thus, there is a pressing need for new computational methods which can handle the challenges specific to to analyzing and interpreting the output of DNA sequencing experiments of cancer genomes.

Inspired by this need, this dissertation focuses on the design of algorithms that enable analysis of high-throughput DNA sequencing data of cancer genomes. In particular, we focus on designing methods that address challenges related to intra-tumor heterogeneity and characterization of complex rearrangements in cancer. Chapter 1 contains necessary background information related to somatic mutations in cancer, high-throughput DNA sequencing and intra-tumor heterogeneity. Chapter 2 describes a method to infer the composition of a heterogeneous tumor sample. Chapter 3 contains work that extends and improves the algorithm introduced in the previous chapter. Chapter 4 contains an approach for inferring the evolutionary history of a tumor when multiple sequenced samples of the tumor are available. Chapter 5 describes a method for determining the complete set of rearrangements that together best describe an entire cancer genome. Chapter 6 describes work on how to determine whether a set of rearrangements arose gradually over time or as the result of a one-time catastrophic event. Finally, Chapter 7 contains discussion and final conclusions derived from this entire body of work.

1.1 Somatic Mutations in Cancer

Cancer is a *genetic disease* resulting from genomic mutations that lead to the disfunction of multiple genes and uncontrolled growth of a collection of cells into a tumor [165]. A cancerous genome was

once a normal genome that has acquired a set of *somatic* mutations that occurred during the lifetime of the individual. These mutations differ from *germline* mutations that are inherited from an individual's parents. There are several different classes of somatic mutations that may occur in cancer.

The simplest type of mutation is a single nucleotide variant (SNV) where the character at one location has been altered. Even such a small mutation may have substantial impact on function of the cell in which the mutation occurred. For instance, the genetic code within a gene consists of 3 letter words called *codons*. Each distinct codon defines a particular amino acid, and the sequence of amino acids described in a gene is used as a blueprint for the cell to build proteins. A mutation causing a change in the coded sequence of amino acids may lead to disregulation of the function of that gene. SNVs that occur in certain genes, such as KRAS, TP53 and PIK3CA, have long been implicated in a number of different types of cancers [166].

Somatic mutations may also affect entire segments of DNA, rather that just a single base. These genomic rearrangements termed *structural aberrations* or *structural variants* (SVs) include deletions, duplications, inversions and translocations, where portions of multiple chromosomes are fused (Figure 1.1). Aberrations that result in a change in the number of copies of a particular segment of DNA within a genome are called *copy number aberrations* (CNAs) and include rearrangements such as deletions and duplications. Furthermore, cancer genomes are often *aneuploid* containing deletions or duplications of entire chromosomes [54].

Genomic rearrangements are extremely common in many types of cancers [3, 68] and are part of the focus of the work in thesis. Rearrangements may amplify genes that promote cancer (oncogenes) or delete genes that inhibit cancer development (tumor suppressor genes). In addition, rearrangements such as translocations and inversions may change gene structure or regulation and create novel fusion genes, with or without concomitant changes in copy number [3]. Classic examples are the BCR-ABL fusion gene in chronic myeloid leukemia and the activation of the MYC oncogene in Burkitt's lymphoma via a translocation [38]. Identification of other common structural aberrations is essential for understanding the molecular basis of cancer and for developing cancer-specific diagnostic markers or therapeutics such as Gleevec that targets BCR-ABL [46] or Herceptin that targets ERBB2 amplification [79]. However, these types of successes are much less common than we might hope.

Not all patients with the same type of cancer will exhibit the same sets of mutations [21]. In

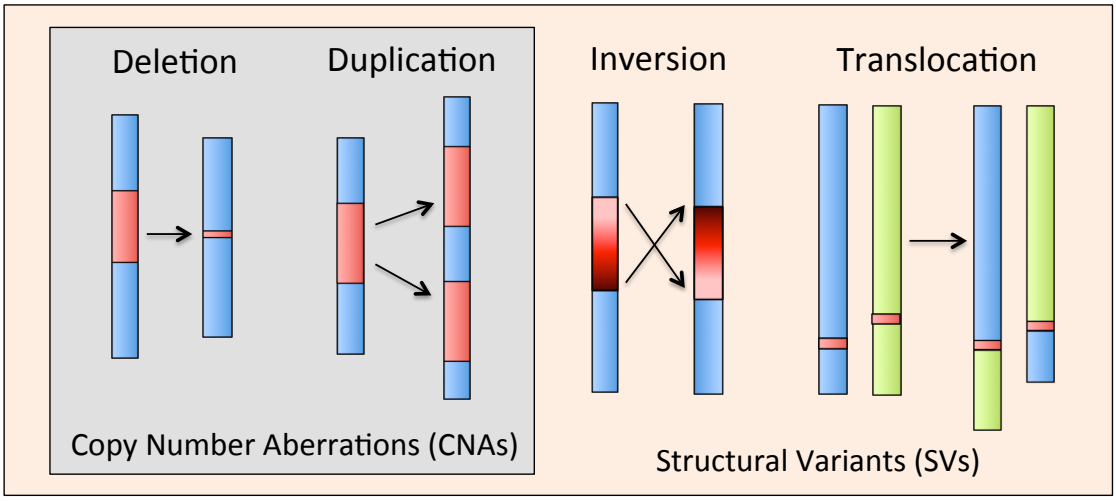


Figure 1.1: **Basic types of genomic rearrangements that occur in cancer.** Deletions include the removal of a segment of DNA, duplications are when a portion of DNA becomes copies (a *tandem duplication* is when the copy is inserted next to the original segment), an inversion is when a segment of DNA if flipped around and translocations are fusions between different chromosomes.

fact, there appears to be an incredible amount of genomic diversity among many types of cancers. This mutational heterogeneity between patients is a driving force behind the burgeoning field of *personalized medicine* or *precision medicine* where a patient may have their treatment plan tailored to their specific genomic architecture [34, 51]. However, in order for such treatments to become mainstream will require processes for accurate determination of the complete set of mutations within a single patient's genome. There are also other subsequent steps required to make personalized medicine an everyday reality, such as determining which mutations are likely driving the growth of cancer or determining whether a mutation is druggable. However, since these downstream steps rely on first identifying the mutations within a tumor, we mainly focus on this goal.

1.2 Measuring Somatic Mutations from DNA Sequencing Data

Measuring the set of somatic mutations that exist within a single tumor sample is not a straightforward task. No technology currently exists that can measure the entire ~3 billion base pair long human genome. Instead, modern high-throughput DNA sequencing technologies provided by

companies such as Illumina, Ion Torrent, Pacific Biosciences, 454 and others, produce many millions-billions of short reads (<1000bp). The *coverage* of a sequencing experiment refers to the average number of times that any position in the genome is sequenced (contained within a short read). Higher coverage sequencing can yield more accurate information about the genome, but with the tradeoff of higher cost. We consider two main categories for sequencing experiments produced by these technologies - whole-genome sequencing (WGS) where the entire genome is sequenced, usually at modest ∼40X coverage, and whole-exome sequencing (WXS) where only the coding region of the genome is targeted but at higher ∼100X coverage. Alternatively, ulta-deep targeted sequencing where only a small portion of a genome is of interested is sequenced to very high coverage may be used. This is typically used as a way to validate mutations predicted using a lower coverage sequencing protocol. These technologies have allowed for the measurement of thousands of cancer genomes through collaborative projects such as The Cancer Genome Atlas (TCGA) [25] or the International Cancer Genome Consortium (ICGC) [74]. However, the challenge of how to extract information from the many short reads produced by these technologies is an ongoing challenge, which requires the continual development of new algorithms.

Tumor samples are generally sequenced in tandem with a matched normal sample for the same patient, which can later be used to distinguish somatic mutations from germline mutations. The most common approach is to sequence a single sample from a tumor. However, several recent studies have sequenced multiple samples of a tumor through either multi-sectioning [52, 185] or by collecting samples at different time points [147]. As cancer genomes are highly rearranged, they are generally analyzed using a *resequencing* approach that relies on alignment of DNA sequence reads to the human reference genome. It is from this alignment data that the presence of somatic mutations in the cancer samples is inferred.

SNVs may be identified by analyzing the set of reads with alignments to the reference genome containing a common loci. Since read alignments generally allow mismatches between the read and reference genome, some fraction of the aligned reads may indicate an alternate allele at the loci. If "enough" reads indicate a different allele, an SNV may be inferred. Somatic SNVs are then identified when the alignment data for the tumor data indicates the presence of a mutation not found in the matched normal sample. A number of algorithms have been developed to infer the presence of somatic SNVs in DNA sequencing data of tumor samples [82, 35, 142, 146].

A *paired-end sequencing* protocol generates reads from both ends of a longer fragment (or insert)

which allows for the detection of all types of somatic structural variants [139, 151, 150, 133, 31]. Paired-end mapping [163, 84], or End Sequencing Profiling [167, 136], aligns paired reads from a cancer genome to the reference human genome. The distance between the aligned reads is computed. If this *aligned distance* is close to the expected length of the sequenced fragments, as determined by the distribution of fragment lengths, the aligned pair of reads is referred to as a *concordant pair*. If the aligned distance is far from the expected fragment length (either shorter or longer) or if the orientation of the aligned reads has changed, then the aligned pair is referred to as a *discordant pair*. Clusters of discordant pairs reveal novel adjacencies (or *breakpoints*) created by somatic structural aberrations [136]. Figure 1.2 shows an overview of DNA sequencing and rearrangement detection from paired-end sequencing.

The process of detecting somatic mutations from aligned reads is potentially confounded by a number of challenges as numerous errors and artifacts are introduced during both the sequencing and alignment processes. These include, but are not limited to: optical PCR duplicates, GC-bias, strand bias (where reads that indicate a possible mutation only align to one strand of the reference), sequencing errors, and alignment artifacts due to regions of low complexity or repetitive DNA sequence [15, 44]. Additional challenges arise for cancer genomes specifically, which we discuss further in the following section.

1.3 Challenges to Detecting and Interpreting Somatic Mutations in Cancer

There are a number of things that make detecting and interpreting somatic mutations in cancer from DNA sequence data particularly difficult. We describe two such challenges in the following subsections.

1.3.1 Intra-Tumor Heterogeneity

One important challenge in identifying and characterizing somatic mutations from high-throughput DNA sequence data is that most tumor samples are heterogeneous, containing admixture with normal cells and potentially multiple distinct populations of tumor cells. The clonal theory of cancer [116] posits that all cancerous cells in a tumor descended from a single cell and that mutations in all tumor

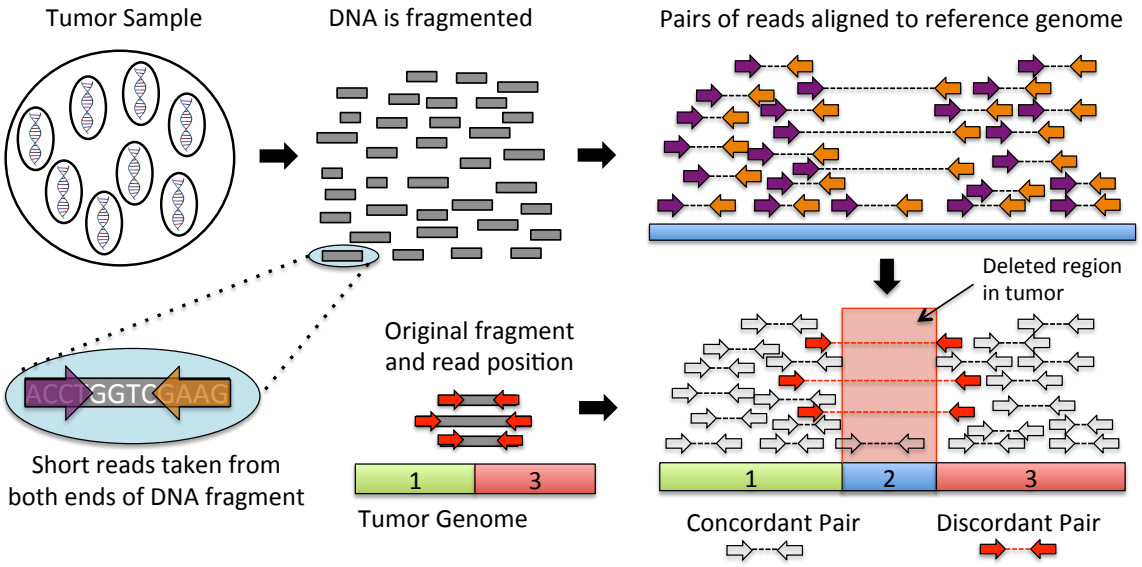


Figure 1.2: **Overview of DNA sequencing and rearrangement (novel adjacency) detection from paired-end sequencing.** DNA from multiple cells is fragmented and reads are sampled from both ends of a fragment and aligned to a reference genome. Read pairs where either the aligned distance is far from the expected fragment length or the orientation of reads is altered are called discordant. Otherwise, read pairs are called concordant. Clusters of discordant pairs indicate rearrangements in the tumor genome, not present in the reference.

cells are either *clonal*, having been inherited from the most recent common ancestor of all tumor cells, or *subclonal* if they occurred later in evolution of the tumor and only exist in a subset of tumor cells (Figure 1.3). Alternatively, subclonal mutations may suggest that the tumor is polyclonal, consisting of subpopulations of cells that are not all descended from a single founder cell [124].

Most cancer genome sequencing studies generate data from a bulk tumor sample that contains both normal cells and one or more subpopulations of tumor cells, each containing a different complement of somatic mutations. This *intra-tumor heterogeneity* complicates the identification of all types of somatic mutations and requires specialized methods to quantify the extent and character of heterogeneity in a sample. The simplest form of intra-tumor heterogeneity is admixture by normal cells and *tumor purity* is defined as the fraction of cells in the sample that are cancerous. A sequenced read from a tumor sample represents the genomic sequence in the cell, or subpopulation of cells, from which the read was derived. Thus, lower tumor purity results in a reduction in the number of sequence reads derived from the cancerous cells, and thus a reduction in the signal to detect somatic mutations. As a result accurate identification of tumor purity may be helpful for identification of all types of mutations and many algorithms either implicitly or explicitly require an

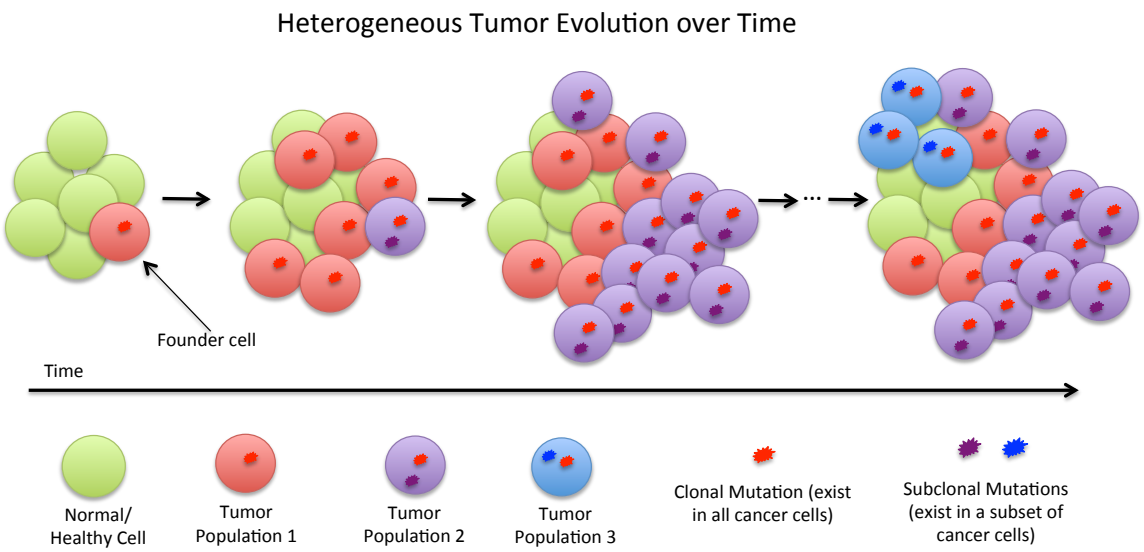


Figure 1.3: **Evolution of a heterogenous tumor.** A tumor evolves after an initial mutation in a founder cell and, in this example, leads to three distinct tumor populations (red, purple and blue) at present time. All tumor populations include a single clonal aberration (red) which exists in all tumor cells. Tumor Population 2 and Tumor Population 3 are subclonal populations, containing the purple and blue subclonal mutations, respectively, which are present in only a subset of tumor cells in the sample.

estimate of tumor purity [82, 145].

Furthermore, it is desirable to identify subclonal aberrations which can provide information on the age or history of the tumor [149] and can yield further insight into tumors that fail treatment or metastasize [149, 56]. As a result there has been major interest in recent years in algorithms that can deconvolve information about intra-tumor heterogeneity from high-throughput DNA sequencing data [177]. There has also been increased interest in not only inferring the different populations of tumor cells that exist within a tumor sample, but also inferring the evolutionary history underlying the tumor which lead to its heterogeneous state [75, 47, 64, 155]. Being able to infer such information about the history a tumor may lead to further insights into how tumors develop.

1.3.2 Complex Genomic Rearrangements

Cancer genomes are aneuploid, contain extensive duplicated sequences, and are highly rearranged compared to the germline genomes from which they were derived (Figure 1.4). The organization of amplified regions in cancer genomes is often highly complex with many high copy amplicons from

distant parts of the reference genome co-localized on the cancer genome [137, 59]. This complex organization is not only difficult to interpret, but also confounds the analysis of other "simpler" mutations such as SNVs [141].

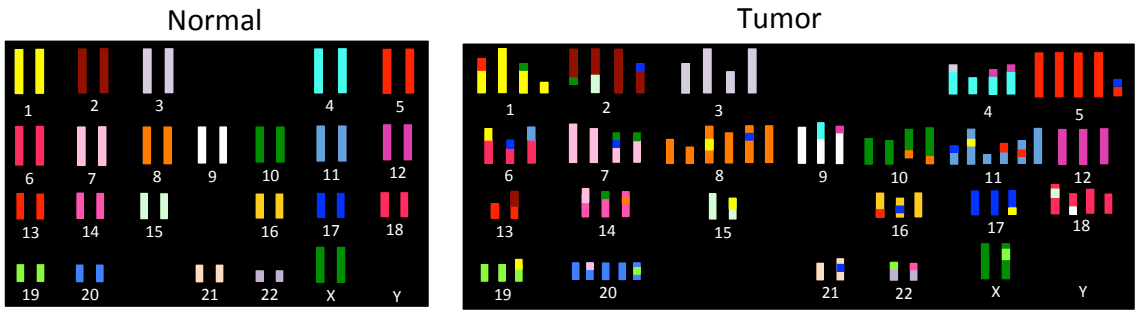


Figure 1.4: **Schematic of the chromosomal structure in both a normal genome and a tumor genome.** In the normal genome (left) every chromosome is "painted" a different color. The tumor genome (right) is colored in the same manner as the normal genome and contains many complex rearrangements including the amplification or loss of entire chromosomes.

Beyond the identification of which mutations exist within a single patient's genome, it can also be important to determine the order in which mutations arose, as earlier mutations may be more likely to have contributed to tumor growth. Closely related is the difficult task of determining whether rearrangements occurred sequentially over many cell divisions, or nearly simultaneously as part of either a *chromothripsis* [154] or *chromoplexy* [11] event where a portion of one or more chromosomes shatter and are randomly stitched back together. The more we are able to understand about how tumors develop, the better prepared we will be when it comes to understanding how to combat cancer.

1.4 Summary of Contributions

In this section we overview the research contributions described in this thesis. All points are expanded in further details in the following chapters.

Chapter 2: Quantifying Intra-Tumor Heterogeneity

In Chapter 2 we present an algorithm called Tumor Heterogeneity Analysis (THetA) [118, 117] which infers the most likely collection of genomes and their proportions in a heterogeneous tumor

sample, for the case where copy number aberrations distinguish tumor subpopulations.

Contributions:

1. We develop a multinomial model of DNA sequence data for a heterogeneous tumor sample that directly incorporates changes in genome length due to deleted or duplicated sequence.
2. Unlike many other methods, our multinomial model allows for *any number* of tumor subpopulations within a single tumor.
3. We present an efficient algorithm for inferring tumor purity and copy number aberrations in the important instance where a tumor sample consists of normal cells and a single tumor population.
4. We apply THetA to simulated data and 3 breast cancer genomes from [114] and demonstrate that THetA not only does favorably compared to other methods but also is able to accurately identify multiple tumor populations from real sequence data.

Chapter 3: An Improved Approach to Quantifying Intra-Tumor Heterogeneity

In Chapter 3 we present THetA2 [121], which extends the THetA algorithm discussed in Chapter 2 in several important directions.

Contributions:

1. We describe a graph-based approach which results in an algorithm that is $> 1000X$ faster than the original for the case of multiple distinct subpopulations in a tumor sample.
2. We extend THetA2 to infer tumor composition for highly rearranged genomes using a two-step procedure where initial estimates are made using high confidence regions of the genome, and then are extended to the entire genome.
3. We devise a probabilistic model of B-allele frequencies, which can be used to solve the identifiability issue when read depth alone is consistent with multiple possible tumor compositions.
4. We extend THetA2 to analyze the cheaper and more widely available whole-exome sequencing data.

Chapter 4: Inferring Tumor Evolution from Multi-Sample Data

In Chapter 4 we present an algorithm called AncesTree [47] that infers the clonal evolution of a tumor from multi-sample sequencing data.

Contributions

1. We formalize the problem of reconstructing the clonal evolution of a tumor using single-nucleotide mutations as the Variant Allele Frequency Factorization Problem (VAFFP).
2. We derive a combinatorial characterization of the solutions to the VAFFP.
3. We derive an integer linear programming solution to the VAFFP in the case of error-free data and extend this solution to real data with a probabilistic model for errors.
4. We apply our AncesTree algorithm both simulated and real sequencing data and demonstrate that our approach outperforms several competing methods.

Chapter 5: Reconstructing Cancer Genome Organization

In Chapter 5 we present an algorithm called Paired-End Reconstruction of Cancer Genome Organization (PREGO) which identifies the most likely set of structural rearrangements that describe an entire cancer genome [120, 119].

Contributions:

1. We formulate the Copy Number and Genome Adjacency Reconstruction problem of determining the most likely sequence of genomic segments from the normal genome that spells out the cancer genome.
2. We introduce the PREGO algorithm which uses two different signals obtained from DNA sequence data to solve a version of the Copy Number and Genome Adjacency Genome Reconstruction problem. PREGO finds the most likely set of variants and their organization that best describe the entire cancer genome, in contrast to other algorithms which consider all variants individually.
3. We demonstrate that PREGO is a polynomial time algorithm through reduction to a network flow problem on a bi-directed graph.

4. We apply PREGO to 5 Ovarian Cancer genomes from The Cancer Genome Atlas (TCGA) and 6 Breast Cancer genomes from [114]. We identify numerous rearrangements and biological events in these genomes.

Chapter 6: Detecting Simultaneous Rearrangements in Cancer Genomes

In Chapter 6 we present analysis of one signature of chromothripsis suggested by Korbel *et al.* [83] followed by two additional measures, Open Adjacency Rate (OAR) and Copy-number Asymmetry Enrichment (CAE), which compute a lower bound on the fraction of rearrangements in a sample that occurred as part simultaneous event [170].

Contributions

1. We present a formal model of chromothripsis and use it to demonstrate potential issues with one signature of chromothripsis suggested by Korbel *et al.* [83].

2. We derive a relationship between the comparative genomics notion of a k-break and simultaneous events in cancer (e.g. chromothripsis, chromoplexy).

3. We introduce the notion of open and closed adjacencies and show that open adjacencies may be used as a proxy for identifying rearrangements that occurred simultaneously.

4. We describe two measures, OAR and CAE, which allow us to compute a lower bound on the fraction of rearrangements in a sample that occurred as part simultaneous event.

Chapter 2

Quantifying Intra-Tumor Heterogeneity

Tumor samples are typically heterogeneous, containing admixture by normal, non-cancerous cells and one or more subpopulations of cancerous cells. Whole-genome sequencing of a tumor sample yields reads from this mixture, but does not directly reveal the cell of origin for each read. In this chapter we present THetA (Tumor Heterogeneity Analysis), an algorithm that infers the most likely collection of genomes and their proportions in a tumor sample, for the case where copy number aberrations distinguish subpopulations.

We apply THetA to simulated data, and three breast cancer genomes that were sequenced and extensively analyzed in [114]. We find that THetA successfully estimates normal admixture and recovers clonal and subclonal copy number aberrations in real and simulated sequencing data. In particular we show that THetA is able to accurately identify subclonal aberrations even from the modest sequencing coverage ($\sim$ 30-40X) that is now the standard in many cancer studies.

Much of the completed work from this chapter is taken from [118] and [117] and was originally presented at the 17^{th} Annual International Conference on Research in Computational Molecular Biology (RECOMB).

2.1 Related Work

As discussed in Chapter 1 a tumor is a heterogeneous population of cells, each cell potentially containing a different complement of somatic mutations (see Figure 1.3). These include both *clonal* mutations which exists in all tumors cells and *subclonal* mutations that exist in a subset of tumor cells. DNA sequencing technologies now enable an unprecedented view of this "intra-tumor" mutational heterogeneity [42]. However, nearly all recent cancer sequencing projects generate DNA sequence from tumor samples consisting of many cells – including both normal (non-cancerous) cells and one or more distinct populations of tumor cells. The *tumor purity* of the sample is the fraction of cells in the sample that are cancerous, and not normal cells. Lower tumor purity reduces the power to detect all types of somatic aberrations in cancer genomes. For example, lower tumor purity attenuates copy number ratios or allele frequencies away from values expected by integral copy numbers. Methods to detect somatic copy number aberrations or loss of heterozygosity (LOH) from SNP array data or array comparative genomic hybridization (aCGH) data must account for this issue [49, 9, 180, 10, 57, 123]. In addition, many algorithms for identifying somatic single-nucleotide mutations from DNA sequence reads implicitly or explicitly rely on an estimate of tumor purity. For example, the VarScan 2 program [82] uses an estimate of tumor purity as input to calibrate the expected number of reads that contain a somatic mutation at a locus.

Traditionally, tumor purity was assessed by visual analysis of tumor cells, either manually by a pathologist or via image analysis [182]. Recently a number of algorithmic methods have been developed to infer tumor purity directly (see Table 2.1). Two widely used methods are ASCAT [164] and ABSOLUTE [27] which were introduced to estimate tumor purity directly from SNP array data, a different technology that measures polymorphisms in a sample. Both of these methods utilize the presence of copy number aberrations in cancer genomes to estimate both tumor purity and *tumor ploidy*, the number of copies of segments of (or entire) chromosomes. Tumor purity and tumor ploidy are intertwined; e.g. a heterozygous deletion of one copy of a chromosome in a 100% pure tumor sample (containing one cancer genome) could also be explained as a homozygous deletion in a 50% pure tumor sample (containing one cancer genome). Thus, it is necessary to estimate tumor purity and ploidy simultaneously, but this is a subtle and difficult problem. ASCAT and ABSOLUTE address this problem by estimating the *average ploidy* over the entire cancer genome. These estimates of tumor purity and average ploidy are then used in a second step to derive copy

number aberrations.

Signal	Method	Datatype	Multiple Tumor Populations	Sample Type
SNV	SciClone [105]	WXS	Yes	Single
	PyClone [143]	Deep	Yes	Single
	EXPANDS [8]	Deep	Yes	Single
	PurBayes [87]	HTS	Yes	Single
	PurityEst [157]	HTS	No	Single
	PhyloSub [75]	Deep	Yes	Multiple
	TrAp [155]	WXS	Yes	Single
	CITUP [98]	Deep	Yes	Multiple
	AncesTree [47]	**WGS, WXS**	**Yes**	**Multiple**
CNA	**THetA [118]**	**WGS**	**Yes**	**Single**
	THetA2 [121]	**WGS, WXS**	**Yes**	**Single**
	ABSOLUTE [27]	SNP Array, WXS	No	Single
	ASCAT [164]	SNP Array	No*	Single
	CNAnorm [62]	WGS (low-pass)	No	Single
CNA,SNV	AbsCN-seq [12]	WXS	No	Single
	CloneHD [48]	HTS	Yes	Multiple
	PhyloWGS [39]	WGS	Yes	Multiple
CNA, LOH	PyLOH [91]	HTS	No	Single

Table 2.1: **Publicly available methods for inferring tumor purity and tumor subpopulations.** Signal refers to the type(s) of signal from an experiment used to estimate tumor purity and populations. Datatype indicates the type of data that the method is designed to use. WGS is whole-genome, WXS is whole-exome and HTS is high-throughput sequencing, which is used when the type is not specified. We also list whether a method is capable of inferring multiple tumor populations and if it works with only single or multiple tumor samples. Methods by the author of this thesis are in bold. *[114] used an extension of ASCAT to consider exactly two tumor subpopulations, but did not generalize the extension to more tumor subpopulations.

Both ASCAT and ABSOLUTE were shown to yield accurate estimates of tumor purity, achieving in some cases better estimates than via pathology or other techniques. However, these methods also have important limitations. First, the mathematical models used by ASCAT and ABSOLUTE are optimized for SNP array data, as we detail below. While these methods may be adapted to run on DNA sequencing data (e.g. for ABSOLUTE see [86] and for ASCAT see below), the underlying mathematical model used by both methods does not adequately describe the characteristics of sequencing data. Second, both of these methods apply various heuristics in their estimation procedures, such as rounding copy numbers to the closest integer [164] and do not directly infer integer copy numbers for each segment of the genome during the estimation. Finally, both methods do not explicitly identify *multiple* tumor subpopulations, and instead infer only a single tumor subpopulation. For example, ABSOLUTE [27] classifies copy number aberrations as outliers if they are not

clonal, but does not refine these outliers into subpopulations. If a tumor sample consists of multiple tumor subpopulations, then considering only a single tumor population may yield inaccurate estimates of tumor purity, as we show below.

High-throughput DNA sequencing data is much higher resolution data than SNP arrays, and provides the opportunity to derive highly accurate estimates of *both* tumor purity *and* the composition of tumor subpopulations. For example, the number of reads containing a somatic single-nucleotide mutation at a locus provides – in principle – an estimate of the fraction of cells in a tumor sample containing this mutation. However, three interrelated factors complicate this analysis: (1) The number of reads supporting a somatic single-nucleotide mutation has high variance, implying that an estimate of the allele frequency will be highly unreliable at the modest coverages (30-40X) employed in nearly all current cancer sequencing projects. (2) Somatic mutations may be present in only a fraction of tumor cells. (3) Somatic copy number aberrations (nearly ubiquitous in solid tumors) alter the number of copies of the locus containing the mutation. While the first issue might be addressed in part by clustering allele frequency estimates across the genome [157, 148, 41, 143] the second and third issues complicate such a clustering. Recent methods for analyzing tumor composition from DNA sequencing data either ignore copy number aberrations [41] or use iterative approaches [62] or other approximations [164, 27], and do not formally model the generation of DNA sequencing data from a mixture of integral copy numbers for each genomic segment.

Beyond the estimation of tumor purity and ploidy, it is desirable to identify subclonal aberrations, which can provide information on the age or history of the tumor [149] and can yield further insight into tumors that fail treatment or metastasize [149, 56, 56]. However, even with a pure tumor sample, characterizing subclonal mutations is a challenge. Tolliver *et al.* [161] infer subclonal copy number aberrations by comparing aberrations across different individuals, thus assuming that the progression of somatic copy number aberrations is conserved across individuals. Gerlinger *et al.* [53] recently demonstrated the extent of subclonal mutations by sequencing multiple (spatially separated) samples from a tumor, complementing earlier studies of heterogeneity using microarray-based techniques [110]. In another approach, [41] used a targeted ultra-deep sequencing (1000X coverage) approach to estimate allele frequencies for relapse mutations in AML. In another recent study, Nik-Zainal *et al.* [114] used a SNP array based estimate of tumor purity [164] followed by extensive manual analysis of somatic mutations to identify a clonal (majority) population and

a number of subclonal populations in each of several breast cancer genomes. Ultimately, single-cell sequencing techniques promise to provide a comprehensive view of cancer heterogeneity [109, 176, 71, 14], but these techniques presently require specialized DNA amplification steps that can introduce artifacts and also incur costs for sequencing many cells. Thus, the problem of *simultaneous* estimation and correction for tumor purity as well as identification of clonal/subclonal mutations will remain a challenge for the majority of cancer sequencing projects.

In this Chapter, we introduce Tumor Heterogeneity Analysis (THetA), an algorithm that infers the most likely collection of genomes and their proportions from high-throughput DNA sequencing data, in the case where copy number aberrations distinguish subpopulations. In contrast to existing methods, we formulate and optimize an explicit probabilistic model for the generation of the observed tumor sequencing data from a mixture of a normal genome and one *or more* cancer genomes, each genome containing integral copy numbers of its segments. Specifically, we derive and solve the Maximum Likelihood Mixture Decomposition Problem (MLMDP) of finding a collection of genomes – each differing from the normal genome by copy number aberrations – whose mixture best explains the observed sequencing data. Thus, we generalize the problem of estimating tumor purity to the problem of determining the proportions of normal cells and *any number* of tumor subpopulations in the sample.

Our formulation and solution of the MLMDP leverages the fact that copy number aberrations create a strong signal in DNA sequencing data: even relatively small copy number aberrations cause deviations in the alignments of thousands-millions of reads. Thus, in contrast to single-nucleotide mutations, where there is high variance in the number of reads at each position, many measurements (reads) are perturbed for each copy number aberration. Thus, each copy number aberration provides many data points for deconvolution of the tumor genome mixture. We show how to solve the MLMDP as a collection of convex optimization problems. THetA is the first algorithm – to our knowledge – that automatically identifies subclonal copy number aberrations in whole-genome sequencing data from mixtures of more than two genomes. Moreover, in the case of admixture between a single (clonal) cancer population and normal cells, THetA runs in polynomial-time, providing the first rigorous and efficient algorithm for simultaneously estimating tumor purity and inferring integral copy numbers.

2.2 The THetA Algorithm

In this section we describe the Tumor Heterogeneity Analysis (THetA) algorithm.

2.2.1 The Maximum Likelihood Mixture Decomposition Problem (MLMDP)

We first formulate the Maximum Likelihood Mixture Decomposition Problem of finding the most likely mixture of tumor cells populations from a sequenced tumor sample. We assume that sequenced reads from a tumor sample are aligned to the reference human genome, the first step in cancer genome sequencing analysis [103, 43]. Typically a matched normal genome is also sequenced to distinguish somatic mutations from germline variants. We focus on copy number aberrations in order to estimate tumor purity and subpopulations. Thus, we assume that a cancer genome differs from the reference genome by gains and losses of segments, or intervals, of the reference genome. These intervals are identified by examining the density, or depth, of reads aligning to each location in the reference [174, 33, 104], and/or by clustering of discordant paired reads that identify the breakpoints of copy number aberrations or other rearrangements [136, 58, 120, 99, 31, 13]. Following this analysis, one obtains a partition the reference genome into a sequence $\mathbf{I} = (I_1, \ldots, I_m)$ of non-overlapping intervals. We represent a cancer genome by an *interval count vector* $\mathbf{c} = (c_1, \ldots, c_m)$, where $c_j \in \mathbb{N}$ is the integer number of copies of interval I_j in the cancer genome. From the sequencing of a tumor sample, we observe a *read depth vector* $\mathbf{r} = (r_1, \ldots, r_m)$, where $r_j \in \mathbb{N}$ is the number of reads with a (unique) alignment within I_j.

A tumor sample is a mixture of cells that contain different collections of somatic mutations, and in particular somatic copy number aberrations. We assume that the tumor sample is a mixture of n subpopulations, including a subpopulation of normal cells and one or more subpopulations of cancer cells. Each subpopulation has a distinct interval count vector representing the genome of the subpopulation. Thus, we represent a tumor sample $\mathcal{T}$ by: (1) an $m \times n$ *interval count matrix* $\mathbf{C} = [c_{jh}]$, where $c_{jh} \in \mathbb{N}$ is the number of copies of interval I_j in the h^{th} distinct subpopulation; and (2) a *genome mixing vector* $\mu \in \mathbb{R}^n$ where μ_h is the fraction of cells in $\mathcal{T}$ from the h^{th} subpopulation. Given a read depth vector $\mathbf{r}$ derived from the sequence of $\mathcal{T}$, our goal is to identify the underlying interval count matrix $\mathbf{C}$ and genome mixing vector μ that best describe $\mathbf{r}$ (Figure 2.1). We formulate the following problem.

Maximum Likelihood Mixture Decomposition Problem (MLMDP)**.** *Given an interval partition* $\mathbf{I}$ *of a reference genome and an associated read depth vector* $\mathbf{r}$ *derived from a tumor sample* $\mathcal{T}$*, find the underlying interval count matrix* $\mathbf{C}$ *and genome mixing vector* μ *that maximize the likelihood* $P(\mathbf{r}|\mathbf{C}, \mu)$.

2.2.2 Intervals and Counts: Probability Model

In this section we derive the probability $P(\mathbf{r}|\mathbf{C}, \mu)$ in the Maximum Likelihood Mixture Decomposition Problem (MLMDP). In brief, under the usual assumptions for DNA sequencing, the probability p_j that a reads that aligns to an interval I_j is equal to the fraction of the total DNA in the sample originating from interval I_j. Hence, the probability $P(\mathbf{r}|\mathbf{C}, \mu)$ of the observed read depth vector $\mathbf{r}$ follows a multinomial distribution determined by the interval count matrix $\mathbf{C}$ and genome mixing vector μ.

Single Genome

Following the usual assumptions (e.g. the Lander-Waterman model), we assume that the starting positions of reads in the cancer genome are uniformly distributed over its length. The probability of a read from the cancer genome aligning to an interval I_j in the reference genome depends on: (i) the number of copies of the interval in the cancer genome; (ii) the length of the interval; (iii) possible difficulties in aligning reads to I_j due to repetitive sequence or other effects. We first describe the model under the simplifying assumption that there are no alignment difficulties and all the intervals in $\mathbf{I}$ are of equal length, which without loss of generality we set to length 1. Below we show how to remove these restrictions by incorporating an *interval weight vector* $\mathbf{w}$ into the model that assigns a weight to each interval in proportion to its length or mappability. Let $\mathbf{c} = (c_1, \ldots, c_m)$ be the (unknown) number of copies of each interval in the cancer genome. Then the probability p_j that a read aligns to I_j satisfies $p_j = \frac{c_j}{|\mathbf{c}|_1}$, where $|\mathbf{c}|_1 = \sum_{j=1}^{m} c_j$ is the ℓ_1-norm of $\mathbf{c}$. We use the notation $\widehat{\mathbf{x}} = \frac{\mathbf{x}}{|\mathbf{x}|_1}$ to denote a normalized vector. Thus, the observed read depth vector $\mathbf{r}$ is the result of $r = \sum_{j=1}^{m} r_j$ independent draws from a multinomial distribution with parameter $\mathbf{p} = \widehat{\mathbf{c}}$; i.e. $\mathbf{r} \sim Mult(r; \widehat{\mathbf{c}})$.

We emphasize that the multinomial distribution models the fact that the number of reads aligning to each interval are *not independent* random variables, but rather are dependent on the number of copies (ploidy) of each interval in the cancer genome(s). As the total number of reads gets

extremely large, the multinomial distribution will converge to independent Poisson distributions in each interval. However, even with the millions-billions of reads produced by high-throughput DNA sequencing, the effects of a finite number of reads are an issue in cancer genome sequencing. This is because large copy number changes – including gain and loss of whole-chromosomes – are common in cancer genomes. A large deletion/duplication will affect the number of reads observed in other intervals; e.g. if we consider the 22 autosomes, a homozygous deletion of chromosome 1 will reduce the effective length of the cancer genome by 8.65%, altering the expected number of reads observed in other intervals. See Appendix A.1.1 for both simulations and real data experiments further motivating the use of the multinomial model.

Mixture of Genomes

Now suppose we sequence a tumor sample $\mathcal{T}$ and align the obtained reads to the reference genome, observing a read depth vector $\mathbf{r} = (r_1, \ldots, r_m) \in \mathbb{N}^m$. Let $\mathbf{C} = [c_{jh}]$ be the (unknown) interval count matrix and μ be the (unknown) genome mixing vector for the tumor sample. Here μ is required to be an element of the unit $(n-1)$-simplex $\Delta_{n-1} = \{(\mu_1, \ldots, \mu_n)^T \in \mathbb{R}^n \mid \sum_{h=1}^{n} \mu_h = 1, \text{ and } \mu_h \geq 0 \text{ for all } h\}$. Then the probability p_j that a read aligns to I_j is the ratio of DNA in $\mathcal{T}$ from I_j compared to the total amount of DNA in the sample. That is, $p_j = \frac{(\mathbf{C}\mu)_j}{|\mathbf{C}\mu|_1}$. Therefore, $\mathbf{r}$ is the result of $r = \sum_{j=1}^{m} r_j$ draws from a multinomial distribution with parameter $\mathbf{p} = \widehat{\mathbf{C}\mu} = \frac{\mathbf{C}\mu}{|\mathbf{C}\mu|_1}$ (Figure 2.1). That is $\mathbf{r} \sim Mult(r; \widehat{\mathbf{C}\mu})$.

In contrast to our probabilistic model for DNA sequencing data, other methods for estimating tumor purity and ploidy [164, 27] do not model the data as an observation from an experiment. Rather, they assume that the observed copy number ratio of an interval (or probe) is the ratio of the expected value of the tumor copy number and the expected value of the normal copy number (see Appendix A.1.2). Thus, they implicitly assume that the observed data is an average over many experiments.

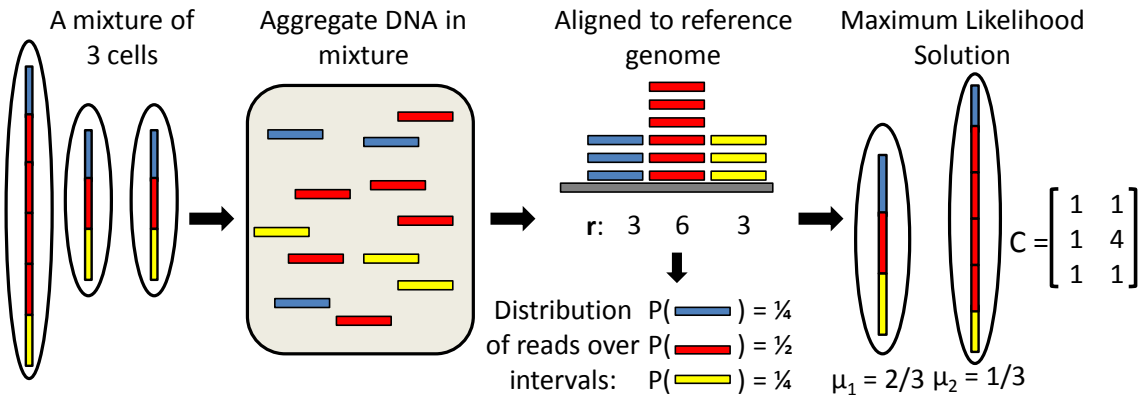


Figure 2.1: **An example mixture of 3 tumor cells.** A mixture of 3 tumor cells with 2 distinct genomes: a normal genome (represented here with one copy of each interval for simplicity), and an aneuploid genome with a duplication of one interval (red). If reads are distributed uniformly over the aggregate DNA in this sample, then the observed distribution of reads over the (blue, red, and yellow) intervals follows a multinomial distribution with parameter $\widehat{\mathbf{C}\mu}$. Here $\mathbf{C}$ is the *interval count matrix* giving the integral number of copies of each interval in each genome in the mixture, and μ is the *genome mixing vector* giving the proportion of each subpopulation in the mixture. We find the pair $(\mathbf{C}, \mu)$ that maximizes the likelihood of the observed read depth vector $\mathbf{r}$.

2.2.3 Solving the MLMDP

We show here how to solve the MLMDP as a disjunction of separate convex optimization problems. The negative log-likelihood of $\mathbf{r}$ as a function of the generic multinomial parameter $\mathbf{p} \in \Delta_{m-1}$ is

$$\mathcal{L}_{\mathbf{r}}(\mathbf{p}) = -\log(P(\mathbf{r}|\mathbf{p})) = -\log\left(\frac{(\sum_{j=1}^{m} r_j)!}{\prod_{j=1}^{m} r_j!}\prod_{j=1}^{m}(p_j)^{r_j}\right) = -\sum_{i=1}^{m} r_i \log(p_i) + \alpha, \tag{2.1}$$

where α is a constant, depending only on $\mathbf{r}$. Finding the multinomial parameter $\mathbf{p}$ that minimizes this negative log-likelihood function is straightforward. Using a Lagrange multiplier to encode the constraint $\mathbf{p} \in \Delta_{m-1}$, one determines that the (unique) value $\mathbf{p}^*$ maximizing $\mathcal{L}_{\mathbf{r}}(\mathbf{p})$ satisfies $p_i^* = \frac{r_i}{\sum_{j=1}^{m} r_j}$. Moreover, if the entries of $\mathbf{r}$ are integers (as they will be for read counts) and $\mathbf{C}$ is permitted to be any integer-valued matrix, then the (unconstrained) solution $p_i^* = \frac{r_i}{\sum_{j=1}^{m} r_j}$ can be written in the form $\mathbf{p} = \widehat{\mathbf{C}\mu}$. In particular, this is true when $\mathbf{C}$ contains some column $\mathbf{c}_j = \mathbf{r}$ and μ is such that $\mu_j = 1$ for some $j \in \{1, \ldots, n\}$. Thus, a solution of the MLMDP is obtained by maximizing the multinomial likelihood over all $\mathbf{p} \in \Delta_{m-1}$.

Constraints on C

In practice, the interval count matrix $\mathbf{C}$ is not allowed to be any integer-valued matrix. There are three natural constraints on the interval count matrix. (1) One component of the tumor sample is the normal genome. Thus, we set the first column $\mathbf{c}_1 = (2, 2, \ldots, 2)^T$, the vector whose entries are all two since humans are diploid. (2) The number n of subpopulations is less than the number m of intervals. (3) The copy numbers of the intervals are integers between 0 and k, inclusive, where $k \geq 2$. We let $\mathcal{C}_{m,n,k}$ denote the set of all matrices satisfying these properties. Thus we define $\Omega_{m,n}$ to be pairs $(\mathbf{C}, \mu)$ where $\mathbf{C}$ satisfies the first two of those conditions.

$$\Omega_{m,n} = \{(\mathbf{C}, \mu) | \mathbf{c}_1 = \mathbf{2}^m, \mathbf{c}_j \in \mathbb{N}^m \text{ for } j > 1, \mu \in \Delta_{n-1}\}. \tag{2.2}$$

Similarly, we define $\Omega_{m,n,k} \subseteq \Omega_{m,n}$ to be the pairs $(\mathbf{C}, \mu)$ where $\mathbf{C}$ satisfies all three of the conditions.

$$\Omega_{m,n,k} = \{(\mathbf{C}, \mu) \mid \mathbf{c}_1 = \mathbf{2}^m, \mathbf{c}_j \in \{0, \ldots, k\}^m \text{ for } j > 1, \mu \in \Delta_{n-1}\}. \tag{2.3}$$

In the following, we will use Ω to refer to *either* $\Omega_{m,n}$ or $\Omega_{m,n,k}$, as appropriate. Given a pair $(\mathbf{C}, \mu) \in \Omega$, we define the negative log-likelihood of the observed read depth vector $\mathbf{r}$ using the multinomial model to be

$$\mathcal{L}_{\mathbf{r}}(\mathbf{C}, \mu) = -\log(P(\mathbf{r}|\mathbf{C}, \mu)) = -\sum_{i=1}^{m} r_i \log((\widehat{\mathbf{C}\mu})_i) + \alpha. \tag{2.4}$$

For an observed $\mathbf{r}$, our goal is to find the $\mathbf{C}$ and μ that minimize (2.4). We define the following optimization problem where the domain of $(\mathbf{C}, \mu)$ can be either of the domains Ω defined above.

$$(\mathbf{C}^*, \mu^*) = \text{argmin}_{(\mathbf{C},\mu)\in\Omega}\, \mathcal{L}_{\mathbf{r}}(\mathbf{C}, \mu) = \text{argmin}_{(\mathbf{C},\mu)\in\Omega}\, \mathcal{L}_{\mathbf{r}}(\widehat{\mathbf{C}\mu}). \tag{2.5}$$

Since all entries of $\mathbf{C}$ are positive integers and all μ_j are positive reals, (2.5) is a mixed integer program. In general, mixed integer *linear* programs (MILP) are NP-hard to solve [67]. In our case, the objective function is a non-linear function of $\mathbf{C}$ and μ meaning that even sophisticated MILP solvers are unlikely to be much benefit for this problem.

A Coordinate Transformation

Rather than attempting to solve the optimization problem (2.5) as a generic MILP, we derive a coordinate transformation that allows us to solve this problem as a constrained optimization problem in $\mathbb{R}^m$. First, note that a pair $(\mathbf{C}, \mu) \in \Omega$ defines a probability distribution $\widehat{\mathbf{C}\mu}$. We define $P_\Omega = \left\{ \widehat{\mathbf{C}\mu} \middle| (\mathbf{C}, \mu) \in \Omega \right\}$ to be the space of all such probability distributions for all $(\mathbf{C}, \mu) \in \Omega$. Note that only $\widehat{\mathbf{C}\mu}$, and not $(\mathbf{C}, \mu)$, is identifiable from the observed data $\mathbf{r}$. We prove the following theorem.

Theorem 2.2.1. *Suppose* $\mathbf{p} \in P_\Omega$*, so* $\mathbf{p} = \widehat{\mathbf{C}\mu}$ *for some* $(\mathbf{C}, \mu) \in \Omega$*. Then there exists* $\mu' \in \Delta_{n-1}$ *such that* $\mathbf{p} = \widehat{\mathbf{C}}\mu'$*, where* $\widehat{\mathbf{C}} = (\widehat{\mathbf{c}_1}, \ldots, \widehat{\mathbf{c}_n})$*.*

Proof. Let $\mu' = (\mu'_1, \ldots, \mu'_n)$ where $\mu'_j = \frac{\mu_j |\mathbf{c}_j|_1}{\sum_{h=1}^n \mu_h |\mathbf{c}_h|_1}$. By definition $\sum_{j=1}^n \mu'_j = 1$ and $\mu'_j \geq 0$, so $\mu' \in \Delta_{n-1}$. We now show that $\widehat{\mathbf{C}}\mu' = \mathbf{p}$. For each $i \in \{1, \ldots, m\}$ we compute that the i^{th} entry of $\widehat{\mathbf{C}}\mu'$:

$$\begin{aligned}
(\widehat{\mathbf{C}}\mu')_i &= \sum_{j=1}^n \widehat{c}_{ij}\mu'_j = \sum_{j=1}^n \frac{c_{ij}}{|\mathbf{c}_j|_1} \frac{\mu_j |\mathbf{c}_j|_1}{\sum_{h=1}^n \mu_h |\mathbf{c}_h|_1} \\
&= \sum_{j=1}^n \frac{\mu_j c_{ij}}{\sum_{h=1}^n \mu_h |\mathbf{c}_h|_1} = \frac{\sum_{j=1}^n \mu_j c_{ij}}{\sum_{h=1}^n \sum_{g=1}^m \mu_h c_{gh}} \\
&= (\widehat{\mathbf{C}\mu})_i = p_i.
\end{aligned}$$

Hence, we see that $\mathbf{p} = \widehat{\mathbf{C}}\mu'$. □

Now suppose the interval count matrix $\mathbf{C}$ is fixed, and let $H(\mathbf{C}) = \{\widehat{\mathbf{C}}\mu \mid \mu \in \Delta_{n-1}\}$ denote the set of convex combinations of the normalized column vectors in $\mathbf{C}$. Then (2.5) reduces to the problem of finding $\operatorname{argmin}_{\mathbf{p} \in H(\mathbf{C})} \mathcal{L}_\mathbf{r}(\mathbf{p})$. Since the objective function $\mathcal{L}_\mathbf{r}(\mathbf{p})$ is separable convex (see Appendix A.1.3) and the domain $H(\mathbf{C})$ is convex, this problem is easy to solve using standard convex optimization routines. Considering all interval count matrices $\mathbf{C} \in \mathcal{C}_{m,n,k}$ gives the following optimization problem.

$$\min \mathcal{L}_\mathbf{r}(\mathbf{p}) \text{ subject to } \mathbf{p} \in \cup_{\mathbf{C} \in \mathcal{C}_{m,n,k}} H(\mathbf{C}). \tag{2.6}$$

Figure 2.2 illustrates the geometry of this optimization problem. Since in general a union of convex sets is not convex, the constraint set in (2.6) is not convex. A brute-force approach to this problem is to enumerate all $\mathbf{C} \in \mathcal{C}_{m,n,k}$, but the number of such matrices is exponential in m and

n. Note that in the Results section THetA demonstrates improved performance in computing $\mathbf{C}$ and μ as the number m of intervals increases in the case where $n = 3$. This is expected from the convex geometry used by our algorithm: for a fixed interval count matrix $\mathbf{C}$, each value $\widehat{\mathbf{C}\mu}$ defines a 2-plane in Δ_{m-1} (see Appendix Figure A.4). These planes become more sparse in Δ_{m-1} as m increases, and thus our algorithm is less prone to overfitting. In the next section, we show that in the $n = 2$ case we can restrict the space of $\mathbf{C}$ matrices to a number that is polynomial in m.

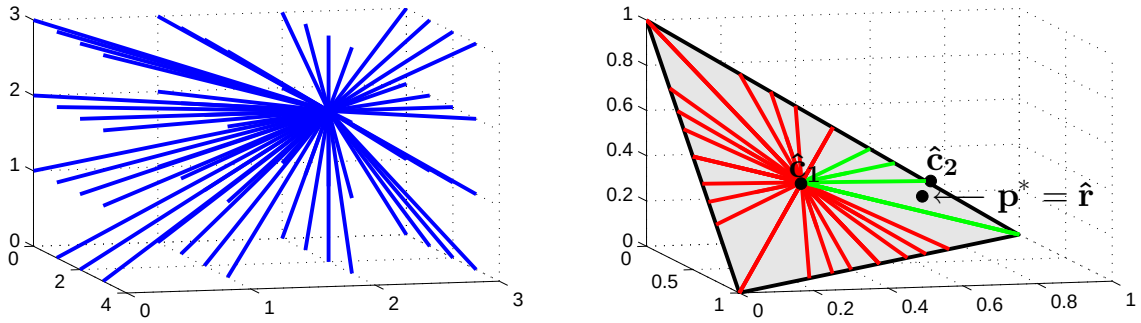


Figure 2.2: **The convex geometry of the MLMDP that is used in the THetA algorithm.** (Left) In the case of a single cancer genome with normal admixture, the interval count vector $\mathbf{c}_2$ of the cancer genome and tumor purity μ define a collection of rays $\mathbf{C}\mu$ for $\mu \in [0, 1]$. (Here we show the space $\Omega_{3,2,3}$). (Right) Normalizing these rays, we obtain the parameter $\mathbf{p} = \widehat{\mathbf{C}\mu}$ used in the multinomial likelihood. These parameters are embedded in the simplex Δ_{m-1} (gray triangle with a black outline) because their entries sum to one. (This is the space $P_{\Omega_{3,2,3}}$). For a fixed interval count matrix $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2)$ the blue ray (left) defined by $\mathbf{C}\mu$ is mapped to the corresponding red/green ray (right) connecting $\widehat{\mathbf{c}_1}$ to $\widehat{\mathbf{c}_2}$ (right), the normalized columns of $\mathbf{C}$, as described in Theorem 2.2.1. For $n > 2$, hyperplanes are mapped to hyperplanes (see Appendix Figure A.4). We show $\mathbf{p}^* = \widehat{\mathbf{r}}$, the maximum likelihood solution when interval counts are not constrained to be integers. Note that this point is not on any of the rays defined by interval count matrices. Rays that satisfy the ordering constraint from Theorem 2.2.2 are in green.

2.2.4 A More Efficient Algorithm for the MLMDP

We derive an algorithm to solve the MLMDP (as formulated in (2.6)) that is polynomial time in m when $n = 2$. This algorithm relies on the observation that it is necessary to consider only a subset of interval count matrices $\mathbf{C}$ whose entries satisfy ordering constraints imposed by the read depth vector $\mathbf{r}$. We say that a vector $\mathbf{a} = (a_1, \ldots, a_m) \in \mathbb{R}^m$ has *compatible order* with another vector $\mathbf{b} = (b_1, \ldots, b_m) \in \mathbb{R}^m$ if for all $1 \leq i, j \leq m$, $b_i \leq b_j$ implies that $a_i \leq a_j$. Note that the vector $x = (s, \ldots, s) \in \mathbb{R}^m$ for any $s \in \mathbb{R}$ has compatible order with all vectors in $\mathbb{R}^m$.

Theorem 2.2.2. *Suppose* $\mathbf{p}^* = \widehat{\mathbf{C}^*\mu^*} = \underset{p \in P_{\Omega_{m,n,k}}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p})$ *and all entries in* $\mathbf{r}$ *are distinct. Then we have the following.*

1. $\mathbf{p}^*$ *has compatible order with* $\mathbf{r}$*.*

2. *If* $n = 2$ *and* $\mu_2^* > 0$*, then* $\mathbf{c}_2^*$ *has compatible order with* $\mathbf{r}$*.*

Proof. We start with part (1) and proceed by contradiction. Suppose $\mathbf{r}$ and $\mathbf{p}$ do not have a compatible order, that is there exist $i, j \in \{1, \ldots, m\}$ such that $r_i \geq r_j$, but $p_i < p_j$. Since all entries in $\mathbf{r}$ are distinct, then necessarily $r_i > r_j$. Without loss of generality assume $i < j$, and let $\mathbf{p}'$ be a point on the simplex obtained from $\mathbf{p}$ by swapping the i^{th} and the j^{th} entries. We first show that $\mathbf{p}' \in P_{\Omega_{m,n,k}}$. Since $\mathbf{p} \in P_{\Omega_{m,n,k}}$, by definition there exists a $(\mathbf{C}, \mu) \in \Omega_{m,n,k}$ such that $\mathbf{p} = \widehat{\mathbf{C}\mu}$. Let $\mathbf{C}'$ be a matrix obtained from $\mathbf{C}$ by swapping the i^{th} and the j^{th} rows. We define $\mathbf{p}' = \widehat{\mathbf{C}'\mu}$. Since the set of entries in $\mathbf{C}'$ is as same as the set of entries in $\mathbf{C}$, we have $(\mathbf{C}', \mu) \in \Omega_{m,n,k}$. Thus $\mathbf{p}' \in P_{\Omega_{m,n,k}}$.

Now, by the theorem assumption $\mathcal{L}_{\mathbf{r}}(\mathbf{p})$ is minimized, and therefore $\mathcal{L}_{\mathbf{r}}(\mathbf{p}) \leq \mathcal{L}_{\mathbf{r}}(\mathbf{p}') \Rightarrow \mathcal{L}_{\mathbf{r}}(\mathbf{p}) - \mathcal{L}_{\mathbf{r}}(\mathbf{p}') \leq 0$. We define $\delta = r_i - r_j > 0$. We have

$$\begin{aligned}
\mathcal{L}_{\mathbf{r}}(\mathbf{p}) - \mathcal{L}_{\mathbf{r}}(\mathbf{p}') &= (-\sum_{i=1}^{m} r_i \log(p_i) + \alpha) - (-\sum_{i=1}^{m} r_i \log(p_i') + \alpha) \\
&= -r_i \log(p_i) - r_j \log(p_j) + r_i \log(p_j) + r_j \log(p_i) \\
&= -(r_j + \delta) \log(p_i) - r_j \log(p_j) + (r_j + \delta) \log(p_j) + r_j \log(p_i) \\
&= -r_j \log(p_i) - \delta \log(p_i) - r_j \log(p_j) + r_j \log(p_j) + \delta \log(p_j) + r_j \log(p_i) \\
&= -\delta \log(p_i) + \delta \log(p_j) \\
&= \delta(\log(p_j) - \log(p_i)) \\
&> 0.
\end{aligned}$$

This is a contradiction. Therefore, it must be the case that $\mathbf{r}$ and $\mathbf{p}$ have compatible order, thus completing the proof of part (1).

Part (2) following directly form part (1) since we assume that the first column of $\mathbf{C}$, denoted $\mathbf{c}_1$ has equal entries. □

Theorem 2.2.2 leads to a more efficient algorithm in the case where $n = 2$. First, a permutation $\pi : \{1, \ldots, m\} \to \{1, \ldots, m\}$ is $\mathbf{r}$-*compatible* if $r_{\pi_1} \leq r_{\pi_2} \leq \ldots, \leq r_{\pi_m}$. For every r-compatible

permutation π, we define $M_{\mathbf{r},\pi}$ to be the set of matrices in $\mathbf{C} \in \mathcal{C}_{m,2,k}$ where $\mathbf{c}_2$ has compatible order with $\mathbf{r}_{\pi(1)}, \ldots, \mathbf{r}_{\pi(m)}$.

We only need to enumerate the elements of sets $M_{\mathbf{r},\pi}$ for each r-compatible permutation π to find the optimal $(\mathbf{C}^*, \mu^*)$. For a specific π the number of elements in $M_{\mathbf{r},\pi}$ is $\binom{m+k}{k}$, and we can enumerate $M_{\mathbf{r},\pi}$ efficiently. If r_i's are all distinct (which is expected for real data) then the r-compatible permutation π is unique, and our search space will drop from $O(k^m)$ to $O\big(\binom{m+k}{k}\big) = O(m^k)$, meaning the exponential size search space reduces to a polynomial (in m) size space.

Not only is $M_{\mathbf{r},\pi}$ polynomial in m in size, but also we can efficiently enumerate all the elements in $M_{\mathbf{r},\pi}$. Suppose the matrices $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2) \in M_{\mathbf{r},\pi}$, where $\mathbf{c}_1 = \mathbf{2}^m$ and $\mathbf{c}_2 = (c_{12}, \ldots, c_{m2})^T$, are ordered based on the lexicographical order of $(c_{\pi(1)2}, \ldots, c_{\pi(m)2})$. Given a matrix $\mathbf{C}' \in M_{\mathbf{r},\pi}$, Algorithm 1 finds the *next* matrix in $M_{\mathbf{r},\pi}$, i.e., the successor of the matrix $\mathbf{C}'$ in $M_{\mathbf{r},\pi}$, in time $O(m)$ which implies that all the matrices in $M_{r,\pi}$ can be enumerated in time $O(m^{k+1})$ since the size of $M_{r,\pi}$ is $O(m^k)$.

Algorithm 1: Enumeration of matrices in $M_{\mathbf{r},\pi}$.

input : A matrix $\mathbf{C} \in M_{\mathbf{r},\pi}$
output: The next matrix $\texttt{next}(\mathbf{C}) \in M_{\mathbf{r},\pi}$

begin
  $(c_{1,2}, \ldots, c_{m,2}) \longleftarrow$ the second column of C;
  $j \longleftarrow$ the largest j such that $c_{\pi(j),2} < k$;
  if j *exists* **then**
    for $i = 1 \to j - 1$ **do**
      $c'_{\pi(i),2} \longleftarrow c_{\pi(i),2}$;
    for $i = j \to m$ **do**
      $c'_{\pi(i),2} \longleftarrow c_{\pi(j),2} + 1$;
  else
    $\texttt{next}(\mathbf{C}) \longleftarrow$ NULL;
  $\texttt{next}(\mathbf{C}) \longleftarrow$ the matrix with columns $\mathbf{2}^m$ and $(c'_{1,2}, \ldots, c'_{m,2})^T$;
  return $\texttt{next}(C)$;

Restricting the Solution Space when $n > 2$

Now suppose $n > 2$. Based on Theorem 2.2.2 if $\mathbf{p}^* = \widehat{\mathbf{C}^*\mu^*} = \underset{p \in P_{\Omega_{m,n,k}}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p})$ and all entries in $\mathbf{r}$ are unique, then $\widehat{\mathbf{C}^*\mu^*}$ has compatible order with $\mathbf{r}$. Therefore, $\forall i, j \in \{1, \ldots, m\}$ we have

$$\mathbf{r}_i \leq \mathbf{r}_j \Rightarrow (\widehat{\mathbf{C}^*\mu^*})_i \leq (\widehat{\mathbf{C}^*\mu^*})_j \Rightarrow (\mathbf{C}^*\mu^*)_i \leq (\mathbf{C}^*\mu^*)_j \Rightarrow \exists\ t \in \{1,\ldots,n\},\ c^*_{i,t} \leq c^*_{j,t}. \tag{2.7}$$

Although this does not impose any total order on columns of $\mathbf{C}$, it still puts some restrictions on entries of $\mathbf{C}$. This leads into a smaller search space with exponential size (in m and n), but is still a useful speed up in practice.

Considering a Fixed Purity

Sometimes an estimate of sample purity (μ_1) is obtained from sample examination by a pathologist or from the output of another computational program. Thus, it is desirable to be able to use this estimate of tumor purity when inferring the size of tumor subpopulations and the copy number aberrations that distinguish them. In this subsection we show that the theoretical results from the previous sections extend to this case.

Let, $p \in [0,1]$ be the previously inferred purity of a tumor sample sample. We define $\Delta^p_{n-1} = \{\mu = (\mu_1,\ldots,\mu_n) \in \Delta_{n-1} \mid \mu_1 = p\}$ and $\Omega_{m,n,k,p} \subseteq \Omega_{m,n,k}$ to be the pairs $(\mathbf{C},\mu)$ where $\mu \in \Delta^p_{n-1}$. We have the following theorem whose proof (appearing in the Appendix) is nearly identical to the proof of Theorem 2.2.2.

Theorem 2.2.3. *Suppose* $\mathbf{p}^* = \widehat{\mathbf{C}^*\mu^*} = \underset{p \in P_{\Omega_{m,n,k,p}}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p})$ *and all entries in* $\mathbf{r}$ *are distinct. Then we have the following:* $\mathbf{p}^*$ *has compatible order with* $\mathbf{r}$.

Theorem 2.2.3 tells us that the optimal solution when considering a fixed level purity will have compatible order with the observed read depth vector $\mathbf{r}$. Lastly, since Δ^p_{n-1} is a convex subset of Δ_{n-1} and $\mathcal{L}_{\mathbf{r}}(\widehat{\mathbf{C}\mu})$ is convex over Δ_{n-1}, we have that $\mathcal{L}_{\mathbf{r}}(\widehat{\mathbf{C}\mu})$ is convex over Δ^p_{n-1}. However, we note that the transformation described in Theorem 2.2.1 cannot be used directly when performing optimization since there is no guarantee that $\mu'_1 = p$.

2.2.5 Intervals of Unequal Length and Mappability

Thus far, we made the simplifying assumptions that all intervals in $\mathbf{I}$ were of equal length and that reads aligned to each interval without any biases from the DNA sequence of the interval. Now we consider the general case where each interval I_j has an associated positive weight w_j. These weights

can model both interval lengths as well as different *mappability* of intervals – i.e. the probability of reads aligning uniquely to an interval in the reference genome can depend on the repeat content of the interval [89]. Let $\mathbf{w} = (w_1, \ldots, w_m)$ be the *interval weight vector*. In practice we use the read depth vector over $\mathbf{I}$ for the paired normal sample as $\mathbf{w}$ which allows us to implicitly incorporate information of interval length, mappability and GC content into the model.

Consider a single cancer genome where $\mathbf{c} = (c_1, \ldots, c_m)$ is the number of copies of each interval in the cancer genome. Then the probability p_j of a read aligning to interval I_j in the reference genome is $\frac{w_j c_j}{\sum_{i=1}^m w_i c_i} = \frac{w_j c_j}{|\mathbf{Wc}|_1}$, where $\mathbf{W}$ is a diagonal matrix such that $\mathbf{W}_{j,j} = w_j$. Therefore, the observed read depth vector $\mathbf{r}$ is obtained by $r = \sum_{j=1}^m r_j$ independent draws from a multinomial distribution with parameter $\mathbf{p} = (\frac{w_1 c_1}{|\mathbf{Wc}|_1}, \ldots, \frac{w_m c_m}{|\mathbf{Wc}|_1})$. We define the linear transformation $\Phi : \mathbb{R}^m \to \mathbb{R}^m$ to be $\Phi(\mathbf{v}) = \widehat{\mathbf{Wv}}$. Thus, $\mathbf{p} = \Phi(\mathbf{c})$ and $\mathbf{r} \sim Mult(r; \Phi(\mathbf{c}))$. As in the unweighted case above, if the entries in $\mathbf{c}$ are allowed to be arbitrary positive integers, then for any integer read depth vector $\mathbf{r}$ and non-negative weight vector $\mathbf{w}$ we can always find the maximum likelihood solution to the corresponding weighted MLMDP (see Appendix A.1.6).

Similarly, if we consider a tumor mixture $\mathcal{T}$ with interval count matrix $\mathbf{C}$ and genome mixing vector μ, the probability p_j of a read aligning to interval I_j satisfies $p_j = \frac{w_j(\mathbf{C}\mu)_j}{\sum_{i=1}^m w_i(\mathbf{C}\mu)_i} = \frac{(\mathbf{WC}\mu)_i}{|\mathbf{WC}\mu|_1} = \Phi(\mathbf{C}\mu)$. Given a read depth vector $\mathbf{r}$ and an interval weight vector $\mathbf{w}$, we formulate the analogous Maximum Likelihood Mixture Decomposition Problem of identifying the underlying interval count matrix $\mathbf{C}$ and genome mixing vector μ that maximize the multinomial likelihood $Mult(\mathbf{r}|\Phi(\mathbf{C}\mu))$.

Theorem 2.2.4 (see Appendix A.1.6 for the proof) relates the optimal $(\mathbf{C}, \mu)$ in the cases of *equal* and *unequal* weighted intervals.

Theorem 2.2.4. *Let* $\Phi^{-1} : \mathbb{R}^m \to \mathbb{R}^m$ *be* $\Phi^{-1}(\mathbf{v}) = \widehat{W^{-1}\mathbf{v}}$. *We have the following set equality,*

$$\text{argmin}_{(\mathbf{C},\mu)\in\Omega_{m,n}} \mathcal{L}_{\mathbf{r}}(\Phi(\mathbf{C}\mu)) = \text{argmin}_{(\mathbf{C},\mu)\in\Omega_{m,n}} \mathcal{L}_{\Phi^{-1}(\mathbf{r})}(\widehat{\mathbf{C}\mu}).$$

Using this theorem, we find the optimal solution in the weighted interval case by solving the unweighted interval case; e.g. using the techniques above. As stated, Theorem 2.2.4 applies to the case where $(\mathbf{C}, \mu) \in \Omega_{m,n}$ (i.e. the entries of $\mathbf{C}$ are unbounded). However, we can still leverage the logic behind this result when we restrict to $\mathbf{C} \in \mathcal{C}_{m,2,k}$. While we do not expect that $\text{argmin}_{(\mathbf{C},\mu)\in\Omega_{m,2,k}} \mathcal{L}_{\mathbf{r}}(\Phi(\mathbf{C}\mu))$ is equal to $\text{argmin}_{(\mathbf{C},\mu)\in\Omega_{m,2,k}} \mathcal{L}_{\Phi^{-1}(\mathbf{r})}(\widehat{\mathbf{C}\mu})$, we may assume that a solution to $\text{argmin}_{(\mathbf{C},\mu)\in\Omega_{m,2,k}} \mathcal{L}_{\mathbf{r}}(\Phi(\mathbf{C}\mu))$ will satisfy the same order constraints as $\mathcal{L}_{\Phi^{-1}(\mathbf{r})}(\widehat{\mathbf{C}\mu})$.

Namely, we expect that the optimal solution will have compatible order with $\Phi^{-1}(\mathbf{r})$ (Theorem 2.2.2). This is because: (1) the unconstrained optima (when $(\mathbf{C}, \mu) \in \Omega_{m,n}$) for the two likelihood functions are equal; (2) the objective function $\mathcal{L}_{\mathbf{r}}(\mathbf{p})$ is well-behaved (separable convex); (3) the transformation Φ is linear. Thus, the optima in the constrained weighted case cannot deviate too much from the optima in the constrained unweighted case, where the ordering conditions hold. Thus, we need only to consider $\mathbf{C} \in \mathcal{C}_{m,2,k}$ where $\mathbf{c}_2$ has compatible order with $\Phi^{-1}(\mathbf{r})$ to find an optimum. We verified this statement empirically over a variety of simulations where both $\mathbf{r}$ and $\mathbf{w}$ were chosen at random. We then verified that the solution sets returned when considering all matrices $\mathbf{C} \in \mathcal{C}_{m,2,k}$ and when only considering $\mathbf{C}$ with compatible order $\Phi^{-1}(\mathbf{r})$ were the same.

2.2.6 Model Selection

W note that the likelihood $P(\mathbf{r}|\mathbf{C}, \mu)$ increases with larger number n of tumor subpopulations in the mixture: indeed the observed read depth vector can be fit "perfectly" by placing each copy number aberration in its own tumor subpopulation. However, mixtures with larger n also have greater model complexity (i.e. more parameters). We use a model selection criterion based on the Bayesian Information Criterion (BIC) to select the model with a balance between higher likelihood and lower model complexity, and avoid overfitting. The standard form of the BIC is $-2\log(L) + a\log(b)$ where L is the likelihood of a solution, a is the number of free parameters in the model, and b is the number of datapoints. We add a slight modification to this, similar to a modification used by the segmentation algorithm BIC-Seq [174] that allows use to more stringently penalize solutions with more free parameters using a new parameter γ. The motivation is that the BIC tends to be too liberal when the model space is large [30] – as is the case here. Values of γ above 1 will more strongly penalize models with more distinct tumor populations, that is, increasing this parameter will more strongly encourage solutions with fewer subpopulations. The default value of γ is 10, and was chosen because we expect to recover a small number of distinct subpopulations from sequencing data - thus making penalization of models with more subpopulations attractive. Additionally, changing γ in either direction (by up to 4) from this default value yields consistent results on the datasets analyzed. Our modified BIC is $-2\log(L) + \gamma a\log(b)$, where $a = (m+1)(n-1)$ and b is the total number of reads in the intervals for both the tumor and normal samples. Since we often run the $n = 3$ version of the algorithm on a subset of the intervals used in the $n = 2$ algorithm, we use the following steps to determine which value of n to select. (1) Run the algorithm for $n = 2$ and $n = 3$

using the subset of intervals and the lower and upper bounds used for $n = 3$ and obtain respective likelihood values. (2) Compute the modified BIC for both values of n and choose the one with the lowest value.

2.2.7 Sets of Maximum Likelihood Solutions

We note that tumor-sequencing data alone does not distinguish between different *optimal* solutions with the *same* maximum likelihood. In mathematical terms, this is because only the parameter of the multinomial distribution is *identifiable* from the observed read depth vector $\mathbf{r}$. Thus, we cannot distinguish between pairs $(\mathbf{C}, \mu)$ and $(\mathbf{C}', \mu')$ of interval count matrices and genome mixing vectors that give the same multinomial parameter (that is, $\widehat{\mathbf{C}\mu} = \widehat{\mathbf{C}'\mu'}$) and therefore have the same likelihood. By default THetA will always output the complete set of maximum likelihood solutions to the MLMDP given the input parameters (e.g. the maximum copy number k to consider). However, THetA has several options that allow a user to input additional information, like sample ploidy, that may be known in advance. One option allows a user to supply an expected ploidy for a sample (e.g. 4 in the case of a tetraploid genome), and the lower and upper bounds considered for all intervals are rescaled to reflect this expected ploidy. Another option allows a user to directly set lower and upper bounds on copy numbers for all intervals in the genome. In either case, THetA will still output the complete set of maximum likelihood solutions that reflect the options supplied by the user.

2.2.8 Virtual SNP Arrays

In the Results section, one means we employ to compare our predictions that differ from those presented in [114] is to look at known germline SNP allele frequencies derived directly from the sequencing data – a "virtual SNP array". We emphasize that this data is not used by our THetA algorithm for computing tumor heterogeneity, and therefore provides an independent data for validation. We look at read coverage and variant allele frequency for the 907,693 SNP positions on the 22 autosomes tested by the Affymetrix 6.0 SNP array (SNP positions and major and minor alleles for hg19 determined using the UCSC genome browser [78]). Read coverage for a SNP position is the number of concordant reads with mapping quality > 30 that have an alignment containing either the major or minor allele at the SNP position. The variant allele fraction, or BAF, is the fraction of such reads that contains the minor allele.

2.3 Results

In this section we describe results from running THetA on both simulated data and a breast cancer dataset from [114].

2.3.1 Simulated Data

In this section we present the results from multiple different simulations and compare THetA to other methods for estimating tumor purity and ploidy. Further details about how simulated data was created, the other algorithms and additional results not presented here are located in Appendix A.2

Normal Admixture: Single Cancer Genome

Using two different sets of simulated data, we compare our THetA algorithm to three other methods for estimating tumor purity and ploidy: ASCAT [164], ABSOLUTE [27], and CNAnorm [62]. ASCAT and ABSOLUTE jointly estimate tumor purity and ploidy, and were originally designed for SNP array data. While both can be adapted to run on DNA sequencing data, they do not formally model this type of data, as noted above. CNAnorm is designed for DNA sequencing data, but rather than allowing tumor purity and tumor ploidy to inform each other, it uses an iterative approach that separately infers purity and copy numbers. In some instances, CNAnorm relies on the user to manually enter the most abundant ploidy.

As noted above, there may be multiple optimal solutions with the same maximum likelihood. CNAnorm [62] and ASCAT [164] use *ad hoc* criteria to return only a single purity estimate, and ABSOLUTE [27] uses external cancer karyotypes to select among multiple possible solutions. To compare THetA to these other methods, we must select a single pair $(\mathbf{C}, \mu)$ from the set returned by THetA as a representative sample reconstruction. For all simulations we choose the pair $(\mathbf{C}, \mu)$ that maximizes the total length of all genomic intervals in the tumor genome with copy number of 2, the expected copy number of the normal genome for humans. We note that this decision applies only to these simulations – for real sequencing data the set of all equally like solutions are returned by THetA from which a user may select one in the presence of additional information about the sample under consideration.

In our first set of simulations we generate a cancer genome consisting of chromosome arm copy

number aberrations. The copy number for each non-acrocentric chromosome arm is chosen uniformly at random from the range 0 (i.e. homozygous deletion) through $k > 2$ (amplification), up to a specified maximum copy number k). While real cancer genomes may have copy numbers larger than the maximum value ($k = 7$) considered in these simulations, such high amplitude amplifications are generally focal events. We emphasize that it is not necessary to use all copy number aberrations to infer the tumor composition; e.g. if there are a sufficient number of arm-level copy number aberrations, these may suffice. We then create a random mixture of this cancer genome and a "matched normal" genome and simulate a read depth vector $\mathbf{r}$ for the mixture, adding noise according the *read depth estimation error* ϕ. The parameter ϕ models errors in the sequencing and analysis process, and we estimate ϕ from real sequencing data to range from 0.01 to 0.04 (see Appendix Figure A.5). Since the ASCAT algorithm uses SNP array data, we also simulate SNP array data from our mixture. In this first set of simulations, we find that our THetA algorithm computes both $\mathbf{C}$ and μ very accurately over a range of copy numbers k (Table 2.2). In particular, THetA outperforms CNAnorm, ABSOLUTE, despite the fact that ASCAT uses additional information (allele frequencies) that THetA does not consider. Even with high amplitude copy number aberrations ($k = 7$) THetA on average estimates tumor purity within 0.5% of the true purity, compared to 6.9%, 10.8% and 14.9% by ASCAT, CNAnorm and ABSOLUTE respectively. Even when THetA does not estimate all copy numbers across the genome correctly, it estimates most copy numbers correctly and estimates the copy number correctly for more segments than the other algorithms (see Appendix Figures A.6, A.7 and A.8).

We also compare THetA, CNAnorm, and ABSOLUTE on a second set of simulated mixtures of tumor and normal cells created using real sequencing data from an acute myeloid leukemia (AML) tumor sample and matched normal sample (TCGA-AB-2965) from The Cancer Genome Atlas (TCGA) [25]. This sample was chosen due to its high purity (∼95% pure) and lack of copy number aberrations. We spike in 10 copy number variants of length 2.5Mb at random non-overlapping positions in Chr20 (excluding the centromere) into the tumor genome. As in the first set of simulations, the copy number for each variant is chosen uniformly at random from the range 0 (i.e. homozygous deletion) through $k > 2$ (amplification), up to a specified maximum copy number $k = 5$. (We did not run ASCAT on this simulated data since this algorithm designed only for microarray data.) We again find that THetA outperforms both CNAnorm and ABSOLUTE on all measures (Figure 2.3). In particular, THetA estimates the sample purity with an order of magnitude better accuracy (using

	% Correct $\mathbf{C}$				Copy number error (median)				Purity error (median)			
k	THetA	ASCAT	CNA	ABS	THetA	ASCAT	CNA	ABS	THetA	ASCAT	CNA	ABS
3	100.0	85.0	40.0	70.0	0.0	0.0	0.103	0.000	0.004	0.040	0.068	0.010
4	90.0	55.0	8.3	50.0	0.0	0.0	0.163	0.013	0.004	0.037	0.064	0.010
5	85.0	50.0	6.7	15.0	0.0	0.013	0.185	0.160	0.004	0.062	0.038	0.075
6	55.0	40.0	0.0	15.0	0.0	0.026	0.291	0.433	0.006	0.063	0.066	0.157
7	30.0	15.0	0.0	10.0	0.031	0.036	0.445	0.471	0.005	0.069	0.108	0.149

Table 2.2: **Performance of THetA, ASCAT, CNAnorm and ABSOLUTE on simulated data with one tumor population ($n = 2$).** Performance of algorithms on simulated datasets with interval count matrix and mixing vector $(\mathbf{C}, \mu) \in \Omega_{39,2,k}$, and read depth estimation error $\phi = 0.03$. 20 simulated datasets were generated for each value of k, the maximum copy number. % Correct $\mathbf{C}$ is the percentage of datasets where the inferred interval count matrix $\mathbf{C}^*$ exactly equals the true simulated matrix $\mathbf{C}$ for the sample. Copy number error is $\frac{1}{m(n-1)}|\mathbf{C} - \mathbf{C}^*|_2$, that is the average error per copy number estimate made, or per entry in $\mathbf{C}$, where error is the euclidean distance between $\mathbf{C}$ and $\mathbf{C}^*$. Purity error is $|\mu_2 - \mu_2^*|$, that is the distance between the true and inferred tumor purity. We only calculate results for CNAnorm where the inferred purity was $< 100\%$ (between 12-15 trials for each k). Appendix Figure A.6 illustrates the results obtained by each algorithm on one of the datasets when $k = 7$.

root mean squared error as a metric of comparison as is done in ABSOLUTE [27]), and consistently identifies more true copy number aberrations than the other algorithms across different purity values and sequencing coverage. In particular, THetA identifies 7.4 and 2.2 more copy number aberrations, on average, than ABSOLUTE and CNAnorm, respectively, across all purity values at 30X sequencing coverage. Even when we relax the requirement that a copy number aberration be predicted with the correct copy number, and instead count any non-normal copy number as correct, THetA still outperforms the other algorithms (see Appendix Figure A.9).

Mixture of Tumor Subpopulations

We next evaluate the performance of THetA on a simulated mixture containing two subpopulations of tumor cells with different copy number aberrations and admixture by normal cells. Thus, there are three distinct subpopulations in the mixture ($n = 3$). Our method for constructing simulated data is the same as the first set of simulations in the previous section with a fixed read depth estimation error of $\phi = 0.02$ along with a few minor changes (see Appendix A.2.1). While the performance of THetA is less precise than in the tumor with normal admixture ($n = 2$) case in estimating all copy numbers (i.e. the entries in $\mathbf{C}$) *exactly*, THetA maintains a good level of accuracy as the estimates are *near* the true interval copy numbers (Table 2.3). In fact, THetA correctly computes on average 94% of the copy numbers across all subpopulations in the mixture when there are $m = 12$ intervals with varying copy in the subpopulations. THetA also estimates the tumor purity with good accuracy (within 3.6% of the true purity when $m = 12$), whereas both CNAnorm and ABSOLUTE gravely misestimate the tumor purity by 30.1% and 54.7% respectively. One possible explanation for these errors is that both these methods do not account for multiple subpopulations in the sample and therefore tend to report tumor purity as either the fraction of the sample representing the largest subpopulation, or as an average of the fractions of all tumor subpopulations. Thus, THetA successfully recovers a complex mixture of two tumor subpopulations and normal cell admixture directly from the observed read depth.

2.3.2 Breast Cancer Sequencing Data

We analyzed Illumina paired-end sequencing data from 3 breast cancer genomes and their matched normal samples from [114]. We downloaded the data from the European Genome-phenome Archive (accession number EGAD00001000138). This includes 2 samples that were sequenced to a depth of

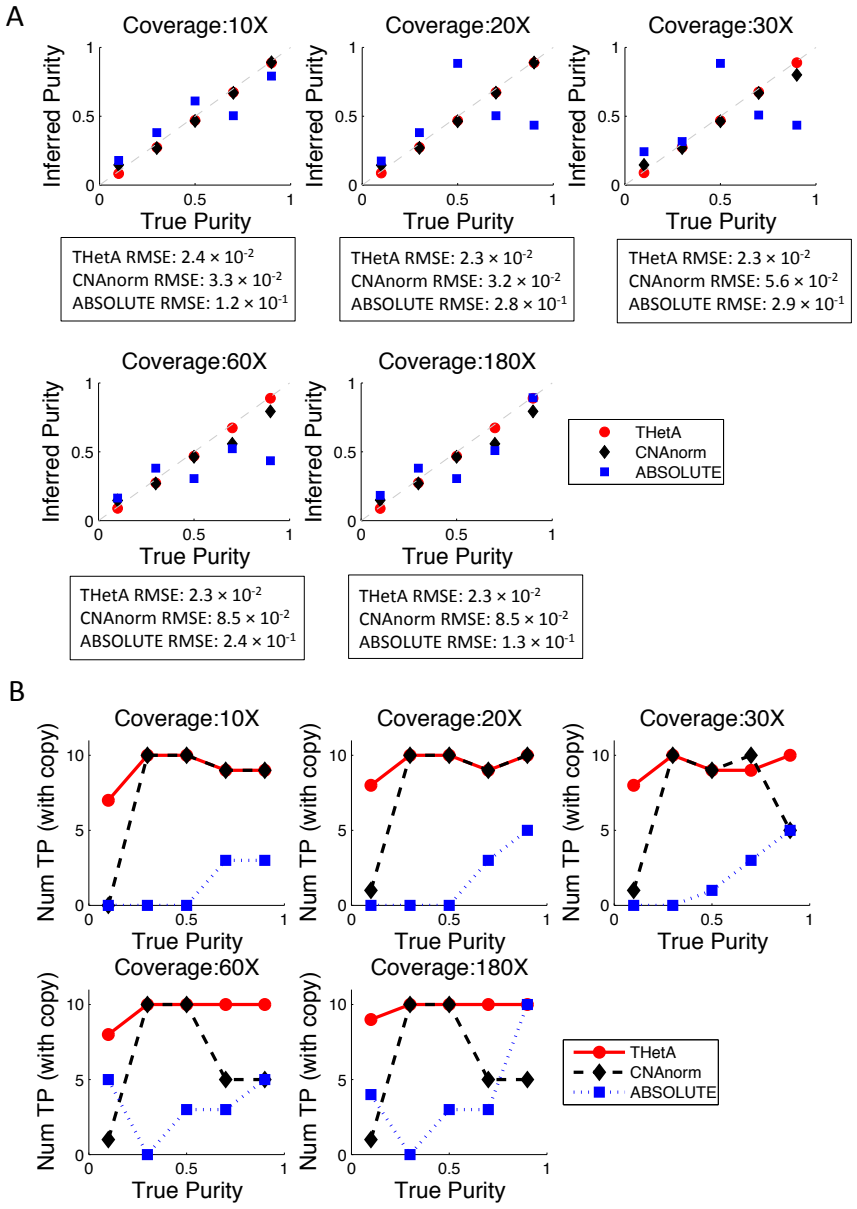


Figure 2.3: **Comparison of THetA to CNAnorm and ABSOLUTE on simulated mixtures from real sequencing data. A.** Comparison of true and inferred tumor purity by THetA, CNAnorm and ABSOLUTE on simulated mixtures of DNA sequencing data from an acute myeloid leukemia (AML) sample and a matched normal sample. Gray dashed line indicates True Purity = Inferred Purity. Below each plot are root mean squared error (RMSE) for each method. **B.** Comparison of the number of copy number aberrations correctly predicted (defined as 50% reciprocal overlap in position and correct integral copy count) by each method for varying tumor purity and sequence coverage. In most cases, THetA outperforms both CNAnorm and ABSOLUTE. Similar results counting aberrations with correct position (with 50% reciprocal overlap) but allowing for difference between true and predicted copy number are in Appendix Figure A.9.

	% Correct $\mathbf{C}$	Copy number error (median)	Purity error (median)		
m	THetA	THetA	THetA	CNA	ABS
6	35.0	0.118	0.081	0.202	0.458
8	45.0	0.075	0.052	0.276	0.477
10	35.0	0.071	0.055	0.177	0.434
12	45.0	0.059	0.036	0.301	0.547

Table 2.3: **Performance of THetA, CNAnorm and ABSOLUTE on simulated data containing two tumor populations ($n = 3$).** Performance of algorithms simulated datasets with interval count matrix and mixing vector $(\mathbf{C}, \mu) \in \Omega_{m,3,3}$, and read depth estimation error $\phi = 0.02$. % Correct $\mathbf{C}$ and Copy number error are defined as in Table 1. We define purity error as the distance between the true and predicted fraction of tumor cells in the sample. Thus, purity error is $|(1 - \mu_1) - (1 - \mu_1^*)|_2$, as the proportion of tumor cells in the sample is $1 - \mu_1$. Since CNAnorm and ABSOLUTE are not able to infer multiple subpopulations, their % Correct $\mathbf{C} = 0$, and we list only their purity estimates. We only calculate results for CNAnorm where the inferred purity was $< 100\%$ (between 14-18 trials for each m).

approximately ∼40X coverage and one sample, PD4120a, that was sequenced with ∼188X coverage. We use the BIC-Seq segmentation algorithm [174] to partition the 22 autosomes into intervals according to read depth. We form an interval count vector from all intervals longer than 50 kb, focusing on these longer genomic intervals because their observed read depth will have lower variance. Most intervals removed in this step are relatively short for the samples analyzed. For the two ∼40X coverage genomes changing this cutoff to 10 kb resulted in the same partition as when 50 kb was used. For the ∼188X genome we only remove 9 intervals from consideration when the threshold is 50 kb and 7 when the threshold is 10 kb. The results from THetA are identical for the two different sets of intervals. We assume that most of the tumor genome does not undergo copy number aberrations, and thus the mode of the read depth vector provides a normal "baseline". We set lower and upper bounds on the copy number for each interval from this baseline. For further details please see Appendix A.3.

Breast Tumor: 188X Sequence Coverage

We analyze the 188X sequenced tumor PD4120a using our THetA algorithm where we consider that the mixture contains normal admixture with a single tumor subpopulation ($n = 2$) and normal admixture with two distinct tumor subpopulations ($n = 3$). This sample was extensively annotated by [114] and thus provides a positive control for THetA. Assuming a single tumor subpopulation admixed with normal cells ($n = 2$), THetA's estimate of tumor purity (65.7%) and inferred copy number aberrations are very close to those obtained by CNAnorm [62] (67.2%), ASCAT (66.0%) [164]

and ABSOLUTE (65%) [27] (Table 2.4). However, all of these estimates are lower than the tumor purity of 70% reported by [114], who identified a second tumor subpopulation in the sample (see below). Because ABSOLUTE, ASCAT, CNAnorm, and THetA (with $n = 2$) do not model multiple tumor subpopulations, their reported tumor purities are an average of the fraction of aberrant cells amongst the different subpopulations in the tumor sample, and thus generally smaller than the tumor purity estimate obtained when we allow more than one tumor subpopulation (see below). In addition, we note that ASCAT uses additional information (B-allele frequencies), while THetA, CNAnorm and ABSOLUTE are using only read depth. The identified aberrations do not distinguish between those in different subpopulations, but do include several previously reported in breast cancer [29, 115, 168, 28, 45, 32]. We also ran THetA using chromosome arms as the intervals, rather than the BIC-Seq intervals. Using chromosome arms, we estimate a similar sample purity of 61.7% and predict the same set of copy number aberrations as with the BIC-Seq intervals.

Assuming $n = 3$ subpopulations – normal cells plus two distinct subpopulations of cancer cells – we analyze a subset of longer intervals that are most informative for copy number analysis. THetA's estimate of 72% tumor purity is slightly higher than the 70% reported by [114]. Moreover, THetA's estimate of tumor purity is higher than the ∼65-67% tumor purity given above for ABSOLUTE, ASCAT, and CNAnorm, three methods which assume only one tumor subpopulation. Our BIC model selection chooses this solution as a better representation of the data (Figure 2.4A), than the solution that only considers a mixture of normal cells and a single tumor population. Using the $n = 3$ model we identify all copy number variants identified above for a single tumor population, plus some additional aberrations including subclonal deletions of chromosomes 8, 11, 12, 14 and 15 not identified under that model (nor by the other algorithms). This demonstrates THetA's ability to identify copy number aberrations occurring in subpopulations of cells. While many of the clonal and subclonal copy number aberrations found by THetA are identical to those reported by [114], there are several notable differences including: a clonal deletion of 16q and different classification of aberrations on chromosomes 1 and 22 as clonal vs. subclonal (Table 2.4). We further analyze these differences below.

We further investigated the following three differences between our analysis and [114]: (1) Clonal deletion of chromosome 16q; (2) Clonal vs. subclonal amplification of chromosome 1q; and (3) Clonal vs. subclonal deletions in chromosome 22q. We analyze these differences using two complementary approaches. First, we analyze the distribution of tumor/normal read depth ratios in 50 kb bins across

the genome. This distribution contains distinct peaks corresponding to copy number aberrations occurring in different subpopulations. We correct read depth ratios using the following 3 steps: (1) Normalize tumor and normal read depth to the same number of reads; (2) Translate distribution so that the mean for a selected set of chromosomes or intervals that are likely to contain no large copy number aberrations is equal to 1; and (3) For a predetermined amount of normal admixture μ, scale each ratio r using the following linear transformation: $(r-1)\frac{1}{1-\mu}+1$. After correcting the read depth ratios, peaks corresponding to clonal aberrations will occur at ratios divisible by 0.5, whereas peaks corresponding to subclonal aberrations will not (Figure 2.4B). Second, we analyze a "virtual SNP array" that we construct from the read counts and the variant allele frequencies derived from aligned reads at known germline SNPs (as described previously). Copy number aberrations occurring in different subpopulations appear as distinct clusters in a scatter plot of read count vs. variant allele frequencies (Figure 2.4C).

The first difference we analyze is our prediction of a clonal deletion of chromosome 16q, which is not reported by [114]. Visual inspection of the virtual SNP array data for chromosome 4 (predicted to be a clonal deletion by both methods) and chromosome 16q shows three distinct clusters - one for regions of normal copy (centered at a variant allele frequency of 0.5) and two clusters (positioned symmetrically around variant allele frequency of 0.5) with lower read count that indicate a deletion (Figure 2.4D). These deletion clusters have virtually identical locations in the scatter plot for chromosomes 4 and 16q – supporting the conclusion that these deletions occur in the same fraction of the tumor sample. Comparing the difference between the observed and expected read depth ratios in these deletions for different aberration fractions (the percent of the sample containing the aberration) reveals that the optimal aberration fraction for both deletions is very similar – additional evidence that these deletions occur in the same fraction of the tumor sample (see Appendix A.4.1 and Appendix Figure A.14). Given the strong evidence for this chromosome 16 deletion, we suspect that its omission from [114] was an oversight rather than a deficiency of the analysis.

The second difference is that we predict chromosome 1q to be amplified in a subclonal population consisting of 61.9% of the cells in the sample, whereas [114] indicate that this aberration is clonal (occurring in 70% cells in the sample). Since this is the only large amplification present in the sample, we cannot compare its variant allele frequencies to a different amplification (as we did with chromosomes 4 and 16 above). Therefore, we more closely examine the read depth data. Visual inspection of read depth ratios after adjusting for our predicted 28% normal admixture (Figure 2.4B)

and the 30% normal admixture predicted by [114] (see Appendix A.13) show that the corrected read depth ratios for chromosome 1q do not match well to a ratio of 1.5 (as would be expected if there was a clonal amplification with copy number 3) – an indication that 1q is a subclonal aberration. Comparison of read depth ratios for 1q to other clonal aberrations supports our prediction that 1q is a sub clonal deletion (see Appendix Figure A.14).

The final difference involves chromosome 22q which we predict to contain both clonal and subclonal deletions, while [114] only report subclonal events. In particular, [114] report that a deletion of a derivative chromosome from a translocation between chromosomes 1 and 22 is the rearrangement responsible for the subclonal deletion observed on 22q. We find that 1p (see Appendix Figure A.14) and the distal portion of 22q (cytogenetic bands 12.2-13.3) appear to be be clonal deletions, while the proximal portion of 22q (cytogenetic bands 11.2-12.1) is a subclonal deletion. In particular, the read-depth/variant-allele plot from the virtual SNP array shows an oblong cluster for chromosome 22 that only partially overlaps with the cluster for chromosome 13, a chromosome predicted by both methods to undergo a subclonal deletion (Figure 2.4E). This evidence supports another possible sequence of rearrangements: (1) A non-reciprocal translocation occurred between chromosomes 1 and 22 (supported by GASV clustering [150] of discordant reads as discussed in Appendix A.4.1) resulting in the clonal loss of 1p and 22q12.2-13.3. Following this translocation, two copies of 22q11.2-12.1 remain. (2) One of these remaining two copies of 22q11.2-12.1 is deleted in a subclonal population (see Appendix Figure A.15).

Breast Tumor: 40X Sequence Coverage

We also analyzed 2 tumor samples from [114] sequenced at ∼40X coverage. For sample PD4088a, our model selection procedure preferred the model of this mixture as a single clonal tumor population with normal admixture fraction 41%. [114] also report this sample as clonal, although do not provide an estimate of tumor purity or copy number aberrations. Further details of the analysis of this sample are in Appendix A.4.2.

We analyzed sample PD4115a, sequenced at ∼40X coverage using THetA again considering the case where the mixture contains normal admixture with a single tumor subpopulation ($n = 2$) and normal admixture with two distinct tumor subpopulations ($n = 3$). Our BIC model selection chooses the model where the sample is a mixture of normal cells and two distinct subpopulations of tumor cells (Figure 2.5A) over the model where the sample contains a single tumor subpopulation with

Sample PD4120a			
Algorithm	% Normal Admixture	Clonal (% Tumor Purity)	Subclonal (%)
THetA, $n = 2$ (segmentation)	34.3%	Del: 1p, 4q, 13, 16q, 22q +1: 1q (65.7%)	-
THetA, $n = 2$ (chromosome arms)	38.3%	Del: 1p, 4q, 13, 16q, 22q +1: 1q (61.7%)	-
CNAnorm * (chromosome arms)	32.8%	Del: 1p, 4q, 13, 16q, 22q +1: 1q (67.2%)	-
ASCAT ** (virtual SNP array)	34%	Del: 1p, 4q, 13, 16q, 22q +1: 1q, 17q, 18, 19, 20 (66.0%)	-
ABSOLUTE *** (segmentation)	35%	(65.0%)	-
THetA $n = 3$	28.0%	Del: **1p**, 4q, **16q**, **22q12.2-13.3** (72.0%)	Del: 13, 22q11.2-12.1 +1: **1q** (61.9%) Del: 8, 11, 12, 14, 15 (10.1%) *Del: 2, 7, **4p**, 6, 9, 18, 21*
Nik-Zainal et al. (2012)	30%	Del: 4q +1: **1q** (70.0%)	Del: 13, **t(1;22)** (47.6%) **Tetraploid** with: Del(-2): 2, 7 Del(-1): 6, 8, 9, 11, 12, 14, 15, 18, 21 (9.8%)

Table 2.4: **Comparison of various algorithms on the 188X coverage breast cancer genome.** Tumor purity and copy number aberrations identified by various algorithms on the 188X coverage breast cancer genome (sample PD4120a). When restricting to a single tumor population with normal admixture ($n = 2$), THetA, CNAnorm, ASCAT, and ABSOLUTE report similar results. When considering a mixture of two distinct tumor subpopulations along with normal admixture ($n = 3$), THetA finds many of the same aberrations as reported in [114], with aberrations in **bold** indicating the differences. Virtual SNP array data and read depth analysis support our predictions for 1p, 1q, 16q, and 22q (See Figure 2.4). This includes aberrations reported by [114] in several chromosomes not considered as part of our analysis (2,6,7,9,18 and 21). Italicized aberrations were not input to the $n = 3$ THetA analysis, and were inferred by examination read depth ratios corrected for normal admixture and tumor cell fractions derived from THetA (please see Appendix A.4.1). *When CNAnorm was run using BIC-Seq intervals the normal admixture was estimated at 6.7%, therefore we report results from CNAnorm using chromosome arms. CNAnorm does not return integer copy numbers – and thus we report aberrations where the returned copy number was within 0.15 of the nearest integer, other aberrations were considered inconclusive. **For ASCAT we use virtual SNP array data as input. ASCAT performs its own segmentation; we list only the large aberrations. ***We report here the maximum likelihood solution returned by ABSOLUTE when considering only Karyotypes. When considering only somatic copy number aberrations (SCNA) or a combination, ABSOLUTE infers a tetraploid solution. For this sample, ABSOLUTE returns copy numbers for only a subset of the input intervals, so we do not report specific copy number aberrations predicted.

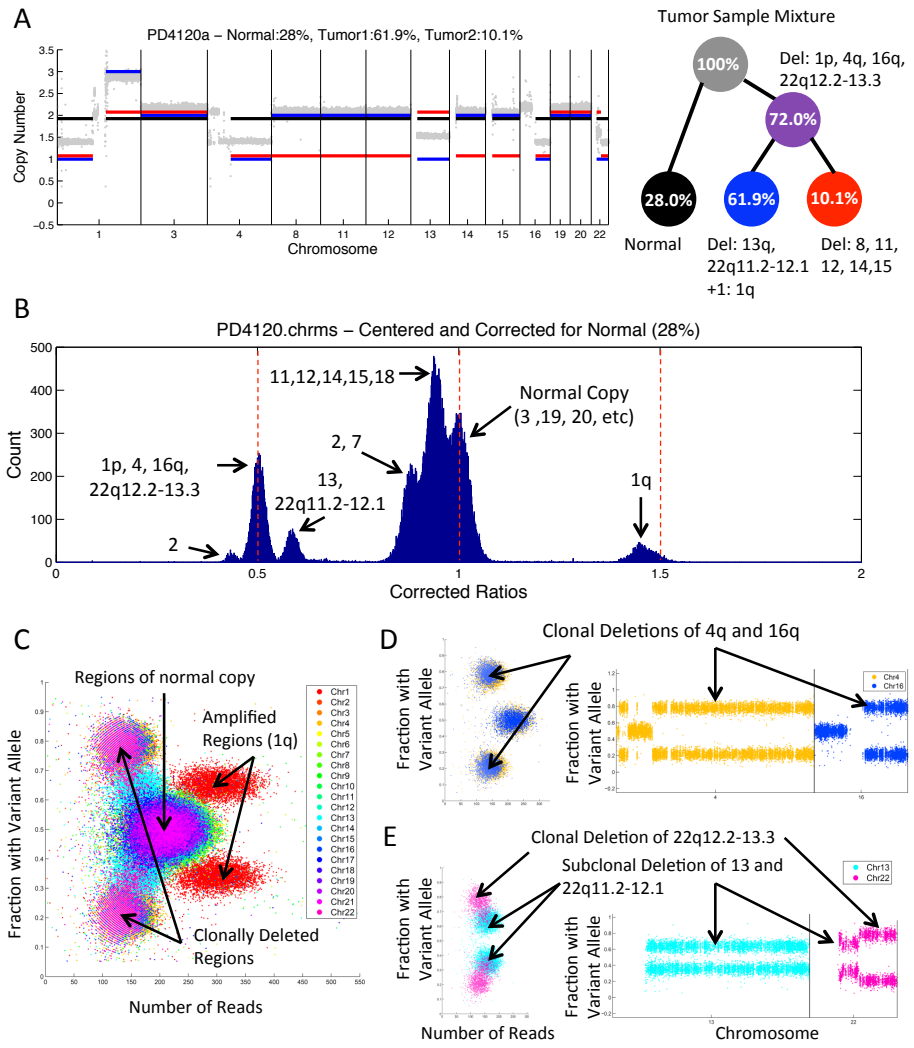


Figure 2.4: **Analysis of the 188X coverage breast tumor PD4120a A.** (Left) Read depth ratios (gray) and the inferred copy number aberrations by our algorithm when $n = 3$ including the normal population (black), dominant (clonal) tumor population (blue), and subclonal tumor population (red). (Right) A reconstruction of the tumor mixture with the inferred aberrations and estimated fraction of cells in each subpopulation. **B.** Read depth ratios in 50 kb intervals after centering so chromosome 3 has a mean of 1 and correcting for 28% normal admixture using a simple linear scaling. **C.** Virtual SNP array results showing distinct clusters of regions according to number of reads containing each SNP and faction of reads supporting variant allele. **D.** Virtual SNP array data comparing variant allele fractions and read counts for chromosomes 4 and 16. This data demonstrate that both chromosomes have undergone the same type of copy number aberration, which we predict to be a clonal deletion in 72% cells in the sample. **E.** Virtual SNP array data for chromosomes 13 and 22. Chromosome 22q11.2-12.1 and chromosome 13 appear to be affected by the same type of aberration, which we predict to be a subclonal deletion in 61.9% cells in the sample. In contrast, 22q12.2-13.3 is different, and the data is consistent with a clonal deletion. See Appendix Figure A.15 for further details.

normal admixture. While [114] provides some analysis of aberrations in this example, they do not provide a complete tumor history as they did for the ∼188X coverage genome. Complete information on our analysis of this sample when we consider it as a mixture of a single tumor subpopulation along with normal admixture is contained in Appendix A.4.3. For the model considering multiple tumor subpopulations, we analyze a subset of longer intervals that are most informative for copy number analysis (further details are in Appendix A.4.3) and estimate a normal admixture of 24% (tumor purity 76%) and two tumor subpopulations of 43.3% and 32.7%, respectively. The presence of these subclonal populations is apparent from visual inspection of corrected read depth ratios after centering the distribution (ratios in chromosome 20 – which is predicted to contain no copy number aberrations – are translated to have a mean ratio of 1) and correction for normal admixture (Figure 2.5B). In particular, a large peak near a corrected ratio of 0.5 represents clonal deletions (Figure 2.5C). In addition, two overlapping, but distinct smaller peaks appear between the clonal deletions and regions of normal copy (Figure 2.5D and 2.5E). These peaks represent two distinct subclonal populations present in the tumor sample. A statistical test of the difference in read depth ratios between these peaks supports the conclusion that these subclonal populations are indeed distinct (see Appendix Figure A.18).

Virtual SNP array analysis of this sample is difficult due to the lower sequence coverage. This leads to overlapping clusters in the read-count/variant-allele plot, as well as distinct banding resulting from the integrality of read counts (Figure 2.6A). The only clearly distinct clusters are for highly amplified regions which exhibit correspondingly higher read counts. Since our analysis for this model used only a subset of chromosome intervals to infer normal admixture and tumor subpopulations, we can use the resulting genome mixing vector μ to analyze other chromosomes that were not used in computing the maximum likelihood solution. We analyze several regions in chromosome 8 (Figure 2.6B), a chromosome with a complicated amplification pattern. After correcting read depth ratios in 50 kb intervals in this region for the estimated normal admixture of 24%, three distinct peaks centered near ratios of 1, 1.5 and 2 are apparent, corresponding to integer copy numbers of 2, 3, and 4, respectively in the tumor sample (Figure 2.6C). The amplifications are clonal aberrations. Interestingly, the variant allele frequencies for germline SNPs in the regions corresponding to the peak at corrected ratio 2 (copy number 4) are centered at 0.5. This implies that both homologs of chr8q13-21 are present at equal copy number in this region; i.e. there is a duplication of both homologs (Figure 2.6D). In addition, we observe that the variant allele frequencies for chromosome

8p are centered at the values of 0 and 1, although this segment of the chromosome is inferred to have copy number 2. This indicates that there was a copy-neutral loss of heterozygosity (LOH) in this region. LOH in 8p have been previously reported in breast cancer [7, 179] and copy neutral LOH in 8p has been reported in cell line data for other cancers [2].

2.4 Discussion

In this Chapter we describe a probabilistic model of DNA sequencing data of heterogeneous tumor sample. Using this model, we formulate the Maximum Likelihood Mixture Decomposition Problem (MLMDP) of finding the most likely collection of genomes and their proportions from high-throughput DNA sequencing data in the case where copy number aberrations distinguish subpopulations. We then introduce Tumor Heterogeneity Analysis (THetA), a convex optimization algorithm that solves the MLMDP. THetA is an efficient (polynomial time) algorithm in the important case where a tumor is a mixture of normal (healthy) cells and a single tumor population.

We show that THetA outperforms three other methods, CNAnorm [62], ASCAT [164] and ABSOLUTE [27], for inferring tumor purity and identifying copy number aberrations in the case of a single tumor cell population admixed with normal (non-cancerous) cells. Moreover, we demonstrate that THetA successfully estimates tumor purity even at low purity (10%) and with modest sequence coverages ($\sim$30X) on both real and simulated data. In contrast to other recent methods [27, 164] that first infer average ploidy across the genome, THetA *simultaneously* estimates tumor purity and computes the integral copy number of each genomic segment/interval. These advantages result from THetA exploiting the large number of data points (reads) that measure copy number aberrations in high-throughput sequencing data – information that is not available from SNP arrays.

We also demonstrate that THetA successfully deconvolves a tumor sample into a normal population and *multiple* tumor subpopulations, inferring the proportion of each subpopulation in the mixture, and partitioning copy number aberrations into clonal and subclonal populations. Other existing methods such as ASCAT [164], ABSOLUTE [27], and CNAnorm [62] do not directly infer multiple subpopulations. Further, we show that these methods can produce highly inaccurate estimates of tumor purity on samples containing multiple subpopulations, and are sometimes unable to identify some copy number aberrations that occur in subpopulations of tumor cells. In addition, THetA reports all possible solutions of interval count matrices $\mathbf{C}$ and genome mixing vectors μ with

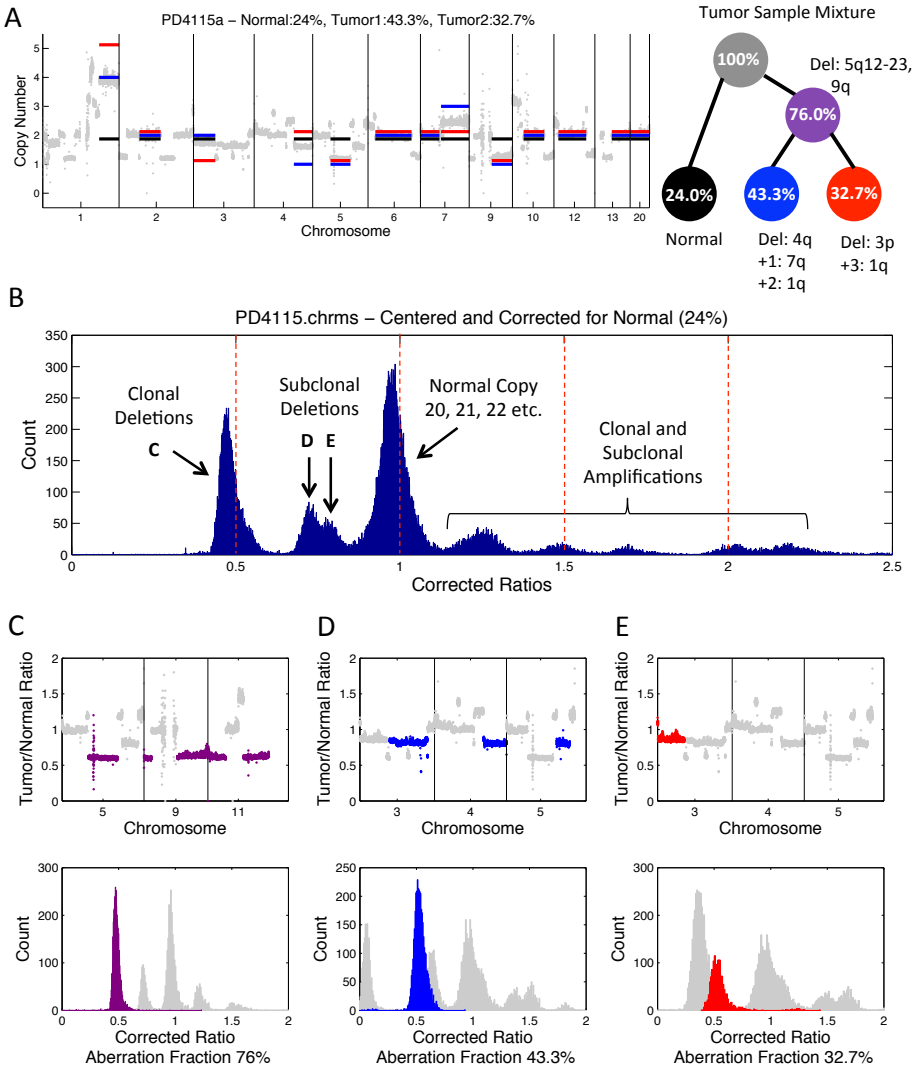


Figure 2.5: **Analysis of the ∼40X coverage breast tumor PD4115a. A.** (Left) Read depth ratios (gray) and the inferred copy number aberrations by our algorithm when $n = 3$ including the normal population (black), dominant (clonal) tumor population (blue), and subclonal tumor population (red). (Right) A reconstruction of the tumor mixture with the inferred aberrations and estimated fraction of cells in each subpopulation. **B.** Distribution of read depth ratios over 50 kb intervals after centering and correction for 24% normal admixture using a simple linear scaling. Several peaks fall near to expected corrected ratios (0.5, 1, 1.5, 2). Two nearby but distinct peaks can be seen indicating multiple sub-clonal deletions in similar proportions. **C.** (Top) Read depth ratios in 50 kb bins for chromosomes 5, 9, and 11, each of which contains a clonal deletion (purple). (Bottom) Distribution of read depth ratios after correction for the aberration fraction of 76% of the sample. **D.** (Top) Read depth ratios in 50 kb bins for chromosomes 3, 4, and 5, each of which contains a subclonal deletion (blue). (Bottom) Distribution of read depth ratios after correction for the aberration fraction of 43.3% of the sample. **E.** (Top) Read depth ratios as in part D, but a different subclonal deletion is highlighted (red). (Bottom) Distribution of read depth ratios after correction for the aberration fraction of 32.7% of the sample.

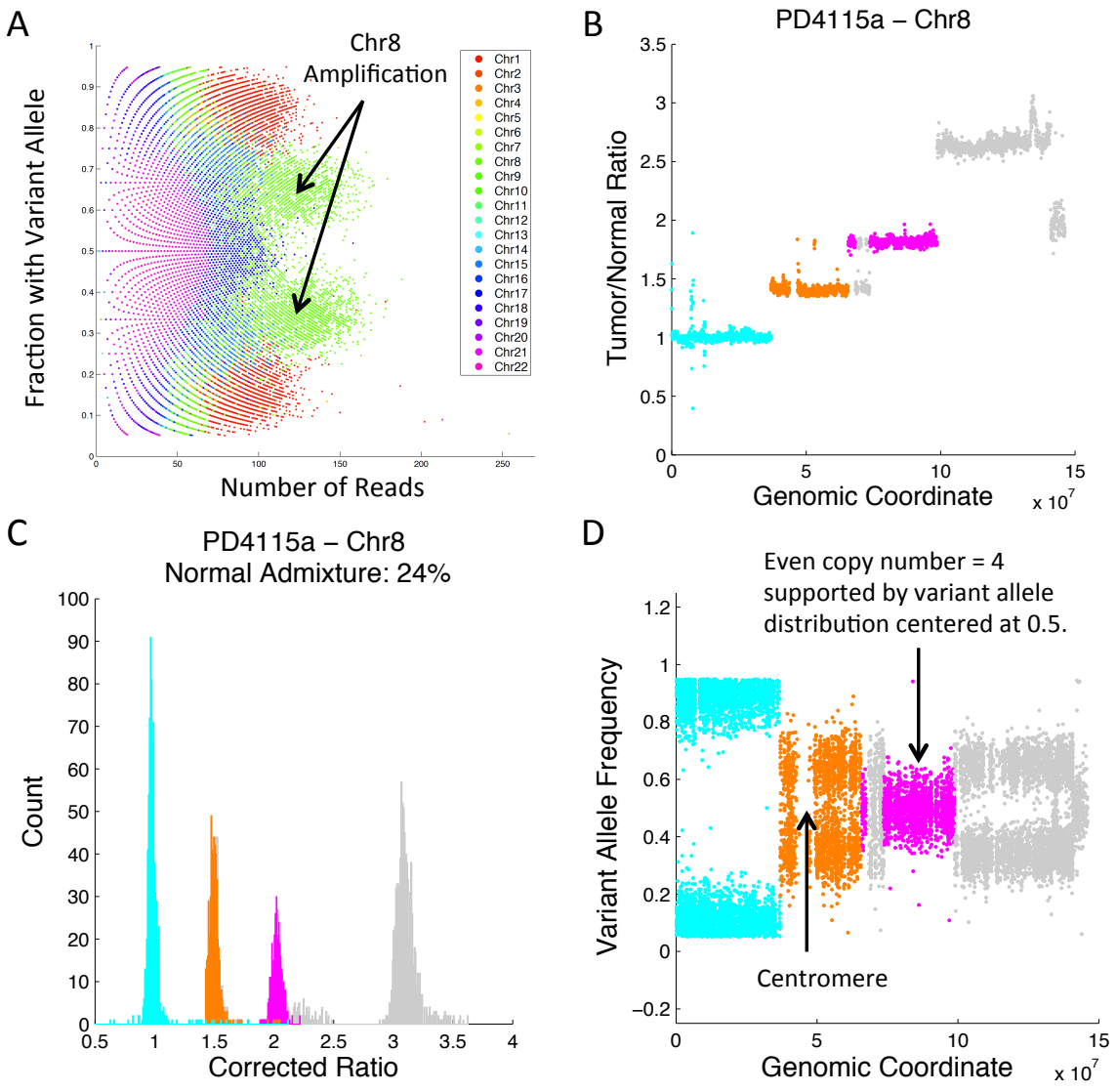


Figure 2.6: **Analysis of chromosome 8 in breast tumor PD4115a. A.** Virtual SNP array data from this sample show few distinct clusters (compare to the 188X sample in Fig 3), with amplification of chromosome 8 (green) being the most prominent. **B.** Read depth ratios for chromosome 8 organized by genomic coordinate. **C.** Histograms of read depth ratios for chromosome 8 corrected for 24% normal admixture, indicating regions of copy 2,3,and 4 (cyan, orange and pink) respectively, with the later two being clonal amplifications. **D.** Variant allele frequencies for chromosome 8. The region with copy number 4 (pink) has variant allele frequencies clustered around 0.5, suggesting duplication of both chromosomal homologs, while the telomeric region with copy number 2 (cyan) has a loss heterozygosity, suggesting a copy neutral LOH event.

the same maximum likelihood, allowing users to explore different maximum likelihood solutions. Thus, THetA is an attractive alternative to these methods.

We show the advantages of THetA on 3 breast cancer genomes sequenced in [114]: one sequenced at ∼188X coverage and two at ∼40X coverage. Nik-Zainal *et al.* [114] showed how a large amount of information about a tumor's evolutional history can be derived by analyzing clonal and subclonal mutations in high-coverage sequencing data. Our THetA algorithm automates some of the manual analysis involved in such reconstructions. For the ∼188X genome, our results are largely concordant with the extensive analysis and annotation of this sample in [114]. THetA automatically recovers nearly all of the copy number aberrations reported in [114], but with some differences in the classification of aberrations of clonal or subclonal. Allele data not used by THetA provides external evidence that support the THetA results in several cases. On one of the ∼40X coverage genomes, we identify two previously unreported tumor subpopulations in nearly equal proportions, as well as 24% normal admixture. These results are supported by statistical comparisons of read depth ratios, and also allow us to identify copy-neutral LOH on chromosome 8q. Thus, we demonstrate that it is possible to successfully identify multiple tumor populations in a single sample when considering a subset of genomic intervals. Further, we do so for a ∼40X sequenced tumor, demonstrating the ability to identify intra-tumor heterogeneity at sequence coverages that are the current standard in cancer sequencing studies.

While we have demonstrated several of the strengths of THetA, the algorithm (as presented here) does have some limitations. First, the reliance on copy number aberrations means that THetA is unable to identify tumor subpopulations that do not contain copy number aberrations. As copy number aberrations are ubiquitous in many types of cancers, particularly solid tumors, we expect that THetA will prove useful for analyzing a wide range of different cancer samples. Second, THetA is designed for use with whole-genome data and is not compatible with other popular datatypes such as whole-exome sequence data. Finally, THetA's computation time increases with more subpopulations, making analysis of genomes with multiple tumor populations more difficult. In the Chapter 3 we present THetA2 which allows us to overcome several of these limitations.

Our focus in the development of THetA was to address rigorously the difficult problem of analyzing tumor purity and subclonal copy number aberrations from DNA sequencing data. A logical next step is to use the output of THetA to help predict single nucleotide mutations in tumor samples and/or assess the clonality of somatic mutations, both challenging problems on their own. Carter

et al. [27] and Nik-Zainal *et al.* [114] show that once tumor purity is correctly estimated, then this value can be used to analyze the clonality/subclonality of somatic mutations. Incorporating the additional signal of variant allele frequencies into the probabilistic model, as well as extending the model to allele-specific copy number changes [40] are important directions for future work. Ultimately, a desirable goal is to integrate into a single probabilistic framework the detection of all types of somatic aberrations (single nucleotide, copy number, and rearrangements) with the estimation of tumor purity and the derivation of tumor subpopulations. Next, further algorithmic improvements in THetA would help in the analysis of more complicated tumor samples that have more intervals (e.g. smaller copy number aberrations), higher amplitude copy number aberrations, more subpopulations, or more complicated rearrangements; e.g. due to breakage/fusion/bridge (B/F/B) cycles [183], chromothripsis [83], or extrachromosomal amplifications [135]. THetA is polynomial time for a mixture of two genomes with intervals of equal weight, but the question of the complexity of the MLMDP for $n > 2$ remains open. Finally, the number and scope of datasets that THetA can analyze would grow significantly if the algorithm were modified to work with other datatypes.

A number of other techniques have recently been used to study intra-tumor heterogeneity. For example [134] uses expression profiles across different individuals to identify differentially expressed genes with respect to healthy cells at the cancer site of origin. Single cell sequencing and multi-region sequencing from a primary tumor are alternative strategies that have been successfully employed [110, 109, 114, 14, 176, 71, 53]. As these technologies improve they will likely further contribute to our understanding of intra-tumor heterogeneity. However, sequencing of primary tumor samples as well as matched tumor/metastasis samples will remain a dominant protocol for some time. Thus, algorithms such as THetA, ABSOLUTE, ASCAT, and others that can derive information about intra-tumor heterogeneity from DNA sequencing of tumor samples provide a useful complement to other technologies/techniques for tumor heterogeneity studies.

Chapter 3

An Improved Approach to Quantifying Intra-tumor Heterogeneity

In Chapter 2 we introduced Tumor Heterogeneity Analysis (THetA) – an algorithm to infer the most likely collection of genomes and their proportions in a single sample from a heterogeneous tumor when copy number aberrations distinguish subpopulations. Since the original paper describing THetA [118] was published, the area of algorithm development for inferring tumor composition from sequencing data of single samples of a tumor has remained extremely active [105, 90, 92, 143, 132, 91]. Data in the form of a single sequenced sample from a tumor remains widely available, especially through large-scale efforts such as TCGA or ICGC. Thus, there is a continued need for better methods to infer tumor composition from a single sample of a tumor.

In this chapter we explore several avenues for improved inference of tumor composition from a single mixed tumor sample. While THetA was able to overcome some of the limitations of previous methods, it still has several drawbacks. In this chapter we introduce THetA2, which extends the original THetA algorithm in a number of important directions. We apply THetA2 to both whole-genome (including low-pass) and whole-exome sequence data from 18 samples from The Cancer Genome Atlas (TCGA). We find that the improved algorithm is substantially faster and able to analyze highly rearranged genomes - identifying numerous tumors with subclonal tumor populations

in the TCGA data. Where available, we compare our purity estimates to published values for another widely used algorithm, ABSOLUTE [27]. While the purity estimates are largely in agreement for higher purity samples, we find cases where ABSOLUTE fails or underestimates purity, but THetA2 identifies multiple tumor subpopulations. These improvements greatly expand the range of sequencing data and tumors for which we can infer tumor composition.

Much of the work in this chapter was a collaboration with Gryte Satas and is taken from [121]. This algorithm was presented at the 2014 Intelligent Systems for Molecular Biology special interest group HitSeq: High-throughput Sequencing Algorithms and was accepted as a platform presentation at the 2014 TCGA Scientific Symposium.

3.1 Related Work

Many recent studies indicate that most tumor samples are a heterogeneous mixture of cells, including admixture by normal (non-cancerous) cells and subpopulations of cancerous cells with different complements of somatic aberrations [114, 53]. As discussed in Chapter 2, characterizing this *intra-tumor heterogeneity* is essential for many reasons. We briefly recap these reasons here. First, an estimate of tumor *purity*, the fraction of cancerous cells in a tumor, is necessary for accurate identification of somatic aberrations of all types in the sample. Most cancer genome sequencing studies use a re-sequencing approach to detect somatic aberrations. Reads from a tumor sample (and usually a matched normal sample) are aligned to the human reference genome. Differences in the sequence of aligned reads, the number of aligned reads, or the configuration of aligned reads (e.g. split reads or discordant pairs) are used to infer the presence of single-nucleotide or other small variants, copy number aberrations, or structural aberrations respectively [103, 43]. However, presence of intra-tumor heterogeneity can dilute the signals required to identify somatic aberrations.

Second, estimates of the *composition* of a tumor sample – including not only the tumor purity, but also the number and fractions of subpopulations of tumor cells – provide useful for understanding tumor progression and determining possible treatment strategies [108, 56]. In particular, *clonal* somatic aberrations that exist in all tumor cells are likely early mutational events and their identification sheds light on the early stages of cancer. Conversely, *subclonal* somatic aberrations might reveal properties shared by a subset of tumor cells such as drug resistance or ability to metastasize. Identification of such aberrations and subpopulations of tumor cells might inform treatment

strategies, and/or help predict metastasis/relapse.

Many methods to infer tumor purity and/or tumor composition have been developed (Table 2.1). Traditionally, these methods generally fall into two categories: (1) methods that use somatic single nucleotide variants (SNVs) and (2) methods that use somatic copy number aberrations. SNV based methods such as EXPANDS [8], PyClone [143], SciClone [105] and many others [75, 87] use clustering of variant allele frequencies to determine tumor populations and frequencies. While these types of methods are able to derive multiple tumor subpopulations, they often require estimates of copy number for each region containing SNVs. Deriving such estimates for highly rearranged, aneuploid tumors is as difficult as the estimation of intra-tumor heterogeneity itself. Moreover, these approaches require high coverage sequencing to overcome the high variance in read counts at individual SNVs. For example both PyClone [143] and PhyloSub [75] explicitly require deeply sequenced data. Thus, less expensive low-coverage sequence data, as generated in TCGA [23] is not amenable to these approaches.

Copy number based methods such as ABSOLUTE [27] and CNAnorm [62] use observed shifts in read depth due to copy number aberrations to predict tumor purity, but do not explicitly consider multiple tumor subpopulations, and therefore may return purity estimates that only reflect a single subpopulation of tumor cells in a sample. In chapter 2 we described the Tumor Heterogeneity Analysis (THetA) algorithm to infer the composition of a tumor sample – including both the percentage of normal admixture and the fraction and content of one *or more* tumor subpopulations that differ by copy number aberrations [118]. While THetA represented an advancement over existing methods at the time of publication, the method as originally published had limitations. For instance, the algorithm required long run-times when considering multiple tumor populations, was only adapted for whole-genome sequencing data (rather than the widely popular whole-exome sequencing protocol) and only utilized read depth information rather than also incorporating alternative data signals.

Finally, several methods that utilize both SNV and CNA data to infer tumor composition have recently been developed [39, 91]. However, integrating these different data signals is not a straight forward process. For instance, PhyloWGS[39] encodes copy number aberrations (already predicted by another method) as SNVs. Thus, PhyloWGS not only requires that CNA be called a priori, but also assumes equal weight for every mutation, regardless of the fact that a CNA may be supported by many more reads than an SNV. Thus, there still remains many questions about how to utilize

multiple types of input data in order to accurately infer tumor composition.

3.2 The THetA2 Algorithm

In this section we describe the THetA2 algorithms, which extends the THetA algorithm (presented in chapter 2) in several important directions. First, we substantially improve the computation for the case of multiple distinct tumor subpopulations in a sample. Second, we extend THetA2 to infer tumor composition for highly rearranged genomes using a two-step procedure where initial estimates are made using high confidence regions of the genome, and then are extended to the entire genome. Third, we devise a probabilistic model of B-allele frequencies, which can be used to solve the identifiability issue when read depth alone is consistent with multiple possible tumor compositions. Lastly, we extend THetA2 to analyze whole-exome sequencing data. Thus, THetA2 is applicable to a much wider array of data than the original algorithm.

3.2.1 Notation and Problem Formulation

We assume that the reference genome is partitioned into a sequence $\mathbf{I} = (I_1, \ldots, I_m)$ of non-overlapping intervals, according to changes in the density, or depth, of reads aligning to each position in the reference [174]. Given $\mathbf{I}$, we define a corresponding *read depth vector* $\mathbf{r} = (r_1, \ldots, r_m) \in \mathbb{N}^m$ where r_j is the number of reads with a (unique) alignment within I_j. A cancer genome is defined by an *interval count vector* $\mathbf{c} \in \mathbb{N}^m$, where c_j is the integer number of copies of interval I_j in the cancer genome.

A tumor sample $\mathcal{T}$ is a mixture of cells that contain different collections of somatic mutations, and in particular somatic copy number aberrations. Each subpopulation has a distinct interval count vector representing the genome of the subpopulation. Following the model introduced in [118] we represent $\mathcal{T}$ by: (1) an *interval count matrix* $\mathbf{C} = [c_{j,k}] \in \mathbb{N}^{m \times n}$ where $c_{j,k}$ is the number of copies of interval I_j in the k^{th} distinct subpopulation; and (2) a *genome mixing vector* $\mu \in \Delta_{n-1} = \{(\mu_1, \ldots, \mu_n)^T \mid \sum_{j=1}^{n} \mu_j = 1, \text{ and } \mu_j \geq 0 \text{ for all } j\}$ where μ_k is the percentage of cells in $\mathcal{T}$ that belong to the k^{th} distinct subpopulation.

Let the interval count matrix $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_n)$, where $\mathbf{c}_j$ is the j^{th} column of $\mathbf{C}$. We assume that $\mathbf{C}$ satisfies three constraints. (1) The first column $\mathbf{c}_1 = 2^m$ so that the first component of the tumor sample is the normal genome. (2) The number n of subpopulations is less than the number m of

intervals. (3) The copy numbers of the intervals are bounded below by 0 and above by a maximum copy number $k \geq 2$. Thus, $\mathbf{C} \in \{0, \ldots, k\}^{m \times n}$. We define $\mathcal{C}_{m,n,k}$ to be the set of all such $\mathbf{C}$, and define $\Omega_{m,n,k} = \{(\mathbf{C}, \mu) \mid \mathbf{C} \in \mathcal{C}_{m,n,k},\ \mu \in \Delta_{n-1}\}$ to be the domain of pairs $(\mathbf{C}, \mu)$ satisfying all constraints.

We model the observed read depth vector $\mathbf{r}$ using a multinomial probability distribution with parameter $\mathbf{p} = (p_1, \ldots, p_m)$, where p_j is the probability that a randomly chosen read will align to interval I_j. A pair $(\mathbf{C}, \mu)$ defines a value for the multinomial parameter $\mathbf{p} = \widehat{\mathbf{C}\mu} = \frac{\mathbf{C}\mu}{|\mathbf{C}\mu|_1}$. Thus, the negative log likelihood $L(\mathbf{C}, \mu|\mathbf{r}) = -\log(\text{Mult}(\mathbf{r}; \widehat{\mathbf{C}\mu}))$ is the negative log of the multinomial probability of observing counts $\mathbf{r}$ in the intervals given the probability of a read aligning to an interval is defined by $\widehat{\mathbf{C}\mu}$. The goal is to find the interval count matrix $\mathbf{C}^*$ and genome mixing vector μ^* that minimize the negative log likelihood:

$$(\mathbf{C}^*, \mu^*) = \text{argmin}_{(\mathbf{C},\mu) \in \Omega_{m,n,k}} L(\mathbf{C}, \mu; \mathbf{r}) \tag{3.1}$$

3.2.2 Interval Count Matrix Enumeration

In this section, we derive an improved procedure to solve the optimization problem (3.1). In [118] we showed that the function $L(\mathbf{C}, \mu; \mathbf{r})$ is a convex function of μ. Thus, for a fixed interval count matrix $\mathbf{C}$, the optimal value of μ can be computed efficiently. In the important special case of a mixture of normal cells and a single tumor population ($n = 2$), we reduce the domain of interval count matrices $\mathbf{C}$ to a set whose size is polynomial in m and guaranteed to contain the optimal $\mathbf{C}^*$. This set is easy to enumerate and we obtain an efficient algorithm. However, when a tumor sample contains multiple tumor subpopulations ($n > 2$) the algorithm in [118] enumerates all $\mathbf{C} \in \mathcal{C}_{m,n,k}$ and checks whether each such $\mathbf{C}$ satisfies a particular ordering constraint that is a necessary, but not sufficient, condition for the optimal $\mathbf{C}^*$.

In this section we derive a algorithm that explicitly enumerates only those matrices $\mathbf{C}$ that satisfy a more restrictive necessary ordering constraint for a mixture of *any number* of tumor genomes. All proofs are contained in Appendix B.

Compatible Order

We say that vectors $\mathbf{v} = (v_1, \ldots, v_m)$ and $\mathbf{w} = (w_1, \ldots, w_m) \in \mathbb{R}^m$ have *compatible order* provided all $1 \leq i, j \leq m$, $v_i \leq v_j$ if and only if $w_i \leq w_j$. In [118] we proved that if $(\mathbf{C}^*, \mu^*)$ is optimal (i.e.

satisfies Equation (3.1)) then $\widehat{\mathbf{C}^*\mu*}$ and $\mathbf{r}$ have compatible order. We define $\mathcal{S}_{m,n,k}$ to be the set of matrices $\mathbf{C} \in \mathcal{C}_{m,n,k}$ that satisfy this ordering constraint: i.e. $\mathcal{S}_{m,n,k} = \{\mathbf{C} \mid \mathbf{C} \in \mathcal{C}_{m,n,k}$ and $\exists \mu \in \Delta_{n-1}$ such that $\widehat{\mathbf{C}\mu}$ is in compatible order with $\mathbf{r}\}$. Thus, to find the optimal solution $(\mathbf{C}^*, \mu^*)$, it is sufficient to examine matrices $\mathbf{C} \in \mathcal{S}_{m,n,k}$.

Without loss of generality, we assume that the read depth vector $\mathbf{r} = (r_1, \ldots, r_m)$ satisfies $r_1 \leq r_2 \leq \cdots \leq r_m$. Thus, the set $\mathcal{S}_{m,n,k} = \{\mathbf{C} \in \mathcal{C}_{m,n,k} | \, (\mathbf{C}\mu)_1 \leq (\mathbf{C}\mu)_2 \leq \cdots \leq (\mathbf{C}\mu)_m$ for some $\mu \in \Delta_{n-1}\}$.

For a matrix $\mathbf{C} \in \mathcal{C}_{m,n,k}$, the set of μ which result in a compatible ordering can be calculated using the function $\Phi(\mathbf{C})$:

$$\Phi(\mathbf{C}) = \bigcap_{j=1}^{m-1} \{\mu | \mu \in \Delta_{n-1} \text{ such that } (\mathbf{C}\mu)_j \leq (\mathbf{C}\mu)_{j+1}\}. \tag{3.2}$$

Thus, a matrix $\mathbf{C} \in \mathcal{S}_{m,n,k}$ if and only if $\Phi(\mathbf{C})$ is not empty. Corollary 3.2.1 follows directly from Equation (3.2).

Corollary 3.2.1. *Suppose $\mathbf{C} \in \mathcal{C}_{m,n,k}$. If there exists an $i \in \{1, \ldots, m-1\}$ such that for all $t \in \{2, \ldots, n\}, c_{i,t} \geq c_{i+1,t}$ and there exists a $t \in \{2, \ldots, n\}$ such that $c_{i,t} > c_{i+1,t}$, then $\Phi(\boldsymbol{C}) = \emptyset$.*

Using a graph to enumerate $\mathcal{S}_{m,n,k}$

We now present an algorithm to enumerate $\mathcal{S}_{m,n,k}$ for $n \geq 2$. Consider a complete (including self loops) directed graph $G_{n,k}$, with a vertex for each possible row in a matrix in $\mathcal{C}_{m,n,k}$. Paths on $G_{n,k}$ of length $m-1$ correspond to matrices in $\mathcal{C}_{m,n,k}$ (See Figure 3.1 and Figure B.1).

To enumerate the subset of paths on $G_{n,k}$ which correspond to matrices in $\mathcal{S}_{m,n,k}$, we use a depth first search. While building paths, we calculate the set Φ for the matrix implied by the current path, and only proceed down branches which do not result in the empty set (see Appendix B). As a result, we are guaranteed to enumerate only the matrices in $\mathcal{S}_{m,n,k}$.

Corollary 3.2.1 allows us to reduce the graph $G_{n,k}$ by showing that there are certain edges which will never appear in paths which correspond to matrices in $\mathcal{S}_{m,n,k}$ and thus can be removed from the graph prior to matrix enumeration. In the case where $n = 3$, the calculation of Φ is reduced to a problem in a single variable, $\frac{\mu_2}{\mu_3}$.

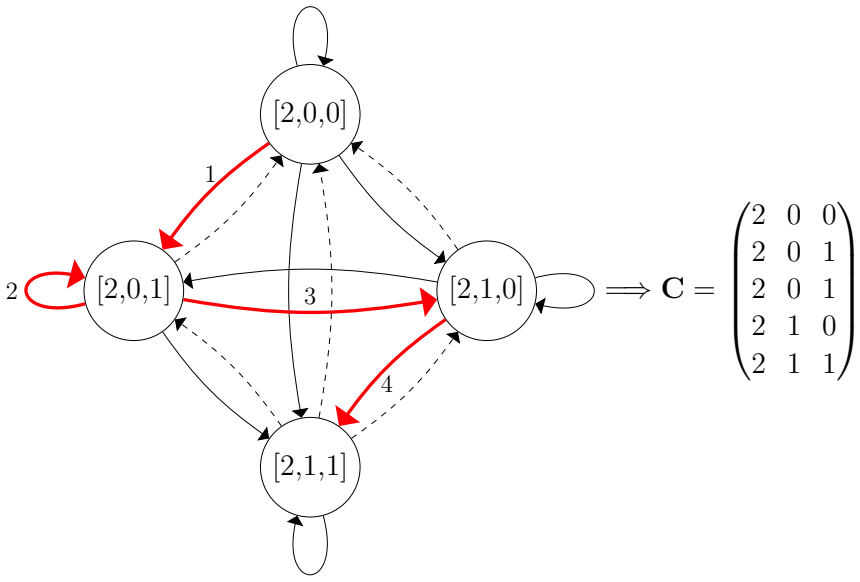


Figure 3.1: **An example of the graph used by THetA2 to enumerate matrices C.** Specifically, the graph $G_{3,1}$ is used to enumerate the matrices $\mathcal{S}_{m,3,1}$ as a subset of the paths of length $m-1$. The dashed edges can be removed by applying Corollary 3.2.1. The highlighted path corresponds to the matrix on the right.

3.2.3 A Two-Step Procedure for Genome-Wide Inference of Copy Numbers

In [118] we inferred tumor composition using a relatively coarse interval partition $\mathbf{I}$ of the reference genome, considering only large copy number aberrations. As a result, the published approach could not readily be applied to highly rearranged genomes that are segmented into many intervals. Moreover, manual selection of a subset of intervals was typically required when analyzing samples containing multiple tumor populations. Even with the improved enumeration procedure described in the previous section, when more than one tumor subpopulation is considered, the number of matrices $\mathbf{C}$ that need to be enumerated is exponential in the number m of intervals. Moreover, the number of matrices $\mathbf{C}$ is also exponential in the maximum copy number state k considered, making analysis of genomes with extensively amplified regions more difficult.

In this section we present a two-step procedure for interval selection that overcomes the limitations stated above, and allows us to infer the composition of highly rearranged genome that are highly fragmented and/or contains amplified segments with more than k copies. Our two-step procedure consists of the following steps: (1) Select a set of high-confidence intervals, and determine

the most likely $\mathbf{C}$ and μ for those intervals. (2) Use the estimates of $\mathbf{C}$ and μ to determine copy numbers for all other intervals in $\mathbf{I}$ not used in the first step, thus allowing for analysis of both highly amplified regions and very fragmented genomes.

Interval Selection

We automate the selection of a subset of high confidence intervals used to determine the optimal $(\mathbf{C}^*, \mu^*)$ for those intervals. Further details are included in Appendix B. Briefly, we partition $\mathbf{I}$ into two sets of intervals: (1) $\mathbf{I}_H$ - high confidence intervals; (2) $\mathbf{I}_L$ - lower confidence intervals. $\mathbf{I}_H$ is selected to contain up to a fixed integer d longest intervals from $\mathbf{I}$ such that each interval selected is longer than a predetermined minimum length and is not obviously amplified beyond the specified max copy number k. $\mathbf{I}_L$ contains all remaining intervals from $\mathbf{I}$. Additionally, $\mathbf{I}_H$ must represent $> 10\%$ of the total length of all provided intervals, otherwise the sample is determined not to be a good candidate for analysis using THetA2. Once $\mathbf{I}_H$ and $\mathbf{I}_L$ have been selected, we use the improved THetA2 algorithm described in the previous section to calculate $\mathbf{C}^*_H$ and μ^*_H for just the intervals in $\mathbf{I}_H$.

Determining Additional Copy Numbers: Single Row

Given $(\mathbf{C}^*_H, \mu^*_H)$ predicted for high confidence intervals $\mathbf{I}_H$, we infer copy numbers for the remaining intervals $\mathbf{I}_L$. We start with the simplifying assumption that $|\mathbf{I}_L| = 1$. We prove the following theorem.

Theorem 3.2.2. *Let $\mathbf{C} = [c_{i,j}]$ be an interval count matrix. $L(\mathbf{C}, \mu|\mathbf{r})$ is a convex function of $c_{i,j}$.*

We use Theorem 3.2.2 to find the optimal real valued solution for the $c_{i,j}$'s corresponding to the single interval $\mathbf{I} \in \mathbf{I}_L$, given $\mathbf{C}^*_H$ and μ^*_H. We then check the surrounding integer values to find the integral solution, which, by convexity, is guaranteed to find the optimal integer solution.

Determining Additional Copy Numbers: Multiple Rows

In the previous section we showed how to find the optimal copy number for a single additional interval in $\mathbf{I}_L$ given optimal values $\mathbf{C}^*_H$, and μ^*_H for a set of high confidence intervals $\mathbf{I}_H$. To estimate copy numbers when $I \in \mathbf{I}_L$ contains more than one interval we estimate the optimal copy numbers for each interval in $I \in \mathbf{I}_L$ when appended to $\mathbf{C}_H$ individually as described in the previous section, and

then jointly append all inferred copy numbers to $\mathbf{C}_H^*$ to obtain a new matrix $\mathbf{C}_{H\cup L}$. We then return the solution $(\mathbf{C}_{H\cup L}, \mu_H^*)$. We note that this approach provides no guarantee for finding the optimal copy numbers across *all* $I \in \mathbf{I}_L$ given $\mathbf{C}_H^*$, and μ_H^*. However, in practice we find that the solutions returned by our procedure are generally very similar to this optimum (Table B.1).

3.2.4 Model Selection

As in [118], we use the Bayesian information criterion (BIC) to select from different sized models (that is, different numbers n of tumor populations) and their corresponding maximum likelihood solutions. We use the standard BIC of $-2log(L) + a\log(b)$ where L is the likelihood of a solution, $a = (m+1)(n-1)$ is the number of free parameters in the model and b is the number of data points (the total number of tumor and normal reads). In contrast, [118] used a modified BIC that more strongly penalized solutions with more tumor populations. Such a modification is not necessary here, as our improved algorithm considers copy number data across the entire genome, rather than only a small number of intervals, reducing the possibility of overfitting. Thus, we are able to more robustly identify samples with multiple subpopulation of tumor cells.

3.2.5 Probabilistic Model of B-allele Frequencies

THetA2 may return multiple equally like pairs $(\mathbf{C}, \mu)$ when using read depth alone. We derive a probabilistic model of B-allele frequencies (BAFs) – the fraction of reads containing the minor allele – that may be used to distinguish between multiple pairs $(\mathbf{C}, \mu)$. Let $\mathbf{v} = (v_1, v_2, \ldots, v_q)$ be the observed BAFs for q heterozygous germline SNPs in the normal sample and $\mathbf{w} = (w_1, w_2, \ldots, w_q)$ be the corresponding BAFs from the tumor genome. We model $\mathbf{w}$ as being drawn from Gaussian distributions whose parameters depend upon $\mathbf{v}$, $\mathbf{C}$ and μ. We then select the $(\mathbf{C}, \mu)$ which maximizes the likelihood of the observed BAFs in the tumor sample:

$$L(\mathbf{C}, \mu|\mathbf{v}, \mathbf{w}) = P(\mathbf{w}|\delta, \sigma^2) = \prod_{i=1}^{q} \mathcal{N}(w_i|sgn(0.5 - w_i)\delta_i, \sigma_i^2) \tag{3.3}$$

Here σ_i^2 is the observed variance for all heterozygous SNPs in $\mathbf{v}$ that lie within interval I_j and δ_j is the expected BAF deviation away from 0.5 given $\mathbf{C}$ and μ. See Appendix B for further details.

3.2.6 Application to Whole-Exome Data

Lastly, we extend THetA2 for whole-exome data, where only the coding regions of the genome have been targeted for sequencing. From whole-exome data we need to infer the following two values: (1) A set of non-overlapping intervals $\mathbf{I} = (I_1, \ldots, I_m)$ in the reference genome; and (2) A corresponding read depth vector $\mathbf{r} = (r_1, \ldots, r_m)$.

To infer the interval partition $\mathbf{I}$ we rely on recently developed algorithms such as ExomeCNV [145] and EXCAVATOR [95] for segmentation and detection of copy number aberrations from whole-exome data. The segmentation returned by one of these algorithms may contain gaps rather than being a complete partition of the reference genome, but still provides a set of non-overlapping intervals that may be used as input to THetA2. We note that some methods utilize normalization procedures for GC content, mappabability and even exon length and this information is therefore implicitly incorporated into the input provided to THetA2.

We compute the read depth vector $\mathbf{r} = (r_1, \ldots, r_m)$ for whole-exome data as follows. Given a set $\mathbf{I}$ of non-overlapping intervals in the reference genome, a set $\mathbf{E}$ of exons in the reference genome and a read length ℓ we set $r_j = \frac{x_j}{\ell}$ where x_j is the total number of sequenced nucleotides that have a unique alignment to some exon $e \in \mathbf{E}$ within interval I_j. Thus, r_j is approximate count of the number of reads aligning to some exon located in interval I_j.

3.3 Results

We ran THetA2 on simulated data, and whole-genome (including low-pass data 5-7X coverage) and whole-exome data from 18 breast carcinoma, ovarian carcinoma, glioblastoma multiforme, kidney renal clear cell and lung squamous cell carcinoma samples from TCGA (Table B.2). Where available, we compare our estimates of tumor purity to the estimates reported by the ABSOLUTE algorithm [27] that estimates purity from SNP array data.

The rest of this section is organized as follows. First we discuss results on simulated data. Second, we demonstrate THetA2's performance on whole-exome data, including comparison of results for samples for which both whole-genome and whole-exome data was available. Next, we present in-depth analysis of several whole-genome samples in order to demonstrate the efficacy of THetA2 on highly rearranged genomes, using both low-pass and moderate coverage sequence data. Lastly, we apply our probabilistic model of B-allele frequencies to one sample and disambiguate between two

equally likely solutions.

3.3.1 Simulated Data

We tested THetA2 on simulated data in order to demonstrate the improvements in THetA2 over the original THetA as well as ABSOLUTE. We created simulated mixtures using real sequencing data from an AML tumor sample and matched normal sample (TCGA-AB-2965) from [26]. This sample was chosen due to its high purity (approximately 95% pure) and lack of copy number aberrations as predicted by array data, providing high confidence that our simulated mixture and implanted copy number aberrations are not confounded by impurity and aberrations in the real data. Simulated mixtures are created by implanting random amplifications and deletions (see Appendix B) to create different tumor populations, and then creating a mixture representing different tumor compositions.

Mixtures with 3 subpopulations

We find that THetA2 computes the optimal solution orders of magnitude faster than the original THetA (Figure 3.2a). Using 30X simulated data THetA2 demonstrates consistent accuracy at estimating μ (error < 0.05) and copy numbers in the larger tumor population (error < 0.1). In addition, the accuracy in estimating copy numbers improves for the smaller tumor population as its proportion increases (Figure 3.2b). Furthermore, THetA2 has increased performance at estimating copy numbers for both populations when only considering longer intervals. For example, when the smaller subpopulation comprises 0.3 of cells in the sample and we only consider intervals longer than 5Mb, the error rate for both populations drops below 0.06. We see similar trends using 7X simulated data, but the lower coverage results in slightly worse copy numbers estimates (Figure B.2).

We also directly compare THetA2 to the original THetA on this simulated data. The two-step method enables THetA2 to infer copy numbers for 100% of the genome compared to only 6-11% of the genome with the original THetA (Figure B.3). The expanded fraction of the genome analyzed also translates into a substantial increase in the fraction of genome with correct copy numbers estimates. In our simulations THetA2 correctly infers copy numbers for the larger and smaller tumor subpopulation in $83\% - 87\%$ and $28\% - 72\%$ more of the genome, respectively, than THetA (Figure B.4).

Using the simulated mixture in Figure 3.3A, we demonstrate that the improved enumeration procedure in combination with the two-step can lead to improved estimates of both μ and $\mathbf{C}$.

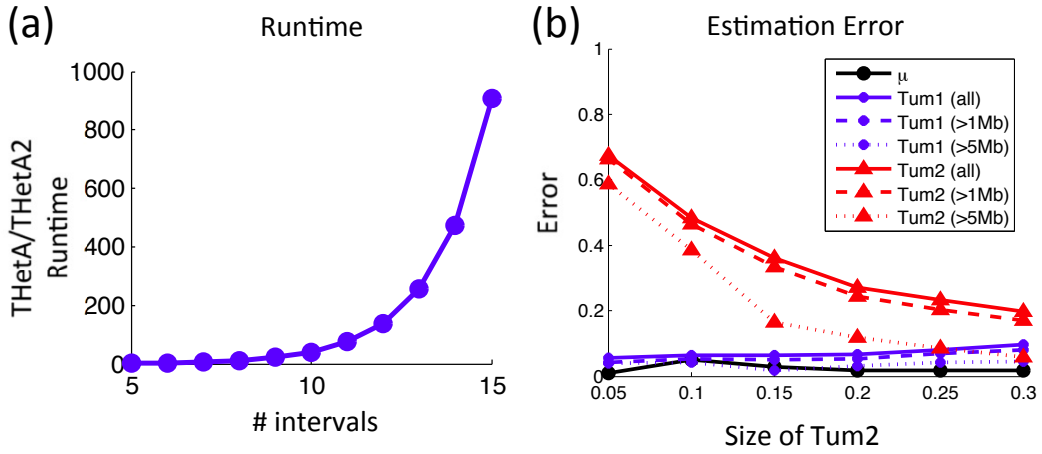


Figure 3.2: **Runtime comparison and estimation error for THetA2 on simulated data containing a mixture of normal cells and two tumor subpopulations**. **(a)** The ratio of runtimes for the old and new enumeration procedures as a function of the number of intervals used in the first step of the algorithm. **(b)** Estimation error for both μ and $\mathbf{C}$ for each tumor population (Tum1 and Tum2) as the proportion of Tum2 increases and the proportion of Tum1 is fixed at 0.5. Error for μ is the euclidean distance from the true value and error for each tumor population is the fraction of the genome for which the copy number is incorrectly inferred for the all copy number estimates, and when only considering intervals that are longer than 1Mb and 5Mb.

On this mixture THetA2 is able to reconstruct both tumor populations with accuracy above 0.87 (Figure 3.3B) across the entire genome. However, because THetA is only able to consider a small fraction of the genome, when applied to this mixture it has increased error at estimating μ and completely misestimates the smaller tumor subpopulation with error of 0.95 across the regions for which copy number estimates were made and error 0.99 across the whole genome (Figure 3.3C). We also applied ABSOLUTE [27] to this mixture, run with default parameters, using the same partition of the genome output by BIC-seq [174]. ABSOLUTE returns a collection of 12 different solutions, each with a different purity and likelihood (Figure 3.3D). The most likely solutions returned by ABSOLUTE underestimate purity by at least 0.28 and estimated a tetraploid solution, while the true sample has mean ploidy 1.75 and 1.77 in the two tumor populations. Further details are located in Appendix B.

Mixtures with 4 subpopulations

To demonstrate the extensibility of the model to greater numbers of subpopulations, we create a simulated $30X$ coverage mixture containing 4 distinct subpopulations. Due to the increased

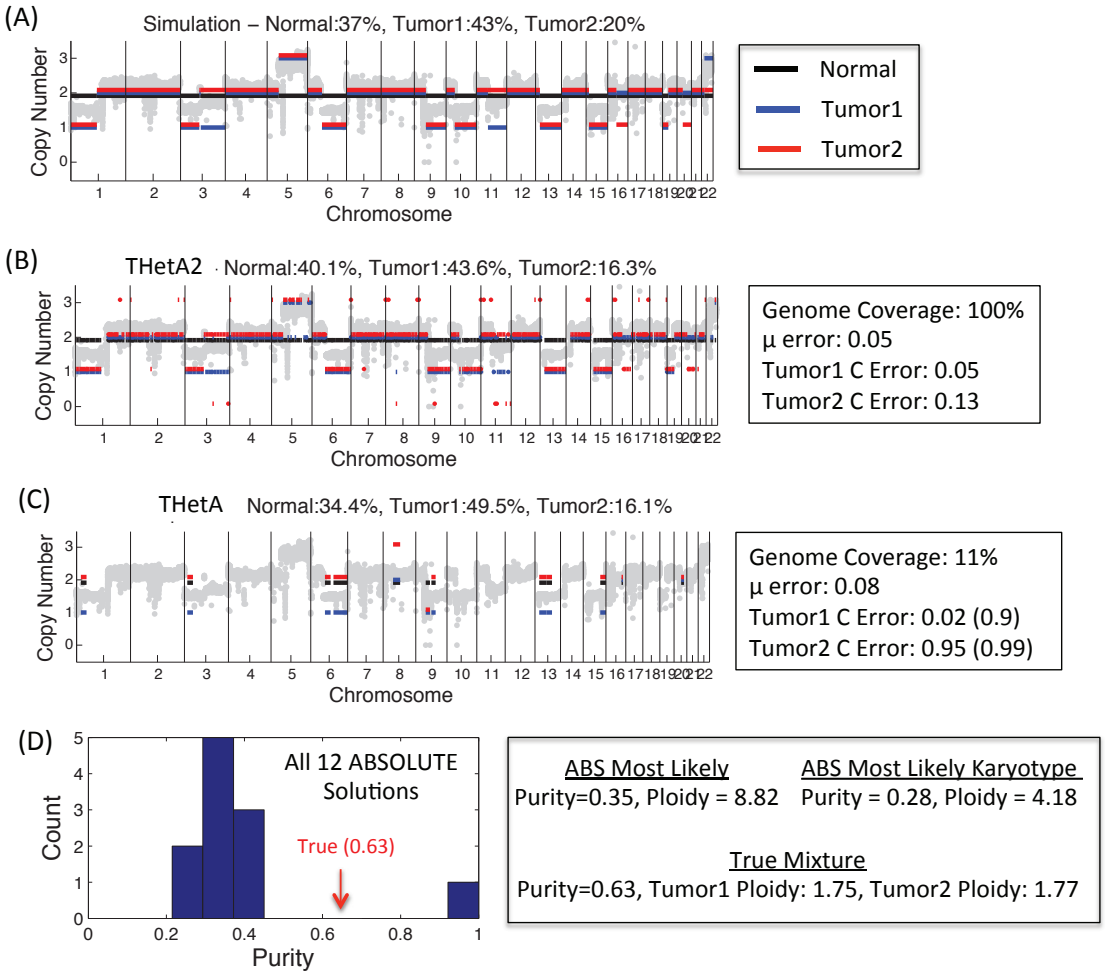


Figure 3.3: **Comparison of THetA2, THetA and ABSOLUTE on a simulated mixture of 3 subpopulations. (A)** True simulated mixture including read depth ratios (gray) over 50 kb bins and the true copy numbers for a mixture of normal cells (black) and two tumor subpopulations (blue and red). **(B)** Tumor composition inferred by THetA2 using default parameters. Genome coverage is the fraction of the genome for which copy number estimates are made. μ error is euclidean distance from the true μ and **C** error is the fraction of the genome with the incorrect copy number estimate. **(C)** Similar to (b) but shows composition inferred by the original THetA and also shows **C** error across both predicted regions and whole-genome. **(D)** (left) Histogram of all 12 purity estimates output by ABSOLUTE. (right) The purity and ploidy reported in the most likely and most likely using only Karyotype solutions output by ABSOLUTE.

runtime when considering larger numbers of subpopulations, we employ an alternative segmentation procedure to reduce the total number of intervals (see Appendix B for details). We find that on this simulation, THetA2 was able to estimate μ with 0.05 error, comparable to the accuracy achieved for smaller numbers of subpopulations, and was able to correctly infer copy number for 99.6% of the intervals considered, with the tradeoff of only considering 87.6% of the total genome (Figure B.5). Further, we demonstrate how the output of THetA2 changes when the number of subpopulations (n) is fixed below the true number of subpopulations. In particular, we show that in this case THetA2 still provides useful information about the true mixture (Figure B.6).

3.3.2 Extension to Whole-Exome Sequencing Data

To demonstrate THetA2's effectiveness on whole-exome data we ran THetA2 on Illumina whole-exome data for the subset of 16 of the 18 tumor samples from TCGA for which whole-exome data was available (Table B.2). For each sample we used both ExomeCNV [145] and EXCAVATOR [95] with default parameters to determine an interval partition $\mathbf{I}$ (see Figure B.7 for the complete whole-exome workflow). If we assume that the tumor sample is a mixture of normal cells and a single tumor population, then the purity estimates obtained by THetA2 on the ExomeCNV and EXCAVATOR interval segmentations were similar for most samples (Figure B.8). The two exceptions were two tumor samples where we find subclonal copy number aberrations (for one example see Figure B.9). We found that the presence of subclonal aberrations can result in estimates of purity that are artificially low. For example, a segmentation may not accurately distinguish all present subclonal aberrations. Thus in the results below, we use the THetA2 solution with higher purity estimate from the ExomeCNV and EXCAVATOR segmentations. Further details are in Appendix B.

Comparison of THetA2 with ABSOLUTE

On most samples THetA2 purity estimates are within 0.08 of the estimates reported by the ABSOLUTE algorithm [27] (Figure 3.4a). One example is glioblastoma sample TCGA-06-0214, for which we estimate purity of 0.67 compared to 0.66 reported by ABSOLUTE. However, while the purity estimates are similar, THetA2 is additionally able to identify two subpopulations of tumor cells, in 46.4% and 20.1% of cells in sample (Figure 3.4b) and determine which copy number aberrations are part of each subpopulation.

There are two samples where THetA2 purity estimates are not in agreement with those reported

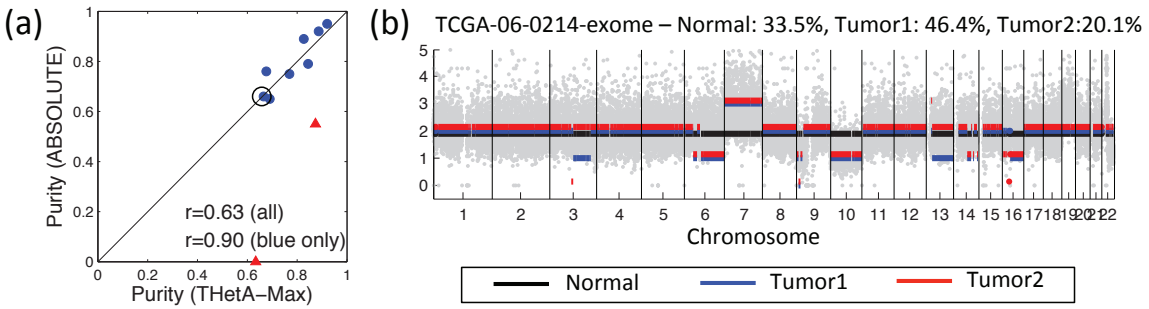


Figure 3.4: **THetA2 results on whole-exome (WXS) data. (a)** Comparison of purity estimates by THetA2 and ABSOLUTE (as reported in [27]). With exception of two outlier samples (red triangles; TCGA-29-1768 and TCGA-06-0188), both approaches predict similar estimates on high purity samples: $r = 0.9$ from Pearson correlation coefficient. Circled sample is TCGA-06-0214, for which both methods agree on sample purity. **(b)** Tumor composition inferred by THetA2 on glioblastoma multiforme sample TCGA-06-0214. Read depth ratios (gray) over 50 kb bins and the copy numbers (for all intervals $>$ 2Mb) inferred by THetA2 for a mixture of normal cells (black) and two tumor subpopulations (blue and red). We detect rearrangements common to glioblastoma multiforme [156] such as amplification of chromosomes 7, and loss of chromosomes 6q, 9p, 10, 13q and 14q.

for the ABSOLUTE algorithm [27] (Figure 3.4a). The first is ovarian carcinoma sample TCGA-29-1768 where we infer multiple tumor subpopulations and report a purity of 0.87 compared to 0.55 reported by [27]. Notably, one of the tumor subpopulations returned by THetA2 is in 54% cells. A possible explanation is that ABSOLUTE reported the purity for the major tumor subpopulation. The second is glioblastoma sample TCGA-06-0188 which we infer to contain two tumor subpopulations consisting of 43.1% and 20.3% cells. In comparison, ABSOLUTE reports that the sample is highly non-clonal and is unable to estimate purity. Our purity estimate of 0.7 is in the range of $0.6 - 0.8$ reported by TCGA histopathology reports. We perform further analysis of this sample and find supporting evidence for our estimated tumor composition (Figure B.10). These results demonstrate that consideration of multiple tumor populations may be important for determining tumor purity, especially for samples with large subclonal populations.

Consistency Across Sequencing Platforms

To further validate the results of THetA2 on whole-exome data we compared results for the 7 of the 18 TCGA samples for which both whole-genome (including low-pass with 5-7X coverage) and whole-exome sequence data was available. For whole-genome samples, we partition the reference

Sample	Path.	ABS	WGS Purity (# populations)	WXS Purity (# populations)	Overlap	CNA Sim
TCGA-06-0185	0.95	0.89	0.87 (3)	0.83 (2*)	0.97	0.91
TCGA-06-0188	**0.6-0.8**	**NA**	**0.70 (3)**	**0.63 (3)**	0.96	0.79, 0.62
TCGA-06-0214[1]	0.25-0.8	0.66	0.67 (3)	0.67 (3)	0.96	0.97, 0.92
TCGA-56-1622	0.9	-	0.68 (3)	0.78 (3)	0.96	0.89, 0.57
TCGA-A2-A0EU	0.9	-	0.77 (3)	0.90 (3)	0.91	0.61, 0.22
TCGA-AO-A0JJ	0.8	-	0.52 (3)	0.52 (2)	0.85	0.67
TCGA-BH-A0W5	0.7	-	0.51 (2*)	0.54 (2*)	0.98	0.97

Table 3.1: **Comparison of THetA2 results on whole-genome and whole-exome data.** Path. are purity estimates reported in TCGA histopathology reports. ABS are ABSOLUTE purity estimates reported by [27] (samples marked with '-' do not have published purity estimates from ABSOLUTE). WGS Purity, WXS Purity and # populations are values predicted by THetA2. Overlap is $\frac{\mathbf{I}^*}{|\mathbf{I}_{WGS}|\cup|\mathbf{I}_{WXS}|}$ where $\mathbf{I}_{WGS}$ and$\mathbf{I}_{WXS}$ are the interval partitions for the whole-genome and whole-exome data, respectively, and $\mathbf{I}^*$ is the set of intervals longer than 100kb contained in both $\mathbf{I}_{WGS}$ and$\mathbf{I}_{WXS}$. CNA Sim is the fraction of $\mathbf{I}^*$ where the copy number estimates are the same between the two data types. * indicates that the sample did not pass the criteria to be considered for multiple tumor populations (see Appendix B). [1]For sample TCGA-06-0214, WGS data was aligned to hg18 and WXS data aligned to hg19. See Table B.3 for purity estimates across all genomes analyzed and results using an additional similarity metric.

genome using the BIC-seq algorithm [174] run with default parameters (see Figure B.7 for whole-genome workflow). We found that purity estimates for whole-exome data to be within 0.04 of purity estimates for whole-genome data for 4 of the 7 samples (Table 3.1).

We also compare the copy number aberrations predicted for the different subpopulations between the whole-exome and whole-genome data using a similarity measure described in Table 3.1 caption. We find that 4 of the genomes have ≥ 0.89 similarity under our measure (Table 3.1) for the major subpopulation. Notably, we find that THetA2 infers 3 subpopulations for sample TCGA-06-0214 on both whole-exome and whole-genome data – selecting the $n = 3$ solution over both $n = 2$ and $n = 4$ for whole-genome data (see Appendix B) and has similarity 0.92 between the datatypes for the minor subpopulation. We also found similar copy number similarity results using a less stringent measure that only considers copy number state rather than exact copy number value (Table B.3). These results demonstrate the consistency of THetA2 – including the inference of multiple tumor subpopulations – across different types of sequencing data.

3.3.3 Analysis of Highly Rearranged and Heterogeneous Genomes

One of the main advantages of THetA2 is the ability to analyze highly rearranged genomes containing many copy number aberrations in one or more tumor subpopulations. We analyze in further detail several highly rearranged genomes that THetA2 predicted to contain subclonal populations from whole-genome data.

Low-Pass Breast Cancer Samples TCGA-A2-A0EU and TCGA-AO-A0JL

We used THetA2 to analyze two breast cancer genomes, TCGA-A2-A0EU and TCGA-AO-A0JL, that were sequenced with low-pass (5-7X) whole-genome sequencing. These are the most rearranged of the breast cancer genomes that we analyzed – containing many intervals in BIC-seq segmentation (493 and 675 intervals respectively) and more predicted copy number aberrations. We attempted to run ABSOLUTE [27] on these genomes using the BIC-seq segmentation. However, despite trying a range of values for the parameters, we obtained purity below 0.3 for both samples. For comparison, we cite the results reported by [177] on these samples, using ABSOLUTE and a different segmentation.

In both samples, THetA2 identifies multiple subclonal populations. We infer that breast cancer sample TCGA-A2-A0EU contains normal admixture with two distinct tumor subpopulations, one with 42.7% cells and another with 34.6% cells (Figure B.11(a)). We note that our estimate of tumor purity (0.77) is below the reported histopathology purity of 0.90 for this sample, but closer than the ABSOLUTE estimate of 0.49. We infer that breast cancer sample TCGA-AO-A0JL contains normal admixture with two distinct tumor subpopulations, one with 57.0% cells and another with 30.5% cells (Figure B.11(b)). Despite being the most rearranged of the breast cancer genomes analyzed, our estimated tumor purity of 0.88 is near the reported histopathology value of 0.80. In comparison, ABSOLUTE inferred purity of 0.50 for this sample. We are also able to identify a number of clonal and subclonal chromosome arm level events for both genomes (see Appendix B), as well as many other small events, thus demonstrating that THetA2 can analyze highly rearranged genomes with low-coverage whole-genome sequencing data.

Lung Squamous Cell Sample TCGA-56-1622

We ran THetA2 on a highly rearranged lung squamous sample TCGA-56-1622, containing 2847 intervals in the segmentation. We note that this genome is so fragmented that ABSOLUTE [27] does not attempt to estimate tumor purity when run with default parameters. Moreover, this sample has so many copy number changes that SNV based algorithms [8, 143] would have extreme difficulty in defining regions of normal copy number to analyze. THetA2 infers that sample contains normal admixture with two distinct tumor subpopulations, one with 50.1% cells and another with 18.1% cells (Figure 3.5a). Using the new two-step procedure, THetA2 also identifies many smaller copy number aberrations (Figure B.12) and we find that the read depth predicted using our reconstruction closely matches the observed read depth (Figure 3.5b).

We examine this sample in further detail using using B-allele frequency (BAF) information not used by THetA2. We constructed a virtual SNP array defining the BAF at a known germline SNP to be the fraction of reads containing the minor allele as described in [118]. In diploid regions of the genome that have not undergone any copy number changes we expect that the BAFs for germline heterozygous SNPs to be near a value of 0.5, as approximately half of the reads should contain the B-allele. In a pure tumor sample a deletion of a segment on a single chromosome will lead to a loss of heterozygosity (LOH) and B-allele frequencies (BAF) at 0 or 1 in a symmetric *double banded pattern* centered around 0.5. As the sample become less pure (i.e. more admixture by normal cells), the double banded pattern will shift closer to 0.5.

In many of the regions where THetA2 predicted a clonal deletion (i.e. in all subpopulations), such as chromosomes 3, 5q and 18 (Figure 3.5b) we observe that the BAFs cluster near 0 and 1, as expected for a deletion occurring in a majority of cells in the sample. Similarly, we find that the shifts in BAF are consistent with THetA2's predictions of subclonal deletions in 50.0% and 18.1% of cells (Figure 3.5b). On chromosome 1p, we observe a discrepancy between THetA2's predictions and BAF. THetA2 predicts that 1p is a clonal deletion; however, the BAFs are clustered tightly around 0.5, indicating an equal number of both parental copies of this region in the tumor sample. One explanation is that 1p is homozygously deleted in one of the tumor subpopulations, rather than a heterozygous deletion in both subpopulations, which would keep the balance of the parental copies of 1p in the tumor sample.

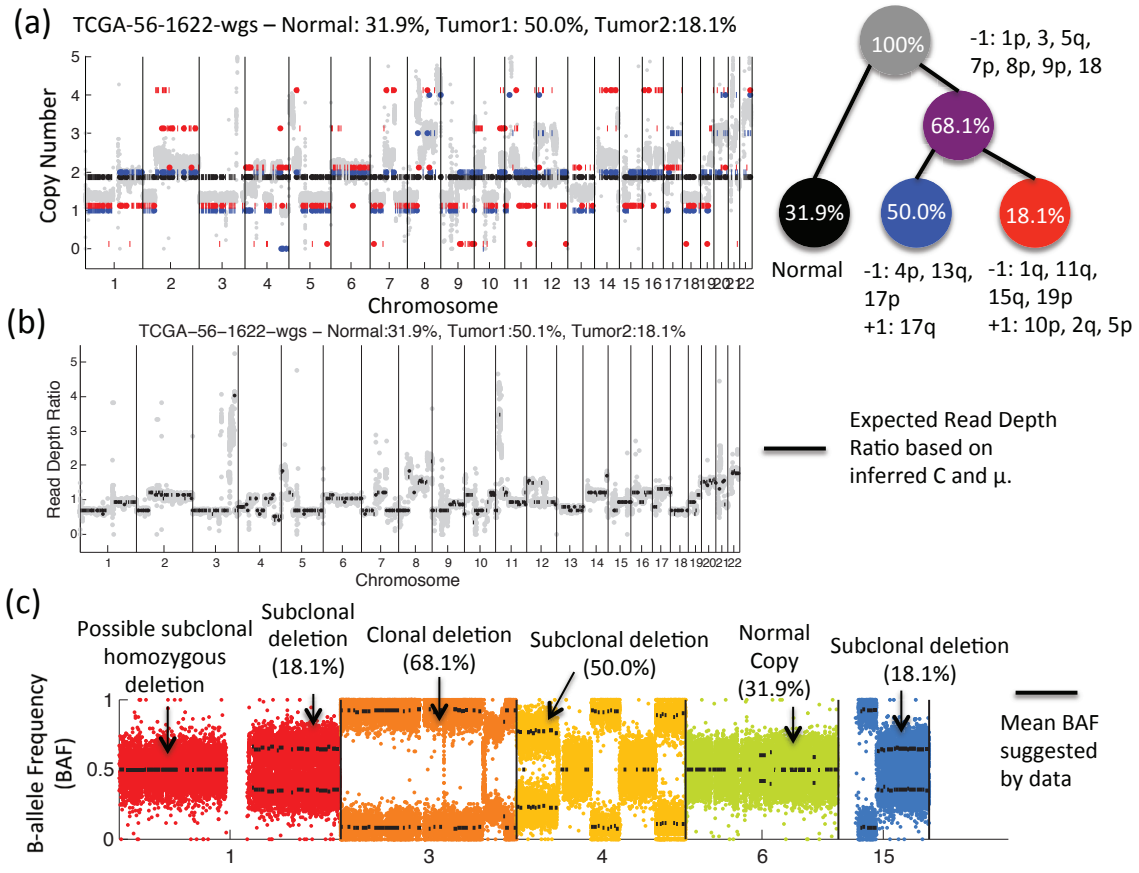


Figure 3.5: **Analysis of squamous cell lung cancer sample TCGA-56-1622. (a)** (Left) Read depth ratios (gray) over 50 kb bins and the inferred copy number aberrations calculated by THetA2 when the tumor is considered to be a mixture of 3 subpopulations: normal cells (black), and two tumor subpopulations (blue and red). (Right) A reconstruction of the tumor mixture along with ancestral clonal population (purple) with the inferred aberrations and estimated fraction of cells in each population (see Appendix B). **(b)** Expected read depth ratios (see Appendix B) for intervals longer than 2Mb based on inferred $\mathbf{C}$ and μ (black) overlaid on observed read depth ratios (gray). **(c)** Virtual SNP array showing B-allele frequencies at germline SNPs on indicated chromosomes and the mean BAF in each segment (see Appendix B).

3.3.4 Using B-allele Frequencies

For glioblastoma sample TCGA-06-0145, THetA2 outputs two possible $(\mathbf{C}, \mu)$ pairs using only read depth – one largely haploid and one largely diploid. We apply our probabilistic model of BAFs and find that the diploid reconstruction, which includes rearrangements characteristic to glioblastoma such as amplification of chr7 and deletion of chr10 [156], is determined to be the more likely tumor composition (Figure B.13).

3.4 Discussion

We introduced an algorithm to infer tumor composition – of highly rearranged genomes from whole-genome (high or low coverage) or whole-exome DNA sequencing data. These are implemented as improvements to our Tumor Heterogeneity Analysis (THetA) algorithm which we call THetA2. The THetA2 algorithm is able to analyze highly rearranged, aneuploid samples that are beyond the scope of existing algorithms that infer tumor heterogeneity. A recently published comparison of algorithms for inferring tumor purity [177] showed that our original THetA algorithm [118] performed well, but sometimes underestimated tumor purity when run to only consider normal cells and one tumor subpopulation. We argue that this purity underestimation is likely a result of not directly considering all tumor subpopulations in the sample. In every sample that we analyzed with the new algorithm, tumor purity was higher when considering multiple tumor subpopulations.

While the improved THetA2 presented here is useful on a wide range of sequencing data from different tumors, some limitations remain. First, THetA2 is unable to distinguish tumor subpopulations that are not differentiated by copy number aberrations. Since copy number aberrations are ubiquitous in most solid tumors [3], we expect that THetA2 will still be applicable to many genomes. However, for some diploid tumors, SNV analysis is preferable. Incorporation of additional information, such as B-allele frequencies for somatic and germline SNPs, directly into the model [8, 143] may also increase the scope of samples for which THetA2 is applicable.

Secondly, while the improvements presented here greatly decrease the computational burden of the algorithm when considering multiple tumor populations, the algorithm remains exponential in the size of the interval partition of the reference genome - making it impractical to infer tumor composition with more than a handful of subpopulations in many cases. Identification of further mathematical restrictions to the domain of interval count matrices, or use of sampling techniques

in place of complete enumeration are future avenues of investigation which may prove useful in this respect. Additionally, when considering multiple tumor of subpopulations, the quality of the results is limited by features of the data including the presence of copy number aberrations that distinguish subpopulations as well as the number of sequence reads available to identify these aberrations, with the later a function of sequencing coverage, aberration length, and proportion of cells that have the aberration.

While the limited number of tumor subpopulations that THetA2 analyzes may not be sufficient to fully analyze tumor progression, THetA2's ability to recover subpopulations with relatively low-coverage sequencing data can provide some insight into tumor subpopulations in cases where methods that rely on high-coverage data [75, 143] cannot. Combining THetA2's output with other methods that do not explicitly consider the phylogenetic history of a tumor such as [75] or [64] may prove a useful avenue of exploration.

The two-step procedure introduced here allows us to infer subclonal copy number aberrations at much smaller scales. However, some care is required to avoid overfitting the data, particularly for small, subclonal copy number aberrations where GC-bias or other sequencing artifacts may lead to incorrect inferences. Incorporating more sophisticated segmentation procedures that account for such effects and appropriately scale read counts [15] are useful directions for future research.

Finally, this work focuses on the important first step of quantifying intra-tumor heterogeneity from a single mixed tumor sample. Downstream analysis including the clinical and functional impact of the inferred tumor composition is an important area for future work.

In conclusion, we present a new algorithm, THetA2, to infer the composition of a tumor sample – including both the percentage of normal admixture and the fraction and content of one *or more* of tumor subpopulations that differ by copy number aberrations. The new algorithm builds upon our Tumor Heterogeneity Analysis (THetA) algorithm [118], and includes several improvements that allow us to analyze highly rearranged genomes from whole-genome (high and low coverage) or whole-exome sequencing data. In addition, the new algorithm is orders of magnitude faster and allows us to use B-allele frequencies to distinguish between different reconstructions.

Chapter 4

Inferring Tumor Evolution from Multi-Sample Data

In this chapter we continue our exploration of methods to infer the composition of heterogeneous tumors. However, in contrast to the work presented in Chapters 2 and 3 we make the following two alterations to the type of input data we consider: (1) We now restrict our attention to data in the form of reads whose alignment indicate the presence of a single nucleotide variant, rather than shifts in read depth due to copy number aberrations; and (2) We consider input data in the form of multiple sequenced samples from the same tumor. Specifically, we formalize the problem of reconstructing the clonal evolution of a tumor as the Variant Allele Frequency Factorization Problem (VAFFP). The input to this problem are the *variant allele frequencies* (VAFs) for individual somatic mutations, i.e. the fraction of tumor cells that contain each mutation, in one or more samples. The problem is to determine the composition of each sample, including the number and proportion of clones in each, and a tree that describes the ancestral relationships between all clones. We prove necessary and sufficient conditions for the VAFFP to have a solution, thus providing a combinatorial characterization of the space of all solutions.

Based on this characterization of the solutions in the case of error-free data, we describe an algorithm called AncesTree to analyze noisy data. AncesTree is fundamentally different from existing approaches because it uses ancestral constraints derived from our combinatorial characterization of solutions to the VAFFP to group mutations rather than directly clustering VAFs. We model errors

in the data using a probabilistic model and derive an integer linear programming solution to the VAFFP. Our AncesTree algorithm is better able to identify ancestral relationships between individual mutations than existing approaches as we demonstrate on both simulated and real sequencing data.

The work from this chapter is joint work with Mohammed El-Kebir and is taken from [47]. The contributions in this chapter were accepted for presentation at the 23^{rd} International Conference on Intelligent Systems for Molecular Biology (ISMB).

4.1 Related Work

Cancer is a disease resulting from *somatic mutations* that accumulate during an individual's lifetime and lead to uncontrolled growth of a collection of cells into a tumor. The clonal theory of cancer [116] predicts that all cells within a tumor have descended from a single founder cell containing a mutation that gave it a selective growth advantage producing a *clonal expansion*. Subsequent advantageous mutations lead to additional clonal expansions.

As a result, the cells within a tumor differ in their complement of somatic mutations, each cell being a descendant of a *clone* from a clonal expansion (Figure 4.1A). High-coverage sequencing of tumor genomes allows one to study this *intra-tumor heterogeneity* by measuring the frequencies of mutations within a tumor [114, 41, 148]. Characterization of intra-tumor heterogeneity and inference of the clonal evolutionary history of somatic mutations within a tumor provide useful insight in the tumor's development and may help inform treatment.

Somatic mutations are typically measured in human solid tumors only at a single time point, when the patient undergoes surgery. Therefore, clonal evolution is not directly observed and one is faced with the problem of inferring the ancestral relationships between cells in a tumor from measurements at one time point. This is the problem of phylogenetic tree reconstruction, a well-studied problem. The direct application of phylogenetic methods requires that we measure mutations in individual cancer cells that correspond to the leaves (species) of the phylogenetic tree. However, due to technical limitations and financial considerations, single-cell sequencing of tumors remains uncommon [111, 169] with nearly all cancer sequencing studies – including large-scale studies such as TCGA and ICGC – sequencing a small number of samples from a bulk tumor, each containing potentially millions of cells. Thus, the data one obtains represents the mutations in a mixture of cells with potentially distinct evolutionary histories.

Given sequencing data from a single sample, methods have been developed to determine the composition – including the fraction of normal cells and one or more populations of cancer cells – of a tumor either by analyzing differences in read depth due to copy number aberrations [118, 121] or by analyzing changes in the variant allele frequencies (VAFs) of single-nucleotide mutations [143, 105]. Inferring the clonal history of a single sample requires additional assumptions about the evolutionary process, such as parsimony [64, 155].

Recently, several studies have conducted multi-sample sequencing from the same tumor [53, 52, 112, 185]. These studies measure somatic mutations in multiple spatially distinct regions from the same tumor at a single time or measure a tumor at multiple time points [147]. Using multiple samples from the same tumor, one can directly apply phylogenetic techniques; e.g. by computing a distance between samples according to the number of shared mutations [52, 185]. However, each sequenced sample is itself a heterogeneous mixture of cells (Figure 4.1B), and thus existing phylogenetic tree reconstruction techniques cannot be applied.

Several methods have recently been introduced to infer tumor composition and evolution from VAFs of somatic mutations in multi-sample sequencing data. Clomial [184] infers the set of clones present in the tumor and their frequencies in each sample, but does not describe the evolutionary relationships between the clones. Three recent approaches infer a tree describing the evolutionary history of a tumor. PhyloSub uses a Bayesian approach to sample trees using a tree-structured process [75]. CITUP [98] enumerates all rooted trees and for each one solves a quadratic program. LICHeE, which recently appeared on the arXiv preprint server [131], uses a graph construction similar to one we describe below, but does not provide a rigorous mathematical justification for it. All of these approaches are data-driven and focus on the construction and optimization of models that minimize the error between the observed and inferred mutation frequencies. However, they do not address the combinatorial structure of the problem. Stated more directly: given error-free VAF data, under what conditions is it possible to reconstruct the clonal evolution of a tumor?

4.2 The AncesTree Algorithm

In this section we first derive a computational formulation for the problem of inferring the clonal evolution of a tumor from multi-sample sequence data. We then show how this problem can be solved in the instance of error-free data and noisy data, leading to our AncesTree algorithm described at

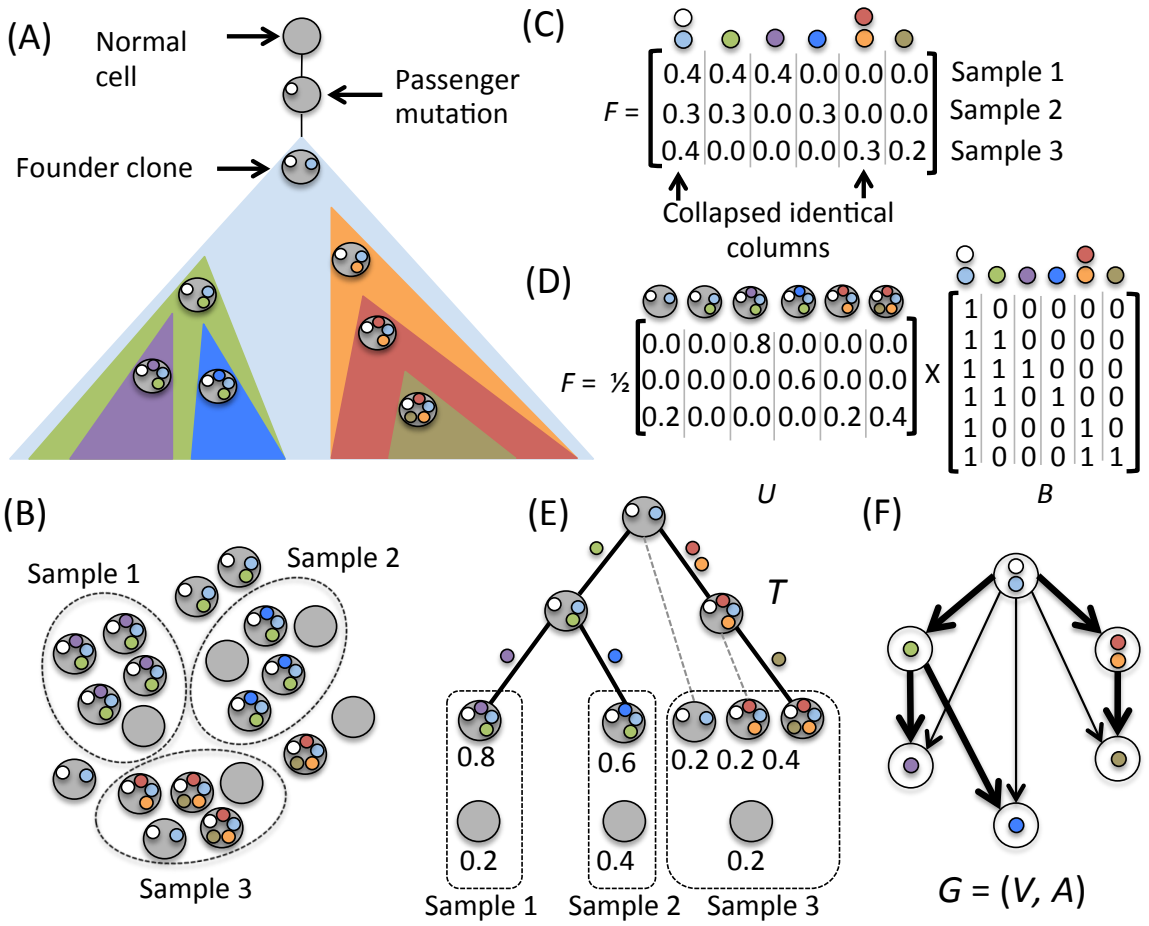


Figure 4.1: **Model for clonal evolution and inference.** (A) An example of the evolution of a tumor containing seven distinct clones. Passenger mutations (white) occurring before the first clonal expansion will be indistinguishable from mutations driving the growth of the founding clone (light blue). Each subsequent mutation (green, purple, dark blue, orange, red, tan) creates a new clone. (B) Three sequenced tumor samples. Some clones may no longer exist at the time of sequencing (orange). Samples 1 and 2 each contain a single clone (purple and blue respectively), while Sample 3 is a mixture of 3 clones (light blue, red and tan). (C) The frequency matrix F observed for the three sequenced samples indicated in part B. (D) The usage matrix U and clonal matrix B that generate F. Even though some clones existing at the current time may not be contained within an sequenced sample (green), their existence in the evolutionary history of the tumor may be recovered. (E) Tree of the inferred tumor clones. Solid black edges are the clonal tree T corresponding to the clonal matrix B. Gray dashed edges indicate internal vertices used in the mixing of some sample. The number next to each clone in each sample indicates the fraction of cells in the sample from that clone. (F) The ancestry graph for the observed data. The bold arcs indicate one spanning arborescence.

the conclusion of this section.

4.2.1 The Variant Allele Frequency Factorization Problem (VAFFP)

We describe a model for the accumulation of single-nucleotide somatic mutations in a tumor, and the generation of sequencing data from the tumor. This leads us to the definition of the Variant Allele Frequency Factorization Problem (VAFFP) at the conclusion of this subsection.

Following the clonal theory of cancer, we assume that all cancer cells in a tumor are descendants of a single founding clone; i.e. the tumor is *monoclonal*. In this work, we model only somatic single-nucleotide mutations, and assume that these are unaffected by copy number aberrations or rearrangements in the cancer genome. We will use *mutation* to refer specifically to these events. We assume, as in previous work [75, 98], that mutations satisfy the *infinite sites assumption*, which states that a mutation occurs at a single genomic position, or locus, at most once during the clonal evolution of the tumor. We encode the state of a specific locus in a clone as a binary value – where 1 indicates a somatic mutation at that position, and 0 indicates no mutation. Thus, each clone corresponds to a binary vector in $\{0,1\}^n$, where n is the total number of loci affected by mutations.

Under the assumptions above the ancestral relationships between clones are described by a phylogenetic tree where: (1) vertices represent different tumor clones that have existed the tumor's evolution; (2) edges represent the direct ancestral relationships between clones and are labeled with the mutation that distinguishes the child from its parent (Figure 4.1A). We assume that each clone is distinguished from its parent by a single mutation (in practice we will group individual mutations into sets that satisfy this assumption). Thus, we describe the mutational process that produced a tumor by an *n-clonal tree* T, which we formally define as follows.

Definition 4.2.1. *A rooted tree T on n vertices is an* n-clonal tree *for a mutation set $[n] = \{1, \ldots, n\}$ provided each edge is labeled with exactly one mutation from $[n]$ and no mutation appears more than once in T. Let $\mathcal{T}_n$ be the set of all n-clonal trees.*

We denote the root vertex of an n-clonal tree by v_r where $r \in [n]$ is the mutation that does not label any edge in T. We denote the remaining vertices by v_j where $j \neq r$ is the mutation on the last edge of the path from v_r to v_j. Note that the set of mutations present in a clone v_j is the set of mutations of all vertices on the path from v_r to v_j. The root vertex v_r contains only mutation r and thus represents the *founding clone*.

Alternatively, we may describe the n-clonal tree T by an $n \times n$ binary matrix B. We label the vertices of T by binary vectors indicating the mutations present in each vertex (clone). Each vertex v_j corresponds to a binary row vector $\mathbf{b}_j^T$ with 1's at the r^{th} position and positions indicated on the edge labels on the unique path from v_r to v_j, and 0's at remaining positions. Let B be the $n \times n$ binary matrix whose j^{th} row is $\mathbf{b}_j$. Since the mutations adhere to the infinite sites assumption, it follows that B is a *perfect phylogeny matrix* [60]. That is, for a column j of B, let $I(j)$ be the positions of the 1 entries. Then for any pair of columns j and k of B either $I(j)$ and $I(k)$ are disjoint, or one contains the other.

Not every $n \times n$ perfect phylogeny matrix corresponds to a n-clonal tree T. For example, a perfect phylogeny matrix may have a row and/or column of all 0's or have duplicated rows and/or columns. We define a subset of $n \times n$ perfect phylogeny matrices, which we call n-clonal matrices that are in 1-1 correspondence with n-clonal trees T.

Definition 4.2.2. *A matrix $B \in \{0,1\}^{n \times n}$ is an* n-clonal matrix *provided:*

1. *There exists exactly one $r \in [n]$ such that $\sum_{j=1}^{n} b_{rj} = 1$.*
2. *For each $j \in [n] \backslash \{r\}$ there exists exactly one $k \in [n]$ such that $\mathbf{b}_k \subseteq \mathbf{b}_j$ and $\sum_{l=1}^{n} (b_{jl} - b_{kl}) = 1$.*
3. *$b_{jj} = 1$ for all $j \in [n]$.*

Let $\mathcal{B}_n$ be the set of all n-clonal matrices.

The second condition above ensures that every n-clonal matrix is a perfect phylogeny matrix. We have the following lemmas which we prove in Appendix C.

Lemma 4.2.1. *There is a one-to-one correspondence between $\mathcal{T}_n$ and $\mathcal{B}_n$.*

Lemma 4.2.2. *Any $B \in \mathcal{B}_n$ has rank n.*

Figure 4.1D and 4.1E show a clonal matrix together with its corresponding clonal tree.

Measurement of Clonal Trees

We do not directly observe the clonal tree T relating the clones in a tumor. Moreover, unless we perform single-cell sequencing, we do not directly measure the presence/absence of mutations in individual clones. Rather each sequenced sample is a mixture of cancer cells (clones) and normal cells. We obtain *variant allele frequencies* (VAFs), or the fraction of reads covering a position that

indicate the variant/mutation, at each of the n mutation sites for each of the m samples. The VAF for a mutation is proportional to the *cellular prevalence*, or fraction of cells in the sample that contain the mutation. Suppose we sequence m samples from a tumor. Our observations are described by an $m \times n$ *frequency matrix* $F = [f_{pi}]$ where f_{pi} indicates the observed VAF in sample p for the i^{th} mutation (Figure 4.1C).

The observed mutation frequencies (entries of F) are related to the tree T by the proportions of normal cells and clones that define the mixture in each sample. We define an $m \times n$ usage matrix $U = [u_{pi}]$, where u_{pi} indicates the fraction of cells in sample p that come from clone v_i, as follows.

Definition 4.2.3. *An $m \times n$ matrix $U = [u_{pj}]$ is a usage matrix provided $u_{pj} \geq 0$ and $\sum_{j=1}^{n} u_{pj} \leq 1$. Let $\mathcal{U}_{m,n}$ be the set of all $m \times n$ usage matrices U.*

Since, each sequenced sample is a mixture of clones from T with proportions defined in the usage matrix U, the observed frequency matrix $F = [f_{pj}]$ satisfies

$$F = \frac{1}{2}UB. \tag{4.1}$$

The coefficient $\frac{1}{2}$ arises because, by the infinite sites assumption, all mutations are heterozygous (affect only one homolog), and thus each $f_{pj} \in [0, 0.5]$. If we are given an error-free F, our goal is to find U and B satisfying (4.1). We define this problem as follows (see Figure 4.1D).

Variant Allele Frequency Factorization Problem. *Given an $m \times n$ frequency matrix F, find a usage matrix $U \in \mathcal{U}_{mn}$ and a clonal matrix $B \in \mathcal{B}_n$ such that $F = \frac{1}{2}UB$.*

Without loss of generality, we assume that the rows and columns of any frequency matrix F are distinct, as duplicated rows or columns can be collapsed.

The Variant Allele Frequency Factorization Problem (VAFFP) represents an important generalization that can be used to describe a wide array of important problems, each with unique optimization or constraint criteria. For example, a special case of the VAFFP with $m = 1$, with the goal of minimizing the number of non-zero entries in U, was previously considered by [64] and by [155] who break ties in favor of solutions whose corresponding clonal trees have minimum depth. Additionally, the Perfect Phylogeny Mixture Problem in [65] can also be described as a variant of the VAFFP where F is binary (a mutation is either observed or not) and additional constraints are placed on the usage matrix U. Throughout the remainder of this chapter, we will make note of when

and how related work may be interpreted in terms of the general VAFFP framework.

4.2.2 Solving the VAFFP

In this section, we derive a characterization of the solutions of the VAFFP (Theorem 4.2.7 below) as constrained spanning arborescences of a directed acyclic graph (DAG) called the ancestry graph (Definition 4.2.4 below). From this characterization, we can show that the VAFFP is NP-complete (Theorem 4.2.8 below) and give an exact algorithm for solving the problem.

A Necessary Condition and the Ancestry Graph

We say that B (or T) *generates* F if and only if there exists a matrix $U \in \mathcal{U}_{mn}$ such that $F = \frac{1}{2}UB$. To obtain a characterization of all solutions of the VAFFP, we first define several properties that relate the observed values of F to any clonal tree T that generates F.

We start by observing that any T induces a partial ordering on the vertices. That is, for $j, k \in [n]$, $j \prec_T k$ if and only if vertex v_j is an ancestor of vertex v_k. Conversely, we say that j and k are *incomparable* if and only if neither v_j nor v_k is an ancestor to the other. Because B is a perfect phylogeny matrix, there is a partial order on the columns of B [61]. That is, for $j, k \in [n]$, we have $j \prec_B k$ if and only if $I(j) \supseteq I(k)$. Similarly, j and k are incomparable if and only if $I(j) \not\supseteq I(k)$ and $I(k) \not\supseteq I(j)$. The following observation follows directly from Lemma 4.2.1.

Observation 4.2.1. *Given an n-clonal tree T and its corresponding clonal matrix B, $j \prec_T k$ if and only if $j \prec_B k$ for all $j, k \in [n]$.*

Since $\prec_B$ and $\prec_T$ are equivalent, we will use $\prec$ to denote either one.

We prove the following proposition.

Ancestry Condition. *If T generates F and $j \prec_T k$ then $f_{pj} \geq f_{pk}$ for all samples $p \in [m]$.*

Proof. Since $j \prec_T k$, by Observation 4.2.1 we have $j \prec_B k$. Therefore $I(j) \supseteq I(k)$. Moreover, since every entry in U is non-negative, we have the following for all samples $p \in [m]$:

$$\begin{aligned} f_{pj} &= \frac{1}{2}\sum_{i=1}^{n} u_{pi} \cdot b_{ij} = \frac{1}{2}\sum_{i \in I(j)} u_{pi} \\ &\geq \frac{1}{2}\sum_{i \in I(k)} u_{pi} = \frac{1}{2}\sum_{i=1}^{n} u_{pi} \cdot b_{ik} = f_{pk}. \end{aligned}$$

□

Applying the contrapositive of the above lemma on two distinct samples yields the following corollary which is equivalent to the "crossing rule" stated in [75].

Corollary 4.2.3. *If T generates F and there exist samples $p, q \in [m]$ and mutations $j, k \in [n]$ such that $f_{pj} > f_{pk}$ and $f_{qj} < f_{qk}$ then j and k are incomparable.*

Definition 4.2.4. *Given an $m \times n$ frequency matrix F, we define the* ancestry graph $G = (V, A)$ *where $V = \{v_1, \ldots, v_n\}$ and $A = \{(v_j, v_k) \mid f_{pj} \geq f_{pk}$, for all $p \in [m]\}$.*

Intuitively, G is the graph whose vertices represent mutations and whose arcs represent possible ancestral relationships consistent with observed variant allele frequencies using the ancestry condition (Figure 4.1F). We note the following observation which will be useful in the following section.

Observation 4.2.2. *If all columns of a frequency matrix F are distinct then its ancestry graph G is a DAG.*

A *spanning arborescence* of the ancestry graph G is a subgraph $G' = (V, A')$ with $A' \subseteq A$ such that there exists a unique path from the root vertex v_r to every vertex $v \in V$.

Lemma 4.2.4. *If T generates F then it is a spanning arborescence of G.*

Proof. Let T be a tree that generates F. We proceed using contradiction. Suppose that T is not a spanning arborescence of G. Thus, there exists an edge (v_j, v_k) in T with $j \prec k$ such that $(v_j, v_k) \notin A$. By definition of A there must exist $p, q \in [m]$ such that $f_{pj} < f_{pk}$ and $f_{qj} > f_{qk}$. By Corollary 4.2.3, j and k are incomparable – a contradiction. Hence, T must be a spanning arborescence of G. □

If an ancestry graph $G = (V, A)$ does not have a spanning arborescence then there exists no tree T that generates F. Checking whether G has a spanning arborescence can be done in $\mathcal{O}(|A|)$ time since by definition A contains all transitive arcs. Figure 4.2A shows an example of a frequency matrix whose ancestry graph has no spanning arborescence. Furthermore, not all spanning arborescences T of G generate F. Figure 4.2B shows such an example, where the matrix U obtained from T and F has negative entries and thus is not a usage matrix. Hence, the existence of a spanning arborescence in G is a necessary but *not* a sufficient condition for a solution to the VAFFP.

The Sum Condition and Sufficiency

In the previous section, we saw that the ancestry condition is not sufficient to produce a solution to the VAFFP. Sufficiency will be obtained through a second condition, which we refer to as the *sum condition*. This condition was stated as the "sum rule" in [75], and also appears in [98], and a special case was called the "children sum to parents" condition in [64]. Given a clonal tree T, $\delta(v_j)$ denotes the children of a vertex v_j in T.

Sum Condition. *If T generates F then for all samples $p \in [m]$ and mutations $j \in [n]$,*

$$f_{pj} \geq \sum_{v_k \in \delta(v_j)} f_{pk}. \tag{4.2}$$

Proof. All $v_k \in \delta(v_j)$ are pairwise incomparable, i.e., there are no $v_k, v_l \in \delta(v_j)$ such that $v_k \prec_T v_l$. Therefore all $I(k)$ with $v_k \in \delta(v_j)$ are pairwise disjoint. Moreover, $I(j) = \bigcup_{k \in \delta(v_j)} I(k) \cup \{j\}$. Since every entry of U is non-negative we thus have

$$\begin{aligned} f_{pj} = \frac{1}{2}\sum_{k=1}^{n} u_{pk} \cdot b_{kj} &= \frac{1}{2}\sum_{k \in I(j)} u_{pk} \\ &= \frac{1}{2}u_{pj} + \frac{1}{2}\sum_{v_k \in \delta(v_j)} \sum_{l \in I(k)} u_{pl} \\ &= \frac{1}{2}u_{pj} + \sum_{v_k \in \delta(v_j)} \frac{1}{2}\sum_{l=1}^{n} u_{pl} \cdot b_{lk} \\ &= \frac{1}{2}u_{pj} + \sum_{v_k \in \delta(v_j)} f_{pk} \\ &\geq \sum_{v_k \in \delta(v_j)} f_{pk} \end{aligned}$$

□

For a clonal tree T, sample p and mutation j, we define the *deficit* $d_{pj} = f_{pj} - \sum_{v_k \in \delta(v_j)} f_{pk}$. Thus, the Sum condition above says that if T generates F, then the deficit d_{pj} is non-negative for all samples p and mutations j. It turns out that the deficits for all samples and mutations determine the matrix U. In particular, we have the following Lemma.

Lemma 4.2.5. *Given an $m \times n$ frequency matrix F and an n-clonal matrix B, the $m \times n$ matrix*

$U = [u_{pj}]$ *defined as*

$$u_{pj} = 2d_{pj} = 2\left(f_{pj} - \sum_{v_k \in \delta(v_j)} f_{pk}\right) \tag{4.3}$$

is the unique matrix such that $F = \frac{1}{2}UB$.

Proof. Lemma 4.2.2 tells us that there is thus a unique $U \in \mathbb{R}^{m \times n}$ such that $F = \frac{1}{2}UB$. It suffices to show that $f_{pj} = \frac{1}{2}\sum_{k=1}^{n} u_{pk} \cdot b_{kj}$ for any sample p and mutation j. Let T be the corresponding clonal tree of B and $j \in [n]$. Since $b_{kj} = 1$ for all $k \in I(j)$, we have

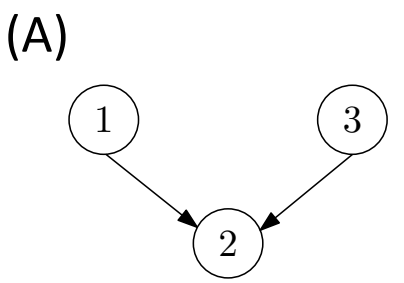
$$\begin{aligned}
\frac{1}{2}\sum_{k=1}^{n} u_{pk} \cdot b_{kj} &= \frac{1}{2}\sum_{k \in I(j)} u_{pk} \\
&= \frac{1}{2}\sum_{k \in I(j)} 2\left(f_{pk} - \sum_{v_l \in \delta(v_k)} f_{pl}\right) \\
&= \sum_{k \in I(j)} \left(f_{pk} - \sum_{v_l \in \delta(v_k)} f_{pl}\right) \\
&= \sum_{k \in I(j)} f_{pk} - \sum_{k \in I(j)} \sum_{v_l \in \delta(v_k)} f_{pl}
\end{aligned}$$

We note that $I(j)$ returns the set of indices of all vertices in the subtree rooted at v_j in T including v_j itself. Since T is a tree, for any $k, l \in I(j)$ with $k \neq l$ we have $\delta(v_k) \cap \delta(v_l) = \emptyset$. Thus, in the last line of the above derivation we subtract f_{pl} exactly once for all vertices $v_l \neq v_j$ that are in the subtree rooted at v_j. The set of such vertices is $I(j) \setminus \{j\}$. Hence,

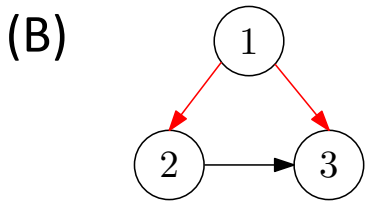
$$\sum_{k \in I(j)} f_{pk} - \sum_{k \in I(j)} \sum_{v_l \in \delta(v_k)} f_{pl} = \sum_{k \in I(j)} f_{pk} - \sum_{k \in I(j) \setminus \{j\}} f_{pk} = f_{pj}.$$

□

Note that by Lemma 4.2.2, any $B \in \mathcal{B}_n$ is full-rank, and therefore invertible. Thus, for any frequency matrix F there is a unique $U \in \mathbb{R}^{m \times n}$ such that $F = \frac{1}{2}UB$, namely $U = 2FB^{-1}$. Thus, the above theorem gives an explicit formula for the entries of $U = 2FB^{-1}$. For the single sample case, a similar formula was derived by [155]. However, instead of using this formula to infer the usage vector the authors use back substitution. [75] also describe a recursive formula relating the frequencies and usages. Moreover, note that $d_{pj} = 0$ is equivalent to the "children sum to parents" condition in [64] and the "non-populated clone" condition in [155]. In this case $u_{pj} = 0$, which



$$F = \begin{pmatrix} 0.4 & 0.2 & 0.25 \\ 0.3 & 0.1 & 0.35 \end{pmatrix}$$



$$F = \begin{pmatrix} 0.15 & 0.1 & 0.06 \\ 0.2 & 0.05 & 0.04 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

$$U = 2FB^{-1} = \begin{pmatrix} -0.02 & 0.2 & 0.12 \\ 0.22 & 0.1 & 0.08 \end{pmatrix}$$

Figure 4.2: **Spanning arborescences of the ancestry graph.** (A) F cannot be factorized as its ancestry graph does not admit a spanning arborescence. (B) Red arcs indicate a spanning arborescence T of the ancestry graph of F with corresponding matrix B. B does not generate F as the matrix $U = 2FB^{-1} \notin \mathcal{U}_{mn}$.

implies that the clone v_j is not present (or mixed) in sample p.

The matrix U defined by Equation (4.3) has non-negative entries whose rows sum to at most 1 and thus is a valid usage matrix precisely when the deficits are non-negative. This in turn happens when F satisfies the sum condition (Equation (4.2)). Combining these results, we obtain the following lemma.

Lemma 4.2.6. *If an $m \times n$ frequency matrix $F = [f_{pj}]$ satisfies Equation (4.2) for the tree T corresponding to $B \in \mathcal{B}_n$, then B generates F.*

Proof. We need only to show that U created according to Equation (4.3) is an element of $\mathcal{U}_{mn}$. Thus, we need to show that $u_{pj} \geq 0$ for all p, j and $\sum_{j=1}^{n} u_{pj} \leq 1$ for all p. The condition that $u_{pj} \geq 0$ follows directly from our assumption that $f_{pj} \geq \sum_{v_k \in \delta(v_j)} f_{pk}$ for all p and j, as defined in Equation (4.2). By definition, column r of B consists of only 1-entries. Moreover, every entry $f_{pj} \leq 0.5$ and thus $f_{pr} = \frac{1}{2}\sum_{k=1}^{n} u_{pk} \cdot b_{kr} = \frac{1}{2}\sum_{k=1}^{n} u_{pk} \leq 0.5$. Hence, $\sum_{k=1}^{n} u_{pk} \leq 1$. □

Using this lemma, we obtain the following characterization of those spanning arborescences of the ancestry graph G that generate F.

Theorem 4.2.7. *T generates $F = [f_{pj}]$ if and only if T is a spanning arborescence of G such that Equation (4.2) holds for all f_{pj}.*

Proof. The forward direction follows from Lemma 4.2.4 and the sum condition (Equation (4.2)). For

the reverse direction, we know that T spans all vertices of G and therefore is a valid clonal tree with corresponding clonal matrix B. Lemma 4.2.6 tells us that B, and hence also T, generate F. □

Thus, there is a 1-1 correspondence between spanning arborescences in G that satisfy the sum condition and solutions to the VAFFP. While a spanning arborescence can be found efficiently, deciding whether G admits a spanning arborescence satisfying the sum condition is NP-complete.

Theorem 4.2.8. *VAFFP is NP-complete.*

Proof. By reduction from Not-All-Equal-3SAT. We direct readers to [47] for the proof. □

In summary, the following procedure gives a solution to the VAFFP for a frequency matrix F: (i) Build the ancestry graph G for F. (ii) Find a spanning arborescence T of G that satisfies the sum condition. (iii) Find the corresponding matrix B and compute U according to Equation (4.3).

An Integer Linear Programming Solution

We formulate an integer linear program (ILP) to find the largest arborescence in an ancestry graph G that adheres to the sum constraint. If this is a spanning arborescence, then we have found a solution to the VAFFP. First, we introduce an artificial root vertex v_r that has has an outgoing arc to every other vertex in V. Let $A' = A \cup \{(v_r, w) \mid w \in V\}$ denote this extended arc set. For $v \in V \cup \{v_r\}$, we define $\delta^+(v) = \{w \in V \mid (v, w) \in A'\}$ to be the set of vertices connected to v by an outgoing arc. Similarly, we define $\delta^-(v) = \{w \in V \mid (w, v) \in A'\}$ to be the set of vertices connected to v by an incoming arc. Let variables $\mathbf{x} \in \{0, 1\}^{|A'|}$ be binary variables indicating the

presence/absence of arcs in a solution.

$$\max \sum_{(v_j,v_k)\in A'} x_{jk} \tag{4.4}$$

$$\text{s.t.} \sum_{v_j\in\delta^+(v_r)} x_{rj} = 1 \tag{4.5}$$

$$x_{kl} \leq \sum_{v_j\in\delta^-(v_k)} x_{jk} \qquad \forall (v_k, v_l) \in A \tag{4.6}$$

$$\sum_{v_j\in\delta^-(v_k)} x_{jk} \leq 1 \qquad \forall v_k \in V \tag{4.7}$$

$$\sum_{v_j\in\delta^-(v_k)} f_{pk}x_{jk} \geq \sum_{v_l\in\delta^+(v_k)} f_{pl}x_{kl} \quad \forall p \in [m],\, v_k \in V \tag{4.8}$$

$$x_{jk} \in \{0,1\} \qquad \forall (v_j, v_k) \in A' \tag{4.9}$$

Constraint (4.5) enforces that the arborescence T has only one root vertex. Constraints (4.6) state that for every outgoing arc (v_k, v_l) in T there is an incoming arc (v_j, v_k) in T. The arborescence constraints (4.7) enforce that every vertex v_k has at most one incoming arc (v_j, v_k) in T. Constraint (4.8) is the sum condition (Equation (4.2)). Thus, these constraints encode that any arborescence satisfying the sum condition is a feasible solution and vice versa. The objective function (4.4) maximizes the number of edges in the arborescence. Therefore, the answer to the VAFFP is 'yes' if and only if the optimal solution has objective value n. In that case the corresponding arborescence T would span the vertices of G and satisfy the sum condition. Note that because G is a DAG (Observation 4.2.2), our formulation of the ILP does not have to consider cycles. Lastly, we observe that this ILP only allows us to determine if a solution to the VAFFP exists, but provides no way to discriminate between multiple solutions – something we consider in the following section.

4.2.3 VAFFP with Errors

Thus far we have assumed that the observed frequency matrix F is error-free. That is, there exists some $B \in \mathcal{B}_n$ and $U \in \mathcal{U}_{mn}$ such that $F = \frac{1}{2}UB$. However, this may not be the case for real sequencing data where the entries of F are obtained from integer read counts, and thus are approximations of the true frequencies. We address this uncertainty in the frequencies by relaxing both the ancestry condition and the sum condition.

Approximate Ancestry Graph

We build an approximate ancestry graph using a probabilistic model for the observed read counts. Let X_{pj} be a random variable describing the variant allele frequency (VAF) for a sample p and mutation j. For any pair of mutations j, k and sample p let $Pr[X_{pj} \geq X_{pk}]$ denote the posterior probability that $X_{pj} \geq X_{pk}$. The sample p with the smallest such probability, represents the weakest evidence that mutation j preceded mutation k in the evolutionary history of the tumor. Thus, we denote the posterior probability that $j \prec k$ as $\min_p Pr[X_{pj} \geq X_{pk}]$.

We build the *approximate ancestry graph* $G = (V, A)$ in two steps: (1) We use a graph clustering procedure to group mutations whose posterior probabilities indicate that they likely occurred together; and (2) We restrict to high-confidence ancestral relationships among the clusters of mutations. For these two steps, we use input parameters α and β, respectively, to control the size of V and A in the graph G. Specifically, our process for building the approximate ancestry graph G is as follows.

We expect to cluster mutations whose posterior probability distributions are similar across all samples. If mutations j and k have identical posterior probability distributions in sample p, then $P[X_{pj} \geq X_{pk}] = P[X_{pk} \geq X_{pj}] = 0.5$. Thus, we define the graph $H = ([n], A_H)$ whose vertices are the mutations and whose edges $A_H = \{(j, k) \mid 0.5 - \alpha \leq \min_p Pr[X_{pj} \geq X_{pk}] \leq 0.5 + \alpha\}$. The edges A_H are those ancestry relationships where the posterior probability that $j \prec k$ and $k \prec j$ is within α of 0.5, for some $\alpha \in [0, 1]$. The resulting graph H may have directed cycles. These directed cycles correspond to sets of mutations whose frequencies suggest than none of the mutations is ancestral to the others. We group such mutations into clusters by computing strongly connected components in H. We then determine ancestry between clusters/components by including a directed edge between two components only if there exists mutations k and l in the corresponding clusters such that the posterior probability that $k \prec l$ is greater than β for all samples, for some β. Formally, let $\mathcal{S} = \{S_1, \ldots, S_t\}$ be the set of strongly connected components in H. We define the approximate ancestry graph $G = (V, A)$ whose vertices $V = \mathcal{S}$ and whose edges $A = \{(i, j) \mid \exists k \in S_i, l \in S_j \text{ s.t. } Pr[X_{pk} \geq X_{pl}] \geq \beta, \text{ for all } p \in [m]\}$. We note there is no theoretical guarantee that the resulting graph G is a DAG because cycles may exist containing one or more edges with posterior probability $> 0.5 + \alpha$. However, since increasing β reduces the number of edges in G, we find in practice that setting β sufficiently large generally produces a DAG.

We compute the distribution of X_{pj} for a sample p and mutation j as the posterior probability of the VAF given the observed read counts. The observed VAF $\tilde{f}_{pj} = \frac{\tilde{c}_{pj}}{(\tilde{c}_{pj}+\tilde{d}_{pj})}$, where $\tilde{c}_{pj}$ and $\tilde{d}_{pj}$ are the number of reads from sample p that cover mutation j and that contain the variant and reference alleles, respectively. The distribution of X_{pj} is the posterior distribution of the binomial proportion when one observes $\tilde{c}_{pj}$ "successes" on $\tilde{c}_{pj} + \tilde{d}_{pj}$ trials. Assuming a flat prior on the proportion, we have $X_{pj} \sim \text{Beta}(\tilde{c}_{pj} + 1, \tilde{d}_{pj} + 1)$. In other words, we use a generative model for variant allele frequencies with $\tilde{c}_{pj} \sim \text{Binomial}(\tilde{c}_{pj} + \tilde{d}_{pj}, q)$ and $q \sim \text{Beta}(1, 1)$. For $j, k \in [n]$, we use the method described in [36] to compute $Pr[X_{pj} \geq X_{pk}]$. Finally, as the vertices in the approximate ancestry graph G correspond to strongly connected components that typically include more than one mutation, we compute the frequency matrix $F = [f_{pj}]$ for the approximate ancestry graph G by combining read counts for all mutations in the same component. That is, for a vertex $v_j \in V$ and sample $p \in [m]$ we define $c_{pj} = \sum_{k \in S_j} \tilde{c}_{pk}$ and $d_{pj} = \sum_{k \in S_j} \tilde{d}_{pk}$. We set $f_{pj} = \frac{c_{pj}}{(c_{pj}+d_{pj})}$.

Note that our approach clusters mutations according to the uncertainty in the ancestry constraints, which in turn is defined by the uncertainty in the frequency of individual mutations, where the latter is computed from the overlap between the posterior distributions of the binomial parameters. This is very different from existing approaches such as CITUP [98], PhyloSub [75], and SciClone [105] that cluster mutations according to VAF alone. Moreover, in some methods the uncertainty in the VAF of each mutation is considered to be fixed, rather than a function of the observed read counts. Our approach allows us to distinguish mutations whose observed VAFs may be similar, but which are likely contained within distinct clones, according to their relationships to other mutations in different samples.

An MILP for Arborescences with Errors

Our construction of the approximate ancestry graph relaxes the ancestral relationships in the case of errors in VAFs. However, errors in the observed VAFs may also result in violations of the sum condition. Thus, we formulate a mixed integer linear program that finds the largest arborescence on the approximate ancestry graph while allowing for the inferred frequencies to differ slightly from the observed frequency values. We create a confidence interval $[f_{ij}^{-}, f_{ij}^{+}]$ as the $(1 - \gamma)$ equal-tailed posterior probability interval of the Beta distribution with parameters $(c_{ij} + 1, d_{ij} + 1)$ where γ is a fixed parameter. This interval will provide lower and upper bounds on the inferred frequency values in the MILP formulation.

G may not contain any *spanning* arborescence that satisfies the sum condition since G only consists of high confidence arcs. Therefore, we choose to return a partial solution to the VAFFP by returning the largest arborescence T in G that adheres to the sum condition. This arborescence represents a subset of mutations for which we can confidently infer the ancestral relationships. We note that this is a departure from other methods such as CITUP [98] and PhyloSub [75] that require that all mutations be placed on a single tree. There may be multiple such maximal trees T in G. Rather than considering all such trees, we return the clonal tree T (corresponding to a clonal matrix B) and and associated usage matrix U which minimizes the average deviation between entries in the inferred frequency matrix $F = \frac{1}{2}UB$ and the observed frequency matrix $\tilde{F}$. Since we have clustered mutations into sets, we need to define a map σ which relates individual mutations, to their respective cluster. That is, $\sigma(j) = k$ when mutation j occurs in cluster k.

The MILP is as follows.

$$\max \sum_{(v_j,v_k)\in A} x_{jk} - \frac{1}{mn}\sum_{p=1}^{m}\sum_{j=1}^{n} |\tilde{f}_{pj} - f_{p,\sigma(j)}| \tag{4.10}$$

$$\text{s.t. } (4.5), (4.6), (4.7), (4.8) \text{ and } (4.9) \tag{4.11}$$

$$f_{pj} \in [f_{pj}^{-}, f_{pj}^{+}] \qquad \text{for all } p \in [m],\ v_j \in V \tag{4.12}$$

We model the absolute value in (4.10) and the product $f_{pk}x_{jk}$ in (4.8) using standard linearization techniques [172]. We call the resulting algorithm AncesTree.

4.3 Results

We implemented AncesTree in C++ using CPLEX v12.6. We analyze 90 simulated datasets and 22 real tumor samples. The real data consists of chronic lymphocytic leukemia (CLL) [147], lung adenocarcinoma [185] and renal cell carcinoma tumors [52]. The lung and renal tumors have undergone multi-section sequencing, while the CLL tumors were sequenced over multiple time points. For 14 of the 22 tumors we have both whole-genome/whole-exome sequencing data and targeted deep resequencing data of either the same or a subset of mutations for all sections of the tumor (Table C.1). For all analyses, we set $\alpha = 0.3$, $\beta = 0.8$ and $\gamma = 0.01$. See the Appendix C for results as α and β are varied.

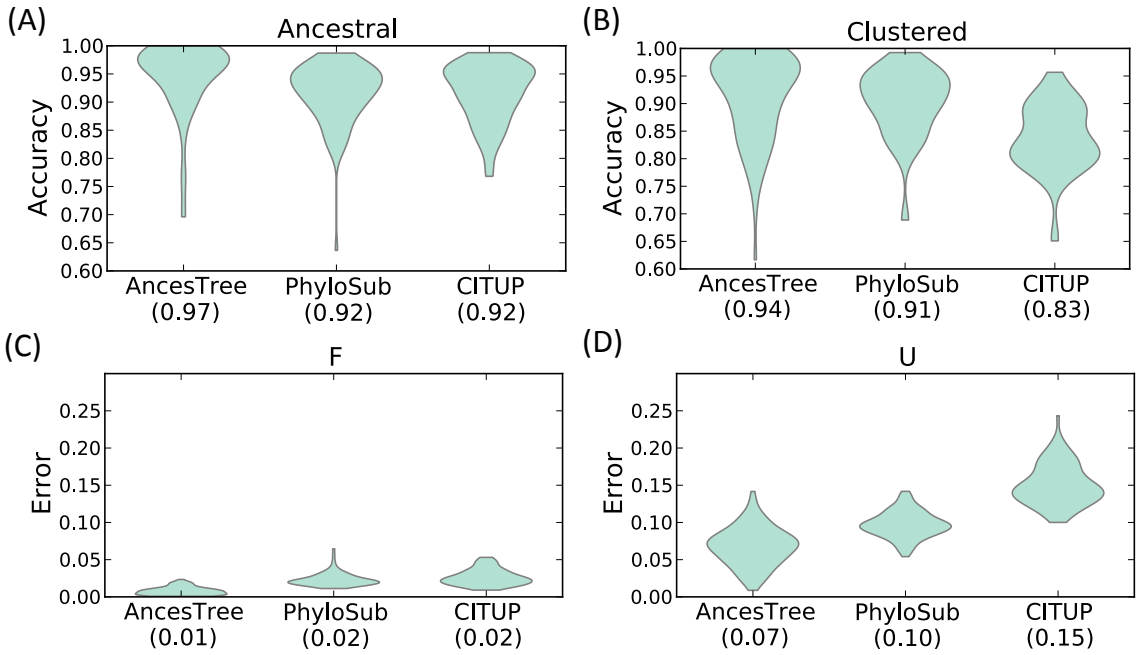


Figure 4.3: **Violin plots comparing AncesTree, PhyloSub and CITUP on simulated data.** (A) Accuracy of each method in predicting when mutations are ancestral to each other or (B) clustered in the same population. (C) Error in the inferred VAF f_{pj} and (D) usage values u_{pj}. Median values are indicated below the names of the algorithms.

4.3.1 Comparison of AncesTree to PhyloSub and CITUP

We compare AncesTree to two other recent algorithms that infer trees from multi-sample sequencing data: PhyloSub [75] and CITUP [98]. We were unable to compare to LICHeE [131] as the software only provides a graphical user interface with no way to easily export results.

We created 90 synthetic tumor datasets. Each dataset contains 100 mutations grouped into 10 clones that accumulated following the infinite sites assumption. For each dataset we simulated between 4-6 samples sequenced at a coverage of 50X, 100X, or 1000X. Further details of the simulated data is contained in Appendix C. We ran AncesTree, PhyloSub and CITUP on each dataset and compared the results using five measures: (1) accuracy of the ancestral relationship (either ancestral or not) between all pairs of mutations (Figure 4.3A); (2) accuracy of the clustering relationship (either clustered in the same clone or not) between all pairs of mutations (Figure 4.3B); (3) accuracy at determining whether all pairs of mutations are *incomparable* (i.e. neither ancestral or clustered) (Figure C.1); (4) the average error $\frac{1}{mn}||\tilde{F} - F||_1$ between the simulated $\tilde{F}$ and inferred frequency matrix F (Figure 4.3C); (5) the error between the simulated usage matrix $\tilde{U}$ and the inferred usage U

using the same metric as [98] (Figure 4.3D). We note that we compute these measures only on the set of mutations that are including in the output of all methods, which equates to the set of mutations output by AncesTree (median of 69 of the 100 total mutations) since CITUP and PhyloSub include all mutations. We find that AncesTree has higher accuracy in determining ancestral, clustered, and incomparable relationships with median accuracy more than 0.05, 0.03 and 0.08, respectively, above the median accuracy of the other methods. Further, we find that AncesTree achieves a median error on F and U that is 0.01 and 0.03 lower than the median error of the other methods. See Appendix C for further details on all five metrics.

We also compare the output of AncesTree, CITUP and PhyloSub on the sequencing data from 22 tumor samples and find that AncesTree produces results that are more consistent with the input data in terms of our probabilistic model (see Appendix C).

4.3.2 Analysis of Whole-exome vs. Deep Sequencing Data

A key difference between AncesTree and other approaches is that we use a graph clustering approach to to group mutations by their putative ancestral relationships across all samples, rather than clustering variant allele frequencies (VAFs) directly. We demonstrate the advantages of this approach on a lung tumor (patient 330 in [185]) that had multiple samples sequenced using both whole-exome and targeted deep sequencing (higher coverage) data. One would expect that deep sequencing data should provide a more accurate measurements of the VAF for each mutation due to the higher read counts. However, in aggregate there is very little difference between the VAF histograms for whole-exome vs. deep sequencing (Figure 4.4A). Thus, methods that first cluster mutations according to their VAF without considering the variance in the VAFs of individual mutations from the observed read counts, including CITUP [98] and LICHeE [131], will not recognize differences in clustering between the low and high coverage data.

Examining the posterior probabilities of ancestral relationships between individual mutations (Figure 4.4B) reveals a striking difference between the low and high coverage datasets. The higher coverage targeted sequencing data has a much clearer distinction in ancestral relationships with many more pairs of mutations having posterior probability $\min_p Pr[X_{pi} \geq X_{pj}]$, the probability that mutation i precedes mutation j, close to 1 or 0, indicating high confidence in the ancestral relationships. The approach used by AncesTree exploits this higher confidence in individual ancestral relationships, both in grouping mutations and in determining the tree. For example, Figure 4.4C

shows the posterior probabilities of the VAF for 3 mutations. With lower coverage whole-exome sequencing, the distributions overlap, and there is no clear ancestral or grouping relationship between the mutations. With deep sequencing data, the variance of VAF for each mutation is smaller, and relationships between the mutations become apparent. The red mutation has a strong probability to be ancestral to both the blue and green mutations as $P(\text{red} \prec \text{green}) = 1.0$ and $P(\text{red} \prec \text{blue}) = 1.0$. In contrast, the blue and green mutations overlap significantly suggesting that these mutations should be clustered together. We find $P(\text{blue} \prec \text{green}) = 0.45$ and $P(\text{green} \prec \text{blue}) = 0.22$, both of which are within the interval $[0.5 - \alpha, 0.5 + \alpha]$ that we use for clustering. Thus, these mutations will be found in the same same strongly connected component when building the approximate ancestry graph.

4.3.3 Uncovering High-Confidence Ancestral Relationships

Fig. 4.5A shows the clonal tree inferred by AncesTree for CLL patient 077 previously analyzed with both PhyloSub and CITUP. The structure of our clonal tree closely resembles the trees reported by the other algorithms (Fig. 4.5C); in particular, both trees have two branching lineages containing mutations in the same genes. Furthermore, AncesTree returns purity estimates within 0.04 and 0.05, respectively of those reported by PhyloSub and CITUP across all 5 tumor samples. However, there are also important differences between the trees. PhyloSub and CITUP group together multiple pairs of mutations that AncesTree separates into successive clones. For instance, PhyloSub and CITUP cluster MAP2K1, HMCN1 and NOD1 into a single clone, while the tree produced by AncesTree shows these mutations as the result of three successive clonal expansions. The extremely high read counts (>450K) for these three mutations across all five samples give high confidence in the posterior probability of the ancestral relationships: the minimum posterior probabilities over all samples are 0.86 and 1 for the two edges. Similarly, $\Pr[\text{PLA2G16} \prec \text{EXOC6B}] = 1$ as is reported in AncesTree's clonal tree (Fig. 4.5B).

In addition to the differences in ancestry, the clonal tree output by AncesTree contains only a subset of the mutations, while the tree output by PhyloSub contains all mutations. We find that three of the missing mutations (in genes BCL2L13, NAMPTL and SAMHD1) have VAFs that are significantly higher than 0.5. Indeed the $1 - \gamma$ confidence interval used by our ILP implementation is strictly larger than 0.5. It is likely that these mutations occur in regions affected by copy number aberrations, thus violating the assumptions of our model. We examined the approximate ancestry

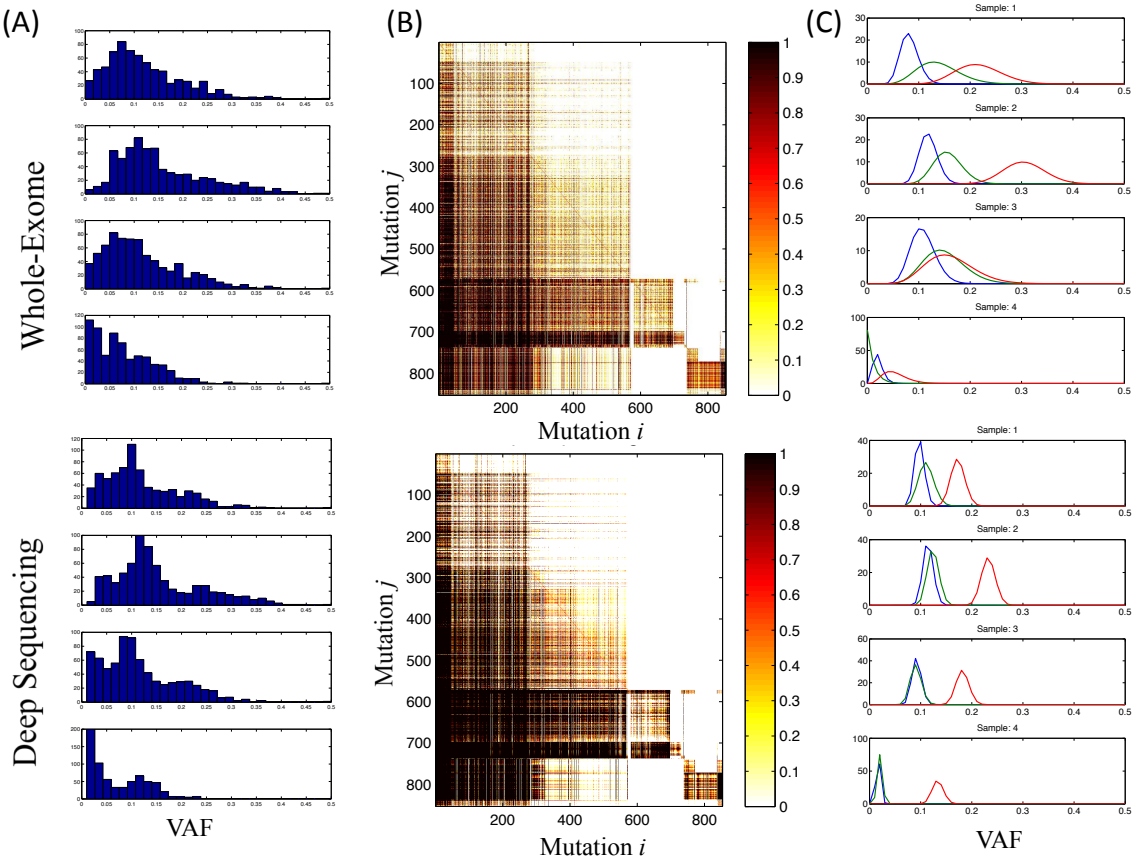


Figure 4.4: **Comparison of whole-exome (Top) and deep sequencing data (Bottom) for lung patient 330.** (A) Histogram of observed variant allele frequencies (VAFs) for all mutations for both datatypes does not reveal a significant difference between lower (201X) coverage (top) and higher (674X) coverage (bottom) sequencing data. (B) A heat map showing the posterior probability that mutation $i \prec j$, that is $\min_p Pr[X_{pi} \geq X_{pj}]$, for all pairs of mutations i and j. The asymmetry in the matrix reveals high confidence ancestral relationships, which become much clearer with higher coverage. (C) The posterior distribution of the VAF for three mutations given the observed read counts. In higher coverage data, the distributions become much tighter, revealing that the red mutation is ancestral to the blue and green mutations.

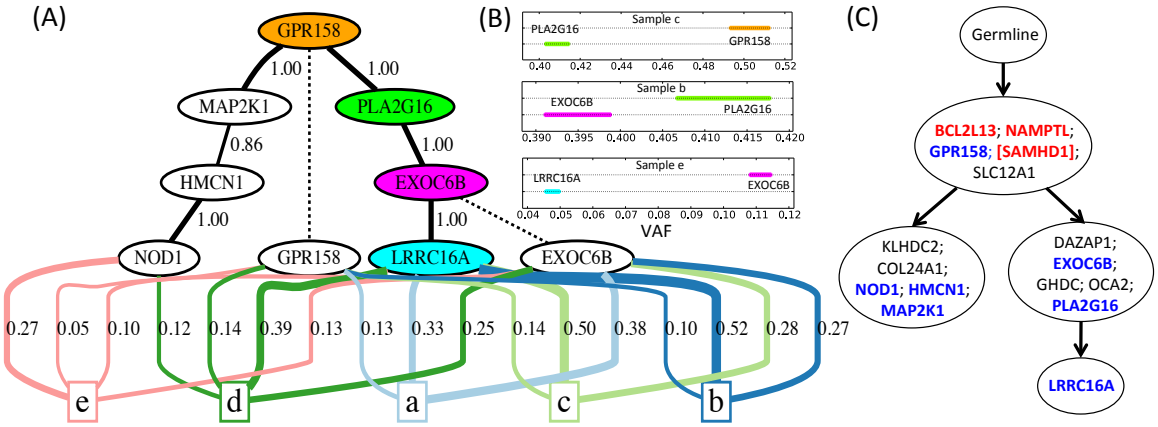


Figure 4.5: **Analysis of CLL patient 077 shows AncesTree's ability to infer successive clonal expansions.** (A) (A) The clonal tree output by AncesTree is indicated by the black solid edges whose weights correspond to the posterior probability of the ancestral relationship. Dashed edges are used to indicate ancestral clones which exist at the time of sequencing. The blocks labeled 'a' through 'e' each represent a sequenced sample, with colored edges indicating the inferred composition of clones and their fraction in each sample (only edges with usage at least 0.05 are shown). (B) The $1 - 10^{-6}$ confidence intervals of VAF for the sample with the weakest ancestral evidence for each of the edges connecting gene GPR158 to LRRC16A. (C) The tree reported by PhyloSub, which is identical to the tree reported by CITUP except for the addition of SAMHD1. Mutations indicated in blue are those present in part A. Mutations indicated in red likely occur in regions affected by copy number aberrations.

graph for this sample (Fig. A7) to determine why other mutations were missing from the tree output by AncesTree. We find that mutations in SLC12A1 and GPR158 only have incoming arcs from the three genes listed above whose VAFs exceed 0.5. Thus, there is no subtree of the ancestry graph that contains both SLC12A1 and GPR158. The other missing mutations (KLHDC2, COL24A1, DAZAP1, GHDC, OCA2) are all descendants of SLC12A in the ancestry graph. Of these missing mutations, all except for GHDC are also descendants of GPR158, but each violates the sum condition if added to the tree output by AncesTree.

4.3.4 Heterogeneity within Samples

Since AncesTree directly computes the usage matrix U, we obtain estimates of the amount of mixing, or intra-tumor heterogeneity, of clones within each analyzed sample. Specifically, for a given sample, the number of clones that are inferred to be mixed in a sample is the number of non-zero entries in the corresponding row of U. For each tumor we compute its *mixing proportion* to be the fraction entries in U that are non-zero (see Table C.1). Using the deep sequencing data we find that the CLL tumors have on average a mixing proportion of 1.0. This is much higher than the renal and lung

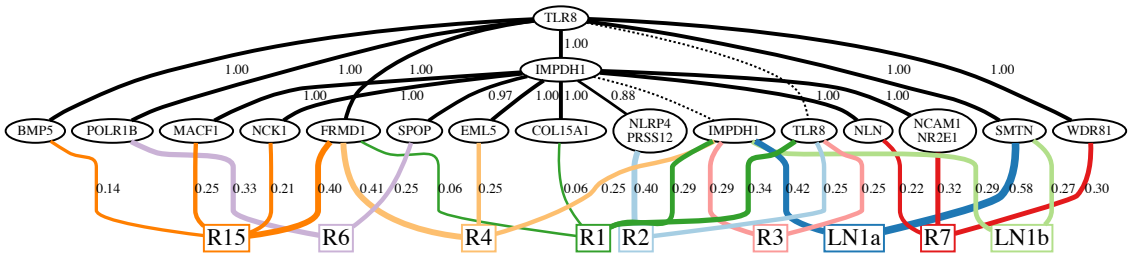


Figure 4.6: **Analysis of renal patient EV006 reveals distinctive sample composition.** The clonal tree output by AncesTree. Some sequenced sections (R6, R7) are mixtures of clones appearing only in those sections. In contrast, other sequenced sections (LN1a, LN1b, R3) are mixtures of clones that each appear in more than one section. In particular, both lymph node samples (LN1a and LN1b) are mixtures of the same two clones, but in different proportions.

tumors which have on average mixing proportions of 0.22 and 0.5 respectively. The higher mixing proportions for CLL are consistent with it being a liquid tumor, where mixing between clones is likely to be more common than in solid tumors.

We further analyzed one renal tumor, EV006, for which we obtained a relatively low mixing proportion of 0.21 (Figure 4.6). Samples R6 and R7 from this tumor were found to be the mixture of two and three distinct clones, respectively, that do not appear in other samples. This shows that AncesTree can infer the composition of individual samples containing clones distinct from all other samples. The remaining samples in this tumor all include a clone that appears in at least one other sample. Notably, the two lymph node samples, LN1a and LN1b, are inferred to be mixtures of the same two clones. The only difference between these two samples appears to be that LN1b contains a higher admixture with normal cells (0.45) than LN1a (< 0.01), and indeed the two lymph node samples are grouped together in the original analysis of this tumor by Gerlinger *et al.* [52].

4.4 Discussion

Reconstructing the evolutionary history of a tumor given VAFs measured in multiple sequenced samples for a single tumor is a challenging task. In this chapter we formalize this problem and present the AncesTree algorithm, which solves this problem. While we have demonstrated the advantages of AncesTree over other methods such as PhyloSub and CITUP, there are a number of ways that our approach may be improved.

First, throughout this work we assume that no measured mutations occur in regions affected by

copy number aberrations. Given the prevalence of such rearrangements in many types of cancers [3], AncesTree may not currently be applicable to some datasets. This represents an important area of future work. Second, AncesTree only outputs the single largest rooted subtree of the approximate ancestry graph that satisfies the sum condition. The algorithm may be applied iteratively by removing the clonal tree found at each step from the ancestry graph and re-running, thus returning a forest. However, it is unclear how the trees in this forest relate to each other or if there is an approach for joining them. Third, there are also ways for which the theoretical grounding of this work may be improved. For instance, what is the hardness for constant m? In practice the number m of samples is much smaller than the number n of mutations, and hence the problem may be fixed-parameter tractable. Finally, our use of the binomial distribution to model read counts may underestimate the variance; e.g. due to factors such as PCR artifacts. More realistic models of read counts may improve the performance of AncesTree.

We also note that the kidney and lung datasets analyzed here contain multiple sections of a solid tumor obtained at a single time point, whereas the CLL datasets contain samples obtained at different times. Future work will include investigation into handling multi-section samples and multi-time-point samples separately in order to account for potential time related dependencies.

In conclusion, we formalize the problem of reconstructing the clonal evolution of a tumor using single-nucleotide mutations as the Variant Allele Frequency Factorization Problem (VAFFP). This problem formalization provides a general framework for which numerous other related work may be interpreted. We derive a combinatorial characterization of the solutions to the VAFFP and show that the problem is NP-complete. We derive an integer linear programming solution to the VAFFP in the case of error-free data and extend this solution to real data with a probabilistic model for errors. The resulting AncesTree algorithm is better able to identify ancestral relationships between individual mutations compared to existing approaches.

Chapter 5

Reconstructing Cancer Genome Organization

A cancer genome is derived from the germline genome through a series of somatic mutations. Somatic structural variants – including duplications, deletions, inversions, translocations, and other rearrangements – result in a cancer genome that is a scrambling of intervals, or "blocks" of the germline genome sequence. In this chapter, we present an efficient algorithm, called Paired-end Reconstruction of Genome Organization (PREGO), for reconstructing this block organization of a cancer genome from paired-end DNA sequencing data.

We apply PREGO to simulated data, five ovarian cancer genomes that were sequenced as part of The Cancer Genome Atlas and six breast cancer genomes from [114]. We identify numerous rearrangements, or structural variants, in these genomes, analyze reciprocal vs. non-reciprocal rearrangements, and identify rearrangements consistent with known mechanisms of duplication such as tandem duplications and breakage/fusion/bridge (B/F/B) cycles. Finally, we demonstrate that PREGO efficiently identifies complex and biologically relevant rearrangements in cancer genome sequencing data.

Most of the work in this chapter is taken from [120], and was originally presented at the Second Annual RECOMB Satellite Workshop on Massively Parallel Sequencing (RECOMB-seq) in 2012. The extension to utilize a matched normal sample was referenced in [119] and along with the results on the breast cancer datasets were presented as part of an invited talk at the Third Workshop

on Computational Advances for Next Generation Sequencing (CANGS 2013) in conjunction with the Third IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS).

5.1 Related Work

Numerous methods have been developed in the past few years to identify structural variants using paired end mapping [150, 70, 31] and clustering of discordant read pairs. Some methods such as [139, 88, 133] have begun to incorporate additional signals such as *split reads* where a read aligns as two consecutive sequences separated by a gap (e.g. when the read directly contains a breakpoint) or *multiple-mappings* where a read may align to many positions in the genome when making variant predictions. In addition, when the sequencing coverage is high, the number of aligned reads [33, 174] or concordant pairs [181] provides an estimate of the number of copies of segments of the cancer genome. Lastly, some methods such as [151, 101] incorporate information from both discordant and concordant read pairs when making predictions. Details of all these methods can be found in several recent review articles [175, 102, 6].

However, most methods for structural variant prediction treat all aberrations as independent events. While this may be a reasonable assumption when analyzing germline variants, where the size and number of rearrangements that differ in comparison to a reference genome is relatively small, it is not appropriate when analyzing a cancer genome where a large portion of the genome may have undergone some type of rearrangement. A few exceptions are CNVer [101], a method designed for analyzing germline genomes rather than cancer, and nFuse [99] which requires the of RNA-seq data in addition to DNA sequencing data.

PREGO addresses the problem of reconstructing the complete organization of the cancer genome(s) present in a cancer DNA sample from the adjacencies and copy number information revealed by the concordant and discordant pairs from a paired-end resequencing approach. We define the Copy Number and Adjacency Genome Reconstruction Problem, a general formulation of the problem which we solve as a convex optimization problem. Our approach adapts and generalizes techniques that have been employed previously in genome assembly [127, 126, 100], ancestral genome reconstruction and genome rearrangement analysis in the presence of duplicated genes [5], and prediction of copy number variants [101]. In contrast to these works, we focus on the particular features and

challenges of cancer genome reconstruction including a broad class of rearrangements, aneuploidy, heterogeneity, and the availability of an "ancestral" reference genome.

5.2 The PREGO Algorithm

In this section we describe the Paired-end Reconstruction of Genome Organization (PREGO) Algorithm.

5.2.1 Intervals, Adjacencies, and Cancer Genome Reconstruction

Suppose the cancer genome is derived from the germline genome through a series of somatic rearrangements. We perform paired-end DNA sequencing on a cancer DNA sample $\mathcal{S}$. We assume that the sample $\mathcal{S}$ contains a genome sequence derived from the reference genome through some series of somatic structural rearrangements of blocks of DNA (we are not considering single nucleotide mutations). From the alignments of paired reads to the reference genome, we derive three pieces of information. First, we derive a partition of the reference genome into a sequence of intervals $\mathbf{I} = (I_1, I_2, \ldots, I_n)$. Each interval $I_j = [s_j, t_j]$ is the DNA segment from the positive strand of the reference genome that starts at coordinate s_j and ends at coordinate t_j. Since intervals also appear in the opposite direction in a cancer genome (e.g. due to an inversion), we denote by $I_{-j} = [t_j, s_j]$ the inverted DNA segment. Second, concurrently with the definition of $\mathbf{I}$, we derive a set $\mathcal{A}$ of novel adjacencies in the cancer genome. Each adjacency (I_j, I_k) indicates that the end t_j of interval I_j is adjacent to the start s_k of interval I_k in the cancer genome. Thus $\mathcal{A} \subseteq \{(I_j, I_k) | j, k \in \{\pm 1, \pm 2, \ldots, \pm n\}\}$. The partition $\mathbf{I}$ and associated set of adjacencies $\mathcal{A}$ are obtained by clustering discordant paired reads whose distance or orientation suggest a rearrangement in the cancer genome [136]. Any existing algorithm can be used to create such input and therefore, the decision about what data to use (i.e. ambiguous reads, split reads, read mapping quality, etc) are part of upstream processing. Third, we derive a read depth vector $\mathbf{r} = (r_1, \ldots, r_n)^T$, where r_j is the number of (paired) reads that align entirely within interval I_j. The read depth vector $\mathbf{r}$ is obtained by counting concordant pairs in each interval [22].

Our goal is to reconstruct the *block organization* of the cancer genome(s) in the cancer DNA sample $\mathcal{S}$ from the interval, adjacency, and copy number information. The block organization corresponds to a sequence $I_{\alpha(1)} I_{\alpha(2)} \ldots I_{\alpha(M)}$ of M intervals where each $\alpha(j) \in \{\pm 1, \ldots, \pm n\}$. We

formulate the following problem.

Copy number and adjacency genome reconstruction problem. *Given an interval vector* $\mathbf{I}$*, a set* $\mathcal{A}$ *of cancer adjacencies, and a read depth vector* $\mathbf{r}$ *derived from a cancer sample* $\mathcal{S}$*, find the cancer genome(s) that are most consistent with these data.*

The statement of this problem does not quantify "most consistent". Defining such a quantitative measure requires the consideration of several complicating factors. First, the measurements of adjacencies $\mathcal{A}$ and the partition $\mathbf{I}$ that they determine may be incomplete or inaccurate. Second, many cancer genomes are *aneuploid*, meaning that the copy number of many intervals is above and below the diploid number of 2, and thus the read depth vector may not accurately represent the actual copy number of each interval in the cancer genome. Finally, a cancer sample $\mathcal{S}$ consists of many tumor cells, and each of these may contain different somatic mutations. However, because most tumors are clonal originating from a single cell, a large fraction of the important somatic mutations will be found in all cells of the cancer sample $\mathcal{S}$. In this paper, we assume that the cancer sample $\mathcal{S}$ is genetically homogenous so that we need only construct the organization of one rearranged cancer genome. Below, we formulate a specific instance of the Copy Number and Adjacency Genome Reconstruction Problem that considers the case of a single cancer genome with errors in the set $\mathcal{A}$ of adjacencies, sequence $\mathbf{I}$ of intervals, and the copy numbers must be inferred from the read depth vector $\mathbf{r}$. We defer the question of heterogeneity to future work. We first consider the case of perfect data.

5.2.2 Perfect Data

We begin with the case that the data is complete and error-free: thus, all cancer adjacencies $\mathcal{A}$ are correctly measured, and we have correctly estimated the copy number of each interval from the read depth vector $\mathbf{r}$. Also, for ease of exposition, we assume that the reference and cancer genomes each contain a single chromosome. Specifically, we define the *interval count vector* $\mathbf{c} = (c_1, c_2, \ldots, c_n)^T$, where c_j indicates how many times the interval I_j occurs in $\mathcal{S}$. Note that in general $\mathbf{c}$ is not directly measured, but rather must be inferred from the data, and we consider this extension in the next section. We have the following problem.

Single chromosome copy number and adjacency genome reconstruction problem. *Given an interval vector* $\mathbf{I}$*, a set* $\mathcal{A}$ *of cancer adjacencies, an interval count vector* $\mathbf{c}$*, and the set* $\mathcal{R} =$

$\{(I_j, I_{j+1}) : j \in \{1, \ldots, n-1\}\}$ *of reference adjacencies, find a cancer genome* $I_{\alpha(1)} I_{\alpha(2)} \ldots I_{\alpha(M)}$ *satisfying:*

1. *For* $j = 1, \ldots, M-1$ *either* $(I_{\alpha(j)}, I_{\alpha(j+1)}) \in \mathcal{A}$ *or* $(I_{\alpha(j)}, I_{\alpha(j+1)}) \in \mathcal{R}$.

2. *For* $k = 1, \ldots, n$, *the total number of indices* j *with* $\alpha(j) = k$ *or* $\alpha(j) = -k$ *is equal to* c_k.

To solve this problem, we introduce the *interval-adjacency* graph, which is derived from the interval vector $\mathbf{I}$ and cancer adjacencies $\mathcal{A}$ (Figure 5.1). The interval-adjacency graph $G = (V, E)$ is an undirected graph with vertices $V = \{s_1, t_1, s_2, t_2, \ldots, s_n, t_n\}$ and edges $E = E_I \cup E_R \cup E_{\mathcal{A}}$. The set $E_I = \{e_I(j) = (s_j, t_j) : j = 1, \ldots, n\}$ of *interval edges* connect s_j to t_j for each j. The set E_R of *reference edges* connect the ends of adjacent intervals in the reference genome; i.e. $E_R = \{(t_j, s_{j+1}) : j \in \{1, \ldots, n-1\}$. The set $E_{\mathcal{A}}$ of *variant edges* connect intervals that are adjacent in the cancer genome, but are not adjacent in the reference genome. These adjacencies are inferred from the set of discordant pairs. Every $a \in \mathcal{A}$ defines a variant edge. The interval, reference, and variant edges in the interval-adjacency graph are analogous to the gray, green, and black edges, respectively, in the breakpoint graph used in genome rearrangement analysis [5]. The interval-adjacency graph represents the set of possible adjacencies of intervals in the reference genome similar to how the gene order graph used in [171] contains possible gene orderings. Although, in that case the nodes of the graph represent genes and edges are gene adjacencies. Note that any $v \in V$ is incident to exactly one interval edge I_j. Thus, we define $e_I(v) \in E_I$ to be the interval edge containing vertex v, and define $e_I(j) \in E_I$ to be the interval edge corresponding to interval I_j. Similarly, we define $e_R(v) \in E_R$ to be the reference edge containing vertex v, if such an edge exists, and $E_{\mathcal{A}}(v) \subseteq E_{\mathcal{A}}$ to be the set of variant edges incident to vertex v.

Now if the data $\mathbf{I}$, $\mathcal{A}$, and $\mathbf{c}$ are generated from an unknown cancer genome generated by a series of rearrangements, duplications and deletions that do not alter the chromosome ends (telomeres) s_1 and t_n, then the block organization of this cancer genome corresponds to an alternating path through G beginning at s_1 and ending at t_n that alternately traverses interval edges and non-interval edges (i.e. reference/variant edges), and where the number of times that each interval I_j is traversed (in either direction) on the path is equal to c_j (Figure 5.1). We require an alternating path since traversal of an interval edge is equivalent to selection of a block from the reference genome, and traversal of a reference/variant edge corresponds to a transition between blocks. Therefore, such an alternating path spells out a sequence of blocks from the reference genome. Formally, if we

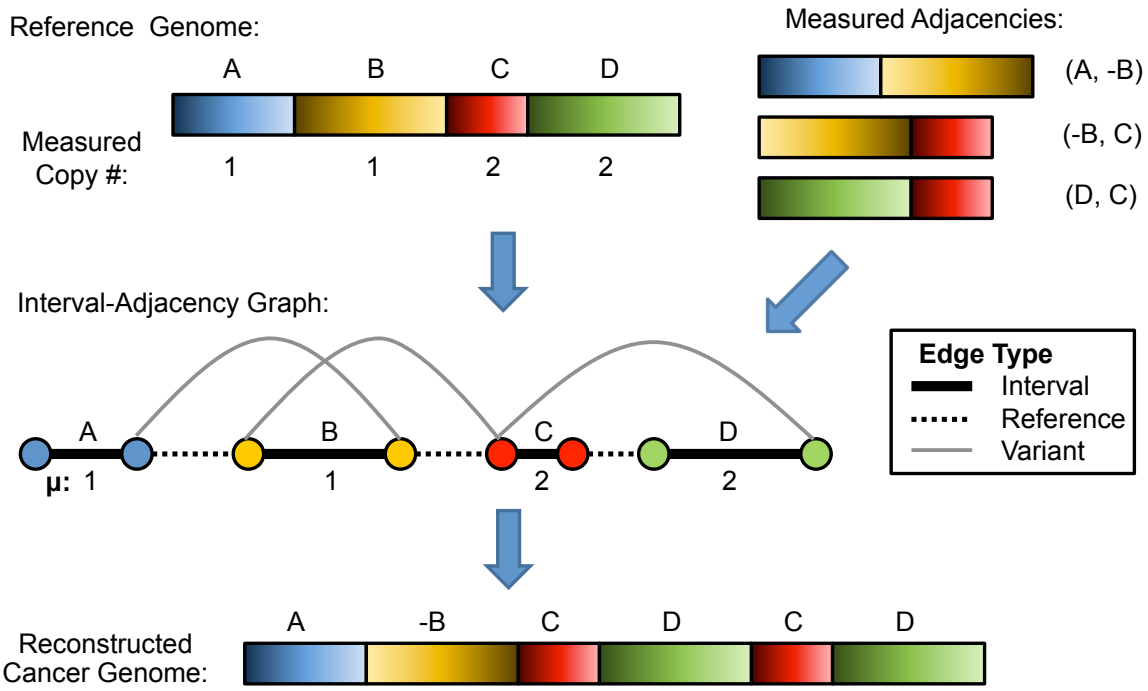


Figure 5.1: **Construction of the interval-adjacency graph.** Paired-end sequencing data partitions a reference genome into intervals A, B, C, and D with associated copy numbers. These intervals and the measured adjacencies are used to build an interval-adjacency graph. Deriving the appropriate multiplicities on this graph results in an Eulerian tour which reconstructs a cancer genome consistent with the input data. Here, a possible reconstruction is A -B C D C D where -B indicates that the block is in the inverse orientation from the reference genome. Another possible reconstruction is A B C D C D which results from the assignment of multiplicity 0 to some variant edges.

transform the interval-adjacency graph into a multigraph where the multiplicity of each edge equals the number of times it is traversed, then the multigraph has an Eulerian tour, as in the repeat graph, or deBruijn graph, in genome assembly algorithms [130, 127].

Conversely, if we are given data $\mathbf{I}$, $\mathcal{A}$,and $\mathbf{c}$ then we would like to infer an integer multiplicity $\mu(e)$ on each edge e such that an alternating Eulerian path from s_1 to t_n exists. We refer to s_1 and t_n as *telomeric vertices* and denote by $\mathcal{T} = \{s_1, t_n\}$ the set of telomeric vertices. Finding such an assignment of multiplicities can be formulated as an integer linear program (ILP). In particular, the restriction that the tour alternates between interval edges and non-interval (reference/variant edges) means that at each non-telomeric vertex v, the multiplicity of the interval edge $e_I(v)$ must equal the sum of the multiplicities of the reference edge $e_R(v)$ and variant edges $e_{\mathcal{A}}(v)$. Telomeric vertices $\mathcal{T} = \{s_1, t_n\}$ are excluded from this requirement since by definition they are only incident

to an interval edge, but not incident to any reference or variant edges. This constraint imposes the following *copy number balance* conditions on the multiplicities.

$$\mu(e_I(v)) = \mu(e_R(v)) + \sum_{a:a\in E_{\mathcal{A}(v)}} \mu(a), \quad \forall v \in V\backslash\mathcal{T}. \tag{5.1}$$

The following theorem follows directly from (5.1) and Kotzig's Theorem for alternating Eulerian paths [85] (see also [129]).

Theorem 5.2.1. *Given a connected interval-adjacency graph $G = (V, E)$, there exists a function $\mu : E \mapsto \mathbb{N}$ satisfying the copy number balance conditions* (5.1) *if and only if there exists a multigraph $G_\mu = (V, E_\mu)$ with edge multiplicities μ containing an alternating Eulerian Tour beginning at s_1 and ending at t_n.*

Finding such a function μ is the Eulerization problem and can be solved in polynomial time [100]. Applying the above result with the additional constraint $\mu(e_I(j)) = c_j$ for $j = 1, \ldots, n$ provides an interval-adjacency multigraph that contains an alternating Eulerian tour, corresponding to a cancer genome consistent with the data $\mathbf{I}$, $\mathcal{A}$, and $\mathbf{c}$. In a later section, we extend Theorem 5.2.1 to the case of multiple chromosomes by finding a set of alternating tours.

In the case of perfect data, there is guaranteed to be a solution to the Eulerization problem: one such solution is the assignment of multiplicities that correspond to the cancer genome. However, there is no guarantee on the uniqueness of the solution, and other solutions – including solutions that do not use all variant edges – are possible. Figure 5.1 gives an example. In the case of perfect data we could require that all variant edges are assigned non-zero multiplicity, thus ensuring that all variant edges from the cancer genome are used. However, in the case of imperfect data addressed below, such constraints are not appropriate as we expect such data to contain missing and false adjacencies due to difficulties in inferring adjacencies (structural variants) from paired-end sequencing data.

5.2.3 Imperfect Data

The previous section considered the case where the intervals $\mathbf{I}$ and adjacencies $\mathcal{A}$ were derived from a cancer genome with no errors, and where the interval count vector $\mathbf{c}$ was known. Now we consider

the situation that is presented by real data, where $\mathbf{c}$ is unknown and the adjacencies $\mathcal{A}$ may be incorrect (with missing adjacencies and/or false adjacencies). Instead of $\mathbf{c}$, we are given a (paired) read depth vector $\mathbf{r} = (r_1, \ldots, r_n)$ derived by the alignment of concordant paired reads to the reference genome. Each entry r_j is the number of concordant pairs of reads that when aligned to the reference genome lie entirely within the interval I_j. We use a probabilistic model to derive the most likely edge multiplicities μ in the interval-adjacency graph.

Specifically, let $L_1, L_2, \ldots, L_n$ be the lengths of intervals $\mathbf{I} = (I_1, I_2, \ldots, I_n)$, and let $L_R = \sum_{i=1}^{n} L_i$ be the length of the reference genome. Let $N = \sum_{i=1}^{n} r_i$ be the total number of concordant pairs that align within these intervals. Following the Lander-Waterman model, we assume that the reads are distributed uniformly on the genome, so that the number of reads that align to each interval follows the Poisson distribution with mean λ_j equal to the expected number of reads that align to an interval I_j. Of course, the Poisson distribution is an idealized assumption, and it has been shown that read depth is more accurately fit by a over-dispersed Poisson or negative binomial model [16, 181]. Nevertheless, the Poisson assumption has proven useful for copy number variant detection [101], and thus we use the Poisson model here, postponing consideration of other distributions to later work. We assume that the length of the cancer genome is approximately equal to the length L_R of the reference genome and $\mu_j = \mu(e_I(j))$ is the integer multiplicity assigned to the interval edge I_j. In a genome without any rearrangements, we expect $\frac{NL_j}{L_R}$ concordant paired reads to align within interval I_j (ignoring end effects). Since humans are diploid, we need to rescale this value to indicate the presence of two copies of interval I_j. Therefore, we introduce a variable τ that represents the expected number of copies of each interval in a non-rearranged sample. Given τ, the expected number of reads that align to an interval I_j appearing μ_j times in the genome is $\lambda_j(\frac{\mu_j}{\tau}) = \frac{NL_j}{L_R} \times \frac{\mu_j}{\tau}$. In general we set $\tau = 2$, but we defer discussion of handling multiple chromosomes until the next section.

We define a convex optimization problem that finds the maximum likelihood assignment of multiplicities $\mu(e)$ to all edges e in the interval-adjacency graph G, subject to the copy number balance conditions discussed in the previous section. The likelihood function is the product over all interval edges I_j of the Poisson probability of the observed number r_j of concordant pairs that align within interval edge I_j, which after taking the negative logarithm and removing constant terms gives us the (negative of) the likelihood function $L_{\mathbf{r}}(\mu) = \sum_j \lambda_j(\frac{\mu_j}{\tau}) - r_j \log(\lambda_j(\frac{\mu_j}{\tau}))$. Thus, we have the

following formulation.

$$\min_{\mu} L_{\mathbf{r}}(\mu) = \sum_{j=1}^{n} \lambda_j(\frac{\mu_j}{\tau}) - r_j \log(\lambda_j(\frac{\mu_j}{\tau})) \tag{5.2}$$

subject to

$$\mu(e_I(v)) - \mu(e_R(v)) - \sum_{a:a\in E_{\mathcal{A}}(v)} \mu(a) = 0, \tag{5.3}$$

$$\forall v \in V \backslash \mathcal{T}$$

Setting $\hat{c}_j = \mu_j$ gives the most likely multiplicity for the interval I_j in the cancer genome.

Note that [101] derives a similar formulation to predict germline copy number variants in human genomes, using a different construction based on bidirected graphs. Since human genomes are diploid, [101] add an additional source/sink vertex σ and add additional constraints that a flow of 2 be conserved across the graph. In contrast, most cancer genomes are aneuploid and might suffer deletions/duplications at the ends of chromosomes, this additional constraint is not applicable. We address this issue in the following section. [101] also show that their formulation reduces to a network flow problem that is solvable in polynomial time. The polynomial time result relies on two properties: (1) the objective function $L_{\mathbf{r}}(\mu)$ is separably convex; (2) the constraints are totally unimodular [69].

The interval-adjacency graph has a corresponding bidirected graph, and assignment of edge multiplicities in the interval-adjacency graph is equivalent to assignment of flow to the corresponding edges in the bidirected graph. Thus, the problem formulation in (5.2) above also reduces to a network flow problem that is solvable in polynomial time. In particular, for an interval-adjacency graph, we obtain a corresponding bidirected graph by adding orientation information to both ends of all edges in the original interval-adjacency graph. Specifically, for all interval edges (s_j, t_j) we assign a positive direction to the end at vertex s_j and a negative direction to the end at vertex t_j. For all reference edges (t_j, s_{j+1}) we assign a positive direction to the end at vertex t_j and a negative direction to the end at vertex s_{j+1}. For all the variant edges (v_1, v_2) we assign a positive direction for all $v \in \{v_1, v_2\}$ such that v is a vertex of the form s_j, and a negative direction if v is a vertex of the form t_j. We directly transfer all constraints on edge multiplicities. The problem formulation in (5.2) can now be equivalently described as a network flow problem on the corresponding bidirected graph since edge multiplicity assignment can be viewed as equivalent to flow assignment. Due to how we orient the bidirected edges, the copy number balance conditions from (5.1) are also equivalent to requiring

that the amount of flow going into each vertex is equal to the flow exiting the vertex.

The formulation above addresses the fact that sequencing data does not directly give copy numbers of intervals, but rather yields read depth, which we use along with adjacencies to estimate copy number simultaneously across all intervals. However, another source of error in the data are incorrect and missing adjacencies in the set $\mathcal{A}$. Incorrect adjacencies will subdivide intervals and alter the read depths in each of these intervals. Because our likelihood function considers both read depth and adjacencies when determining edge multiplicities, our algorithm is somewhat robust to the presence of incorrect adjacencies. Incorrect adjacencies that do not alter the estimated copy numbers of intervals are likely not to be used (i.e. the adjacency will be assigned multiplicity $\mu = 0$). Missing adjacencies will also affect the local structure of the interval-adjacency graph near the missing variant. In particular, all interval edges incident to the missing variant will be concatenated, and the corresponding variant edge will not be present. In most cases, we expect that the resulting reconstruction will simply not contain the missing adjacency. However, in other cases the missing adjacency may lead to additional errors in the reconstruction: for example the cases where the missing adjacency leads to large differences in the estimated copy number of the merged interval, or where the missing adjacencies overlaps with other variants. Our objective function (5.2) does not attempt to maximize the usage of variant edges, instead allowing the copy number estimates to determine whether variant edges are used are not. Defining an appropriate objective function that includes both copy number balance and scoring of variant edges is left for future work.

5.2.4 Multiple Chromosomes and Telomere Loss

We generalize the formulation above to handle two additional features of real data: (1) the reference and cancer genomes have multiple chromosomes, and (2) ends of chromosomes (telomeres) may be deleted in the generation of the cancer genome. First, to address the case of multiple chromosomes, we build a multichromosomal interval-adjacency graph $G = (V, E)$ where the interval and reference edges are the union of interval and reference edges in the unichromosomal interval-adjacency graph, respectively. The variant edges $E_{\mathcal{A}}$ are derived from the set $\mathcal{A}$ of adjacencies that connect intervals that are adjacent in the cancer genome, but not in the reference genome. These adjacencies are inferred from the discordant pairs, and now can include adjacencies between different chromosomes; e.g. those resulting from a translocation. The set $\mathcal{T}$ of telomeric vertices is the union of telomeric vertices of each chromosome, and consequently $|\mathcal{T}|$ is even. We now revise Theorem 5.2.1 to

multi-chromosomal genomes, where we now decompose the interval-adjacency graph into a set of alternating tours.

Theorem 5.2.2. *Given an multichromosomal interval-adjacency graph $G = (V, E)$ with telomeric vertices $\mathcal{T}$, there exists a function $\mu : E \mapsto \mathbb{N}$ satisfying the copy number balance condition* (5.1) *for all $v \in V/\mathcal{T}$ if and only if there exists a multigraph $G_\mu = (V, E_\mu)$ with edge multiplicities μ containing a set of edge-disjoint alternating tours that each begin and end at vertices in $\mathcal{T}$, and whose union is E_μ.*

A second feature of cancer genome data is that telomeres of the reference genome may be lost. In this case, the set $\mathcal{T}$ of telomeric vertices contains vertices other than the starts and ends of each chromosome of the reference genome. *De novo* telomere loss does not produce novel adjacencies in the cancer genome, and thus requires examining the read depth along the genome to find changes in concordant coverage, as used in read depth methods for copy number variant prediction [181]. Additionally, non-reciprocal translocations or breakage/fusion/bridge cycles produce novel adjacencies in the cancer genome and thus the drop in concordant coverage will be apparent over adjacent intervals in $\mathbf{I}$. We use a heuristic which determines the relative ratio of concordant reads to interval length between intervals to determine these drops in concordant coverage, and if at least one such case is found, we add an additional vertex σ to the interval-adjacency graph and to the set $\mathcal{T}$ of telomeric vertices. We also add variant edges from σ to the incident interval edge of the loss.

5.2.5 Utilizing a Matched Normal Sample

In most cancer sequencing experiments, in addition to sequencing a sample from a tumor, a matched normal sample (usually from a blood sample in the case of a solid tumor) is also sequenced. We now consider the case where a matched normal sample is also available. As before, let $L_1, L_2, \ldots, L_n$ be the lengths of intervals $\mathbf{I} = (I_1, I_2, \ldots, I_n)$, $L_R = \sum_{i=1}^{n} L_i$ be the length of the reference genome, and $N = \sum_{i=1}^{n} r_i$ be the total number of concordant pairs that align within these intervals. We are now able to observe a corresponding read depth vector $\mathbf{s} = (s_1, \ldots, s_n)$ where s_j is the read depth observed for interval I_j in the matched normal sample.

Previously we assumed that in a genome without any rearrangement we would expect to $\frac{NL_j}{L_R}$ concordant pairs to align within interval I_j. This relies on the assumption that reads will be uniformly distributed across all intervals in the genome based only upon the length of the interval.

In many instances this may not be the case due to factors such as GC content (fraction of bases that are a G or a C) or mappability of the region. Given a matched normal, we use the observed number of concordant pairs s_j that align to interval I_j to implicitly account for such biases. Therefore, given $\tau = 2$, the expected number of reads that align to interval I_j appearing μ_j times in the tumor genome is $s_j \frac{\mu_j}{\tau}$. We now can replace the previous estimate of λ_j with s_j and obtain the following objective function (while keeping the same constraints as (5.3).

$$\min_{\mu} L_{\mathbf{r}}(\mu) = \sum_{j=1}^{n} s_j(\frac{\mu_j}{\tau}) - r_j \log(s_j(\frac{\mu_j}{\tau})) \tag{5.4}$$

subject to

$$\mu(e_I(v)) - \mu(e_R(v)) - \sum_{a:a \in E_{\mathcal{A}}(v)} \mu(a) = 0, \tag{5.5}$$

$$\forall v \in V \backslash \mathcal{T}$$

5.3 Results

We ran our PREGO algorithm on both simulated data and real sequencing data from both ovarian and breast cancer genomes. We solve the convex optimization formulation in Equation (5.2) with CPLEX 12.1, using a piecewise linear approximation of the log term in the objective function, thus transforming the problem into an Integer Linear Program (ILP). Note, we use CPLEX rather than the efficient network flow algorithm discussed in a previous section as there is no good implementation of the later for bi-directed graphs.

5.3.1 Simulated Data

We tested our algorithm on simulated data to determine how robust the reconstructed interval-adjacency graphs are to various errors in the input data. Errors in the input data arise from a number of sources, and we studied the effect of two types of errors on the performance of a simulated sequence: sample contamination and read depth estimation error. We begin by constructing a cancer genome $C = I_{\alpha(1)} I_{\alpha(2)} \dots I_{\alpha(M)}$ consisting of 200 novel adjacencies: 100 homozygous deletions and 100 heterozygous deletions distributed over 22 autosomes (similar to the ovarian cancer genomes we analyzed in the next section). The lengths of the deletions are sampled from a normal distribution

with mean 10Kb and standard deviation 1Kb. From C we identify the sequence of intervals $\mathbf{I}$. We introduce 50 additional "false" adjacencies, where each false adjacency simply partitions an interval in $\mathbf{I}$ into three subintervals and adds a corresponding false deletion adjacency to the set $\mathcal{A}$. We then simulate 30X physical coverage of paired-end sequencing by sampling uniformly from C the starting positions of intervals, called *read-intervals*. We sample the length of these intervals from a normal distribution with mean 200 and standard deviation 10. We compute the resulting read depth r_j for each interval I_j.

Tumor samples are often a mixture of cells from the tumor itself and cells from non-cancerous cells. To model this type of error, we sample some proportion ρ of the read-intervals from the corresponding reference genome (i.e. the sequence of intervals $I_1 I_2 \dots I_n$), and sample $(1-\rho)$ of the read-intervals from the cancer genome C. Additional noise in the read depth estimation occurs due to experimental error (such as sequencing errors and alignment errors due to repetitive sequences in the reference genome) when estimating r_j. Thus, we add Gaussian noise to each r_j drawn from $\mathcal{N}(0, \phi r_j)$. We use ϕr_j rather than a single variance parameter to adjust the noise model for intervals with different read depths.

We ran our algorithm on the simulated datasets with error parameters ρ and ϕ and counted the number of edges in the interval-adjacency graph where the predicted multiplicity is the same as the correct multiplicity and averaged the results over 10 trials (Figure 5.2). The percent of correct edges drops by at most by 40%. Most of the errors made as the read depth variance ϕ increases are that heterozygous deletions are incorrectly called either homozygous no deletion (Figure 5.2).

5.3.2 Ovarian Cancer Sequencing Data

We analyzed DNA sequencing data from 5 ovarian cancer genomes and matched normal samples that were sequenced as part of The Cancer Genome Atlas (TCGA) (Table 5.1).Each sample was sequenced at 30X coverage using Illumina paired end technology with read length of 36bp. We downloaded the BAM files containing aligned reads from TCGA Data portal, and used the GASV algorithm [150] to cluster discordant pairs from each sample and from the matched normal using only those paired reads with mapping quality ≥ 30 in the BAM file. We then removed any clusters of discordant pairs that contain paired reads from both the tumor sample and the matched normal. In this way, we focus on somatic rearrangements. We also require that the discordant clusters are: (1) at least 1Mb away from the centromeres as annotated in the UCSC Genome Browser; (2) that

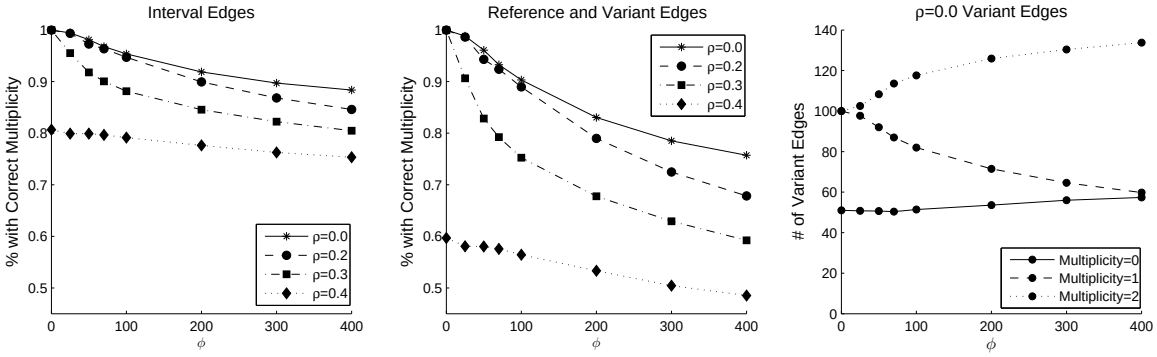


Figure 5.2: **Effect of sample contamination and read depth estimation errors on a simulated cancer genome.** ϕ is a scaling factor for the variance; for example $\phi = 400$ means that the noise model has a standard deviation 20 times r_j for interval I_j. We show the average percent of interval edges (left) and reference and variant edges (middle) correctly estimated over 10 trials. (Right) At $\rho = 0$, as ϕ increases most of the errors result from variant edges moving from the correct multiplicity of 1 (heterozygous deletion) to a multiplicity of 2 (homozygous deletion).

Dataset	ID	# Var Edges (Used)
OV1	TCGA-13-0890	771 (499)
OV2	TCGA-13-0723	562 (268)
OV3	TCGA-24-0980	311 (172)
OV4	TCGA-24-1103	340 (218)
OV5	TCGA-13-1411	389 (255)

Table 5.1: **Overview of ovarian cancer datasets.** Statistics of inferred interval-adjacency graphs for 5 ovarian genomes when a minimum of 5 discordant pairs are required to add a variant edge to the graph. A variant edge e is used if $\mu(e) > 0$.

they have a minimum number (either 5 or 10 as indicated below) of supporting discordant pairs; (3) introduce intervals no smaller than 8Kb in the interval sequence **I**. Restricting the lengths of the intervals in **I** allows for a better estimation of read depth, which is obtained by counting the number of concordant pairs within each interval I_j. We also restricted our analysis to the 22 autosomes. Table 1 gives the results of our algorithm when the cancer adjacencies $\mathcal{A}$ are restricted to those with at least 5 discordant pairs supporting each adjacency. The possible number of variants is quite large, and given the high rates of false positives with structural variant prediction [102, 106] many of these are not likely to be real variants. Since we are lacking a set of validated structural variants for these ovarian cancer genomes, we examine in the next section features of the interval-adjacency graph that might help distinguish true variants.

Reciprocal vs. Non-reciprocal Variants

Each measured adjacency in $A \in \mathcal{A}$ represents the result of cutting the reference genome at two locations, resulting in four free "ends" of two pairs I_p:I_{p+1} and I_q:I_{q+1} of interval edges. Two of these ends are then pasted together in the cancer genome. In some cases, e.g. an inversion or a reciprocal translocation, there is a corresponding partner adjacency A' that joins together the other two free ends of the intervals. Note that the GASV algorithm [150] clusters discordant pairs to identify partner adjacencies, when present. Thus, we distinguish two types of variant edges in the interval-adjacency graph: non-reciprocal edges, and (pairs of) reciprocal edges. Figure 5.3 shows examples of both types of edges, including reciprocal and non-reciprocal inversions and translocations. Moreover, following the cytogenetic nomenclature, we distinguish two types of translocations: classical translocations that preserve the orientation of both chromosomes and Robertsonian translocations that switch the orientation of one chromosome.

Thus, as a first step in evaluating the solutions produced by our algorithm, we examined the frequency with which reciprocal edges were used in the resulting interval-adjacency graph (i.e. the corresponding variant edge has inferred multiplicity > 0) versus the frequency with which non-reciprocal edges were used (Table 2). Note that reciprocal edges may be used in the following "trivial" way. If the inferred multiplicities on the two variant edges are both equal (i.e. $\mu(A) = \mu(A') = k$) and the inferred multiplicities of each pair of interval edges surrounding the corresponding breakpoints are also equal (i.e. $\mu(I_p) = \mu(I_{p+1})$ and $\mu(I_q) = \mu(I_{q+1})$) then the objective function (5.2) of the ILP is unchanged if one sets $\mu(A) = \mu(A') = 0$ and increases the edge multiplicity of the incident reference edges by k, thus removing the variant edges from the graph (Figure 5.3). We define reciprocal variant edges that satisfy this condition as *trivial* and those that do not satisfy this condition as *non-trivial*. Note that non-reciprocal variant edges have no equivalent trivial definition as altering the multiplicity assigned to a non-reciprocal variant edge would force a corresponding change in the multiplicity assigned the incident reference edges to maintain the copy number balance condition at the vertices of the variant edge. This change, however will cause the vertices at either end of the references edges to become unbalanced.

We analyzed the output of our algorithm for reciprocal (non-trivial) edges and non-reciprocal variant edges. For each type of reciprocal variant (inversions, classical translocations and Robertsonian translocations) we tested whether there was an association between a variant edge being

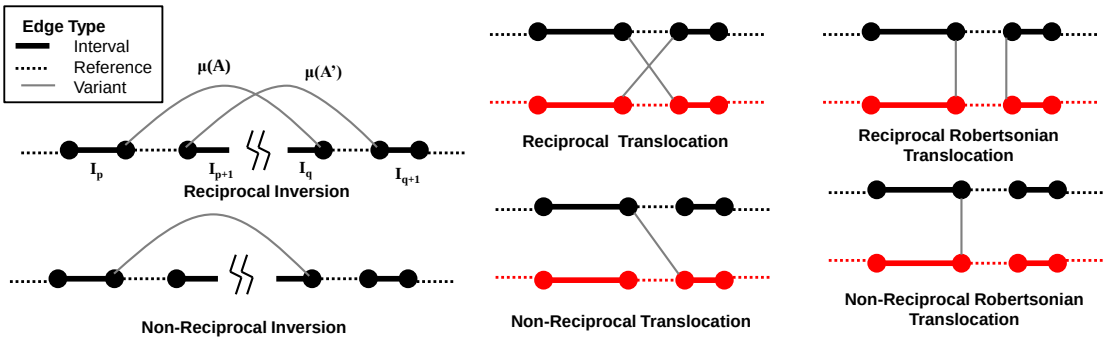


Figure 5.3: **Two classes of variant edges in the interval-adjacency graph.** Reciprocal variants are pairs of variant edges incident to the same four interval edges (I_p, I_{p+1}, I_q, I_{q+1}), while non-reciprocal variants are the cases where only a single variant edge is incident to the four interval edges defined by the variant edge. A trivial reciprocal variant has equal inferred multiplicities: $\mu(A) = \mu(A')$, $\mu(I_p) = \mu(I_{p+1})$, and $\mu(I_q) = \mu(I_{q+1})$.

used vs. unused, and reciprocal vs. non-reciprocal, using Fisher's exact test. We find that in most cases there is a statistically significant association, with a larger fraction of (non-trivial) reciprocal variant edges being used than non-reciprocal variant edges (Table 5.2). We surmise that the observed significant association between reciprocal variants and their use in the solution obtained by our method is an indication that it may be easier to satisfy the copy number balance conditions for vertices associated with a reciprocal variant. In particular, we may only use a non-reciprocal variant if additionally the concordant coverage on the surrounding intervals is indicative of a possible change in copy number. In this respect, non-reciprocal variant edges that are used may represent structural variants whose signature is supported by both read depth and discordant read pairs.

Reconstructed Variants

In this section, we give several examples of reconstructed variants in the OV genomes. First, we show two cases of reciprocal translocations, one trivial and one non-trivial, demonstrating that in some cases we may infer possible ordering of rearrangements - for example a translocation preceding a duplication (Figure 5.4).

We also find subgraphs of the interval-adjacency graph that suggest particular mechanisms of aberrant DNA repair in cancer genomes. In particular, Figure 5.5 shows part of the interval-adjacency graph of the proximal arm of chromosome 18 in sample OV2. We identify highly amplified

Reciprocal vs. Non Reciprocal Variant Edges								
Dataset	VariantType	R(all)	$\overline{R}$(all)	R(non-triv)	$\overline{R}$(non-triv)	NR	$\overline{NR}$	p-Val
OV1	T	179	41	75	13	9	58	$<$ 1E-15
OV1	I	46	20	16	12	2	29	3.46E-5
OV1	TO	210	46	70	16	9	38	2.79E-12
OV2	T	77	51	41	23	12	49	5.17E-7
OV2	I	21	15	9	5	10	21	0.057
OV2	TO	96	64	46	18	15	44	2.63E-7
OV3	T	61	13	19	3	6	30	2.111E-7
OV3	I	19	13	5	5	2	13	0.075
OV3	TO	58	26	22	8	7	28	1.92E-5
OV4	T	74	16	40	6	12	35	1.54E-9
OV4	I	10	0	2	0	3	12	0.073
OV4	TO	48	22	22	10	12	26	0.0036
OV5	T	93	19	29	7	8	37	2.30E-8
OV5	I	12	8	2	0	6	13	0.13
OV5	TO	82	26	22	8	7	34	2.29E-6

Table 5.2: **Statistical tests for variant edges.**Results of Fisher's exact test showing that non-trivial reciprocal edges are more likely to be used (assigned a multiplicity $\mu > 0$) in the interval-adjacency graph than non-reciprocal variant edges when a minimum of 5 discordant pairs is required to add a variant edge to the graph. Variant edges are classified as Inversion (I), Translocation (T), and Robertsonian Translocation (TO). Each variant edge is also classified as either reciprocal or not and by whether it is used ($\mu > 0$) or not used ($\mu = 0$). We report the number of edges of the following types: used reciprocal edges (R(all)), non used reciprocal edges ($\overline{R}$(all)), used reciprocal non-trivial (R(non-triv)), not used reciprocal non-trivial ($\overline{R}$(non-triv)), used non-reciprocal (NR), and not used non-reciprocal ($\overline{NR}$)

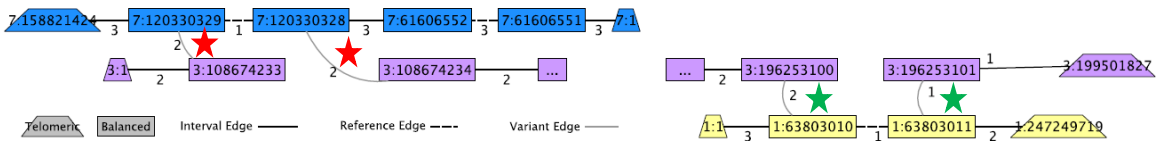


Figure 5.4: **Examples of reciprocal translocations in ovarian cancer sample OV5.** The Chr3/Chr7 translocation (left) has the same multiplicity on the variant edges (red stars) as well as on the corresponding pairs of incident interval edges making it trivial. The Chr1/Chr3 translocation (right) has different multiplicities on the variant edges (green stars) and is therefore non-trivial. In the Chr1/Chr3 translocation there is a single copy of Chr1 that does not use any variant edges, suggesting that only one copy of Chr1 is involved in the translocation, and that duplication of one of the translocated chromosomes occurs subsequent to the translocation.

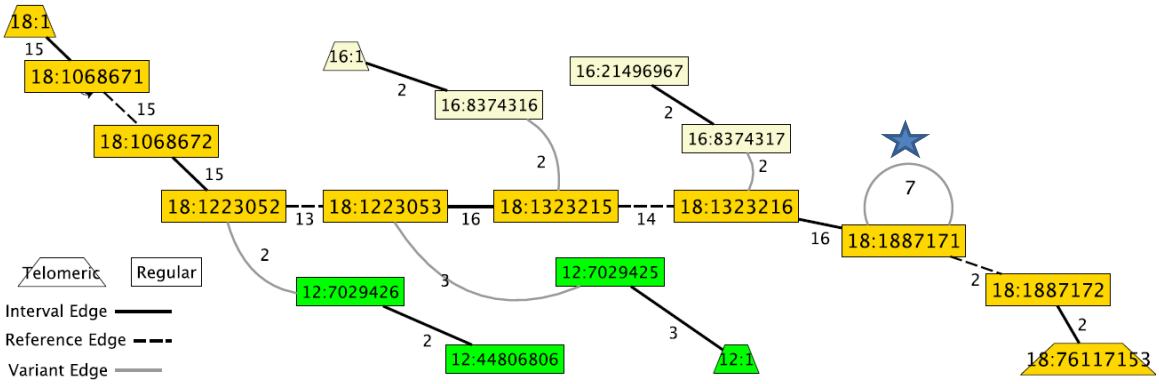


Figure 5.5: **Example of a Breakage/Fusion/Bridge Cycle on Chr18 in ovarian cancer sample OV2.** The first two Mb of Chr18 (starting in the upper left) is highly amplified, and this high multiplicity continues until the self-loop at Chr18:1887171 (blue star), which indicates an inverted repeat.

intervals that are incident to a loop variant edge that also has high multiplicity. Loops in the interval-adjacency graph are indication of inverted duplications, a signature of breakage/fusion/bridge cycles, a known source of genome instability in cancer genomes [58]. Oncogenes YES1 and TYMS appear in this amplified region, and both have been implicated in ovarian cancer [153, 80].

We also find tandem duplications on Chr2 of both OV2 and OV3 (Figure 5.6). Recently, a tandem duplication signature was reported in SNP data from Ovarian TCGA samples as well as in a pair of cell lines [113]. In particular, the cell line data included tandem duplications on Chr2. In the interval-adjacency graph, the location of these tandem duplications on the homologs of Chr2 are ambiguous. For example, OV2 has two copies of the variant edge, which may be one tandem duplication present on both copies of Chr2 or two tandem duplications present on one copy of Chr2. OV3 has two different locations where tandem duplications occur, one of which is within 2Mb of the duplicated region on OV2. All three of these tandem duplications occur with 4Mb of a duplication reported in [113] and one duplicated region in OV2 includes several cancer associated genes including PLB1, PPP1CB, ALK [107, 158, 76].

5.3.3 Breast Cancer Sequencing Data

We also analyzed DNA sequencing data from 6 breast cancer genomes and matched normal samples that were sequenced in [114] (Table 5.3).Most samples were sequenced at $\sim 30-40X$ coverage with Illumina paired-end sequencing except for sample PD4120, which was sequence at $\sim 188X$

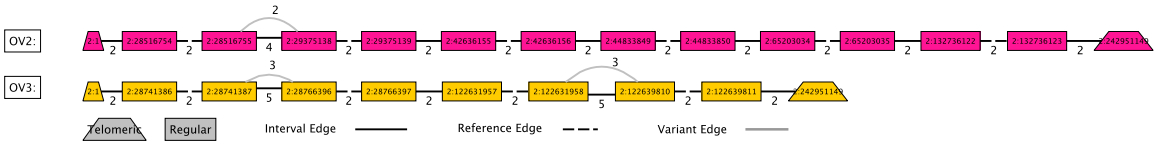


Figure 5.6: **Tandem duplications found on Chr2 in ovarian cancer samples OV2 and OV3.** OV2 has a single site of tandem duplication, while OV3 has two sites of tandem duplication. Note that the region duplicated in OV2 is much larger than the region duplicated on OV3, and the duplicated region in OV2 contains several cancer associated genes including PLB1, PPP1CB, ALK [107, 158, 76].

coverage. Data processing of these genomes and their corresponding BAM files was as described in the previous section except that we require that discordant clusters have at least 20 reads for all samples. For all genomes we ran PREGO with 4 different parameter settings: (1) Use matched normal and allow telomere deletions, (2) Use matched normal and don't allow telomere deletions, (3) Only use tumor data and allow for telomere deletions, and (4) Only use tumor data and don't allow telomere deletions. Unless otherwise indicated, all results reported are for the parameter settings that both use the matched normal and allow for telomere deletion.

Number of Edges		
Sample	Interval	Variant
PD3890	100	45
PD3904	315	160
PD3905	165	85
PD4005	148	75
PD4120	647	395
PD4199	123	57

Table 5.3: **Overview of breast cancer datasets.** The number of interval and variant edges for the 6 breast cancer datasets analyzed.

Reconstructed Variants

We identify numerous rearrangements in these genomes, including many previously reported by [114]. For instance, we recover the gain of 8q and the loss of 4p in sample PD3890, loss of Chr4 in PD4199, and the trisomy of Chr 1q (a hallmark for breast cancer [19]) in sample PD4005 – all predicted by [114]. Lastly, in the $\sim 188X$ sample, which we analyze in further detail in Chapter 2, we recover the trisomy of Chr 1q as well as the deletion of Chr 1p, predicted by both us [118] and [114]

We also infer many other previously unreported mutations, such as 3 tandem duplications in Chr1

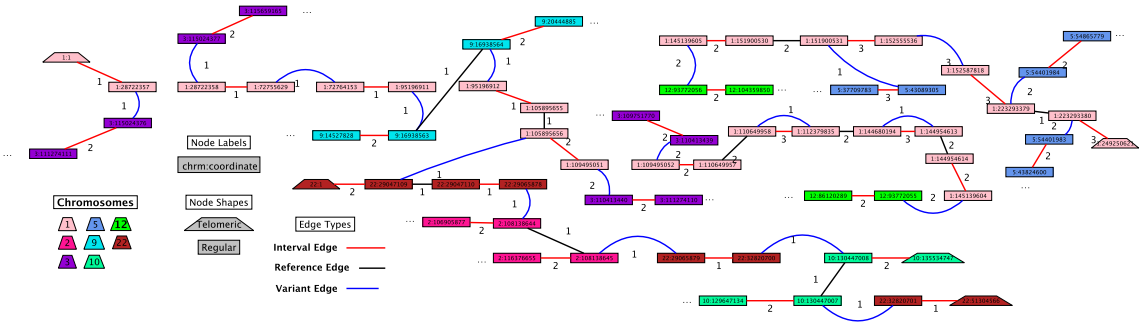


Figure 5.7: **A portion of the interval adjacency graph for breast cancer sample PD4120.** The interval adjacency graph for chromosome 1 and other connected chromosomes for PD4120, including a trisomy of Chr 1q and a deletion in Chr 1p.

of sample PD4005, including one that contains the gene PDE4B which has been shown to be up-regulated by the oncogene KRAS in colorectal cancer [162]. In sample PD4199a we find two regions of the genome that exhibit a large amount of amplification. The first occurs in a complex rearrangement on Chr17 (Figure 5.8) where extensive amplification occurs over two originally separate segments (hg19 positions 28019260-28046479, and 37207568-37958521). Several aberrations have caused these segments become adjacent in the cancer genome. The first segment is amplified up to 15 copies and contains the cancer related gene SSH2 [94], which has recently been suggested as a therapeutic drug target. The second segment contains between 40-53 copies and contains the well known ERBB2 gene [77] as well as the breast cancer related gene MED1 [37]. [114] report an amplification of ERBB2, but do not state the number of copies of the amplification, nor do they discuss the exact configuration of the organization of the chromosome. Another chromosome that exhibits an interesting structure in sample PD4199 is Chr12 which contains a highly amplified region (9 copies) adjacent to a self loop (Figure 5.9). We also note that this chromosome appears to have lost at least one copy of its telomere, the first step in a breakage/fusion/bridge (B/F/B) event. This configuration is the same as the one we (and others) postulated to be a signature of a B/F/B event [120, 58], as discussed in the previous section.

5.4 Discussion

The PREGO algorithm presented here combines copy number and adjacency information from paired-end sequencing data to infer cancer genome organization. However, the algorithm does not

consider all the issues involved in real cancer sequencing data. In particular, we assume that structural variants can be identified by mapping of discordant paired reads, but this is difficult for structural variants in repetitive regions of the human genome [70, 133]. Thus, there may be missing or incorrect adjacencies in the data. Similarly, estimates of read depth are difficult to obtain in repetitive regions [181]. While some of these issues may be addressed computationally, the more difficult cases will require longer reads and/or longer fragments for paired reads.

Beyond the issues with data quality are limitations on the inferred organization. While we derive multiplicities on the edges using adjacency and copy number data, we do not resolve the resulting paths through the interval-adjacency graph, except in simple cases. In many datasets, there will be many such paths and therefore many reconstructions of the cancer genome that are consistent with the data. Even the solution for the estimated edge multiplicities may not be unique. Resolving such longer paths requires additional information about connections between consecutive adjacencies, and such information is generally not available unless the distance between consecutive adjacencies is within the length of a read/fragment. In addition, the interval-adjacency graph does not contain allele-specific information about copy number variants, as considered in other work [58]. Finally, we assume that a cancer sample contains a single genome, when in fact most cancer samples contain DNA from a mixture of tumor cells, each with potentially different somatic mutations. It is possible that some of this intra-tumor heterogeneity could be resolved computationally. Alternatively, DNA sequencing of single cells, or smaller pools of cells, will minimize these effects.

In summary, we formulated the Copy Number and Adjacency Genome Reconstruction Problem of reconstructing a rearranged cancer genome and developed an efficient algorithm, called Paired-end Reconstruction of Genome Organization (PREGO), for a particular instance of this problem. We designed an optimization problem on the interval-adjacency graph, which is related to the breakpoint graph used in genome rearrangement studies. We applied our algorithm to simulated data, ovarian cancer genomes sequenced as part of The Cancer Genome Atlas (TCGA) and breast cancer data from [114] and reconstruct structural variants in these genomes. We analyzed the patterns of reciprocal vs. non-reciprocal rearrangements, and identified rearrangements consistent with known mechanisms of duplication such as tandem duplications and breakage/fusion/bridge cycles.

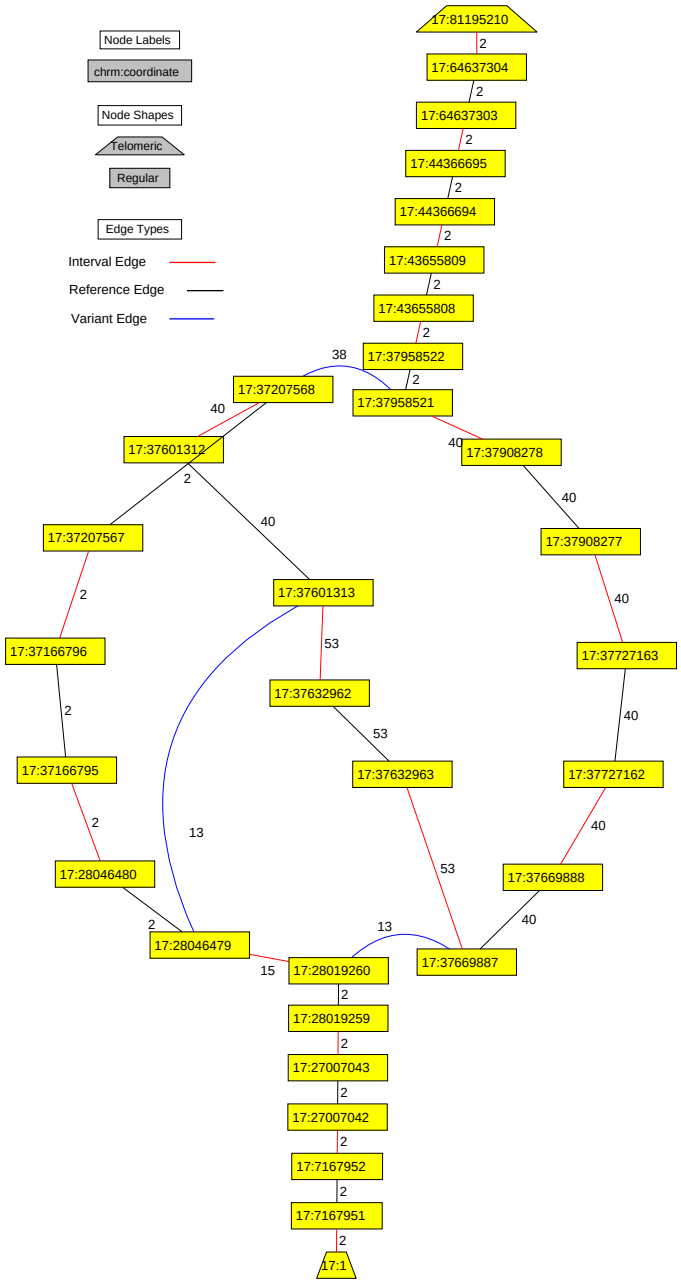


Figure 5.8: **The interval adjacency graph along with assigned edge counts for Chr17 in breast cancer sample PD4199.** A series of complex structural variants have resulted in large amplifications in two distinct segments of the chromosome. One decomposition of this graph implies one normal copy of Chr17 and one copy that contains multiple, nested duplications.

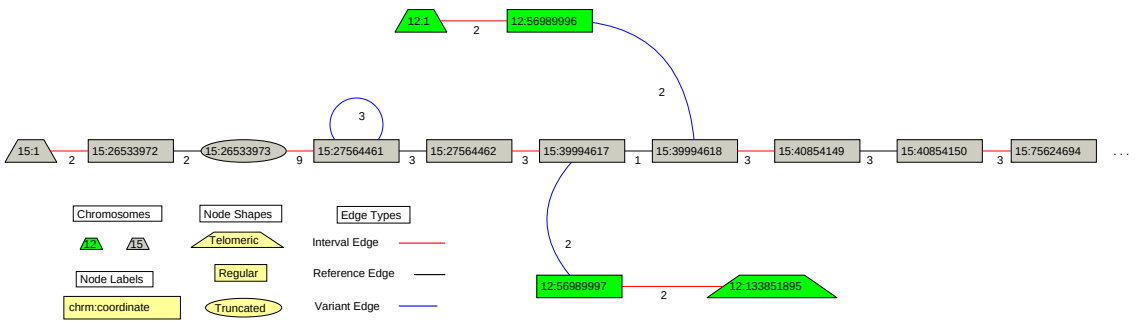


Figure 5.9: **The interval adjacency graph along with assigned edge counts for Chr12 and Chr15 in breast cancer sample PD4199.** The self loop which is adjacent to a region of high copy (9 copies here) is a signature of a breakage/fusion/bridge event.

Chapter 6

Detecting Simultaneous Rearrangements in Cancer Genomes

The evolution of a cancer genome has traditionally been described as a sequential accumulation of individual mutations – including chromosomal rearrangements – over long period of time. Several recent studies report that some rearrangements exhibit a complicated structure involving multiple, closely located breakpoints. In 2011 Stephens *et al.* [154] proposed a novel mechanism of chromosomal rearrangement in cancer termed *chromothripsis* in order to explain extreme cases of this phenomenon. The chromothripsis model posits that in some instances a small portion of the genome in a cancer cell undergoes a cataclysmic event resulting in a shattering of genetic material that is subsequently pieced back together in apparently random order. Under this hypothesis, many mutations are acquired *simultaneously* in contrast to the long standing *sequential* model.

Since the chromothripsis model was proposed, a number of studies – using a myriad of different criteria for determining the presence of chromothripsis – have reported varying rates for this phenomenon across different cancer types [154, 97, 138]. However, there is a paucity of formal mathematical models or descriptions of how such cataclysmic events would manifest themselves in the face of noisy sequencing data, making it unclear how to unbiasedly distinguish whether or not a chromothripsis (or another simultaneous event) has actually occurred.

In this chapter we present work related to building a formal mathematical model of simultaneous events from high-throughput DNA sequencing data. This takes the form of a rigorous analysis of one signature of chromothripsis proposed by [83] and two measures we developed that place a lower bound on the fraction of rearrangements in a sample that were likely caused simultaneously. The analysis of the signature suggested by [83] was presented in the form of a platform presentation at the 2013 Wellcome Trust Scientific Conferences/Cold Spring Harbor Laboratory Conference on Genome Informatics and at a poster session at the 2013 Microsoft Research Computational Aspects of Biological Information Conference. The two measures of simultaneous rearrangements were part of a collaboration with an undergraduate student, Caleb Weinreb, and were presented at the 2014 RECOMB-CG Satellite Workshop [170].

6.1 Related Work

Cancer is driven by somatic mutations in a population of cells [116]. These somatic mutations range in scale from single nucleotide mutations to large-scale chromosomal rearrangements. Traditionally, the evolution of a cancer genome has been described as a sequential accumulation of such mutations over many cell divisions. In 2011, however, Stephens *et al.* [154] suggested that cancer genomes may also acquire tens to hundreds of genomic rearrangements simultaneously as part of a one-time catastrophic event termed *chromothripsis*. It was proposed that during a chromothripsis event a portion of one, or a few chromosomes, shatter into many fragments. DNA repair mechanisms then stitch together some subset of these genomic fragments into a mosaic chromosome (Figure 6.1). Fragments not included in the reconstructed chromosome are lost, and therefore appear as interspersed deletions throughout the region.

The Chromothripsis hypothesis was formed as a means of describing observations in data that seemingly could not be described using the standard sequential model of genome rerrangements. A related phenomenon reported by Berger *et al.* [17] was later named *chromoplexy* by Baca *et al.* [11]. Both chromothripsis and chromoplexy involve simultaneous breakage and repair at multiple genomic locations, although with slight differences: e.g. chromoplexy is proposed to favor inter-chromosomal over intra-chromosomal rearrangements.

Simultaneous breakage and repair at multiple genomic locations has not yet been measured *in*

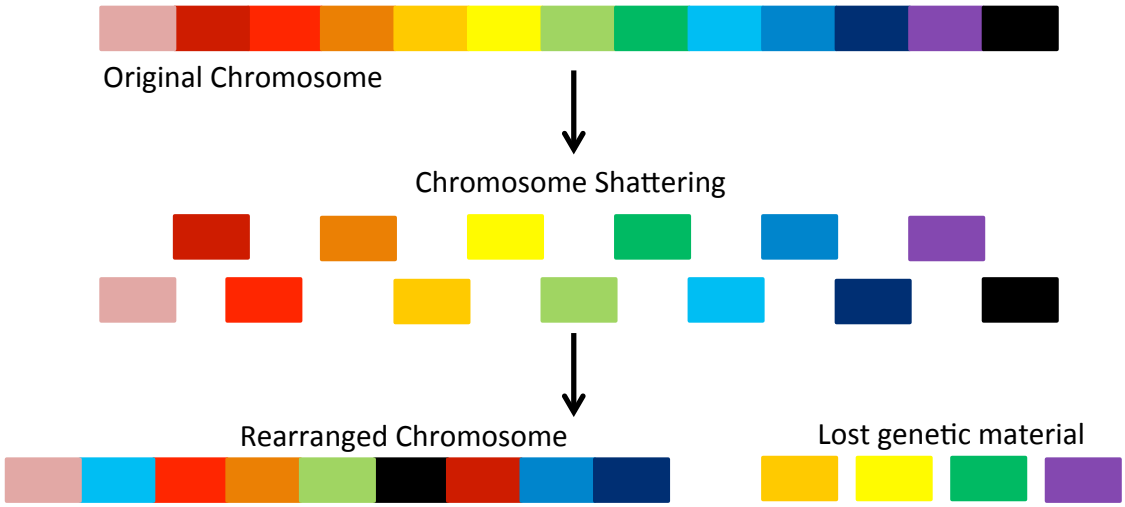


Figure 6.1: **A diagram depicting a chromothripsis event on a single chromosome.**

vivo. Thus, to infer that such an event has occurred one must argue that simultaneous rearrangement is a more plausible explanation for the observed sequencing data than sequential accumulation of rearrangements. Several different signatures have been proposed as the defining characteristics of chromothripsis [83, 96] including clustering of rearrangement breakpoints and a small number of oscillating copy number states. Other suggested signatures, including "the ability to walk the derivative chromosome" are not well defined, making them difficult to interpret. While these signatures may be suggestive of a simultaneous, or *one-off*, rearrangement event, they do not conclusively establish the occurrence of such an event. In addition, there is variability in how these criteria are implemented [154, 97, 139] making it unclear how to interpret or compare results across different studies.

The lack of formal models and definitions for detecting chromothripsis and chromoplexy has led to a growing debate about whether these are true phenomena [152, 81]. For instance, Sorzano *et al.* [152] suggest that the observed clustered rearrangement breakpoints do not exist in every cell, but rather reflect heterogeneity in the tumor population as a result of an event such as breakage-fusion-bridge (B/F/B) cycle. The fundamental question underlying this debate is how to identify *simultaneous* acquisition of rearrangements – the defining feature of chromothripsis/chromoplexy – in a cancer genome, given sequence data from a tumor sample and matched normal.

The original chromothripsis publication [154] used Monte Carlo simulations to demonstrate that

it was unlikely to observe only a few copy number states under a sequential model. While variations on this approach have been adopted in several other studies [97, 139], recent reports have questioned the conclusions drawn from this approach. For example, [81] demonstrate that a small but significant proportion (3.9%) of simulated datasets with sequential accumulation of $50-55$ breakpoints exhibit three or fewer copy states, thus showing a high false positive rate with this approach. Recently, other methods for identifying simultaneously formed rearrangement clusters have been proposed. ShatterProof [55] provides a framework for combining the various proposed criteria of chromothripsis [83] to generate a composite likelihood score. ChainFinder [11] detects chromoplexy using a graph based model which identifies closed chains of rearrangements that are unlikely to have arisen independently.

6.2 Analysis of a Proposed Signature of Chromothripsis

In this section we present a formal model of chromothripsis using strings which we then use to analyze one of the signatures of chromothripsis suggested by Korbel *et al.* [83], the ability to walk the derivative chromosome. Furthermore, we quantify the exact instances when this signature would arise under perfect data and demonstrate that this signature degrades quickly when noise is added. This further motivates the need for alternative methods to detecting simultaneous events from sequencing data of cancer genomes. In the following section, we present several other measures which may be more capable of detecting the presence of simultaneously obtained rearrangements from noisy data.

6.2.1 A Formal Model of Chromothripsis

We first present a simple mathematical model of chromothripsis using strings and which will serve as the basis for further investigations. We begin by describing a mathematical model of unichromosomal genome G that undergoes a chromothripsis event. Suppose that we label consecutive intervals, or genomic segments, along the original genome G using the characters $1, 2, \ldots, n$. An interval $g \in \{1, \ldots, n\}$ that subsequently appears in the reverse orientation can then be denoted as $-g$. Any linear/circular sequence of characters from the set $\{\pm 1, \ldots, \pm n\}$ therefore represents a possible rearrangement of segments from the original genome G. Using this notation, we can now define which such configurations may result from a chromothripsis event.

Definition 6.2.1. *We define a linear string C to be a* chromothripsis string for G *if C is a signed permutation of 2 or more characters from the set $\{1, \ldots, n\}$.*

We now present some further notation related to this definition of a chromothripsis string that will be useful in later sections. Each $g \in \{1, \ldots, n\}$ can be denoted as an interval with two extremities: $[g_t, g_h]$ where g_t is the *tail extremity* and g_h is the *head extremity* of the interval denoted by the character g. We define the *extremity set* $V = T \cup H$ where $T = \{g_t \mid g \in \{1, \ldots, n\}\}$ and $H = \{g_h \mid g \in \{1, \ldots, n\}\}$. An interval $g \in \{1, \ldots, n\}$ that appears in the reverse orientation is denoted as $-g = [-g_t, -g_h] = [g_h, g_t]$. Therefore, any interval $g \in \{\pm 1, \pm 2, \ldots, \pm n\}$ can be written as an ordered pair of extremities from V where one extremity is from H and the other from T. Notice that once a single extremity for an interval g is defined, the other extremity, or *obverse extremity* is completely predetermined. Therefore we define an *adjacency* to be an unordered pair of extremities from V, indicating an adjacency between two intervals from $\{\pm 1, \ldots, \pm n\}$. Suppose $g, g' \in \{\pm 1, \ldots, \pm n\}$, the adjacency between (g, g') is defined in Equation (6.1).

$$\mathcal{A}(g, g') = \begin{cases} (g_h, g'_t), & g > 0, g' > 0 \\ (|g|_t, |g'|_h), & g < 0, g' < 0 \\ (g_h, |g'|_h), & g > 0, g' < 0 \\ (|g|_t, g'_t), & g < 0, g' > 0 \end{cases} \tag{6.1}$$

Any linear/circular sequence of characters C from the set $\{\pm 1, \ldots, \pm n\}$, representing a possible rearrangement of segments from the original genome G, can be represented as a set of such unordered pairs of extremities from V. Suppose that $C = c_1 c_2 \ldots c_m$ is a linear/circular string where each $c_j \in \{\pm 1, \pm 2, \ldots, \pm n\}$. We define the *adjacency set* of C as $\mathcal{A}(C) = \{\mathcal{A}(c_j, c_{j+1}) : j = 1, \ldots, m-1\}$ if C is linear and $\mathcal{A}(C) = \{\mathcal{A}(c_j, c_{j+1}) : j = 1, \ldots, m-1\} \cup \{(c_m, c_1)\}$ if C is circular. We also note any such sequence of characters C can easily be depicted using a graph $G = (V, E)$ similar to the interval-adjacency graph introduced in Chapter 5 where $V = T \cup H$ is just extremity set defined above and $E = E_{\mathbf{I}} \cup \mathcal{A}(C)$ is the union of interval edges $E_{\mathbf{I}} = \{(g_t, g_h) | g = 1, \ldots, n\}$ and and adjacency edges $\mathcal{A}(C)$ as defined as above.

We define $\mathcal{T}(C)$, the *terminal* set of C, as the set of extremities from V appearing in some adjacency in $\mathcal{A}(C)$ but where the obverse extremity for the interval does not appear in any adjacency in $\mathcal{A}(C)$. Note that that an extremity in the terminal set indicates that the associated character

(or interval) must appear at the end of the string C, and in terms of genomes may be interpreted as the telomere. We define the *sorted adjacency string* $\pi(C) = \pi_1\pi_2\dots\pi_p$ to be the permutation of unique extremities appearing in some element of $\mathcal{A}(C)$, to be listed in sorted order according to their position in G. We say that a string $\pi(C)$ is *H/T alternating* if its characters alternate between being members of the sets H and T. Figure 6.2(A-C) shows three tumor genomes along with their strings C, adjacency sets $\mathcal{A}(C)$, corresponding graph representations, terminal sets $\mathcal{T}(C)$ and the sorted adjacency string $\pi(C)$. Figure 6.2(A-B) correspond to chromothripsis strings while Figure 6.2C does not.

6.2.2 H/T Alternating

Korbel *et al.* [83] suggest that one proposed signature of chromothripsis (the ability to walk the derivative chromosome) is defined by an alternating head/tail pattern observed when measured tumor adjacencies are sorted according to their position in the reference genome. In our model, this corresponds to the string $\pi(C)$ having a H/T alternating pattern. In Theorem 6.2.1 we explicitly quantify 2 necessary and sufficient conditions that determine when a chromothripsis string C will exhibit a H/T alternating pattern. In particular, the first condition corresponds to when a chromothripsis event occurs somewhere in the middle of a chromosome, leaving the telomeres in place (Figure 6.2A), while the second condition corresponds to a degenerate case where both ends of the derivative chromosome originate in a particular configuration from the interior of the chromosome. It is important to note that the cases detailed in Theorem 6.2.1 do not include the important case where a chromothripsis event includes a telomere (Figure 6.2B). Further, Figure 6.2C shows one example where a genome that has not undergone a chromothripsis event exhibits H/T alternating pattern. Lastly, Figure 6.2D shows the relationship of the three example genomes in terms of being chromothripsis strings and H/T alternating, thus demonstrating that the H/T alternating pattern suggested by [83] does not completely capture chromothripsis events.

Theorem 6.2.1. *Suppose that C is a chromothripsis string for G. $\pi(C)$ is H/T alternating if and only if the terminal set $\mathcal{T}(C)$ is one of the following:*

1. *$\mathcal{T}(C) = \{\pi_1, \pi_p\}$ where $p = |\pi(C)|$, $\pi_1 \in H$, and $\pi_p \in T$.*
2. *There exists some k such that $\mathcal{T}(C) = \{\pi_k, \pi_{k+1}\}$ where $\pi_k \in T$, and $\pi_{k+1} \in H$.*

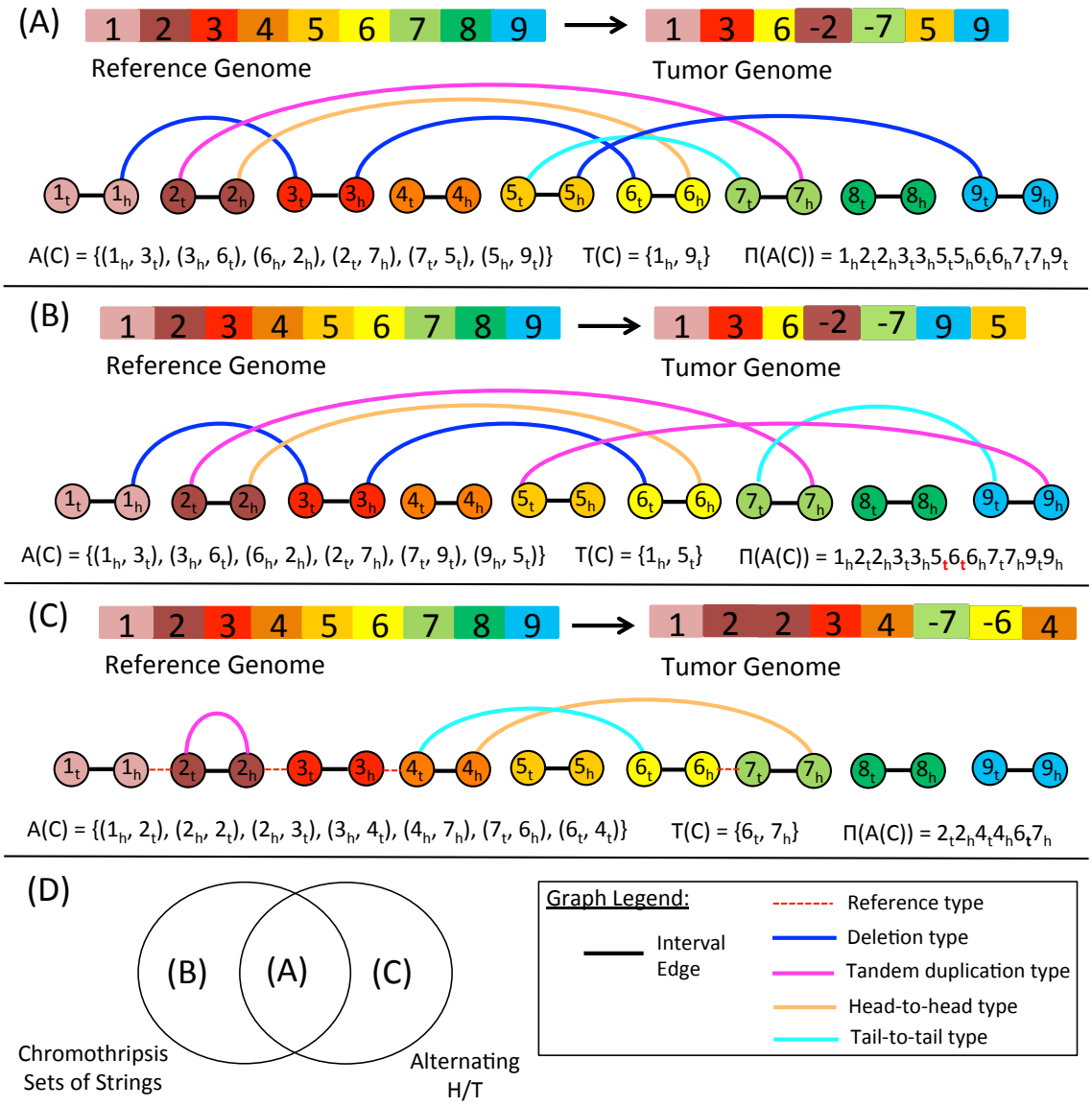


Figure 6.2: **Graph representations and H/T alternating status of several chromosomes that have undergone chromothripsis.**(A)-(C) Three different tumor genomes obtained as a rearrangement of blocks from the reference genome along with their corresponding graph representation (where different types of adjacency edges in $\mathcal{A}(\mathbf{C})$ are indicated using different colors), adjacency sets, terminal sets and sorted adjacency string. (D) Shows the relationship of the previous three examples in terms of being a chromothripsis string or exhibiting a H/T alternating pattern.

Proof. Let C be a chromothripsis string for G.

($\Rightarrow$) Assume that $\pi(C)$ is H/T alternating. We will proceed by contradiction. Assume that none of the above conditions about $\mathcal{T}(C)$ are true. In particular, we can also assume that there exists

some $k' \in \{2, \ldots, p\}$ such that $\pi_{k'} \in \mathcal{T}(C)$ but $\pi_{k'-1}, \pi_{k'+1} \notin \mathcal{T}(C)$. Therefore, there must exist some $g, g' \in \{1, \ldots, n\}$ such that $\pi_{k'-1} = g_h$ and $\pi_{k'+1} = {g'}_t$. However, if $\pi_k \in H$, this implies that $\pi(C)$ does not alternate. Similarly, if $\pi_k \in T$, this implies that $\pi(C)$ does not alternate – a contradiction. Hence, one of the above conditions about $\mathcal{T}(C)$ must be true.

($\Leftarrow$) We will consider each possible telomere set $\mathcal{T}(C)$ separately and show that for each it is true that $\pi(\mathcal{A}(C))$ is alternating.

Assume that $\mathcal{T}(C) = \{\pi_1, \pi_p\}$ where $p = |\pi(C)|$, $\pi_1 \in H$, and $\pi_p \in T$. We will proceed by contradiction. Assume that $\pi(C)$ is not alternating. This implies (without loss of generality) that there exists some $k \in \{1, \ldots, p-1\}$ and $g, g' \in \{1, \ldots, n\}$ such that $\pi_k = g_h, \pi_{k+1} = {g'}_h$ (that is $\pi_k, \pi_{k+1} \in H$). This implies that ${g'}_t \notin \pi(C)$ and therefore ${g'}_h = \pi_{k+1} \in \mathcal{T}(C)$. And since $k+1 > 1$, it must be the case that ${g'}_h = \pi_{k+1} = \pi_p$, therefore contradicting our assumption that $\pi_p \in T$. The argument for $\pi_k = g_t, \pi_{k+1} = {g'}_t$ is similar. Hence, $\pi(C)$ must be H/T alternating.

Assume there exists some k such that $\mathcal{T}(C) = \{\pi_k, \pi_{k+1}\}$ where $\pi_k \in T$, and $\pi_{k+1} \in H$. We will proceed by contradiction. Assume that $\pi(C)$ is not alternating. This implies (without loss of generality) that there exists some $k' \in \{1, \ldots, p-1\}$ and $g, g' \in \{1, \ldots, n\}$ such that $\pi_{k'} = g_h, \pi_{k'+1} = {g'}_h$ (that is $\pi_{k'}, \pi_{k'+1} \in H$). This implies that ${g'}_t \notin \pi(C)$ and therefore ${g'}_h \in \mathcal{T}(C)$. There are only two possible values of k' such that $\pi_{k'+1} \in \mathcal{T}(C)$. The first possibility is that $k' = k-1$. If $k' = k-1$, then $\pi_k = {g'}_h$, a contradiction with our assumption that $\pi_k \in T$. The second possibility is that $k' = k$. If $k' = k$, then $\pi_k = g_h$, a contradiction to our assumption that $\pi_k \in T$. The argument for $\pi_{k'} = g_t, \pi_{k'+1} = {g'}_t$ is similar. Hence, $\pi(C)$ must be H/T alternating. □

In addition to showing that not all chromothripsis strings exhibit the H/T alternating property, we have also proven the following Theorem (proved in Appendix D) showing the probability that a randomly chosen chromothripsis string will exhibit the H/T alternating property and that this probability only depends on the number m of characters in the chromothripsis string.

Theorem 6.2.2. *Suppose that C is a chromothripsis string of length m derived from a reference genome G composed of n intervals. The probability that $\pi(C)$ is H/T alternating is $\frac{1}{2(m-1)}$.*

Given that Theorem 6.2.1 allows us to know exactly which chromothripsis strings will exhibit the H/T alternating property, we explore how robust this signature is when noise is present, as is expected from real data. Specifically, we add noise by randomly adding and removing a predetermined number of adjacencies from $\mathcal{A}(C)$ and determine whether or not the resulting data retained

the alternating H/T property. We find that the H/T alternating signature degrades quickly as noise is added (Figure 6.3). Specifically, the random removal of just a single edge only displayed the H/T alternating signature in less than 4% of the 10,000 simulations. We also analyzed 17 genomes formally predicted to have undergone chromothripsis/chromoplexy [154, 97, 139] and none of them exhibit the H/T alternating signature. Thus, as proposed by Korbel *et al.* [83] the H/T alternating signature may not be useful in practice as a means of measuring "the ability to walk the derivative chromosome". This conclusion is supported by the findings in a recent paper [93] which noted that this signature was not applicable to their findings of chromothripsis.

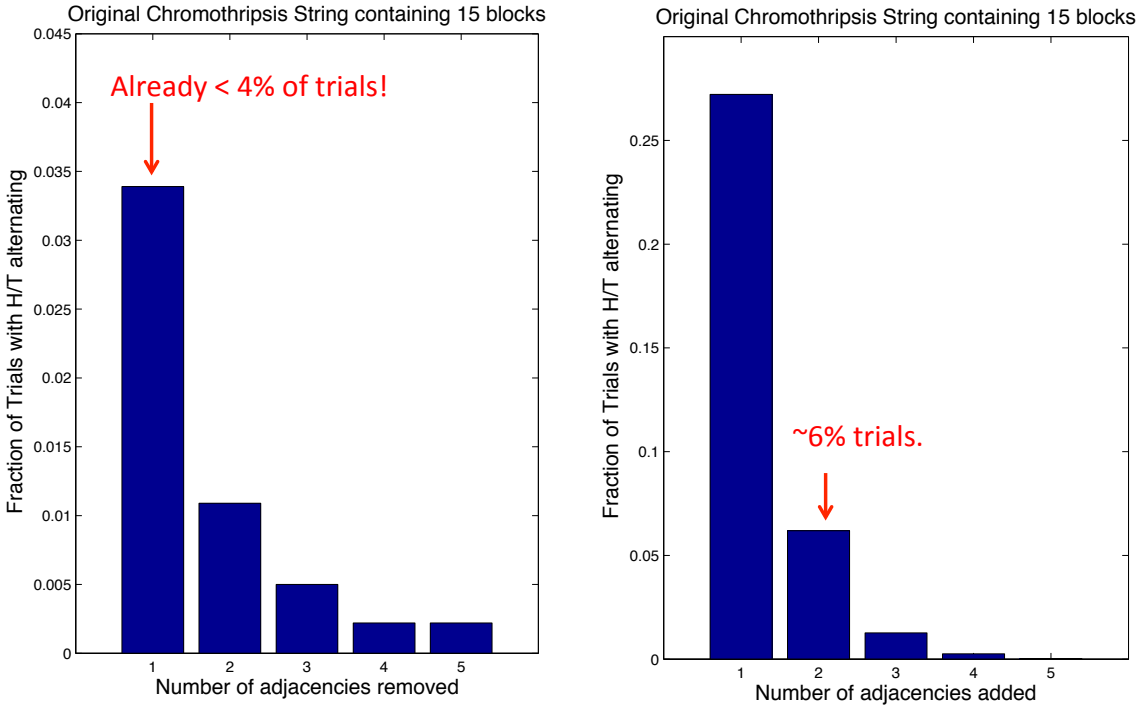


Figure 6.3: **Simulations demonstrating that the H/T alternating property degrades quickly with noise.** We created 10,000 random H/T alternating chromothripsis strings C containing 15 characters each (from a set of 50 possible characters) and randomly added and removed adjacencies from the from the corresponding adjacency set $\mathcal{A}(C)$. We found that the H/T alternating signature degraded quickly as edges were either removed or added.

6.3 Two Alternative Measures of Simultaneous Rearrangements

In the previous section we demonstrated that one of the signatures of chromothripsis proposed by Korbel *et al.* [83] may not be useful in practice. Thus, here we introduce two alternative measures of chromothripsis/chromoplexy based on the properties of the adjacencies and copy number changes that are measured by high-throughput sequencing. Since the defining characteristic of chromothripsis/chromoplexy is the simultaneity of breakpoint formation, we define the *open adjacency rate* (OAR) and *copy-number asymmetry enrichment* (CAE) in order to assess the prevalence of simultaneously formed breakpoints. In terms of the models introduced in the genome rearrangement community, genome rearrangements can be modeled as double cut and join (DCJ) operations, where two double-stranded breaks (DSBs) are introduced and repaired in an aberrant configuration [178]. Simultaneous breakage and repair at multiple sites is an operation with more than two cuts, and can be modeled as a k-break [4]. We note that in general, a k-break may be equivalent to a sequence of DCJ operations. However, under certain conditions described below an observed k-break with $k > 2$ cannot be equivalently described by a sequence of DCJ operations. Thus, chromothripsis/chromoplexy is the occurrence of one or more k-breaks with $k > 2$. The OAR and the CAE use different data as input, but both aim to provide an estimate in answer to the following question: given a genome, what proportion of the observed breakpoints were formed in k-breaks with $k > 2$?

6.3.1 Definitions and Preliminaries

We consider a *derivative genome* to be a genome that is formed from the normal, or *reference genome* through a series of k-breaks. A k-break is an operation that cuts the genome at k locations and joins the resulting free ends together [4]. k-breaks are a general purpose model for structural variation in cancer, since they formally describe a diverse set of rearrangement types including balanced rearrangements such as translocations, inversions and transpositions as well as deletions.

Formally, we define a *breakend* to be an oriented position on the genome, representing one side of a break (e.g. $x = (\text{chr17:}105227, +)$). Thus, each k-break produces $2k$ breakends, which are then joined together in an aberrant configuration in the derivative genome. Note that 2-breaks are equivalent to double cut and join (DCJ) operations [178]. Depending on how the resulting breakends are joined, a 2-break models either a translocation, an inversion, or creates a new circular

chromosome (Figure 6.4A). In the last case, if the breakends are on the same chromosome and this circular chromosome is lost, the result is a deletion of the intervening segment. Pairs of breakends that were separate before the breakage but connected after the repair (i.e. in the derivative genome) are called *adjacent*. An unordered pair $A = \{x, y\}$ of adjacent breakends is called an *adjacency*. Adjacencies are the signal left by k-breaks in the derivative genome. Pairs of breakends connected before the breakage (i.e. in the reference genome) are called *counterparts*. We denote counterpart breakends using a prime, so that if x is a breakend, x' is its counterpart. For example, a break occurring between nucleotides n and $n+1$ will generate counterpart breakends $x = (n, +)$ and $x' = (n+1, -)$.

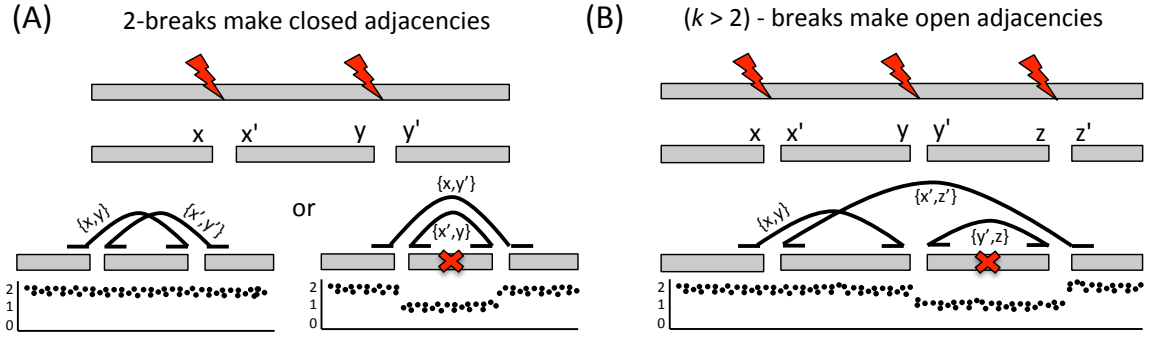


Figure 6.4: **Examples showing a 2-break and 3-break. (A)** In a 2-break, two breaks produce four breakends, organized into counterpart pairs x, x' and y, y', Aberrant repair leads to an inversion/translocation (left) with adjacencies $\{x, y\}$ and $\{x', y'\}$ or a closed loop that is then lost resulting in a deletion (red X, right). In both cases, all adjacencies are closed. This can be detected as counterpart-symmetry for the inversion adjacencies ($\{x, y\}, \{x', y'\}$, left) and copy-number symmetry for the deletion adjacency ($\{x, y'\}$, right), since due to copy number loss $\Delta(x) = \Delta(y') = -1$. **(B)** In a $(k > 2)$-break, k breaks produce $2k$ breakends which are aberrantly repaired. Closed loops formed in this process can result in deletions (red X). The resulting adjacencies are open, since for each adjacency A the counterparts of the two breakends in A are not themselves adjacent. For example, x and y are adjacent but x' and y' are not. This can be detected using counterpart-asymmetry (e.g. $\{x, y\}$, since x' is adjacent to z' but $z' \neq y'$) or copy-number asymmetry (e.g. $\{x', z'\}$, since $\Delta(x') = 0$ while $\Delta(z') = -1$).

6.3.2 Modeling Cancer Genomes with k-breaks

We model the process of genome rearrangements in cancer as follows. Each tumor begins as a non-mutated founder cell containing the reference genome. Over time, a sequence of k-breaks occur in the founder cell's lineage, eventually forming the derivative genome which is revealed at the time of sequencing. k-breaks occur according to two assumptions:

1. There is no breakpoint reuse; i.e. breaks never occur in the same location twice.

2. All breakends are fused; i.e. no new telomeres are formed. Note, the formation of new closed loops of DNA is allowed

The "no breakpoint reuse" assumption is a subtle issue in evolutionary comparisons [144, 128] where the breakends of genome rearrangements are determined as boundaries of synteny blocks from sequence alignments. These boundaries may be ambiguous due to subsequent mutations and/or repetitive sequences at the boundaries, leading to the identification of *breakpoint regions* rather than precise breakends. This lack of resolution is less of an issue in cancer data from high-throughput sequencing where we expect that any breakpoint that is detected is also localized precisely (within a few hundred nucleotides), as there has been little time for subsequent mutations to obscure this breakpoint.

6.3.3 Open and Closed Adjacencies

Let $\mathcal{A}$ be the set of all adjacencies produced by a sequence of k-breaks that transform the reference genome into a derivative genome. $\mathcal{A}$ should be thought of as a complete 'record' of all the somatic rearrangements that occurred, and not only those that can be measured in the derivative genome; i.e. $\mathcal{A}$ contains adjacencies that may be removed by subsequent deletions in the creation of the derivative genome. Chromothripsis and chromoplexy are putative rearrangement mechanisms in which many breaks occur simultaneously followed by aberrant repair of the resulting breakends, and thus is modeled as the occurrence of one or more k-breaks with large k. Under the "no breakpoint reuse" and "all breakends fused" assumptions listed above, the occurrence of a k-break with $k > 2$ will leave a specific signature in the set $\mathcal{A}$.

Let $A \in \mathcal{A}$ be an adjacency with breakends x and y. From x, we infer that at some time a DNA break occurred at x's location. This break would have produced an additional breakend x', the counterpart of x. Similarly the break at y would have generated a counterpart breakend y'. Since adjacencies (hence breakends) are never removed from $\mathcal{A}$, both x' and y' can be found in adjacencies $B, C \in \mathcal{A}$. We now ask, when does $B = C$? The answer depends on k. If A was produced by a 2-break, then no other breakends would have been present at the time, forcing x' and y' to form an adjacency (Figure 6.4A). On the other hand, if $k > 2$, then additional breakends would have been available for fusion with x' and y' (Figure 6.4B). To distinguish between these scenarios we make

the following definition.

Definition 6.3.1. *Given the set $\mathcal{A}$ of adjacencies produced by a sequence of k-breaks, $A = \{x, y\} \in \mathcal{A}$ is* closed *if $\{x', y'\} \in \mathcal{A}$; otherwise A is* open.

Every k-break generates k adjacencies. When $k = 2$, these adjacencies must be closed. Conversely, every open adjacency must have come from a k-break with $k > 2$. For a given adjacency set $\mathcal{A}$, let $\mathcal{A}^2$ the subset of adjacencies produced by 2-breaks and let $\mathcal{A}^k$ be the subset produced by $(k > 2)$-breaks, so that $\mathcal{A} = \mathcal{A}^2 \cup \mathcal{A}^k$. Let $\mathcal{O}(\mathcal{A})$ be the set of open adjacencies in $\mathcal{A}$. We have the following.

Observation 6.3.1. *For every adjacency set $\mathcal{A}$, $\mathcal{O}(\mathcal{A}) \subset \mathcal{A}^k$.*

Two Signatures of Open Adjacencies

Our goal is to detect chromothripsis/chromoplexy by inferring the history of k-breaks that gave rise to an observed set of adjacencies and copy number aberrations. In particular, we are interested deriving a lower bound for the number of adjacencies produced in k-breaks with $k > 2$. As described above, this can be accomplished by counting open adjacencies. However, "open" and "closed" are theoretical categories, describing the etiology of an adjacency, rather than its structure in the derivative genome. In particular, subsequent rearrangements or experimental error may obscure whether adjacencies are open or closed. Thus we need to define signatures of open adjacencies that can be robustly applied to real data. We define two such signatures below: (1) *counterpart-asymmetry*; and (2) *copy-number asymmetry.*

Let $\mathcal{A}$ be the complete set of adjacencies produced by a sequence of k-breaks and $\tilde{\mathcal{A}} \subseteq \mathcal{A}$ be the subset observed in genome sequencing data from the derivative genome. Consider an adjacency $A = \{x, y\} \in \tilde{\mathcal{A}}$. If A is open then the counterpart breakends x' and y' must belong to separate adjacencies in $\mathcal{A}$, say $\{x', w\}$ and $\{y', z\}$ where $w \neq y'$ and $z \neq x'$. Based on the assumption of no breakpoint reuse, observing either $\{x', w\}$ or $\{y', z\}$ in the derivative genome precludes the existence of $\{x', y'\}$, and demonstrates that A is open. We call this signature *counterpart-asymmetry* (Figure 6.4B).

Definition 6.3.2. *Given a set $\tilde{\mathcal{A}}$ of experimentally detected adjacencies, $A = \{x, y\} \in \tilde{\mathcal{A}}$ has* counterpart-asymmetry *if there exists a breakend w such that $w \neq y'$ and $\{x', w\} \in \tilde{\mathcal{A}}$ or there exists a breakend z such that $z \neq x'$ and $\{y', z\} \in \tilde{\mathcal{A}}$.*

The second signature of open adjacencies relies on copy number. We represent copy number as an integer-valued function $\mathcal{N}$ on genomic coordinates. Assuming the k-break model of rearrangement, each discontinuity in $\mathcal{N}$ occurs at a site of breakage and results in a distinct copy number state over each breakend in a counterpart pair. Thus, if a break between nucleotides n and $n+1$ produces a pair of counterpart breakends: $x = (n, +)$ and $x' = (n+1, -)$, $\mathcal{N}(x)$ represents the absolutel copy number state immediately upstream of the break and $\mathcal{N}(x')$ the copy number downstream. In addition to the absolute copy number at a breakend, we wish to characterize the change in copy number change *across* a breakend. Thus, we define $\Delta(x) := \mathcal{N}(x') - \mathcal{N}(x)$ where x' is the counterpart of x.

In this formulation, breakends flanking deleted regions have negative Δ values. For example, suppose the adjacency $A = \{x, y\}$ resulted from a heterozygous (single copy) deletion. Then x', the counterpart of x, must lie within the deleted region, meaning $\mathcal{N}(x') = \mathcal{N}(x) - 1 \implies \Delta(x) = -1$. A similar argument implies that $\Delta(y) = -1$. In this case, the changes in copy number are symmetric at the two breakends of the adjacency. Alternatively, an adjacency $A = \{x, y\}$ may exhibit different copy number changes across both its breakends. Such an occurrence is our second signature of open adjacencies called *copy-number asymmetry*, which we define as follows.

Definition 6.3.3. *Given a set $\tilde{\mathcal{A}}$ of experimentally detected adjacencies, $A = \{x, y\} \in \tilde{\mathcal{A}}$ has* copy-number asymmetry *provided* $\Delta(x) \neq \Delta(y)$.

It is not immediately clear that an adjacency with copy-number asymmetry is necessarily an open adjacency, so we prove the following.

Proposition 1. *If an adjacency A has copy-number asymmetry, then it is open.*

Proof. Suppose that $A = \{x, y\}$ is a closed adjacency formed by a k-break at some time t_0. This means that the pairs of breakends $\{x, x'\}$ and $\{y, y'\}$ were connected before time t_0, and the pairs $\{x, y\}, \{x', y'\}$ are connected after time t_0. Since we assume there is no breakpoint reuse, x and y must have been 'untouched' before time t_0. Thus, $\mathcal{N}(x) = \mathcal{N}(x')$ and $\mathcal{N}(y) = \mathcal{N}(y')$ before t_0. After t_0, these counterpart breakend pairs are no longer fused, meaning their copy numbers can change independently. However, the newly adjacent breakend pairs are now 'locked' to each other and their copy numbers must rise and fall together. For example, once x and y are adjacent, a copy number decrease over x implies a copy number decrease over y. Indeed their copy numbers could only change differentially if they were re-broken, violating the assumption that breakpoints are not reused. This

implies that in the derivative genome, $\mathcal{N}(x') - \mathcal{N}(x) = \mathcal{N}(y') - \mathcal{N}(y) \implies \Delta(x) = \Delta(y)$, which means closed adjacencies cannot be copy-asymmetric. Conversely, adjacencies with copy-number asymmetry must be open. □

We emphasize here the importance of analyzing the differences $\Delta(x)$ and $\Delta(y)$ in copy number *across* breakends to define copy-number asymmetry rather than absolute copy numbers $\mathcal{N}(x)$ and $\mathcal{N}(y)$ *at* breakends. $\mathcal{N}(x)$ and $\mathcal{N}(y)$ can be unequal even when the adjacency $\{x, y\}$ is closed, as a change in copy number for either breakend may have occurred *prior* to the formation of the adjacency $\{x, y\}$. After the formation of the adjacency $\{x, y\}$, however, copy number changes that affect x must also apply to y since the two breakends are fused. Thus assuming $\{x, y\}$ is closed, we expect to find $\Delta(x) = \Delta(y)$ even when $\mathcal{N}(x) \neq \mathcal{N}(y)$. Critically, this argument rests on our assumption that there is no breakpoint reuse, since a second break at x or y (on the originally rearranged chromosome or its homologue) would allow $\Delta(x)$ and $\Delta(y)$ to vary independently.

Detecting a Range of Open Adjacencies.

Since counterpart-asymmetry relies on the presence of counterpart breakends and copy-number asymmetry implicitly relies on their absence, the two signatures in combination can identify a broader set of open adjacencies than each can on its own. This is illustrated in the following two examples.

First, let $A = \{x, y\}$ be a closed adjacency. This implies that x' and y' were fused in the k-break that created A. Clearly, observing the adjacency $\{x', y'\}$ in the derivative genome would demonstrate that A is closed, but what if $\{x', y'\}$ is not observed? There are two possible explanations: either $\{x', y'\}$ exists in the derivative genome but was not detected, or the genomic segment containing $\{x', y'\}$ is deleted. In the latter case, the deletion would have a occurred at the same time as the creation of A (i.e. A was created by a deletion) or subsequent to the creation of A. Since the deletion of an adjacency entails a copy number drop at its constituent breakends and our no breakpoint reuse assumption implies that any subsequent copy number changes would produce coordinated copy number changes across x and y, we have that $\Delta(x) = \Delta(y)$. Thus, A would show counterpart symmetry if $\{x', y'\}$ were retained and copy-number symmetry if it were deleted. In either case, A will be considered a closed adjacency according to our definitions (Figure 6.4A).

Next, suppose $A = \{x, y\}$ is an open adjacency. This means that the counterpart x' was fused to a breakend $w \neq y'$, producing an adjacency $\{x', w\}$. Observing the adjacency $\{x', w\}$ in the

derivative genome would demonstrate that A is open through counterpart-asymmetry. On the other hand, if the DNA supporting $\{x', w\}$ were deleted, then there would be a copy number change at x ($\Delta(x) \neq 0$). Since y' is not adjacent to x' or w, it is unlikely that y' is also deleted at the same time. If we also assume that y' does not experience an independent change in copy number at another time, then we have $\Delta(y) = 0$. Under these conditions $\Delta(x) \neq \Delta(y)$, giving A copy-number asymmetry. Therefore, A would look open to our signatures if either $\{x', w\}$ were retained and measured or if $\{x', w\}$ were deleted and y' were retained (Figure 6.4B).

6.3.4 Open Adjacency Rate (OAR)

Given a collection of measured adjacencies $\tilde{\mathcal{A}}$ and a copy number profile $\mathcal{N}$, we identify the adjacencies that exhibit counterpart-asymmetry or copy-number asymmetry and form a putative set of open adjacencies $\mathcal{O} \subset \tilde{\mathcal{A}}$. Note that $\tilde{\mathcal{A}}$ may represent all measured adjacencies, or a subset of adjacencies that suspected to reflect a chromothripsis-like or chromoplexy-like event. To estimate the proportion of adjacencies in $\tilde{\mathcal{A}}$ formed by ($k > 2$)-breaks, we define the *open adjacency rate* (OAR)

$$\text{OAR}(\tilde{\mathcal{A}}, \mathcal{N}) =: \frac{|\mathcal{O}|}{|\tilde{\mathcal{A}}|}. \tag{6.2}$$

In real data, not all open adjacencies will display copy-number asymmetry or counterpart-asymmetry. For example, if only a sparse set of adjacencies is detected, then counterparts will be rare. However, those adjacencies which do show either signature can be called open with high-confidence. Hence the total number of adjacencies exhibiting counterpart/copy-number asymmetry bounds the true number of open adjacencies from below. Thus, if there is no experimental error generating false-positive open adjacencies then it follows from Observation 1 that $OAR(\tilde{\mathcal{A}}, \mathcal{N}) < |\tilde{\mathcal{A}}^k|/|\tilde{\mathcal{A}}|$.

6.3.5 Copy-number Asymmetry Enrichment (CAE)

For two breakends to be considered counterparts, they must satisfy several criteria, including that they lie close together on the genome. Therefore, in regions that exhibit a dense clustering of breakends it can become difficult to disambiguate breakends that are close because they are counterparts from those that are close due to other factors. Thus, adjacencies which are densely clustered may occasionally appear open due to false positive counterpart breakend calls, artificially enhancing the open adjacency rate. Since adjacency sets representing putative chromothripsis/chromoplexy events

are often formed on the basis of breakend clustering [83], it is desirable to develop a measure which ignores the relative positions of breakends and allows one to separate the contribution of breakend clustering from other factors when assessing whether the given adjacencies were formed during a one-off event. We introduce a second measure, *copy-number asymmetry enrichment* (CAE), that imputes the open adjacency rate using only relative copy number changes at adjacent breakends.

Consider an adjacency set $\tilde{\mathcal{A}}$ produced by k-breaks with $k \geq 2$. Let $\tilde{\mathcal{A}}^2$ be the set of adjacencies from 2-breaks and $\tilde{\mathcal{A}}^k$ be the set from $(k > 2)$-breaks, so that $\tilde{\mathcal{A}} = \tilde{\mathcal{A}}^2 \cup \tilde{\mathcal{A}}^k$. Further, let $\mathcal{C} \subseteq \tilde{\mathcal{A}}$ denote the subset of copy-number asymmetric adjacencies. We wish to estimate the fraction of adjacencies in $\tilde{\mathcal{A}}$ that came from $(k > 2)$-breaks using copy-number asymmetry alone; i.e. to estimate $|\tilde{\mathcal{A}}^k|/|\tilde{\mathcal{A}}|$ from $|\mathcal{C}|$. Proposition 1 tells us that $|\mathcal{C}| \leq |\tilde{\mathcal{A}}^k|$. Turning this lower bound into a direct estimate requires quantifying the degree to which $|\tilde{\mathcal{A}}^k|$ exceeds $|\mathcal{C}|$. This depends critically on the fraction of breakends in $\tilde{\mathcal{A}}$ that co-locate with changes in copy number.

Let p_Δ be the fraction of breakends x in $\tilde{\mathcal{A}}$ such that $\Delta(x) \neq 0$ (i.e. the fraction of breakends co-locating with a change in copy number). To derive an expected relationship between $|\mathcal{C}|$, p_Δ and $|\tilde{\mathcal{A}}^k|$, we treat the copy number changes $\Delta(x)$ as random variables and make the following assumptions: (1) For each breakend x, $\Delta(x)$ is always -1 or 0 (deletion or non-deletion); (2) For each adjacency $\{x, y\} \in \tilde{\mathcal{A}}^2$, $\Delta(x)$ and $\Delta(y)$ are equal (dependent) and Bernoulli distributed with $P(\Delta(x) = \Delta(y) \neq 0) = p_\Delta$; (3) For each adjacency $\{x, y\} \in \tilde{\mathcal{A}}^k$, $\Delta(x)$ and $\Delta(y)$ are independent and Bernoulli distributed with $P(\Delta(x) \neq 0) = P(\Delta(y) \neq 0) = p_\Delta$. It follows from these assumptions that $\tilde{\mathcal{A}}^2 \cap \mathcal{C} = \emptyset$ and that for an adjacency $\{x, y\} \in \tilde{\mathcal{A}}^k$ chosen uniformly at random, $P(\{x, y\} \in \mathcal{C}) = P(\Delta(x) \neq \Delta(y)) = P(\Delta(x) = 0, \Delta(y) = -1) + P(\Delta(x) = -1, \Delta(y) = 0) = 2p_\Delta(1 - p_\Delta)$. It follows that $\mathbb{E}(|\mathcal{C}|) = 2p_\Delta(1 - p_\Delta)|\tilde{\mathcal{A}}^k|$, allowing us to approximate $|\tilde{\mathcal{A}}^k| \approx |\mathcal{C}|/(2p_\Delta(1 - p_\Delta))$. Thus, we can estimate $(|\tilde{\mathcal{A}}^k|/|\tilde{\mathcal{A}}|)$, the fraction of $(k > 2)$-breaks, by the *copy-number asymmetry enrichment* (CAE) ratio, defined as

$$\text{CAE}(\tilde{\mathcal{A}}) := \frac{|\mathcal{C}|}{2p_\Delta(1 - p_\Delta)|\tilde{\mathcal{A}}|}. \tag{6.3}$$

6.3.6 Application of OAR and CAE to Real Data

Detecting open adjacencies in real sequencing data requires: (1) a set $\tilde{\mathcal{A}}$ of measured adjacencies along with an annotation of the corresponding breakends for membership in counterpart pairs; (2) a copy number profile $\mathcal{N}$ across the genome that maps copy number changes to breakends. The procedures we use to collect this data are described below.

We assume that a collection of rearrangements, or structural aberrations, has been identified in the derivative genome by analyzing paired-read or split read data using one of the many algorithms for this purpose [13, 151, 138]. The output of these algorithms is a collection $\mathcal{V}$ of pairs of breakends $\{x, y\}$ representing novel adjacencies in the derivative genome, where x and y are oriented genomic coordinates in the reference genome. We form the the adjacency set $\tilde{\mathcal{A}}$ from $\mathcal{V}$ by identifying counterpart breakend pairs $\{x, x'\}$ such that $x, x' \in \mathcal{V}$, $x \leq x'$, and the following criteria are satisfied: (1) $x' - x \leq D$ for a small integer D; (2) x has positive orientation and x' has negative orientation; i.e. the pair (x, x') has convergent $(+, -)$ orientation; (3) $\{x, x'\} \notin \mathcal{V}$; (4) no other breakends in $\mathcal{V}$ lie between x and x'. In principle, counterpart breakends occupy adjacent nucleotides, so that we expect $x' - x = 1$, indicating a distance threshold of $D = 1$ in criterion (1) above. However, higher values of D may be used in practice since many structural aberration algorithms do not identify breakends to single nucleotide resolution. In addition, counterpart breakends may be separated by a small distance due to microdeletions or "deletion bridges" [11] that occur at rearrangement breakpoints.

One may compute the OAR on the full set of novel adjacencies; i.e. build $\tilde{\mathcal{A}}$ from $\mathcal{V}$. Alternatively, one may evaluate a subset of detected adjacencies, for example a spatially clustered set of adjacencies or a collection previously implicated as representing a chromothripsis-like event, by building $\tilde{\mathcal{A}}$ from a subset of $\mathcal{V}$. We use the later approach in our analyses below.

To create a copy profile $\mathcal{N}$ which maps changes in copy to breakends, we analyze a whole-genome segmentation as follows. First, we match the ends of copy number segments (indicating a change in copy number) to nearby breakends. This is done by creating a breakpoint interval I with length L around the boundary of each copy number segment For each breakend x and breakpoint interval I, we declare a match if: (1) x lies within I; (2) x is the only breakend occupying this interval. Since determination of absolute copy number in tumors is challenging due to heterogeneity [118], we assign change in copy values Δ to breakends using a step function: $\Delta(x) = 1$ for breakends matched to intervals indicating positive copy change; $\Delta(x) = -1$ for breakends matched to intervals indicating negative copy change; $\Delta(x) = 0$ for breakends without a matched copy change.

Results on Real Data

In [170] we compute the OAR and CAE on 121 cancer genomes from two datasets that were previously screened for chromothripsis/chromoplexy [97, 11]. We refer the reader to [170] for the

complete results. In short, we find that both measures correlate well with the predicted classifications of chromothripsis/chromoplexy versus sequential ($p < 10^{-3}$ on data from [97] and $r = 0.73$ on data from [11]), but differ on a small subset of genomes. Visual inspection of the genomes for which OAR makes differing predictions suggest that they have been mis-classified in the published analyses. Appendix D also contains preliminary experimental results on real data that motivated the work appearing in [170].

6.4 Discussion

The definition of rigorous criteria to distinguish events such chromothripsis/chromoplexy which result in the simultaneous acquisition of multiple rearrangements from the stepwise accumulation of rearrangements using DNA sequencing data from a single time point is challenging task [154, 11, 83, 97, 55, 81, 91]. We first provide a rigorous model of chromothipsis which we use to analyze one signature of chromothipsis proposed by [83]. We use this analysis to demonstrate that noise in sequencing data may quickly degrade the signal associated with this signature. This motivates the need for measures of simultaneous events which are more robust to noise. Thus, we introduced two measures, the open adjacency rate (OAR) and copy-number asymmetry enrichment (CAE), to quantify the occurrence of simultaneous rearrangements, or k-breaks [4] with $k > 2$, in the formation of a derivative genome. We showed that the OAR and CAE measures correlate well with previously published analyses [97, 11] of chromothripsis/chromoplexy, but that our measures also reveal some potential misclassifications in these studies.

While our results demonstrate that the OAR and CAE are useful measures, they both have limitations. The OAR and CAE are *local* measures that estimate the proportion of ($k > 2$)-break adjacencies by considering each adjacency in turn, rather than examining their global configuration. While some information is lost in this approach, robustness to experimental error is gained. Indeed, measures of chromothripsis/chromoplexy that rely solely on the global configuration, such as ChainFinder [11] may be affected by a single missing adjacency. Combining information from global configurations with local measures such as the OAR is therefore an important area for future investigation. In addition, recent studies suggest that chromothripsis/chromoplexy events do not occur in isolation [91]. Thus, flexible measures, such as the OAR and CAE, may be better able to distinguish the available signal of a one-time event from the noise of sequential rearrangements in

the same region.

The ability to detect chromothripsis/chromoplexy using OAR, CAE, or related measures is impacted by the extent of intra-tumor heterogeneity within a sample. If a chromothripsis/chromoplexy event exists in only a fraction of cells in the sample, then the power to detect the adjacencies and copy number changes that characterize this event is diminished. Recently developed methods to characterize intra-tumor heterogeneity within a single sample [118, 143, 63] or new single cell sequencing approaches [169], may provide better data for measures such as OAR.

Ultimately, *in vivo* or *in vitro* studies of chromothripsis/chromoplexy are necessary to further quantify the causes and prevalence of these events. In the interim, analytical methods to predict k-breaks from high-throughput sequencing data will remain useful tools, with the caveat that for some samples such *post hoc* analysis may insufficient to determine reliability whether a chromothripsis/chromplexy event occurred.

Chapter 7

Conclusions

"More people each year die of cancer in the United States than all the Americans who lost their lives in World War II. This shows us what is at stake. It tells us why I sent a message to the Congress the first of this year, which provided for a national commitment for the conquest of cancer, to attempt to find a cure."

– President Richard M. Nixon, December 23, 1971

It has been over 40 years since President Richard Nixon declared a "War on Cancer" by signing the National Cancer Act of 1971. In that time, increased awareness and funding opportunities have led to extensive research investigating both causes of and treatments to cancer. As a result, we have taken important steps forward in our understanding and approach to treatment of some types of cancer. For example, the advent of the drug Gleevec, introduced in 2001, has helped to increase the 5 year survival rate for patients with Chronic Myeloid Leukemia (CML) from 31% to 59% [73]. Despite the progress that has been made, we are still a long way from the ultimate goal of curing cancer.

The advent of new high-throughput DNA sequencing technologies in recent years have ushered in a new era of cancer research. The prospect of sequencing a tumor sample from a patient recently diagnosed with cancer would hardly have been imaginable just 10 years ago, but is quickly becoming reality. The ability to probe and measure the human genome in this manner opens up incredible

potential for better understanding of cancer, but more importantly provides the initial step needed to embark down the path of personalized medicine. To fully realize this ultimate goal of personalized treatment will require continued advancement not only in the technology for sequencing genomes, but also in the algorithms used to analyze that data [125].

There are still many challenges related to analyzing high-throughput DNA sequence data. The basic task of identifying the set of somatic mutations in tumor, a necessary prerequisite for personalized medicine to become a reality, is still plagued by many challenges. Some of these challenges are specific to the DNA sequencing technology and others are a result of genomic anomalies specific to cancer. New methods developed to handle to intricacies of DNA sequence data of cancer genomes have incredible potential to change our fundamental understanding of cancer.

7.1 Summary of Contributions

This dissertation focuses on the design of algorithmic methods that address challenges specific to analyzing DNA sequence data of cancer genomes; in particular intra-tumor heterogeneity and complex genomic rearrangements.

7.1.1 Intra-Tumor Heterogeneity

In Chapter 2 we present Tumor Heterogeneity Analysis (THetA), an algorithm to infer the collection of tumor genomes that differ by copy number aberrations and their prevalences within a single tumor sample [118]. THetA addressed several important limitations of previous methods for this task. First, THetA utilizes a probabilistic model of DNA sequencing data (as opposed to previous methods which were designed for a different technology). Second, THetA allows *any number* of tumor populations to be inferred as opposed to previous methods which were only able to infer a single tumor population. Finally, we show that THetA is an efficient algorithm in the important case when a tumor contains a single tumor population along with admixture with normal cells.

In Chapter 3 we present THetA2 which extends the original THetA algorithm in a number of important directions [121]. This includes a new optimization procedure which reduces the runtime of the algorithm when considering multiple tumor populations by a factor of >1000X. We also formulate a probabilistic model of B-allele frequencies (an alternative type of data not used in the original THetA algorithm) that allows THetA2 to distinguish between solutions which are equally

likely using read depth information alone. This improvement helps to address an identifiability issue with the original THetA. Lastly, we extend THetA2 to work with whole-exome sequencing data. These improvements greatly increase the breadth of tumors that may be analyzed with THetA2 – thus making it a more attractive program for biologists to use when analyzing sequencing data.

In Chapter 4 we formalize the problem of reconstructing the clonal evolution of a tumor using single-nucleotide mutations as the Variant Allele Frequency Factorization Problem (VAFFP) and derive a combinatorial characterization of the solutions to this problem [47]. We also describe an integer linear programming solution to the VAFFP in the case of error-free data and extend this solution to real data with a probabilistic model for errors. We call the resulting algorithm presented here AncesTree.

7.1.2 Complex Genomic Rearrangements

In Chapter 5 we present the Paired-End Reconstruction of Genome Organization (PREGO) algorithm which uses complementary data signals (read depth and adjacencies) to infer the most likely collection of rearrangements that together best describe the cancer genome [120]. We demonstrate that PREGO is a polynomial time algorithm through reduction to a network flow problem on a bi-directed graph. We obtain biologically meaningful results when applying PREGO to both breast and ovarian cancer datasets.

In Chapter 6 we address the question of whether it can be determined from sequencing data if a set of observed rearrangements occurred sequentially overtime or part of a one-time catastrophic event such as Chromothripsis [154] or Chromoplexy [11]. We first define a rigorous mathematical model of chromothripsis on strings. We then use this model to critique one of the signatures of chromothripsis as suggested by Korbel *et al.* [83]. We then introduce two new measures on the fraction of rearrangements that occurred simultaneously (as is predicted by Chromothripsis/Chromoplexy) called the Open Adjacency Rate (OAR) and the Copy-number Asymmetry Enrichment (CAE) [170].

7.2 Future Work

If the past is a valid predictor of the future, DNA sequencing technologies will continue to evolve quickly. Concurrent with these changing technologies will be new algorithmic challenges. Thus, there is will be a continuing need for the development of novel computational algorithms into the

foreseeable future. While it is impossible to predict how the available technologies will change in the coming years, we can make use of past trends to help guide our visions of what may come. Furthermore, the motivating factors behind necessary algorithmic development is not only driven by the type of available data, but also it's scale and scope. Given the ongoing efforts of large scale collaborations such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), it seems likely that in the near future the amount of available data will continue to increase in terms of raw numbers and types of cancers.

The work presented in this dissertation is aimed at addressing challenges related to analyzing DNA sequence data from cancer samples. Specifically, we have addressed challenges related to intra-tumor heterogeneity and complex genomic rearrangements. These have been extremely active research areas, especially intra-tumor heterogeneity, in recent years. While the algorithms presented in this dissertation, and the numerous others that have been published in recent years, have helped to scratch the surface on our understanding of the complex facets of cancer, we still have a long way to go. We now discuss possible future directions around the areas of intra-tumor heterogeneity, complex genomic rearrangements and other related directions.

7.2.1 Intra-Tumor Heterogeneity

Despite the volume of recent methods aimed at handling and detecting intra-tumor heterogeneity from a single DNA sequenced sample [105, 143, 8, 12, 63, 91, 92, 48], including THetA presented in Chapters 2 and THetA2 presented in Chapter 3, there are a number of challenges that remain in this space. The integration of multiple data signals, including but not limited to somatic SNVs, germline SNPs, CNAs, and other structural variants such as inversions or translocations, may be advantageous when inferring tumor composition. While some methods have begun to integrate a subset of these data signals [48, 91, 12], it is still unclear how to weight the contribution of the different signals and if all can be efficiently combined. Related, is the challenge of how to proceed if the different signals actually provide contradictory evidence. Appropriate integration of these data signals, and perhaps others derived from alternative data sources such as RNA-Seq or Methylation data can only improve our ability to detect intra-tumor heterogeneity.

Another challenge that still need to be more thoroughly addressed is how to accurately and automatically identify and incorporate genome-wide events, such as whole-genome doubling when

inferring tumor composition. While the method ABSOLUTE [27] does detect whole-genome duplication events, it is intended to be used with a trained analyst annotating and selecting the final results. Many methods, including THetA [118] and THetA2 [121], rely on the assumption that much of the genome maintains the normal diploid state, thus making prediction of whole-genome events difficult. This is one challenge where the incorporation of additional data signals may in fact make inference of such whole genome events more plausible.

Most methods aimed at predicting the evolutionary history of a heterogeneous tumor sample [75, 155, 64, 39] including the AncesTree algorithm presented in Chapter 4 are constrained by a number of rigid assumptions. These include the use of only SNV data in analysis and the assumption that all observed SNVs occur in copy neutral regions. The one exception is PhyloWGS [39] which can be given pre-computed copy number estimates. However, this only sidesteps the task of being able to simultaneously infer both SNVs and CNAs in the context of tumor evolution. Furthermore, The the rigid assumption that observed SNVs occur in copy neutral regions is likely not valid for many tumors, especially for those cancer types such as breast or ovarian cancer [50, 24] where copy number aberrations are extremely common. Thus, there is a need for further exploration into how copy number aberrations can be incorporated into such methods. Such exploration would likely need to define assumptions, similar to the infinite sites assumption for SNVs, on how copy number aberrations evolve over time.

Lastly, one ongoing challenge that affects all current and future methods is how to perform validation. On real data the true tumor composition or evolutionary history is not known, thus validation of predictions is a tricky task. Many approaches utilize results obtained on simulated data as a means of verifying the accuracy of the method. However, there is extreme variability in how such simulated datasets are created. A few large scale simulated datasets are now becoming available through challenges related to benchmarking existing algorithms [72]. While this may partially help by providing standardized datasets for comparison, there are other issues that need to be addressed. For instance, it is difficult to create simulated datasets that accurately reflect the intricacies of real datasets. In fact, it is unclear if the datasets created for these benchmarking challenges are truly indicative of the amount and quality of noise contained within real data. Thus, there is a need for standardized and realistic comparison datasets that can be used to analyze and compare methods.

7.2.2 Complex Genomic Rearrangements

The ability to detect and analyze complex genomic rearrangements is closely tied to available DNA sequencing technologies. For instance the PREGO method presented in Chapter 5 relies on using a set of rearrangements detected using any structural variation prediction method (e.g. [151, 133, 139], etc.). Unfortunately, most such methods have high false positive rates and produce variable results on the same datasets [1, 122]. However, as read lengths continue to grow, the ability to detect novel rearrangements, especially in repetitive regions, by clustering discordant pairs of reads should improve. This will certainly help with downstream prediction of complex genomic rearrangements. Of course as read lengths grow, the probability that an individual read contains information about multiple aberrations increases. Thus, methods that are able to handle multiple breaks will be needed [140].

Furthermore, methods for analyzing and interpreting complex genomic rearrangements including PREGO presented in Chapter 5, OAR and CAE presented in Chapter 6 and various others [99, 55, 11] do not utilize information about intra-tumor heterogeneity. This represents just part of a larger issue where information about intra-tumor heterogeneity ought to be included when analyzing DNA sequence data that likely represents a collection of different tumor genomes. The omission of information such as tumor composition from many algorithms is a multi-faceted issue. In some instances this is because it has only been in the past few years that methods for inferring intra-tumor have really started to appear. In others, this omission may be the result of simplifying assumptions made about the underlying biology. There is certainly a gap between the theory underlying many algorithmic methods and their true usefulness in practice. Figuring out how to better close this gap, or to identifying what information is truly lost when simplifying assumptions are made, is an important area of future work.

Finally, the question of whether simultaneous events truly occur *in vivo* is still an open debate. In the absence of irrefutable evidence that such events are real, there will be a continued need for rigorous analytical methods to asses what signal may be left by such events in DNA sequence data and if that signal can be confidentially distinguished from noise. The OAR and CAE presented in Chapter 6 provide a partial glimpse of such analysis, but further work is needed.

7.2.3 Other Related Directions

Recent advances in high-throughput DNA sequencing have revolutionized our ability to measure genomes. These advancements have led to falling sequencing costs and as a result, large-scale efforts such as the 1000 Genomes project, The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC), etc. have been producing an ever growing plethora of publicly available data. With sequencing data becoming available for large cohorts of patients, there are questions about tumors and their development that only now can be posed (and potentially answered). For example, are there particular molecular pathways (sets of genes) that are more likely to be mutated in clonal tumor subpopulations (existing in all tumor cells) or in a subclonal tumor subpopulation (existing in only a subset of tumor cells)? Such information may yield further insight into the order and progression of pathways mutated in cancer. The availability of datasets such as those through the TCGA or ICGC will make such analysis possible. One algorithmic challenge with this goal is to develop a method that does not rely on previously characterized pathways but instead, allows for the discovery of novel pathways.

Parts of this dissertation have focused on the problem of intra-tumor heterogeneity, where a tumor sample contains a mixture of different cancer genomes. There are a number of other contexts where a sequenced sample may contain a mixture of distinct genomes and it may be useful to be able to deconvolve the mixture into its constituent components. One example of a context where a mixture of DNA exists is blood samples of a patient with cancer. Recent studies have shown that when a tumor cell dies, it may ultimately release fragments of its genomic content into the blood stream, and this may happen at such a rate as to be detectable. Thus, sequencing data from a blood sample of a cancer patient may ultimately contain a mixture of tumor DNA along with normal (healthy) DNA [18]. Another specific example where a genomic mixture is measured, and deconvolution of the mixture is desirable, is *metagenomics*. Here, an environmental sample contains a mixture of multiple distinct, but potentially related, species [173]. Both of these examples have their own unique sets of challenges (number of components in the mixtures, relative similarity of mixture components, etc.) that need to be addressed in order to deconvolve the mixed DNA sample. Overall, it may be useful to study the distinct challenges associated with identifying the composition of mixed genomic samples in a generic setting or to see if the lessons learned in these different contexts can be shared across the domains.

Appendix A

Quantifying Intra-Tumor Heterogeneity

A.1 Additional Algorithmic Details

In this section we present additional algorithmic details related to the THetA algorithm not included in Chapter 2.

A.1.1 Motivating the Multinomial Model

The multinomial model that we use in our likelihood function does not assume that the observed read depths in different intervals are independent. Even though we assume that reads are distributed uniformly on the cancer genome, large copy number aberrations (e.g. gain and loss of whole chromosomes) will cause the observed number of aligned reads in an interval I_j to deviate from expected *even* when the interval I_j itself is not affected by a copy number aberration. The reason is because the number of reads is fixed; thus, for example the lack of reads aligning to one part of the genome due to a deletion will mean that there will be more reads observed from the non-deleted parts of the genome. We see shifts in read depth ratios between tumor and normal samples due to the change in the length of the tumor genome in both simulation and real data (Figures A.1 and A.2).

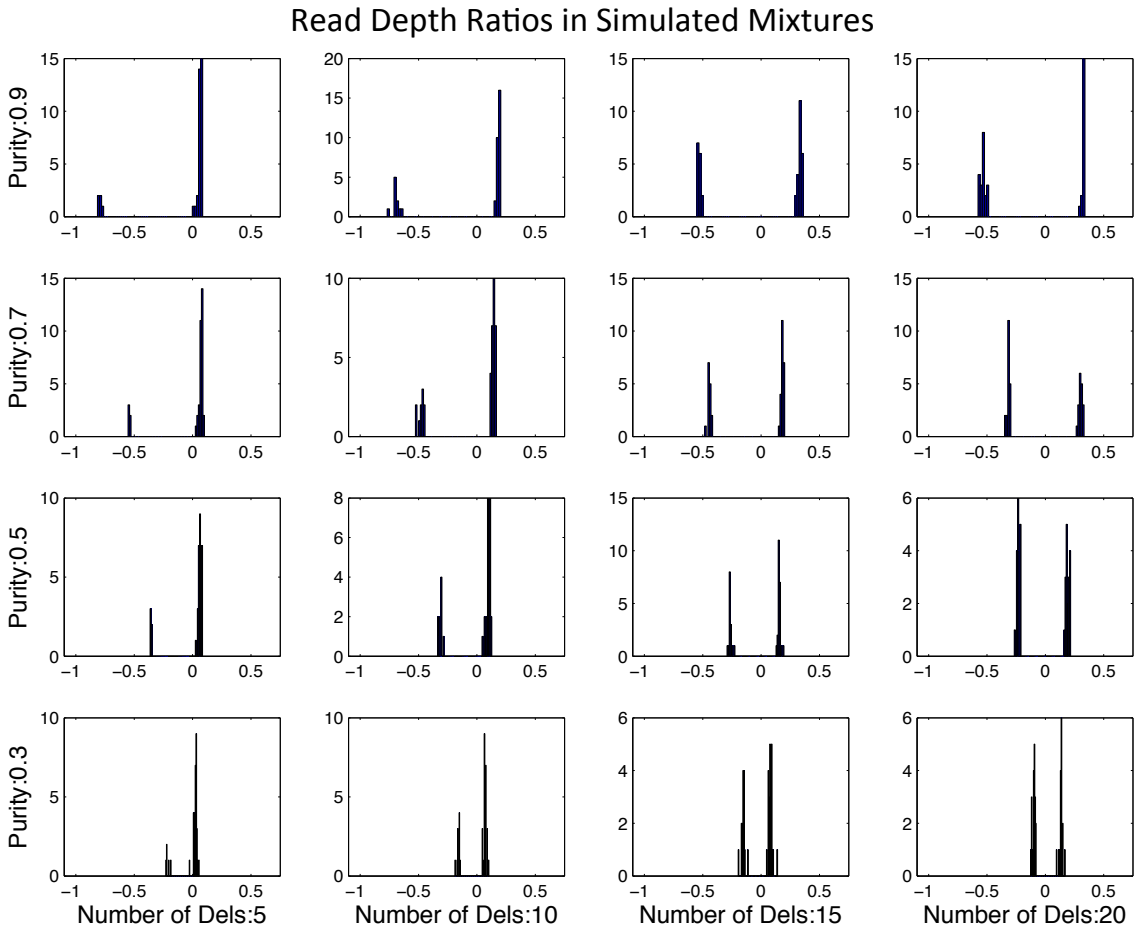


Figure A.1: **Simulations demonstrating that large deletions can affect read depth across the entire genome.**In simulated data containing a mixture of normal cells and one tumor population with a specified number of chromosome arm deletions we observe shifts in read depth ratios across the entire genome. As more of the genome becomes deleted, the ratios (even for regions of normal copy) shift to the right. As the sample purity decreases, all ratio peaks become compressed together. For many of these mixtures simply rounding read depth ratios to obtain integer copy numbers will result in errors.

A.1.2 Derivation of Equations Used by ASCAT and ABSOLUTE

In this section we show how Equation 1 from [27] and the log term in Equation 1 from [164] can be written as a function of expected values of observations in our probabilistic model. Let $\mathbf{I} = (I_1, \ldots, I_m)$ be a partition of the reference genome into m intervals. Using our notation, these equations can be written directly as $\frac{2\mu_1 + c_{j2}\mu_2}{2\mu_1 + \rho\mu_2}$ where ρ is average ploidy in the cancer genome. Suppose we sequence a tumor sample $\mathcal{T}$ with P reads. Let X_{pq} be a random variable such that

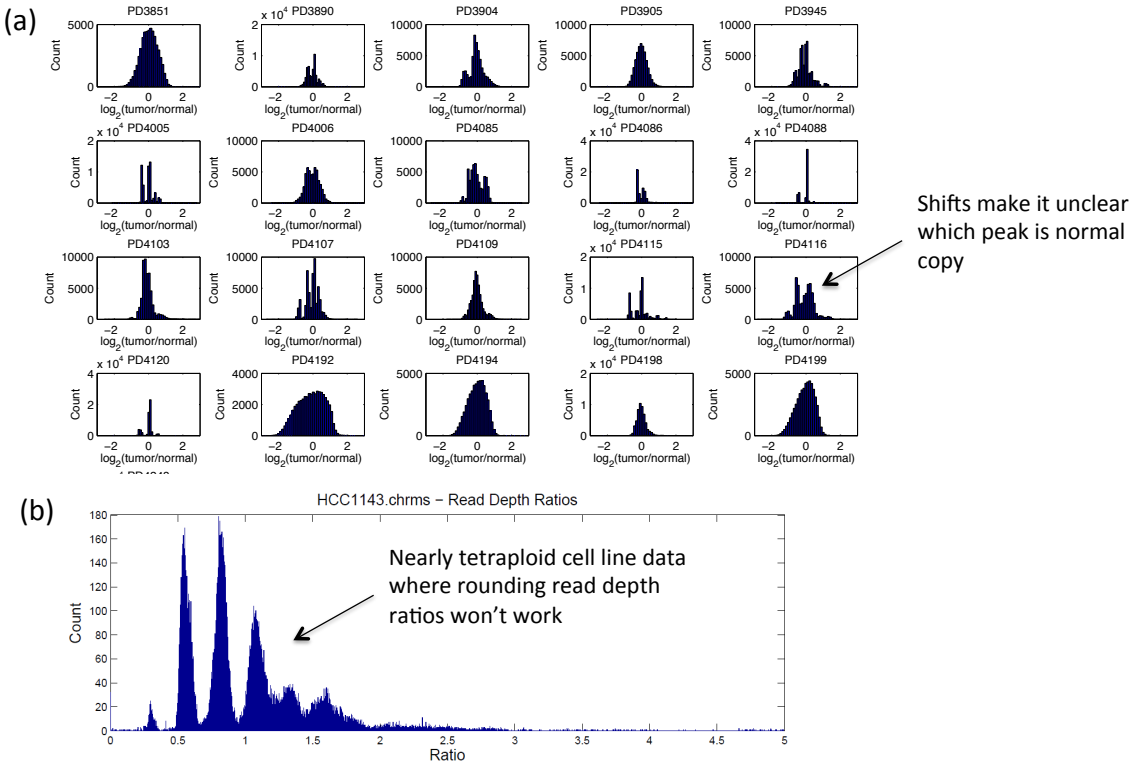


Figure A.2: **Read depth ratios across 50kb bins for 20 breast cancer genomes and one cell line.(a)**Tumor/Normal read depth ratios over 50kb bins for 20 of the breast caner genomes from [114]. In several instances we can observe shifting in the read depth ratios away from 0, perhaps due to changes in the genome length, thus motivating the use of the multinomial model. **(b)** Tumor/Normal read depth ratios over 50kb gins for breast cancer cell line HCC1143, which is nearly tetraploid, making simple rounding of read depth ratios to obtain copy number estimates highly inaccurate.

$X_{pq} = 1$ if the p^{th} read from $\mathcal{T}$ aligns to I_q and 0 otherwise. Therefore $X_q = \sum_{n=1}^{N} X_{pq}$ is the number of reads from $\mathcal{T}$ that align to I_q. Let Y_q be a similar random variable, but for a matched normal sample $\mathcal{N}$. We can now calculate the expected number of reads aligning to interval I_q.

$$E[X_q] = E[\sum_{p=1}^{P} X_{pq}] = \sum_{p=1}^{P} E[X_{pq}] = \sum_{p=1}^{P} Prob(X_{pq} = 1) = P \times Prob(X_{pq} = 1)$$

Now assume that $\mathcal{T}$ has a true underlying $\mathbf{C}$ and μ where $n = 2$ and the first column of $\mathbf{C}$ is set to 2 (the normal component). Let $\rho = \frac{1}{m} \sum_{k=1}^{m} c_{k2}$, that is ρ is the average value of entries in the second column of $\mathbf{C}$. Using our multinomial model, we can directly calculate $Prob(X_{pq} = 1)$.

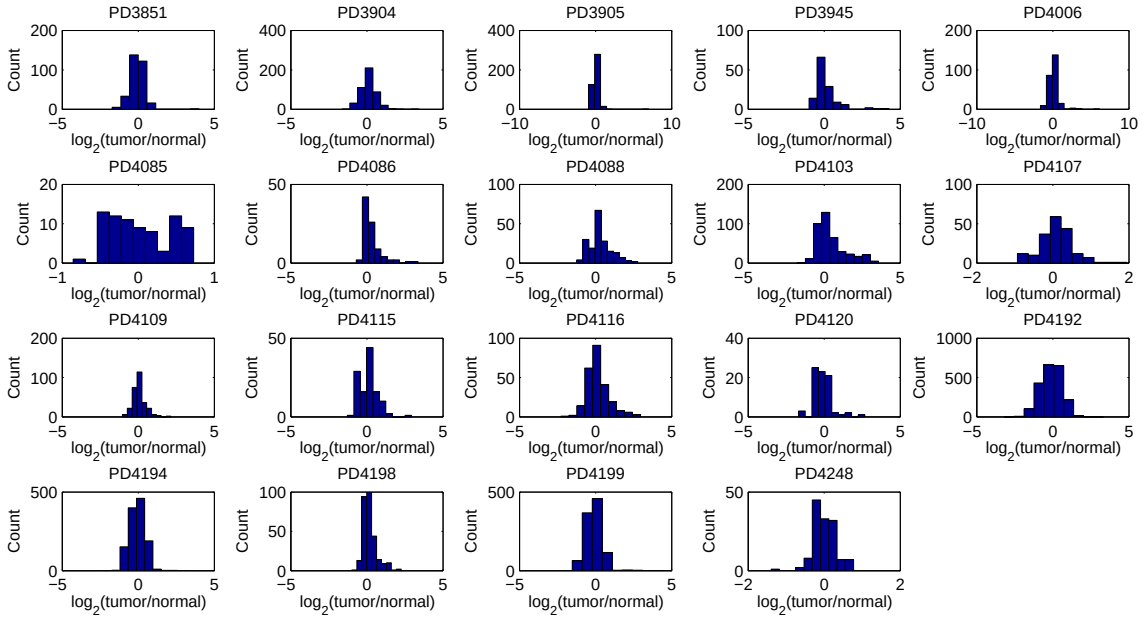


Figure A.3: **Read depth ratios using intervals determined using BIC-Seq for 19 breast cancer samples.** BIC-Seq run with $\lambda = 100$.

$$Prob(X_{pq} = 1) = (\widehat{\mathbf{C}\mu})_q = \frac{(\mathbf{C}\mu)_q}{\sum_{k=1}^{m}(\mathbf{C}\mu)_k} = \frac{2\mu_1 + c_{q2}\mu_2}{\sum_{k=1}^{m}(2\mu_1 + c_{k2}\mu_2)}$$
$$= \frac{2\mu_1 + c_{q2}\mu_2}{2m\mu_1 + \mu_2\sum_{k=1}^{m} c_{k2}} = \frac{2\mu_1 + c_{q2}\mu_2}{2m\mu_1 + \mu_2 m\rho} = \frac{2\mu_1 + c_{q2}\mu_2}{m(2\mu_1 + \mu_2\rho)}$$

Since $\mathcal{N}$ is just the normal sample and has all copy 2, we similarly calculate $Prob(Y_{pq} = 1) = \frac{1}{m}$. We can now see how to derive the equation used by both [27] and [164].

$$\frac{E[X_q]}{E[Y_q]} = \frac{P \times Prob(X_{pq} = 1)}{P \times Prob(Y_{pq} = 1)} = \frac{Prob(X_{pq} = 1)}{Prob(Y_{pq} = 1)}$$
$$= \frac{\frac{2\mu_1 + c_{j2}\mu_2}{m(2\mu_1 + \mu_2\rho)}}{\frac{1}{m}} = \frac{2\mu_1 + c_{j2}\mu_2}{2\mu_1 + \mu_2\rho}$$

A.1.3 Separable Convexity of $\mathcal{L}_{\mathbf{r}}(\mathbf{p})$

In this section we show that our objective function is separable convex.

Lemma A.1.1. *$\mathcal{L}_{\mathbf{r}}(\mathbf{p}) = -\sum_{j=1}^{m} r_j \log(p_j) + \alpha$ is separable convex for $\mathbf{p} \in P_m$.*

Proof. To show that $\mathcal{L}_{\mathbf{r}}(\mathbf{p})$ is a separable convex function for $\mathbf{p} \in P_m$, we show (i) P_m is a convex space; (2) the functions $\ell_j(p_j) = -r_j \log(p_j)$ are convex. Finally since $\mathcal{L}_{\mathbf{r}}(\mathbf{p}) = \sum_{j=1}^{m} \ell_j(p_j) + \alpha$ we conclude the separable convexity of $\mathcal{L}_{\mathbf{r}}(\mathbf{p})$.

Suppose $\lambda \in [0,1]$ and $\mathbf{p}, \mathbf{q} \in P_m$ are arbitrarily chosen. Let $\mathbf{s} = \lambda\mathbf{p} + (1-\lambda)\mathbf{q}$. By definition $s_j = \lambda p_j + (1-\lambda)q_j \geq 0$ since $p_j, q_j, \lambda \geq 0$, and $\sum_{j=1}^{m} s_j = 1$ since

$$\sum_{j=1}^{m} s_j = \lambda \sum_{j=1}^{m} p_j + (1-\lambda) \sum_{j=1}^{m} q_j = \lambda + (1-\lambda) = 1 \Rightarrow \mathbf{s} \in P_m, \text{ and } P_m \text{ is convex.}$$

Now, for any $j \in \{1, \ldots, m\}$

$$\begin{aligned}
\ell_j(s_j) = \ell_j(\lambda p_j + (1-\lambda)q_j) &= -r_j \log(\lambda p_j + (1-\lambda)q_j) \\
&\leq -r_j(\lambda \log(p_j) + (1-\lambda)\log(q_j)) \qquad \text{(By Jensen's Inequality)} \\
&= -r_j \lambda \log(p_j) - r_j(1-\lambda)\log(q_j) \\
&= \lambda \ell_j(p_j) + (1-\lambda)\ell_j(q_j) \\
&\Rightarrow \text{the function } \ell_j(p_j) \text{ is a convex function.}
\end{aligned}$$

Finally, since α is a constant and $\mathcal{L}_{\mathbf{r}}(\mathbf{p}) = \sum_{j=1}^{m} \ell_j(p_j) + \alpha$, the function $\mathcal{L}_{\mathbf{r}}(\mathbf{p})$ is separable convex.

□

A.1.4 Proof of Theorem 2.2.3

Theorem 2.2.3. *Suppose $\mathbf{p}^* = \widehat{\mathbf{C}^*\mu^*} = \underset{p \in P_{\Omega_{m,n,k,p}}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p})$ and all entries in $\mathbf{r}$ are distinct. Then we have the following: $\mathbf{p}^*$ has compatible order with $\mathbf{r}$.*

Proof. We proceed with proof by contradiction. Suppose $\mathbf{r}$ and $\mathbf{p}$ do not have a compatible order, that is there exist $i, j \in \{1, \ldots, m\}$ such that $r_i \geq r_j$, but $p_i < p_j$. Since all entries in $\mathbf{r}$ are distinct, then necessarily $r_i > r_j$. Without loss of generality assume $i < j$, and let $\mathbf{p}'$ be a point on the

simplex obtained from $\mathbf{p}$ by swapping the i^{th} and the j^{th} entries. We first show that $\mathbf{p}' \in P_{\Omega_{m,n,k,p}}$. Since $\mathbf{p} \in P_{\Omega_{m,n,k,p}}$, by definition there exists a $(\mathbf{C}, \mu) \in \Omega_{m,n,k,p}$ such that $\mathbf{p} = \widehat{\mathbf{C}\mu}$. Let $\mathbf{C}'$ be a matrix obtained from $\mathbf{C}$ by swapping the i^{th} and the j^{th} rows. We define $\mathbf{p}' = \widehat{\mathbf{C}'\mu}$. Since the set of entries in $\mathbf{C}'$ is as same as the set of entries in $\mathbf{C}$, we have $(\mathbf{C}', \mu) \in \Omega_{m,n,k,p}$. Thus $\mathbf{p}' \in P_{\Omega_{m,n,k,p}}$.

By our original assumption $\mathcal{L}_{\mathbf{r}}(\mathbf{p})$ is minimal, and thus $\mathcal{L}_{\mathbf{r}}(\mathbf{p}) \leq \mathcal{L}_{\mathbf{r}}(\mathbf{p}') \Rightarrow \mathcal{L}_{\mathbf{r}}(\mathbf{p}) - \mathcal{L}_{\mathbf{r}}(\mathbf{p}') \leq 0$. We define $\delta = r_i - r_j > 0$. We have

$$\begin{aligned}
\mathcal{L}_{\mathbf{r}}(\mathbf{p}) - \mathcal{L}_{\mathbf{r}}(\mathbf{p}') &= (-\sum_{i=1}^{m} r_i \log(p_i) + \alpha) - (-\sum_{i=1}^{m} r_i \log(p'_i) + \alpha) \\
&= -r_i \log(p_i) - r_j \log(p_j) + r_i \log(p_j) + r_j \log(p_i) \\
&= -(r_j + \delta) \log(p_i) - r_j \log(p_j) + (r_j + \delta) \log(p_j) + r_j \log(p_i) \\
&= -r_j \log(p_i) - \delta \log(p_i) - r_j \log(p_j) + r_j \log(p_j) + \delta \log(p_j) + r_j \log(p_i) \\
&= -\delta \log(p_i) + \delta \log(p_j) \\
&= \delta(\log(p_j) - \log(p_i)) \\
&> 0.
\end{aligned}$$

This is a contradiction to $\mathcal{L}_{\mathbf{r}}(\mathbf{p})$ as minimal. Therefore, it must be the case that $\mathbf{r}$ and $\mathbf{p}$ have compatible order. Thus, if we consider a fixed purity, then we need only to consider the same set of matrices $\mathbf{C}$ as if we were considering all purities.

□

A.1.5 Further Details Related to Theorem 2.2.1

As described in the main text, Theorem 2.2.1 allows us to solve separate convex optimization problems in the space P_{Ω}. Figure A.4 shows an example of the geometry of the problem when $m = 4$ and $n = 3$.

A.1.6 Proof of the Theorem 2.2.4

In this section we present the proof for Theorem 2.2.4. To do so, we first prove two important properties of the function Φ, and next we show there is a unique $\mathbf{p} \in \Phi(\mathbf{C}\mu)$, where $(\mathbf{C}, \mu) \in \Omega_{m,n}$, such that $\mathcal{L}_{\mathbf{r}}(\mathbf{p})$ is minimized.

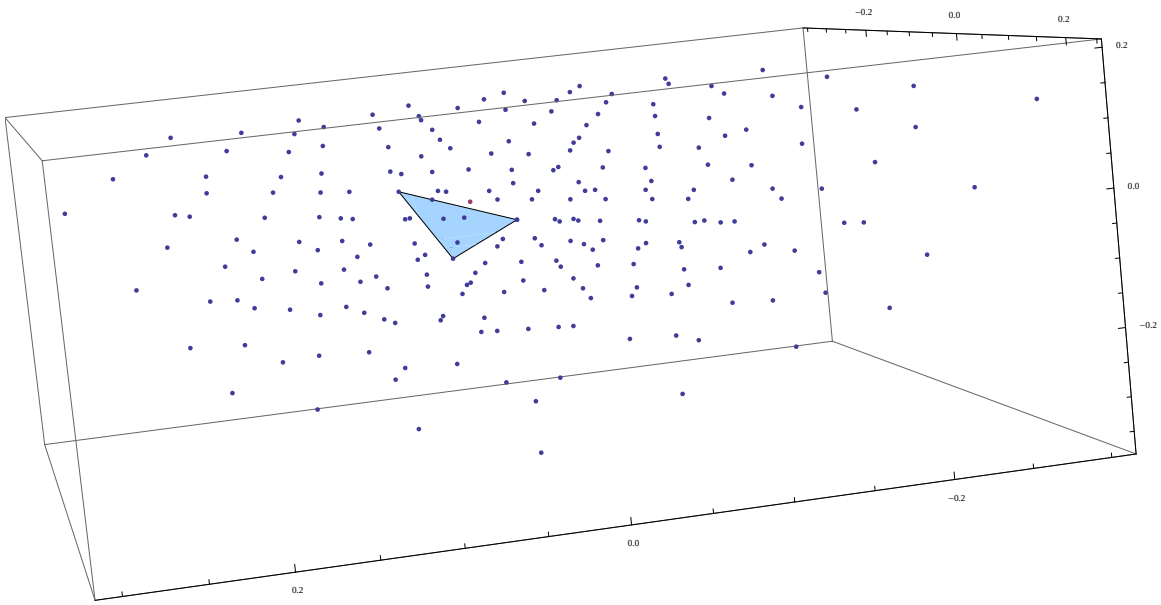


Figure A.4: **The convex geometry of the MLMDP that is used in the THetA algorithm in the instance when $n = 3$.** When $n = 3$, convex combinations of normalized columns in $\mathbf{C}$ form planes, or 2-simplices, which are embedded in Δ_{m-1} (as described in Theorem 2.2.1). Here we show an example with $m = 4$ where one such plane is highlighted in blue.

Lemma A.1.2. *The function Φ has the following properties:*

(i) $\Phi(\mathbf{v}) = \Phi(\widehat{\mathbf{v}})$, for all $\mathbf{v} \in \mathbb{R}^m$, and

(ii) restrictions of Φ and Φ^{-1} to Δ_{m-1} are inverse of each other.

Proof. First we show that $\Phi(\mathbf{v}) = \Phi(\widehat{\mathbf{v}})$:

$$\Phi(\mathbf{v}) = \widehat{\mathbf{W}\mathbf{v}} = \frac{\mathbf{W}\mathbf{v}}{|\mathbf{W}\mathbf{v}|_1} = \frac{\frac{\mathbf{W}\mathbf{v}}{|\mathbf{v}|_1}}{\frac{|\mathbf{W}\mathbf{v}|_1}{|\mathbf{v}|_1}} = \frac{\mathbf{W}\frac{\mathbf{v}}{|\mathbf{v}|_1}}{|\mathbf{W}\frac{\mathbf{v}}{|\mathbf{v}|_1}|_1} = \frac{\mathbf{W}\widehat{\mathbf{v}}}{|\mathbf{W}\widehat{\mathbf{v}}|_1} = \Phi(\widehat{\mathbf{v}}). \tag{A.1}$$

By definition of Φ^{-1} we have $\Phi^{-1}(\mathbf{q}) = \widehat{\mathbf{W}^{-1}\mathbf{q}}$. Let $\mathbf{p} \in \Delta_{m-1}$:

$$\Phi^{-1}(\Phi(\mathbf{p})) = \Phi^{-1}(\widehat{\mathbf{W}\mathbf{p}}) = \Phi^{-1}(\frac{\mathbf{W}\mathbf{p}}{|\mathbf{W}\mathbf{p}|_1}) = (\widehat{\mathbf{W}^{-1}\frac{\mathbf{W}\mathbf{p}}{|\mathbf{W}\mathbf{p}|_1}}) = \frac{\mathbf{W}^{-1}\frac{\mathbf{W}\mathbf{p}}{|\mathbf{W}\mathbf{p}|_1}}{|(\mathbf{W}^{-1}\frac{\mathbf{W}\mathbf{p}}{|\mathbf{W}\mathbf{p}|_1})|_1}$$
$$= \frac{\frac{\mathbf{p}}{|\mathbf{W}\mathbf{p}|_1}}{\frac{|\mathbf{p}|_1}{|\mathbf{W}\mathbf{p}|_1}} = \frac{\mathbf{p}}{|\mathbf{p}|_1} = \mathbf{p},$$

where the last equation comes from the fact that $\mathbf{p} \in \Delta_{m-1}$ and $|\mathbf{p}|_1 = 1$. Using the same argument $\Phi(\Phi^{-1}(\mathbf{p})) = \mathbf{p}$, and the proof is complete. □

Theorem A.1.3. $\left| \underset{\mathbf{p}=\Phi(\mathbf{C}\mu),(\mathbf{C},\mu)\in\Omega_{m,n}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p}) \right| = 1.$

Proof. Let $\mu^* = (1, 0, 0, \ldots, 0)^T \in \mathbb{R}^n$ and $\mathbf{C}^*$ be a matrix whose first column is $(c_1^*, \ldots, c_m^*)^T$, where $c_i^* = (\prod_j w_j) \cdot \frac{r_i}{w_i} \in \mathbb{Z}$, and other entries of $\mathbf{C}$ are arbitrary integers in $\mathbb{N}$. Define $\mathbf{p}^* = \Phi(\mathbf{C}^*\mu^*)$. We claim that $\mathbf{p}^*$ is the only element in the set $\underset{\mathbf{p}=\Phi(\mathbf{C}\mu),(\mathbf{C},\mu)\in\Omega_{m,n}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p})$.

First, $\mathbf{p}^* \in \underset{\mathbf{p}=\Phi(\mathbf{C}\mu),(\mathbf{C},\mu)\in\Omega_{m,n}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p})$: by definition we have

$$p_i^* = \Phi(\mathbf{C}^*\mu^*)_i = \frac{c_i^* \cdot w_i}{\sum_j c_j^* \cdot w_j} = \frac{(\prod_j w_j)\frac{r_i}{w_i} \cdot w_i}{\sum_h \frac{r_h \prod_j w_j}{w_h} \cdot w_h} = \frac{(\prod_j w_j) r_i}{(\prod_j w_j)\sum_h r_h} = \frac{r_i}{\sum_h r_h} = (\widehat{\mathbf{r}})_i.$$

This implies that $\mathbf{p}^* \in \underset{\mathbf{p}\in\Delta_{m-1}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p})$, and since $\{\Phi(\mathbf{C}\mu) \mid (\mathbf{C}, \mu) \in \Omega_{m,n})\} \subset \Delta_{m-1}$ we have $\mathbf{p}^* \in \underset{\mathbf{p}=\Phi(\mathbf{C}\mu),(\mathbf{C},\mu)\in\Omega_{m,n}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p})$.

Now suppose $\mathbf{p}' \in \underset{\mathbf{p}=\Phi(\mathbf{C}\mu),(\mathbf{C},\mu)\in\Omega_{m,n}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p})$. Thus, $\mathcal{L}_{\mathbf{r}}(\mathbf{p}') = \mathcal{L}_{\mathbf{r}}(\mathbf{p}^*)$, and $\mathbf{p}' \in \underset{\mathbf{p}\in\Delta_{m-1}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p})$, as we showed $\mathbf{p}^* \in \underset{\mathbf{p}\in\Delta_{m-1}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p})$. But $\mathbf{p}' = \mathbf{p}^*$, since there is a unique optimal point for $\underset{\mathbf{p}\in\Delta_{m-1}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p})$, i.e., $|\underset{\mathbf{p}\in\Delta_{m-1}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\Phi(\mathbf{p}))| = 1$, which completes the proof for uniqueness of $\mathbf{p}^*$. □

Theorem 2.2.4. *Let $\Phi^{-1} : \mathbb{R}^m \to \mathbb{R}^m$ be $\Phi^{-1}(\mathbf{v}) = \widehat{W^{-1}\mathbf{v}}$. We have the following set equality,*

$$\underset{(\mathbf{C},\mu)\in\Omega_{m,n}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\Phi(\mathbf{C}\mu)) = \underset{(\mathbf{C},\mu)\in\Omega_{m,n}}{\operatorname{argmin}} \mathcal{L}_{\Phi^{-1}(\mathbf{r})}(\widehat{\mathbf{C}\mu}).$$

Proof. By Theorem A.1.3 $\underset{\mathbf{p}=\mathbf{C}\mu,(\mathbf{C},\mu)\in\Omega_{m,n}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\mathbf{p})$ has a unique point $\mathbf{p}^*$, where $p_i^* = \frac{r_i}{\sum_{j=1}^m r_j}$. There is a unique optimal point $\mathbf{q}^*$ in $\underset{(\mathbf{C},\mu)\in\Omega_{m,n}}{\operatorname{argmin}} \mathcal{L}_{\Phi^{-1}(\mathbf{r})}(\widehat{\mathbf{C}\mu})$, where $q_i^* = \frac{\frac{r_i}{w_i}}{\sum_{j=1}^m \frac{r_j}{w_j}}$. Now we have

$$(\mathbf{C}^*, \mu^*) \in \underset{(\mathbf{C},\mu)\in\Omega_{m,n}}{\operatorname{argmin}} \mathcal{L}_{\mathbf{r}}(\Phi(\mathbf{C}\mu)) \Leftrightarrow \mathcal{L}_{\mathbf{r}}(\Phi(\mathbf{C}^*\mu^*) = \mathcal{L}_{\mathbf{r}}(\mathbf{p}^*)$$

$$\text{(using uniqueness of } \mathbf{p}^*\text{, and Lemma A.1.2)} \Leftrightarrow \Phi(\widehat{\mathbf{C}^*\mu^*}) = \Phi(\mathbf{C}^*\mu^*) = \mathbf{p}^* = \frac{\mathbf{r}}{\sum_j r_j} = \widehat{\mathbf{r}}$$

$$\text{(applying } \Phi^{-1} \text{ and using Lemma A.1.2)} \Leftrightarrow \widehat{\mathbf{C}^*\mu^*} = \Phi^{-1}(\widehat{\mathbf{r}})$$

$$\text{(by Lemma A.1.2)} \Leftrightarrow \widehat{\mathbf{C}^*\mu^*} = \Phi^{-1}(\mathbf{r})$$

$$\text{(since } \Phi^{-1}(\mathbf{r}) \in \Delta_{m-1},\ |\Phi^{-1}(\mathbf{r})|_1 = 1) \Leftrightarrow \widehat{\mathbf{C}^*\mu^*} = \frac{\Phi^{-1}(\mathbf{r})}{|\Phi^{-1}(\mathbf{r})|_1} = \left(\frac{\frac{r_1}{w_1}}{\sum_{j=1}^m \frac{r_j}{w_j}}, \ldots, \frac{\frac{r_m}{w_m}}{\sum_{j=1}^m \frac{r_j}{w_j}}\right)$$

$$\Leftrightarrow \mathcal{L}_{\Phi^{-1}(\mathbf{r})}(\widehat{\mathbf{C}^*\mu^*}) = \mathcal{L}_{\Phi^{-1}(\mathbf{r})}(\mathbf{q}^*)$$

$$\Leftrightarrow (\mathbf{C}^*, \mu^*) \in \underset{(\mathbf{C},\mu)\in\Omega_{m,n}}{\operatorname{argmin}} \mathcal{L}_{\Phi^{-1}(\mathbf{r})}(\widehat{\mathbf{C}\mu}).$$

□

A.2 Simulated Data

In this section we describe our methods for data simulation of sequencing data, corresponding SNP array data, and synthetic mixtures of real sequencing data. We also present additional simulation results.

A.2.1 Simulated Data Creation

Simulated Sequencing Data with $n = 2$ We randomly simulate sequencing data using for a mixture of tumor and normal cells ($n = 2$) using the following the following procedure. For the interval partition $\mathbf{I}$ we use the 39 autosomal chromosome arms excluding the 5 acro-centric p-arms on chromosomes 13,14,15,21, and 22. We sample an interval count matrix $\mathbf{C}$ uniformly at random from $\mathcal{C}_{39,2,k}$. For each simulated tumor sample, we draw a value μ_2 uniformly from the interval $[0.5, 0.95]$, values of tumor fraction that are reasonable for real cancer sequencing data. The expected distribution of reads is then $\mathbf{p} = \Phi(\mathbf{C}\mu)$, where the weight vector ω, is obtained from paired end sequencing data from 9 normal human genomes from [114]. ω is determined by first counting the number of concordant read pairs (with mapping quality ≥ 30) that align to the interval partition $\mathbf{I}$ for all 9 genomes. We then average the observed distribution of reads over these intervals to obtain

a mean observed weight vector ω which we then normalize to obtain a valid multinomial parameter.

Under perfect conditions, the read depth vector $\mathbf{r}$ and weight vector $\mathbf{w}$ are drawn directly from the multinomial distributions with parameters $\mathbf{p}$ and ω. We simulate errors in the sequencing and analysis process by adding noise to $\mathbf{r}$ and $\mathbf{w}$. These errors occur for a variety of reasons. First, copy number aberrations that change the length of the cancer genome, but are not appropriately represented in $\mathbf{I}$, may *globally* alter the observed read depth over all intervals. We model this noise by drawing vectors $\mathbf{r}_0$ and $\mathbf{w}_0$ from a Dirichlet prior with parameters proportional to $\mathbf{p}$ and ω such that the expected number of reads corresponds to 30X coverage of the normal genome. Second, additional sources of noise in read depth estimation occur due to sequencing errors or alignment errors caused by repetitive regions. To model these errors, we add Gaussian noise to each entry of r_i of $\mathbf{r}_0$ (resp. w_i of $\mathbf{w}_0$) using a Gaussian distribution with mean 0 and standard deviation ϕr_i (resp. ϕw_j). ϕ is estimated from real data as described below.

Simulated Sequencing Data with $n = 3$ We construct simulated sequencing data for a mixture of normal cell and two cancer subpopulations ($n = 3$ genomes in the mixture) using the same procedure as for $n = 2$ genomes with the following changes: (1) We require the content of each tumor component to be greater than 20%; (2) our set of intervals is just the first m q-arms and (3) we randomly sample $\mathbf{C}$ such that it contains a fixed number of amplifications and a random number of heterozygous deletions (similar to the real genomes we analyze). We set individual lower and upper bounds on the copy number for each interval using the same heuristic we use for real data (Supplemental Material Section A.3).

Simulated SNP Array To compare our algorithm for sequencing data to the algorithm ASCAT [164], designed for SNP data, we devised a method for simulating sequencing data and then converting that data to SNP array data in the format required by ASCAT. We initially create read depth data using the process described in the previous section. For both the tumor and normal genome we then create LogR and B allele-frequencies (BAF) values (necessary values for running ASCAT) for the 907,693 SNP positions on the 22 autosomes queried by the Affymetrix 6.0 SNP array. The LogR value for the tumor sample is $\log_2$ of the ratio of the tumor to normal read counts for the interval containing the SNP location. The LogR ratio for the normal sample is 0, indicating a copy number of 2. We randomly determine A_N and B_N, the number of A and B alleles for each SNP in

the normal genome. We then draw a total read count for the SNP from a poisson distribution with parameter equal to the observed coverage for the genomic interval containing the SNP. We then simulate the number reads with the variant allele by making draws from a binomial distribution with parameter equal to the expected variant allele fraction $\frac{B_N}{A_N+B_N}$. The number of reads with the variant allele divided by the total number of reads gives the observed BAF. BAFs for the tumor sample are created in a similar manner after randomly determining the number of copies of each parental chromosome for each interval in the tumor cells. If the total number of copies for an interval is ≥ 2 we require that at least one copy of each parental chromosome is retained in the tumor cells (otherwise this implies the more complicated situation where multiple events would have occurred to the same interval). Using the data from the matched normal, we calculate A_T and B_T the number A and B alleles for each SNP in the tumor genome. We define μ_N to be the fraction of normal cells in the sample, and μ_T to be the fraction of tumor cells in the sample ($\mu_N + \mu_T = 1$). The expected variant allele fraction in the tumor sample is calculated using the following equation:

$$\begin{aligned} BAF = &(\frac{B_N}{A_N + B_N})\frac{\mu_N(A_N + B_N)}{\mu_N(A_N + B_N) + \mu_T(A_T + B_T)} + \\ &(\frac{B_T}{A_T + B_T})\frac{\mu_T(A_T + B_T)}{\mu_N(A_N + B_N) + \mu_T(A_T + B_T)} \end{aligned} \tag{A.2}$$

Lastly, by default ASCAT assumes that LogR values have been scaled by a platform dependent parameter γ. We multiply all LogR values by $\gamma = 0.55$, which [164] reports as default for Illumina. We also ran all experiments with $\gamma = 1$ and find that ASCAT performs much better when $\gamma = 0.55$ and therefore present only those results.

Simulated Mixtures of Real Data In this section we describe how we created simulated mixtures of tumor cells with normal cell admixture by using real sequencing data from a matched tumor and normal AML samples. We first obtained BAM files for sample TCGA-AB-2965 obtained from CG-Hub [25]. We choose this sample because it is estimated to have 95% purity and zero copy number aberrations (as determined using array data) [25] which allows us to create datasets with realistic sequencing noise by spiking in copy number variants and mixing reads in different proportions from the tumor and normals.

We first identify all concordant read pairs where each read has mapping quality ≥ 30 and use

this data to create simulated mixtures. We spike in 10 copy number variants of a fixed length (for the reported simulations we used variants of length 2.5Mb) at random non-overlapping positions in Chr20 (excluding the centromere) by up/down sampling concordant pairs. The copy number of the variant is uniformly at randomly determined to be either a deletion (heterozygous or homozygous) or an amplification (up to copy 5). We then create mixtures of this tumor genome and the matched normal by mixing together concordant pairs sampled from each. Concordant pairs are sampled with probability such that the expected total number of reads sampled from the tumor and normal samples respectively reflects a uniform sampling from all DNA in the sample given the mixing percentages and lengths of the tumor and normal genomes (while maintaining the original coverage of the original tumor sample). We then create samples of different coverage by randomly up/down sampling from this mixture. Interval partitions for Chr20 is then determined by running BIC-Seq with parameter $\lambda = 10$. We run with a smaller value of λ than we use on real data since we are only considering a single chromosome and can therefore allow for a finer partition of the reference genome into intervals. The read depth vectors $\mathbf{r}$ and $\mathbf{w}$ used as input to THetA and CNAnorm is just the count of the number of concordant pairs aligning within the derived intervals from the mixed sample and matched normal. The input to ABSOLUTE is the $\log_2$ ratio of these counts.

Read Depth Estimation Error ϕ In simulation we use an empirically observed value of the read depth estimation error ϕ by using the distribution of read depth, after normalizing for coverage, over the non-acrocentric chromosome arms of 9 normal samples from [114]. Over all intervals we find a mean value of $\phi = 0.037$ and median of $\phi = 0.029$. When we exclude the genome with the lowest actual coverage we observe a mean value of $\phi = 0.021$ and median of $\phi = 0.013$ (see Figure A.5).

A.2.2 Details of Other Algorithms

In this section we explain other details pertaining to the other algorithms that we compare THetA against.

CNAnorm Details Since CNAnorm first determines the set of copy numbers expressed in a sample, before estimating the sample purity, errors in the first step may result in estimated purity values above 100% (as discussed in [62]). Therefore, in simulation we only considered trials where purity was inferred $\leq 100\%$. Additionally, since CNAnorm returns non-integer copy numbers, we

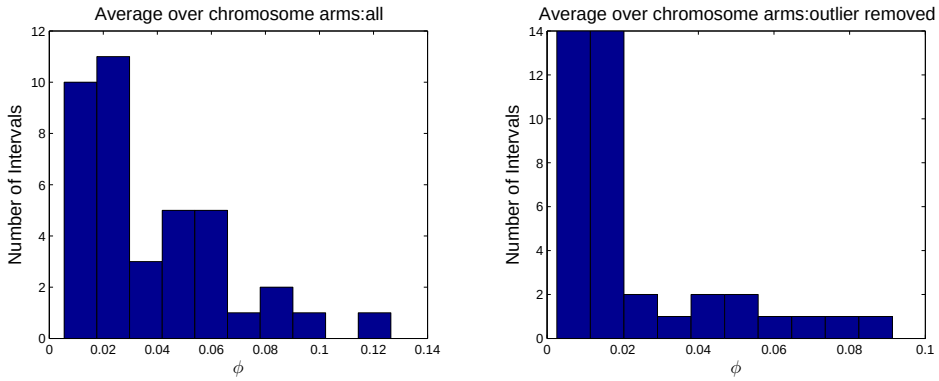


Figure A.5: **Distributions of observed read depth estimation error ϕ.** The distribution of the average observed ϕ over the non-acrocentric chromosome arms for 9 normal genomes (left). This distribution has mean 0.037 and median 0.029. The distribution of the average observed ϕ over the non-acrocentric chromosome arms for 8 normal genomes (right) after removal of the genome with the lowest coverage. This distribution has mean 0.021 and median 0.013.

round their results to the nearest integer value when comparing the interval count matrices $\mathbf{C}$.

ABSOLUTE Details ABSOLUTE does not return a single solution, but instead returns three sets of solutions: (1) Solutions based on somatic copy number aberrations (SCNA); (2) Solutions based on recurrent Karyotypes; and (3) Solutions based on combined SCNAs and Karyotypes. Since our simulations are based on randomly generated cancer genomes, we select the ABSOLUTE solution with the highest likelihood from the SCNA category as the returned solution to compare against.

When running ABSOLUTE on simulated data, we set the maximum possible ploidy to the maximum possible value of copy number aberrations in the simulated data rather than the default value of 10. In all cases this limits the set of possible solutions considered by ABSOLUTE. All other parameters are set to their default values as described in the ABSOLUTE documentation.

A.2.3 Additional Simulation Results

Here we present a more detailed analysis comparing the results of THetA, CNAnorm [62], ASCAT [164] and ABSOLUTE [27]. Figure A.6 shows the true interval count matrix $\mathbf{C}$ and genome mixing vector μ and the inferred values by the above algorithms on one of the trials from the first set of simulations in the main text. The figure also includes calculations of the Copy number error and Purity error for these solutions (as defined in the main text). THetA has the most accurate purity estimate of all algorithms - being with 0.5% of the true purity and only misestimate the copy of

three segments (each where the copy number estimate is off from the true value) whereas CNAnorm, ASCAT and ABSOLUTE misestimate $17, 4$ and 37 segments respectively (out of 39 possible). The large number of copy number estimates by ABSOLUTE show the dependence between copy number estimates and purity estimates as ABSOLUTE gravely underestimates the purity of this sample.

We also ran additional experiments comparing our algorithm to CNAnorm for varying read depth estimation error ϕ, a parameter that is not relevant for the SNP array data used by ASCAT. We compare to three versions of CNAnorm [62]: (1) run with default parameters (CNAnorm), (2) run with an optional smoothing step (CNAnorm-S), and (3) run where each chromosome arm was broken into 100 intervals (CNAnorm-M) with equal read depth in each interval. In all cases, our algorithm consistently outperforms CNAnorm by a large margin – beating CNAnorm by up to 50 percentage points at estimating $\mathbf{C}$ correctly (Figure A.7). Figure A.8 shows the complete results for our method and CNAnorm when we analyze how well each estimates sample purity.

We also include additional simulation results for the experiments using real sequencing data to create varying mixtures of tumor and normal admixture. These results are for the simulated data presented in the main manuscript, but we provide here results where a true positive for predicting a copy number variant only requires a 50% reciprocal overlap with a true variant and a non-normal copy number predicted without requiring that the copy number be exactly correct. We find that the results for THetA change only slightly - indicating that when THetA predicts a copy number variant, it often predicts the true copy number accurately. Whereas we see significant changes in the accuracy of both CNAnorm and ABSOLUTE - indicating that these algorithms, especially for low sample purity, incorrectly estimate the overall ploidy of the tumor genomes in the mixture.

A.3 Heuristics Applied to Real Data

We assume that most of the tumor genome does not undergo focal copy number aberrations. Thus, the mode of the read depth vector provides a normal "baseline" and allows us to set tighter lower and upper bounds on the copy number for each interval. This allows us to use the following heuristics when analyzing real sequencing data.

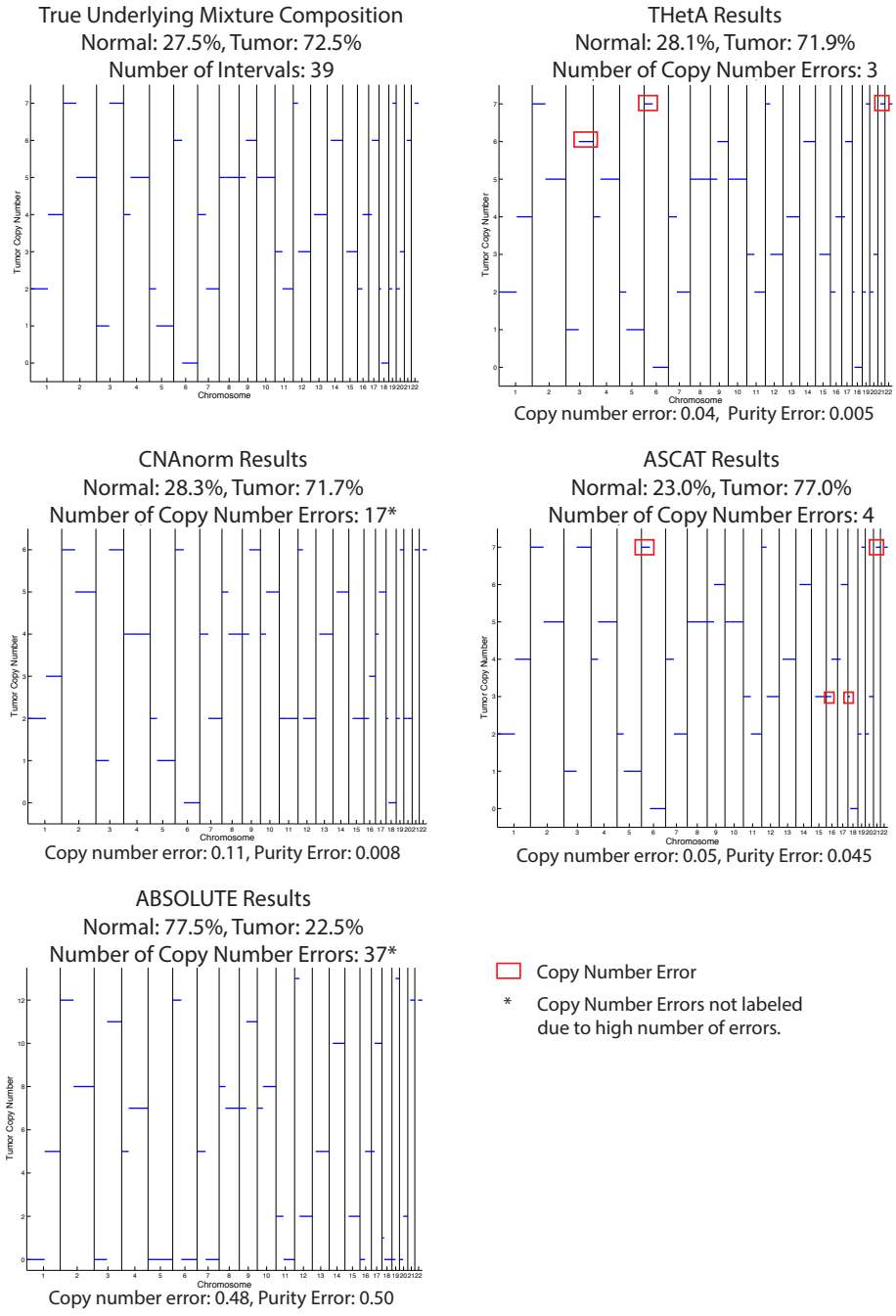


Figure A.6: **The true underlying interval count matrix C and genome mixing vector μ for one simulation along with sample reconstructions by THetA, CNAnorm, ASCAT and ABSOLUTE.** Copy number error is $\frac{1}{m(n-1)}|\mathbf{C}-\mathbf{C}^*|_2$, that is, the average error per copy number estimate made, or per entry in $\mathbf{C}$, where error is the euclidean distance between $\mathbf{C}$ and $\mathbf{C}^*$. Purity error is $|\mu_2 - \mu_2^*|$, that is the distance between the true and inferred sample purity.

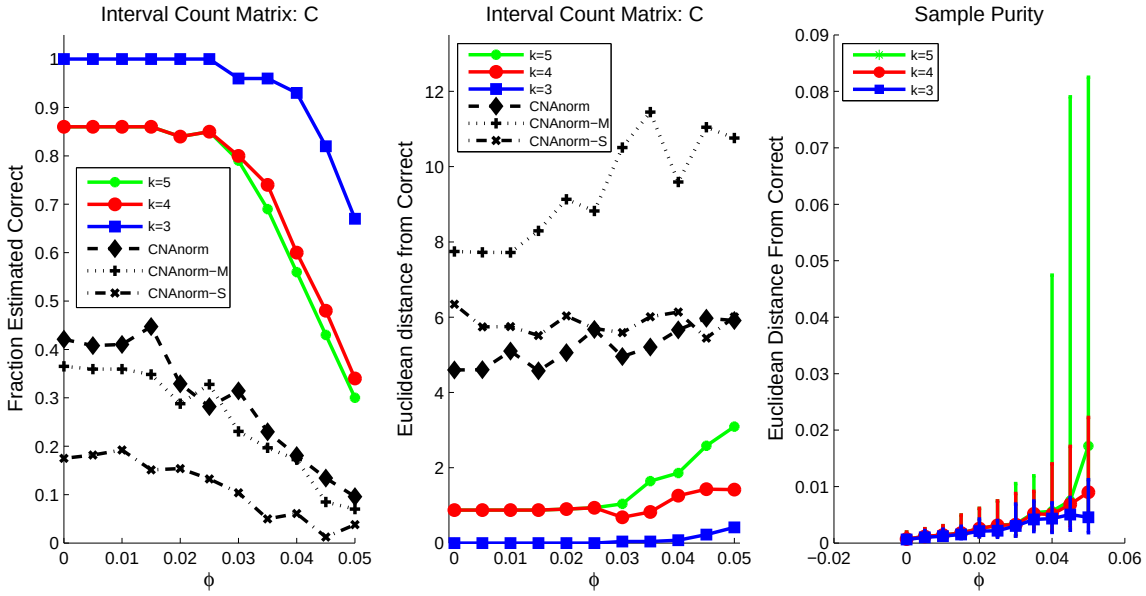


Figure A.7: **Comparison of THetA and CNAnorm on simulated tumor samples.** Simulations contain normal admixture between 5% and 50% and true underlying interval count matrix **C** sampled uniformly at random from $\mathcal{C}_{39,2,3}$. We perform 100 random trials (between 57 and 86 of which CNAnorm returns purity $< 100\%$) across varying read depth estimation error (ϕ). We also vary the maximum copy number (k) considered by THetA (beyond the true maximum copy number in the sample). We outperform CNAnorm in the metrics: (left) average number of trials where the inferred interval count matrix **C** exactly matches the true underlying interval count matrix for the sample ; (middle) average euclidean distance of between the inferred and true integer count vector $\mathbf{c}_2$ of the tumor genome; (right) median of the euclidean distance between the inferred and true genome mixing vector μ. Error bars represent the 25 and 75 percentiles. In the right figure, values for CNAnorm are not shown as the corresponding median values are between 0.07 and 0.13 – outside the range of the plot when error bars are included. (Appendix Figure A.8 gives all CNAnorm results).

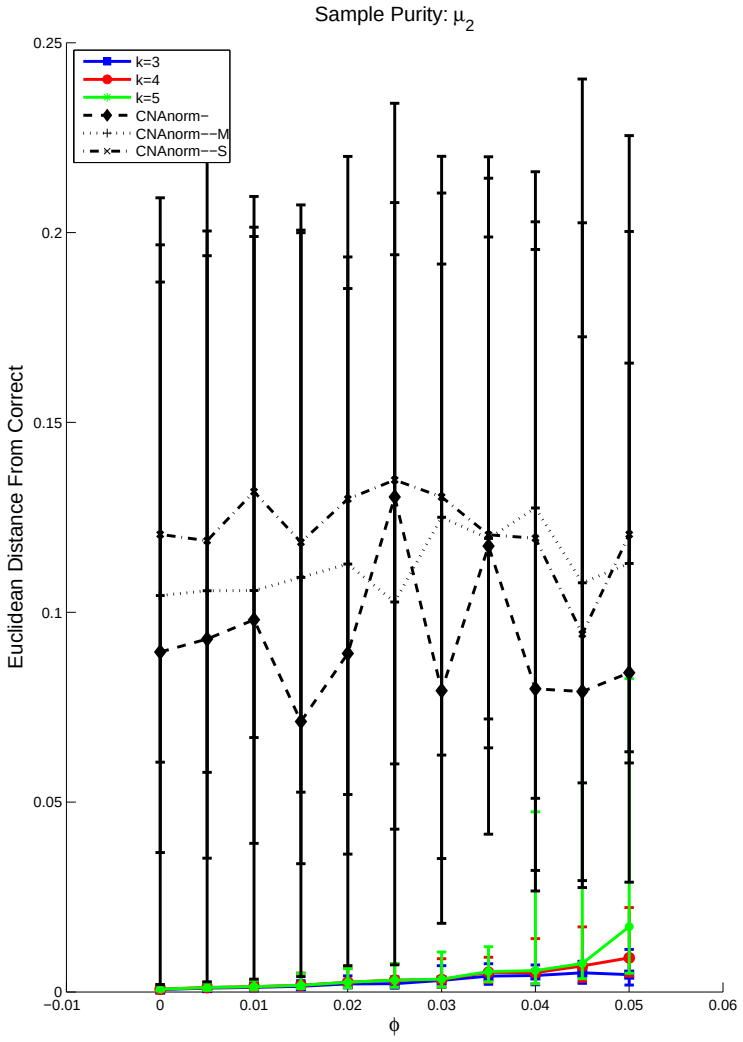


Figure A.8: **Effect of read depth estimation errors on a simulated data.** Simulations have underlying pair $(\mathbf{C}, \mu) \sim \Omega_{39,2,3}$. ϕ is a scaling factor for the variance of Gaussian noise added to each interval. We show results for our algorithm when run with $k = 3, 4, 5$ and for three variations of the CNAnorm algorithm. We show the median of the euclidean distance of the estimated sample purity from the true underlying sample purity (where error bars represent the 25 and 75 percentiles).

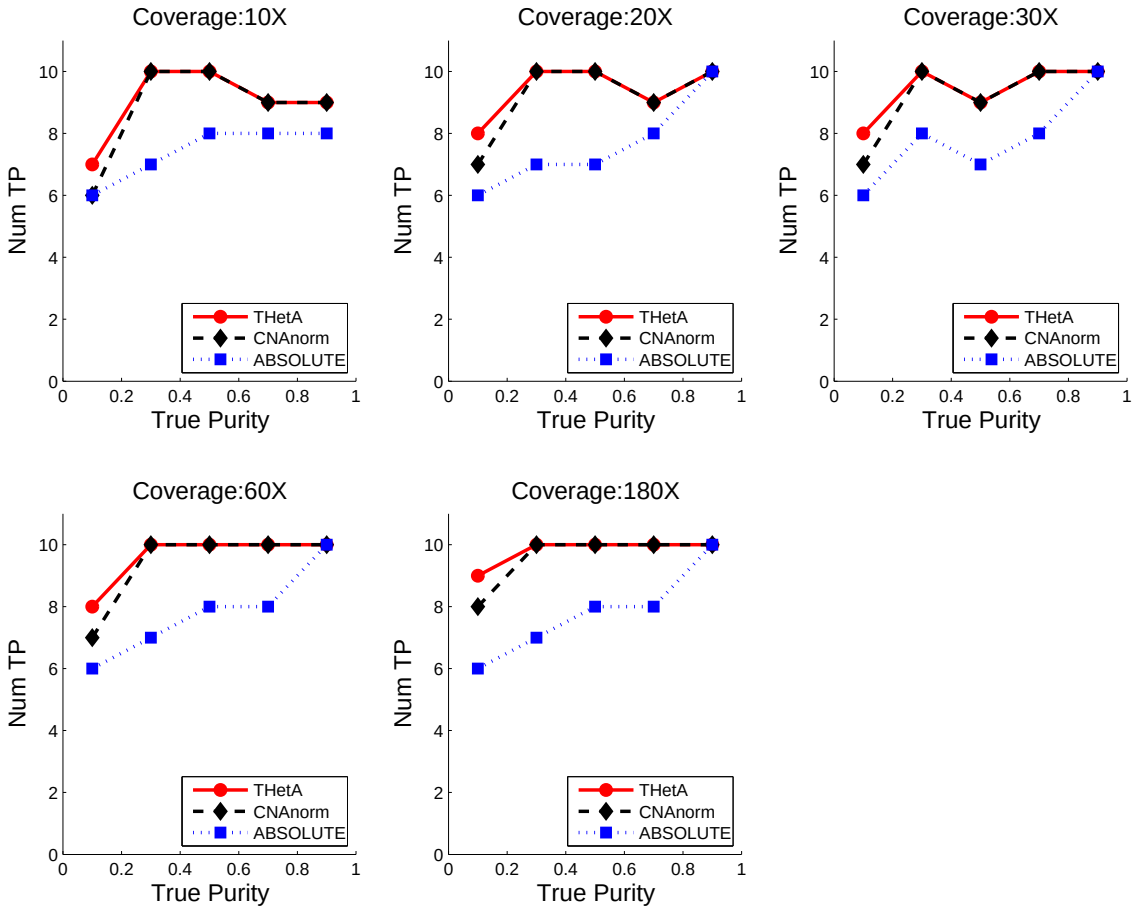


Figure A.9: **Comparison of copy number aberrations predicted by THetA, CNAnorm and ABSOLUTE.** Plots show the number of true copy number aberrations predicted (using 50% reciprocal overlap) across different purity levels.

A.3.1 Copy Number Bounds: $n = 2$

We introduce an optional heuristic that uses the distribution of $\frac{\widehat{r_j}}{\widehat{w_j}}$ over all intervals I_j to partition intervals into two groups: (1) intervals that are either deleted or have an unchanged copy number; and (2) intervals that are potentially amplified. We set different lower and upper bounds depending on which group an interval falls into.

For real sequencing data we initially make the assumption that most of the tumor genome will retain the normal expected copy number. This assumption allows us to further restrict the set of copy number values we consider for each interval. Rather than using an individual global maximum copy number value k for all intervals, we set individual lower and upper bounds for each interval using the following process. Let $\mathbf{r}$ and $\mathbf{w}$ be the observed read depth vectors for the tumor sample and matched normal samples respectively. For each interval I_j we calculate the ratio $\frac{\widehat{r_j}}{\widehat{w_j}}$, and then determine the mean x and standard deviation y for all of these ratios. For all intervals I_j where $\frac{\widehat{r_j}}{\widehat{w_j}} \leq x + by$ we set the lower bound to 0 and the upper bound to the expected normal copy number τ, where $b \in \mathbb{R}^+$ is a user supplied positive value. For all intervals I_j where $\frac{\widehat{r_j}}{\widehat{w_j}} > x + by$ we set the lower bound to $\max(\tau, \lceil \frac{\widehat{r_j}}{\widehat{w_j}} \rceil - 1)$ and the upper bound to $\max(k, \lceil \frac{\widehat{r_j}}{\widehat{w_j}} \rceil + 1)$. Setting b too small increases the the size of the search space, and setting b too large may lead to true amplification being missed. In practice we find that $b = 0.5$ provides a reasonable balance between these tradeoffs in many cases. However, the best value for b does change some depending on the number of copy number variants included in a sample.

In cases where most of the genome does not retain the normal expected copy number, THetA allows for the bounds set using the above heuristic to be rescaled to reflect the expected average ploidy of the sample. Additionally, since THetA returns the set of all maximum likelihood solutions, this set of solutions can contain different sample reconstructions representing different ploidys.

A.3.2 Copy Number Bounds: $n = 3$

We use a heuristic in the $n = 3$ to set bounds on the copy number for each interval that is similar to the $n = 2$ heuristic described above. For each interval in the subset of intervals that are considering for $n = 3$ we use the same lower and upper bounds on copy number estimated using the heuristic for the $n = 2$ case, with the exception that for any interval that previously had a lower bound of 0, we set a new lower bound of 1. This is required to limit the size of the search space, and is a

reasonable assumption when none of the intervals in the subset being considered were predicted to be homozygously deleted by the $n = 2$ algorithm. In simulations we also set an upper bound of $k = 3$ as the genomes do not contain high-level amplifications.

A.3.3 Normal Admixture

Since we expect tumor samples to be relatively pure we add an additional constraint to our model where we require the inferred purity of the sample to be greater than 50%.

A.4 Breast Cancer Sequencing Data

The breast cancer samples analyzed were sequenced using Illumina paired-end technology with read length of 100bp or 108bp. We downloaded the BAM files from the European Genome-phenome Archive (`http://www.ebi.ac.uk/ega/`, accession number EGAD00001000138). The read depth vector is derived from the number of concordant paired reads where each read has a mapping quality ≥ 30.

We set the interval weight vectors $\mathbf{w}$ equal to to the read depth vector for the matched normal sample. This is a reasonable weight vector use; for example, if $\frac{w_i}{w_j} = 2$ we expect the number of reads originating from a single copy of i^{th} interval to be as twice as the number of reads originating from a single copy of j^{th} interval. That is, $\frac{r_i}{r_j} \approx 2$, for a large enough number of reads and intervals i and j are not amplified or deleted in the tumor sample.

A.4.1 Sample PD4120a

In this section we present some additional processing details, results and further analysis of the sample PD4120a from [114] not included in the main text.

$n = 2$ analysis

We present here the results of when we run our algorithm with $n = 2$ on sample PD4120a. We use all genomic intervals derived following BIC-Seq segmentation ($\lambda = 100$) after removal of all intervals less then 50 kb in length. We infer that the sample contains 34.3% normal cells and 65.7% tumor cells, compared to the 30% and 70% respectively reported by [114]. Figure A.10 shows the complete set of copy number aberrations predicted using this analysis.

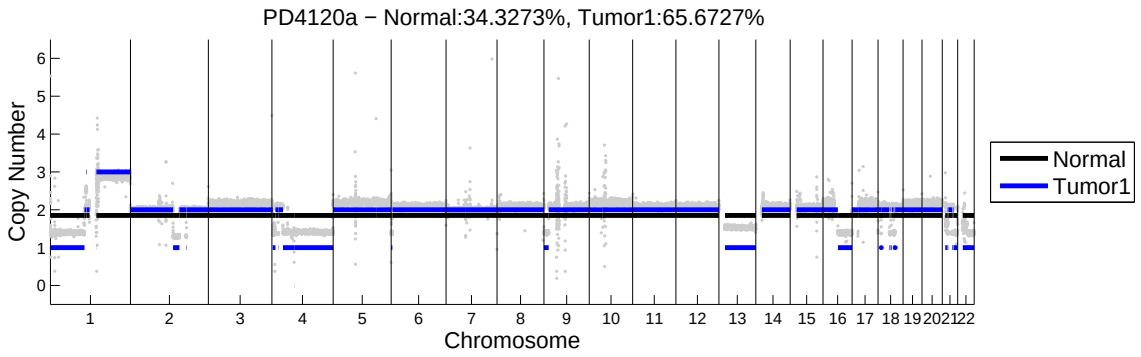


Figure A.10: **Results of running THetA on breast cancer sample PD4120a when only considering a single tumor population.** We identify several aberrations including the trisomy of 1q, and deletions of 1p, 4q, 13q, 16q and 22q. Copy numbers for the normal genome are in black and copy numbers for the tumor genome are in blue. Read depth ratios (gray) are over 50 kb bins across the entire genome.

$n = 3$ analysis

We performed analysis using our algorithm with $n = 3$ on a subset of intervals in sample PD4120a. We select a subset of the intervals used in the $n = 2$ analysis by choosing: (1) all chromosomes that BIC-Seq partitioned into a single interval; (2) all intervals > 22 Mb that were reported as abnormal copy number ($\neq 2$) in the $n = 2$ analysis. We use only the longest such interval per chromosome if the number of such intervals is large. We later add all intervals from chromosome 22 to this subset in order to resolve differences between our results and those presented in [114]. Since the results for both subsets are extremely similar, we present in the main text the results for the larger subset (including chromosome 22). Here we present the similar results obtained when using the original smaller set of intervals.

Our algorithm estimates a normal admixture of 27.96% and two tumor populations comprising of 62.19% and 9.84% of the cells in the sample (Figure A.11). Again, we recover the fully clonal loss of 4q, and also estimate loss of 16q and part of 22q as clonal aberrations – the later [114] reports as subclonal. We also estimate the trisomy of 1q as subclonal - which [114] reports as clonal. We find a subclonal loss of 13q in 62.19% of the total cells in the sample. In comparison, [114] report a similar subclonal deletion in 47% of all cells in the sample (68% of the tumors cells in a 70% pure sample). Lastly, we identify subclonal deletions of chromosomes 8,11,12,14,and 15 in 9.84% of cells in the sample. [114] report aberrations for these chromosomes in 9.8% of the cells in the sample (14% of a 70% pure sample). However they claim that these aberrations are chromosomal loss after

a genome duplication, a scenario that is not found by our analysis.

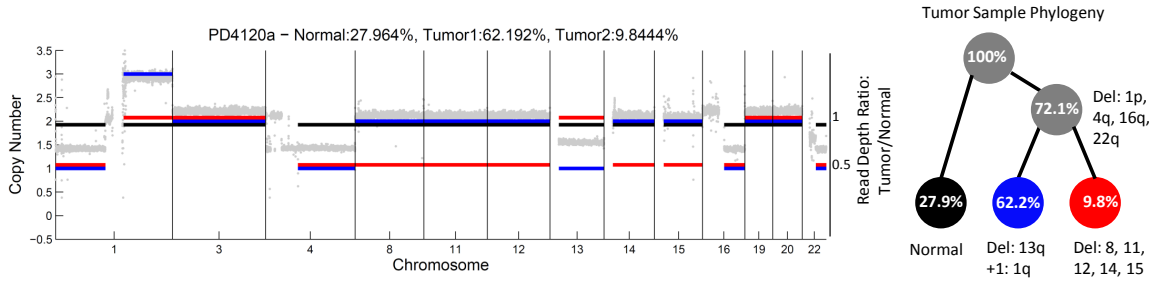


Figure A.11: **Analysis of copy number aberrations in breast tumor PD4120a when considering a smaller subset of intervals than presented in the main text.** Results shown are for $n = 3$. (Left) We show a copy number profile of of the normal population (black), dominant (clonal) tumor population (blue), subclonal tumor population (red) and read depth ratios (gray). (Right) A reconstruction of the tumor phylogeny with the estimated fraction of cells in each subpopulation. We identify copy number changes in chromosomes 1p, 4q, 16q, and 22q as clonal, the deletion of 13q and trisomy of 1q as part of the 62.2% subclonal population and deletions of chromosomes 8, 11,12,14,15 as part of the 9.8% subclonal population.

Additionally, we use estimates of normal admixture to correct read depth ratios over all chromosomes in order to predict subclonal events in chromosomes that were not used directly in the $n = 3$ analysis. We can predict the existence of a subclonal event if a peak in the read depth distribution occurs at a value that is not near a multiple of 0.5. Using this method we are able to predict the existence of subclonal deletions of chromosomes 2, 4p, 6, 7, 9, 18 and 21 (Figure A.12). The deletions in chromosomes 2 and 7 appear in a similar fraction of the sample, while the deletions in chromosomes 4p, 6, 7, 9, 18 and 21 appear in a similar fraction of the chromosomes – perhaps as part of the same subclonal population as the deletions of chromosomes 8,11,12,14, and 15 predicted directly by THetA.

Comparisons between Our Predictions and [114]

In this section we provide additional details of differences between our predictions and those presented in [114]. One of the techniques we use to validate our predictions is to use corrected read depth ratios in 50 kb intervals over entire genomes or specific genomic intervals. Figure A.13(B) shows that such binning provides a more discerning view of read depth ratios than considering ratios over individual SNP positions (Figure A.13(A)).

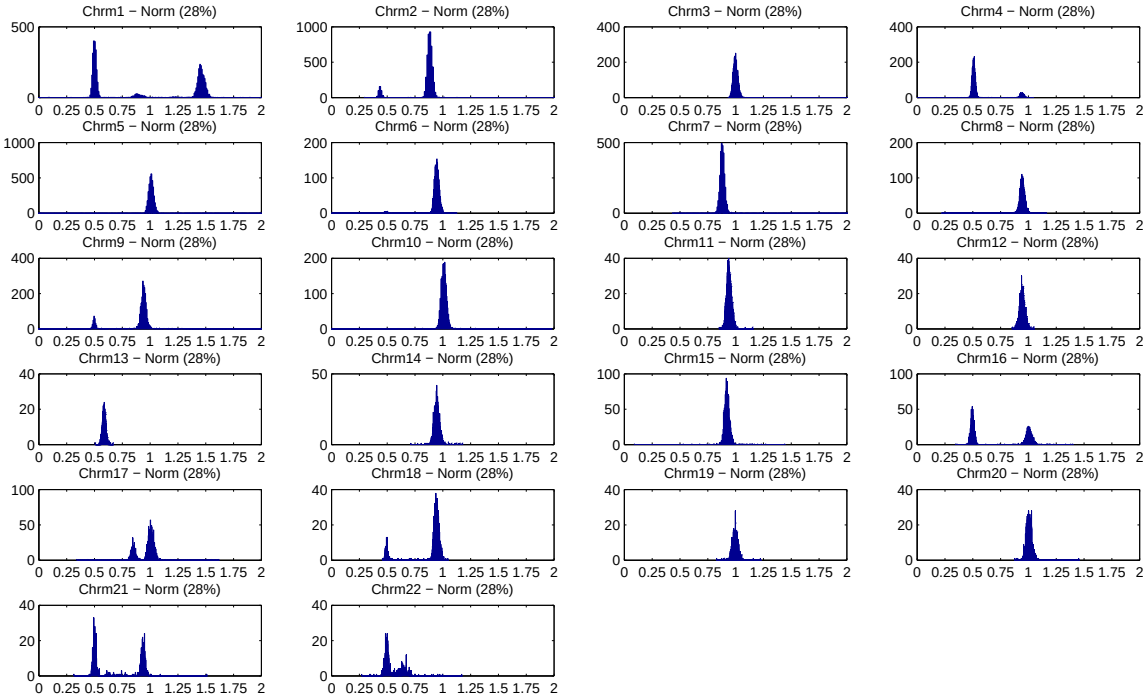


Figure A.12: **The distributions of read depth ratios over 50 kb intervals in the 22 autosomes of breast cancer sample PD4120a.** Results are plotted after centering and correction for 28% normal admixture for each using a simple linear scaling. Peaks that occur not at a ratio that is not divisible by 0.5 indicate the presence of a subclonal aberration.

Chromosome 1q We performed further analysis of chromosome 1q for sample PD4120a. [114] indicated that the amplification of 1q to 3 copies is a clonal mutation (occurring in 70% cells), whereas our analysis finds that 1q is a subclonal mutation (occurring in 62% cells). We partitioned the genome into 50 kb intervals and determined read depth vectors over all intervals for both the tumor sample and matched normal sample. We determined read depth ratios by normalizing the total number of reads to be the same for both the tumor and normal sample to be equal. Both our analysis and that done by [114] agree that chromosome 3 does not contain large copy number aberrations. We therefore center the ratios so that the average ratio in chromosome 3 is set to a ratio of 1 (Figure A.13(A)). We also look at how the distribution changes when we correct for 28% normal admixture (as estimated by our model) and 30% normal admixture (as predicted in [114]). In both of these cases, we see that the amplified intervals (which are those intervals in chromosome 1q) appear to be amplified in a smaller proportion of the cells than the clonal deletions (Figure A.13(B) and Main Text Figure 3B). We analyze this data further by focusing on the intervals contained in the amplified portion of 1q and testing how far the corrected ratios are from the expected ratio of 1.5 (for an amplification to copy 3). We find that average ratio for intervals in 1q with normal admixture of 27.96% is 1.456 and with normal admixture of 30.0% is 1.47). While the later estimate is closer to the expected value of 1.5 (given copy number 3), the observed distributions have p-values of $4.7E-80$ and $2.9E-39$ using a a t-test when compared to the expected mean value of 1.5 – an indication that this aberration may indeed be subclonal.

We also perform further statistical analysis by comparing the read depth ratios for 1q to other intervals we predict to be clonally deleted (1p, 4q, and 16q). For all 4 intervals we use a linear scaling (see Main Text Methods) to correct read depth ratios for a range of possible *aberration fractions* – the fraction of the sample containing the aberration. For each aberration fraction and interval pair, we calculate a Z-score equal to $\frac{x-\mu}{\sigma}$ where x is the expected corrected ratio (0.5 for deletions, 1.5 for amplification of a single copy) for the interval and μ and σ are the mean and standard deviation of the corrected ratios. For all pairwise combinations such interval and aberrations fraction pairs, we calculate a pairwise Z-score by summing the absolute value of the corresponding Z-scores. Lower pairwise Z-scores indicate a better fit to the data. We investiage all pairwise Z-scores in Figure A.14 and find evidence that the aberrations in 1p, 4q and 16q are likely to occur in similar fractions of the sample. Whereas, the amplification in 1q appears to occur in a smaller fraction of the sample than the deletions in 1p, 4q and 16q. This evidence supports our hypothesis that the amplification

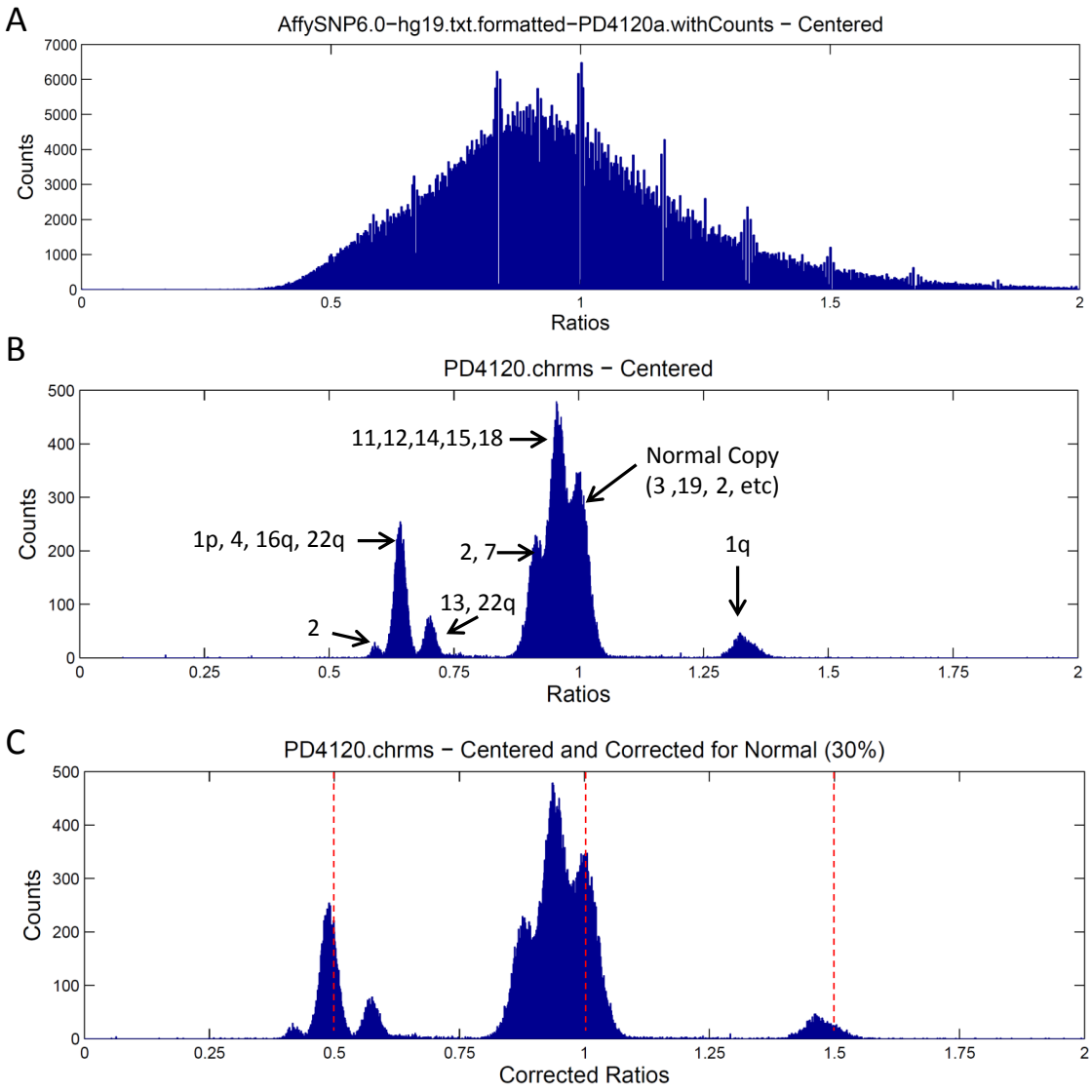


Figure A.13: **Distribution of all read depth ratios for sample PD4120a.** All plots are centered so that the mean ratio for Chromsome 3 (which doesn't contain larger copy number aberrations) is set to 1. **A.** Read depth ratios determined by counting the number of reads with an alignment including known germline SNP positions. **B.** Read depth ratios in 50 kb bins. **C.** Read depth ratios after correction for 30% normal admixture using a simple linear scaling. The peak for chromosome 1q is less than 1.5, as would be expected if this aberration were clonal.

of 1q was a subclonal event.

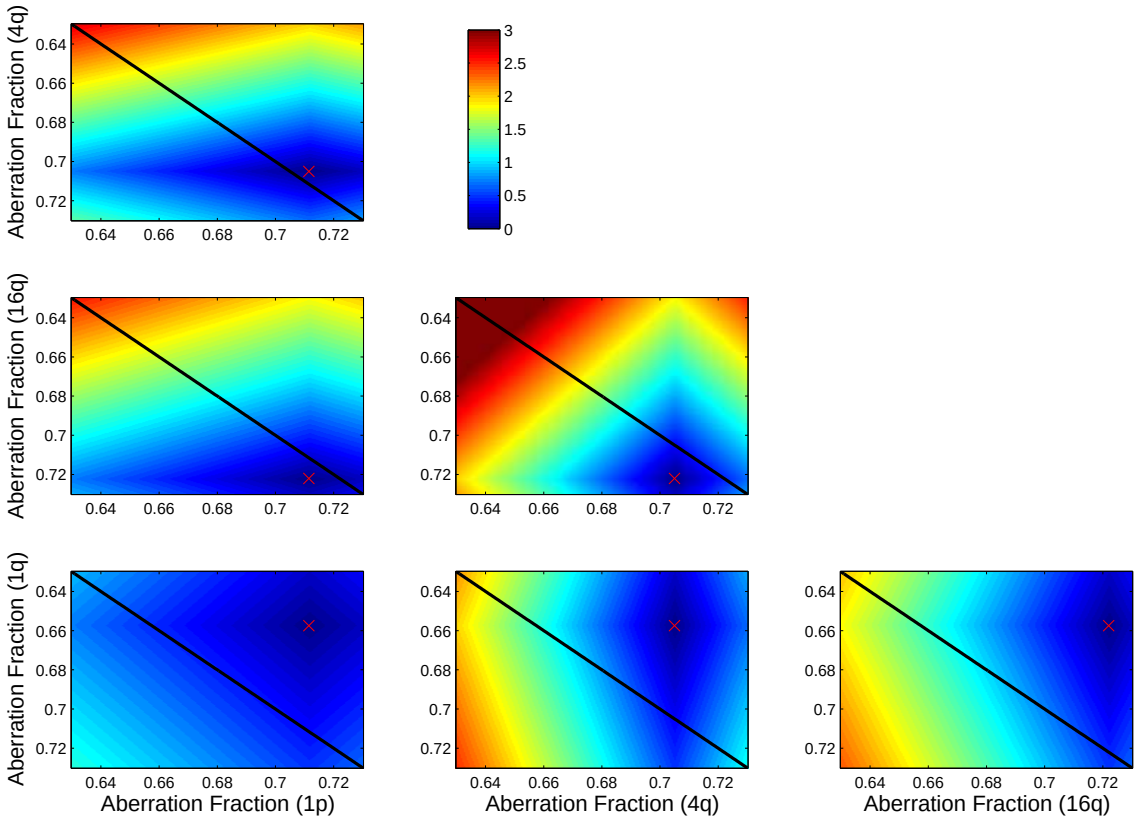


Figure A.14: **Pairwise Z-scores for corrected read depth ratios in aberrations occurring in sample PD4120a.** A lower intensity indicates a lower Z-score, and an overall better fit with the data. The black diagonal marks equal aberration fractions between the intervals under consideration. The red X marks the pair of aberration fractions with the lowest pairwise Z-score. The closer that the red X is to the black diagonal, the more likely that the aberrations occur in a similar fraction of the sample and hence are in the same subclonal/clonal population.

Chromosome 22 We performed further analysis of Chm22 in sample PD4120a which our analysis in the main paper indicates that 22q contains a clonal deletion as well as a subclonal deletion. [114] posit that a translocation between Chr1 and Chr22 is subclonally deleted. We used the algorithm GASV [150] to cluster discordant pairs where one read aligned to Chr1 and the other read aligned to Chr22. We find a cluster of 41 discordant supporting a single non-reciprocal translocation between Chr1p21 and Chr22q12. A visual representation [159] of these discordant read pairs is located in Figure A.15(B). This finding supports another potential sequence of events, where the non-reciprocal translocation between Chr1 and Chr22 results in clonal deletion of parts of 1p and 22q and a dicentric

chromosome. This is later followed by a subclonal deletion of a part of 22q which was not previously deleted in the translocation (Figure A.15(A)). Dicentric chromosomes are known to be unstable, and so this deletion might occur on this chromosome.

Analysis of PD4120a using ASCAT

We ran the ASCAT algorithm [164] on synthetic SNP array data we created from sequencing data for sample PD4120a [114]. For the 907,693 SNP locations on chromosomes 1 - 22 queried by the Affymetrix 6.0 SNP array, we created the two types of input used by ASCAT: 1) LogR values; and 2) B allele frequency (BAF). For each SNP position we counted the number of reads containing different alleles for both the tumor and matched normal samples. BAF values were calculated as the fraction of such reads containing the variant allele. LogR ratios for the were calculated by using quantile normalization [20] over the read counts for the matched tumor and normal sample and then taking the log_2 of the ratio of number of reads containing the SNP location from the tumor sample over the number from the normal genome. By default, ASCAT assumes that LogR values have been scaled by a platform dependent parameter $\gamma = 0.55$, so we scaled all LogR values by this γ. LogR ratios for the normal sample were all set to 0. ASCAT inferred sample purity of 66% using this synthetic SNP data as input, a value similar to our estimate when $n = 2$ of 65.6%. The copy number aberrations predicted by ASCAT agree with both our analysis and that in [114] on several aberrations like the trisomy of 1q and the monosomy of 4q, but differs by predicting that chromosomes 17, 18, 19 and 20 are amplified.

A.4.2 Sample PD4088a

In this section we provide analysis of sample PD4088a, not contained in the main text. [114] report that sample PD4088a contains little subclonal copy number variation, (although they do not provide tumor purity estimates or copy number aberrations) making this sample a good candidate to analyze with our efficient method for inferring a clonal population. Following BIC-Seq segmentation ($\lambda = 200$), we find that chromosome 17 is extensively fragmented containing 25% of the intervals. To avoid overfitting to this chromosome, we remove all chromosome 17 intervals from consideration. Our algorithm run with $n = 2$ reports that PD4088a contains 41% normal cells and 59% tumor cells and identifies a loss of chromosomes 3, 10q, 11q, 18q, and 22q in the tumor population (Figure A.16(A)). Setting $n = 3$ and selecting a subset of intervals using the rules discussed in the

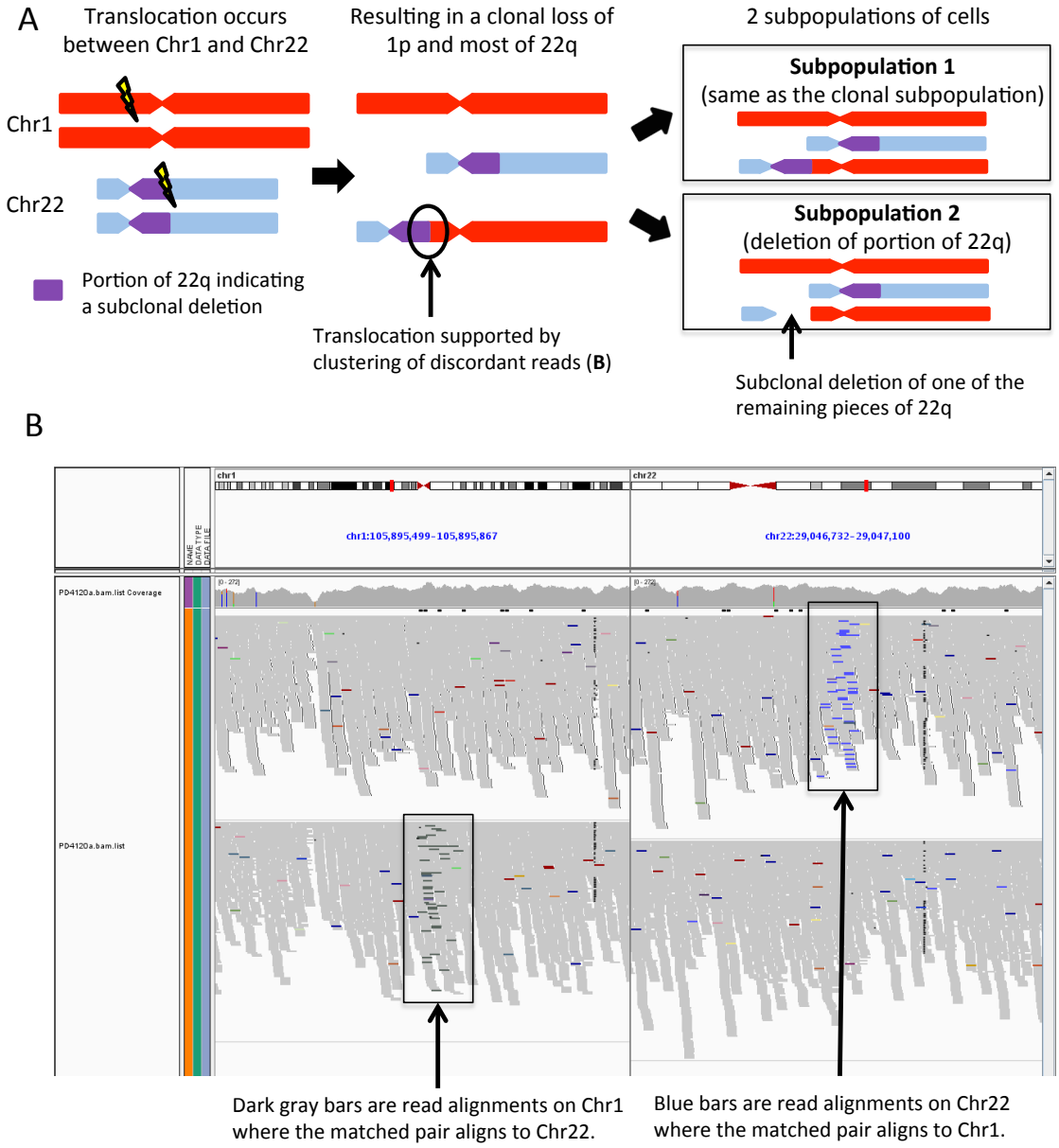


Figure A.15: **Additional analysis of rearrangements on chromosome 22q in breast cancer sample PD4120a.A.** A potential sequence of events describing the clonal and subclonal deletions observed in 22q for sample PD4120a. A translocation (whose position is indicated by yellow markers) occurs between Chr1 (red) and Chr22 (blue) leading to a clonal loss of 1p and a portion of 22q. Then one of the two remaining copies of a portion of 22q (purple) is later subclonally deleted. **B.** Visual display of the discordant read pairs that support the translocation between Chr1 and Chr22.

main text (with an interval length lower bound of 40 Mb) also indicates that this sample is mostly clonal (Figure A.16(B)). This analysis indicates a normal admixture of 40.2% with two subclonal population comprising 58.6% and 1.2% of cells in the sample. Aberrations reported in the major subclonal population including a deletion of chromosomes 3, 10q, 11q, and 22q are the same as those reported in the clonal population in our analysis when $n = 2$. When we compare the $n = 2$ and $n = 3$ solutions using our version of the BIC model selection we choose the $n = 2$ solution indicating that this sample is mostly clonal.

Read depth ratio analysis also supports our conclusion that this sample is mainly clonal. Figure A.16(C) shows the original distribution of tumor/normal read count ratios over 50kb bins across the genome. Figure A.16(D) shows this distribution after correction for normal admixture using our estimates of the $n = 2$ analysis. The clearly defined peaks at corrected ratios of 0.5, 1 and to some extent at 1.5 indicated that this sample is indeed mostly clonal (at least with respect to copy number variants).

A.4.3 Sample PD4115a

In this section we present some additional processing details, results and further analysis of the sample PD4115a from [114] not included in the main text.

$n = 2$ analysis

We present here the results when we run our algorithm with $n = 2$ on sample PD4115a. We use all genomic intervals derived from BIC-Seq segmentation ($\lambda = 200$) after removal of all intervals 50 kb in length. On initial analysis we infer that the sample contains a high normal admixture and most copy numbers returned were equal to 1. Under our model, an equally likely solution is obtained by adding 1 to all copy numbers, thus translating the mode to copy number 2, and redetermining normal admixture. Using that sequence of steps, our algorithm reports that PD4115a contains 32.33% normal cells and 67.67% tumor cells (Figure A.17(A)). We identify several amplifications including 1q (+3), 7q (+1) and part of 8q (+5). We also identify multiple deletions including 9q, 11q, and 14q. Figure A.17(B) shows the distribution of read depth ratios over the entire genome after centering using chromosome 20 (determined by the above analysis to have normal copy number of 2). Figure A.17(C) shows the same distribution of read depth ratios but includes a correction for the estimate normal admixture of 32.33%. Other than the normal copy peak at a corrected ratio of

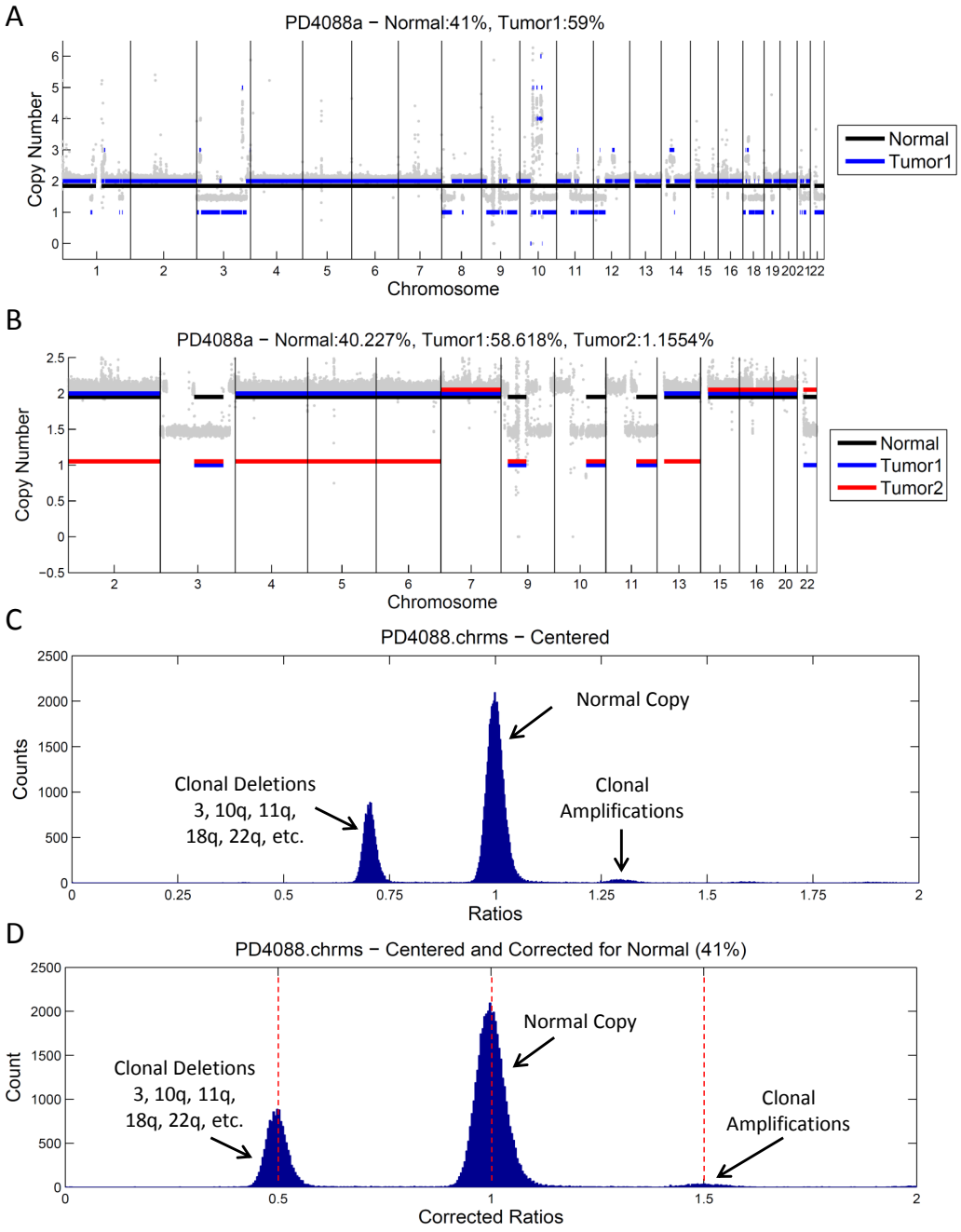


Figure A.16: **Analysis of the ∼40X coverage breast tumor PD4088a. A.** Read depth ratios (gray) and the inferred copy number aberrations by our algorithm when $n = 2$ including the normal population (black), and clonal tumor population (blue). **B.** Read depth ratios (gray) and the inferred copy number aberrations by our algorithm when $n = 3$ including the normal population (black), major tumor population (blue) and subclonal population (red). **C.** Read depth ratios in 50kb intervals after centering so chromosomes 4,5,6 and 7 have a mean of 1. **D.** Read depth ratios in 50kb intervals after centering so chromosomes 4,5,6 and 7 have a mean of 1 and correcting for 41% normal admixture using a simple linear scaling. The only visible peaks fall near to expected corrected ratios (0.5, 1, 1.5), indicating that this sample is mostly clonal (with respect to copy number variants).

1, these corrected ratios do no align well to increments of 0.5, as would be expected if the sample was clonal. This supports our analysis that this sample contains several subclonal populations.

$n = 3$ analysis

We find that PD4115a contains many apparent copy number aberrations with the segmentation used for $n = 2$ containing 102 intervals (compared to only 69 intervals for sample PD4120a above). In addition, this sample also includes several highly amplified regions, and no chromosome was segmented into a single interval. Thus, we ran THetA for $n = 3$ on subset of the longest intervals in the BIC-Seq partition (since no chromosome was partitioned into a single interval). Due to the greater amount of fragmentation of this genome, we used a parameter of 0.4 for our heuristic for setting lower and upper bounds. Our analysis indicated that the sample contained two distinct subclonal populations in near similar proportions. We compare these subclonal populations using Z-scores in the same manner as described in the previous section. The only difference is that we compare a deleted interval in chromosome 3p simultaneously to the deleted intervals in 3q, 4q and 5q (Figure A.18). We find that the deletion in 3p does appear to be deleted in a different fraction of cells than the other deletions.

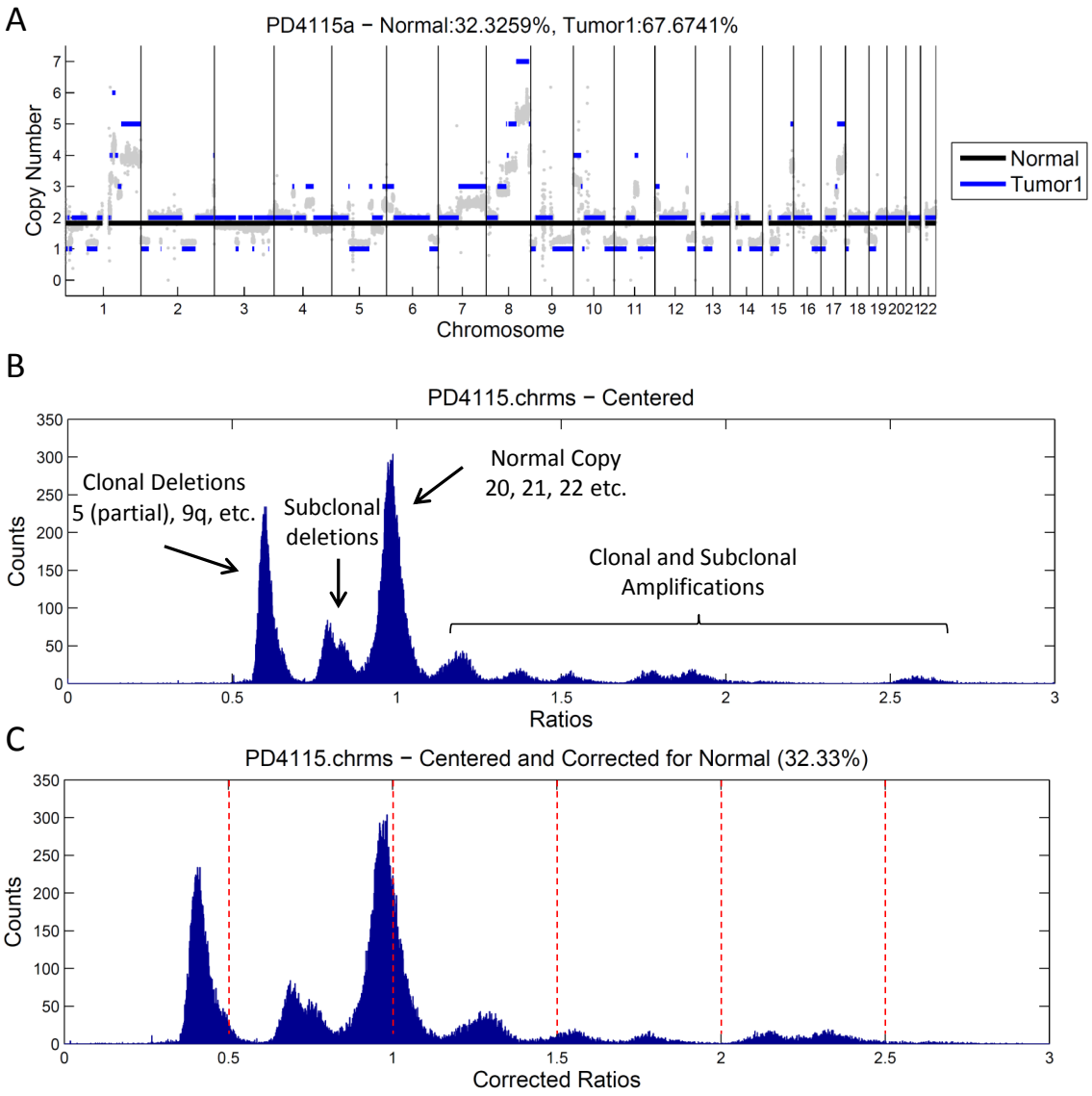


Figure A.17: **Distribution of read depth ratios in 50 kb intervals for breast cancer sample PD4115a.** (A) Read depth ratios when distribution is centered so that the mean ratio in Chromosome 20 is set to a ratio of 1. (B) Ratios after centering, and correction for 32.33% normal admixture using a simple linear scaling. No peaks fall near expected corrected ratios - supporting our analysis that this sample contains subclonal aberrations. (C) Inferred copy number aberrations by our algorithm when $n = 2$ including the normal population (black), clonal tumor population (blue), and read depth ratios (gray).

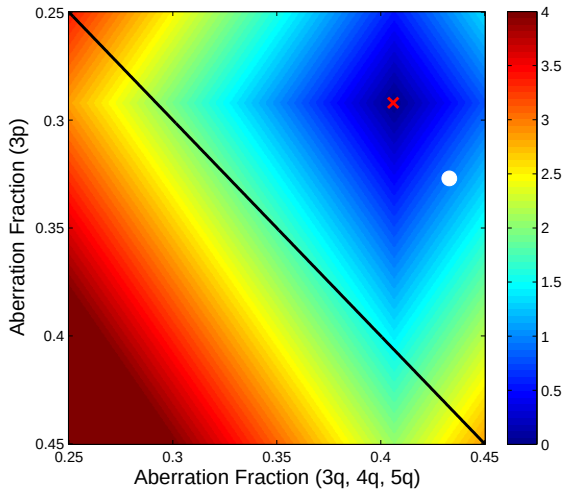


Figure A.18: **Pairwise Z-scores for corrected read depth ratios between different subclonal deletions in breast cancer sample PD4115a.** A lower intensity indicates a lower Z-score, and an overall better fit with the data. The black diagonal marks equal aberration fractions between the intervals under consideration. The red X marks the pair of aberration fractions with the lowest pairwise Z-score. The closer that the red X is to the black diagonal, the more likely that the aberrations occur in a similar fraction of the sample and hence are in the same subclonal/clonal population. Here the red X is quite far from the black line - an indication that these deletions do occur in different subclonal populations. The white dot indicates the aberration fractions estimated by our algorithm and is near (within $\sim$1) the optimal pairwise Z-score within for this subset of deletions.

Appendix B

Improved Approaches to Quantifying Intra-tumor Heterogeneity

B.1 Proofs Omitted from the Main Text

Corollary 3.2.1. *Suppose* $\mathbf{C} \in \mathcal{C}_{m,n,k}$. *If there exists an* $i \in \{1, \ldots, m-1\}$ *such that for all* $t \in \{2, \ldots, n\}, c_{i,t} \geq c_{i+1,t}$ *and there exists a* $t \in \{2, \ldots, n\}$ *such that* $c_{i,t} > c_{i+1,t}$, *then* $\Phi(\mathbf{C}) = \emptyset$.

Proof. Now, we will proceed by contradiction. Assume the above conditions hold and $\Phi(\mathbf{C}) \neq \emptyset$. Then there exists some μ such that $(\mathbf{C}\mu)_i \leq (\mathbf{C}\mu)_{i+1}$. Since, $\mu_k > 0$ for all $k \in \{2, \ldots, n\}$ this implies that $c_{i,s}\mu_s \geq c_{j,s}\mu_s$ for all $s \in \{2, \ldots, n\}$ and $c_{i,t}\mu_t > c_{j,t}\mu_t$ for some $t \in \{2, \ldots, n\}$. However, this implies $(\mathbf{C}\mu)_i > (\mathbf{C}\mu)_{i+1}$, a contradiction. Hence, $\Phi(\mathbf{C}) = \emptyset$. □

Theorem 3.2.2. *Let* $\mathbf{C} = [c_{i,j}]$ *be an interval count matrix.* $\mathcal{L}_{\mathbf{r}}(\mathbf{C}, \mu)$ *is a convex function of* $c_{i,j}$.

In order to prove Theorem 3.2.2 we first need to build up some notation. We will then prove Theorem 3.2.2 in the situation where $n = 2$, which is then easily generalizable to $n > 2$. We start with the following real valued function defined when $n = 2$. Given a read depth vector $\mathbf{r} = (r_1, \ldots, r_{m+1})$, and a pair $(\mathbf{C}, \mu) \in \Omega_{m,2,k}$ we define $\mathcal{L}_{\mathbf{r},\mathbf{C},\mu} : [0, \infty) \longrightarrow \mathbb{R}$ such that $\mathcal{L}_{\mathbf{r},\mathbf{C},\mu}(x) = \mathcal{L}_{\mathbf{r}}(\widehat{\mathbf{X}\mu} | \mathbf{X} = [\mathbf{C}; 2, x])$. Here $\widehat{\mathbf{X}\mu} = \frac{\mathbf{X}\mu}{|\mathbf{X}\mu|_1}$ is just the normalized version of $\mathbf{X}\mu$. To prove Theorem 3.2.2 for the

case when $n = 2$, we just need to show that $\mathcal{L}_{\mathbf{r}}(\widehat{\mathbf{X}\mu}|\mathbf{X} = [\mathbf{C}; 2, x])$ is convex in x. Before we do so, we first need the following lemmas.

Lemma B.1.1. *$\mathcal{L}_{\mathbf{r}}(\mathbf{p}) = -\sum_{j=1}^{m} r_j \log(p_j) + \alpha$ is separable convex for $\mathbf{p} \in \Delta_{m-1}$.*

Proof. See the supplement of [118]. □

Lemma B.1.2. *Let $(\mathbf{C}, \mu) \in \Omega_{m,n,k}$ and $[a, b]$ be a non-negative real valued interval. The set $\mathcal{X} = \{\widehat{\mathbf{X}\mu}|\mathbf{X} = [\mathbf{C}; 2, x], x \in [a, b]\}$ is a convex subset of Δ_m.*

Proof. We show that every element in $\mathcal{X}$ can be written as a convex combination of two particular elements of $\mathcal{X}$ and therefore defines a line in $\mathbf{R}^{m+1}$ (embedded in Δ_m) which is by definition a convex set. Let $\mathbf{A} \in \mathcal{S}$ such that $\mathbf{A} = [\mathbf{C}; 2, a]$ and let $\mathbf{B} \in \mathcal{X}$ such that $\mathbf{B} = [\mathbf{C}; 2, b]$. Notice that for any $x \in [a, b]$ there exists some λ where $0 \leq \lambda \leq 1$ such that $x = \lambda a + (1 - \lambda)b$. Therefore, any $\mathbf{X} \in \mathcal{X}$ can be written as $\mathbf{X} = \lambda\mathbf{A} + (1 - \lambda)\mathbf{B}$ for some λ where $0 \leq \lambda \leq 1$. We note the following two observations which can easily be verified for any $\mathbf{X} \in \mathcal{X}$ and corresponding λ.

(1) $|\mathbf{X}\mu|_1 = |(\lambda\mathbf{A} + (1 - \lambda)\mathbf{B})\mu|_1 = |\lambda\mathbf{A}\mu + (1 - \lambda)\mathbf{B}\mu|_1 = \lambda|\mathbf{A}\mu|_1 + (1 - \lambda)|\mathbf{B}\mu|_1$.
(2) $(\mathbf{X}\mu)_i = \lambda(\mathbf{A}\mu)_i + (1 - \lambda)(\mathbf{B}\mu)_i$ for all $i \in \{1, \ldots, m + 1\}$.

We now show that for any $\mathbf{X} \in \mathcal{X}$ there exists some $\alpha = (\alpha_1, \alpha_2) \in \Delta_1$ such that $\widehat{\mathbf{X}\mu} = \alpha_1\widehat{\mathbf{A}\mu} + \alpha_2\widehat{\mathbf{B}\mu}$. Set $\alpha_1 = \frac{\lambda|\mathbf{A}\mu|_1}{\lambda|\mathbf{A}\mu|_1 + (1-\lambda)|\mathbf{B}\mu|_1}$ and $\alpha_2 = \frac{(1-\lambda)|\mathbf{B}\mu|_1}{\lambda|\mathbf{A}\mu|_1 + (1-\lambda)|\mathbf{B}\mu|_1}$. By definition $\alpha_1 + \alpha_2 = 1$ and $\alpha_1, \alpha_2 \geq 0$, so $\alpha \in \Delta_1$. We now show that $\widehat{\mathbf{X}\mu} = \alpha_1\widehat{\mathbf{A}\mu} + \alpha_2\widehat{\mathbf{B}\mu}$. For each $i \in \{1, \ldots, m+1\}$ we compute the i^{th} entry:

$$
\begin{aligned}
(\alpha_1\widehat{\mathbf{A}\mu} + \alpha_2\widehat{\mathbf{B}\mu})_i &= \alpha_1 \frac{(\mathbf{A}\mu)_i}{|\mathbf{A}\mu|_1} + \alpha_2 \frac{(\mathbf{B}\mu)_i}{|\mathbf{B}\mu|_1} \\
&= \frac{\lambda|\mathbf{A}\mu|_1}{\lambda|\mathbf{A}\mu|_1 + (1-\lambda)|\mathbf{B}\mu|_1} \frac{(\mathbf{A}\mu)_i}{|\mathbf{A}\mu|_1} + \\
&\quad \frac{(1-\lambda)|\mathbf{B}\mu|_1}{\lambda|\mathbf{A}\mu|_1 + (1-\lambda)|\mathbf{B}\mu|_1} \frac{(\mathbf{B}\mu)_i}{|\mathbf{B}\mu|_1} \\
&= \frac{\lambda(\mathbf{A}\mu)_i + (1-\lambda)(\mathbf{B}\mu)_i}{\lambda|\mathbf{A}\mu|_1 + (1-\lambda)|\mathbf{B}\mu|_1} \\
&= \frac{(\mathbf{X}\mu)_i}{|\mathbf{X}\mu|_1} \text{ (Using both of the above observations.)} \\
&= (\widehat{\mathbf{X}\mu})_i
\end{aligned}
$$

Hence, we see that $\widehat{\mathbf{X}\mu} = \alpha_1\widehat{\mathbf{A}\mu} + \alpha_2\widehat{\mathbf{B}\mu}$, and is therefore any $\widehat{\mathbf{X}\mu} \in \mathcal{X}$ is a convex combination of $\widehat{\mathbf{A}\mu}, \widehat{\mathbf{B}\mu} \in \mathcal{X} \subseteq \Delta_m$. And therefore $\mathcal{X}$ must be a convex subset of Δ_m. □

We now can prove Theorem 3.2.2.

Proof. We start by considering the case where $n = 2$. Lemma B.1.2 tells us that for a fixed $(\mathbf{C}, \mu) \in \Omega_{m,n,k}$ and closed positive real valued interval $[a, b]$ the set $\mathcal{X} = \{\widehat{\mathbf{X}\mu} | \mathbf{X} = [\mathbf{C}; 2, x], x \in [a, b]\}$ is a convex subset of Δ_m. Notice that $[0, \infty) = \cup_{i=1}^{\infty}[0, i]$ is the union of a non-decreasing sequence of convex intervals. Let $\mathcal{X}_i = \{\widehat{\mathbf{X}\mu} | \mathbf{X} = [\mathbf{C}; 2, x], x \in [0, i]\}$ and $\mathcal{X} = \cup_{i=1}^{\infty}\mathcal{X}_i$. From the proof of Lemma B.1.2, it is clear that $\mathcal{X}_i \subset \mathcal{X}_{i+1}$ for all $i \geq 1$. Hence, each $\mathcal{X}_i$ is a non-decreasing sequence of convex subsets of Δ_m and therefore $\mathcal{X}$ is a convex subset of Δ_m where $\mathcal{X} = \{\widehat{\mathbf{X}\mu} \mid \mathbf{X} = [\mathbf{C}; 2, x], x \in [0, \infty)\}$.

Since $\mathcal{X}$ is a convex subset of Δ_m we can apply the result from Lemma B.1.1 to prove that $\mathcal{L}_{\mathbf{r}}(\mathbf{p})$ is separable convex for $\mathbf{p} \in \mathcal{X}$. Since $\mathcal{L}_{\mathbf{r},\mathbf{C},\mu}(x) = \mathcal{L}_{\mathbf{r}}(\widehat{\mathbf{X}\mu} \mid \mathbf{X} = [\mathbf{C}; 2, x])$ there is a one-to-one correspondence between x and $\mathbf{p} \in \mathcal{X}$, we have shown that $\mathcal{L}_{\mathbf{r},\mathbf{C},\mu}(x)$ is convex in x.

We have therefore shown that $\mathcal{L}_{\mathbf{r}}(\mathbf{C}, \mu)$ is convex in $c_{i,j}$ when $n = 2$. The proof easily extends to the case when $n > 2$ by determining an appropriate pair $(\mathbf{X}', \mu')$ where $\mathbf{X}' = (\mathbf{x}_1, \mathbf{c}_j) \in \mathbf{R}^{m,2}$ and $\mu' = (1 - \mu_j, \mu_j)$ and $\widehat{\mathbf{X}'\mu'} = \widehat{\mathbf{C}\mu}$ and the proof from the case of $n = 2$ can be directly applied. □

B.2 Using a Graph to Enumerate $\mathcal{S}_{m,n,k}$

In this section we provide further details and pseudocode on our algorithm for using a graph to enumerate $\mathcal{S}_{m,n,k}$.

The algorithm depends on being able to calculate and efficiently union the $\mu Set\,(v, w)$, i.e the set of values for μ for which $v\mu \leq w\mu$. In the case where $n = 3$, this set is defined by the single variable, $\frac{\mu_2}{\mu_3}$. In particular, in the case where $v_2 > w_2$, the upper bound on $\frac{\mu_2}{\mu_3}$ is $\frac{w_3 - v_3}{w_2 - v_2}$. Likewise, in the case where $v_2 < w_2$, the lower bound on $\frac{\mu_2}{\mu_3}$ is $\frac{w_3 - v_3}{w_2 - v_2}$. The case where $v_2 = w_2$ doesn't restrict the values of $\frac{\mu_2}{\mu_3}$.

B.3 Interval Selection

In this section we discuss how interval selection is done during the first step of our two-step procedure for different values of n.

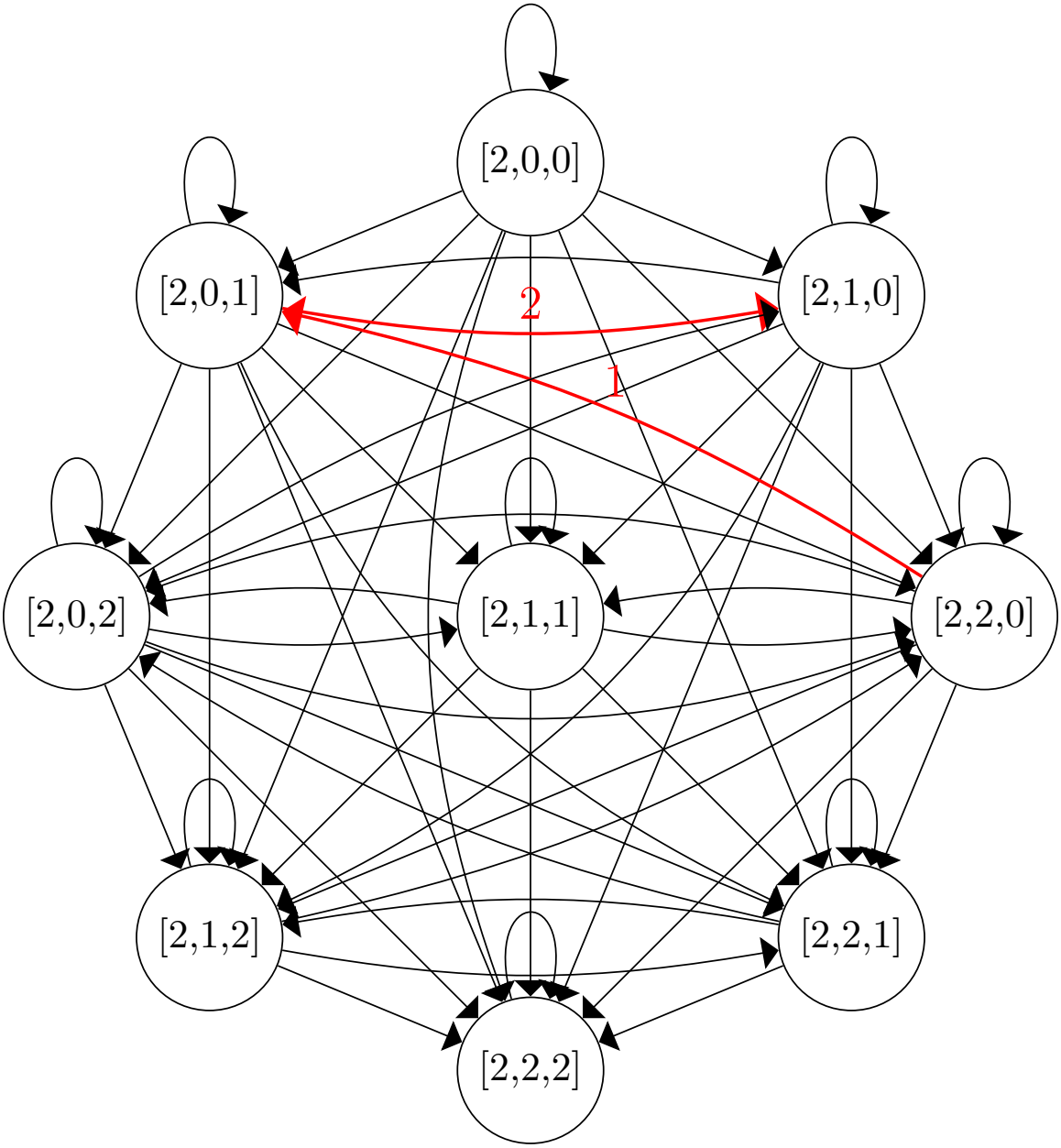


Figure B.1: **Enumeration Graph for k=2**. Edges that cannot possibly result in valid matrices have been removed from the fully connected graph. However, a simple enumeration of the paths on this graph would still result in matrices which do not satisfy the compatible ordering condition (i.e. the set $\Phi(\mathbf{C})$ is empty). One example is the path highlighted in red. To account for this, the enumeration algorithm keeps track of Φ as paths are being enumerated and does not explore paths that cannot lead to valid matrices $\mathbf{C}$.

Algorithm 2: Enumerate $\mathcal{S}_{m,n,k}$ using modified depth-first search on $G_{3,k}$. $\mu Set\,(v, w)$ is the set of values for μ for which $v\mu \leq w\mu$.

Input: $G_{n,k}$, m
Output: The set $\mathcal{S}_{m,n,k}$
procedure `Setup`($G_{n,k}$, m)
  $\mathcal{S} \leftarrow \emptyset$
  $\mathbf{C} \leftarrow n \times m$ matrix
  $\Phi \leftarrow \emptyset$
  $V, E \leftarrow G_{n,k}$
  for $v \in V$ **do**
    $\mathbf{C}[1:] \leftarrow v$
    $\mathcal{S} \leftarrow \mathcal{S} \cup$ `Enumerate`(C, *1*, Φ, m, $G_{n,k}$)
  return $\mathcal{S}$
procedure `Enumerate`(C, i, m, Φ, m, $G_{n,k}$)
  if $i = m$ **then**
    return $\mathcal{S} \cup \mathbf{C}$
  $V, E \leftarrow G_{n,k}$
  $v \leftarrow \mathbf{C}[i:]$
  for $(v, w) \in E$ **do**
    $\Phi \leftarrow \Phi \cup \mu Set(v, w)$
    if $\Phi \neq \emptyset$ **then**
      $\mathbf{C}[i+1:] \leftarrow w$
      $\mathcal{S} \leftarrow \mathcal{S} \cup$ `Enumerate`(C, $i+1$, Φ, m, $G_{n,k}$)
  return $\mathcal{S}$

B.3.1 Mixtures of normal and one tumor subpopulation ($n = 2$)

For the first step in our two step procedure, we need a way to select a subset of high confidence intervals that will be used to infer $(\mathbf{C}^*, \mu^*)$ for just those intervals. Since we are modeling a sequencing experiment as a probabilistic model where reads are distributed according to a multinomial model, intervals with larger read depths are a natural candidate for selection. However, this may be confounded for intervals that are extremely amplified, thus resulting higher read counts, but where precise estimates of copy number are difficult to make. Therefore, we choose the intervals that have the longest length in the reference genome as a compromise between these competing interests. For a fixed integer d, we select up to the d longest intervals such that: (1) The number of tumor reads (t_j) and normal reads (n_j) aligning to interval I_j is non-zero; (2) The length of interval I_j is longer that $1Mb$; and (3) If T is the total number of tumor reads, N is the total number of normal reads and k is the provided max copy number parameter, then the following holds: $\frac{t_j/T}{n_j/N} < \frac{k+1}{2}$. This final constraint forces the observed copy number ratio to not be too high beyond the specified max copy number k. Additionally, if the set of selected intervals must represent $> 10\%$ of the total length of

all provided intervals, otherwise the sample is determined to not be a good candidate for analysis using THetA. By default we set $d = 100$.

B.3.2 Mixtures of normal and two tumor subpopulations ($n = 3$)

When considering a tumor to be a mixture of multiple distinct tumor subpopulations ($n \geq 3$) we rely upon the results obtained from considering the tumor to only contain a single tumor population ($n = 2$) to find the set of intervals that allow us to best be able to measure events that have occurred in a sub-population of tumor cells. In particular, we include intervals determined by the $n = 2$ analysis to have normal copy ($c^*_{j2} = 2$) as well as intervals determine to contain copy number aberrations ($c^*_{j2} \neq 2$). We also limit the copy number aberrations used to have either been predicted to be a deletion or an amplification of a single copy, as these intervals have the most reliable signal for predicting multiple tumor populations. For a fixed integer d (we use 20 by default) the interval selection process goes as follows:

- Select the top $a = \lceil d \times 0.75 \rceil$ longest intervals such that: (1) The length of the interval I_j is longer than $5Mb$, (2) $c^*_{j2} \neq 2$, and (3) $c^*_{j2} < 4$. If a such intervals do not exists, the genome is determined to not be a good sample for multiple tumor population analysis.
- Select the top $d - a$ longest intervals such that: (1) The length of the interval I_j is longer than $5Mb$, and (2) $c^*_{j2} = 2$. If $d - a$ such intervals do not exists, the genome is determined to not be a good sample for multiple tumor population analysis.

B.4 Determining Additional Copy Numbers: Multiple Rows

Individually estimating optimal copy numbers for low confidence intervals is not guaranteed to find optimal solution as if all intervals were jointly estimated. In order to test how well the procedure does in practice, we ran the two-step algorithm and inferred copy numbers for all intervals on three less fragmented (< 200 interval) whole genome and exome samples, as well as a single chromosome of a more fragmented sample (See Table B.1). We then fixed the values of the high confidence intervals used in Step 1, and the estimated μ value, and through brute-force enumeration, found the true optimal value for the low confidence intervals. We find that for the less fragmented whole genome and exome samples, the step two procedure correctly inferred the optimal copy number for

all intervals. On the single chromosome sample, the procedure was correct for all but one interval.

ID	Data Type	# Intervals (Total)	# Intervals (Step 2)	# Intervals Incorrect
TCGA-06-0137	Exome	163	75	0
TCGA-AO-A0JF	WGS (low)	129	29	0
TCGA-BH-A0W5	WGS (low)	53	3	0
TCGA-56-1622 (Chrm 1)	WGS	159	92	1

Table B.1: **Analysis of two step method with respect to optimality.** Comparison of copy numbers inferred during step two of the two-step procedure to the optimal values.

B.5 Probabilistic Model of BAFs

In this section we describe in more detail our probabilistic model of BAFs. Let $\mathbf{s} = (s_1, s_2, \ldots, s_q)$ be a set of genomic coordinates for germline heterozygous SNPs in a patient, and let $\mathbf{v} = (v_1, v_2, \ldots, v_q)$ be the observed BAFs across all $s \in \mathbf{S}$ in the normal sample and $\mathbf{w} = (w_1, w_2, \ldots, w_q)$ be the BAFs for the corresponding tumor sample. We use a probabilistic model to describe $\mathbf{w}$.

Assuming that reads are generated uniformly at random across all DNA in a sample, we first calculate the expected deviation in BAF away from 0.5 for an interval I_j given a pair $(\mathbf{C}, \mu) \in \Omega$. In order to calculate this deviation, we need to know the number of copies of I_j for both parental chromosomes (and hence the number of copies of each allele for any $s_i \in I_j$. We do not need to know which copy number pertains to which allele, just the pair of integer values. We note that if we assume that if we make the simplifying assumption that no region is deleted, followed by a gain (and vice versa), we can exactly determine these values for regions of total copy $0, 1, 2$ and 3. Therefore, for the remainder of this section we assume that no entry in $\mathbf{C}$ is greater that 3. We define a function ϕ that given total copy number of an interval, returns the number of copies of the more common parental chromosome. That is $\phi(0) = 0$, $\phi(1) = 1$, $\phi(2) = 1$, and $\phi(3) = 2$. We now can define a value δ_j that gives the deviation away from 0.5 expected for interval I_j given a pair $(\mathbf{C}, \mu)$:

$$\delta_j = \frac{\sum_{k=1}^{n} \phi(c_{jk})\mu_k}{\sum_{k=1}^{n} c_{jk}\mu_k} - 0.5. \tag{B.1}$$

That is, δ_j is the fraction of total copies of interval I_j that contain the major (or more common) allele for any germline SNP located in I_j. For example, if interval I_j has not undergone any copy number events $c_{jk} = 2$ for all k then $\delta_j = 0$. Let $\delta = (\delta_1, \delta_2, \ldots, \delta_m)$ be the expected deviation away

from 0.5 for all intervals in $\mathbf{I}$. Note that if $\delta_j \neq 0$ we expect that the BAFs in interval I_j will be double banded, containing two clusters around $0.5 \pm \delta_j$.

We define a map $\pi : \{1, \ldots, q\} \longrightarrow \{1, \ldots, m\}$ where $I_{\pi(i)} \in \mathbf{I}$ is the genomic interval that contains SNP s_i. Let $\sigma^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$ where σ_j^2 is the observed variance around 0.5 for all heterozygous SNPs in interval I_j in the matched normal genome. That is $\sigma_j^2 = \frac{\sum_{i=1}^{q} \mathbb{1}(\pi(i),j)(v_i - 0.5)^2}{\sum_{i=1}^{q} \mathbb{1}(\pi(i),j)}$ where $\mathbb{1}$ is the identity function. Lastly, we define a sign function $sgn(x)$ such that $sgn(x) = 1$ if $x \geq 0$ and $sgn(x) = -1$ if $x < 0$. We now present a probabilistic model using a collection of gaussians for observed BAFs $\mathbf{w}$ given a pair $(\mathbf{C}, \mu) \in \Omega$ and observed BAFs in the matched normal $\mathbf{v}$ as a product of draws from different normal distributions.

$$P(\mathbf{w}|\mathbf{C}, \mu, \mathbf{v}) = P(\mathbf{w}|\delta, \sigma^2) = \prod_{i=1}^{q} P(w_i|\delta, \sigma^2) = \prod_{i=1}^{q} \mathcal{N}(w_i; 0.5 + sgn(w_i - 0.5)\delta_{\pi(i)}, \sigma_{\pi(i)}^2) \qquad \text{(B.2)}$$

Given multiple pairs $(\mathbf{C}, \mu)$ with the same likelihood using only read depth, we may select the pair that maximizes the likelihood in Equation (B.2) to select the reconstruction most consistent with observed BAF data.

B.6 Simulated Data

B.6.1 Simulation Procedure

We create a simulated mixture of a specified number of tumor subpopulations along with normal admixture using real sequencing data from an AML tumor sample and matched normal sample (TCGA-AB-2965) from The Cancer Genome Atlas [26]. This sample was chosen due to its high purity (approximately 95% pure) and lack of copy number aberrations. We create tumor subpopulations similar to the glioblastoma genomes analyzed in the next sections by using up/down sampling to randomly spike in chromosome arm deletions and amplifications (we excluded the p-arms of the acrocentric chromosomes 13, 14, 15, 21 and 22 from consideration). For each mixture we ensure that some aberrations are shared by different populations and that some are unique to the subpopulations. We then created a mixture by selecting reads uniformly at random from the original tumor genome and the created subpopulations to create a simulated mixtures. We then used the true matched normal sample as the normal sample in the simulation. Using up and down sampling

we can create mixtures of different coverages.

We run our simulated data through the same pipeline as real data, including interval partitioning determine by using BIC-seq [174]. We note that BIC-seq recommends using a parameter setting of $\lambda = 2$ for low-coverage genomes and $\lambda = 4$ for higher coverage genomes. We adhere to these recommendations for these simulations.

B.6.2 Additional Simulation Results

Mixtures made with Normal Only

We note that in the main manuscript and this supplement, we include simulated data where a mixture was created by spiking in deletions and amplifications into a tumor sample which are then mixed with the original tumor sample and compared against the normal sample. As validation we also created similar mixtures by using the normal sample for all steps. We note that the data created by such a procedure will not include variation present in real data such as batch effects. We find that mixtures created using only the matched normal sample are segmented into many fewer intervals (<100) than when the mixture is created using the tumor sample (1000's intervals). As a result, we also find that THetA2 is able to perfectly reconstruct both μ and $\mathbf{C}$ for the mixtures created using only the normal sample (perhaps due to the fewer number of intervals). Therefore, we find that such simulations are valid for demonstrating that the implementation of THetA2 works as expected, but do not represent a realistic simulation given what we would expect to find in real sequencing data.

7X Coverage

We also generated simulated data with 7X coverage. We find similar trends in 7X sequence coverage data as we see with 30X sequence coverage. Namely, we find good performance at estimating μ, the larger tumor population (Tum1) and increased performance at estimating the copy numbers in the smaller tumor population (Tum2) as it increases in size. We also see increased performance at estimating copy number aberrations in both tumor populations when only considering longer intervals (Figure B.2a). We comparing to 30X coverage simulations we see similar results, except with improved performance at estimating copy numbers for longer intervals (Figure B.2b).

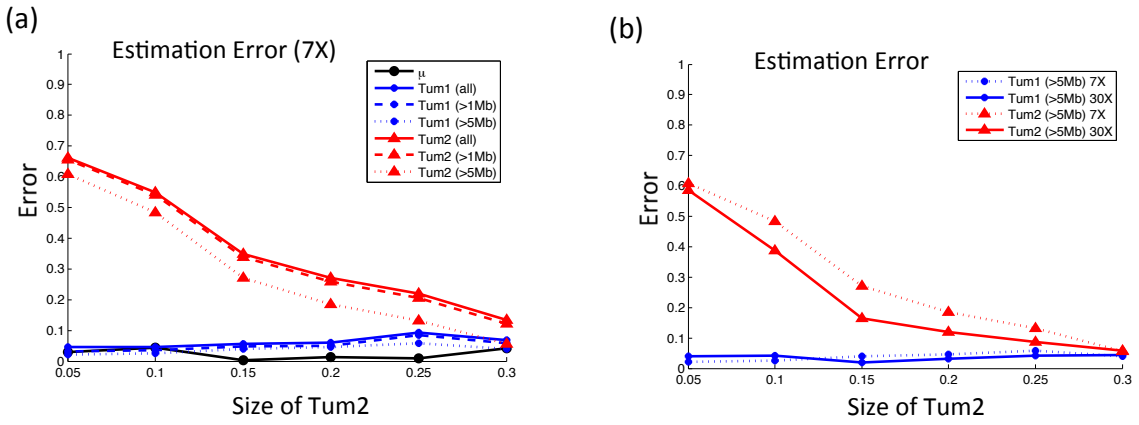


Figure B.2: **Results from running THetA2 on simulations with 7X Coverage and Comparison to 30X. (a)** Estimation error for both μ and $\mathbf{C}$ for each tumor population (Tum1 and Tum2) as the size of Tum2 increases and the size of Tum1 is fixed at 0.5 for 7X coverage. Error in μ is euclidean distance from the true μ and error for each tumor population is the fraction of the genome for which the copy number is incorrectly inferred. We also report error rates for estimating copy numbers in both populations when we only restrict consideration to longer intervals. **(b)** Comparison of 30X to 7X coverage when considering intervals longer than 5Mb.

Comparison of THetA to THetA2

We include here additional results from comparing THetA to THetA2 on simulated data. We use simulated mixtures of 3 subpopulations where the proportion of the sample in the larger tumor subpopulation is fixed at 0.5 and the proportion of the sample in the smaller subpopulation varies from 0.05 to 0.3. Figure B.3 shows a comparison between what fraction of the genome THetA and THetA2 make copy number estimates. The two-step procedure allows THetA2 to consider all of the genome while THetA only considers less than 10% of the genome.

On these same simulations we also compare the accuracy of estimating both μ and $\mathbf{C}$. When no copy number prediction is made for a region, we consider this to be an incorrect prediction.

Mixture of 4 subpopulations

We generated a mixture containing one normal and three tumor subpopulations. Subpopulation sizes were chosen to be sufficiently distinct from one another (20%, 30%, 40%). Whole arm deletions and amplifications were spiked into the mixture.

Due to the additional runtime requirements for 4 subpopulations, an alternate segmentation procedure was used. The simulated sample was divided into 50 kb intervals. We filtered out intervals which were likely to be noisy or lower quality: ones within the centromeres, ones that contained less

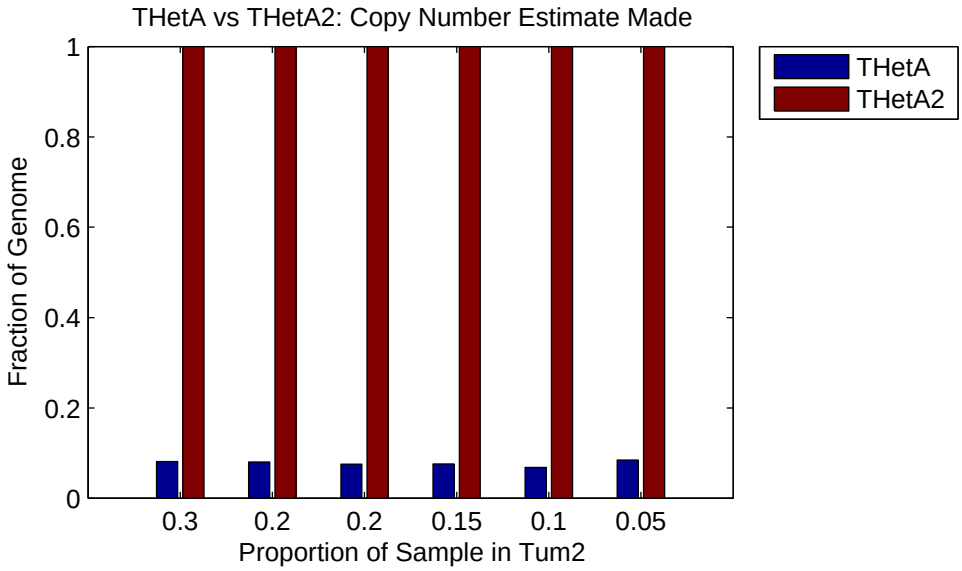


Figure B.3: **THetA vs THetA2: Fraction of Genome Considered**. A bar plot showing the fraction of the genome for which copy number estimates are made for both THetA and THetA2.

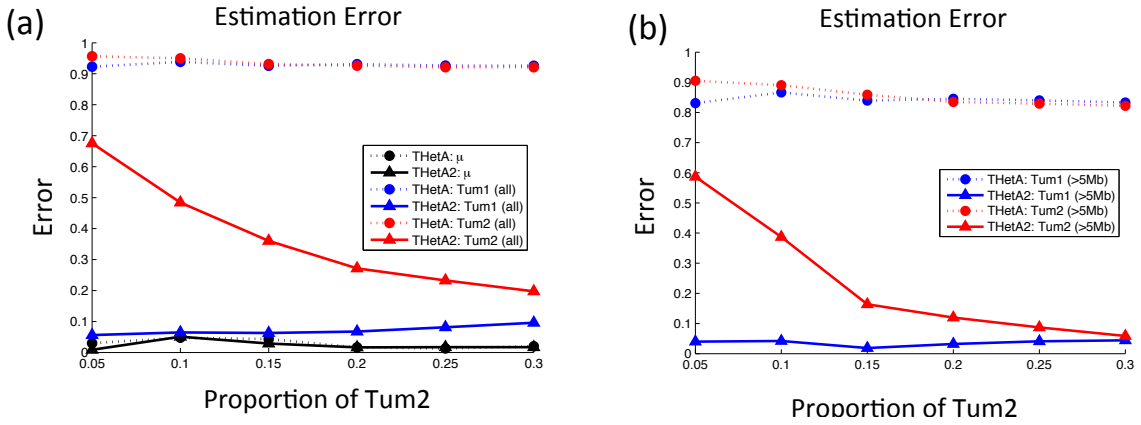


Figure B.4: **Comparison of THetA vs THetA2 on simulated data**. (Left) Comparison of error at estimation μ (measured as euclidean distance from true) and $\mathbf{C}$ (measured as fraction of genome incorrectly estimated) between THetA and THetA2. (Right) Comparison of error at estimating $\mathbf{C}$ between THetA and THetA2 when only restricting consideration to intervals longer than $5Mb$.

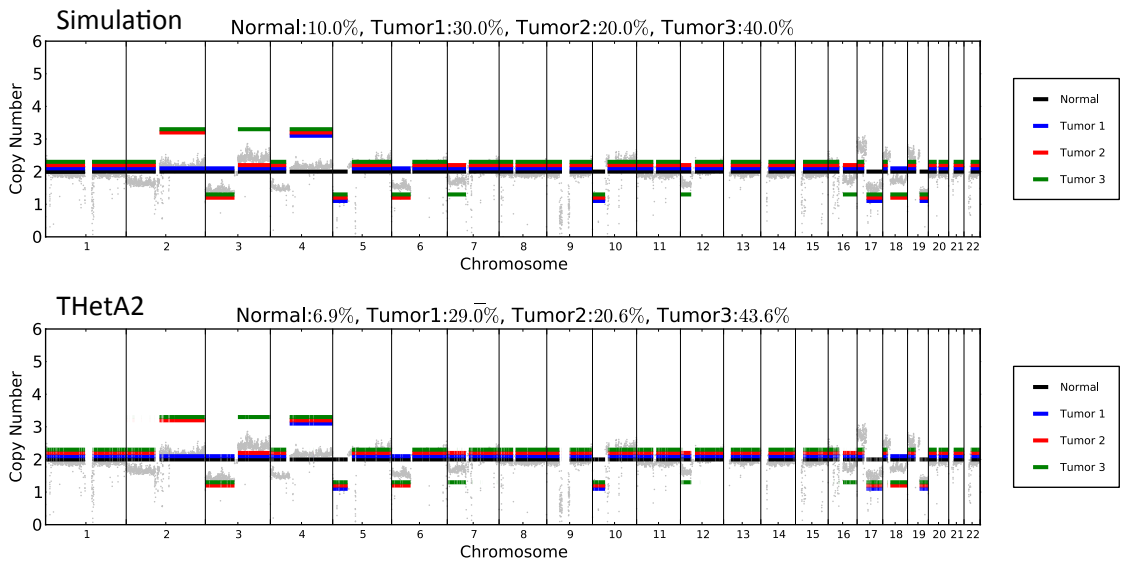


Figure B.5: **Simulation results with** 4 **subpopulations.** The simulated mixture was created by spiking in chromosome arm deletions and amplifications to create three distinct tumor populations and mixing with a matched normal genome. Due to runtime concerns for $n = 4$, an alternative segmentation algorithm was used to obtain the intervals used. The figure shows the true mixture (above) and the solution obtained by THetA2 (below).

than 2000 reads from the normal sample, and ones for which the ratio of tumor to normal reads was greater than 10% different from both of its neighbor intervals were filtered out, leaving 87.6% of the genome. For each chromosome, kernel density estimation of the distribution of tumor to normal read ratios was used to cluster intervals into larger intervals that we expect to contain the same copy number, then these large intervals were merged with intervals from other chromosomes which display similar tumor to normal read ratios.

Figure B.5 shows the results of running THetA2 on these intervals. THetA2 was able to infer μ with 4.9% estimation error (using euclidian distance from the true μ. THetA2 was also able to infer the correct copy number values for 99.9%, 99.9%, and 99.6% of the intervals that were considered for the 3 tumor subpopulations respectively, which cover ∼87.6% of the whole genome.

Underestimating Number of Subpopulations

We also investigated THetA2's behavior when the number of tumor subpopulations is incorrectly estimated. We considered six different mixtures of 3 subpopulations and evaluated the results returned by THetA2 when the number of subpopulations was fixed at two ($n = 2$). We find that

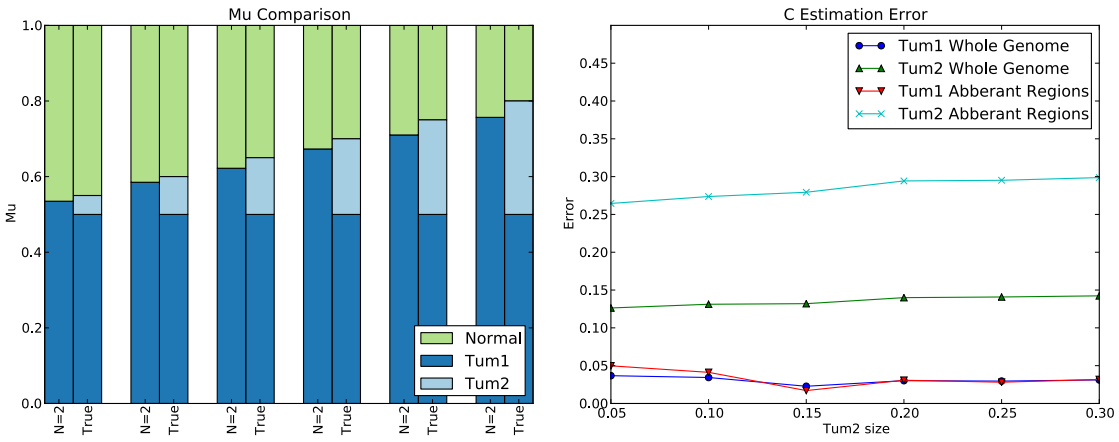


Figure B.6: **THetA2 results when underestimating the number of subpopulations.** We ran THetA2 with the number of subpopulations fixed at two ($n = 2$) on six simulated 30X mixtures of 3 subpopulations. (a) For each mixture, the predicted μ is shown next to the true underlying μ. We find that the THetA2 tends to slightly underestimate the tumor purity when considering fewer subpopulations than exist in the true underlying mixture. (b) For each mixture, the copy number profile **C** predicted was compared to the true copy number profile for the large and small subpopulation. The fraction of the genome estimated incorrectly is shown, for both the whole genome and the aberrant regions (those that contain an amplification or deletion in at least one subpopulation). We find that when considering fewer subpopulations than exist in the true mixture, THetA2 copy number predictions tend to resemble those in the largest true subpopulation.

THetA2 consistently underestimates tumor purity, but only by 0.027 on average (Figure B.6). We also compared the values in the integer count matrix **C** returned by THetA2 to the true **C** values for the large and small subpopulations. We find that in this case, THetA2 was able to accurately estimate the copy number profile of the major subpopulation: on average 97.0% of the whole genome, and 96.6% of aberrant regions (regions which contain an amplification or deletion in at least one subpopulation). Thus, THetA2 may return useful information about a sample's purity and copy number profile, even if runtime constraints force THetA2 to underestimate the true number of subpopulations.

B.7 Real Data Processing

B.7.1 Whole-Exome Data

BAM files for each sample were obtained from CGHub (`https://cghub.ucsc.edu/`) and only reads with a mapping quality ≥ 30 were used in our analysis. We determined exon positions **E** using

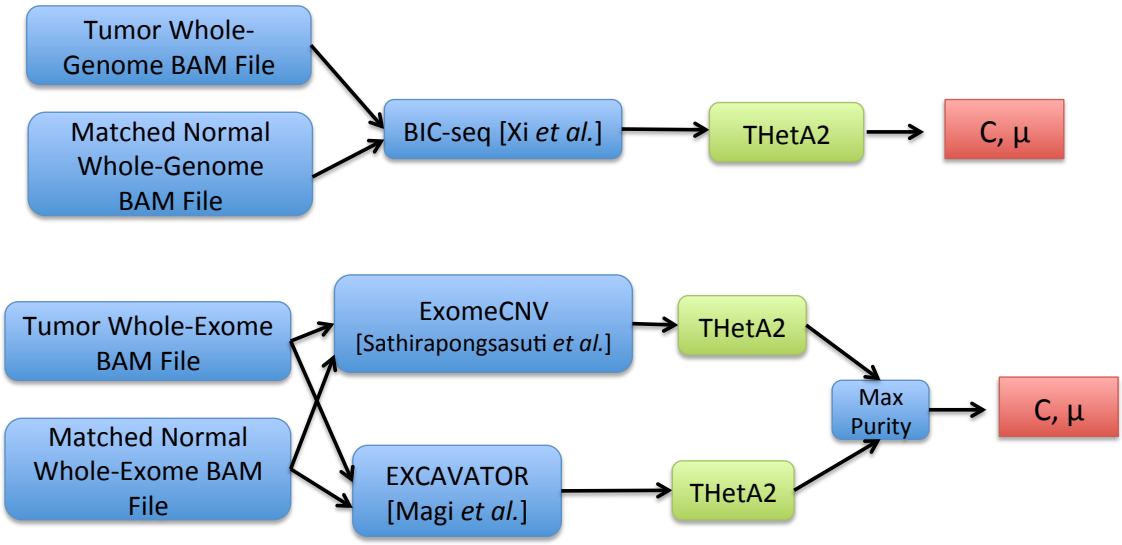


Figure B.7: **THetA2 workflow for whole-genome and whole-exome datasets.**

the UCSC genome browser [78] and merging any overlapping exonic intervals. For each sample we used both ExomeCNV [145] and EXCAVATOR [95] run with default parameters to determine an interval partition $\mathbf{I}$. ExomeCNV directly provides a segmentation $\mathbf{I}$. Whereas, EXCAVATOR only provides regions that were predicted to contain non-normal copy, so $\mathbf{I}$ was determined to be the set of returned intervals and all genomic segments located between returned intervals. Read depth $\mathbf{r}$ over these intervals and the set of exons $\mathbf{E}$ was calculated as described in the methods for both the tumor and normal genomes. A diagram describing the workflow for whole-exome data is shown in Figure B.7.

B.7.2 Whole-Genome Data

BAM files for each sample were obtained from CGHub (`https://cghub.ucsc.edu/`) and concordant reads (as determined by using the GASV pre-processing utility [150]) with a mapping quality ≥ 30 were used in our analysis. For each sample we used both BIC-seq [174] run with default parameters ($\lambda = 2$) to determine an interval partition $\mathbf{I}$. A diagram describing the workflow for whole-genome data is shown in Figure B.7.

B.7.3 Virtual SNP Array

See our previous publication [118] for details of how a virtual SNP array is created. When comparing results obtained from THetA2 to data from a virtual SNP array, we calculate an observed mean BAF. This value is calculated for each interval in an interval partition of the genome obtained from either the BIC-Seq [174], ExomeCNV [145] or EXCAVATOR [95]. We only report values for intervals that are longer than 2Mb and contain at least 10 heterozygous SNPs in the matched normal sample. We calculate the standard deviation in the observed B-allele frequencies (BAFs) for all germline SNPs occurring in the specified interval in both the tumor (σ_t) and matched normal (σ_n) samples. If the $\sigma_t < 1.5\sigma_n$, then we report the mean as 0.5, as would be expected in a non-rearranged interval. However, if $\sigma_t >= 1.5\sigma_n$, then we report two mean values - the mean of all BAFs in the interval that are greater than 0.5 and the mean of all BAFs in the interval that are less than 0.5. These values represent the mean BAF suggested by the data as reported using black bars in all BAF plots.

B.7.4 Tree Construction

We describe how the trees associated with the results from a run of THetA2 are constructed. First, this tree should not be interpreted as a phylogenetic tree, but rather as a tree representing the nested partitioning of inferred subpopulations and the aberrations whose population frequencies place them in each subpopulation. This construction of a binary tree partition is defined formally and studied in [64]. We only create such trees when they can be constructed unambiguously. This will always be the case for mixtures of three or fewer subpopulations, but since THetA2 makes no "perfect phylogeny" assumption about the subpopulations that it infers, such a tree may not be constructible with four or more subpopulations.

Each tree is constructed as follows. Each subpopulation is a leaf and is annotated with the fraction of the tumor mixture that was predicted to account for that population. For any pair of tumor subpopulations that share aberrations we add a parent node connecting them and label the node with the total fraction of cells in the sample that are part of either subpopulation. We iterate this process up the tree until we can join all remaining populations with a root node. The aberrations labeled on leaf nodes are unique to that subpopulation. Any aberrations that are shared among the tumor subpopulations are labeled on their parent node, rather than labeling each leaf node. An aberration is listed as a whole-arm event when more than a fixed proportion (> 0.7) of the

ID	Cancer Type	ABS	WXS	WGS	WGS (low)
TCGA-06-0137	GBM	X	X		
TCGA-06-0145	GBM	X	X		
TCGA-06-0171	GBM	X	X		
TCGA-06-0174	GBM	X	X		
TCGA-06-0185	GBM	X	X	X	
TCGA-06-0188	GBM	X	X	X	
TCGA-06-0214	GBM	X	X	X	
TCGA-06-0219	GBM	X	X		
TCGA-06-2557	GBM		X		
TCGA-56-1622	LUSC		X	X	
TCGA-A2-A0EU	BRC		X		X
TCGA-AO-A0JF	BRC		X		X
TCGA-AO-A0JJ	BRC		X		X
TCGA-AO-A0JL	BRC				X
TCGA-BH-A0W5	BRC		X		X
TCGA-13-1500	OV	X	X		
TCGA-29-1768	OV	X	X		
TCGA-A3-3324	KIRC			X	

Table B.2: **Genomes and associated datatype analyzed with THetA2.** A list of the genomes analyzed, the cancer type and what type of datasets were available for sample purity analysis. ABS refers to ABSOLUTE results obtained from SNP array data as reported in [27].

chromosome arm was predicted to be either deleted or amplified in a single subpopulation. Finally the root of the tree represents the complete collection of cells in the sample.

B.8 TCGA Samples: Additional Results

Table B.2 contains a complete list of genomes analyzed broken down by TCGA sample ID and the available datatypes and purity estimates for each. Table B.3 contains the complete purity estimation results across all samples, including TCGA histopathology results and purity estimates reported for the ABSOLUTE algorithm [27].

B.8.1 Whole Exome Sequencing Data

When considering a tumor to be a mixture of normal cells and a single tumor population we find that THetA2 purity estimates obtained from both the ExomeCNV and EXCAVATOR interval segmentations to be similar for most genomes (Figure B.8) with a few outliers. While the Pearson correlation coefficient between the purity estimates obtained from the different segmentations is 0.47, most of

Sample	Path.	ABS	WGS Purity (# populations)	WXS Purity (# populations)	Overlap	CNA Sim1	CNA Sim2
TCGA-06-0137	0.85-0.9	0.92	-	0.89 (2*)	-	-	-
TCGA-06-0145	0.8-0.9	0.79	-	0.84 (2*)	-	-	-
TCGA-06-0171	0.3-0.5	0.76	-	0.68 (3)	-	-	-
TCGA-06-0174	0.8-0.9	0.95	-	0.92 (3)	-	-	-
TCGA-06-0185	0.95	0.89	0.87 (3)	0.83 (2*)	0.97	0.91	0.91
TCGA-06-0188	0.6-0.8	0	0.70 (3)	0.63 (3)	0.96	0.79, 0.62	0.80, 0.70
TCGA-06-0214[1]	0.25-0.8	0.66	0.67 (3)	0.67 (3)	0.96	0.97, 0.92	0.97, 0.94
TCGA-06-0219	0.8-0.95	0.65	-	0.69 (3)	-	-	-
TCGA-06-2557	1.0	-	-	0.58 (3)	-	-	-
TCGA-56-1622	0.9	-	0.68 (3)	0.78 (3)	0.96	0.89, 0.57	0.91, 0.77
TCGA-A2-A0EU	0.9	-	0.77 (3)	0.90 (3)	0.91	0.61, 0.22	0.64, 0.31
TCGA-AO-A0JF	0.7	-	0.52 (2*)	1.00 (2*)	0.97	0.98	0.98
TCGA-AO-A0JJ	0.8	-	0.52 (3)	0.52 (2)	0.85	0.67	0.68
TCGA-AO-A0JL	0.8	-	0.87 (3)	-	-	-	-
TCGA-BH-A0W5	0.7	-	0.51 (2*)	0.54 (2*)	0.98	0.97	0.97
TCGA-13-1500	0.89	0.75	-	0.77 (3)	-	-	-
TCGA-29-1768	0.25-0.5	0.55	-	0.87 (3)	-	-	-
TCGA-A3-3324	0.3-0.45	-	0.58 (2*)	-	-	-	-

Table B.3: **Comparison of THetA2 results on whole-genome and whole-exome data.** Path. are purity estimates reported in TCGA histopathology reports. ABS are ABSOLUTE purity estimates reported by [27]. WGS Purity, WXS Purity and # populations are values predicted by THetA.* indicates that the sample did not pass the criteria to be considered for multiple tumor populations (see Appendix B - Interval Selection). Overlap is $\frac{\mathbf{I}^*}{|\mathbf{I}_{WGS}|\cup|\mathbf{I}_{WXS}|}$ where $\mathbf{I}_{WGS}$ and $\mathbf{I}_{WXS}$ are the interval partitions for the whole-genome and whole-exome data, respectively, and $\mathbf{I}^*$ is the set of intervals longer than 100kb contained in both $\mathbf{I}_{WGS}$ and $\mathbf{I}_{WXS}$. CNA Sim1 is the fraction of $\mathbf{I}^*$ where the copy number estimates are the same between the two data types. CNA Sim2 is the fraction of $\mathbf{I}^*$ where the copy number estimates are of the same type (deletion, amplification, normal) between the two data types. [1]For sample TCGA-06-0214, WGS data was aligned to hg18 and WXS data aligned to hg19. We also compared to WGS data aligned to hg19, but found it contained a much larger variance in read depth than the hg18 data.

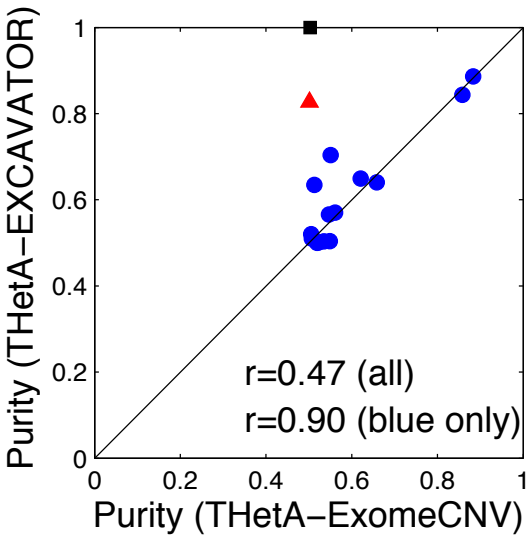


Figure B.8: **Comparison of purity estimates obtained for two whole-exome segmentation methods when considering a tumor to be a mixture of normal cells and one tumor population.** The sample indicated by the red triangle is TCGA-06-0185. The sample indicated by the black square is TCGA-AO-A0JF. Values of r shown are the Pearson correlation coefficient over either all the datapoints, or the indicated subset.

this error comes from two samples, TCGA-06-0185 and TCGA-AO-A0JF, and the correlation increases to 0.9 when these two samples are excluded. THetA2 infers multiple tumor subpopulations in sample TCGA-06-0185, so we surmise that the discrepancy between the purity estimates is due to the presence of subclonal copy number aberrations. We infer that sample TCGA-AO-A0JF contains copy number aberrations in a small subpopulation (Figure B.9) by running THetA with parameters that allow for normal contamination up to 100% cells (rather than using the default settings). We believe this leads to the discrepancy in purity estimates between the two segmentation methods when run with the default parameters. We therefore exclude this sample from further analysis.

B.8.2 Consistency Across Sequencing Platforms

For the 7 genomes for which we have both whole-exome and whole-genome data, we compare THetA2 results across both data types. To compare copy number predictions, we use two different similarity metrics (see Table B.3). For similarity metric 1 (CNA Sim 1) we calculate the fraction genomic intervals in $\mathbf{I}^*$ where THetA2 returns the same copy number for the whole-genome and whole-exome data. For similarity metric 2 (CNA Sim 2) we relax the assumption that THetA2 returns the same integer copy number in the whole-genome and whole-exome data, and instead calculate the fraction

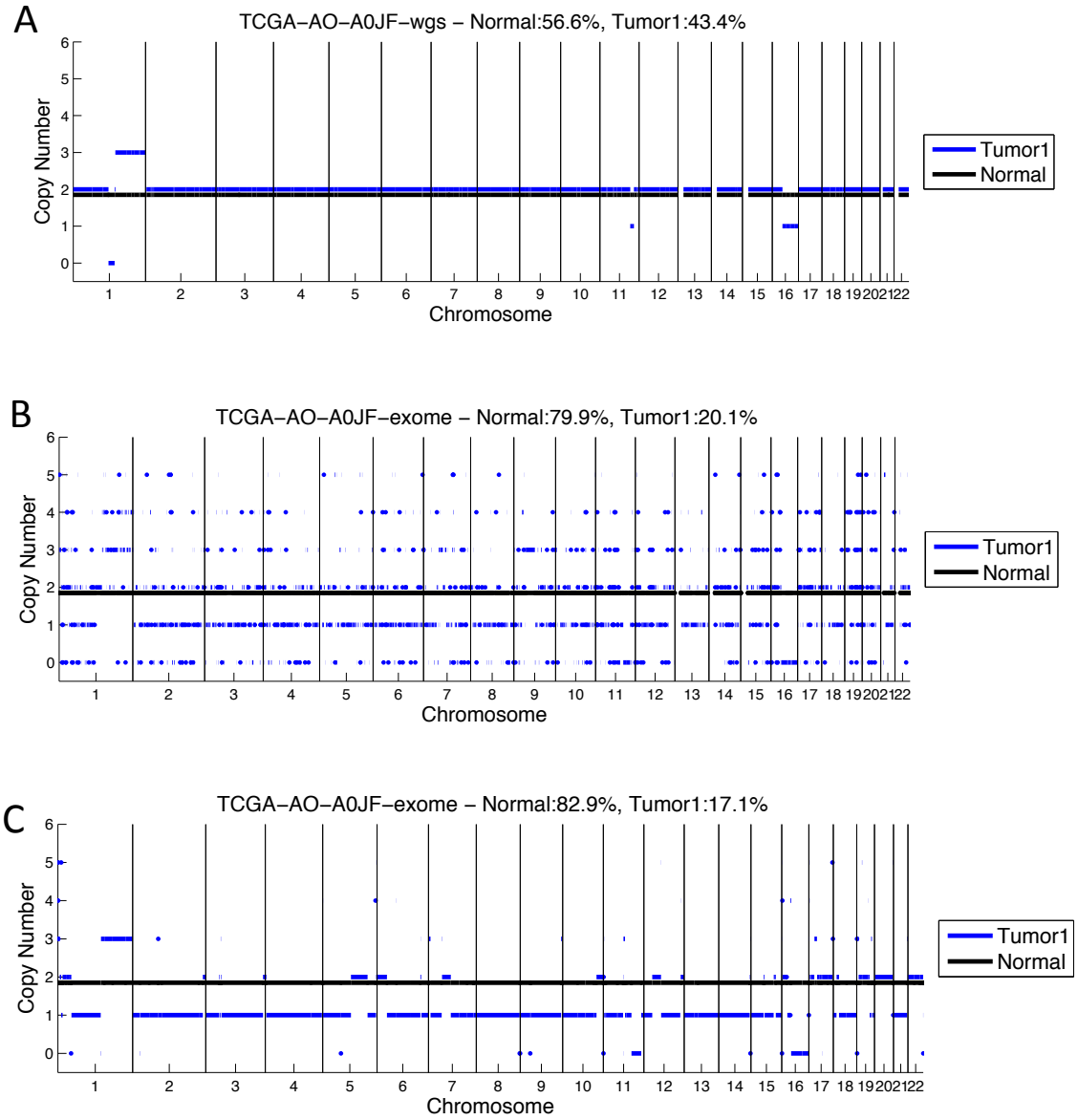


Figure B.9: **THetA2 results when analyzing whole-genome and whole-exome data for sample TCGA-AO-A0JF and considering normal contamination up to 100% cells. A.** Results using the BIC-Seq partition from whole genome data. **B.** Results using the ExomeCNV partition on whole-exome data. **C.** Results using the EXCAVATOR partition on whole-exome data. All 3 indicate that sample purity is < 0.5.

of intervals in $\mathbf{I}^*$ where the copy state (normal, deleted, amplified) is the same for both datatypes. To account for different numbers of populations predicted from the different datatypes (either due to different estimates or one datatype not passing all criteria of multiple population analysis), we report similarity between the two largest subpopulations, and when applicable, the similarity between the two smaller subpopulations.

B.8.3 Sample TCGA-06-0188

We perform additional analysis on GBM sample TCGA-06-0188 which reported ABSOLUTE results [27] indicate as non-clonal and therefore was unable to determine sample purity. TCGA histopathology reports this sample as having purity between 0.6-0.8. Both whole-genome and whole-exome data from TCGA was available for this sample. THetA results on whole-genome data indicate that the sample contains 30% normal cells and two tumor populations in 43.2% cells and 26.8% cells (Figure B.10A). Results from applying THetA to whole-exome data are similar and indicate that the sample contains 36.6% normal cells and two tumor populations in 43.1% cells and 20.3% cells (Figure B.10B). Notably, both purity estimates are within the range indicated by histopathology. A number of large copy number aberrations are predicted from both data types. Virtual SNP array analysis appears to indicate the existence of aberrations predicted by both data types, such as clonal deletion of 13q and subclonal deletion of 10 as well as other aberrations inferred from the whole-exome data such as clonal amplification of chromosome 7, clonal deletion of chromosome 22q and subclonal deletion of 17p (Figure B.10C).

B.8.4 Low-Pass Breast Cancer Genomes

We include here the sample composition inferred by THetA2 for two of the low-pass breast cancer genomes, TCGA-A2-A0EU and TCGA-AO-A0JL (Figure B.11), for which we infer multiple distinct tumor subpopulations. Both genomes appear highly rearranged and we predict a number of chromosome arm events. We note that our inferred purity values of 0.77 and 0.88 are near the reported histopathology purity values of 0.9 and 0.8 for these samples.

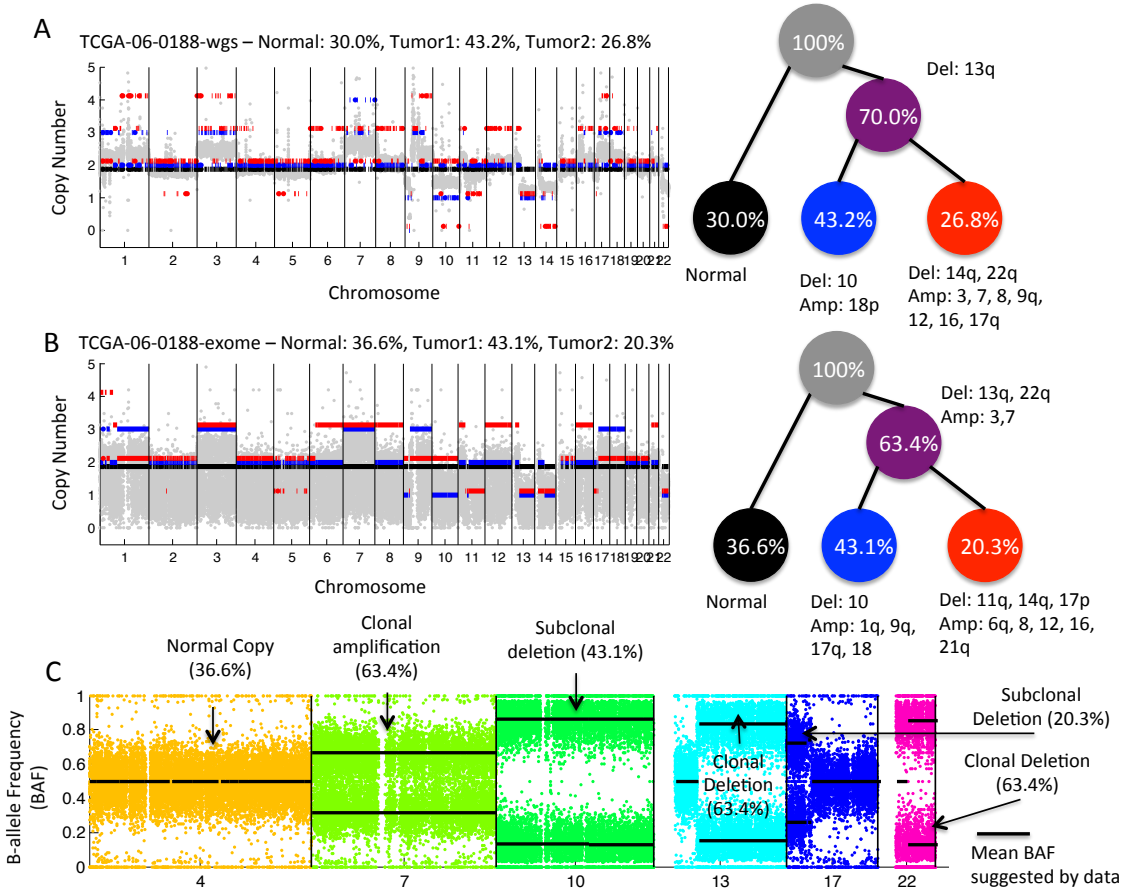


Figure B.10: **THetA2 results when analyzing whole-genome and whole-exome data for sample TCGA-06-0188. A.** (Left) Read depth rations (gray) over 50 kb bins and the inferred copy number aberrations for intervals > 2 Mb calculated by THetA2 applied to whole-genome data when the tumor is considered to be a mixture of 3 subpopulation: normal cells (black), and two tumor subpopultions (blue and red). (Right) A reconstruction of the tumor mixture along with ancestral clonal population (purple) with the inferred aberrations and estimated fraction of cells in each population. THetA2 results when analyzing whole-genome data for sample TCGA-06-0188. **B.** Same as the previous part, but applied to whole-exome data. **C.** Virtual SNP array showing B-allele frequencies for chromosomes 4, 7, 10, 13, 17 and 22.

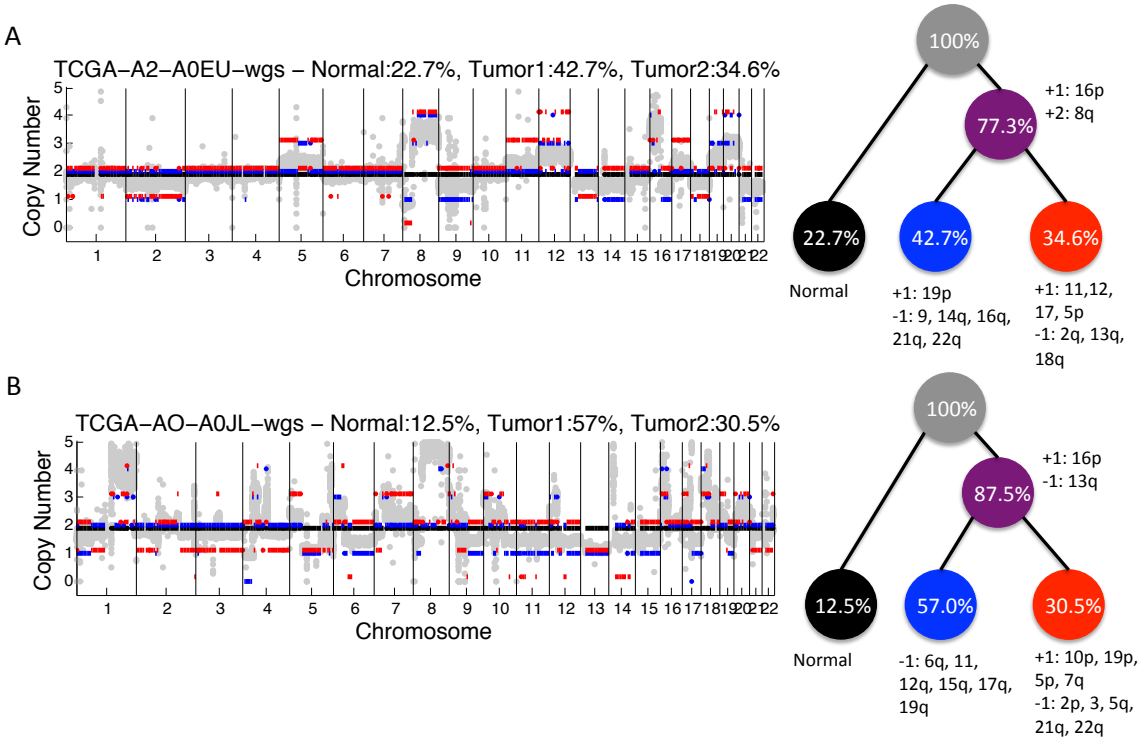


Figure B.11: **THetA2 results when analyzing low pass whole-genome data for two breast cancer samples predicted to have 3 subpopulations from low pass whole-genome data. A.** (Left) Read depth ratios (gray) over 50 kB bins and the inferred copy numbers (for all intervals $>$ 2Mb) for a mixture of normal cells (black) and two distinct tumor subpopulations (blue and red) inferred by THetA2 for sample TCGA-A2-A0EU. (Right) A reconstruction of the tumor mixture with the inferred aberrations and estimated fraction of cells in each subpopulation. **B.** Same as for part A, but for tumor sample TCGA-AO-A0JL.

ID	CNA Type	Tumor 1 (50%)	Tumor 2 (18.1%)
3q26	Amplification	X	X
5p13-14	Amplification		X
8q23	Amplification		X
8q24	Amplification	X	X
3p21	Deletion	X	X
8p21	Deletion	X	X
9p21-22	Deletion	X	X
13q22	Deletion	X	
17p12-13	Deletion	X	

Table B.4: **A list of CNAs identified in squamous cell lung cancer sample TCGA-56-1622 by THetA2 which have been reported as recurrent CNAs in lung cancers [66].**

B.8.5 Sample TCGA-56-1622

We present further results for squamous cell lung cancer sample TCGA-56-1622. First, Figure 5(b) in the main manuscript shows the observed read depth over 50kb bins as well as the predicted read depth as determined by the inferred tumor composition for this sample. For a given vector $\mathbf{x}$ we define $\widehat{\mathbf{x}} = \frac{\mathbf{x}}{|\mathbf{x}|}$. Predicted read depth is calculated using $\mathbf{C}$, μ and normal read depth vector $\mathbf{w}$. Let W be the square matrix with entries along the main diagonal equal to $\mathbf{w}$ and all other entries 0. The predicted read depth ratio for interval j is: $\frac{(\widehat{W\mathbf{C}\mu})_j}{(\widehat{\mathbf{w}})_j}$.

Second, using THetA2 results, we are able to identify several deletions and amplifications that have been reported as recurrent CNAs in lung cancers [66] (see Table B.4). In particular, we see a high amplification in 3q26 (see Figure B.12). Amplification in this region has been reported to be particularly common in squamous cell carcinoma genomes, and contains several genes which have been identified as potential oncogenic drivers in squamous cell carcinoma, including PI3KCA, SOX2, p63, SSCRO/DCUND1, and TERC [66].

B.8.6 Sample TCGA-06-0214

For this sample, we ran THetA2 with $n = 2, 3, 4$ on the whole-genome data. We find that after correcting for model size using the BIC, the $n = 2$ and the $n = 4$ solutions have a lower likelihood than the $n = 3$ solution.

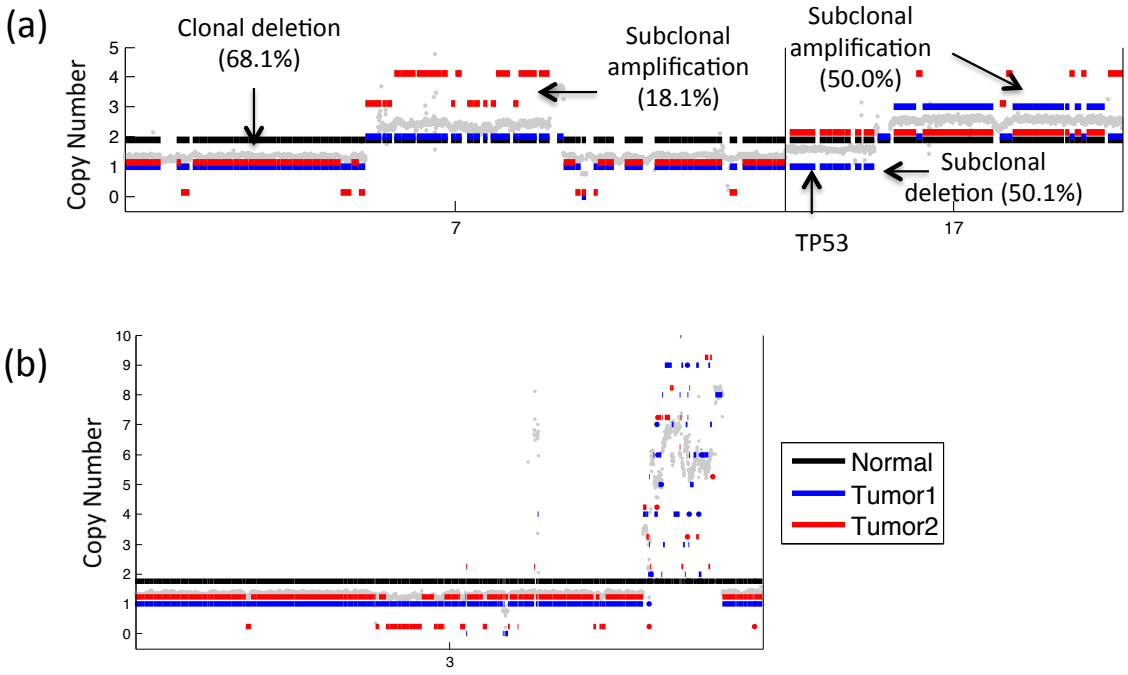


Figure B.12: **Zoomed in versions of THetA2 results when analyzing whole-genome data for sample TCGA-56-1622.** **(a)** Zoomed in view of chromosomes 7 and 17 where we identify several copy number aberrations including a subclonal deletion containing TP53. **(b)** Zoomed in view of chromosome 3. We are able to identify several CNAs common in squamous cell lung cancer, including deletion in 3p21, and amplification in 3q26.

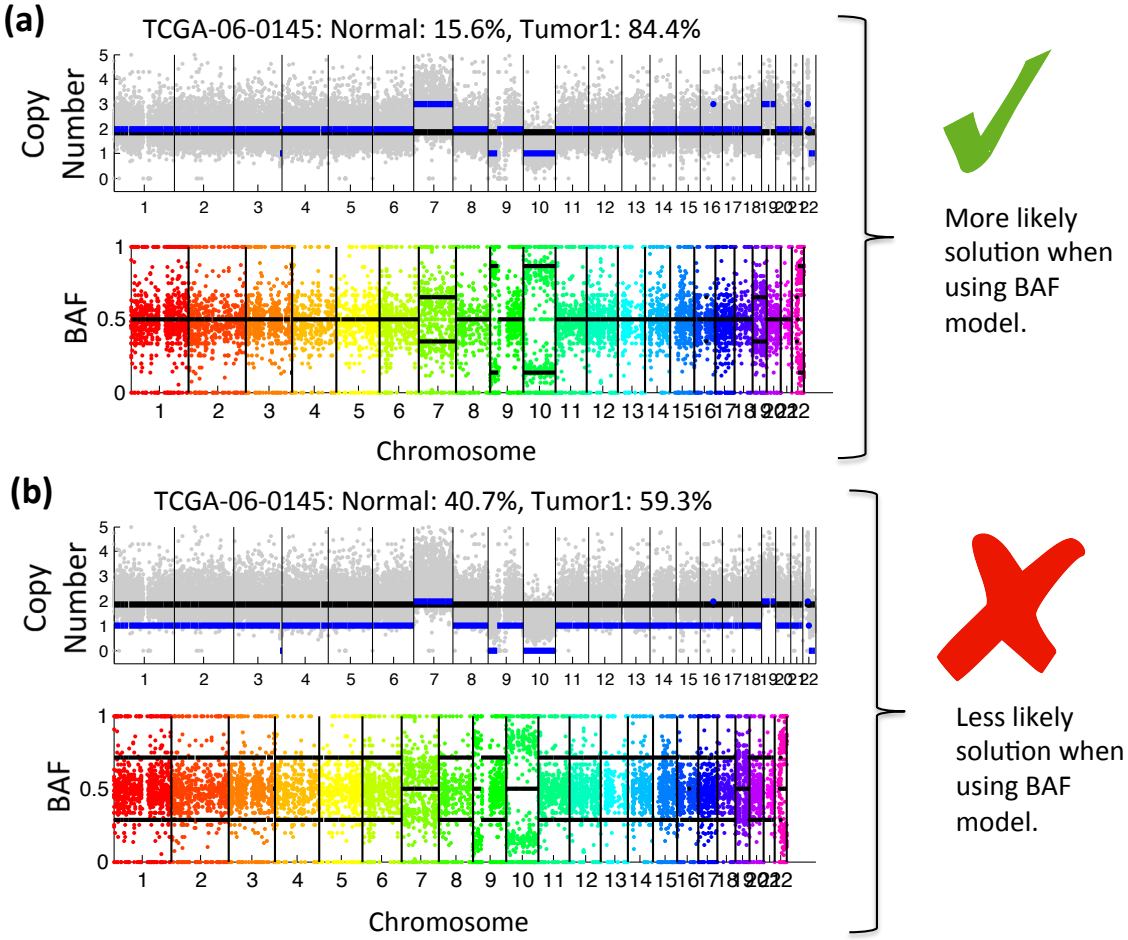


Figure B.13: **Analysis of two equally likely solutions returned by THetA2 for GBM sample TCGA-06-0145. (a)** One reconstruction returned by THetA2. (Top) Read depth ratios over 50kb bins (gray) and inferred copy numbers for normal genome (black) and one cancer genome (blue). (Bottom) Observed BAF for the genome along with expected BAF calculated using $(\mathbf{C}, \mu)$. Under the BAF model described in Equation (B.2) this reconstruction is determined to be more likely. **(b)** Same as (a) but for the second solution returned by THetA2. Under the BAF model described in Equation (B.2) this reconstruction is determined to be less likely.

B.8.7 Sample TCGA-06-0145

For glioblastoma sample TCGA-06-0145, THetA outputs two possible $(\mathbf{C}, \mu)$ pairs using only read depth – one largely haploid and one largely diploid. We apply our probabilistic model of BAFs described previously and find that the diploid reconstruction, which includes rearrangements characteristic to glioblastoma such as amplification of chr7 and deletion of chr10 [156], is determined to be the more likely tumor composition (see Figure B.13).

Appendix C

Inferring Tumor Evolution from Multi-Sample Data

C.1 Proofs Omitted from the Main Text

In this section we include proofs that were omitted from the main text.

Lemma 4.2.1. *There is a one-to-one correspondence between $\mathcal{T}_n$ and $\mathcal{B}_n$.*

Proof. First we define a function $\Phi : \mathcal{T}_n \longrightarrow \{0,1\}^{n \times n}$ such that $\Phi(T) = X$ with $x_{jk} = 1$ if and only if either $k = r$ or mutation k exists on the unique path from v_r to v_j in T. Intuitively, the j^{th} row of X is a binary vector indicating which mutations exist in clone v_j. It suffices to show that for any tree $T \in T_n$ we have $\Phi(T) \in \mathcal{B}_n$, and that for any matrix $B \in B_n$ there exists some $T \in T_n$ such that $\Phi(T) = B$.

Let $T \in T_n$. We need to show that $\Phi(T) = X \in \mathcal{B}_n$. We do so by showing that X adheres to the three conditions in the definition of an n-clonal matrix.

1. Let v_r be the root node of T. By definition of Φ, row r of X must only have a single non-zero entry at x_{rr} and therefore $\sum_{k=1}^{n} x_{rk} = 1$. For all other $j \in \{1, \ldots, n\} \setminus \{r\}$ we have $x_{jr} = 1$ and the path from v_r to v_j must have length at least one, and thus $\sum_{k=1}^{n} x_{jk} > 1$.

2. For each $j \in \{1, \ldots, n\} \setminus \{r\}$ pick $k \in \{1, \ldots, n\}$ such that T contains an edge from v_k to v_j labeled with j. Since the unique path in T from v_r to v_j traverses the unique path from v_r to

v_k, we have $x_{jl} = x_{kl}$ for all $l \neq j$. Hence $\mathbf{x}_k \subseteq \mathbf{x}_j$ and $\sum_{p=1}^{n}(b_{jp} - b_{kp}) = 1$.

3. Let $j \in \{1, \ldots, n\}$. If $j = r$ then $x_{rr} = 1$ by definition of Φ. If $j \neq r$, then by the definition of n-clonal trees the label j exists on the unique path from v_r to v_j and hence $x_{jj} = 1$.

Let $B \in B_n$. We will show that there exists some $T \in \mathcal{T}_n$ such that $\Phi(T) = B$. Let $\mathbf{b}_r$ be the unique row vector of B such that $|b_r| = 1$. By definition for every row vector $\mathbf{b}_j$ where $j \in \{1, \ldots, n\} \setminus \{r\}$ there exists a unique row vector $\mathbf{b}_k$ such that $\sum_{p=1}^{n}(b_{jp} - b_{kp}) = 1$. Given such a pair $(\mathbf{b}_j, \mathbf{b}_k)$ we define $\pi(j) = k$. We now show how to build a tree $T \in \mathcal{T}_n$ such that $\Phi(T) = B$. The vertices of T are $v_1, \ldots, v_n$ and the vertices v_j with $j \in \{1, \ldots, n\}$ correspond to row vector $\mathbf{b}_j$. Tree T is rooted at vertex v_r. For any $j, k \in \{1, \ldots, n\}$ there is an edge (v_k, v_j) if $\pi(j) = k$ and this edge is labeled by mutation j. The resulting rooted tree has n vertices and adheres to the required constraints for T to be an n-clonal tree: All edges are labeled with a unique mutation from the set $\{1, \ldots, n\} \setminus \{r\}$. Hence $T \in \mathcal{T}_n$. By construction, it is also evident that $\Phi(T) = B$. □

Lemma 4.2.2. *Any $B \in \mathcal{B}_n$ has rank n.*

Proof. We claim that the row echelon form of B is the $n \times n$ identity matrix I_n. We obtain this form by performing Gaussian elimination using a post-order traversal of all non-root vertices of the clonal tree T corresponding to B. That is, we visit a vertex only when all of its children have been visited. Let $v_j \neq v_r$ be a vertex of T whose parent v_i has not been visited yet. By Definition 4.2.1, we have that (v_i, v_j) is labeled with mutation j. We perform the elementary row operation $\mathbf{b}_j := \mathbf{b}_j - \mathbf{b}_i$ on the row vector $\mathbf{b}_j$ corresponding to vertex v_j.

By definition row $\mathbf{b}_r$ has exactly one 1-entry which occurs in column r. Let $j \in \{1, \ldots, n\} \setminus \{r\}$ and suppose v_i is the parent of vertex v_j in T. Since vertex v_i is unvisited (as we are using post-order traversal), we have that b_i has not been changed yet. Therefore subtracting b_i from b_j results in a row vector with exactly one 1-entry at column j. After the post-order traversal every row j of B will have $b_{jj} = 1$ and $b_{jk} = 0$ for $k \neq j$. Hence, the row echelon form of B is I_n, which implies that the rank of B is n. □

C.2 Details related to CITUP and PhyloSub

CITUP was run with an error rate of 0.03 and the number of clusters was kept at the default of 15. For CLL006 we set the number of clusters to 10 as the instance had fewer than 15 mutations. We

note that CITUP limits the number of vertices in the inferred trees to at most 5. We used CITUP's BIC criterion for selecting the most likely tree. PhyloSub was run with default parameters. That is, $\mu_r = 0.999$, $\delta_r = 1$, $\mu_v = 0.5$, $\delta_v = 1$. The number of MCMC samples was set to 100 and the number of Metropolis-Hastings iterations was set 5000.

C.3 Simulated Data

In this section we include some additional details and results on the simulated data.

C.3.1 Simulation Procedure

All simulated datasets were created using the following set of procedures. First, given a specified number of mutations m and number of clones n, a clonal tree was constructed by first partitioning the m mutations into n sets uniformly at random. Each set represents mutations that are clustered together, and thus first appear in the same clone in the clonal tree. A tree was then built by first randomly selecting one of the n sets of mutations to be those appearing in the root. The remainder of the tree is then constructed iteratively by randomly selecting a set of mutations and an existing vertex in the tree. A child clone is added to the selected vertex and the selected set of mutations first appear in this clone.

The corresponding usage matrix was then created row by row where each row represents the usage of a sequenced sample. Usage was determined by first selecting the number c of clones mixed in each sample by uniformly at random selecting a value between 1 and 4. Then, usage was determined by randomly sampling a value from the c simplex and applying the first c values to c randomly selected clones. We used rejection sampling over this whole process to ensure that only simulations where all mutations were included in at least two samples were created.

The simulation process described above implicitly creates a pair of matrices B and U. Let $F = [f_{pi}] = \frac{1}{2}UB$. We simulate sequenced read counts for this tumor and sequenced samples as follows. Given some expected sequencing coverage value a, mutation i and sample p we first draw the number of reads containing mutation i in sample p as $n_{pi} \sim Poiss(a)$. We then determine the number of reads containing the variant allele as $x_{pi} \sim Binomial(n_{pi}, f_{pi})$. Thus, the number of reads containing the variant allele is reported as x_{pi} and the number of reads containing the reference allele is reported as $n_{pi} - x_{pi}$.

C.3.2 Details of Comparison Metrics

In the main text we briefly describe 5 different metrics we use to compare the results of AncesTree to PhyloSub and CITUP. In particular this includes three accuracy metrics. These metric are computed using the following definition of accuracy: $(TP + TN)/(TP + FP + TN + FN)$ where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

We also compute two different error metrics. This includes one error that computes the error between the simulated usage matrix U and the inferred usage $\tilde{U}$ using a metric similar to that used by [98]. In particular, since U and $\tilde{U}$ may have a different number of columns corresponding to clones containing different subsets of mutations, we need to map the columns of $\tilde{U}$ to U. We do so, by computing a minimum-cost perfect matching in a complete bipartite graph whose two vertex sets correspond to the columns of U and $\tilde{U}$. The cost of matching column j of U to column j' of $\tilde{U}$ is the size of symmetric difference between the mutation sets of j and j' – i.e. the number of mutations that are unique to either set. To deal with potentially different sizes of U and $\tilde{U}$, we add dummy nodes to the smaller of the two vertex sets and label these dummy nodes by the empty set.

C.3.3 Additional Simulation Results

In the main text we report results comparing AncesTree to both CITUP and PhyloSub for several metrics. We include in Figure C.1 the results of one additional metric, accuracy at determining mutation pairs that are incomparable, not reported in the main text. We see that AncesTree has better accuracy, by at least 0.08, than the other approaches. We also show the distribution of the fraction of mutations included in the results returned by AncesTree over all 90 simulations in Figure C.2.

C.4 Real Data

In this section we include some additional details and results on the real data.

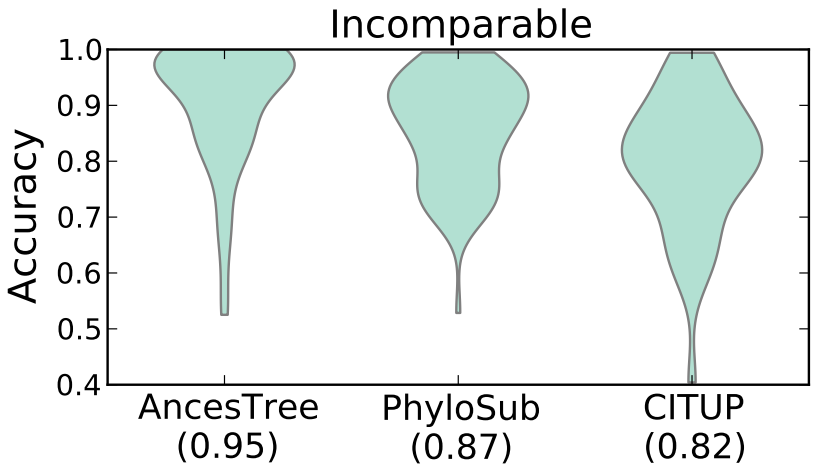


Figure C.1: **AncesTree demonstrates better accuracy at predicting mutations that are incomparable than CITUP and PhyloSub on simulated data.** Violin plots show the accuracy for each method at determining mutations that are incomparable over 90 different simulations. Values in parenthesis are the median value obtained over all simulations.

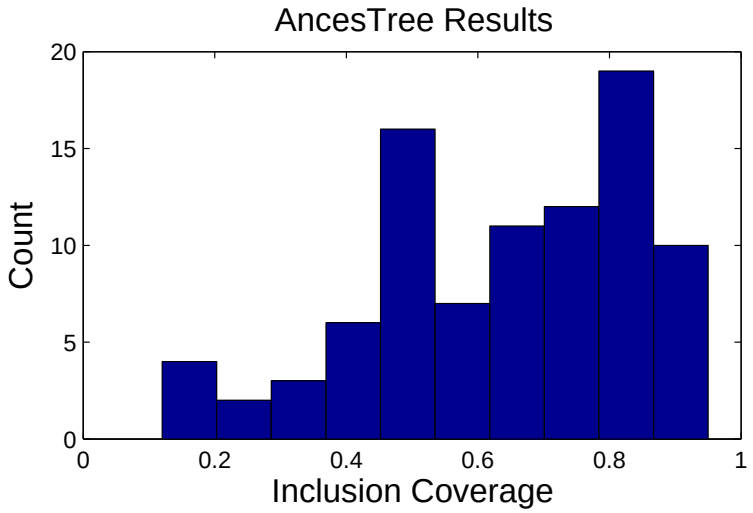


Figure C.2: **Distribution of the fraction of mutations included by AncesTree over the 90 simulations.**

C.4.1 Data Acquisition and Processing

For the lung and CLL tumors, we obtained a list of all called mutations and reference/variant allele read counts for all samples from the supplementary materials of the corresponding publication [147, 185]. For the renal tumors, we obtained the list of called mutations from the supplementary materials for [52] and obtained read counts from the authors (M. Gerlinger, personal communication). We exclude, from our analysis any mutation which is reported to be an indel or occurs in a region affected by a copy number aberration as indicated in the original publication. Thus, across all samples, we analyze 7621 of the 7684 originally reported mutations (Table C.1). For all analyses, we set $\alpha = 0.3$, $\beta = 0.8$ and $\gamma = 0.01$.

C.4.2 Overview of Results

Table C.1 contains an overview of the AncesTree results over all 22 samples analyzed.

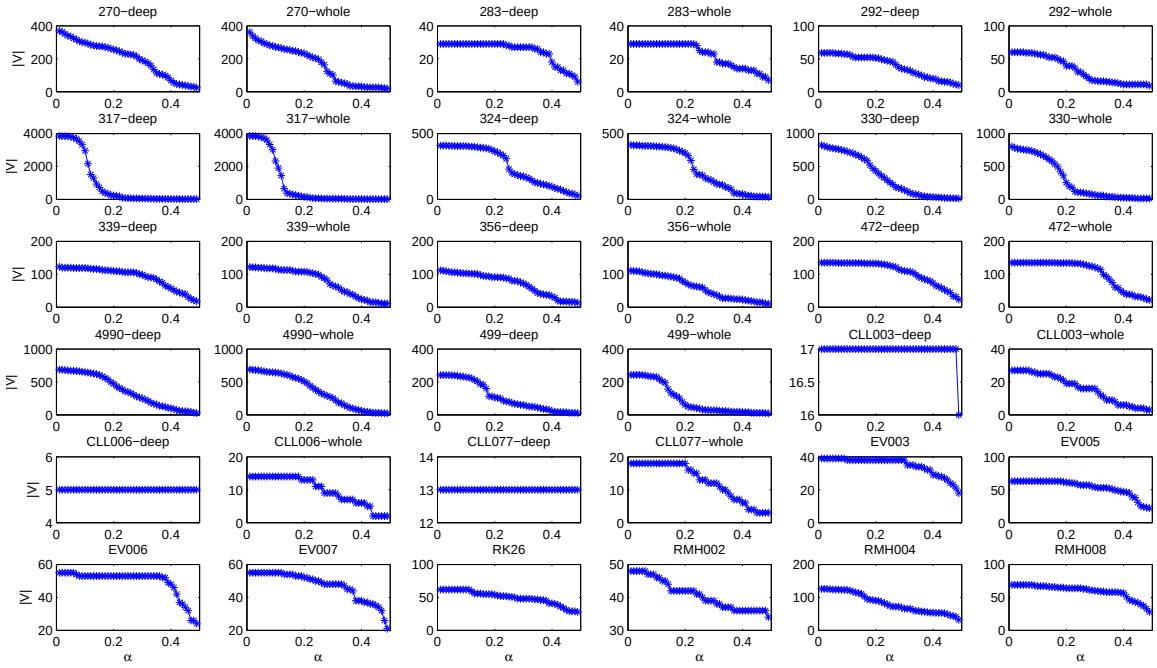


Figure C.3: **Relationship between α and $|V|$ for all datasets.** The number of vertices included in the ancestry graph $G = (V, A)$ increases with α.

Cancer	Patient	m	coverage		n		n'		$\|V\|$		$\|T\|$		depth		inner nodes		usage		gap		time (s)	
			d	w	d	w	d	w	d	w	d	w	d	w	d	w	d	w	d	w	d	w
Renal	EV003	8	1035	-	40	-	8	-	38	-	8	-	2	-	0.13	-	0.28	-	0%	-	7	-
	EV005	7	3204	-	64	-	13	-	54	-	13	-	4	-	0.23	-	0.23	-	0%	-	16	-
	EV006	9	790	-	57	-	15	-	53	-	15	-	3	-	0.13	-	0.21	-	0%	-	15	-
	EV007	8	836	-	56	-	9	-	48	-	9	-	2	-	0.11	-	0.21	-	0%	-	3	-
	RK26	11	1580	-	62	-	15	-	48	-	15	-	3	-	0.13	-	0.20	-	0%	-	5	-
	RMH002	5	3084	-	48	-	10	-	39	-	10	-	2	-	0.10	-	0.26	-	0%	-	3	-
	RMH004	6	1095	-	126	-	19	-	66	-	19	-	4	-	0.21	-	0.20	-	0%	-	12	-
	RMH008	8	3107	-	71	-	13	-	60	-	13	-	2	-	0.08	-	0.19	-	0%	-	12	-
CLL	CLL003	5	71861	77	20	30	5	9	17	16	5	4	4	2	0.60	0.25	1.00	0.65	0%	0%	1	0
	CLL006	5	93027	38	10	16	5	9	5	9	5	4	5	3	0.80	0.50	1.00	0.80	0%	0%	0	0
	CLL077	5	151858	41	16	20	7	10	13	12	7	5	4	2	0.71	0.20	1.00	0.64	0%	0%	0	0
Lung	270	5	537	156	396	396	202	311	190	101	42	40	11	9	0.60	0.55	0.40	0.39	12%	0%	43210	7391
	283	5	599	198	29	29	16	22	27	23	14	16	5	4	0.57	0.38	0.64	0.59	0%	0%	26	12
	292	3	735	211	61	61	35	58	34	17	20	15	8	6	0.70	0.47	0.67	0.69	0%	0%	46	0
	317	4	842	265	3890	3890	3822	3836	54	45	33	33	6	5	0.61	0.39	0.61	0.52	0%	0%	356	8607
	324	5	674	244	438	438	78	302	175	132	52	45	10	9	0.60	0.56	0.43	0.45	20%	6%	43208	43203
	330	4	674	201	853	853	514	825	139	69	55	45	11	5	0.65	0.42	0.44	0.50	12%	0%	43204	2239
	339	4	636	229	124	124	59	90	99	66	42	37	9	7	0.57	0.49	0.42	0.47	7%	0%	43216	5021
	356	4	673	200	112	112	56	91	72	39	23	24	7	5	0.65	0.46	0.42	0.44	0%	0%	9417	36
	472	5	791	249	135	135	58	46	110	121	40	41	10	6	0.58	0.46	0.45	0.44	17%	5%	43230	43216
	499	4	685	171	243	243	379	433	249	255	57	60	9	9	0.61	0.57	0.38	0.35	33%	16%	43214	43222
	4990	5	568	229	702	702	175	231	60	28	19	20	4	5	0.37	0.25	0.63	0.61	0%	0%	2733	10

Table C.1: **Overview of datasets and results from running AncesTree with parameters $\alpha = 0.3$ and $\beta = 0.8$.** We analyze multi-section sequencing data for 8 patients with renal cancer [52] and 11 lung cancer patients [185]. We also analyze multiple time point data for 3 patients with CLL [147]. For the CLL and lung datasets we analyzed both whole-exome/whole-genome data (w) and targeted deep sequencing (d). For the renal datasets we only analyze ulta-deep exome sequencing data (d). We report the following values for all datasets: the number of sequenced samples (m), the average number of reads (coverage), the total number of mutations identified across all samples (n), the number of mutations we use in our evolutionary reconstruction (n'), the number of vertices in the ancestry graph ($|V|$), the size of the clonal tree we infer ($|T|$), the length of the longest path in T (depth), the relative number of ancestral populations inferred (inner nodes), the average deviation between the assigned and observed frequencies (ΔF), the fraction of entries in the inferred usage matrix U that are non-zero (usage), the optimality gap (UB-LB)/LB (gap) and the running time in seconds (time).

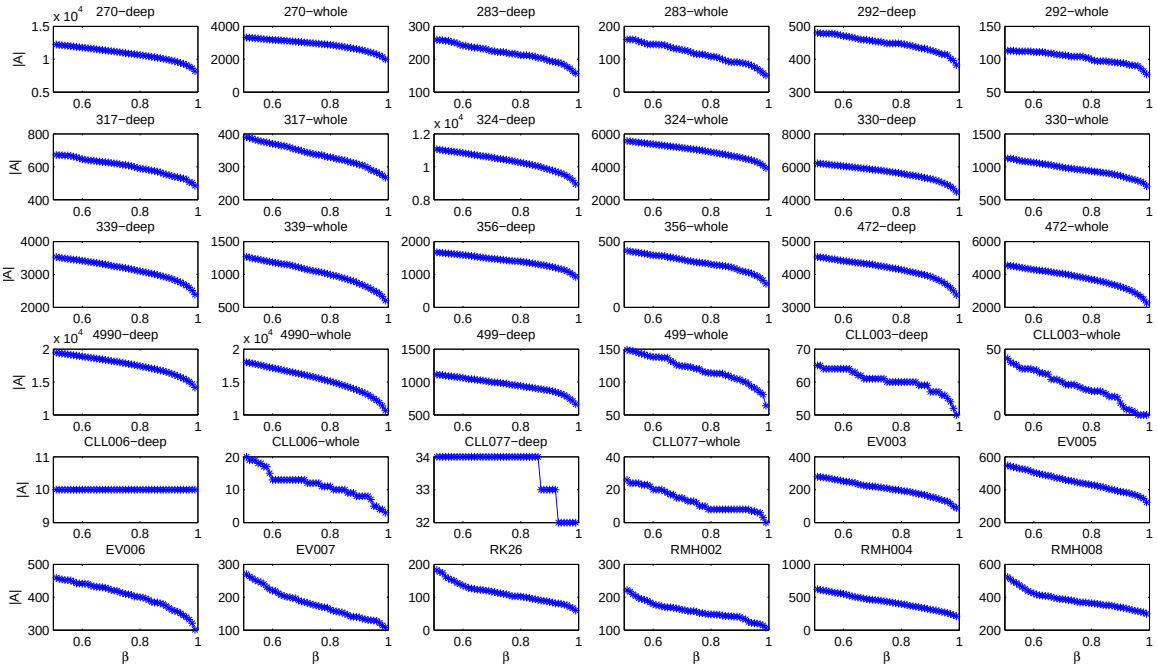


Figure C.4: **Relationship between β and $|A|$ for all datasets.** For fixed $\alpha = 0.1$, the number of edges included in the ancestry graph $G = (V, A)$ decreases as β increases.

C.4.3 Effect of α and β Parameters on G

The parameters α and β alter the resulting relaxed ancestry graph $G = (V, A)$. The α parameter controls how many mutations are clustered together and therefore controls $|V|$. Small values of α allow more mutations to satisfy the relaxed ancestry constraint, resulting in a smaller number of strongly connected components; i.e. there is more clustering of mutations when α is small. This is indeed the pattern we see as we vary α for all datasets (Figure C.3).

The parameter β controls how confident we must be in the ancestral relationship between two genes (or two sets of clustered genes) to include the corresponding edge in the relaxed ancestry graph. Thus, as β increases the number of edges $|A|$ in the ancestry graph decreases (Figure C.4).

C.4.4 Comparison of AncesTree to PhyloSub and CITUP on Real Data

We compare AncesTree to CITUP and PhyloSub [75, 98] – run with default parameter settings (see Section C.2) – on the 22 sequenced tumors. We do not compare to LICHeE [131] as the associated program provides only a graphical user interface, but no way to easily export results. We use our probabilistic model to assess the consistency of the inferred trees with the observed read counts. For

Cancer	Patient	CITUP				Phylosub				AncesTree			
		F error	MCP	MAP	MIP	F error	MCP	MAP	MIP	F error	MCP	MAP	MIP
Renal	EV003	0.027	0.01	0.07	0.00	0.030	0.02	0.64	0.00	0.000	-	1.00	0.00
	EV005	0.042	0.00	0.18	0.00	0.045	0.00	0.86	0.00	0.002	0.28	1.00	0.00
	EV006	0.029	0.00	1.00	0.00	0.031	0.00	1.00	0.00	0.003	0.43	1.00	0.00
	EV007	0.033	0.00	0.14	0.00	0.037	0.02	0.39	0.00	0.002	0.44	1.00	0.00
	RK26	0.025	0.00	0.93	0.00	0.030	0.00	0.73	0.00	0.002	0.38	1.00	0.00
	RMH002	0.035	0.00	0.51	0.00	0.031	0.00	0.49	0.00	0.003	0.43	1.00	0.00
	RMH004	0.039	0.00	0.50	0.00	0.041	0.00	0.59	0.00	0.004	0.24	1.00	0.00
	RMH008	0.053	0.00	0.08	-	0.035	0.00	0.22	0.00	0.002	0.35	1.00	0.00
CLL	CLL003_d	0.075	0.00	0.00	-	0.024	0.00	0.98	0.00	0.000	-	1.00	0.00
	CLL003_w	0.155	0.12	0.30	-	0.065	0.15	0.49	0.00	0.026	0.35	0.79	0.00
	CLL006_d	0.186	0.00	1.00	-	0.102	0.00	1.00	-	0.000	-	1.00	-
	CLL006_w	0.162	0.17	0.80	-	0.102	0.15	0.70	-	0.036	0.19	0.89	0.10
	CLL077_d	0.164	0.00	1.00	-	0.038	0.00	1.00	0.00	0.000	-	1.00	0.00
	CLL077_w	0.195	0.22	0.55	-	0.065	0.20	0.65	0.00	0.012	0.28	0.93	0.02
Lung	270_d	0.094	0.00	1.00	-	0.041	0.00	1.00	0.00	0.004	0.36	1.00	0.00
	270_w	0.110	0.00	1.00	-	0.051	0.00	1.00	0.00	0.016	0.15	0.97	0.00
	283_d	0.015	0.07	1.00	-	0.019	0.07	1.00	0.28	0.009	0.28	1.00	0.07
	283_w	0.027	0.09	0.96	-	0.028	0.10	0.95	0.08	0.021	0.24	0.97	0.08
	292_d	0.029	0.09	1.00	-	0.019	0.10	1.00	0.01	0.009	0.24	1.00	0.05
	292_w	0.037	0.14	0.99	-	0.033	0.14	0.93	0.08	0.032	0.16	1.00	0.09
	317_d	x	x	x	x	0.029	0.09	1.00	0.07	0.065	0.09	1.00	0.00
	317_w	x	x	x	x	0.036	0.08	0.96	0.06	0.070	0.08	1.00	0.00
	324_d	0.092	0.03	1.00	-	0.035	0.05	1.00	0.00	0.005	0.30	1.00	0.00
	324_w	0.115	0.04	1.00	-	0.044	0.05	1.00	0.00	0.031	0.07	1.00	0.00
	330_d	0.024	0.01	1.00	0.00	0.028	0.01	1.00	0.00	0.009	0.12	1.00	0.00
	330_w	0.031	0.03	0.96	0.00	0.036	0.03	0.99	0.00	0.041	0.06	0.90	0.00
	339_d	0.049	0.00	1.00	-	0.028	0.01	1.00	0.00	0.008	0.38	1.00	0.00
	339_w	0.054	0.02	1.00	-	0.034	0.03	1.00	0.01	0.017	0.15	0.99	0.00
	356_d	0.021	0.00	1.00	-	0.023	0.05	1.00	0.00	0.004	0.29	1.00	0.00
	356_w	0.024	0.03	0.79	-	0.027	0.04	0.99	0.01	0.011	0.15	0.94	0.00
	472_d	0.025	0.01	1.00	-	0.027	0.01	1.00	0.00	0.007	0.20	1.00	0.00
	472_w	0.036	0.01	0.98	-	0.038	0.02	0.99	0.00	0.012	0.25	1.00	0.00
	4990_d	0.085	0.00	1.00	-	0.036	0.02	1.00	0.00	0.009	0.09	1.00	0.00
	4990_w	0.114	0.00	1.00	-	0.043	0.01	1.00	0.00	0.015	0.08	0.88	0.00
	499_d	0.053	0.11	1.00	-	0.021	0.10	1.00	0.00	0.013	0.13	0.92	0.03
	499_w	0.030	0.13	0.57	0.08	0.035	0.12	0.65	0.06	0.037	0.13	0.62	0.01

Table C.2: **Analysis of CITUP, Phylosub, and AncesTree on real data.** The F error, median clustered probability (MCP), median ancestral probability (MAP) and median incomparable probability (MIP) of the trees inferred by CITUP, Phylosub, and AncesTree are listed above.

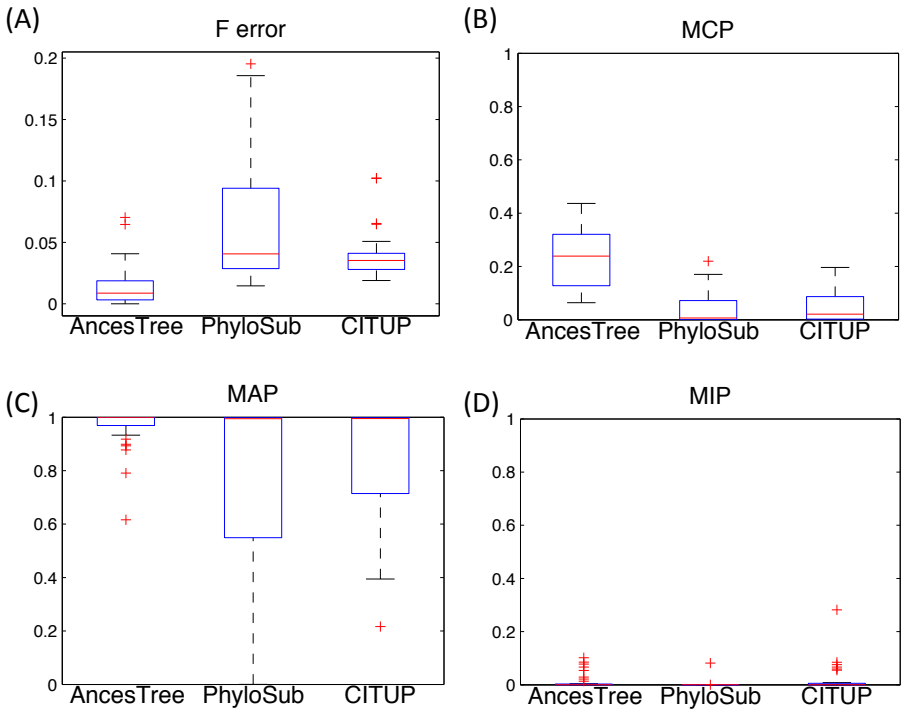


Figure C.5: **Comparison of inferred trees by AncesTree, PhyloSub and CITUP on 22 real sequenced tumors.** (A) Shows the error between the inferred and observed frequencies. (B) Shows the median clustered probability (MCP) between pairs of mutations that are clustered. (C) Shows the median ancestral probability (MAP) between pairs of mutations where one mutation is inferred to be ancestral to the other mutation. (D) Shows the median incomparable probability (MIP) between pairs of mutations that occur in different branches of the inferred tree.

a fair comparison, we only consider the data sets for which all three methods computed a solution (28 patients out of 36 patients). Certain methods did not compute a solution on certain data sets because of the size of data sets and limitations of different methods. Using four different metrics, we find that AncesTree outperforms both CITUP and PhyloSub. The first metric, ΔVAF, measures the error between inferred and observed frequencies (Figure C.5(A)) in the same way as the F error computed in the main text. The second metric measures likely errors in the clustering of mutations by identifying the fraction of pairs of clustered mutations for which there is evidence that one mutation is ancestral to the other with high probability across all samples (Figure C.5(B)). The third metric identifies the fraction of pairs of clustered mutations that are likely pairwise incomparable, i.e., there exist at least two samples with incompatible ancestry relationships between these mutations (Figure C.5(C)). The last metric measures whether the mutations connected by an arc of the inferred tree have read counts that support the implied ancestral relationship. To do so, we compute the fraction of arcs that admit a pair of mutations whose ancestral probability, calculated using the observed read counts under our posterior model, is at least β (Figure C.5(D)). See Table C.2 for further details.

C.4.5 Ancestry Graph for CLL 077

Figure C.6 contains the ancestry graph for CLL sample 077.

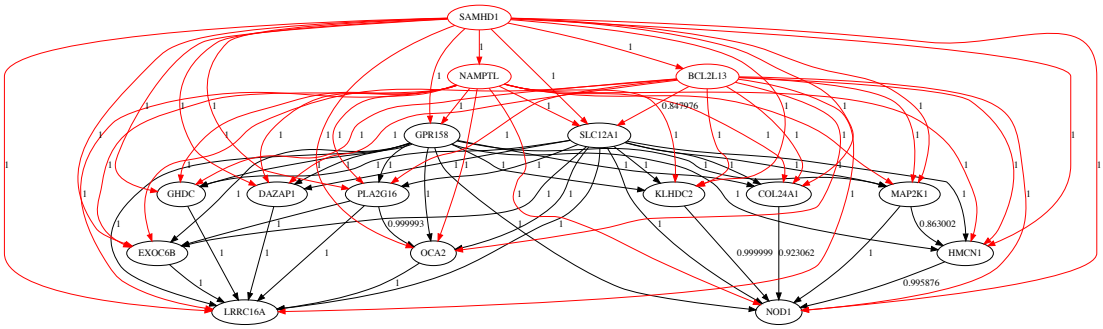


Figure C.6: **The ancestry graph for CLL007.** Edges are annotated with their posterior probability under our probabilistic model. Genes indicated in red are those whose VAFs are much higher than 0.5. In fact, their $1 - \gamma$ confidence intervals lie entirely above 0.5. It is likely that these mutations occur in a region affected by a copy number aberration, thus violating the assumptions of our model.

Appendix D

Detecting Simultaneous Rearrangements in Cancer Genomes

D.1 Proof of Theorem 6.2.2

In this section we present the necessary background to prove Theorem 6.2.2 which shows how to compute the probability that a chromothripsis string C of length m, constructed using blocks from a reference genome G of n blocks, drawn uniformly at random from the set of such strings has H/T alternating $\pi(C)$. First, we calculate $|G(m,n)|$, the total number of chromothripsis strings of m blocks given a reference genome of n blocks. This is straightforward using the following equation.

$$|G(m,n)| = \prod_{i=0}^{m-1}(2n-2i) = \prod_{i=0}^{m-1} 2(n-i) = 2^m \prod_{i=0}^{m-1}(n-i). \tag{D.1}$$

Next we calculate $|A(m,n)|$, the number of chromothripsis strings of m blocks given a reference genome of n blocks that are H/T alternating. First, we note that the selection of any two blocks (telomeres) in a reference genome $G = 1 \ldots n$ defines a partition of the remaining $n-2$ blocks into two sets: (1) blocks that lie between the two chosen telomere blocks in G; and (2) blocks that lie outside the chosen telomere blocks in G. By construction the two cases from Theorem 6.2.1 are

mutually exclusive. In Case 1 all non-telomere blocks in the derivative chromosome lie in between the telomeres and in Case 2 they must lie outside the telomeres. Therefore, we consider these two cases separately.

We begin with Case 1 from Theorem 6.2.1. Suppose we choose two blocks at random to be the telomeres for a genome. A genome with those telomeres can only fall into Case 1 if there are i blocks in between. For each value of i there are $n-1-i$ possible pairs of telomeres where $i \geq m-2$. In such an instance, there are two possible configurations of the telomeres - both in normal orientation with the smaller interval as the start of the derivative chromosome, or both in reverse with the larger interval as the start of the derivative chromosome. We now only have to choose $m-2$ blocks, along with their order and orientation from the i blocks between the telomeres. Case 2 from Theorem 6.2.1 is similar to Case 1, except now we let i represent the number of blocks that lie outside the telomeres of which there are $i+1$ paris with $i \geq m-2$. There are still two possible configurations of the telomores - both in reverse orientation with the smaller interval as the start of the derivative chromosome or both in normal orientation with the larger interval as the start of the derivative chromosome. These insights form the basis for the following equation which we simplify.

$$\begin{aligned}
|A(m,n)| &= \sum_{i=m-2}^{n-2} (2)(i+1)\left(\prod_{j=0}^{m-3}(2i-2j)\right) + \sum_{i=m-2}^{n-2} (2)(n-1-i)\left(\prod_{j=0}^{m-3}(2i-2j)\right) \\
&= 2^{m-1}\sum_{i=m-2}^{n-2} (i+1)\left(\prod_{j=0}^{m-3}(i-j)\right) + 2^{m-1}\sum_{i=m-2}^{n-2} (n-1-i)\left(\prod_{j=0}^{m-3}(i-j)\right) \\
&= 2^{m-1}\sum_{i=m-2}^{n-2} (i+1+n-1-i)\left(\prod_{j=0}^{m-3}(i-j)\right) \\
&= 2^{m-1}n\sum_{i=m-2}^{n-2}\left(\prod_{j=0}^{m-3}(i-j)\right)
\end{aligned}$$

We can now directly calculate the probability of a sequence of m blocks from an n block reference genome having an H/T alternating alignment and we do so in Theorem 6.2.2. But first we must state and prove Lemma D.1.1 which is used in the proof for Theorem 6.2.2.

Lemma D.1.1. $\sum_{i=1}^{n}\prod_{j=0}^{m-1}(i+j) = \frac{1}{m+1}\prod_{j=0}^{m}(n+j)$

Proof. We use proof by induction on the variable n. We start with the base case $n = 1$.

$$\sum_{i=1}^{1}\prod_{j=0}^{m-1}(i+j) = \prod_{j=0}^{m-1}(1+j) = \frac{1}{m+1}\prod_{j=0}^{m}(1+j)$$

We now assume the property holds for values up to $n-1$ and want to prove for generic n.

$$\begin{aligned}
\sum_{i=1}^{n}\prod_{j=0}^{m-1}(i+j) &= \sum_{i=1}^{n-1}\prod_{j=0}^{m-1}(i+j) + \prod_{j=0}^{m-1}(n+j) \\
&= \frac{1}{m+1}\prod_{j=0}^{m}(n-1+j) + \prod_{j=0}^{m-1}(n+j) \\
&= \frac{n-1}{m+1}\prod_{j=0}^{m-1}(n+j) + \prod_{j=0}^{m-1}(n+j) \\
&= \frac{n+m}{m+1}\prod_{j=0}^{m-1}(n+j) \\
&= \frac{1}{m+1}\prod_{j=1}^{m}(n+j)
\end{aligned}$$

Therefore we have proven the Lemma for general n. □

Theorem 6.2.2. *Suppose that C is a chromothripsis string of length m derived from a reference genome G composed of n intervals. The probability that $\pi(C)$ is H/T alternating is $\frac{1}{2(m-1)}$.*

Proof. The probability that $\pi(C)$ is alternating is just the ratio of the number of possible H/T alternating chromothripsis strings C of length m divided by all chromothripsis strings C of length m.

$$
\begin{aligned}
P(\pi(C) \text{ alternates } |m,n) &= \frac{|A(m,n)|}{|G(m,n)|} \\
&= \frac{2^{m-1} n \sum_{i=m-2}^{n-2} \left(\prod_{j=0}^{m-3} (i-j) \right)}{2^m \prod_{i=0}^{m-1} (n-i)} \\
&= \frac{\sum_{i=m-2}^{n-2} \left(\prod_{j=0}^{m-3} (i-j) \right)}{2 \prod_{i=1}^{m-1} (n-i)} \\
&= \frac{\sum_{i=1}^{n-m+1} (\prod_{j=0}^{m-3} (i+j))}{2 \prod_{i=1}^{m-1} (n-i)} \\
&= \frac{\frac{1}{m-1} \prod_{j=0}^{m-2} ((n-m+1)+j)}{2 \prod_{i=1}^{m-1} (n-i)} \qquad \text{(by Lemma D.1.1)} \\
&= \frac{\frac{1}{m-1} \prod_{j=1}^{m-1} (n-j)}{2 \prod_{i=1}^{m-1} (n-i)} \\
&= \frac{1}{2(m-1)}
\end{aligned}
$$

□

D.2 Preliminary Results on Real Data

In this section we present some preliminary results on real data which led to some of the ideas and results presented in [170]. Here, we classify observed breakpoints into categories in order to see if a genome has a configuration that under a sequential model would require breakpoint re-use, and therefore may be better explained by a one-time chromothripsis event. Novel tumor adjacencies detected using high-throughput sequencing are generally reported as a pair of breakpoints, where each breakpoint is defined by a chromosome, a genomic position on that chromosome and an orientation $(+/-)$ which indicates which strand of the genome the reads supporting the adjacency align to. An adjacency between such a pair of breakpoints indicates that the corresponding portions of the genome have been "glued" together. An orientation of $+$ indicates that the block to the left of the breakpoint position is involved in the gluing operation. Similarly an orientation of $-$ indicates that the block to the right of the breakpoint position is involved in the gluing operations.

We classify all observed breakpoints into three separate categories: (1) unpaired, (2) paired, and

(3) other. Since DNA sequence data is imperfect, we rarely expect to exactly measure the exact location of a breakpoint and therefore need to allow for some variation in reported breakpoint locations. Therefore, we define an *unpaired* breakpoint as an observed breakpoint where no other breakpoints are observed within a fixed *breakpoint distance* L in any direction on the same chromosome. A *paired breakpoint* is a breakpoint that where exactly one other breakpoint occurs within a fixed breakpoint distance L (and no other breakpoint occurs within distance L of that breakpoint) and the observed breakpoints align to opposite strands of the genome (one with orientation + and one with $-$). Any other breakpoint is classified as a *other*. Lastly, we can also break the paired breakpoints into two separate classes: (1) mixed paired - where the two breakpoints are part of two separate adjacencies measured and (2) simple paired - where the two breakpoints are the ends of a single adjacency (i.e. a deletion that is smaller than L).

A large number of mixed paired breakpoints, which can not be explained by simple events such as a reciprocal inversion would be an indication of chromothripsis, as these types of breakpoints indicate that either breakpoint re-use occurred or that these adjacencies occurred simultaneously. We examined 24 samples from 5 studies [114, 154, 97, 139, 183] which had been been previous classified as chromothripsis or step-wise (including B/F/B which is a sequential mechanism of rearrangement [183]). No reciprocal inversions were included in this analysis. We compared the distribution of breakpoints in these samples that were classified as mixed paired versus unpaired and found that this was able to distinguish several of the genomes called as chromothripsis (Figure D.1).

Further refinement of this signature is necessary to determine if some version may be appropriate for determine whether or not chromothripsis was likely to have occurred.

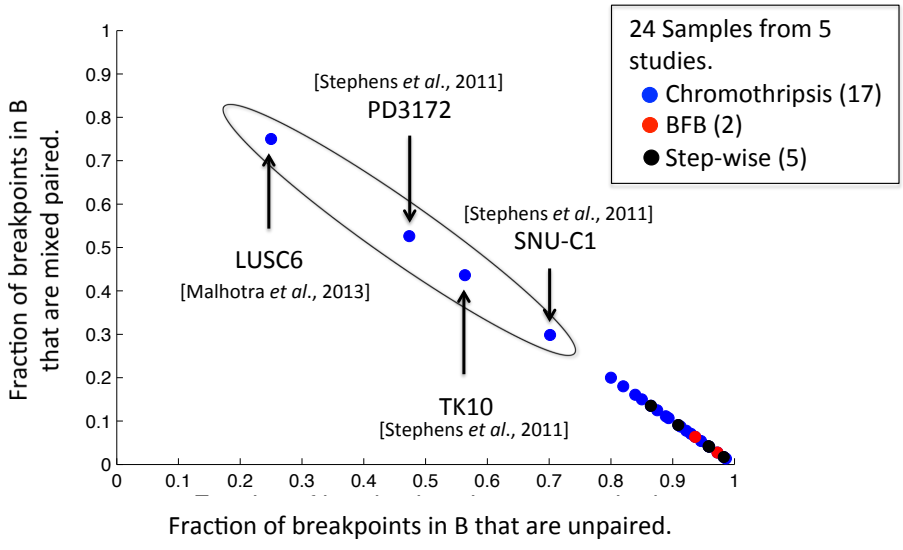


Figure D.1: **A comparison of the fraction of observed breakpoints classified as mixed paired versus unpaired for 24 genomes.** This includes, 17 previously classified as chromothripsis (blue), 2 previously classified to contain a B/F/B cycle (red) and 5 that have previously classified as the result of a step-wise sequence of events (black). The genomes with the largest fraction of mixed paired breakpoints (circled in gray) are all genomes that have been previously indicated to contain a chromothripsis event.

Bibliography

[1] Haley J Abel and Eric J Duncavage. Detection of structural dna variation from next generation sequencing data: a review of informatic approaches. *Cancer Genet*, 206(12):432–40, Dec 2013.

[2] Jacqui Adams, Sarah V Williams, Joanne S Aveyard, and Margaret A Knowles. Loss of heterozygosity analysis and dna copy number measurement on 8p in bladder cancer reveals two mechanisms of allelic loss. *Cancer Res*, 65(1):66–75, Jan 2005.

[3] Donna G Albertson, Colin Collins, Frank McCormick, and Joe W Gray. Chromosome aberrations in solid tumors. *Nat Genet*, 34(4):369–76, Aug 2003.

[4] Max A. Alekseyev and Pavel A. Pevzner. Multi-break rearrangements and chromosomal evolution. *Theoretical Computer Science*, 395(2–3):193 – 202, 2008. {SAIL} – String Algorithms, Information and Learning: Dedicated to Professor Alberto Apostolico on the occasion of his 60th birthday.

[5] Max A Alekseyev and Pavel A Pevzner. Breakpoint graphs and ancestral genome reconstructions. *Genome Res*, 19(5):943–57, May 2009.

[6] Can Alkan, Bradley P Coe, and Evan E Eichler. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–76, May 2011.

[7] R Anbazhagan, H Fujii, and E Gabrielson. Allelic loss of chromosomal arm 8p in breast cancer progression. *Am J Pathol*, 152(3):815–9, Mar 1998.

[8] Noemi Andor, Julie V Harness, Sabine Müller, Hans W Mewes, and Claudia Petritsch. Expands: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics*, 30(1):50–60, Jan 2014.

[9] Guillaume Assié, Thomas LaFramboise, Petra Platzer, Jérôme Bertherat, Constantine A Stratakis, and Charis Eng. Snp arrays in heterogeneous tissue: highly accurate collection of both germline and somatic genetic information from unpaired single tumor samples. *Am J Hum Genet*, 82(4):903–15, Apr 2008.

[10] Edward F Attiyeh, Sharon J Diskin, Marc A Attiyeh, Yaël P Mossé, Cuiping Hou, Eric M Jackson, Cecilia Kim, Joseph Glessner, Hakon Hakonarson, Jaclyn A Biegel, and John M Maris. Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Res*, 19(2):276–83, Feb 2009.

[11] Sylvan C Baca, Davide Prandi, Michael S Lawrence, Juan Miguel Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, Naoki Kitabayashi, Theresa Y MacDonald, Mahmoud Ghandi, Eliezer Van Allen, Gregory V Kryukov, Andrea Sboner, Jean-Philippe Theurillat, T David Soong, Elizabeth Nickerson, Daniel Auclair, Ashutosh Tewari, Himisha Beltran, Robert C Onofrio, Gunther Boysen, Candace Guiducci, Christopher E Barbieri, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Gordon Saksena, Douglas Voet, Alex H Ramos, Wendy Winckler, Michelle Cipicchio, Kristin Ardlie, Philip W Kantoff, Michael F Berger, Stacey B Gabriel, Todd R Golub, Matthew Meyerson, Eric S Lander, Olivier Elemento, Gad Getz, Francesca Demichelis, Mark A Rubin, and Levi A Garraway. Punctuated evolution of prostate cancer genomes. *Cell*, 153(3):666–77, Apr 2013.

[12] Lei Bao, Minya Pu, and Karen Messer. Abscn-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. *Bioinformatics*, Jan 2014.

[13] Ali Bashir, Stanislav Volik, Colin Collins, Vineet Bafna, and Benjamin J Raphael. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput Biol*, 4(4):e1000051, Apr 2008.

[14] Timour Baslan, Jude Kendall, Linda Rodgers, Hilary Cox, Mike Riggs, Asya Stepansky, Jennifer Troge, Kandasamy Ravi, Diane Esposito, B Lakshmi, Michael Wigler, Nicholas Navin, and James Hicks. Genome-wide copy number analysis of single cells. *Nat Protoc*, 7(6):1024–41, Jun 2012.

[15] Yuval Benjamini and Terence P Speed. Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic Acids Res*, 40(10):e72, May 2012.

[16] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, Jonathan M Boutell, Jason Bryant, Richard J Carter, R Keira Cheetham, Anthony J Cox, Darren J Ellis, Michael R Flatbush, Niall A Gormley, Sean J Humphray, Leslie J Irving, Mirian S Karbelashvili, Scott M Kirk, Heng Li, Xiaohai Liu, Klaus S Maisinger, Lisa J Murray, Bojan Obradovic, Tobias Ost, Michael L Parkinson, Mark R Pratt, Isabelle M J Rasolonjatovo, Mark T Reed, Roberto Rigatti, Chiara Rodighiero, Mark T Ross, Andrea Sabot, Subramanian V Sankar, Aylwyn Scally, Gary P Schroth, Mark E Smith, Vincent P Smith, Anastassia Spiridou, Peta E Torrance, Svilen S Tzonev, Eric H Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D Alam, Carole Anastasi, Ify C Aniebo, David M D Bailey, Iain R Bancarz, Saibal Banerjee, Selena G Barbour, Primo A Baybayan, Vincent A Benoit, Kevin F Benson, Claire Bevis, Phillip J Black, Asha Boodhun, Joe S Brennan, John A Bridgham, Rob C Brown, Andrew A Brown, Dale H Buermann, Abass A Bundu, James C Burrows, Nigel P Carter, Nestor Castillo, Maria Chiara E Catenazzi, Simon Chang, R Neil Cooley, Natasha R Crake, Olubunmi O Dada, Konstantinos D Diakoumakos, Belen Dominguez-Fernandez, David J Earnshaw, Ugonna C Egbujor, David W Elmore, Sergey S Etchin, Mark R Ewan, Milan Fedurco, Louise J Fraser, Karin V Fuentes Fajardo, W Scott Furey, David George, Kimberley J Gietzen, Colin P Goddard, George S Golda, Philip A Granieri, David E Green, David L Gustafson, Nancy F Hansen, Kevin Harnish, Christian D Haudenschild, Narinder I Heyer, Matthew M Hims, Johnny T Ho, Adrian M Horgan, Katya Hoschler, Steve Hurwitz, Denis V Ivanov, Maria Q Johnson, Terena James, T A Huw Jones, Gyoung-Dong Kang, Tzvetana H Kerelska, Alan D Kersey, Irina Khrebtukova, Alex P Kindwall, Zoya Kingsbury, Paula I Kokko-Gonzales, Anil Kumar, Marc A Laurent, Cynthia T Lawley, Sarah E Lee, Xavier Lee, Arnold K Liao, Jennifer A Loch, Mitch Lok, Shujun Luo, Radhika M Mammen, John W Martin, Patrick G McCauley, Paul McNitt, Parul Mehta, Keith W Moon, Joe W Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M Novo, Michael J O'Neill, Mark A Osborne, Andrew Osnowski, Omead Ostadan, Lambros L Paraschos, Lea Pickering, Andrew C Pike, Alger C Pike, D Chris Pinkard, Daniel P Pliskin, Joe Podhasky, Victor J

Quijano, Come Raczy, Vicki H Rae, Stephen R Rawlings, Ana Chiva Rodriguez, Phyllida M Roe, John Rogers, Maria C Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K Roth, Natalie J Rourke, Silke T Ruediger, Eli Rusman, Raquel M Sanches-Kuiper, Martin R Schenker, Josefina M Seoane, Richard J Shaw, Mitch K Shiver, Steven W Short, Ning L Sizto, Johannes P Sluis, Melanie A Smith, Jean Ernest Sohna Sohna, Eric J Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L Tregidgo, Gerardo Turcatti, Stephanie Vandevondele, Yuli Verhovsky, Selene M Virk, Suzanne Wakelin, Gregory C Walcott, Jingwen Wang, Graham J Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C Mullikin, Matthew E Hurles, Nick J McCooke, John S West, Frank L Oaks, Peter L Lundberg, David Klenerman, Richard Durbin, and Anthony J Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–9, Nov 2008.

[17] Michael F Berger, Michael S Lawrence, Francesca Demichelis, Yotam Drier, Kristian Cibulskis, Andrey Y Sivachenko, Andrea Sboner, Raquel Esgueva, Dorothee Pflueger, Carrie Sougnez, Robert Onofrio, Scott L Carter, Kyung Park, Lukas Habegger, Lauren Ambrogio, Timothy Fennell, Melissa Parkin, Gordon Saksena, Douglas Voet, Alex H Ramos, Trevor J Pugh, Jane Wilkinson, Sheila Fisher, Wendy Winckler, Scott Mahan, Kristin Ardlie, Jennifer Baldwin, Jonathan W Simons, Naoki Kitabayashi, Theresa Y MacDonald, et al. The genomic complexity of primary human prostate cancer. *Nature*, 470(7333):214–20, Feb 2011.

[18] Chetan Bettegowda, Mark Sausen, Rebecca J Leary, Isaac Kinde, Yuxuan Wang, Nishant Agrawal, Bjarne R Bartlett, Hao Wang, Brandon Luber, Rhoda M Alani, Emmanuel S Antonarakis, Nilofer S Azad, Alberto Bardelli, Henry Brem, John L Cameron, Clarence C Lee, Leslie A Fecher, Gary L Gallia, Peter Gibbs, Dung Le, Robert L Giuntoli, Michael Goggins, Michael D Hogarty, Matthias Holdhoff, Seung-Mo Hong, Yuchen Jiao, Hartmut H Juhl, Jenny J Kim, Giulia Siravegna, Daniel A Laheru, Calogero Lauricella, Michael Lim, Evan J Lipson, Suely Kazue Nagahashi Marie, George J Netto, Kelly S Oliner, Alessandro Olivi, Louise Olsson, Gregory J Riggins, Andrea Sartore-Bianchi, Kerstin Schmidt, le-Ming Shih, Sueli Mieko Oba-Shinjo, Salvatore Siena, Dan Theodorescu, Jeanne Tie, Timothy T Harkins, Silvio Veronese, Tian-Li Wang, Jon D Weingart, Christopher L Wolfgang, Laura D Wood, Dongmei Xing, Ralph H Hruban, Jian Wu, Peter J Allen, C Max Schmidt, Michael A Choti, Victor E Velculescu, Kenneth W Kinzler, Bert Vogelstein, Nickolas Papadopoulos, and Luis A

Diaz, Jr. Detection of circulating tumor dna in early- and late-stage human malignancies. *Sci Transl Med*, 6(224):224ra24, Feb 2014.

[19] Graham R Bignell, Chris D Greenman, Helen Davies, Adam P Butler, Sarah Edkins, Jenny M Andrews, Gemma Buck, Lina Chen, David Beare, Calli Latimer, Sara Widaa, Jonathon Hinton, Ciara Fahey, Beiyuan Fu, Sajani Swamy, Gillian L Dalgliesh, Bin T Teh, Panos Deloukas, Fengtang Yang, Peter J Campbell, P Andrew Futreal, and Michael R Stratton. Signatures of mutation and selection in the cancer genome. *Nature*, 463(7283):893–8, Feb 2010.

[20] B M Bolstad, R A Irizarry, M Astrand, and T P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–93, Jan 2003.

[21] Rebecca A Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–45, Sep 2013.

[22] Peter J Campbell, Philip J Stephens, Erin D Pleasance, Sarah O'Meara, Heng Li, Thomas Santarius, Lucy A Stebbings, Catherine Leroy, Sarah Edkins, Claire Hardy, Jon W Teague, Andrew Menzies, Ian Goodhead, Daniel J Turner, Christopher M Clee, Michael A Quail, Antony Cox, Clive Brown, Richard Durbin, Matthew E Hurles, Paul A W Edwards, Graham R Bignell, Michael R Stratton, and P Andrew Futreal. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*, 40(6):722–9, Jun 2008.

[23] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, Oct 2012.

[24] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–15, Jun 2011.

[25] Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*, 368(22):2059–74, May 2013.

[26] Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*, 368(22):2059–74, May 2013.

[27] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, Rameen Beroukhim, David Pellman, Douglas A Levine, Eric S Lander, Matthew Meyerson, and Gad Getz. Absolute quantification of somatic dna alterations in human cancer. *Nat Biotechnol*, 30(5):413–21, May 2012.

[28] A Castells, J F Gusella, V Ramesh, and A K Rustgi. A region of deletion on chromosome 22q13 is common to human breast and colorectal cancers. *Cancer Res*, 60(11):2836–9, Jun 2000.

[29] L R Cavalli, L M Cavaliéri, L A Ribeiro, I J Cavalli, R Silveira, and S R Rogatto. Cytogenetic evaluation of 20 primary breast carcinomas. *Hereditas*, 126(3):261–8, 1997.

[30] Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

[31] Ken Chen, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, Xiaoqi Shi, Robert S Fulton, Timothy J Ley, Richard K Wilson, Li Ding, and Elaine R Mardis. Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*, 6(9):677–81, Sep 2009.

[32] T Chen, A Sahin, and C M Aldaz. Deletion map of chromosome 16q in ductal carcinoma in situ of the breast: refining a putative tumor suppressor gene region. *Cancer Res*, 56(24):5605–9, Dec 1996.

[33] Derek Y Chiang, Gad Getz, David B Jaffe, Michael J T O'Kelly, Xiaojun Zhao, Scott L Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S Lander. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods*, 6(1):99–103, Jan 2009.

[34] Lynda Chin, Jannik N Andersen, and P Andrew Futreal. Cancer genomics: from discovery science to personalized medicine. *Nat Med*, 17(3):297–303, Mar 2011.

[35] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe,

Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*, 31(3):213–9, Mar 2013.

[36] John Cook. Exact calculation of beta inequalities, 2005.

[37] Jiajun Cui, Katherine Germer, Tianying Wu, Jiang Wang, Jia Luo, Shao-chun Wang, Qianben Wang, and Xiaoting Zhang. Cross-talk between her2 and med1 regulates tamoxifen resistance of human breast cancer cells. *Cancer Res*, 72(21):5625–34, Nov 2012.

[38] R Dalla-Favera, M Bregni, J Erikson, D Patterson, R C Gallo, and C M Croce. Human c-myc onc gene is located on the region of chromosome 8 that is translocated in burkitt lymphoma cells. *Proc Natl Acad Sci U S A*, 79(24):7824–7, Dec 1982.

[39] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: Reconstructing subclonal composition and evolution from whole genome sequencing of tumors. *Genome Biol*, In Press.

[40] Ninad Dewal, Yang Hu, Matthew L Freedman, Thomas Laframboise, and Itsik Pe'er. Calling amplified haplotypes in next generation tumor sequence data. *Genome Res*, 22(2):362–74, Feb 2012.

[41] Li Ding, Timothy J Ley, David E Larson, Christopher A Miller, Daniel C Koboldt, John S Welch, Julie K Ritchey, Margaret A Young, Tamara Lamprecht, Michael D McLellan, Joshua F McMichael, John W Wallis, Charles Lu, Dong Shen, Christopher C Harris, David J Dooling, Robert S Fulton, Lucinda L Fulton, Ken Chen, Heather Schmidt, Joelle Kalicki-Veizer, Vincent J Magrini, Lisa Cook, Sean D McGrath, Tammi L Vickery, Michael C Wendl, Sharon Heath, Mark A Watson, Daniel C Link, Michael H Tomasson, William D Shannon, Jacqueline E Payton, Shashikant Kulkarni, Peter Westervelt, Matthew J Walter, Timothy A Graubert, Elaine R Mardis, Richard K Wilson, and John F DiPersio. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–10, Jan 2012.

[42] Li Ding, Benjamin J Raphael, Feng Chen, and Michael C Wendl. Advances for studying clonal evolution in cancer. *Cancer Lett*, 340(2):212–9, Nov 2013.

[43] Li Ding, Michael C Wendl, Daniel C Koboldt, and Elaine R Mardis. Analysis of next-generation genomic data in cancer: accomplishments and challenges. *Hum Mol Genet*, 19(R2):R188–96, Oct 2010.

[44] Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic Acids Res*, 36(16):e105, Sep 2008.

[45] K Driouch, F Dorion-Bonnet, M Briffod, M H Champéme, M Longy, and R Lidereau. Loss of heterozygosity on chromosome arm 16q in breast cancer metastases. *Genes Chromosomes Cancer*, 19(3):185–91, Jul 1997.

[46] B J Druker, M Talpaz, D J Resta, B Peng, E Buchdunger, J M Ford, N B Lydon, H Kantarjian, R Capdeville, S Ohno-Jones, and C L Sawyers. Efficacy and safety of a specific inhibitor of the bcr-abl tyrosine kinase in chronic myeloid leukemia. *N Engl J Med*, 344(14):1031–7, Apr 2001.

[47] Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field, and Benjamin J. Raphael. Reconstruction of clonal trees and tumor composition from multi-sample cancer sequencing data. *ISMB*, To Appear.

[48] Andrej Fischer, Ignacio Vázquez-García, Christopher J R Illingworth, and Ville Mustonen. High-definition reconstruction of clonal composition in cancer. *Cell Rep*, 7(5):1740–52, Jun 2014.

[49] Jane Fridlyand, Antoine M. Snijders, Dan Pinkel, Donna G. Albertson, and Ajay N. Jain. Hidden markov models approach to the analysis of array {CGH} data. *Journal of Multivariate Analysis*, 90(1):132 – 153, 2004. Special Issue on Multivariate Methods in Genomic Data Analysis.

[50] Jane Fridlyand, Antoine M Snijders, Bauke Ylstra, Hua Li, Adam Olshen, Richard Segraves, Shanaz Dairkee, Taku Tokuyasu, Britt Marie Ljung, Ajay N Jain, Jane McLennan, John Ziegler, Koei Chin, Sandy Devries, Heidi Feiler, Joe W Gray, Frederic Waldman, Daniel Pinkel, and Donna G Albertson. Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer*, 6:96, 2006.

[51] Levi A Garraway. Genomics-driven oncology: framework for an emerging paradigm. *J Clin Oncol*, 31(15):1806–14, May 2013.

[52] Marco Gerlinger et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet*, 46(3):225–33, Mar 2014.

[53] Marco Gerlinger, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, Ignacio Varela, Benjamin Phillimore, Sharmin Begum, Neil Q McDonald, Adam Butler, David Jones, Keiran Raine, Calli Latimer, Claudio R Santos, Mahrokh Nohadani, Aron C Eklund, Bradley Spencer-Dene, Graham Clark, Lisa Pickering, Gordon Stamp, Martin Gore, Zoltan Szallasi, Julian Downward, P Andrew Futreal, and Charles Swanton. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*, 366(10):883–92, Mar 2012.

[54] David J Gordon, Benjamin Resio, and David Pellman. Causes and consequences of aneuploidy in cancer. *Nat Rev Genet*, 13(3):189–203, Mar 2012.

[55] Shaylan K Govind, Amin Zia, Pablo H Hennings-Yeomans, John D Watson, Michael Fraser, Catalina Anghel, Alexander W Wyatt, Theodorus van der Kwast, Colin C Collins, John D McPherson, Robert G Bristow, and Paul C Boutros. Shatterproof: operational detection and quantification of chromothripsis. *BMC Bioinformatics*, 15:78, 2014.

[56] Mel Greaves and Carlo C Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–13, Jan 2012.

[57] Chris D Greenman, Graham Bignell, Adam Butler, Sarah Edkins, Jon Hinton, Dave Beare, Sajani Swamy, Thomas Santarius, Lina Chen, Sara Widaa, P Andy Futreal, and Michael R Stratton. Picnic: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, 11(1):164–75, Jan 2010.

[58] Chris D Greenman, Erin D Pleasance, Scott Newman, Fengtang Yang, Beiyuan Fu, Serena Nik-Zainal, David Jones, King Wai Lau, Nigel Carter, Paul A W Edwards, P Andrew Futreal, Michael R Stratton, and Peter J Campbell. Estimation of rearrangement phylogeny for cancer genomes. *Genome Res*, 22(2):346–61, Feb 2012.

[59] Christopher Greenman, Philip Stephens, Raffaella Smith, Gillian L Dalgliesh, Christopher Hunter, Graham Bignell, Helen Davies, Jon Teague, Adam Butler, Claire Stevens, Sarah Edkins, Sarah O'Meara, Imre Vastrik, Esther E Schmidt, Tim Avis, Syd Barthorpe, Gurpreet Bhamra, Gemma Buck, Bhudipa Choudhury, Jody Clements, Jennifer Cole, Ed Dicks, Simon Forbes, Kris Gray, Kelly Halliday, Rachel Harrison, Katy Hills, Jon Hinton, Andy Jenkinson, David Jones, Andy Menzies, Tatiana Mironenko, Janet Perry, Keiran Raine, Dave Richardson, Rebecca Shepherd, Alexandra Small, Calli Tofts, Jennifer Varian, Tony Webb, Sofie West, Sara Widaa, Andy Yates, Daniel P Cahill, David N Louis, Peter Goldstraw, Andrew G Nicholson, Francis Brasseur, Leendert Looijenga, Barbara L Weber, Yoke-Eng Chiew, Anna DeFazio, Mel F Greaves, Anthony R Green, Peter Campbell, Ewan Birney, Douglas F Easton, Georgia Chenevix-Trench, Min-Han Tan, Sok Kean Khoo, Bin Tean Teh, Siu Tsan Yuen, Suet Yi Leung, Richard Wooster, P Andrew Futreal, and Michael R Stratton. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–8, Mar 2007.

[60] Dan Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21(1):19–28, 1991.

[61] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology*. Cambridge University Press, 1997.

[62] Arief Gusnanto, Henry M Wood, Yudi Pawitan, Pamela Rabbitts, and Stefano Berri. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics*, 28(1):40–7, Jan 2012.

[63] Gavin Ha, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M Prentice, Nataliya Melnyk, Andrew McPherson, Ali Bashashati, Emma Laks, Justina Biele, Jiarui Ding, Alan Le, Jamie Rosner, Karey Shumansky, Marco A Marra, C Blake Gilks, David G Huntsman, Jessica N McAlpine, Samuel Aparicio, and Sohrab P Shah. Titan: Inference of copy number architectures in clonal cell populations from tumor whole genome sequence data. *Genome Res*, Jul 2014.

[64] Iman Hajirasouliha, Ahmad Mahmoody, and Benjamin J Raphael. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, 30(12):i78–i86, Jun 2014.

[65] Iman Hajirasouliha and Benjamin J. Raphael. Reconstructing mutational history in multiply sampled tumors using perfect phylogeny mixtures. In *Algorithms in Bioinformatics - 14th International Workshop, WABI 2014, Wroclaw, Poland, September 8-10, 2014. Proceedings*, pages 354–367, 2014.

[66] Rebecca S Heist, Lecia V Sequist, and Jeffrey A Engelman. Genetic changes in squamous cell lung cancer: a review. *J Thorac Oncol*, 7(5):924–33, May 2012.

[67] R. Hemmeke, M. Koppe, J. Lee, and R. Weismantel. Nonlinear integer programming. In Michael Junger, Thomas M. Liebling, Denis Naddef, George L. Nemhauser, William R. Pulleyblank, Gerhard Reinelt, Giovanni Rinaldi, and Laurence A. Wolsey, editors, *50 Years of Integer Programming 1958-2008*, pages 561–618. Springer Berlin Heidelberg, 2010.

[68] James Hicks, Alexander Krasnitz, B Lakshmi, Nicholas E Navin, Michael Riggs, Evan Leibu, Diane Esposito, Joan Alexander, Jen Troge, Vladimir Grubor, Seungtai Yoon, Michael Wigler, Kenny Ye, Anne-Lise Børresen-Dale, Bjørn Naume, Ellen Schlicting, Larry Norton, Torsten Hägerström, Lambert Skoog, Gert Auer, Susanne Månér, Pär Lundin, and Anders Zetterberg. Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res*, 16(12):1465–79, Dec 2006.

[69] D. S. Hochbaum and J. George Shanthikumar. Convex separable optimization is not much harder than linear optimization. *J. ACM*, 37(4):843–862, October 1990.

[70] Fereydoun Hormozdiari, Can Alkan, Evan E Eichler, and S Cenk Sahinalp. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res*, 19(7):1270–8, Jul 2009.

[71] Yong Hou, Luting Song, Ping Zhu, Bo Zhang, Ye Tao, Xun Xu, Fuqiang Li, Kui Wu, Jie Liang, Di Shao, Hanjie Wu, Xiaofei Ye, Chen Ye, Renhua Wu, Min Jian, Yan Chen, Wei Xie, Ruren Zhang, Lei Chen, Xin Liu, Xiaotian Yao, Hancheng Zheng, Chang Yu, Qibin Li, Zhuolin Gong, Mao Mao, Xu Yang, Lin Yang, Jingxiang Li, Wen Wang, Zuhong Lu, Ning Gu, Goodman Laurie, Lars Bolund, Karsten Kristiansen, Jian Wang, Huanming Yang, Yingrui Li, Xiuqing Zhang, and Jun Wang. Single-cell exome sequencing and monoclonal evolution of a jak2-negative myeloproliferative neoplasm. *Cell*, 148(5):873–85, Mar 2012.

[72] https://www.synapse.org/#!Synapse:syn2813581/wiki/. ICGC-TCGA DREAM somatic mutation calling challenge - tumor heterogeneity and evolution.

[73] http://www.cancer.net/cancer-types/leukemia-chronic-myeloid cml/statistics. Leukemia - chronic myeloid - cml: Statistics.

[74] International Cancer Genome Consortium, Thomas J Hudson, Warwick Anderson, Axel Artez, Anna D Barker, Cindy Bell, Rosa R Bernabé, M K Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, Alan Guttmacher, Mark Guyer, Fiona M Hemsley, Jennifer L Jennings, David Kerr, Peter Klatt, Patrik Kolar, Jun Kusada, David P Lane, Frank Laplace, Lu Youyong, Gerd Nettekoven, Brad Ozenberger, Jane Peterson, T S Rao, Jacques Remacle, Alan J Schafer, Tatsuhiro Shibata, Michael R Stratton, Joseph G Vockley, Koichi Watanabe, Huanming Yang, Matthew M F Yuen, Bartha M Knoppers, Martin Bobrow, Anne Cambon-Thomsen, Lynn G Dressler, Stephanie O M Dyke, Yann Joly, Kazuto Kato, Karen L Kennedy, Pilar Nicolás, Michael J Parker, Emmanuelle Rial-Sebbag, Carlos M Romeo-Casabona, Kenna M Shaw, Susan Wallace, Georgia L Wiesner, Nikolajs Zeps, Peter Lichter, Andrew V Biankin, Christian Chabannon, Lynda Chin, Bruno Clément, Enrique de Alava, Françoise Degos, Martin L Ferguson, Peter Geary, D Neil Hayes, Thomas J Hudson, Amber L Johns, Arek Kasprzyk, Hidewaki Nakagawa, Robert Penny, Miguel A Piris, Rajiv Sarin, Aldo Scarpa, Tatsuhiro Shibata, Marc van de Vijver, P Andrew Futreal, Hiroyuki Aburatani, Mónica Bayés, David D L Botwell, Peter J Campbell, Xavier Estivill, Daniela S Gerhard, Sean M Grimmond, Ivo Gut, Martin Hirst, Carlos López-Otín, Partha Majumder, Marco Marra, John D McPherson, Hidewaki Nakagawa, Zemin Ning, Xose S Puente, Yijun Ruan, Tatsuhiro Shibata, Michael R Stratton, Hendrik G Stunnenberg, Harold Swerdlow, Victor E Velculescu, Richard K Wilson, Hong H Xue, Liu Yang, Paul T Spellman, Gary D Bader, Paul C Boutros, Peter J Campbell, Paul Flicek, Gad Getz, Roderic Guigó, Guangwu Guo, David Haussler, Simon Heath, Tim J Hubbard, Tao Jiang, Steven M Jones, Qibin Li, Nuria López-Bigas, Ruibang Luo, Lakshmi Muthuswamy, B F Francis Ouellette, John V Pearson, Xose S Puente, Victor Quesada, Benjamin J Raphael, Chris Sander, Tatsuhiro Shibata, Terence P Speed, Lincoln D Stein, Joshua M Stuart, Jon W Teague, Yasushi Totoki, Tatsuhiko Tsunoda, Alfonso Valencia, David A Wheeler, Honglong Wu, Shancen Zhao, Guangyu Zhou, Lincoln D Stein, Roderic Guigó, Tim J Hubbard, Yann Joly, Steven M Jones, Arek Kasprzyk, Mark Lathrop, Nuria

López-Bigas, B F Francis Ouellette, Paul T Spellman, Jon W Teague, Gilles Thomas, Alfonso Valencia, Teruhiko Yoshida, Karen L Kennedy, Myles Axton, Stephanie O M Dyke, P Andrew Futreal, Daniela S Gerhard, Chris Gunter, Mark Guyer, Thomas J Hudson, John D McPherson, Linda J Miller, Brad Ozenberger, Kenna M Shaw, Arek Kasprzyk, Lincoln D Stein, Junjun Zhang, Syed A Haider, Jianxin Wang, Christina K Yung, Anthony Cros, Anthony Cross, Yong Liang, Saravanamuttu Gnaneshan, Jonathan Guberman, Jack Hsu, Martin Bobrow, Don R C Chalmers, Karl W Hasel, Yann Joly, Terry S H Kaan, Karen L Kennedy, Bartha M Knoppers, William W Lowrance, Tohru Masui, Pilar Nicolás, Emmanuelle Rial-Sebbag, Laura Lyman Rodriguez, Catherine Vergely, Teruhiko Yoshida, Sean M Grimmond, Andrew V Biankin, David D L Bowtell, Nicole Cloonan, Anna deFazio, James R Eshleman, Dariush Etemadmoghadam, Brooke B Gardiner, Brooke A Gardiner, James G Kench, Aldo Scarpa, Robert L Sutherland, Margaret A Tempero, Nicola J Waddell, Peter J Wilson, John D McPherson, Steve Gallinger, Ming-Sound Tsao, Patricia A Shaw, Gloria M Petersen, Debabrata Mukhopadhyay, Lynda Chin, Ronald A DePinho, Sarah Thayer, Lakshmi Muthuswamy, Kamran Shazand, Timothy Beck, Michelle Sam, Lee Timms, Vanessa Ballin, Youyong Lu, Jiafu Ji, Xiuqing Zhang, Feng Chen, Xueda Hu, Guangyu Zhou, Qi Yang, Geng Tian, Lianhai Zhang, Xiaofang Xing, Xianghong Li, Zhenggang Zhu, Yingyan Yu, Jun Yu, Huanming Yang, Mark Lathrop, Jörg Tost, Paul Brennan, Ivana Holcatova, David Zaridze, Alvis Brazma, Lars Egevard, Egor Prokhortchouk, Rosamonde Elizabeth Banks, Mathias Uhlén, Anne Cambon-Thomsen, Juris Viksna, Fredrik Ponten, Konstantin Skryabin, Michael R Stratton, P Andrew Futreal, Ewan Birney, Ake Borg, Anne-Lise Børresen-Dale, Carlos Caldas, John A Foekens, Sancha Martin, Jorge S Reis-Filho, Andrea L Richardson, Christos Sotiriou, Hendrik G Stunnenberg, Giles Thoms, Marc van de Vijver, Laura van't Veer, Fabien Calvo, Daniel Birnbaum, Hélène Blanche, Pascal Boucher, Sandrine Boyault, Christian Chabannon, Ivo Gut, Jocelyne D Masson-Jacquemier, Mark Lathrop, Iris Pauporté, Xavier Pivot, Anne Vincent-Salomon, Eric Tabone, Charles Theillet, Gilles Thomas, Jörg Tost, Isabelle Treilleux, Fabien Calvo, Paulette Bioulac-Sage, Bruno Clément, Thomas Decaens, Françoise Degos, Dominique Franco, Ivo Gut, Marta Gut, Simon Heath, Mark Lathrop, Didier Samuel, Gilles Thomas, Jessica Zucman-Rossi, Peter Lichter, Roland Eils, Benedikt Brors, Jan O Korbel, Andrey Korshunov, Pablo Landgraf, Hans Lehrach, Stefan Pfister, Bernhard Radlwimmer, Guido Reifenberger, Michael D Taylor, Christof von Kalle, Partha P Majumder, Rajiv Sarin, T S Rao, M K Bhan, Aldo Scarpa, Paolo

Pederzoli, Rita A Lawlor, Massimo Delledonne, Alberto Bardelli, Andrew V Biankin, Sean M Grimmond, Thomas Gress, David Klimstra, Giuseppe Zamboni, Tatsuhiro Shibata, Yusuke Nakamura, Hidewaki Nakagawa, Jun Kusada, Tatsuhiko Tsunoda, Satoru Miyano, Hiroyuki Aburatani, Kazuto Kato, Akihiro Fujimoto, Teruhiko Yoshida, Elias Campo, Carlos López-Otín, Xavier Estivill, Roderic Guigó, Silvia de Sanjosé, Miguel A Piris, Emili Montserrat, Marcos González-Díaz, Xose S Puente, Pedro Jares, Alfonso Valencia, Heinz Himmelbauer, Heinz Himmelbaue, Victor Quesada, Silvia Bea, Michael R Stratton, P Andrew Futreal, Peter J Campbell, Anne Vincent-Salomon, Andrea L Richardson, Jorge S Reis-Filho, Marc van de Vijver, Gilles Thomas, Jocelyne D Masson-Jacquemier, Samuel Aparicio, Ake Borg, Anne-Lise Børresen-Dale, Carlos Caldas, John A Foekens, Hendrik G Stunnenberg, Laura van't Veer, Douglas F Easton, Paul T Spellman, Sancha Martin, Anna D Barker, Lynda Chin, Francis S Collins, Carolyn C Compton, Martin L Ferguson, Daniela S Gerhard, Gad Getz, Chris Gunter, Alan Guttmacher, Mark Guyer, D Neil Hayes, Eric S Lander, Brad Ozenberger, Robert Penny, Jane Peterson, Chris Sander, Kenna M Shaw, Terence P Speed, Paul T Spellman, Joseph G Vockley, David A Wheeler, Richard K Wilson, Thomas J Hudson, Lynda Chin, Bartha M Knoppers, Eric S Lander, Peter Lichter, Lincoln D Stein, Michael R Stratton, Warwick Anderson, Anna D Barker, Cindy Bell, Martin Bobrow, Wylie Burke, Francis S Collins, Carolyn C Compton, Ronald A DePinho, Douglas F Easton, P Andrew Futreal, Daniela S Gerhard, Anthony R Green, Mark Guyer, Stanley R Hamilton, Tim J Hubbard, Olli P Kallioniemi, Karen L Kennedy, Timothy J Ley, Edison T Liu, Youyong Lu, Partha Majumder, Marco Marra, Brad Ozenberger, Jane Peterson, Alan J Schafer, Paul T Spellman, Hendrik G Stunnenberg, Brandon J Wainwright, Richard K Wilson, and Huanming Yang. International network of cancer genome projects. *Nature*, 464(7291):993–8, Apr 2010.

[75] Wei Jiao, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, 15:35, 2014.

[76] Yeonjoo Jung, Pora Kim, Yeonhwa Jung, Juhee Keum, Soon-Nam Kim, Yong Soo Choi, In-Gu Do, Jinseon Lee, So-Jung Choi, Sujin Kim, Jong-Eun Lee, Jhingook Kim, Sanghyuk Lee, and Jaesang Kim. Discovery of alk-ptpn3 gene fusion from human non-small cell lung carcinoma cell line using next generation rna sequencing. *Genes Chromosomes Cancer*, 51(6):590–7, Jun

2012.

[77] O P Kallioniemi, A Kallioniemi, W Kurisu, A Thor, L C Chen, H S Smith, F M Waldman, D Pinkel, and J W Gray. Erbb2 amplification in breast cancer analyzed by fluorescence in situ hybridization. *Proc Natl Acad Sci U S A*, 89(12):5321–5, Jun 1992.

[78] Donna Karolchik, Galt P Barber, Jonathan Casper, Hiram Clawson, Melissa S Cline, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, Rachel A Harte, Steve Heitner, Angie S Hinrichs, Katrina Learned, Brian T Lee, Chin H Li, Brian J Raney, Brooke Rhead, Kate R Rosenbloom, Cricket A Sloan, Matthew L Speir, Ann S Zweig, David Haussler, Robert M Kuhn, and W James Kent. The ucsc genome browser database: 2014 update. *Nucleic Acids Res*, 42(1):D764–70, Jan 2014.

[79] Päivikki Kauraniemi, Sampsa Hautaniemi, Reija Autio, Jaakko Astola, Outi Monni, Abdel Elkahloun, and Anne Kallioniemi. Effects of herceptin treatment on global gene expression patterns in her2-amplified and nonamplified breast cancer cell lines. *Oncogene*, 23(4):1010–3, Jan 2004.

[80] Linda E Kelemen, Marc T Goodman, Valerie McGuire, Mary Anne Rossing, Penelope M Webb, Australian Cancer Study (Ovarian Cancer) Study Group, Martin Köbel, Hoda Anton-Culver, Jonathan Beesley, Andrew Berchuck, Sony Brar, Michael E Carney, Jenny Chang-Claude, Georgia Chenevix-Trench, Australian Ovarian Cancer Study Group, Daniel W Cramer, Julie M Cunningham, Richard A Dicioccio, Jennifer A Doherty, Douglas F Easton, Zachary S Fredericksen, Brooke L Fridley, Margaret A Gates, Simon A Gayther, Aleksandra Gentry-Maharaj, Estrid Høgdall, Susanne Krüger Kjaer, Galina Lurie, Usha Menon, Patricia G Moorman, Kirsten Moysich, Roberta B Ness, Rachel T Palmieri, Celeste L Pearce, Paul D P Pharoah, Susan J Ramus, Honglin Song, Daniel O Stram, Shelley S Tworoger, David Van Den Berg, Robert A Vierkant, Shan Wang-Gohrke, Alice S Whittemore, Lynne R Wilkens, Anna H Wu, Joellen M Schildkraut, Thomas A Sellers, Ellen L Goode, and Ovarian Cancer Association Consortium. Genetic variation in tyms in the one-carbon transfer pathway is associated with ovarian carcinoma types in the ovarian cancer association consortium. *Cancer Epidemiol Biomarkers Prev*, 19(7):1822–30, Jul 2010.

[81] Marcus Kinsella, Anand Patel, and Vineet Bafna. The elusive evidence for chromothripsis. *Nucleic Acids Research*, 2014.

[82] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*, 22(3):568–76, Mar 2012.

[83] Jan O Korbel and Peter J Campbell. Criteria for inference of chromothripsis in cancer genomes. *Cell*, 152(6):1226–36, Mar 2013.

[84] Jan O Korbel, Alexander Eckehart Urban, Jason P Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M Kim, Dean Palejev, Nicholas J Carriero, Lei Du, Bruce E Taillon, Zhoutao Chen, Andrea Tanzer, A C Eugenia Saunders, Jianxiang Chi, Fengtang Yang, Nigel P Carter, Matthew E Hurles, Sherman M Weissman, Timothy T Harkins, Mark B Gerstein, Michael Egholm, and Michael Snyder. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–6, Oct 2007.

[85] Anton Kotzig. Moves without forbidden transitions in a graph. *Matematický časopis*, 18(1):76–80, 1968.

[86] Dan A Landau, Scott L Carter, Petar Stojanov, Aaron McKenna, Kristen Stevenson, Michael S Lawrence, Carrie Sougnez, Chip Stewart, Andrey Sivachenko, Lili Wang, Youzhong Wan, Wandi Zhang, Sachet A Shukla, Alexander Vartanov, Stacey M Fernandes, Gordon Saksena, Kristian Cibulskis, Bethany Tesar, Stacey Gabriel, Nir Hacohen, Matthew Meyerson, Eric S Lander, Donna Neuberg, Jennifer R Brown, Gad Getz, and Catherine J Wu. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, 152(4):714–26, Feb 2013.

[87] Nicholas B Larson and Brooke L Fridley. Purbayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics*, 29(15):1888–9, Aug 2013.

[88] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. Lumpy: a probabilistic framework for structural variant discovery. *Genome Biol*, 15(6):R84, 2014.

[89] Hayan Lee and Michael C Schatz. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics*, 28(16):2097–105, Aug 2012.

[90] Bo Li and Jun Z Li. A general framework for analyzing tumor subclonality using snp array and dna sequencing data. *Genome Biol*, 15(9):473, 2014.

[91] Yi Li and Xiaohui Xie. Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity. *Bioinformatics*, Apr 2014.

[92] Yi Li and Xiaohui Xie. Mixclone: a mixture model for inferring tumor subclonal populations. *BMC Genomics*, 16 Suppl 2:S1, 2015.

[93] Yilong Li, Claire Schwab, Sarra L Ryan, Elli Papaemmanuil, Hazel M Robinson, Patricia Jacobs, Anthony V Moorman, Sara Dyer, Julian Borrow, Mike Griffiths, Nyla A Heerema, Andrew J Carroll, Polly Talley, Nick Bown, Nick Telford, Fiona M Ross, Lorraine Gaunt, Richard J Q McNally, Bryan D Young, Paul Sinclair, Vikki Rand, Manuel R Teixeira, Olivia Joseph, Ben Robinson, Mark Maddison, Nicole Dastugue, Peter Vandenberghe, Claudia Haferlach, Philip J Stephens, Jiqiu Cheng, Peter Van Loo, Michael R Stratton, Peter J Campbell, and Christine J Harrison. Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature*, 508(7494):98–102, Apr 2014.

[94] X Lu, U Boora, L Seabra, E M Rabai, J Fenton, A Reiman, Z Nagy, and E R Maher. Knockdown of slingshot 2 (ssh2) serine phosphatase induces caspase3 activation in human carcinoma cell lines with the loss of the birt-hogg-dubé tumour suppressor gene (flcn). *Oncogene*, 33(8):956–65, Feb 2014.

[95] Alberto Magi, Lorenzo Tattini, Ingrid Cifola, Romina D'Aurizio, Matteo Benelli, Eleonora Mangano, Cristina Battaglia, Elena Bonora, Ants Kurg, Marco Seri, Pamela Magini, Betti Giusti, Giovanni Romeo, Tommaso Pippucci, Gianluca De Bellis, Rosanna Abbate, and Gian Franco Gensini. Excavator: detecting copy number variants from whole-exome sequencing data. *Genome Biol*, 14(10):R120, 2013.

[96] Christopher A Maher and Richard K Wilson. Chromothripsis and human disease: piecing together the shattering process. *Cell*, 148(1-2):29–32, Jan 2012.

[97] Ankit Malhotra, Michael Lindberg, Gregory G Faust, Mitchell L Leibowitz, Royden A Clark, Ryan M Layer, Aaron R Quinlan, and Ira M Hall. Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Res*, 23(5):762–76, May 2013.

[98] Salem Malikic, Andrew W McPherson, Nilgun Donmez, and Cenk S Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, Jan 2015.

[99] Andrew McPherson, Chunxiao Wu, Alexander W Wyatt, Sohrab Shah, Colin Collins, and S Cenk Sahinalp. nfuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Res*, 22(11):2250–61, Nov 2012.

[100] Paul Medvedev and Michael Brudno. Maximum likelihood genome assembly. *J Comput Biol*, 16(8):1101–16, Aug 2009.

[101] Paul Medvedev, Marc Fiume, Misko Dzamba, Tim Smith, and Michael Brudno. Detecting copy number variation with mated short reads. *Genome Res*, 20(11):1613–22, Nov 2010.

[102] Paul Medvedev, Monica Stanciu, and Michael Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods*, 6(11 Suppl):S13–20, Nov 2009.

[103] Matthew Meyerson, Stacey Gabriel, and Gad Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, 11(10):685–96, Oct 2010.

[104] Christopher A Miller, Oliver Hampton, Cristian Coarfa, and Aleksandar Milosavljevic. Readdepth: a parallel r package for detecting copy number alterations from short sequencing reads. *PLoS One*, 6(1):e16327, 2011.

[105] Christopher A Miller, Brian S White, Nathan D Dees, Malachi Griffith, John S Welch, Obi L Griffith, Ravi Vij, Michael H Tomasson, Timothy A Graubert, Matthew J Walter, Matthew J Ellis, William Schierding, John F DiPersio, Timothy J Ley, Elaine R Mardis, Richard K Wilson, and Li Ding. Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol*, 10(8):e1003665, Aug 2014.

[106] Ryan E Mills, Klaudia Walter, Chip Stewart, Robert E Handsaker, Ken Chen, Can Alkan, Alexej Abyzov, Seungtai Chris Yoon, Kai Ye, R Keira Cheetham, Asif Chinwalla, Donald F Conrad, Yutao Fu, Fabian Grubert, Iman Hajirasouliha, Fereydoun Hormozdiari, Lilia M Iakoucheva, Zamin Iqbal, Shuli Kang, Jeffrey M Kidd, Miriam K Konkel, Joshua Korn, Ekta Khurana, Deniz Kural, Hugo Y K Lam, Jing Leng, Ruiqiang Li, Yingrui Li, Chang-Yun Lin, Ruibang Luo, Xinmeng Jasmine Mu, James Nemesh, Heather E Peckham, Tobias Rausch,

Aylwyn Scally, Xinghua Shi, Michael P Stromberg, Adrian M Stütz, Alexander Eckehart Urban, Jerilyn A Walker, Jiantao Wu, Yujun Zhang, Zhengdong D Zhang, Mark A Batzer, Li Ding, Gabor T Marth, Gil McVean, Jonathan Sebat, Michael Snyder, Jun Wang, Kenny Ye, Evan E Eichler, Mark B Gerstein, Matthew E Hurles, Charles Lee, Steven A McCarroll, Jan O Korbel, and 1000 Genomes Project. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, Feb 2011.

[107] Siver A Moestue, Eldrid Borgan, Else M Huuse, Evita M Lindholm, Beathe Sitter, Anne-Lise Børresen-Dale, Olav Engebraaten, Gunhild M Maelandsmo, and Ingrid S Gribbestad. Distinct choline metabolic profiles are associated with differences in gene expression for basal-like and luminal-like breast cancer xenograft models. *BMC Cancer*, 10:433, 2010.

[108] Charles G Mullighan, Letha A Phillips, Xiaoping Su, Jing Ma, Christopher B Miller, Sheila A Shurtleff, and James R Downing. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science*, 322(5906):1377–80, Nov 2008.

[109] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W Richard McCombie, James Hicks, and Michael Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–4, Apr 2011.

[110] Nicholas Navin, Alexander Krasnitz, Linda Rodgers, Kerry Cook, Jennifer Meth, Jude Kendall, Michael Riggs, Yvonne Eberling, Jennifer Troge, Vladimir Grubor, Dan Levy, Pär Lundin, Susanne Månér, Anders Zetterberg, James Hicks, and Michael Wigler. Inferring tumor progression from genomic heterogeneity. *Genome Res*, 20(1):68–80, Jan 2010.

[111] Nicholas E Navin. Cancer genomics: one cell at a time. *Genome Biol*, 15(8):452, 2014.

[112] Daniel E Newburger, Dorna Kashef-Haghighi, Ziming Weng, Raheleh Salari, Robert T Sweeney, Alayne L Brunner, Shirley X Zhu, Xiangqian Guo, Sushama Varma, Megan L Troxell, Robert B West, Serafim Batzoglou, and Arend Sidow. Genome evolution during progression to breast cancer. *Genome Res*, 23(7):1097–108, Jul 2013.

[113] Charlotte K Y Ng, Susanna L Cooke, Kevin Howe, Scott Newman, Jian Xian, Jillian Temple, Elizabeth M Batty, Jessica C M Pole, Simon P Langdon, Paul A W Edwards, and James D

Brenton. The role of tandem duplicator phenotype in tumour evolution in high-grade serous ovarian cancer. *J Pathol*, 226(5):703–12, Apr 2012.

[114] Serena Nik-Zainal, Peter Van Loo, David C Wedge, Ludmil B Alexandrov, Christopher D Greenman, King Wai Lau, Keiran Raine, David Jones, John Marshall, Manasa Ramakrishna, Adam Shlien, Susanna L Cooke, Jonathan Hinton, Andrew Menzies, Lucy A Stebbings, Catherine Leroy, Mingming Jia, Richard Rance, Laura J Mudie, Stephen J Gamble, Philip J Stephens, Stuart McLaren, Patrick S Tarpey, Elli Papaemmanuil, Helen R Davies, Ignacio Varela, David J McBride, Graham R Bignell, Kenric Leung, Adam P Butler, Jon W Teague, Sancha Martin, Goran Jönsson, Odette Mariani, Sandrine Boyault, Penelope Miron, Aquila Fatima, Anita Langerød, Samuel A J R Aparicio, Andrew Tutt, Anieta M Sieuwerts, Åke Borg, Gilles Thomas, Anne Vincent Salomon, Andrea L Richardson, Anne-Lise Børresen-Dale, P Andrew Futreal, Michael R Stratton, Peter J Campbell, and Breast Cancer Working Group of the International Cancer Genome Consortium. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, May 2012.

[115] Silje H Nordgard, Fredrik E Johansen, Grethe I G Alnaes, Elmar Bucher, Ann-Christine Syvänen, Bjørn Naume, Anne-Lise Børresen-Dale, and Vessela N Kristensen. Genome-wide analysis identifies 16q deletion associated with survival, molecular subtypes, mrna expression, and germline haplotypes in breast cancer patients. *Genes Chromosomes Cancer*, 47(8):680–96, Aug 2008.

[116] P C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–8, Oct 1976.

[117] Layla Oesper, Ahmad Mahmoody, and Benjamin J. Raphael. Inferring intra-tumor heterogeneity from high-throughput dna sequencing data. In Minghua Deng, Rui Jiang, Fengzhu Sun, and Xuegong Zhang, editors, *RECOMB*, volume 7821 of *Lecture Notes in Computer Science*, pages 171–172. Springer, 2013.

[118] Layla Oesper, Ahmad Mahmoody, and Benjamin J Raphael. Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biol*, 14(7):R80, Jul 2013.

[119] Layla Oesper and Benjamin J Raphael. Workshop: Reconstructing the organization of cancer

genomes. In *Computational Advances in Bio and Medical Sciences (ICCABS), 2013 IEEE 3rd International Conference on*, pages 1–1, June 2013.

[120] Layla Oesper, Anna Ritz, Sarah J Aerni, Ryan Drebin, and Benjamin J Raphael. Reconstructing cancer genomes from paired-end sequencing data. *BMC Bioinformatics*, 13 Suppl 6:S10, 2012.

[121] Layla Oesper, Gryte Satas, and Benjamin J Raphael. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, 30(24):3532–40, Dec 2014.

[122] Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R Speicher, Johannes Zschocke, and Zlatko Trajanoski. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform*, 15(2):256–78, Mar 2014.

[123] Fabio Parisi, Stephan Ariyan, Deepak Narayan, Antonella Bacchiocchi, Kathleen Hoyt, Elaine Cheng, Fang Xu, Peining Li, Ruth Halaban, and Yuval Kluger. Detecting copy number status and uncovering subclonal markers in heterogeneous tumor biopsies. *BMC Genomics*, 12:230, 2011.

[124] Barbara L Parsons. Many different tumor types have polyclonal tumor origin: evidence and implications. *Mutat Res*, 659(3):232–47, 2008.

[125] David Patterson. Computer scientists may have what it takes to help cure cancer. *The New York Times*, 2011.

[126] P A Pevzner and H Tang. Fragment assembly with double-barreled data. *Bioinformatics*, 17 Suppl 1:S225–33, 2001.

[127] P A Pevzner, H Tang, and M S Waterman. An eulerian path approach to dna fragment assembly. *Proc Natl Acad Sci U S A*, 98(17):9748–53, Aug 2001.

[128] Pavel Pevzner and Glenn Tesler. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci U S A*, 100(13):7672–7, Jun 2003.

[129] Pavel A. Pevzner. Dna physical mapping and alternating eulerian cycles in colored graphs. *Algorithmica*, 13(1/2):77–105, 1995.

[130] Pavel A Pevzner, Paul A Pevzner, Haixu Tang, and Glenn Tesler. De novo repeat classification and fragment assembly. *Genome Res*, 14(9):1786–96, Sep 2004.

[131] Victoria Popic, Raheleh Salari, Iman Hajirasouliha, Dorna Kashef Haghighi, Robert B. West, and Serafim Batzoglou. Fast and scalable inference of multi-sample cancer lineages. *CoRR*, abs/1412.8574, 2014.

[132] Yi Qiao, Aaron R Quinlan, Amir A Jazaeri, Roeland Gw Verhaak, David A Wheeler, and Gabor T Marth. Subcloneseeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biol*, 15(8):443, 2014.

[133] Aaron R Quinlan, Royden A Clark, Svetlana Sokolova, Mitchell L Leibowitz, Yujun Zhang, Matthew E Hurles, Joshua C Mell, and Ira M Hall. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res*, 20(5):623–35, May 2010.

[134] Gerald Quon and Quaid Morris. Isolate: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics*, 25(21):2882–9, Nov 2009.

[135] Benjamin J Raphael and Pavel A Pevzner. Reconstructing tumor amplisomes. *Bioinformatics*, 20 Suppl 1:i265–73, Aug 2004.

[136] Benjamin J Raphael, Stanislav Volik, Colin Collins, and Pavel A Pevzner. Reconstructing tumor genome architectures. *Bioinformatics*, 19 Suppl 2:ii162–71, Oct 2003.

[137] Benjamin J Raphael, Stanislav Volik, Peng Yu, Chunxiao Wu, Guiqing Huang, Elena V Linardopoulou, Barbara J Trask, Frederic Waldman, Joseph Costello, Kenneth J Pienta, Gordon B Mills, Krystyna Bajsarowicz, Yasuko Kobayashi, Shivaranjani Sridharan, Pamela L Paris, Quanzhou Tao, Sarah J Aerni, Raymond P Brown, Ali Bashir, Joe W Gray, Jan-Fang Cheng, Pieter de Jong, Mikhail Nefedov, Thomas Ried, Hesed M Padilla-Nash, and Colin C Collins. A sequence-based survey of the complex structural organization of tumor genomes. *Genome Biol*, 9(3):R59, 2008.

[138] Tobias Rausch, David T W Jones, Marc Zapatka, Adrian M Stütz, Thomas Zichner, Joachim Weischenfeldt, Natalie Jäger, Marc Remke, David Shih, Paul A Northcott, Elke Pfaff, Jelena Tica, Qi Wang, Luca Massimi, Hendrik Witt, Sebastian Bender, Sabrina Pleier, Huriye Cin, Cynthia Hawkins, Christian Beck, Andreas von Deimling, Volkmar Hans, Benedikt Brors,

Roland Eils, Wolfram Scheurlen, Jonathon Blake, Vladimir Benes, Andreas E Kulozik, Olaf Witt, Dianna Martin, Cindy Zhang, Rinnat Porat, Diana M Merino, Jonathan Wasserman, Nada Jabado, Adam Fontebasso, Lars Bullinger, Frank G Rücker, Konstanze Döhner, Hartmut Döhner, Jan Koster, Jan J Molenaar, Rogier Versteeg, Marcel Kool, Uri Tabori, David Malkin, Andrey Korshunov, Michael D Taylor, Peter Lichter, Stefan M Pfister, and Jan O Korbel. Genome sequencing of pediatric medulloblastoma links catastrophic dna rearrangements with tp53 mutations. *Cell*, 148(1-2):59–71, Jan 2012.

[139] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, Sep 2012.

[140] Anna Ritz, Ali Bashir, Suzanne Sindi, David Hsu, Iman Hajirasouliha, and Benjamin J Raphael. Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics*, 30(24):3458–66, Dec 2014.

[141] Nicola D Roberts, R Daniel Kortschak, Wendy T Parker, Andreas W Schreiber, Susan Branford, Hamish S Scott, Garique Glonek, and David L Adelson. A comparative analysis of algorithms for somatic snv detection in cancer. *Bioinformatics*, 29(18):2223–30, Sep 2013.

[142] Andrew Roth, Jiarui Ding, Ryan Morin, Anamaria Crisan, Gavin Ha, Ryan Giuliany, Ali Bashashati, Martin Hirst, Gulisa Turashvili, Arusha Oloumi, Marco A Marra, Samuel Aparicio, and Sohrab P Shah. Jointsnvmix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, 28(7):907–13, Apr 2012.

[143] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. Pyclone: statistical inference of clonal population structure in cancer. *Nat Methods*, 11(4):396–8, Apr 2014.

[144] David Sankoff and Phil Trinh. Chromosomal breakpoint reuse in genome sequence rearrangement. *J Comput Biol*, 12(6):812–21, 2005.

[145] Jarupon Fah Sathirapongsasuti, Hane Lee, Basil A J Horst, Georg Brunner, Alistair J Cochran, Scott Binder, John Quackenbush, and Stanley F Nelson. Exome sequencing-based

copy-number variation and loss of heterozygosity detection: Exomecnv. *Bioinformatics*, 27(19):2648–54, Oct 2011.

[146] Christopher T Saunders, Wendy S W Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–7, Jul 2012.

[147] Anna Schuh, Jennifer Becq, Sean Humphray, Adrian Alexa, Adam Burns, Ruth Clifford, Stephan M Feller, Russell Grocock, Shirley Henderson, Irina Khrebtukova, Zoya Kingsbury, Shujun Luo, David McBride, Lisa Murray, Toshi Menju, Adele Timbs, Mark Ross, Jenny Taylor, and David Bentley. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*, 120(20):4191–6, Nov 2012.

[148] Sohrab P Shah, Andrew Roth, Rodrigo Goya, Arusha Oloumi, Gavin Ha, Yongjun Zhao, Gulisa Turashvili, Jiarui Ding, Kane Tse, Gholamreza Haffari, Ali Bashashati, Leah M Prentice, Jaswinder Khattra, Angela Burleigh, Damian Yap, Virginie Bernard, Andrew McPherson, Karey Shumansky, Anamaria Crisan, Ryan Giuliany, Alireza Heravi-Moussavi, Jamie Rosner, Daniel Lai, Inanc Birol, Richard Varhol, Angela Tam, Noreen Dhalla, Thomas Zeng, Kevin Ma, Simon K Chan, Malachi Griffith, Annie Moradian, S-W Grace Cheng, Gregg B Morin, Peter Watson, Karen Gelmon, Stephen Chia, Suet-Feung Chin, Christina Curtis, Oscar M Rueda, Paul D Pharoah, Sambasivarao Damaraju, John Mackey, Kelly Hoon, Timothy Harkins, Vasisht Tadigotla, Mahvash Sigaroudinia, Philippe Gascard, Thea Tlsty, Joseph F Costello, Irmtraud M Meyer, Connie J Eaves, Wyeth W Wasserman, Steven Jones, David Huntsman, Martin Hirst, Carlos Caldas, Marco A Marra, and Samuel Aparicio. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403):395–9, Jun 2012.

[149] Darryl Shibata. Cancer. heterogeneity and tumor history. *Science*, 336(6079):304–5, Apr 2012.

[150] Suzanne Sindi, Elena Helman, Ali Bashir, and Benjamin J Raphael. A geometric approach for classification and comparison of structural variants. *Bioinformatics*, 25(12):i222–30, Jun 2009.

[151] Suzanne S Sindi, Selim Onal, Luke C Peng, Hsin-Ta Wu, and Benjamin J Raphael. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol*, 13(3):R22, 2012.

[152] Carlos Oscar Sánchez Sorzano, Alberto Pascual-Montano, Ainhoa Sánchez de Diego, Carlos Martínez-A, and Karel H M van Wely. Chromothripsis: breakage-fusion-bridge over and over again. *Cell Cycle*, 12(13):2016–23, Jul 2013.

[153] Angela A Steinhardt, Mariana F Gayyed, Alison P Klein, Jixin Dong, Anirban Maitra, Duojia Pan, Elizabeth A Montgomery, and Robert A Anders. Expression of yes-associated protein in common solid tumors. *Hum Pathol*, 39(11):1582–9, Nov 2008.

[154] Philip J Stephens, Chris D Greenman, Beiyuan Fu, Fengtang Yang, Graham R Bignell, Laura J Mudie, Erin D Pleasance, King Wai Lau, David Beare, Lucy A Stebbings, Stuart McLaren, Meng-Lay Lin, David J McBride, Ignacio Varela, Serena Nik-Zainal, Catherine Leroy, Mingming Jia, Andrew Menzies, Adam P Butler, Jon W Teague, Michael A Quail, John Burton, Harold Swerdlow, Nigel P Carter, Laura A Morsberger, Christine Iacobuzio-Donahue, George A Follows, Anthony R Green, Adrienne M Flanagan, Michael R Stratton, P Andrew Futreal, and Peter J Campbell. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40, Jan 2011.

[155] Francesco Strino, Fabio Parisi, Mariann Micsinai, and Yuval Kluger. Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res*, 41(17):e165, Sep 2013.

[156] Dominik Sturm, Sebastian Bender, David T W Jones, Peter Lichter, Jacques Grill, Oren Becher, Cynthia Hawkins, Jacek Majewski, Chris Jones, Joseph F Costello, Antonio Iavarone, Kenneth Aldape, Cameron W Brennan, Nada Jabado, and Stefan M Pfister. Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. *Nat Rev Cancer*, 14(2):92–107, Feb 2014.

[157] Xiaoping Su, Li Zhang, Jianping Zhang, Funda Meric-Bernstam, and John N Weinstein. Purityest: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics*, 28(17):2265–6, Sep 2012.

[158] S Takakura, T Kohno, R Manda, A Okamoto, T Tanaka, and J Yokota. Genetic alterations

and expression of the protein phosphatase 1 genes in human cancers. *Int J Oncol*, 18(4):817–24, Apr 2001.

[159] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Brief Bioinform*, 14(2):178–92, Mar 2013.

[160] Sarah A Tishkoff and Kenneth K Kidd. Implications of biogeography of human populations for 'race' and medicine. *Nat Genet*, 36(11 Suppl):S21–7, Nov 2004.

[161] David Tolliver, Charalampos Tsourakakis, Ayshwarya Subramanian, Stanley Shackney, and Russell Schwartz. Robust unmixing of tumor states in array comparative genomic hybridization data. *Bioinformatics*, 26(12):i106–14, Jun 2010.

[162] Toshiyuki Tsunoda, Takeharu Ota, Takahiro Fujimoto, Keiko Doi, Yoko Tanaka, Yasuhiro Yoshida, Masahiro Ogawa, Hiroshi Matsuzaki, Masato Hamabashiri, Darren R Tyson, Masahide Kuroki, Shingo Miyamoto, and Senji Shirasawa. Inhibition of phosphodiesterase-4 (pde4) activity triggers luminal apoptosis and akt dephosphorylation in a 3-d colonic-crypt model. *Mol Cancer*, 11:46, 2012.

[163] Eray Tuzun, Andrew J Sharp, Jeffrey A Bailey, Rajinder Kaul, V Anne Morrison, Lisa M Pertz, Eric Haugen, Hillary Hayden, Donna Albertson, Daniel Pinkel, Maynard V Olson, and Evan E Eichler. Fine-scale structural variation of the human genome. *Nat Genet*, 37(7):727–32, Jul 2005.

[164] Peter Van Loo, Silje H Nordgard, Ole Christian Lingjærde, Hege G Russnes, Inga H Rye, Wei Sun, Victor J Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, Charles M Perou, Anne-Lise Børresen-Dale, and Vessela N Kristensen. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A*, 107(39):16910–5, Sep 2010.

[165] Bert Vogelstein and Kenneth W Kinzler. Cancer genes and the pathways they control. *Nat Med*, 10(8):789–99, Aug 2004.

[166] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, Jr, and Kenneth W Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–58, Mar 2013.

[167] Stanislav Volik, Shaying Zhao, Koei Chin, John H Brebner, David R Herndon, Quanzhou Tao, David Kowbel, Guiqing Huang, Anna Lapuk, Wen-Lin Kuo, Gregg Magrane, Pieter De Jong, Joe W Gray, and Colin Collins. End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc Natl Acad Sci U S A*, 100(13):7696–701, Jun 2003.

[168] C B Vos, N T ter Haar, C Rosenberg, J L Peterse, A M Cleton-Jansen, C J Cornelisse, and M J van de Vijver. Genetic alterations on chromosome 16 and 17 are important features of ductal carcinoma in situ of the breast and are associated with histologic type. *Br J Cancer*, 81(8):1410–8, Dec 1999.

[169] Yong Wang, Jill Waters, Marco L Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam, and Nicholas E Navin. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, Jul 2014.

[170] Caleb Weinreb, Layla Oesper, and Benjamin J Raphael. Open adjacencies and k-breaks: detecting simultaneous rearrangements in cancer genomes. *BMC Genomics*, 15 Suppl 6:S4, 2014.

[171] Roland Wittler, Ján Maňuch, Murray Patterson, and Jens Stoye. Consistency of sequence-based gene clusters. *J Comput Biol*, 18(9):1023–39, Sep 2011.

[172] L.A. Wolsey. *Integer programming*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 1998.

[173] John C Wooley, Adam Godzik, and Iddo Friedberg. A primer on metagenomics. *PLoS Comput Biol*, 6(2):e1000667, Feb 2010.

[174] Ruibin Xi, Angela G Hadjipanayis, Lovelace J Luquette, Tae-Min Kim, Eunjung Lee, Jianhua Zhang, Mark D Johnson, Donna M Muzny, David A Wheeler, Richard A Gibbs, Raju Kucherlapati, and Peter J Park. Copy number variation detection in whole-genome sequencing data using the bayesian information criterion. *Proc Natl Acad Sci U S A*, 108(46):E1128–36, Nov 2011.

[175] Ruibin Xi, Tae-Min Kim, and Peter J Park. Detecting structural variations in the human genome using next generation sequencing. *Brief Funct Genomics*, 9(5-6):405–15, Dec 2010.

[176] Xun Xu, Yong Hou, Xuyang Yin, Li Bao, Aifa Tang, Luting Song, Fuqiang Li, Shirley Tsang, Kui Wu, Hanjie Wu, Weiming He, Liang Zeng, Manjie Xing, Renhua Wu, Hui Jiang, Xiao Liu, Dandan Cao, Guangwu Guo, Xueda Hu, Yaoting Gui, Zesong Li, Wenyue Xie, Xiaojuan Sun, Min Shi, Zhiming Cai, Bin Wang, Meiming Zhong, Jingxiang Li, Zuhong Lu, Ning Gu, Xiuqing Zhang, Laurie Goodman, Lars Bolund, Jian Wang, Huanming Yang, Karsten Kristiansen, Michael Dean, Yingrui Li, and Jun Wang. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, 148(5):886–95, Mar 2012.

[177] Vinod Kumar Yadav and Subhajyoti De. An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Brief Bioinform*, Feb 2014.

[178] Sophia Yancopoulos, Oliver Attie, and Richard Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, 2005.

[179] M L Yaremko, W M Recant, and C A Westbrook. Loss of heterozygosity from the short arm of chromosome 8 is an early event in breast cancers. *Genes Chromosomes Cancer*, 13(3):186–91, Jul 1995.

[180] Christopher Yau, Dmitri Mouradov, Robert N Jorissen, Stefano Colella, Ghazala Mirza, Graham Steers, Adrian Harris, Jiannis Ragoussis, Oliver Sieber, and Christopher C Holmes. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol*, 11(9):R92, 2010.

[181] Seungtai Yoon, Zhenyu Xuan, Vladimir Makarov, Kenny Ye, and Jonathan Sebat. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res*, 19(9):1586–92, Sep 2009.

[182] Yinyin Yuan, Henrik Failmezger, Oscar M Rueda, H Raza Ali, Stefan Gräf, Suet-Feung Chin, Roland F Schwarz, Christina Curtis, Mark J Dunning, Helen Bardwell, Nicola Johnson, Sarah Doyle, Gulisa Turashvili, Elena Provenzano, Sam Aparicio, Carlos Caldas, and Florian Markowetz. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci Transl Med*, 4(157):157ra143, Oct 2012.

[183] Shay Zakov, Marcus Kinsella, and Vineet Bafna. An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. *Proc Natl Acad Sci U S A*, 110(14):5546–51, Apr 2013.

[184] Habil Zare, Junfeng Wang, Alex Hu, Kris Weber, Josh Smith, Debbie Nickerson, ChaoZhong Song, Daniela Witten, C Anthony Blau, and William Stafford Noble. Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput Biol*, 10(7):e1003703, Jul 2014.

[185] Jianjun Zhang, Junya Fujimoto, Jianhua Zhang, David C Wedge, Xingzhi Song, Jiexin Zhang, Sahil Seth, Chi-Wan Chow, Yu Cao, Curtis Gumbs, Kathryn A Gold, Neda Kalhor, Latasha Little, Harshad Mahadeshwar, Cesar Moran, Alexei Protopopov, Huandong Sun, Jiabin Tang, Xifeng Wu, Yuanqing Ye, William N William, J Jack Lee, John V Heymach, Waun Ki Hong, Stephen Swisher, Ignacio I Wistuba, and P Andrew Futreal. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science*, 346(6206):256–9, Oct 2014.