UNIVERSITY OF CALIFORNIA
SANTA CRUZ

ANALYSIS OF GENOMIC REARRANGEMENTS IN CANCER FROM HIGH THROUGHPUT SEQUENCING DATA.

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOINFORMATICS

by

Tracy J. Ballinger

September 2015

The Dissertation of Tracy J. Ballinger
is approved:

Professor David Haussler, Chair

Professor Kevin Karplus

Professor Lindsay Hinck

Professor Ting Wang

Dean Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Analysis of Genomic Rearrangements in Cancer from High Throughput Sequencing Data.

by

Tracy J. Ballinger

In the last century cancer has become increasingly prevalent and is the second largest killer in the United States, estimated to afflict 1 in 4 people during their life. Despite our long history with cancer and our herculean efforts to thwart the disease, in many cases we still do not understand the underlying causes or have successful treatments. In my graduate work, I've developed two approaches to the study of cancer genomics and applied them to the whole genome sequencing data of cancer patients from The Cancer Genome Atlas (TCGA). In collaboration with Dr. Ewing, I built a pipeline to detect retrotransposon insertions from paired-end high-throughput sequencing data and found somatic retrotransposon insertions in a fifth of cancer patients.

My second novel contribution to the study of cancer genomics is the development of the CN-AVG pipeline, a method for reconstructing the evolutionary history of a single tumor by predicing the order of structural mutations such as deletions, duplications, and inversion. The CN-AVG theory was developed by Drs. Haussler, Zerbino, and Paten and samples potential evolutionary histories for a tumor using Markov Chain Monte Carlo sampling. I contributed to the development of this method by testing its accuracy and limitations on simulated evolutionary histories. I found that the ability to reconstruct a history decays exponentially with increased breakpoint reuse, but that we can estimate how accuratly we reconstruct a mutation event using the likelihood scores of the events. I further designed novel techniques for the application of CN-AVG to whole genome sequencing data from actual patients and applied these techniques to search for evolutionary patterns in glioblastoma multiforme using sequencing data from TCGA. My results show patterns of two-hit deletions, as we would expect, and amplifications occurring over several mutational events. I also find that the CN-AVG method frequently makes use of whole chromosome copy number changes following by localized deletions, a bias that could be mitigated through modifying the cost function for an evolutionary history.

To my parents,

Allen and Marcel Ballinger,

who have given up telling me what to do and just supported me.

Also to Elinor Velasquez,

who is one of the smartest, craziest, and most courageous people that I know.

Acknowledgments

I want to thank my committee for great ideas and encouragement. I would also particularly like to thank Adam Ewing for his expertise and help on the retrotransposon project, and Benedict Paten and Daniel Zerbino for supervising my work on the CN-AVG project.

Chapter 1

Introduction

Cancer is a complex, often fatal disease, estimated to afflict one in four people in the United States in their lifetimes. More background about the biology of cancer can be found in Section 2.1, but suffice to say, it is notoriously difficult to cure for two reasons. One, most cancers are caused by a unique mutational combination, so essentially each cancer is a unique disease requiring a unique cure, and two, it is a moving target even within a single patient. Through studying how cancer develops and progresses, rather than simply how it exists at the time of diagnosis or operation, researchers may develop better preventive strategies and, by anticipating the cancer's next evolutionary step, they may develop better treatments to block its progression. In this thesis, I attempt to elucidate the evolution of cancer at the level of individual patients by predicting the order of structural rearrangements in a tumor genome. As I will describe in greater detail in Chapter 2, this pipeline is unique in that it can account for multiclonality within a tumor, focuses on the ordering of structure rearrangement, and is able to predict parsimonious evolutionary histories for single patients.

Drs. Haussler, Zerbino, and Paten have developed a mathematical framework, the Copy Number Ancestral Variation Graph (CN-AVG) to predict the order of structural rearrangements using copy number changes and breaks or links between distant loci in the genome indicative of structural rearrangements [173]. In Chapters 3 and 4 I will describe my work of testing the method using simulated data and novel methods for applying the framework to whole genome sequencing data from TCGA. This new CN-AVG pipeline has the advantage of being able to order large-scale mutational events for a single patient and, therefore, is capable of biological discovery even without large sample sizes. Additionally, it may tell us the chronological order of secondary and later mutations, illuminating how the cancer continued to develop, perhaps giving insight into metastasis and drug resistance processes. I anticipate that this type of analysis, the reconstruction of the evolutionary history of a tumor, will be an important asset in future cancer genomics studies. It could give us greater insight to cancer biology by revealing patterns in how cancers develope. All patients will benefit if their initiating and driving mutations could be determined regardless of how common those mutations are in other patients.

The CN-AVG method, as novel and powerful as it is, had not been applied to actual data nor tested to determine how well it could reconstruct evolutionary histories prior to my work. My aim has been to make this exciting new method accessible and the results interpretable to other researchers, which I will describe in Chapter 4. Furthermore, for others to believe that this method works, it needed to be tested, so this was another aim of my thesis work. Although, ideally, I would test the method against actual evolving tumor genomes, this type of experimental data is currently unavailable with whole genome sequencing. Therefore, I tested the method against evolutionary histories generated *in silico*. In testing, it was necessary to create ways for combining CN-AVG results and compare these to the known history. I describe my methods for doing so in Chapter 3.

Several large projects are undertaking to sequence the DNA of many patients' cancers with the goal of finding mutation patterns associated with cancers. The Cancer Genome Atlas Project, funded by the National Cancer Institute, has sequenced thousands of patients across twenty cancer types since 2006 [27]. The International Cancer Genome Consortium is an international effort studying fifty cancer types [77]. The Wellcome Trust Sanger Institute has its Cancer Genome Project [36], and even the entertainment industry raised funds for studying cancer genomes with the Stand Up to Cancer project [19, 99]. Most cancer genetics studies examine only single nucleotide variants (SNVs) or copy number alterations (CNAs) because these are: 1) the easiest to detect from high throughput sequencing studies, 2) the most frequent, and 3) thought to have the largest cumulative phenotypic effect. In genomic studies, researchers may find genes linked to cancer based on the frequency that a gene is mutated across patients [158, 88].

A patient may appear to lack common mutations because gene function is being altered by an undetected type of mutation, such as inversions, fusions or small indels. For my third aim, I attempt to fill part of this detection gap by developing a pipeline to find retrotransposon insertions from paired-end high throughput sequencing data. As I will discuss in Chapter 5, retrotransposons are particularly interesting because the biological mechanism behind them is well understood and linked to a specific pathway. Activity in this normally dormant pathway is evidenced by somatic retrotransposition. We can generate more accurate and statistically powerful results from population studies using better and more complete mutation detection algorithms. Furthermore, we can better understand the biological mechanisms affecting tumors through studying this unique type of insertional mutagen.

Even with additional mutation-detection algorithms, it is likely that many patients will still lack mutations in known cancer genes, so researchers need other strategies for understanding these unique cases. One is to look for significantly mutated pathways, networks of genes responsible for certain functions such as apoptosis or other cancer hallmarks [92, 28] since a gene pathway or network may be significantly perturbed in a population although any individual gene in the pathway is not mutated at a significant level by itself. This method may explain some cancers lacking mutated oncogenes, but there will likely still be patients lacking mutated oncopathways. Furthermore, these methods do not necessarily elucidate what initially caused the cancer. Fortunately, researchers have taken yet another approach to glean insight to an individual cancer by reconstructing the order that mutations happen from a single tumor sample. Cancer develops gradually over time [162], and although the transition from a pre-cancerous cell to a cancerous one is hard to define, early occurring mutations are more likely to have played a role in initiating cancer than later ones. Further insight to the cause of cancer and therefore how to prevent or treat it may come through knowing the order that mutations happened in addition to knowing what mutations are present.

Chapter 2

Cancer history and evolution

The bulk of my thesis work involved building up the CN-AVG pipeline in order to study the evolution of cancer. In this chapter, I will give a brief history of cancer in order to illustrate the complexity of the disease and why current efforts are focused on studying cancer genomes. Section 2.2 reviews past studies of cancer evolution, Section 2.3 reviews previous theories for reconstructing structural rearrangements in genomes, including an overview of the CN-AVG method.

2.1 A brief history of cancer

The earliest evidence of a human tumor was found on a skeleton in southeastern Africa dating to two million years ago, and ancient Egyption medical writings from 2500 B.C. contain a detailed description of, "a bulging mass of the breast" for which there is no treatment" [109]. Throughout human history, theories on the causes and best treatments for cancer have naturally progressed along with medical technologies and biological knowledge, ranging from a belief that the disease is caused by black bile in the first century (AD 160) [58], to a theory that it is contagious or caused by lympth stagnation[59], to the current belief that genetic mutations are the underlying instigators [162]. Many early theories of cancer were not entirely wrong. Although a contagious human cancer has yet to be found, the tasmanian devil population has recently been plagued by a contagious tumor transmissible through biting [122], and some human cancers have pathogenic origins in the form of oncoviruses and even bacteria. For example, HPV causes cervical cancer [31], and *Helicobacter pylori* can cause stomach cancer through prolonged inflammation [159]. With stomach cancer as a case in point, tumors can also form as a result of chronic inflammation [61] as Hermann Boerhaave and Jean Astruc hypothesized in their lymph stagnation theory [59]. Additionally, some cancer types are hereditary, while others are linked to environmental factors such as x-rays or chemical carcinogens [162]. The classification of cancer as a genetic disease, therefore, does not necessarily discount other theories, but rather explains them. For example, Temin showed that oncoviruses instigate cancer through incorporating their DNA into a cell's genome [153], Knudson hypothesized a genetic link through the study of inherited retinoblastoma [85], and Ames showed that some carcinogens increase mutation rates [1]. Genetic instability and mutation are not one of the hallmarks of cancer, as defined by Hanahan and Weinberg, but they are considered "enabling characteristics" of cancer, along with tumor-promoting inflation, indicating that they are a means through which cells acquire malignant properties [61]. Although tumorigenesis is a complex process driven by diverse mechanisms, it is useful to think of cancer as a genetic disease for several reasons: all tumors have some level of mutations [158]; these mutations may be the root cause of the disease; and we currently have the technology to sequence and study genomes comprehensively.

Frustratingly, although new technologies better our understanding of cancer, they have brought cures to only a subset of disease types (Figure 2.1). For example, there is now a vaccine for HPV, the virus that causes cervical cancer [31] and chronic myelogenous leukemia with the fusion gene BCR-ABL is now treatable with Imatinib [132, 39]. Other types of treatment include surgery, chemotherapy, and radiation, often in combination. Surgery is the oldest remedy and is most successful if the tumor can be safely extracted and has not spread throughout the body [162, 109]. Radiotherapy began in the early twentieth century following the discover of X-rays and consists of blasting tumors with focused electromagnetic radiation [162, 109]. Chemotherapy involves giving patients chemical compounds that block or damage DNA in an effort to selectively kill rapidly growing cells. Unfortunately, healthy cells are killed alongside cancer cells, causing brutal side effects to the patient, so the essential aim of chemotherapy becomes to kill the cancer cells before killing the patient, a difficult target [162, 109]. Surgery and radiation can also have harmful side affects and may instigate future disease through prolonged inflammation as part of the post-surgical healing process, or through off-target mutagenic affects of X-rays. Today, researchers are pursuing immunotherapy, the activation of the native immune system against cancer cells [133, 128], and targeted therapy, which are drugs that inhibit the cancer-specific form of an oncoprotein [162], as ways to specifically kill only cancer cells. Hopefully, these treatments will be less detrimental to patients and more successful overall.

Incurable patients either do not respond to existing treatments, or respond for a short time and then relapse, with the recurring tumor often growing more aggressively than the initial one [162]. One theory of cancer recurrence is that the tumor acquires genetic mutations that confer treatment resistance. In essence, the cancer is changing in response to its environment until it has the right mutational combination to allow it to survive [162]. Hence, today cancer can be understood as an evolutionary disease, a constantly changing disease, difficult to eradicate because of its ability to evolve.

True to its changing form, cancer develops over a long period of time because a cell must acquire multiple *hallmarks* such as immortality, activation of proliferation, and resistance to cell death before it becomes cancerous or, in the final stages, metastatic [61, 60]. Because these hallmarks involve different genes and signaling pathways, multiple genes need to be mutated. Cancer is thought to happen stepwise; a cellular lineage gradually gains random mutations as the cells duplicate and divide. The multiple mutation theory was first described by Nordling in 1952, in which he and others estimate that at least 7 mutations are needed, based on the age incidence of cancer [114]. Decades later, Vogelstein found evidence of mutation accumulation coinciding with cancer progression by sequencing 4 known common mutations in patients suffering from various stages of colon neoplasias [157]. More recently researchers used genomic studies to calculate the number of mutations needed to generate cancer, and their estimate, two to eight, is similar to the estimate of fifty years ago [158]. Mutations accumulate slowly and steadily in a cell until the cell gains a mutation that confers

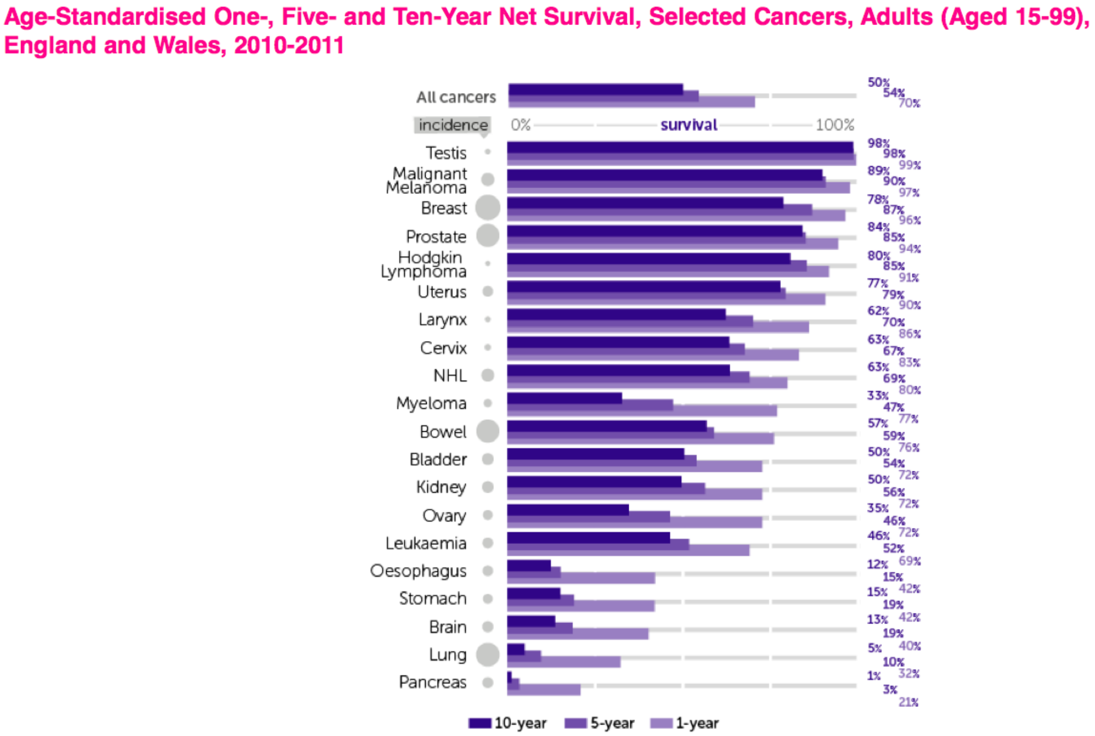


Figure 2.1: Cancers have differing levels of mortality, as measured by various survival times, but few cancers have near 100% survival. Breast is for female only and prostate is for male only. Figure taken from Cancer Research UK, http://www.cancerresearchuk.org/cancer-info/cancerstats/survival/common-cancers/, March 2015.

genetic instability or hypermutation, an increased rate of mutation, at which point the cell can change more quickly, potentially rapidly advancing toward malignancy. Of note, childhood cancers and lymphomas have very few mutations [158] and therefore do not exhibit hypermutation or genomic instability, but these are the few exceptions. Also, different types of cancer exhibit different types of mutation patterns, from C-T transversions frequent in melanomas [89, 162], to microsattellite instability seen in familial colon cancers [124].

It takes years for a cell to obtain a combination of mutations that will allow the cell to grow uncontrollably, and it takes additional time for the cell to spawn enough progeny to become a detectible mass or exhibit symptoms in a patient. At this point of detection, the initiating mutations are often lost amongst many secondary others. The secondary mutations may be passengers, benign and simply carried along with the tumor, affecting neither its growth nor survival [158, 162]. Conversely, these secondary mutations, along with the initiating ones, may be drivers, mutations allowing and encouraging uncontrolled cellular growth. Driver mutations can occur in oncogenes, genes which stimulate cell growth, or in tumor suppressors, genes responsible for monitoring and inhibiting cell growth [162]. Oncogenes such as PIK3CA or EGFR are often amplified or exist in an mutated overactive form, whereas tumor suppressors such as RB1 are often deleted or have disabling mutations [162, 104, 158].

Previous methods and studies designed to study the evolution of cancer will be reviewed in Section 2.2, and a review of methods for reconstructing the evolutionary history of a genome in terms of genome rearrangements, including the method used in this dissertation, appears in Section 2.3.

2.2 Cancer evolution

In the evolutionary process, some mutations are positively selected for because they confer a survival advantage to a cell or organism, others negatively selected, while most mutations have no effect on fitness. The same principles apply to cancer evolution. Peter Nowell first described the theory of cancer evolution in 1976, where, based on cytogentic, X-inactivation, and immunoglobulin studies, he concluded that cancer begins in a single cell [115] (Figure 2.2). The cell gradually acquires genetic mutations allowing it to grow uncontrollably and mutate further, spawning genetically diverse subclones. A select few mutations lead to the increased and uncontrolled growth of the cell, and these are commonly called *driver* mutations. In contrast, most mutations are benign and commonly called *passenger* mutations, because they are merely carried along in the cancer's genome. Nearly all adult solid tumor cancer types have an increased rate of mutation, or *hypermutation* [158, 162], which is advantageous for the cancer as it allows more rapid evolution, but disadvantageous to cancer geneticists, because it generates large numbers of passenger mutations, effectively increasing the noise in the tumor genome. The main goal in cancer genetic research is to distinguish drivers from passengers, because abrogating the effects of driving mutations will end the disease, while targeting passenger mutations is extremely difficult if there are many passengers and will have no curative effect. The increased noise from genetic instability makes it

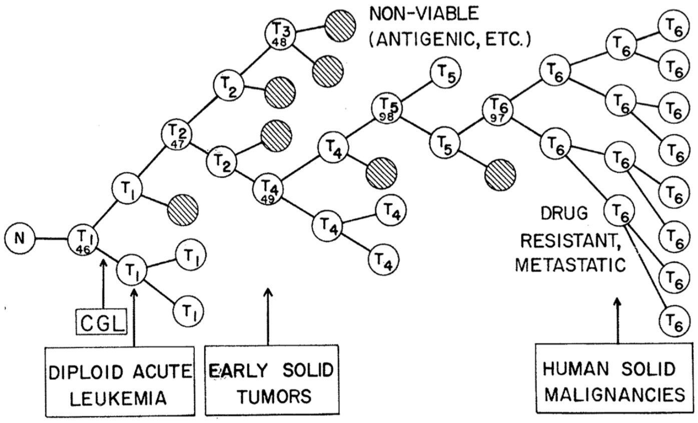


Figure 2.2: Illustration of the theory of cancer evolution from Nowell [115].

more difficult to sift the wheat, drivers, from the chaff, passengers.

Since all tumor genomic data is acquired after a cancer has been diagnosed, cancer evolution cannot be studied from its inception, only from the point at which it has already evolved from a single normal cell to a massive heterogenous population. Several experimental and analytical techniques attempt to look back in time and see the cancer genome near its genesis. Most of these focus on single nucleotide variants (SNVs), ignoring large scale structural rearrangements, and are valuable complementary methods to the CN-AVG method I employ. Methods for determining the rearrangement sequence between two genomes do exist, but have only been applied to cancer genomes in a few cases [3].

2.2.1 Retrospective studies

Researchers have studied cancer evolution retrospectively by looking at the allele frequency of single nucleotide variants (SNVs) within a heterogenous tumor. SNV allele frequencies are estimated by the number of sequencing reads that contain the variant out of all of the sequencing reads covering the variant locus. Mutations with high allele frequencies were probably present in the founding cell or occurred early in the history of the cancer and therefore may be driving mutations. A low allele frequency indicates that the SNV is found in a subset of cells and is likely to have occurred more recently in the history of the cancer. This strategy has been used in several studies to determine the subclonal population structure of tumors [138, 163, 32, 113].

In the last several years, three methods were published for the purpose of rigorously estimating clonality using variant allele frequencies and copy number (CN) estimates generated from read coverage [25, 107, 129], and one method using the CN estimate alone [116]. From these mutation clusters, researchers infer the subclonal structure of a tumor and parts of the evolutionary history. For example, Carter et al. [25] used their ABSOLUTE method to estimate ploidy of cancer cells and thus distinguish whole genome duplication events, or *genome doublings*, from localized copy number changes, or *somatic copy-number alterations* (SCNAs). Through finding similar SCNA profiles with varying ploidy estimates across multiple samples, they concluded that genome doublings frequently occur after SCNAs in many cancer types. Durinck, et al. [41] also timed CN changes relative to each other by looking at the frequency of SNVs within amplified regions. A region of early amplification would acquire more SNVs than a late amplification, assuming SNVs are equally likely across the genome and across time. Furthermore, they timed a SNV relative to a CN change in the same region by looking for loss of heterozygosity (LOH) or amplification of the SNV along with the CN. A SNV that occured before LOH would acquire high allelic frequency with the LOH event, whereas one that occurred after LOH would not.

Researchers use SNVs and SCNAs not only to time copy number rearrangements relative to each other, but also to infer the phylogeny of cells in a tumor and, in turn, the order that these alterations occured [21, 116, 164]. Several studies have used deep sequencing, or next generation sequencing at high coverage, to generate highly accurate SNV allelic counts and construct accurate subclonal estimates [138]. It should be noted that although the clonal structure can be estimated from SNVs or SCNAs as previously described, the clonal architecture of a tumor does not necessarily determine its phylogeny. For example, a subclone constituting 20% of a tumor may be a descendent of an 80% primary clone or it may come from an independent lineage. An accurate phylogenetic tree could be determined given linkage information between SNVs, or which variants are on the same DNA strands, but high throughput sequencing reads are too short to provide this information.

Single-cell sequencing can provide the linkage information needed to disambiguate the phylogeny, and researchers have employed several methods to determine the DNA sequence of single tumor cells. Campbell et al. [21] performed deep long read (250bp) sequencing of the IGH locus in chronic lymphocytic leukemia patients, while

Tsao et al. [155] performed single-cell sequencing of the colorectal tumor microsatellites. Two other groups [112, 75] applied high-throughput sequencing technology to single cells of a breast cancer patient and a patient with essential thrombocythemia, a myeloproliferative disorder. Others obtain single cell sequences through clonal expansion of a single cell in culture or in xenograft, injecting the cell into a host animal to allow it to expand [38]. The additional linkage information shows diverse evolutionary histories of tumors, with some cancers gradually acquiring mutations and others rapidly mutating over short time spans [155]. Furthermore, tumors have diverse and complex phylogenies, often composed of multiple clonal expansions [112, 75, 163] and complex family trees [21, 112]. Phylogenetic studies of tumors are important since the cellular diversity of a tumor may determine its adaptability and be a good predictor of patient survival.

2.2.2 Multiple time point studies

Researchers also study cancer evolution in real time by sampling a tumor at multiple time points [139, 37, 32] or locations [168, 38, 22] from a single patient. Although these studies cannot determine how the cancer began, they may reveal how the cancer resists treatment or becomes metastatic. Is it acquiring novel mutations that are driving drug resistance; does the treatment simply fail to eradicate the primary tumor in the first place; or is the metastatic or recurring tumor formed from a subclone distinct from the primary clone of the tumor? Studies comparing mutations from a single patient's primary and metastic tumors find that both the metastic and recurrent tumors contain mutations not found in the primary tumor or found at low frequency in the primary tumor [139, 38], indicating that metastasis are, indeed, offshoots of the clonally diverse primary tumor. Studies of patients pre- and post-treatment find that the recurrent tumor closely resembles the primary tumor, indicating that the treatment did not eradicate the original source of the tumor. In cases where the recurrent tumor has acquired new mutations it is inconclusive whether these new mutations are responsible for the relapse or are merely a side effect of hypermutability or chemotherapy, since unfortunately, chemotherapies are mutagens themselves [37, 32]. These and other future studies tracing genomic mutations in a tumor over time will yield valuable insights into how a primary tumor evolves on a course to metastasis or in response to treatment. However, this type of study requires more resources per patient and currently, large cancer genomics projects sample a single timepoint. Methods that can retrospectively study cancer evolution, such as those described in the previous section and the CN-AVG method I use in this study, will remain valuable tools.

2.3 Evolutionary reconstruction with structural variants

Recent studies have focused on ordering structural variation in cancer, by, as previously described, ordering SCNAs relative to each other and to SNVs [41, 125]. Still others have developed algorithms to link rearrangements in cancer evolution, which I will describe in Section 2.3.1. Additionally, bioinformaticians have developed mathematical theories to solve the problem of transforming one genome into another via structural rearangements in the interest of modeling genome evolution. I will explain how the CN-AVG method works and briefly review some of the theory behind it in Section 2.3.3.

2.3.1 Reconstructing structural rearrangements in cancer

In addition to the methods described in Section 2.2.1, several recent cancer genomics studies have provided understanding of cancer evolution by examining complex genomic structures common to cancer such as *double minutes*, which are regions of massive amplification of short discontinuous segments of DNA [54, 130, 145, 12, 127], and *neochromosomes*, which are large chromosome-like structures also known as *supernumerary chromsosomes* [95]. Sometimes these studies use special sequencing techniques [54, 12] or FISH experiments to verify or supplement their reconstruction of the cancer genome [57, 130, 145]. They also use high-throughput sequencing data, identifying breakpoints through changes in read coverage, which indicate a change in copy number, and through discordant read pairs. I will refer to the connections between distant regions of the reference genome as evidenced through discordant read pairs as *adjacencies*, and will describe these studies in more detail in the following paragraphs.

Chromothripsis was first defined by Stephens et al. as an event in which “tens to hundreds of genomic rearrangements occur in a one-off cellular crisis” [145] . In other words, it is a shattering of a chromosomal region following by randomly piecing the shattered segments back together with the loss of some genomic content and reordering of the surviving pieces. The hallmarks of chromothripsis are a high density of adjacencies in a genomic region, a copy number state that alternates between only two or three values, and a corresponding alternating state between heterozygous and homozygous regions. These characteristics are unlikely to occur through series of single double-strand break and amplification events and are better explained as a single catastrophic event that shatters a genomic region. Chromothripsis was originally recognized in chronic lympocytic leukaemia, but since then it has been detected in many other cancer types [50, 127, 12]. The discovery of chromothripsis does not tell the full story of how cancer evolves, but it does show that cancer cells may evolve in small leaps, punctuated evolution, rather than gradual steps, and illustrates the immense value in characterizing structural rearrangements indepth.

Another algorithm, ChainFinder, links breakpoints in a tumor genome based on the *adjacency probability* of the breakpoints [3]. The *adjacency probability* is the probability of two breakpoints being in close proximity, given the genome wide breakpoint densities determined from the cohort and the number of breaks for the genome under study. After linking breakpoints together, Baca, et al. search for chains of connected breakpoints and use the resulting long chains as evidence for *chromoplexy* events, or complex rearrangements involving simultaneous and interdependent deletions, inversions, or other rearrangements. Chromoplexy typically includes fewer rearrangements between more dispersed regions of the genome and more separate chromosomes than chromothripsis, which inhibits more breakpoints in a focal region involving only one or two chromosomes. Along with chromothripsis, it supports the idea that complex structural rearrangements seen in cancers happen not gradually, but in spurts, or as "puctuated evolution".

Garsed et al. isolate and sequence neochromosomes from liposarcoma cell lines in order to uncover the evolutionary processes behind their formation [54]. They test whether the copy number and adjacency patterns match a history of breakage fusion bridge cycles or chromothripsis events by comparing the observed profiles to those generated from simulations of various evolutionary scenarios. Breakage fusion bridge (BFB) cycles are thought to play a role in cancer as they can lead to highly amplified genomic regions [12, 55, 96], and they begin with the fusion of two chromosome ends due to eroded telomeres. The fused chromosomes are repeatedly broken and refused together every round of mitosis until they somehow aquire new telomeres [101], and the end result is a series of inverted segmental duplications and a step-like copy number profile. Garsed, et. al determine that the most likely evolutionary path of a neochromosome is the initial formation of a ring chromosome through a chromothripsis event, then amplification of the circlular chromosome, a combination of BFB and chromothripsis events, and finally telomere capture and linearlization of the circular chromosome.

The studies just described uncover some of the underlying processes taking part in cancer evolution, but they do not attempt to predict the exact sequence of structural rearrangements, both large and small. Greenman et al. do make such an attempt through the construction of *allelic graphs* and *somatic graphs* [57]. They use SNVs, copy number, and adjacencies to represent phased mutated genomes which can be converted to these specialized graphs, the components of which represent a single type of structural mutation. To order these events, they search for a parsimonious series of structural rearrangements that generate the known phenotype. While a powerful method, it relies on the assumption of single breakpoint use and uses a limited set of rearrangements which, understandably, does not include chromothripsis.

None of the methods I've just described offer a robust and scalable solution to the reconstruction of both the evolutionary history of a tumor and novel genomic structures such as double minutes or neochromosomes. Some studies rely on a predetermined genome rearrangement [54, 57] from which they make evolutionary claims, while other methods focus on determining the possible novel genome structures from sequencing data, a challenge in and of itself due to noisy sequencing data, ambiguities that arise from the polyploid nature of cancer genomes, and the multiclonality of tumors [3, 130]. For example, Geenman et al. determine a series of specific structural rearrangement operations that can best explain a given copy number and SNP profile, but they apply this method only to small rearrangement clusters for which they can reliably build the allelic graphs and, as previously stated, do not model chromothripsis, an important process in cancer evolution.

The methods have various levels of scalability as well. For example, the sequence of a double minute, a small circular DNA structure typically containing oncogenes and present in many copies in a tumor cell [6, 134, 81], was painstakingly done by hand for some studies [127, 130], and the reconstruction of neochromosomes required isolating the neochromosome before sequencing to prevent other chromosomes from clouding the sequencing data.

The ChainFinder algorithm has been applied to larger cohorts, and in fact, relies on multiple samples for the statistical power to link breakpoints together into chains, but this yields little insight into how the chromoplexy events occur since they are not explainable by BFB cycles or chromothripsis events. Despite the diversity of these methods, they all point to complex structural rearrangement events as playing key roles in cancer evolution, and they often predict these events to happen simultaneously or in close succession.

The CN-AVG method is more general than previous methods because it allows for multiclonality within the DNA sequencing sample and is not limited to a predetermined set of rearrangements. It allows any type of complex rearrangement to play a part in the evolutionary history, from simple segmental duplications, inversions, or deletions, to the simulataneous deletion, amplification, and rearrangement as a single rearrangement event, as in chromothripsis. Furthermore, I will demonstrate the application of this method to widely available WGS data and offer an automated interpretation of its results so that the method can easily be applied at an industrial scale.

2.3.2 Graph-based methods to reconstruct structural rearrangements

Determining the evolutionary path from one genome to another via structural rearrangements has been a long standing problem in computational biology and has a rich theoretical background. The problem can be thought of as a puzzle. One starts with a sequence of lettered blocks on a string, representing genes in a genome, and the object is to get the blocks in the same order and orientation as a second string, or genome, with as few cuts to the string as possible. The problem is easier or more difficult depending on whether the blocks are unique and whether adding blocks, gene duplication, is allowed.

The evolutionary distance between two genomes is often estimated by looking at the number of single nucleotide differences between them, as determined by aligning the sequences of two or more genomes. In 1992 Sankoff devised an evolutionary distance measure between two genomes based on rearrangements rather than small genetic variants gathered from sequence alignments [131]. He used different formulas to measure distance based on gene content (deletions and insertions) or gene order (inversions and transpositions) and computed the final distance as the sum of these two unrelated values. Several years later, Hannenhalli and Pevzner developed the first algorithm to find a series of inversions that could explain the differences between simple genomes in polynomial time and applied it to the mitochondrial sequence of cabbages and turnips [67]. They expanded their theory to include genomes with multiple chromosomes and fusions, fissions, and translocations in addition to inversion operations [66]. Yancopoulos, et.al introduced a double-cut-and-join (DCJ) operation, which can also represent inversions, translocations, fissions, and fusions, and found an algorithm that can determine the transformative sequence between genomes more efficiently than previous works [169]. The DCJ concept becomes important in future studies, as it offers an alternative model for measuring evolutionary distances with rearrangements. It unifies rearrangements affecting gene content with those affecting synteny and represents a unit of measurement for the distance between two genomes in evolutionary time, also called the edit distance. Just as maximum parsimony evolutionary trees are built to minimize the overall number of SNVs between genomes, the evolutionary history of a genome via structural rearrangements is built to minimize the number of DCJ operations between it and its ancestor. Furthermore, unlike SNVs, which are completely independent events until they are placed on a phylogenetic tree, DCJ operations may have an inherent ordering regardless of a phylogenetic tree. For example, although a single base may change and then has a 1 in 3 chance of reverting back to the original nucleotide, a gene cannot be reappear as a duplication after the original has been deleted. The minimal, or most parsimonious, set of DCJ operations tells us both the evolutionary distance between two genomes as well as the evolutionary path.

Having made inroads to the problem of transforming genomes of equal content, the next hurdle would be to find the evolutionary path between genomes of unequal content. Hence, later studies incorporate duplications into their algorithms, one stipulating that only the descendent is allowed duplicated gene content while the ancestor must not [4], another handling duplications by discounting whichever copy made the evolutionary history less parsimonious [170], and a third allowing insertions and deletions but not duplicate regions [16]. Shao and Lin are the first to allow true duplications and also show that, with this allowance, computing the edit distance between two genomes is NP hard.[1] Although an exact solution cannot be computed, they are able to compute a range for the edit distance [140].

I have given this brief history of the problem of reconstructing evolutionary

[1] Problems in computer science are classified based on their "time comlexity", or the function of time it takes to compute a solution given inputs of varying lengths. In bioinformatics the input is often the length of a DNA sequence, and in the case of evolutionary rearrangement history, it is the number of conserved DNA segments. NP stands for "nondeterministic polynomial time" and essentially means that the solution to a problem cannot be found in finite time. Problems for which the solution can be determined in real time belong in complexity class P, which stands for polynomial time, or more specifically, deterministic polynomial time.

histories in order to illustrate that it is a very difficult problem. In fact, computing the rearrangement cost between two genomes cannot be done in a realistic amount of time, let alone determining the exact sequence of rearrangement events. (Unless, of course, the edit distance is very small and the answer very simple.) The CN-AVG method models duplications, deletions, and other complex rearrangements, putting bounds on the minimum rearrangement cost, as Shao and Lin have done [140]. Additionally, the CN-AVG method generates sets of DCJ operations capable of transforming an ancestor genome to the evolved form and is able to model a multiclonal population, essentially assigning each DCJ operation to a subclone. Reconstructing cancer evolution through DCJ operations can be attempted now that amplifications, one of the most significant phenomena in cancer evolution, and multiclonality, present in all tumor DNA samples to some extent, are included in the framework.

2.3.3 The Copy Number Ancestral Variation Graph (CN-AVG) theory

The Copy Number Ancestral Variation Graph (CN-AVG), as previously stated, was developed by Drs. Zerbino, Paten, and Haussler, building on work described in the previous section and on the concept of a bilayered directed history graph, also developed by them [121] and which I explain on the following page. In this section, I will give an detailed overview of the CN-AVG theory, but I refer the reader to Dr. Zerbino's paper for the proofs and further details [173].

The Copy Number Ancestral Variation Graph (CN-AVG) represents a genome as a *sequence graph*, a graph where nodes represent breakpoint locations, and edges represent either genomic sequence, called *segments*, or the bonds connecting those segments, called *adjacencies*. The relationship between bilayered history graphs and sequence graphs is illustrated in Figure 2.3. Each edge has a copy number weighting, a value associated with it representing the number of times that sequence or bond is seen in the genome. As an example, a tandem duplication would be represented as a single segment edge with a weight of two and an adjacency would exist connecting the end of that segment to its beginning, with a weight of one or more than one if the segment is highly amplified. There are restrictions on the weights of the sequence graph. Namely, the total weight of segment edges at every node must equal the total weight of adjacency edges, which Zerbino calls the *balance condition*. This condition ensures that the model is biologically accurate because a single DNA molecule cannot split and connect to two different DNA molecules at one end, nor can there be free DNA ends. With DCJ operations, for every break in a genome, there must be a splicing back together. Zerbino defines *flows* as the set of values for the sequence graph that maintain this balance.

A *history graph* also represents genomes as graphs with conserved sequence blocks as directed nodes and the bonds between them as adjacencies, almost identical to the sequence graph described above. A third type of edge, a branch, connects a sequence block to its orthologous block in a second genome, which forms the second layer of the bilayered directed history graph. The history graph may have many layers which, in sequence, represent an evolutionary path from the utmost ancestor, the reference genome in this case, to the last descendent, the tumor genome. To reconstruct the

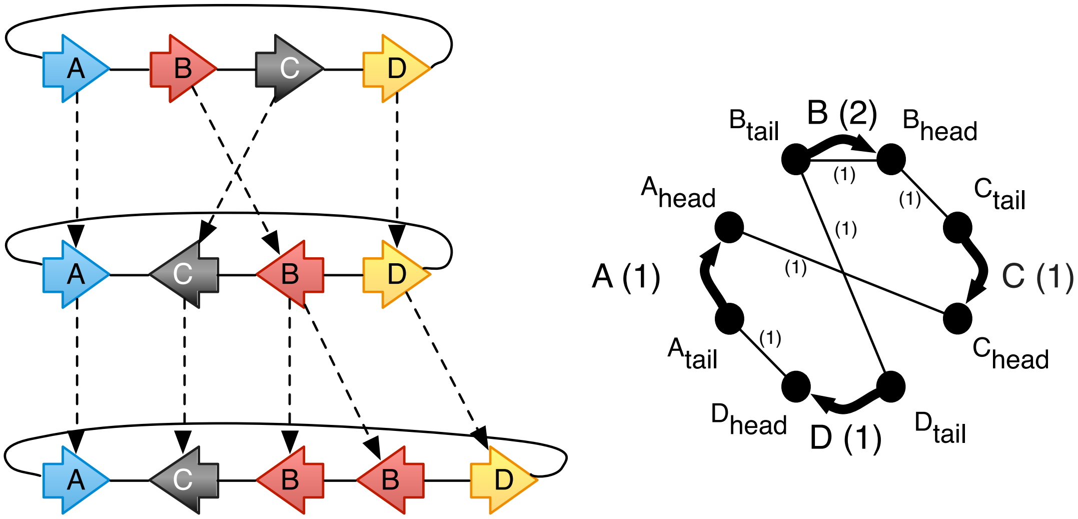


Figure 2.3: On the left is a layered directed history graph, with directed arrows representing conserved and continuous blocks of genomic sequence. Each layer is a genomic sequence at a point in evolutionary time. On the right is the corresponding sequence graph, with the arrows in the sequence graph now represented as directed edges. The numbers in parenthesis represent the copy number weights for the bottom sequence in the history graph. (Figure made by Dr. Zerbino.)

evolutionary history of a tumor, we must go from a sequence graph that we've generated from high-throughput sequencing data, to a bilayered history graph, which will tell us the evolutionary history of the tumor.

To begin, the reference genome is represented by the flow, or set of edge weights, in the sequence graph where every segment in the sequence graph has a weight of one and adjacencies connect them in the proper order. Likewise, the tumor genome can be represented by its own flow, which may traverse some edges multiple times and others not at all. We don't know the exact flow of the tumor genome since we don't know the exact sequence of the tumor genome, but we can estimate its flow using copy number and adjacency information gathered from the sequencing data. Other flows may represent genomes that are intermediate states between the reference and the tumor genome, layers in the bilayered history graph we are trying to construct, or they may represent genomic configurations that never happened in the history of our tumor. We want to find a sequence of flows that best represents the transitionary states and, because the sequence graph is a balanced bi-edge-colored graph, this turns out to be fairly easy to do using a greedy algorithm. Furthermore, Zerbino shows that the edit distance, or rearrangement cost of a flow sequence, is the sum of the cost between each of the flows in the sequence and that the cost between any two flows, f_i and f_{i+1}, falls within the limits given by the lower complexity, $\mathcal{C}^l_{f_i,f_{i+1}}$ and the upper complexity, $\mathcal{C}^u_{f_i,f_{i+1}}$.

A single flow sequence represents a believable evolutionary history, but there may be other flow sequences, alternative evolutionary histories, that are better, more parsimonious. We've already seen that finding the most parsimonious evolutionary history is an NP-hard problem, so we know that it is not feasible to generate all possible flow sequences. We wish to have a methodological approach for finding a more parsimonious solution, one that allows incremental changes to a flow sequence so that, combined with an efficient sampling algorithm such as Markov Chain Monte Carlo [105], better histories can be found over time. The CN-AVG method is able to do this through the use of flow changes and primary flows.

As a reminder, a flow simply represents a transitionary state in our evolutionary history; it is a snapshot of our tumor genome at some point in time. The difference between two flows, or *flow change*, takes a slightly different form than the sequence graphs I have already described. A flow change does not necessarily represent a genome because it may not contain any segments. For example, an inversion in a history graph presents as two flows with identical segments and different adjacencies. The flow change contains only the edges that are different between two flows (or have different weights), so the flow change in this case will have only adjacencies. These adjacencies must form a cycle that alternates between created and deleted edges since every break must be rejoined and the only thing for the broken ends to reconnect to is each other. A flow change representing an amplification would also be cyclic, with a newly created segment connected to an adjacency, and a deletion would be the same except the cycle would have a negative weight since the edges are being deleted. These concepts are illustrated in Figure 2.4.

In the CN-AVG, primary flows were constructed to be the building blocks of

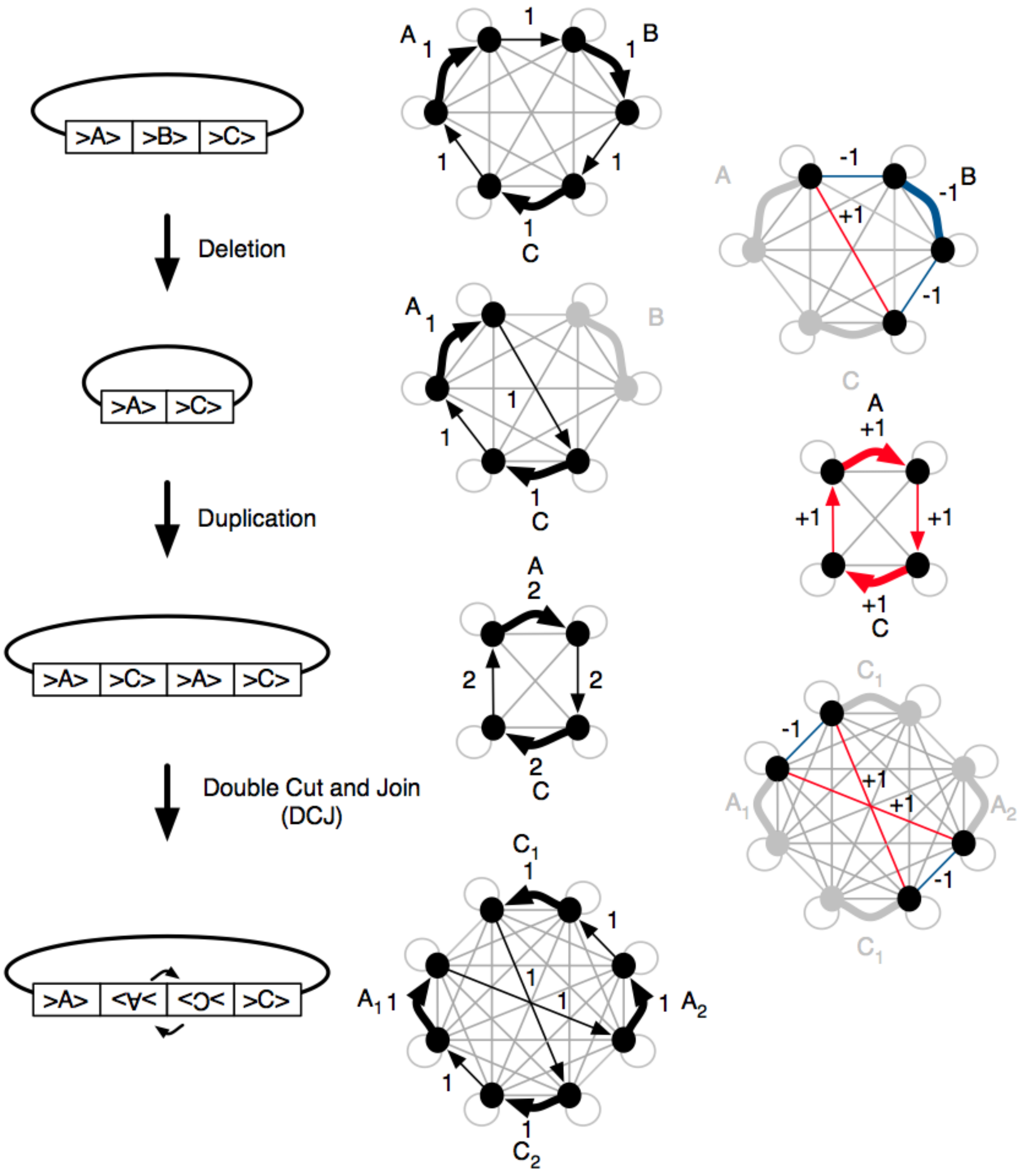


Figure 2.4: An simple evolutionary history in 3 different representations. An intuitive block representation is on the left, the sequence graph in the middle, and flow transitions on the right. CN-AVG works by finding sets of primary flow transitions. (Figure made by Dr. Zerbino.)

the evolutionary histories. Each *primary flow* represents a single rearrangement events such as a duplication, deletion, inversion, fusion, or a complex mutational event that is a combination of those. The manipulation of primary flows is what allows incremental changes in the CN-AVG history, making it amenable to iterative sampling techniques, as previously mentioned. The following paragraphs gives a more indepth description of primary flows and how they are constructed from the CN-AVG.

Since primary flows are the building blocks of the reconstructed evolutionary history, a primary flow is defined as a flow on the CN-AVG that cannot be decomposed as the sum of two flows without increasing the L_1 norm, defined as the absolute sum of the flows over every edge: $\|f\|_1 = \sum_{e \in E} |f(e)|$. The L_1 norm is essentially a measure of how much DNA changed over the course of the evolutionary history, and this is what we want to minimize.

CN-AVG uses a method for finding primary flows that allows edge reuse, since the most parsimonious histories often have edge reuse. To help us explore the space of all possible flows we describe flows using cycles on the graph. First, because it is easy to find alternating cycles in a balanced bi-edge-colored graph [123], we want to transform the sequence graph such that flow changes are alternating on a balanced bi-edge-colored graph. I previously stated that for inversions flow changes alternate between created and deleted adjacencies, so they are alternating cycles if the graph is bi-edge colored in regards to created and deleted adjacencies. However, flow changes representing copy changes do not alternate between created and deleted edges unless we change the graph in some way. Fortunately, there is a simple solution, what Zerbino calls a conjugate transformation, in which we invert the sign of adjacencies in the graph. Now both copy-neutral flow changes and copy-number alterations alternate between positive and negative values, and we can easily transform a graph back to its original state by flipping the signs of adjacency edges back. This is illustrated in Figure 2.5. It should be noted that changing the sign of adjacency edges is arbitrary, and we could just as easily decided to invert the segment edges. Now we have a graph that is balanced and bi-edge-colored such that we can find alternating cycles that represent flow changes. This set of flow changes will be more encompassing than the set generated through finding valid flow sequences. It will allow breakpoint and edge reuse, but it may also allow deletion and recreation of a genomic segment, an unrealistic scenario biologically. In order to correct for this, a penalty cost of two is added to the rearrangement cost of a CN-AVG history if the copy number of a segment falls below zero at any point in the history. The cost of two accounts for a model in which the ghost segment exists as a free-floating sequence that arose from a simultaneous duplication and fission event.

Given this set of alternating flow changes, Zerbino proves that primary flows can be generated from any *nested* and *synchronized* cycle, using the conjugate of the alternating weighting of that cycle. I refer the reader to Dr. Zerbino's paper for a detailed description and proof of these types of cycles [173]. Suffice it to say, both non-nested or non-synchronized cycles can be split into synchronized and nested cycles, so these special, minimal cycles are a representation of primary flows, and they can be merged and resplit to produce different evolutionary histories. In this way, the solution space for an evolutionary history can be explored systematically using Markov

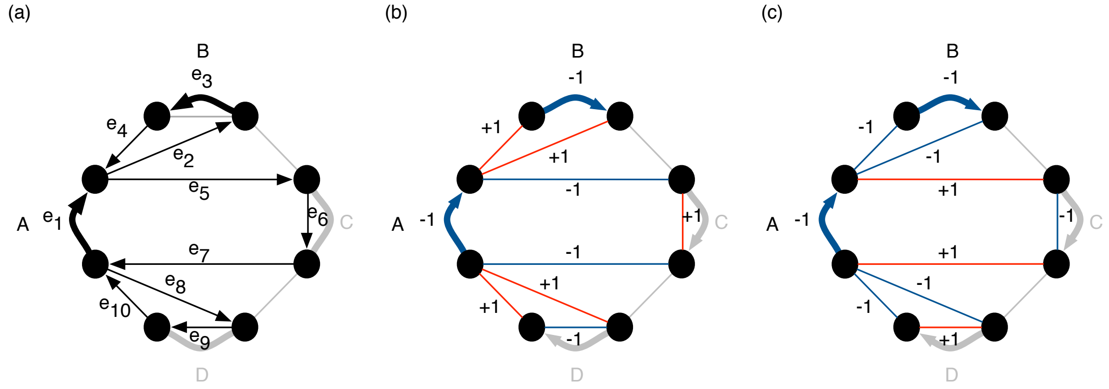


Figure 2.5: This illustrates the concepts of the conjugate transformation. a) an even length cycle traversal. b) The alternating weighting, alternating between created (positive) and deleted (negative) edges. This representation allows us to find primary flows, the building blocks of the history we are constructing. c) The conjugate of the alternating weightings, which satisfy the condition that the total weight of segment edges (thick lines) equals the total weight of adjacency edges (thin lines) at each node, and therefore represents a valid genome configuration. (Figure made by Dr. Zerbino.)

Chain Monte Carlo (MCMC) [105] sampling. MCMC sampling works by starting with a random solution, randomly changing part of the solution, and then keeping the new solution with some probability. We use an exponential probability function with an acceptance ratio: $\alpha_i = e^{k(\bar{C}_i - \bar{C}_{i-1})}$ where $\bar{C}_i$ is the average of the upper and lower complexity costs, each of which is the sum of the upper and lower complexities costs for each flow change in the history ($C_l = \sum_{j=1}^{n} C_{l,j}$). This function is arbitrary, and we assign k a value of 1, but decrease k if the sampling appears to be stuck and is rejecting all other solutions. We allow this flexibility to prevent the sampling from being stuck in a local minimum.

After many iterations, the solution can be approximated as the average of all the randomly generated values, and I will describe how I combine many evolutionary histories into a consensus evolutionary history in Chapter 4.

To summarize, the CN-AVG theory builds relationships between sequence graphs, history graphs, flows, flow changes, and primary flows in order to demonstrate that evolutionary histories can be constructed through generating sets of primary flows from a sequence graph. In essence, it allows us to go from a sequence graph, which we can construct using high throughput whole genome sequencing data for a tumor, to a bilayered history graph, which represents an evolutionary history from the normal genome to the tumor. Through the manipulation of primary flows, the CN-AVG pipeline iteratively approaches the most parsimonious evolutionary histories using MCMC sampling.

Chapter 3

Testing the CN-AVG method with simulations

Before applying the CN-AVG method to real patient data, I needed to determine whether the method worked and to explore its limitations. As described in the previous chapter on CN-AVG theory, the CN-AVG method generates increasingly more parsimonious evolutionary histories through MCMC sampling. In practice, the CN-AVG pipeline takes genomic copy number profiles and novel adjacencies as input and generates multiple evolutionary histories as it goes through MCMC sampling. In order to test the CN-AVG method, I needed a way to test how well the evolutionary histories predicted by the CN-AVG pipeline recapitulated a known "truth", so I first generated random evolutionary histories in silico to serve as my known or "true" histories. These histories involved only simple two- or four-edge cycles representing single duplications, deletions, and inversions of various sizes and copy number changes, and they were applied to single chromosome genomes composed of different lengths and unkown sequence. Although rich models of genome evolution exist (http://www.drive5.com/evolver/), this simple evolution model served the purpose of benchmarking the CN-AVG method, which is ignorant of sequence composition or gene content. To produce CN-AVG predicted histories, I ran the CN-AVG pipeline on the copy number profiles and adjacencies resulting from the final configuration of the random evolutionary histories. As previously stated, the CN-AVG pipeline generates multiple potential histories given a genome's evolved copy number profile and adjacencies, so I needed to develop methods for comparing the true evolutionary histories to the set predicted by the CN-AVG method. Last, I used these methods to calculate accuracy statistics.

3.1 Methods

In the following, I will refer to primary flows as *events* since each primary flow represents a single structural rearrangement event. The event may be a simple inversion, tandem duplication, or deletion, requiring only one DCJ (double-cut-and-join), or it may be a more complex rearrangement requiring multiple DCJs. Also, to review, events have values associated with them that represent the copy number change for that event and the prevalence. In a CN-AVG history, prevalence refers to the estimated fraction of tumor cells that have a certain event and is used to infer the order or timing of events within an evolutionary history. This is exactly what is done in evolutionary studies using SNVs; the variant allele frequency represents the prevalence of that SNV in the population of tumor cells, and is used to estimate the relative timing that SNVs occurred. An evolutionary tree must have the prevalence of any node (representing a mutational event, a structural rearrangement or SNV) be greater than the total prevalence of all of the immediate children of that node. Therefore, events that have a high prevalence are assumed to have occurred before events with low prevalence. For example, if event A has a prevalence of 0.8, and event B has a prevalence 0.2, we assume that the 20% of cells containing event B are a subset of the 80% of cells with event A, and therefore, that event A occurred before event B. Although we could just as easily claim that the 20% event B cells are a separate, unrelated subclone to the 80% event A cells, we make the simplifying assumption that they are subsets because this assumption is consistent with any prevalence values of event A and event B. For example, if event B now has 30% prevalence, it cannot be a distinct subclone, but must be a subset of the 80% clone.

Just as in cancer evolution studies using SNVs, accurate phylogenetic trees cannot be constructed without linkage information. In SNVs, this information comes from single-cell sequencing, and for our study, two events can be linked through parsimony. For example, a deletion overlapping an amplified region can be assumed to have happened after the amplification if more than two copies are deleted since a completely deleted region cannot be subsequently amplified. This type of linkage will only tell us whether event B depends on event A (in this case, the deletion of multiple copies, B, depends on the prior amplification, A), which would indicate that the event B subclone would be a descendent of the event A subclone. If event B is completely independent of event A, for example, if the deletion does not overlap the amplified region, we still have no way to tell if the event B subclone is distinct or part of the event A clone. In this case, we use a simplified linear evolutionary model in which the ordering of mutation events is directly related to prevalence, and assign B to be a descendent of A if the prevalence of B is less than A. Alternatively, we could generate all possible consistent phylogenetic trees to represent event relations, but this would add further complexity and computational toil to the CN-AVG sampling method.

As previously described, the CN-AVG histories are sets of primary flows, cycles in a graph representing structural rearrangement events. Each primary flow consists of a set of edges and nodes which contain information about the genomic location of the rearrangment, a prevalence value representing the fraction of tumor cells that have that rearrangement, and an integer representing the change in copy number associated with that rearrangement. CN-AVG histories can be represented less precisely as the sets of edges constituting the set of primary flows or as cycles or edge sets with only copy number change, ignoring prevalence values, or only a direction (positive or negative) of copy number change, ignoring prevalence values and copy number amplitude. For example, two primary flows may have identical breakpoints, segments, and adjacencies, but different copy number values associated with them; one history may have a three-fold copy change and another five-fold. In the evolution of cancer, I may want to consider these events equivalent because they may both represent an amplification/deletion of a genomic region or the creation of a breakpoint or fusion point in the genome. Hence, reduced representations of the CN-AVG histories allows more meaningful comparisons between them.

This type of simplification will be useful for comparing histories across different patients in which a gene may be frequently amplified to varying degrees (see Chapter 4), but it degrades the resolution of the evolutionary history within a single patient. Within a single patient, multiple amplifications of the same genomic region at different times during the evolution of the tumor should be distinguishable from a single amplification event, so they should not be combined or considered equivalent to each other. Restrictions can be put on event equivalence through enforcing matching copy number or prevalence values. Prevalence and copy number are related, so in most cases, restricting the value of one restricts that of the other. For example, an overall copy number of 3 in the sequencing data may present as a copy number of 3 in 100% of the cells or a copy number of 6 in 50% of the cells and so on for the CN-AVG histories. Since it is easier to test integer copy number values for equivalency than to attempt to define equivalence over a range of floating point prevalence values, for the following analysis, I consider two events or edges to be identical if they have the same structure and the same copy number value.

Having determined various metrics that I can use for comparing CN-AVG evolutionary histories, I used these to compare known true evolutionary histories, generated in silico, to the evolutionary histories predicted by CN-AVG. The purpose of this portion of my thesis is both to measure the accuracy of the CN-AVG method and to find an optimal sampling strategy.

3.2 Results

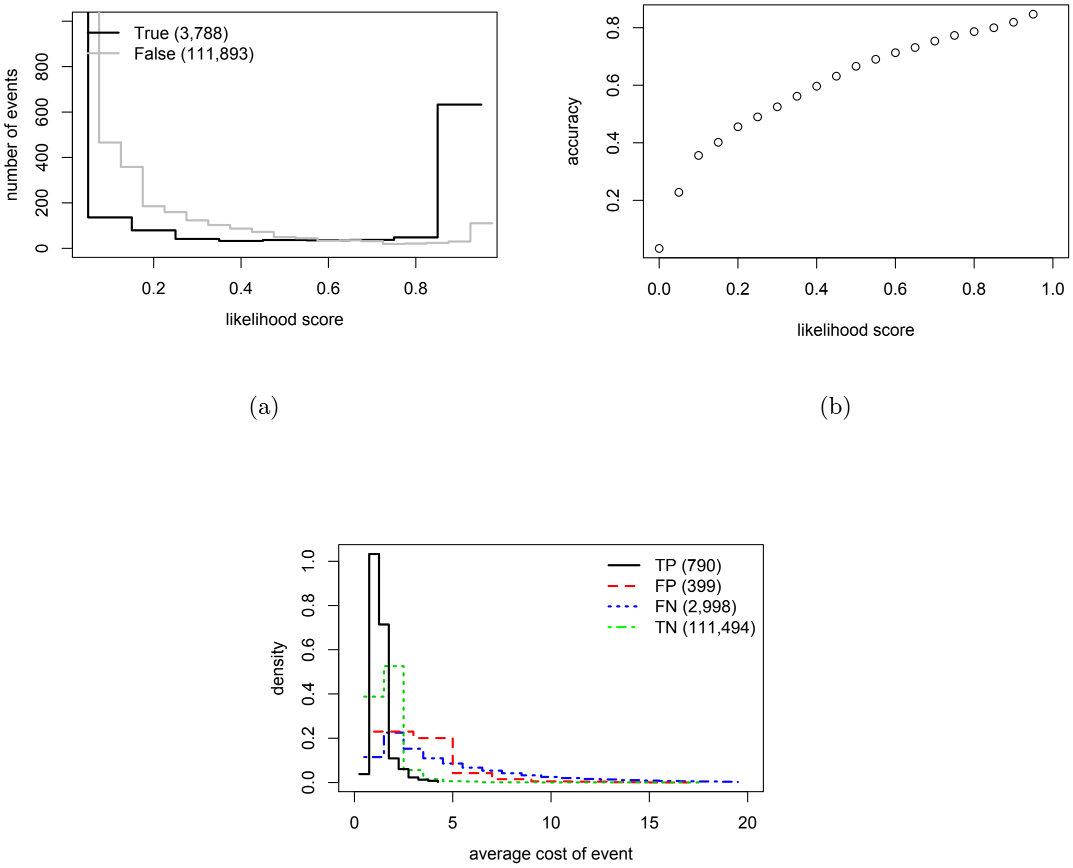
I generated 90 simulated evolutionary histories using CN-AVG code. As previously described, each simulation begins with a single chromosomal genome broken into 100, 150, or 200 blocks and subsequently rearranges the segmental blocks via a series of structural rearrangement events. Only inversions, deletions, or tandem amplifications are applied, with the number of events varying from 5 to 125, and the prevalence of each event given a value of 0.8, 0.7, or 0.6. This simulates a cellular population consisting of 3 subclones at 80, 70, and 60%. Each of these 90 evolutionary histories results in a new copy number profile and associated novel adjacencies, which are subsequenctly input to the CN-AVG pipeline. As previously described, the CN-AVG method uses MCMC sampling to generate sets of potentialy evolutionary histories. I combine all 9,000 evolutionary histories from 1,000 iterations across 9 independent sampling runs into a set of events. The likelihood score of each event is the frequency of that event across all 9,000 histories.

I examined the likelihood score of all events from the combined 90 simulations to see if the likelihood score could be used to distinguish true events, those that matched the true history, from false events, those that did not. Since the true histories are limited to certain simple types of events, events in the predicted histories that were equivalent to linear combinations of true events were also counted as true. I found that most events are false, but that events with high likelihood are more likely to be true, with 789/1185 (66.6%) events with a greater than 0.5 likelihood representing true events (Figure 3.1a).

In some cases, the CN-AVG found an evolutionary history that was more parsimonious than the true history, in which case we would not expect the true history and its events to have high likelihood scores. I did the same analysis excluding those simulations for which the predicted evolutionary cost was less than the true evolutionary cost and found that the accuracy of predicted events with likelihood greater than 0.5 improves only slightly to 69.6% (284/408).

I expected the incorrect or FP events to be more complex events not modeled by our simple simulated history evolutions. In order to verify this, I calculated the average cost of each event across all histories and plotting the distribution of event costs for each type of error classification (Figure 3.1c). Indeed, the events that did not match the simulated histories but had high likelihood (FP) had higher average costs than the TP events. The FN events, true events that we either detected with low likelihood or were unable to reconstruct, had a high cost, indicating that they were probably large events that overlapped many other previous or subsequent events. The cost of an event within a single history changes depending on the context of the rest of the history, so the structure of an event may be the same, but its individual cost will change depending on the rest of the history.

Not only does the cost of an individual event within a history depend on the set of other events constituting the history, but it depends on their ordering. The prevalence value plays a large role in the ordering of events, but I sought to test how well the ordering could be predicted based only on parsimony, so for this experiment, I simulated monoclonal evolutionary histories. For these, the prevalence of every event is 1



(a)

(b)

(c)

Figure 3.1: a) The likelihood score of each predicted event is a good predictor of whether the event exists in the true history. Although most predicted events do not match the true history, using a cutoff of 0.5, we achieve 67% accuracy. b) The probability of an event being true increases with the likelihood score of the event. c)The incorrect or FP events that we predict are higher cost events, indicating that they may be more complex events than are modeled in our simulations.

and therefore cannot inform the ordering of events. I generated 90 evolutionary histories, ran CN-AVG sampling on them with 10 independent runs of 1,000 iterations, and used the events from the lowest cost history as the ideal history for further analysis. The ideal set of events was randomly reordered, or shuffled, 1,000 times and the complexity cost recomputed for each reordering. To reiterate, in this second step, the set of events in the reconstructed history does not change, only the order of the events. In order to test how well the predicted histories reconstructed not only the set of events in the known history, but also their ordering, I looked at whether each pair of events was correctly ordered relative to the known evolutionary history. I used pairwise event orders rather than a rank analysis of the overall order of events because it is a more precise way of detecting codependency between events. For example, two events in a history may be constrained in their ordering, while all other events in the history are completely independent of eachother. In a rank analysis, it would be difficult to detect this ordered relationship amidst all other randomly ordered events.

Each pair of events can be classified in 3 ways. A pair is correct if both events are in the true simulated history and they are in the same order in the predicted history as they are in the true history. It is incorrect if both events are in the true history, but in a different order in the predicted history. It is unknown if either event is not in the true history. For events that we are able to reconstruct correctly, the ordering between them has no effect on the overall complexity score of the history (Figure 3.2). This result falls in line with the fact that we are able to predict well only the very simple, non-overlapping events, which are independent of all others.

In CN-AVG sampling, events are assigned a random order to start, and throughout subsequent iterations only the events which are changed (either merged or split) are reordered, but the entire order of events is never completely reordered or shuffled. Having confirmed that the ordering of events does affect the cost of the overall history, I wanted to see whether CN-AVG was finding the optimal history through selective reordering or if more shuffling should be incorporated into the sampling. I compared the minimum history cost from the purely shuffled histories to the minimum history cost of the regular CN-AVG sampling and found that in 7 out of 90 simulations, shuffling was able to find a lower cost history. However, the complexity cost improved by only 1 DCJ with shuffling versus without shuffling, indicating that regular CN-AVG sampling finds optimal or near optimal orderings through limited reordering (Figure 3.3).

As previously stated, equivalent events have equivalent copy number values, but the prevalence values, although related to copy number, may vary. Therefore, for each event I assign as the final or consensus prevalence the average prevalence of the event across all histories. CN-AVG accurately predicts the prevalence values for events and recapitulates the subclonal structure of the simulated tumor genome with peaks at 80, 70, and 60%, shown in figure 3.4. The additional peaks at 10 and 20% can be explained as the changes between the individual subclones. For example, an amplification that occurred in the 80% subclone could also be predicted to have happened in the 60% subclone and then again in a 20% clone. This is supported by the fact that events that did happen at 80% in the true history are more likely to be split into a 10% or 20% subclone. Figure 3.4 includes all true events, regardless of the likelihood, and in

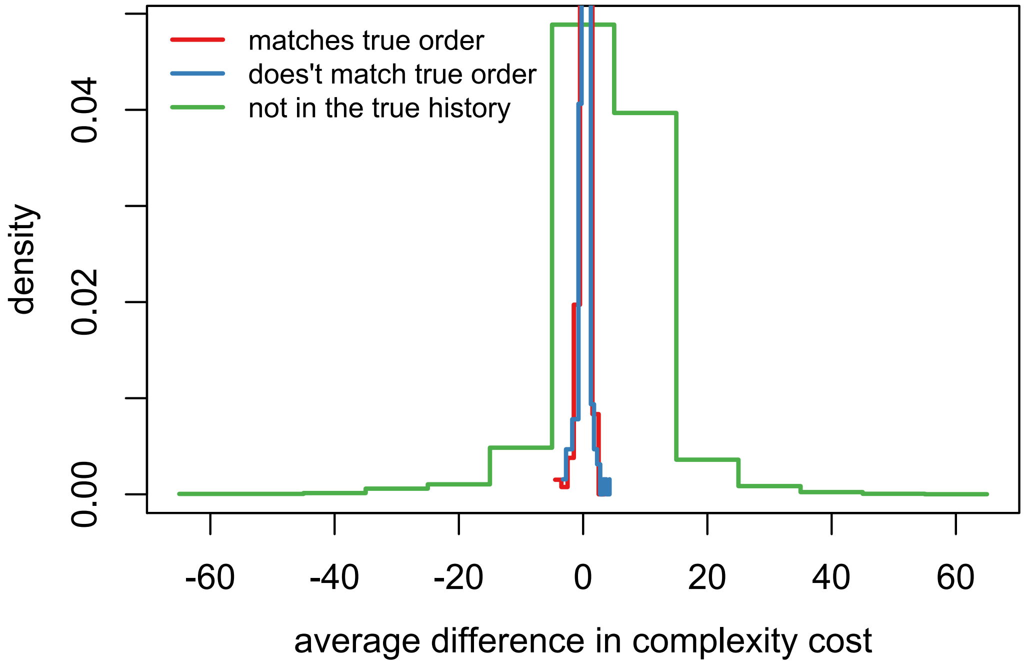


Figure 3.2: For each pair of events, A and B, the average complexity cost was calculated for histories in which A came before B and for histories in which B came before A. Also, each pair in which both events matched the true history was classified as matches true order if their order matched the true history or doesn't match true order if not. For most true event pairs, the ordering between them does not affect the overall history cost. Only for events which do not match the true history does the ordering between them affect the overall history cost in some cases.

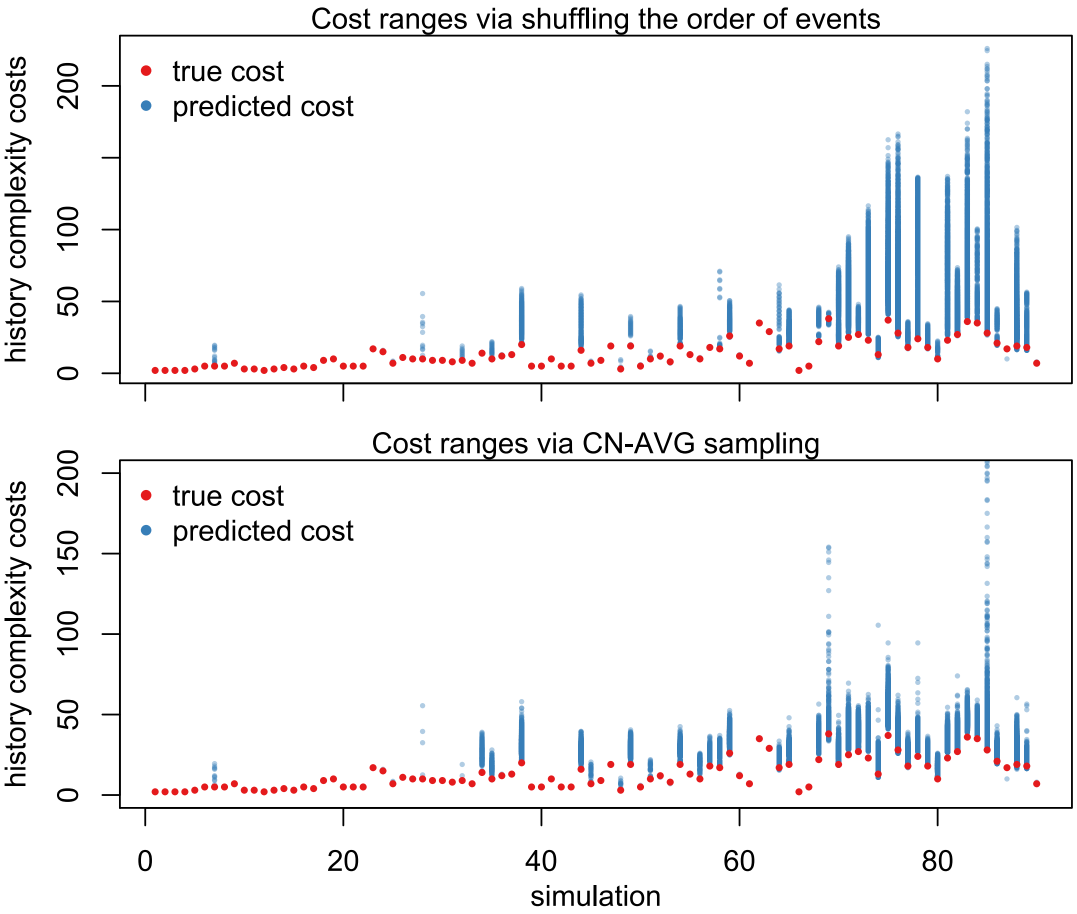


Figure 3.3: Here are 90 simulated evolutionary histories ordered by increasing breakpoint reuse. The lower panel shows the complexity costs for histories generated through normal CN-AVG sampling in which primary flows are merged or split and then assigned a new random order within the new CN-AVG history. The maximum cost for simulation 85 is 491. The upper panel shows the complexity costs for histories in which only the order of events has been changed. In only 7 out of the 90 cases does reordering the entire history find a lower cost than normal CN-AVG sampling, in which only restructured (split or merged primary flows) are reordered.

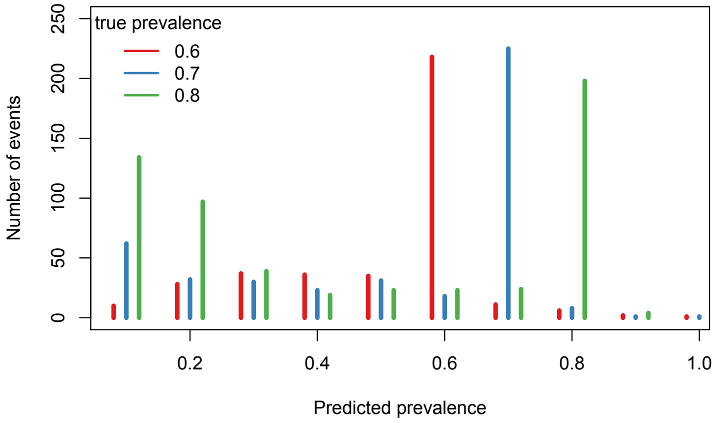


Figure 3.4: CN-AVG is able to recapitulate the subclonal structure of the tumor but creates smaller subclones at 10% and 20%, most likely representing the difference between the larger clones at 60,70, and 80%.

this case, we predict the correct prevalence value only 17% of the time. However, for events with likelihood of greater than 0.5, we predict the correct prevalence with 79.9% accuracy.

I generated statistics for each of the 90 simulations to see how the accuracy changed with different concentrations of events. Simple histories in which there are only a few structural rearrangements will be easier to reconstruct than those containing many rearrangements. The more events there are, the more likely it is that any two of them will overlap and result in alternative histories. In many cases, as previously mentioned, the true history is not the most parsimonious and may be impossible to reconstruct; for example, any rearrangement taking place in a genomic region that is subsequently deleted will be impossible to find. For simple histories with few events, we achieve 100% accuracy, but as the breakpoint reuse increases the accuracy drops exponentially

(Figure 3.5). I use the connectivity of the true history, defined as the number of edges in the graph divided by the number of nodes, to measure the breakpoint resuse. For histories with a connectivity above two, the average accuracy is 17.8%, which makes sense since at this connectivity on average there are two edges connected to each node, creating ambiguity in the reconstruction of every primary cycle representing a structural rearrangement event. Within the 1 to 2 connectivity range, we achieve an average accuracy of 62%. I measure accuracy by the F1 score, where $F_1 = 2 * \frac{precision*recall}{precision+recall}$, which is the harmonic mean of specificity and sensitivity.

I next divided events into three categories: amplifications, deletions, and copy neutral rearrangements, in order to see if we were able to predict certain types of events better than others. Events with only amplified segments were classified as amplifications, similarly for deletions, and events with no copy number change were classified as copy neutral. Some events contain both an amplified and a deleted segment, and these were excluded from this analysis. I found that CN-AVG did not reconstruct large amplifications as well as short ones, as shown in figure 3.6b, most likely because large amplifications are more likely to overlap and be broken up by smaller rearrangements. CN-AVG predicts copy neutral structural rearrangements most consistently with a 36.9% F1 score (Figure 3.6a), which makes sense considering that, although the genomic region being inverted may be large, the footprint of inversions is very small, only at the site of the breakpoints, so other rearrangements would have to overlap a precise region to change the footprint. How well CN-AVG predicts rearrangement events depends not only on the overall complexity of the evolutionary history, but also on the

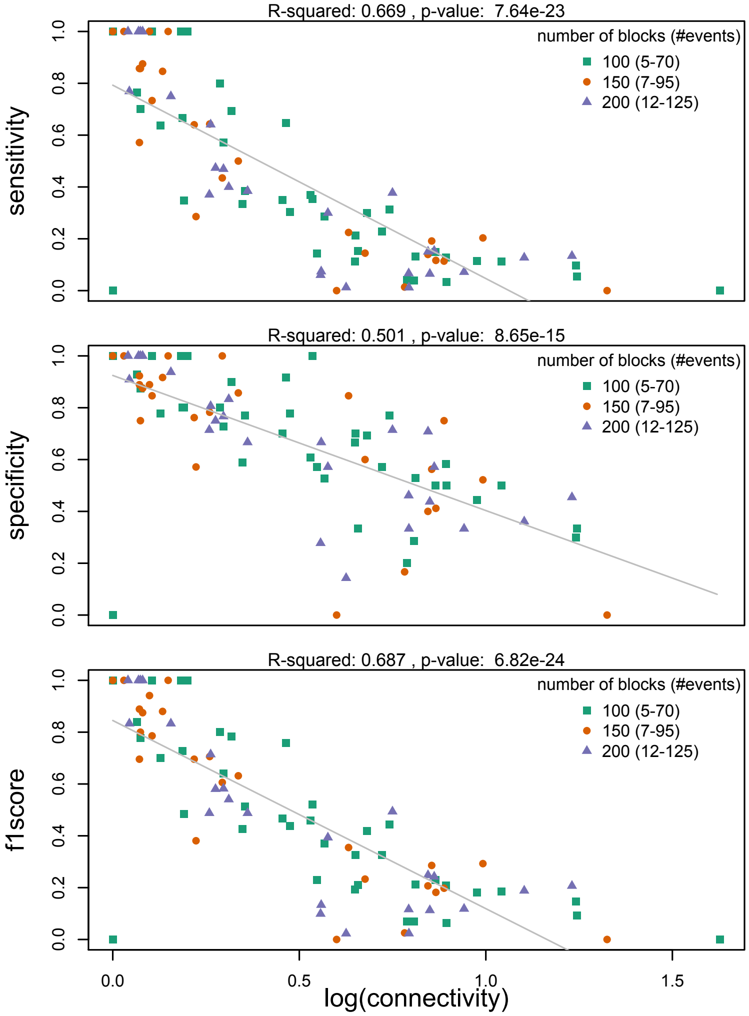
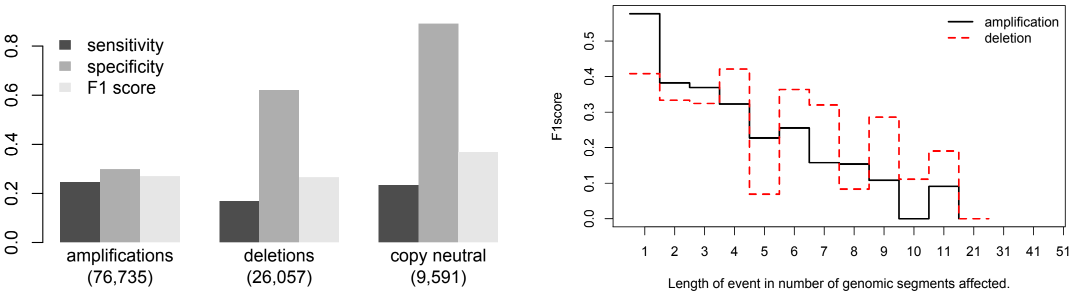


Figure 3.5: Specificity, sensitivity, and F1 score, the harmonic mean of the sensitivity and specificity, for each of the 90 simulations by the log of the connectivity of the true history. Connectivity is the number of edges in the true CN-AVG history divided by the number of nodes and represents breakpoint reuse in the evolutionary history. The line of best fit is shown in gray.



(a) (b)

Figure 3.6: a) CN-AVG reconstructs copy neutral events more reliably than amplifications or deletions. b) Smaller amplifications and deletions are also easier to predict because they are less likely to overlap other structural rearrangements.

type and size of the event.

Last, I wanted to examine the accuracy at various sampling time points to determine an optimal sampling strategy. I calculated the sensitivity and specificity for combined histories from up to 1,000 iterations and 10 independent runs, using a 0.5 likelihood cutoff for events to count as true. I found that sensitivity decreases as more independent sampling runs and iterations are done, but this is largely due to an artifact of the experiment (figure 3.7). Our simulated histories are generated using only simple DCJ events (events creating only two break points) and the sampling begins by finding the smallest cycles possible, so the initial sampled histories will better represent the true history simply because they are both favoring small cycles, or simple events. As the sampling proceeds, the later histories will consist of larger, more complex events, which will automatically be incorrect by our measure. In fact, the sensitivity falls between the narrow range of 19.1% and 21.06% (the total being the total number of events across all simulations, 3,693). The specificity increases with more sampling because the complex events, by default incorrect, also tend to be ambiguous and therefore inconsistent over sampling time. Since we are interested not simply in whether the predicted histories correctly recapitulate our biased simulations but in whether our sampling is sufficient to give consistent results, a better measure of our sampling quality is simply whether the high likelihood scoring events are consistent as we increase the sampling. We see that the number of consistent events changes only slightly above 4 independent runs and above 900 iterations (Figure 3.7), indicating that more sampling, either through independent runs or through longer runs, will not change the consensus CN-AVG history and that our sampling is sufficient.

3.3 Conclusions and discussion

Using simulated evolutionary histories, I have shown that CN-AVG is able to reconstruct simple independent rearrangement events very well and the likelihood score of an event reliably estimates if that event represents part of the true evolutionary history. The sampling method could be improved through periodically randomly reordering all events, but the improvement would be minimal and not worth the extra computational cost. The CN-AVG method can predict parts of the evolutionary history and can determine which parts cannot be reconstructed reliably because they are too complex.

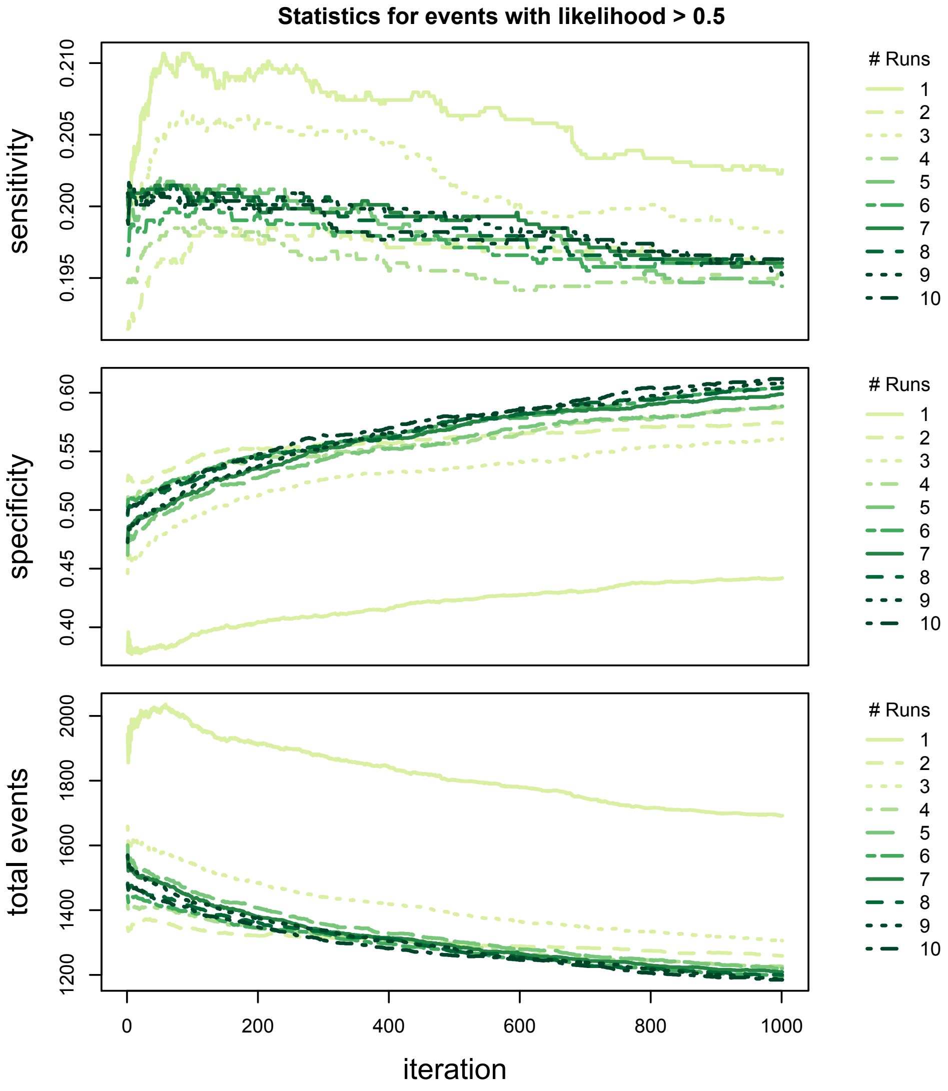


Figure 3.7: Using a likelihood score of 0.5 as a cutoff, the sensitivity and specificity of the CN-AVG predicted histories was calculated. The sensitivity of the method drops slightly with increased sampling due to a bias in the simulations. Sampling CN-AVG histories has diminishing returns, as shown in panel 3, with relatively few new events generated after 900 iterations.

Chapter 4

Applying the CN-AVG analysis method to glioblastoma

Although brain cancer is a relatively uncommon form of cancer, it is one of the most lethal (Figure 2.1) with an average survival time of less than a year. Glioblastoma (GBM) is the most common type of brain cancer [147], and current treatment includes surgery followed by radiotherapy. It is particularly difficult to treat GBM with chemotherapy because the blood-brain barrier bars many drugs from reaching the brain, but some chemotherapeutic agents can be beneficial in some cases and are undergoing clinical trials [52, 147, 70, 44]. In 5% of GBM cases, the brain cancer is linked to a genetically inherited syndrome, such as Li-Frameni or Von-Hippel-Lindau, that predisposes a person to various cancers, but in the majority of cases the cause is unknown. Linking environmental factors or infectious agents to GBM is an area of ongoing research [156, 135].

Clues about the temporal order of mutations in GBM have been gathered from genetically engineered mice and multi-sample analysis of single tumors. Zhu et al. used recombination technology to knock out p54 and NF1 in mice, inducing astrocytomas in cases where p53 was knocked out before NF1, but not vice versa [174]. Sottoriva et al. studied glioblastoma evolution by taking multiple spatially separated samples of single tumors and comparing somatic copy number alterations (SCNAs) between them. They inferred common or shared copy changes between distinct regions as being textitearly phase and unique changes as being *late phase* in GBM development. They also built phylogenetic trees using long single-molecule reads from each separate sample and found evidence for multiple subclones and complex hierarchies [142]. Single-cell RNA-seq was applied to GBM a year later, and again showed that GBM tumors are heterogenous, as the tumor cells showed a variety of expression patterns. They also generated RNA-seq profiles for nine normal cells and found much less variety their expression patterns [119].

GBM is a good candidate for application of our CN-AVG method because it has already been relatively well-characterized genomically. TCGA has published several studies of GBM, detecting single-nucleotide variants (SNVs) and somatic copy number alterations (SCNAs) across hundreds of GBM patients and uncovering significally mutated cellular pathways [51, 102, 17, 23]. These studies have shown that GBM contains relatively few SNVs compared to other cancers [158], indicating that copy number changes and chromosome instability may play a large role in the development of this cancer. Furthermore, the number of copy number changes found in most GBM patients may be “just right” for the CN-AVG method. Unlike ovarian or colorectal cancer which typically exhibit extensive aneuploidy, abnormal numbers of chromosomes, and chromosomal instability or blood cancers which typically have no SCNAs, most GBM tumors contain a handful of focal copy number changes and little aneuploidy [11]. As described in Chapter 3, the evolutionary history of genomic regions with a high density of structural rearrangements and copy number changes, as would be found in ovarian or colorectal tumors, often cannot be accurately reconstructed, and, obviously, tumors with few or no structural rearrangements would yield uninteresting evolutionary histories with few or no structural rearrangement events.

4.1 Methods

For each patient's tumor and normal, or blood, whole genome sequencing sample, bam[1] files were processed using bambam, a custom mutation and copy number calling algorithm developed by Dr. Sanborn [130]. Bambam generates copy number values across the genome by calculating the RPKM[2] values every several thousand reads, hence the size of the genomic region per copy value varies. Bambam also finds locations of possible novel adjacencies based on discordant read pairs.

The CN-AVG pipeline segments the genome, creating breakpoints where the copy number of the tumor changes or where bambam predicts a novel adjacency. The bambam copy number data is also smoothed using CBS [118], a method specifically

[1] bam stands for binary alignment file. It is the standard format for representing genomic alignments of sequencing reads from high throughput sequencing experiments.

[2] RPKM is Reads Per Kilobase Million, or the number of sequencing reads per kilobase divided by the total number of reads mapped to the genome in millions. It is a coverage measure designed to correct for the size of the region of interest and the total number of reads generated in the experiment.

designed for smoothing and segmenting noisy genomic copy number data. To briefly review the CN-AVG pipeline, as described in Section 2.3.3, CN-AVG produces a cactus graph [120] using the copy number profile and novel adjacency information, which it decomposes into cycles, or primary flows, representing sets of structural rearrangements. Each set of primary flows represents a potential evolutionary history for the tumor sample, and the pipeline performs MCMC sampling to generate and search for optimal evolutionary histories. For the GBM patients, 5,000 sampling iterations were done for 10 independent MCMC runs, producing a total of 50,000 possible evolutionary histories per patient.

Again to review, the CN-AVG uses MCMC sampling because it is impossible to examine every evolutionary history for a tumor. Instead, MCMC sampling progressively generates better solutions by comparing each new history with the previous one and keeping the new solution with a probability based on how much better the new solution is from the previous one. Although a single best solution may not be found, the output will represent the set of best solutions. To determine a consensus evolutionary history for a patient, I combine the 50,000 evolutionary histories into a large set of events, again using the frequency of the events across all 50,000 histories as the likelihood of that event. Identical events have the same edges and the same copy number, and the prevalence for a consensus event is the average prevalence over all instances of the event across the various histories containing it. If the same event occurs twice in a single history, it is counted as two separate events. For example, an amplification with the same amplitude that is predicted to occur twice within a history will be in the combined history twice as well, with the prevalence of the first instance being the average prevalence of the first instances and the prevalence of the second instance the average prevalence of the second instances. The combined history is used in downstream processing, which I will describe in the following sections. (See Figure 4.1 for a flow diagram of the complete CN-AVG analysis pipeline.)

4.1.1 Sampling diagnostics

Zerbino et al. demonstrated that the CN-AVG method is ergodic, so that any possible evolutionary history can be reached if given a long enough sampling run. Since our sampling is done in finite time, I wanted to test how much the evolutionary histories changed over time as a way to measure how quickly independent solutions could be reached. To do this I looked at both the cost function over time and the similarities between histories over time. One danger of MCMC sampling is that the sampling may fall into a local minimum, never progressing to a globally minimal solution. To prevent this from happening in our sampling we gradually increase the melting temperature allowing a worse solution to be kept by downweighting the history cost in the probability function. Specifically, we decrease k in the acceptance ratio, $\alpha_i = e^{k(\bar{C}_i - \bar{C}_{i-1})}$, by 1% each time a newly generated solution is rejected. This amounts to temporarily flattening the solution space to allow its further exploration. For most patients, the cost function reaches a minimum range over time, indicating that the sampling runs reached an optimal solution space but are not stuck on a single minimal solution. There are a few patients for whom the cost function does not reach a steady state, indicating that

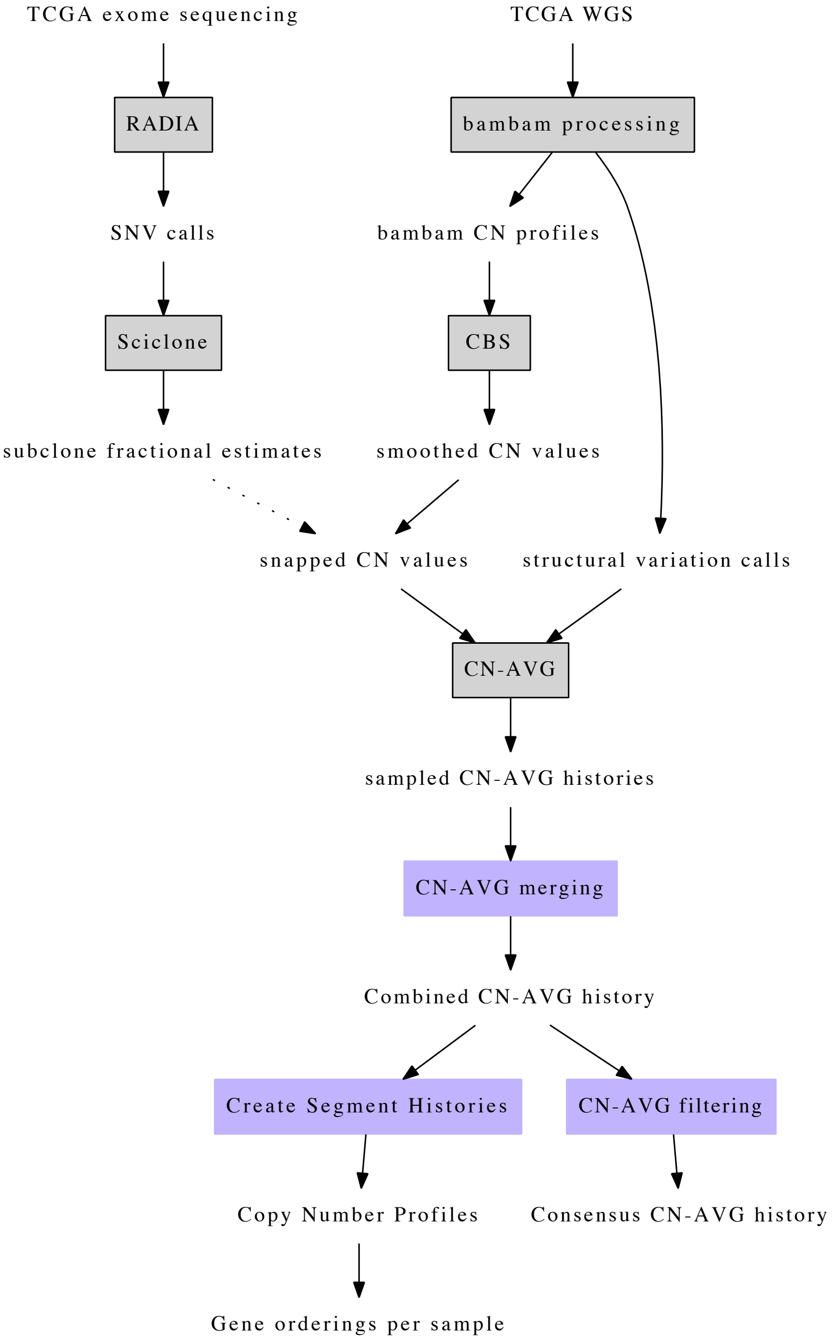


Figure 4.1: A diagram of the CN-AVG pipeline with program modules in boxes. Analysis steps I developed for my thesis are highlighted in purple and not outlined. The dashed arrow indicated an optional input to the pipeline.

we should continue sampling these cases to find their optimal solutions. Further work could be done to automatically optimize the sampling time, but even so, there must be a sampling limit due to computational time. For the purpose of this study, the limit was 50,000 iterations or 24 hours of compute time per run (each run uses a single compute node).

Two sampling runs may reach the same minimal cost while generating very different histories at that minimal cost. To test whether independent sampling runs were converging to the same solution, I looked at the similarity between histories of different sampling runs. The history similarity is simply the intersection of events between two histories divided by the union of those event sets. The initial solutions are very similar, but this is because of the heuristic used by the CN-AVG to initialize the sampling run which finds primary flows via the shortest paths in the CN-AVG. This heuristic results in high cost histories composed of large sets of smaller, more simple events. For most samples, the similarity between independent sampling runs does not increase over time, indicating that they are not converging to similar evolutionary histories and that the optimal solution space is large.

A drawback of MCMC sampling is that sequential solutions are dependent, and in high dimensional solutions such as CN-AVG histories, this dependence can span hundreds of iterations and can bias the end result. In other words, when each new solution closely resembles the previous one, the end result may be largely shaped by the random starting configuration rather than configurations within the optimal solution space because the sampling requires many iterations to reach the optimal solution space.

We correct for this bias by doing multiple sampling runs, essentially using multiple random starting points. However, another correction strategy would be to use only a subset of the solutions, discarding the majority that are shadows of a previous solution. In order to test how many iterations were needed to reach independence, I compared histories within the same sampling run over time. Two solutions within a sampling run should be as similar to each other as two solutions from different runs if they are truly independent. I found that in most cases, independence was never reached, indicating that each sampling run was limited to a subspace of solutions, and therefore independent runs are an important and more efficient way to find reliable consensus evolutionary histories.

In the previous chapter, I showed that the sampling strategy of ten runs and 1,000 iterations was sufficient to generate a consistent set of histories for the artificial histories generated in silico. It will be important in future use of the CN-AVG to run the same analysis on the results from real patient data to test for adequate sampling. Depending on the tumor, more or less sampling might be required, so I created a series of diagnostic plots designed to inform a researcher about the quality of the CN-AVG results. In addition to a histogram of the likelihood scores and the number of events with high likelihood scores per iteration, I showed the history cost and history similarities per iteration (Figure 4.2). Although not biologically informative, these graphs are important for judging the quality of CN-AVG results.

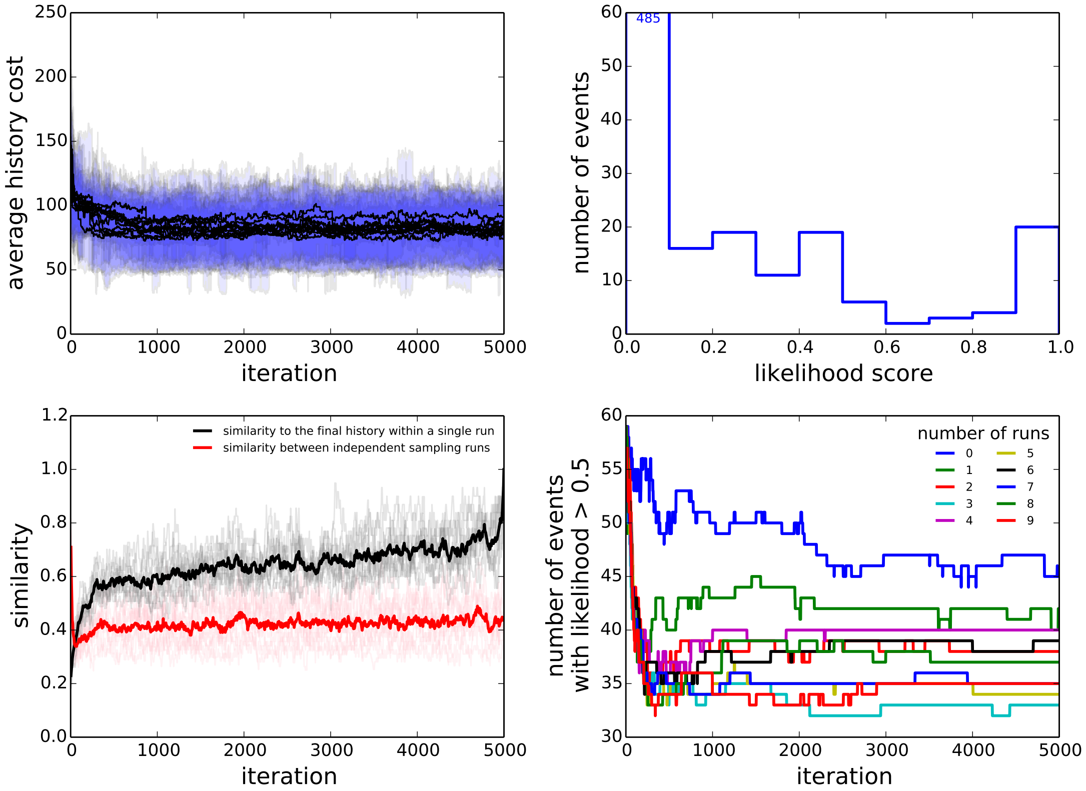


Figure 4.2: A series of plots designed to inform a researcher of the quality of CN-AVG sampling. The complexity cost of the histories should decrease and then reach a minimal range over time (top left panel). The similarity between independent sampling runs indicates how large the solution space is. For this patient, only 40% of the histories are identical, indicating that 60% of the predicted history is ambiguous (bottom left panel). The likelihood score distribution can be used to determine a cutoff for and estimate the accuracy of the final consensus history (top right panel). As with the complexity cost, the number of high-likelihood events should reach a plateau once enough sampling has been done. The bottom right panel shows the number of consistent events for increasing numbers of runs and iterations. For this sample, there are between 35 and 40 consistent events, and these events are found after seven independent runs and approximately 2,000 iterations.

4.1.2 Transforming CN-AVG output to biologically meaningful results

After producing a combined evolutionary history from CN-AVG sampling, the history ought to be translated into something interpretable by other researchers. Namely, the primary flows need to be annotated, so the evolutionary histories can be read as a series of gene perturbations. I did this in two ways. The first was to annotate events with the genes that overlapped a copy number change or that intersected a breakpoint within that event. To speed up the annotation process, rather than annotate every event in the combined history, I discarded events with likelihood scores of less than 0.1 from annotation and further analysis. This remaining set of highly likely events formed an evolutionary summary which could be compared across samples and used to find evolutionary patterns. Additionally, it could be visualized as a scatterplot (Figure 4.3). This type of visualization could aid biological discovery by giving researchers an easily understood summary of the tumor's evolutionary history.

For example, Figure 4.3 shows the summary evolutionary history for TCGA glioblastoma patient no. 0145. This patient has an amplification of EGFR, a commonly amplified oncogene that stimulates cell growth, of approximately 80 copies followed by smaller amplifications of CDK4. The amplification of EGFR is not shown in the plot because there are many different amplification events that could generate this amplification, and none of them have a likelihood score of greater than 0.1. (The likelihood score cutoff of 0.1 rather than 0.5 was chosen in order to display more events and give more information about the evolutionary history of the tumor, even though some events will not represent part of the “true” history.) Although we don’t see an amplification of EGFR at 80 copies, we do see a copy neutral structural rearrangement affecting EGFR (the yellow circle at ten copies near prevalence of 1). This copy neutral rearrangement is affecting around 10 copies of EGFR, indicating that at least ten copies of EGFR must be present at nearly 100% prevalence. Hence EGFR must have undergone an amplification of at least ten prior to this copy neutral event. We also see an early deletion of CDKN2A at around 0.9 prevalence, but this event is not as likely as the later deletion of CDKN2A at 0.2 prevalence, as indicated by the smaller circle size for the early event. CDKN2A is a tumor suppressor gene on chr9, and in this patient there is a chromosomal deletion of chr9, indicated by the purple vertical bar near 0.3 prevalence, and a separate deletion on a smaller region of chr9 at around 0.2 prevalence.

Driver mutations can happen to oncogenes, genes that stimulate cell growth, and to tumor suppressors, genes that control cell growth. In this patient, it appears that the EGFR amplification responsible for driving growth may also have been responsible for early tumor formation, while the tumor suppressor malfunction occurred later. This makes intuitive sense since a cell with increased growth would generate more progeny capable of acquiring more mutations. A cell with a defective tumor suppressor and no growth advantage would not necessarily overtake the cellular population of a tissue before acquiring further mutations necessary for tumor formation. Instead, such a cell is more likely to simply die off because tumor suppressors are often involved in repairing DNA damage. Increased mutations without increased cellular growth may just as readily lead to cell death as to cell growth over time. The display and annotation of CN-AVG

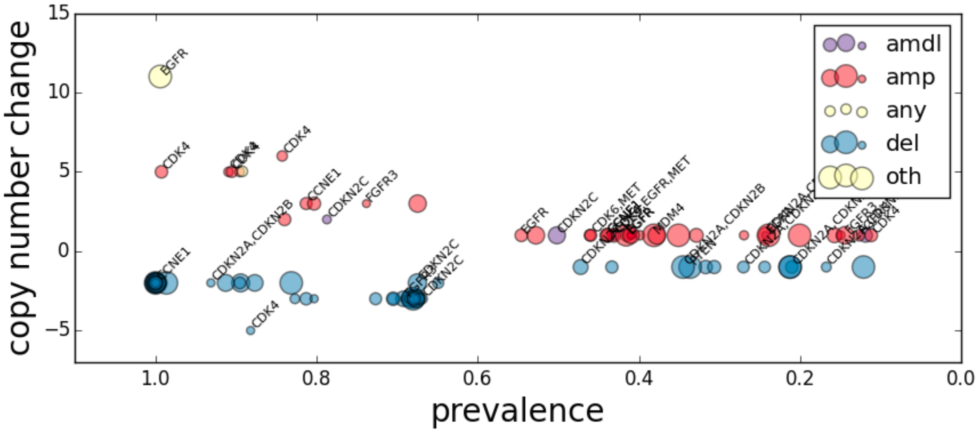


Figure 4.3: Each point represents an event, labeled with known GBM genes that they affect. Only events with likelihood greater than 0.1 are shown, and the size of the dot represents the likelihood of the event, while the color represents the type of event. Events may appear to be duplicates because they are very similar and yet occur in separate evolutionary paths. The genes used for annotation were selected from [17].

consensus histories helped piece together a story of how this patient's disease began and could give further insights to cancer biology as it is applied to more cases.

The second way of translating CN-AVG histories into gene perturbations is to generate consensus histories at the genomic level rather than the event level by essentially painting each primary cycle, or event, onto the genome. Because different CN-AVG histories may generate equivalent localized histories, this allows for a different kind of consensus history to emerge. For example, a region may be amplified as part of a whole chromosome arm duplication in one history, and as a small tandem duplication in another history. Although these are two very different amplification events, the effect on the genomic segment of interest is the same. I call the genomic-level consensus histories *segment histories* since they represent the evolutionary history of a single segment of the genome as a series of breaks and copy number changes.

For each unbroken genomic segment, I generate a copy number profile from each evolutionary history. Each profile is a sequence of copy number (CN) changes with associated prevalence values, and two CN profiles are identical if they have the same sequence of CN changes. The prevalence values associated with each CN change are simply the average prevalences over all histories that exhibit the same CN change sequence. The likelihood of any CN profile for a segment is the frequency of that profile across all CN-AVG histories.

Some genomic regions have no CN changes but may be affected by copy-neutral rearrangements. Copy-neutral rearrangements may incapacitate a gene by splitting it apart, or they may deregulate a gene by fusing a different promoter upstream of it, potentially leading to its hyperactivatity. Copy-neutral rearrangements can also form fusion genes, the most famous case being the BCR-ABL gene [8] that causes leukemia. To include these rearrangements in localized gene-level analysis, I generate breakpoint histories at the breakpoints of all primary flows. Instead of a CN change, a breakpoint history for a given evolutionary history is the number of times the breakpoint gets used within the history. For example, a breakpoint may be used once in one history, twice in another, and so on. The type of event, whether inversion, amplification, or deletion, does not matter in distinguishing breaks. Although the events using the breakpoint in the two latter histories may be very different, the history of that breakpoint will be the same for the latter two CN-AVG histories because it gets used twice, and a separate breakpoint history containing only one break would result from the first CN-AVG history. As with

CN profiles, I assign each break in the sequence the average prevalence of the break across all identical breakpoint profiles. In the previous example, the prevalence of the first break in the two-break histories would be the average prevalence of the first breaks in the two histories that use that breakpoint twice. The likelihood of a breakpoint sequence is its frequency across all histories, so again in my example, the breakpoint of interest has two breakpoint histories: a single break with a 33% likelihood, and a double break with a 66% likelihood.

As with the consensus histories just described, segment histories can also be visualized in an intuitive way, allowing other researchers to explore the data and generate hypotheses, as demonstrated in Figure 4.4. Here I show the same patient as in Figure 4.3 with the highly amplified region around EGFR apparent at the start of disease development, and the deletion of CDKN2A most likely occuring in a two step process afterward. Segment histories have an advantage over consensus events because they are able to summarize even the part of the evolutionary history that is too complex for CN-AVG to predict with any certainty. In this patient, for example, each evolutionary path of EGFR has a low likelihood, presumably because the large amplification of EGFR can be explained through numerous different amplification events or series of events. Nevertheless, we can see in the segment histories that the evolutionary paths of EGFR are all fairly consistent because they all show large amplifications at high prevalence. This is in contrast to CDKN2A for which there are fewer and more diverse segment histories, or evolutionary paths, and one that is clearly more likely than the others.

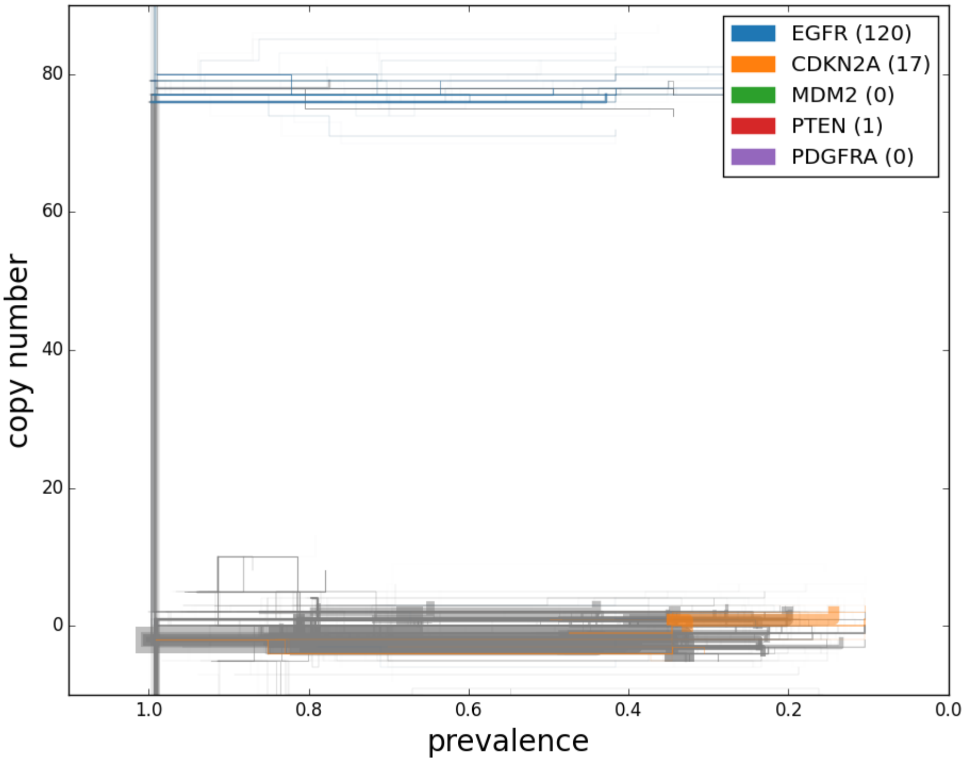


Figure 4.4: Shown are copy number profiles for all segments of the genome as the tumor evolves, using prevalence to estimate the order of somatic copy number alterations. The width represents the likelihood of that profile, and lines are colored by genes of interest that overlap the segment. Grey lines represent all other segments. The number of copy number profiles for each gene is noted in the legend. For this sample, we see a large amplificaiton of EGFR at high prevalence (early), while CDKN2A loss occurrs at a lower prevalence (later). See Figure 4.5 for a zoomed in look.

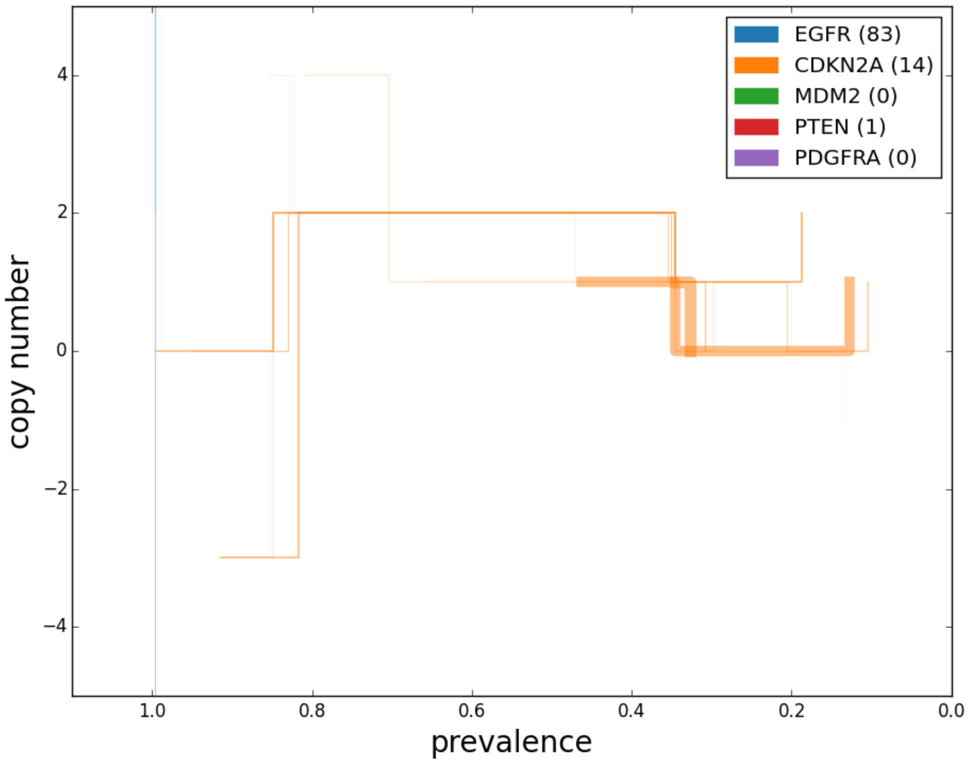


Figure 4.5: This is a zoomed in plot of figure 4.4 showing only the labeled segments to make the copy number profile of CDKN2A more clear.

4.2 Combining evolutionary histories across multiple patients

Having developed methods for researchers to view and easily understand and interpret CN-AVG for individual patient data, I next wanted ways to combine CN-AVG histories across patients. Both the consensus histories composed of the events with likelihood greater than 0.1 and the segment and breakpoint histories can be used to produce gene-level evolutionary summaries that can be combined across samples. The combined patient data from these two representations will take different forms and be amenable to different types of downstream analysis.

Representing patients' diseases via consensus evolutionary histories, I not only annotated each event by the genes that overlap it, but by the effect that the event has on the gene. An event may have four types of effects on a gene: an amplification, a deletion, a simultaneous amplification and deletion of different parts of the gene, or a "break," a rearrangement in which a breakpoint overlaps the gene but does not affect the copy number of the gene. For example, suppose an event represents a complex rearrangement, such as chromothripsis, in which there is a simultaneous deletion, amplification, and inversion. All of the genes that overlap or partially overlap a region of the genome that was deleted in this event would receive a "deletion" effect, genes overlapping an amplified region an "amplification" effect, and a gene that did not overlap any copy number alterations but that overlapped a breakpoint from a novel adjacency would have a "break" effect. In the end, each patient is described as a series of gene effects, each with an associated prevalence and likelihood inherited from the primary flow that caused the effect.

I combined these gene effects across all patients by clustering each set using the prevalence values. For example, amplifications of EGFR across all patients composed the set of “EGFR amplifications” to be clustered, while the set of all of deletions of EGFR were clustered separately (see Figure 4.6). I initially weighed each gene effect the same regardless of likelihood or patient source, and since I used only the prevalence values, the clustering amounted to finding the optimal intervals segregating gene effects. I used the Jenks optimization method to do this, implemented as part of the “classInt” package in R (http://cran.r-project.org/web/packages/classInt/index.html) [79]. Jenks clustering is essentially a one-dimensional k-means clustering. It attempts to bin datapoints to minimize the variance within each bin and maximize the variance between bins. I specified the number of bins to be the average number of gene effects per patient in the cohort. For example, EGFR is amplified by three separate amplification events on average across all patients, so there are three bins for EGFR amplification. In a secondary step, I assigned each gene effect to its respective interval, allowing only one gene effect per patient per bin. In cases where a patient has multiple “hits” in the same interval, I picked the one with the highest likelihood score. Last, I determined the center of each interval as the average prevalence of all of the gene effects in the interval, and noted the number of patients contributing to each bin. Through this method, researchers may detect overall patterns in how certain cancer types develop.

I applied gene-effect clustering to the cohort of 16 TCGA patients with GBM,

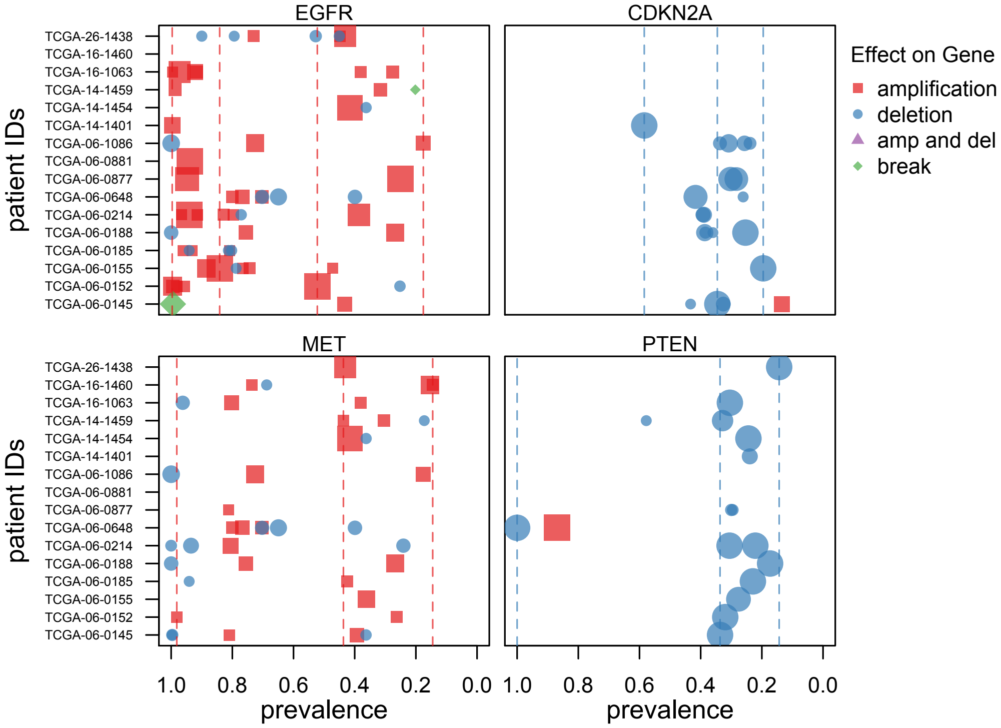


Figure 4.6: The effect of structural rearrangement events as predicted by CN-AVG on four genes frequently undergoing somatic copy number alterations in GBM. The size of the points represents the likelihood score of the event causing the copy number change or copy neutral break within the gene. The dashed lines represent the boundaries of clusters created from the most common gene effect per gene. For example, the top right plot shows gene effects for CDKN2A with the cluster boundaries of CDKN2A deletions as blue dashed lines.

using 29 genes that were known to have significant copy number alterations from a previous TCGA study (see Figure 1B from [17]) as my gene set. The results are shown in Figure 4.7. The gene with the most hits was EGFR, which had an average of three amplifications per patient. Genes near one another or on the same chromosome have similar evolutionary histories, as expected, due to whole chromosome amplification and deletion events.

Forming clusters of "gene effects" is useful for looking for evolutionary patterns, but a single summary statistic per gene rather than multiple clusters would allow the CN-AVG results to be used in traditional types of analysis applied to gene-level data, such as patient stratification from gene expression measurements. Since the purpose of the CN-AVG pipeline is to predict an ordering of mutational events, a natural summary statistic per gene per patient would be the timing of when the gene is first affected by a mutational event. As previously explained, timing can be inferred from the prevalence value of an event, so I used the prevalence values as summary statistics for genes. I assigned each gene a prevalence value using segment histories because, unlike events (primary flows), segment histories can readily be ordered genomically, thereby easily intersected with any other genome annotation.

For each gene, I calculated the total likelihood of all segment and breakpoint histories overlapping the gene. If this total is less than 0.5, this means that the gene is unaffected in most predicted evolutionary histories, so the summary score for the gene is zero. Interpreted another way, a prevalence score of zero means that the gene is unaffected in all cells of the tumor, and in terms of timing, this means the gene has

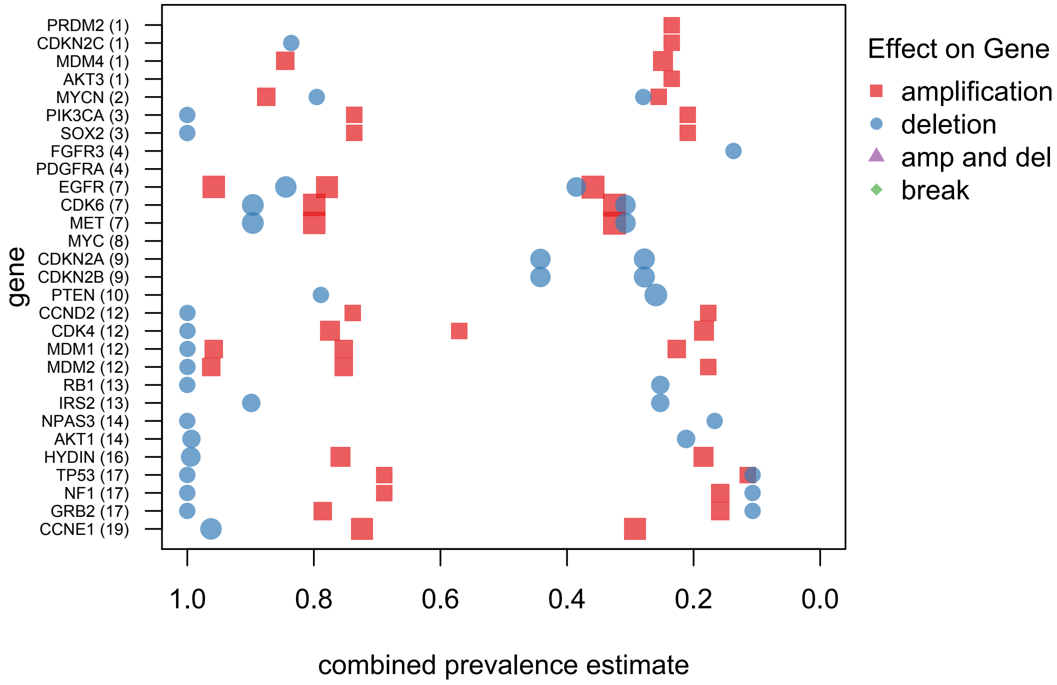


Figure 4.7: Each point represents the average prevalence of all points in a cluster of gene effects (see Figure 4.6 for an illustration). Clusters containing only one patient are not shown. Genes are ordered by genomic location, with the chromosome noted in parenthesis. From these analysis, we can see patterns in how glioblastoma forms, with large deletions on chromosomes 12 and 17 occuring early in the tumor, followed closely by amplification events of several growth factor receptors such as EGFR, MET, and the proto-oncogenes MDM1 and MDM2. We also see that gene deletions tend to happen at two separate times, as we would expect, to eliminate the two copies of the gene, while amplifications may happen more than twice.

yet to be perturbed. If the total likelihood score of all segment histories overlapping a gene is greater than 0.5, I took the segment or breakpoint history with the highest likelihood score associated with that gene as representative of that gene's history. Last, I took the highest prevalence value of all copy number alterations or breaks within that history as the summary value for that gene. This may result in gene histories that are inconsistent with any single CN-AVG history, but it is a useful way of summarizing the large set of possible histories for a single gene.

Distilling evolutionary histories into single values per gene allows for easy downstream analysis of large patient cohorts. As with examining mutational data across large cohorts, through this representation of CN-AVG results, researchers could discover potential oncogenes by detecting genes that are affected early (have high prevalence values) in a small subset of patients. Whereas mutational data alone may not be able to distinguish passenger versus driver mutations for genes mutated in a small subset of patients, the additional ordering information could signify oncogenic potential for early-mutated genes or passenger status for late-mutating ones. Furthermore, representing CN-AVG output as single values per gene allows patient cohorts to be easily merged into a single matrix, so patients can be clustered and classified by not only their gene expression patterns, but by the evolutionary history of their tumors. Combining multiple data types such as expression and copy number data across many tumor types for large patient cohorts led to new insights in a recent study which found that some cancers with different tissues of origin had similar molecular classifications [72]. Patient classification may be useful in determining treatment programs, and the single-gene summary values I've presented will allow CN-AVG to be part of that process.

4.3 Discussion

CN-AVG results contain two potentially problematic features for generating trustworthy evolutionary histories. One feature is the bias of favoring lower amplificaitons with high prevalence over higher amplifications with low prevalence. Highly amplified regions are common in GBM and other cancers in the form of double minute chromosomes or homogenous staining regions (HSRs) [130, 6], but the CN-AVG method is unlikely to be able to accurately predict when these amplifications occurred in tumor development.

The second feature of CN-AVG is a result of allowing whole chromosome deletions and amplifications for "free". Since only double cut and join (DCJ) events affect the complexity cost of an evolutionary history, an amplification could occur in two, essentially cost equivalent ways. The first would simply be an amplification of the region, and the second would be a "free" duplication of the entire chromosome containing the amplified region followed by a deletion of the part of the chromosome that is not amplified. This turns up in the combined histories in Figure 4.7 as deletions and amplificaitons immediately following each other. Figure 4.8 shows this phenomena for MET1 in TCGA patient 0145. MET1 is on the same chromosome as EGFR, chr7, and, although EGFR is highly amplified in this patient, MET1 is only slightly amplified. However, a significant number of evolutionary histories predict large copy number changes for

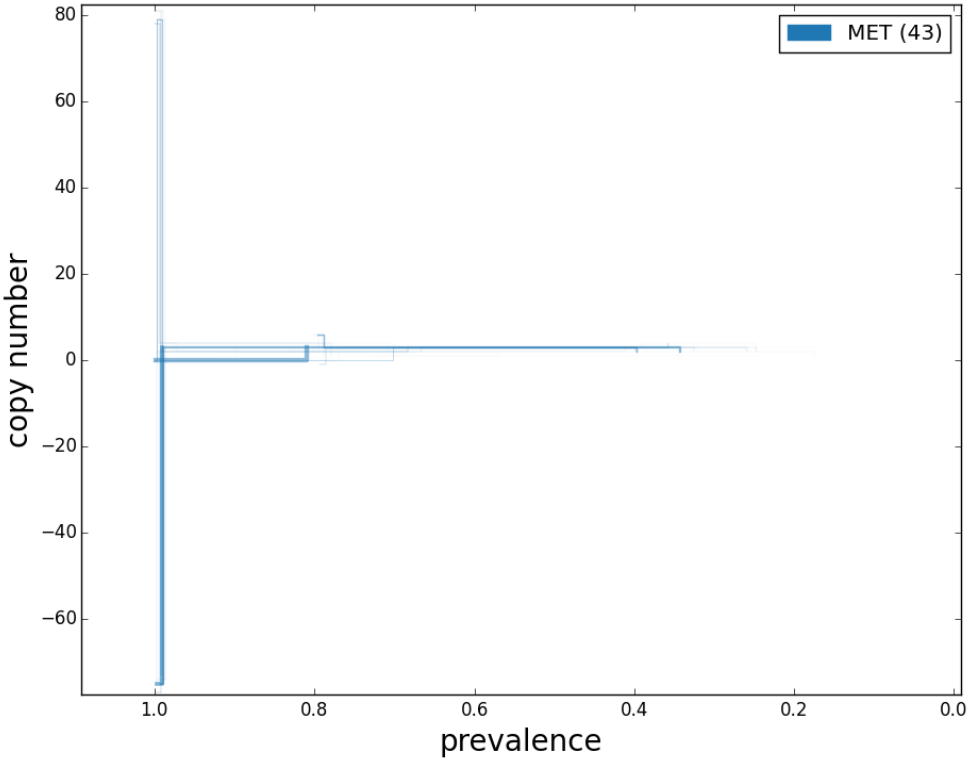


Figure 4.8: The copy number profile of MET1 in patient 0145 shows large sequential amplifications and deletions on the order seen for EGFR which is highly amplified on the same chromosome. This illustrates a phenomena in CN-AVG where entire chromosomes are amplified to account for a local amplification, and then select regions are deleted.

MET1 in close succession.

I attempted to filter such zero sum copy changes by filtering events in a CN-AVG history which were equivalent aside from inverted copy number values. This strategy did not work well because the events causing the rapid copy number changes are often different primary flows in addition to having different copy numbers. Very rarely is the same primary flow used twice in a CN-AVG history. I also attempted to filter events that did not overlap regions of true copy number change in the data, but abandoned this strategy because copy number changes can have various boundaries. Also, to limit the CN-AVG results to predetermined regions would potentially generate histories inconsistent with the true copy number values of the data.

Although the large zero sum copy number changes could be seen as a fault of the method, they also give us insight into cancer biology through understanding why they are there. For example, the MET1 amplification on chr7 (see Figure 4.8) may have been part of the larger amplification on chr7 and been subsequently deleted as part of a chromothripsis event. Furthermore, chromosomal instability, the duplication or deletion of whole chromosomes and whole chromosome arms, is part of the evolution of cancer, so in actuality, it may be a likely phenomena for a cell to acquire a duplicated segment through the deletion of a large portion of an extra chromosome.

4.4 Conclusion

I've developed methods for visualizing CN-AVG output so that it can be intuitively understood and used by other researchers to learn about individual patients' tumors, and I've transformed CN-AVG results to gene-level information for studying large cohorts of patients. The CN-AVG method for generating evolutionary histories is available at https://github.com/dzerbino/cn-avg, and the modules for combining and post processing the output can be found at https://github.com/TracyBallinger/cnavgpost. These expansions of the CN-AVG pipeline could form a basis for future work in the area of tumor genome evolution analysis. The displays I've developed could be made interactive, displaying detailed information about an event or segment history as a scientist clicks on it, incorporating zoom and span ability, and additional mutational information such as SNVs and indels could be displayed alongside.

I applied the CN-AVG pipeline to a cohort of 16 GBM patients from TCGA, and I will apply this to a larger cohort of 54 GBM patients from TCGA in the next couple months. With this larger dataset, I will look for genes or gene sets that mutate significantly early or late across the entire patient cohort by applying GSEA [148] to the data matrix of gene prevalence values. As discussed in Section 4.2, I will also apply heirarchical clustering to this larger patient cohort to test if the prevalence values for rearrangements are able to classify patients in a meaningful way by predicting survival, drug response, or other clinically useful information.

The CN-AVG theory is a powerful method for building evolutionary histories from whole genome sequencing tumor data, and I've demonstrated in this chapter how it can be used to study real patients and generate new biological hypothesis. Information gained from this work could also help guide patient treatment in the future.

Chapter 5

Retrotransposition in cancer genomes

The Cancer Genome Atlas project was initiated by the National Cancer Institute in order to characterize the genomes of hundreds of tumors of various cancer types. While much effort has been put into detecting SNVs in these data, transposable element insertions have not yet been studied. Transposable elements are mobile DNA sequences; thought to be remnants of ancient viruses, they can copy and paste themselves elsewhere in a genome and manifest as repetitive sequences [13, 87]. Transposable elements (TEs) are of particular interest in cancer because of several cases in which a TE insertion is directly linked to cancer formation [106, 65], because of growing evidence for somatic retrotransposition in the human genome [110, 29, 5], and because of the global DNA hypomethylation characteristic of tumors which is hypothesized to lead to retrotransposon reactivation [149, 68].

For this study, I worked with Dr. Adam Ewing to developed a computational pipeline to detect non-reference mobile element insertions from high-throughput paired-end whole genome sequencing data. Using this, I analyzed 86 whole genome tumor datasets with paired normal samples from TCGA across 7 different cancer types. I found the number of cancer specific insertions varied by cancer type, with most epithelial cancers having some degree of somatic retrotransposition, but AML, GBM, and ovarian cancers having none. I detected between 0 and 64 cancer-specific insertions per sample, and in total I detected 157 cancer-specific LINE insertions, 21 cancer-specific Alu insertions, and 1 cancer-specific SVA insertion.

5.1 Retrotransposon background

Retrotransposons are found in all eukaryotic genomes that we know of [64]. They are observed as repetitive DNA elements due to their capacity to insert new copies of themselves into the host DNA through a copy and paste process using an RNA intermediate [13]. They are categorized as either long terminal repeat (LTR) or non-LTR and further into families based on sequence similarity to other elements and by their mechanism of mobilization. The non-LTR retrotransposons that inhabit mammalian genomes are likely to mobilize through a mechanism known as target-primed reverse transcription [97]. Numerous retrotransposon copies have been detected in the human genome, comprising at least 45% of its DNA [87] and perhaps over two-thirds when highly sensitive TE detection methods are applied [34]. The most prolific retroelements in the human genome include LINE-1 and Alu sequences, comprising 17% and 8% of the assembled reference genome, respectively. During primate evolution, the general pattern of retroelement activity has been for one family of LINE element to be active at a time, suggesting either competition for a host factor or adaptation to evade one [83]. LINE-1 elements are autonomous retrotransposons encoding two proteins [136] responsible for both their own mobilization *in cis* and the mobilization of non-autonomous Alu elements [35], SVA elements [62], and processed pseudogenes [43] *in trans.* The activity of the human-specific LINE element, termed L1HS, was first recognized *in vivo* due to its ability to disrupt exons and cause Mendelian disease [82]. Since then, transposable elements have been linked to a variety of diseases, including cancer, through insertional mutagenesis of exons and regulatory regions near genes, disrupting gene function or regulation (see Hancks et al. 2012 for review). For example, in one case, an exonic L1 insertion was found in the APC tumor suppressor gene in colon cancer tissue but not in the normal tissue of the same patient [106]. Intronic retroelement insertions are known to affect splicing by providing 5' or 3' splice sites or disrupting sequence at the branch point [10, 63, 150]. Recent estimates place the rate of L1 retrotransposition in human genomes at one new insertion per every 100 to 150 live births [45, 76]. Since retroelements may clearly have an impact on phenotype and disease, it is important that retrotransposon insertion polymorphisms (RIPs) and mutations be characterized in genomic studies. A plethora of recent studies provides various means to document retrotransposon insertion polymorphisms (RIPs) segregating in human populations [9, 45, 73, 76, 78, 165, 47, 146], including one report of 9 somatic retrotransposon insertions across six lung tumors [78].

Cancer progression depends on the accumulation of somatic mutations, and recent evidence suggests that retrotransposition also occurs in some somatic tissues such as neuronal stem cells [110, 29, 5]. The observation of somatic retrotransposition in specific tissue types suggests tissue-specific regulation, either through known regulators such as APOBEC3 proteins [84, 108, 144, 24], germline piRNAs [2], and DNA methylation [171, 14], or through novel mechanism(s) not yet ascribed to transposable elements. Other lines of evidence for somatic retrotransposition include the aforementioned disease-causing insertions, observations of varying levels of transgenes originating through somatic retrotransposition in transgenic mice [80, 26, 40], and somatic R2 insertions in *Drosophila simulans* [42]. In addition to mutagenizing both somatic and germline genomes through new insertions, transposable elements play an important role in shaping gene regulatory networks by providing binding sites for transcription factors, including those highly important for cancer progression such as *TP53* and *SOX2* [161, 15, 86, 69]. Furthermore, genome-wide methylation status is often assessed through analysis of CpG islands located in the 5' UTRs of LINE-1 elements, which are typically heavily methylated [166], contributing to their quiescence in most somatic tissue types. Through this assay, a wide variety of cancers are found to be hypomethylated [117], leading me to speculate that retrotransposition rates may be substantially increased in certain cancer types or samples.

In order to test this hypothesis, I took advantage of whole-genome sequence data available through The Cancer Genome Atlas (TCGA). TCGA is an ongoing multi-institutional effort that will eventually include whole genome sequence data for hundreds of tumors and corresponding normal samples for over 20 different cancer types. Here, I consider transposable element insertions in the genomes of seven cancer types: acute myeloid leukemia (AML), breast cancer (BRCA), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), glioblastoma multiforme (GBM) [151], ovarian serous cystoadenocarcinoma (OV) [152], and colon/rectal adenocarcinoma (COAD/READ) [111], and present evidence for substantially increased retrotransposition in colorectal adenocarcinoma.

5.2 Methods

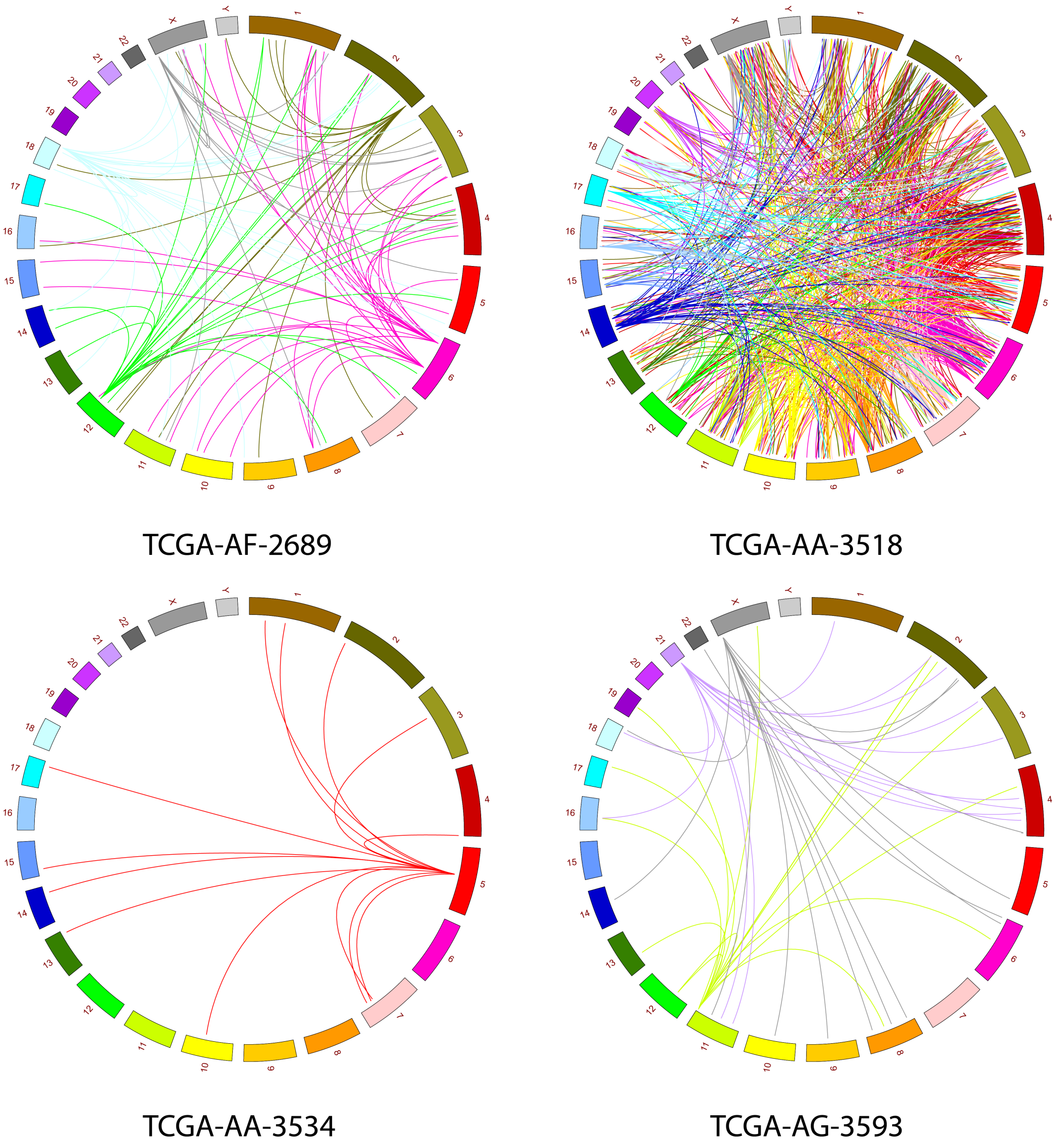
A number of successful computational methods have been devised capable of detecting transposable element insertions from whole-genome sequence data including VariationHunter2 [74], T-lex [48], RetroSeq (https://github.com/tk2/RetroSeq), HYDRA-SV [126], Tea [91], and most recently, TranspoSeq [71]. The approach outlined here, implemented as discord-retro (http://github.com/adamewing/discord-retro), has the advantage of working directly from the ubiquitous .bam sequence alignment format (as do RetroSeq and TranspoSeq) with minimal need for additional mapping apart from that required to identify insertion breakpoints. Here, I give a high-level overview of the discord-retro method. This method was developed in collaboration with Dr. Ewing. I created tools to read the .bam files directly and efficiently filter out discordant read pairs and split reads, using the samtools code base and UCSC's kent source code. This filtering was faster than using Pysam, the python interferface to .bam files, because it is written in C, and it is more efficient that using the SAMtools command line commands because it is able to intersect read mappings with multiple entries in a bed file, rather than one at a time. Intersection with a genome annotation was necessary to find discordant read pairs in which one read mapped to a known transposable element in the reference genome. My .bam filtering tools are also able to filter out reads that partially map, or split reads. Split reads were used to detect exact breakpoints of insertion sites.

Dr. Ewing created the discord-retro tool, which clusters discordant read pairs, maps them against a library of transposable element sequences, and make a prediction for the exact breakpoint based on target site duplicated sequence. Sequence data analyzed in this study was generated on the Illumina platform and aligned to a human reference assembly (NCBI36 or GRCh37) by TCGA Research Network members at TCGA Genome Sequencing Centers.

Paired-end reads can be classified based on how they map to the reference genome. A read pair is called *concordant* if both reads map the proper distance apart and in the correct orientation for the insert size and procedure used in the library preparation and sequencing, and *discordant* if these conditions are not met. For example, ends of a discordant paired read may map to different chromosomes, too far apart, too close together, or in the wrong orientation. A second type of improperly paired reads are ones in which one read maps to the reference, but its pair does not. These reads are referred to as one-end-anchored (OEA). Reads are called *soft-clipped* when part of the read aligns to the reference sequence, but either or both ends of the read do not.

I selected all discordant reads from both the tumor and normal sequencing data of a patient where one read of the pair maps to a unique portion of the genome, called the *anchored* read, and the other end maps to a repeatmasker annotation elsewhere in the genome. See Figure 5.1 for an illustration of this concept. I will refer to these types of read-pairs as one-end-repeat (OER) reads. I filtered elements corresponding to AluS and LTR elements from the results due to an overabundance of calls with no corresponding breakpoint predictions in some samples. Regions where the uniquely mapped ends of the OER reads clustered in two peaks with opposite orientation were considered consistent with an insertion existing between the two clusters of OER reads. I require there to be eight OER read pairs within a 500bp window, and for there to be at least two uniquely mapped or anchored reads on either strand. The requirement that both breakpoints, at the 5' and 3' junctions, be covered by paired reads reduces the chance of incorrectly annotating a segmental duplication, translocation, or inversion as a transposable element insertion.

The selection of clustered discordant OER reads yields a set of 20-50bp windows as predicted transposable element insertion sites. These were annotated as *germline* if there were discordant reads in both the tumor and normal tissue samples, as *somatic/cancer* if there were contributing discordant reads only in the tumor tissue, and as *somatic/normal* if there were contributing discordant reads only in the normal tissue. Insertion loci are cross-referenced against retrotransposon insertion polymorphisms (RIPs) cataloged in previous studies [160, 9, 45, 73, 78, 165, 47, 146] and against each other. As breakpoint resolution varies across studies, insertions within the same 500bp window were considered overlapping.



TCGA-AF-2689

TCGA-AA-3518

TCGA-AA-3534

TCGA-AG-3593

Figure 5.1: Discordant read "one-end repeat (OER)" mappings for the 4 colorectal adenocarcinoma samples with tumor-specific retrotranposon activity. Links are shown in the color of the chromosome where the insertion occurred. Figure made by Dr. Ewing.

5.2.1 Breakpoint refinement

For the breakpoint refinement, I extracted soft-clipped reads from the .bam files and Dr. Ewing used them to predict the breakpoints. Soft-clipped reads mapped using bwa [93] could be used to pinpoint a breakpoint in the insertion site. For each of the 14 samples aligned with bwa (Table S1), soft-clipped reads mapping within 500bp of each of the predicted insertion sites and which had greater than 10 bp clipped from the read were used to find a consensus breakpoint where a majority of soft-clipped read ends occurred at the same nucleotide in the reference genome. When breakpoints for both the 5' and 3' junctions between the element and the reference genome were detected, I identified target site duplications when the breakpoint on the forward strand occured 3-50bp downstream of the breakpoint on the reverse strand.

I also used a local assembly and realignment strategy to determine breakpoints for all samples. All discordant and soft-clipped reads within 500bp of a predicted insertion site were assembled using Velvet [172] with a k-mer size of 31, the shortPaired option, and insert length of 300. If the reads assembled into 5 contigs or less, these contigs were mapped back to the reference genome using BLAT. A cutoff of 5 contigs was chosen because when more contigs were present, they were generally too short to be more informative than the original reads. After mapping the assembled contigs back to the reference assembly, breakpoints present as the point where a contig no longer matches the reference sequence and begins matching a reference retroelement sequence. Target site duplications could be ascertained in cases where two assembled contigs had overlapping alignments to the predicted insertion site on opposite strands.

5.2.2 Simulation

To measure the accuracy and sensitivity of our pipeline, Dr. Ewing inserted 100 LINE, SINE, and SVA sequences randomly into the euchromatic sequence of chr22 from hg19/GRCh37. The retroelement sequences were randomly truncated on the 5' end up to 75% of the original element length for LINEs and SVAs, and 25% for Alus. Truncation was done because transposable elements are often found truncated in genomes, reflecting incomplete reverse transcription. Poly(A) tails between 20 and 70bp in length were added to the 3' end, and 12bp of the target insertion site was duplicated on the 5' junction to mimic target site duplications. Paired Illumina reads were simulated via wgsim (https://github.com/lh3/wgsim) to generate paired 75bp reads at 30x coverage. The wgsim tools is a very basic read simulator that uses a simple uniform substitution error model, and does not model indel error or chimeric reads. These reads were mapped back to the reference genome using bwa [93] with the following parameters: `-q 5 -l 32 -k 2 -t 4 -o 1`, alignments were processed with samtools [94] and used as input to discord-retro.

5.2.3 Assessing sequence similarity to reference elements

I performed local sequence assembly as described in Section 5.2.1 to generate contigs corresponding to inserted sequences. BLAT alignments of the contigs were carried out to find the most closely related elements in the reference genome.

Repeatmasker-annotated elements were scored by the sum of the products of the percent identity of a BLAT alignment times the length of the alignment for each contig that overlapped the repeat masked element. The element with the highest score was predicted to be the source element for the new insertion, excluding elements within 1000bp of the insertion site. Elements scoring within 20 of the highest score were considered as potential progenitors as well. In cases where there were several repeat elements tied for the highest score or very close to the highest score, the progenitor is considered ambiguous. I ranked the repeat-masked elements by the number of times they were predicted to be a progenitor for an somatic insertion, whether ambiguous or not, and examined the top 10 elements for retrotransposition capability.

5.2.4 Calculating COSMIC gene enrichment

I used UCSC Known Genes as the list of total genes in the reference genome. The longest transcript was chosen for each gene when multiple transcripts were listed with the same gene identifier, and genes that overlapped were combined into one genic region to give a total of 20,438 non-overlapping genic regions. I used the Complete working list (http://www.sanger.ac.uk/genetics/CGP/Census/), a curated list of genes that have been implicated in cancer [53], as COSMIC genes. I found 427 out of 487 COSMIC gene symbols matched a gene name in UCSC known genes, and disregarded the 60 gene symbols for which we could not find a UCSC known gene annotation. Again, I merged overlapping genes together resulting in 418 non-overlapping genic regions. I found 2,667 non-overlapping genic regions and 87 COSMIC genic regions which contained a non-reference insertion. Fold enrichment was calculated as $\frac{87/2{,}667}{418/20{,}438}$ and calculated a p-value using the hypergeometric distribution (R command: `phyper(87, 418, 20438-418, 2667, lower.tail=FALSE)`).

5.3 Results

Dr. Ewing and I developed a computational pipeline to detect non-reference retrotransposon insertions from paired end whole genome sequencing data by using mate-pair information from discordantly mapped read pairs (see Methods in Section 5.2). This pipeline, discord-retro, is available online at http://github.com/adamewing/discord-retro. Dr. Ewing measured the detection characteristics of our application by repeatedly inserting 100 retrotransposons into the euchromatic portion of human chr22 at random positions, generating paired reads, mapping to the GRCh37 reference sequence, and applying discord-retro (see Section 5.2 for details). He observed 87.9% sensitivity with perfect specificity when insertions into other insertions of the same class (eg., LINE into a LINE or Alu into an Alu) are discarded, and 94.5% sensitivity and perfect specificity if these insertions are allowed.

Using discord-retro, I analyzed 86 high-coverage (>30 times) tumor and normal genome pairs produced by TCGA, and identified retrotransposon insertions not found in the human reference genome (NCBI36 or GRCh37). For high-coverage data, the tumor and patient-matched normal paired-end sequence data were combined in order to distinguish between a non-reference germline insertion, which would be found in both tissues, from a somatic insertion, one found only in the tumor (or normal) DNA. Refinement of junctions using local assembly and analysis of soft-clipped reads allowed breakpoint-resolution on one or both ends of a predicted insertion. In many cases, target site duplications (TSDs) were identified, a 10-15bp duplication of sequence around the insertion site that occur as a byproduct of target-primed reverse transcription. I found that L1HS insertions had a high rate of TSD detection as well as members of the AluY and SVA subclasses (Figure 5.2), which is what I would expect given that these are known active retrotransposon families.

5.3.1 Germline Insertions

Across all 86 patients, I found 6,029 non-reference germline or polymorphic retrotransposon insertions which were not detected in a previous study [9, 46, 47, 73, 78, 146, 160, 165]. Of these new insertions, 645 were LINE elements, 5,081 were Alu, and 298 were SVA. The number of non-reference LINE, SINE, and SVA insertions found per patient follow approximately normal distributions with means of 100, 769, and 34, respectively (Figure 5.3). Ewing calculated an average of 152 L1 insertions present in an individual genome but not in the reference genome [45], and Huang estimated greater than 100 non-reference L1(Ta) per individual [76]. Both of these studies are more sensitive than studies using WGS to detect retrotransposon insertions, because they used targeted resequencing of L1 to detect polymorphic elements. Hormozdiari et al. characterized Alu RIPs using next-generation whole genome sequencing data, and estimated an expected 1400 non-reference Alu insertions per individual, with 27.1x

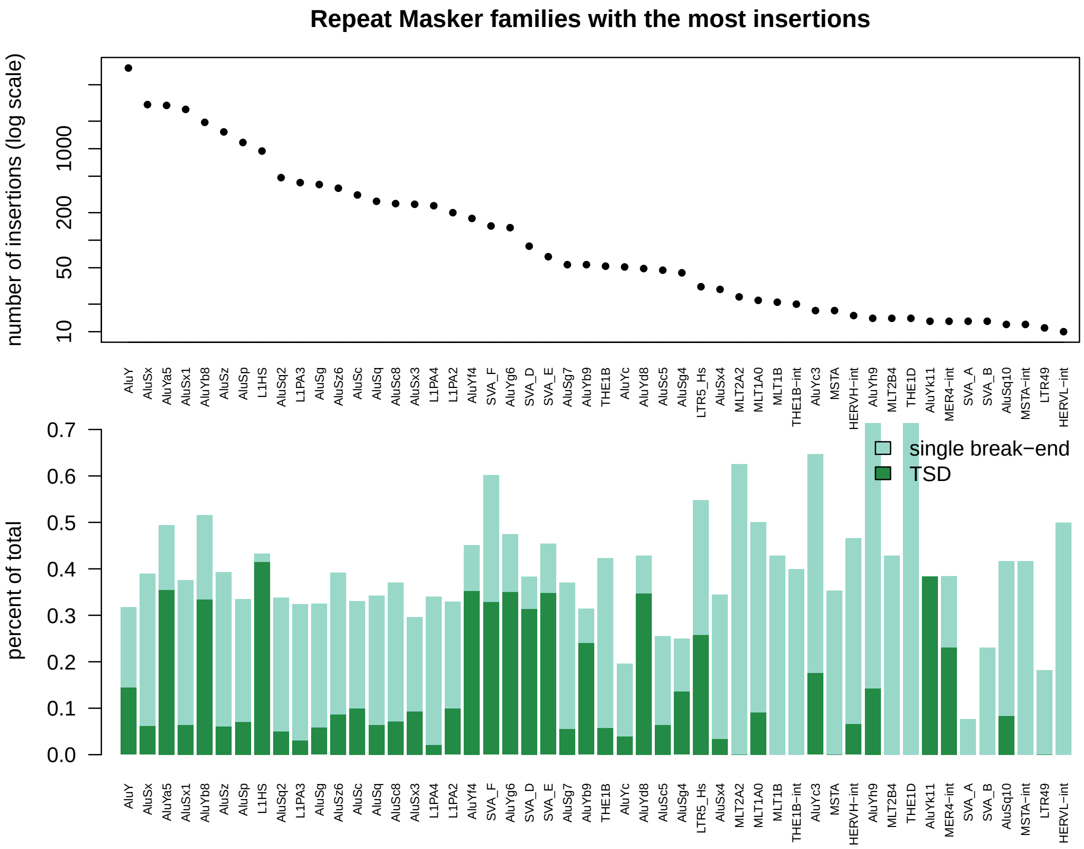


Figure 5.2: The most commonly detected type of novel insertions are the AluY and AluS subclasses, the most recent Alu lineages and the most abundant in the human genome. For each insertion, I show whether a target site duplication (TSD), a 10-15bp duplication of sequence around the insertion site, was identified. TSD provides evidence that the insertion is a results of retrotransposition as opposed to another type of mutation mechanism such as a tandem duplication. TSD detection is highest in the AluY and L1HS subclasses, the known active retrotransposon families, as expected.

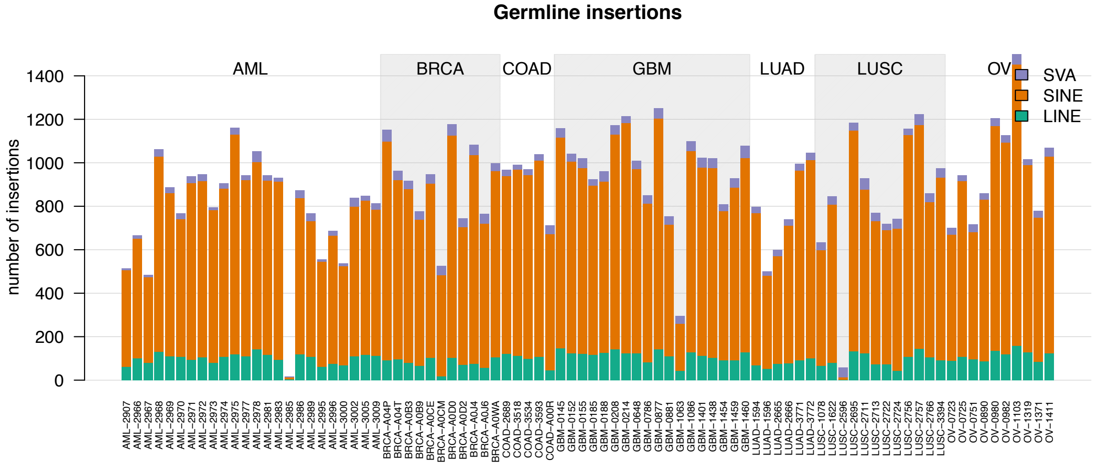


Figure 5.3: The number of non-reference germline insertions varied per patient, but I found an average of 98 LINE insertions, 751 SINE insertions, and 34 SVA insertions across all 86 patients.

coverage Alu (see Table 1 of [73]), and a study of RIPs from the 1000 Genomes Project estimates 224 (120-329) L1 differences, 1570 (1310-1870) Alu differences, and 80 (48-113) SVA differences between individuals [146]. Again, these estimates are higher than what we find, indicating that our criteria for calling an insertion are more stringent than those used in other studies. This is supported by the fact that our false positive rate derived from simulations is 0%, whereas the false discovery rate for the 1000 Genomes Project study is 2-5%. For this study, I preferred specificity over sensitivity in order to generate a high-confidence set of tumor-specific insertions that could be used in downstream analysis and in forming hypotheses about a giving tumor. For example, if I found an insertion in an oncogene of a particular patient, I wanted to be confident that it represented a true somatic insertion rather than an artifact before claiming that the insertion played a role in tumor formation.

Of the 6,029 previously undetected germline insertions found in TCGA samples, 42% were found within a gene, and a similar fraction of previously detected germline RIPs (41.5%) were found within a gene. I compared the set of genes containing germline retrotransposon insertions with a set of oncogenes catalogued by COSMIC [49] and found a 1.6-fold enrichment of oncogenes (3E-6 p-value). This indicates that some RIPs, similarly to some segregating SNPs, may confer a predisposition towards cancer to individuals who carry the mutation. This has been seen in the BRCA1 and BRCA2 genes in which there are germline Alu insertions in families with a predisposition to breast cancer [154, 98]. As more cancer patient genomes are sequenced and characterized, population-scale studies can be done comparing the genomes of individuals with cancer to a normal or healthy population to find associations between cancer and individual RIPs.

5.3.2 Somatic Insertions

Somatic insertions are those occuring exclusively in either the tumor or normal sample of a patient-matched pair of genomes and also not present in any other sample or catalogue of retrotransposon insertions from a previous study. Furthermore, because I combine discordant read pairs across both the tumor and normal tissue for an individual, I can be sure that if a tumor-specific or normal-specific call is made, not a single read that could indicate the presence of the insertion exists in the other sample. I found 162 somatic insertions across all patients. The number of insertions varies by tissue type with colorectal adenocarcinoma (COAD) and lung squamous (LUSC) displaying higher retrotransposon activity, with averages of 14.4 and 5.4 insertions per patient, as compared to breast (BRCA) and lung adenocarcinoma (LUAD), which have averages of 1.1 and 2.2 insertions per patient. (It should also be noted that colon sample COAD-A00R had an abnormally high number of normal-specific predictions which I believe to be an artifact. This sample is also characterized as hypermutated by TCGA.) I detected no somatic insertions in leukemia (AML), glioblastoma multiforme (GBM), or ovarian cancers (OV) (Figure 5.4). This is consistent with the study by Iskow, et al. in which they examined 20 lung tumors and found 9 somatic L1 insertions across 6 of them and none in 10 brain tumors [78]. Of the 162 somatic insertions we found, 154 were L1, 8 were AluY, and TSDs were detected in 27 of the L1 insertions, indicating that they arose through retrotransposition rather than another process such as homologous recombination. I found that 64 of the somatic insertions were within a gene, and 4 were within an exon or UTR. The genes containing somatic insertions within an exon, specifically, RBM4, PBLD, PPP1R1C, and SORCS3, are not well characterized.

5.3.2.1 Age and mutation rate as potential factors

I compared the number of other types of mutations such as SNPs and LOH for each sample to the number of somatic insertions in order to see if retrotransposon activity is simply a byproduct of hypermutation that occurs in tumors or if it may play a more active role in tumor formation. Retrotransposon activity that happens in the late stages of cancer may be happening simply because the machinery needed to silence mobile elements is no longer functional, and in this case, the somatic insertions

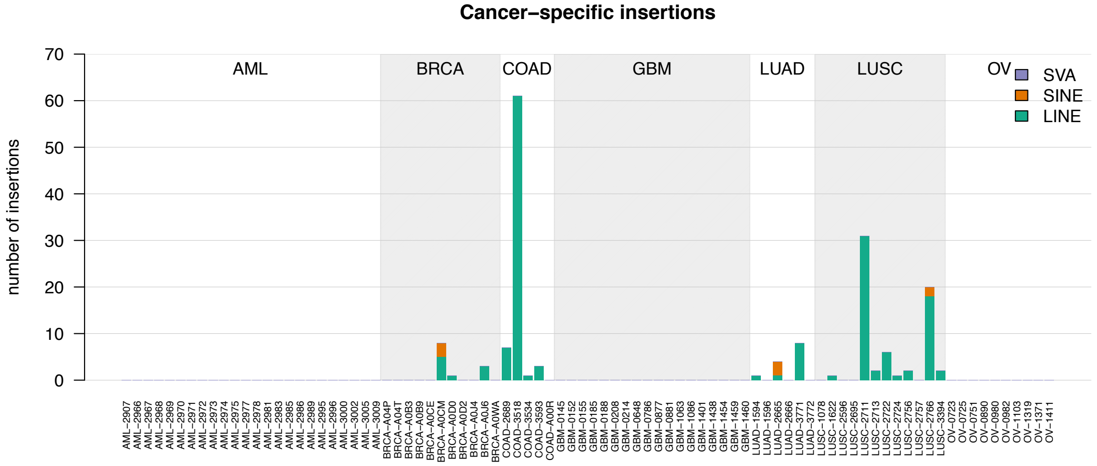


Figure 5.4: The number of retrotransposon insertions detected only in the tumor tissue for each of 86 patients analyzed. No cancer specific insertions were found in ovarian carcinoma (OV), glioblastoma multiforme (GBM), or acute myeloid leukemia (AML).

are likely to be passenger mutations. Cases in which retrotransposon activity happens early on in the tumor formation would present as more somatic insertions than expected as compared with other types of mutations, and would indicate that retrotransposition may play more of a driver role in the formation of the tumor. I used an in-house mutation caller, bambam [130], to detect SNVs, LOH events, copy number alterations (CNAs), and indels. I found a significant correlation between somatic mobile element insertions and other types of mutations (0.433, pvalue 8.0E-8 with Kendall's tau), and there seems to be a threshold below which samples do not have mobile element insertions (Figure 5.5).

As previously mentioned, retrotransposition has been shown to occur in the germline [167, 45, 76] and in neuronal precursor cells [5, 29, 110], but the baseline rate of retrotransposition in other types of somatic tissue is unknown. In order to

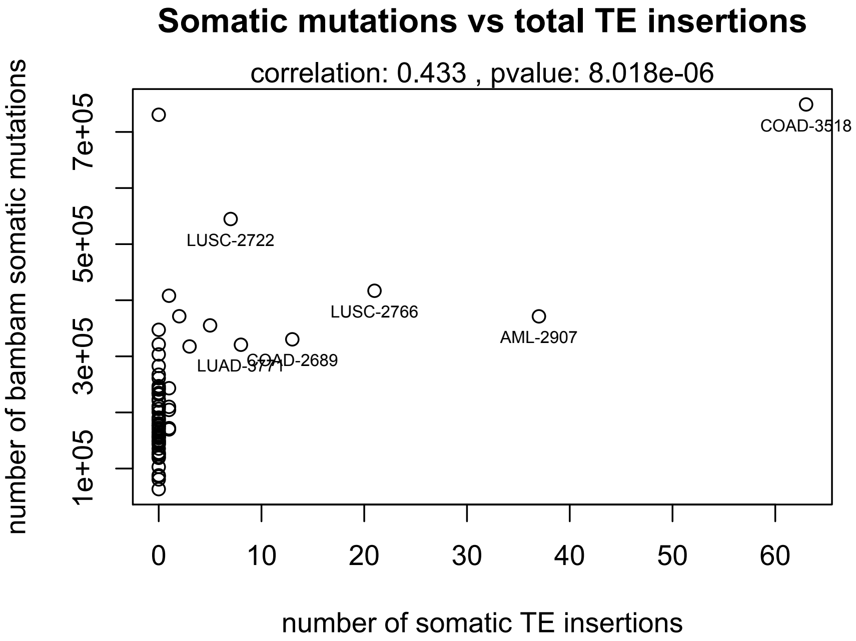


Figure 5.5: There is a significant correlation between the number of cancer-specific insertions and other types of mutations found in patients such as SNVs, CNAs, and small indels. The correlation was calculated using Kendall's tau.

test whether the retrotransposition I found in various tumors is a result of cancer or simply somatic variation, I looked for a correlation between the age of the patient at the initial diagnoses [56] and the number of somatic insertions. If retrotransposition is happening at some level in healthy tissue, I would expect older patients to have more somatic or cancer-specific insertions than younger patients. I did not find a significant correlation (Figure 5.6, but this does not rule out the possibility that retrotransposition was occurring in healthy, normal tissue. Regardless of whether the tumor-specific TE insertions that I found are a result of the tumor environment or simply the result of “normal” somatic retrotransposition, these somatic insertions may play a role in tumorogenesis.

5.3.2.2 Similarity to reference elements

After acquiring a set of insertion predictions for each sample, I sought to determine the closest element in the reference genome in terms of sequence similarity, as this may represent an element similar to the active progenitor element. In general, it is unlikely that the true progenitor can be identified through sequence similarity alone, as the active elements in the human reference genome diverge from one another by less than 1% [18, 137, 9]. That said, I performed this analysis for 72 tumor-specific L1 insertions in 4 colon cancer cases (Table 5.1) and found an enrichment of full-length, intact, human L1 elements, some of which are known to be active elements. This substantiated the claim that the cancer-specific novel insertions that I found are derived from active L1 elements.

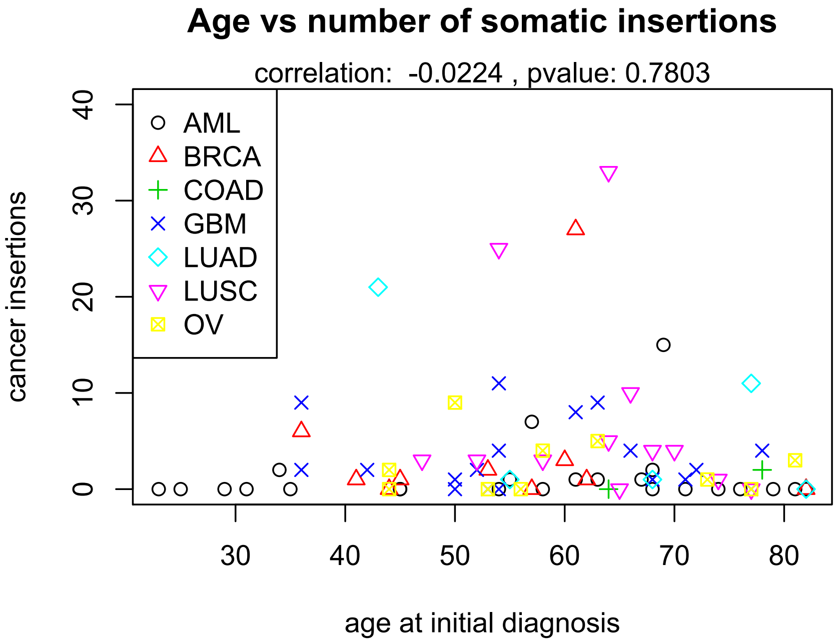


Figure 5.6: I did not find a significant correlation between age and the number of somatic insertions, indicating that the insertions that I found are cancer-related rather than occurring in normal somatic tissue. Correlation was measured using Kendall's tau.

LINE elements most closely related to inserted sequences

		Repeat		Best Related	Related	
	Location	Family	Length	Insertions	Insertions	Characteristics
1	chr10:107127095-107133125	L1HS	6030	14	24	ORF1 broken, ORF2 intact
2	chr11:92793801-92799845	L1HS	6044	12	33	intact
3	chr11:24306074-24312123	L1HS	6049	12	28	intact
4	chr17:65966693-65972723	L1HS	6030	10	26	ORF1 intact, ORF2 broken
5	chr11:60608423-60610418	L1HS	1995	10	18	truncated
6	chr7:49690411-49696442	L1HS	6031	10	18	intact
7	chr16:22618776-22619548	L1HS	772	8	14	truncated
8	chr18:43440743-43446771	L1HS	6028	7	19	ORF1 intact, ORF2 broken
9	chr4:182113842-182114873	L1HS	1031	7	11	truncated
10	chr5:12250459-12251104	L1HS	645	7	9	truncated

Table 5.1: Contigs of inserted sequence were assembled for some cancer-specific transposable element (TE) insertions and aligned back to the reference genome using BLAT. Repeat-masker-annotated elements from the reference genome are listed according to the number of times they have the highest sequence similarity to an insertion's contigs compared to all other repeat-masker-annotated elements (Best Related Insertions). The number of times an element has a high sequence similarity to an insertion's contig, as defined as having 20 or fewer mismatches to the insertion (in essence, is the most similar or a close second), is also listed (Related Insertions).

5.4 Discussion

As TCGA and others continue to sequence more cancer and paired normal cases across a wider variety of cancer types, we may uncover clear driver mutations caused by transposable elements and other cancer types that exhibit high levels of insertional mutagenesis by transposable elements. Notably, several other studies have recently been published that detect retrotransposon insertions in cancer genomes from high-throughput sequencing data, similar to this one [90, 141, 71, 30]. Helman [71] recently produced a very comprehensive study of somatic retrotransposition in cancer, including 11 different cancer types and a total of 967 patients using whole genome and exome sequencing from TCGA. A nearly submitted version of our study that includes only OV, GBM, and COAD cancer types is available on arXiv (http://arxiv.org/abs/1501.04268).

New technologies and the decreasing cost of sequencing will likely provide new insights into somatic retrotransposition in the future as studies examine multiple tissues from a single donor. This is an exciting time for transposable element biology given our improving ability to explore entire genomes. In this case, whole-genome paired-end sequencing has allowed us to detect somatic retrotransposition in cancer genomes, an observation that opens up many new questions regarding the role of mobile DNA in carcinogenesis and tumor molecular biology. As sequencing technologies and our ability to detect structural variants improve, so will our ability to characterize new TE insertions and their parent elements, perhaps gaining further insight into what leads to tissue or disease specific TE activation.

Chapter 6

Conclusions and future directions

I created two new data analysis pipelines for analyzing high-throughput whole genome sequencing datasets with the goal of enabling researchers to uncover tumor histories and causes. The discord-retro pipeline detects retrotransposon insertions by finding read pairs where one end maps to an annotated retrotransposon element and the other maps to a distant unique location in the genome. I applied this method to a 86 TCGA patients and seven cancer types and found that colon, breast, and lung cancers showed evidence of somatic retrotransposition, while acute myeloid leukemia, ovarian cancer and brain cancers did not. This type of analysis has been applied to an even larger TCGA cohort by Helman et al. [71], and it should be incorporated as part of all genomic analysis pipelines in order to give a more complete picture of a patients genome. It is not enough to detect only SNVs, indels, and copy number changes when we know other classes of mutations can cause disease, too.

Additionally, I developed a pipeline for analyzing copy number changes and novel adjacencies using the CN-AVG method, a framework invented by Drs. Zerbino, Paten, and Haussler [173]. This pipeline uses the CN-AVG method to find series of structural rearrangements, or potential evolutionary histories, that give rise to the mutated cancer genome configuration from the germ line genome of a patient. I tested the method on artificial evolutionary histories generated in silico to calculate its accuracy and applied the method to a cohort of 16 glioblastoma patients from TCGA.

I built up the CN-AVG data analysis pipeline so that other researchers may use it to learn about cancer evolution and uncover testable biological hypothesis. For example, the importance of mutational order for certain cancer types can be tested directly using mouse models in which mice are genetically engineered to acquire mutations at specific time points. Zhu et al. used such a model to discover that astrocytomas form only when p53 loss comes before or concurrently with NF1 loss [174]. As demonstrated in Section 4.1, the CN-AVG pipeline can be used to gain insight to individual patients' tumors and therefore would be a useful tool for clinicians or gene therapists. Additionally, I created ways for combining CN-AVG across patients, so the pipeline can be used by researchers studying large patient cohorts as well. My thesis work is the first step towards making the CN-AVG method available and useful to the medical research community.

In the future, I would like to apply the CN-AVG method to whole genome sequencing data taken at multiple time points from the same patient and test whether events that appear early in the CN-AVG predicted history of later time points also appear in the early time points. This could provide further evidence that the CN-

AVG method is able to accurately reconstruct the order of structural rearrangements. Conversely, such an experiment may show that cancer develops in unexpected ways. For example, a rapidly mutating tumor may exhibit genomes at the later time point that are too diverged from the first sampling to predict the early events accurately using CN-AVG. Currently, chronological whole genome sequencing experiments are not widely available and may be difficult to generate. Studies that do sample tumors multiple times focus on a small set of mutations rather than whole genome sequencing, presumably due to limited amounts of tissue. Furthermore, it would be unreasonable to collect multiple samples from patients at controlled time points for research purposes because nearly all tumors are inaccessible without surgery, which is typically done only once or a few times only at the patient's benefit for obvious reasons. Although sampling blood cancers is a less invasive procedure and could potentially be done in a time series, these cancers tend to have very few genomic rearrangments or mutations, making them poor models for studying genome evolution.

Multiple samples gathered at a single time point but from different regions of the tumor or from different metastatic lesions may also be used to study cancer evolution, as discussed in Chapter 2. Barrett oesophagus, a premalignant condition that leads to oesophageal adenocarcinoma, is an ideal candidate to study in this way because the diseased tissue spreads gradually upward in the oesophagus as it progresses, creating a spatially segregated evolutionary path [7, 100]. Although TCGA has not collected sequential whole genome sequencing (WGS) data for oesophageal adenocarcinoma, it has obtained WGS samples of metastatic and primary tumors for ten GBM patients.

For future work I will run these data through the CN-AVG pipeline, using the metastatic as the later time point and the primary tumor as the early time point.

Researchers study the evolution of cancer not only out of desire to better understand the disease and improve patient treatment, but because cancer is a great model for genome evolution. Not only is there a plethora of sequencing data available for the disease, but tumors evolve on timescales shorter than human lifespans, allowing the potential to study their evolution in real time. However, the CN-AVG pipeline could be used to study genomic evolution in contexts outside of cancer, as well. It could be used to study evolution of species by predicting series of structural rearrangements between distant species. Additionally, since the method accounts for multi-clonality in a sample, it can be used in the study of mixed populations of microscopic organisms, or metagenomics. Without a reference genome for the whole population, one could still detect copy number changes in contigs within the population and model the rearrangements of these contigs with CN-AVG. This may help track how resistance to antibacterial drugs develops over time or how the microbial ecosystems of the soil or oceans change in response to environmental perturbations.

Currently, CN-AVG only models evolution of structural rearrangements, leaving out single nucleotide variants (SNVs) and indels, nor does it take biological biases into account when modeling genome evolution. For example, due to sequence composition or chromatin structure, breakpoint reuse may be favored in evolution, and this could be incorporated into the model by lowering the cost function for rearrangements that use breakpoints from prior events. SNVs and indels could be incorporated into the

CN-AVG framework by modeling them as separate sequence paths through the graph, with copy numbers assigned according to the allelic fraction of each SNV. This strategy may have a drawback of generateing highly complex graph structures and no consensus evolutionary path. Alternatively, the evolutionary history of these types of SNVs could be estimated using allele frequency estimates, as described in Section 2.2. The SNV information could then be easily overlayed with the history of structural rearrangements and both mutations types could be further analyzed together.

Yet another strategy to improve the CN-AVG method using SNV data would be to phase genomes prior to application of the CN-AVG method, providing 46 copy number profiles as input rather than 23, and setting the default copy number as one rather than two. Phasing is difficult to do with short read sequencing data because SNVs cannot be linked unless they coexist on the same reads, in which case they must be within a few hundred bases of each other. Incorporating single nucleotide polymorphism (SNP) linkage information from population studies can aid phasing by increasing the number of linked SNVs. For example, if two somatic variants co-occur with SNPs belonging to the same haplotype and hence on the same DNA strand, it is likely that the two somatic variants are likewise on the same strand. Although potentially allowing for a more accurate picture of a genome, SNP information may be cumbersome to generate and incorporate into an analysis pipeline. It is more likely that improved sequencing technology will yield longer reads, making phased genomic sequence the norm. A more accurate and detailed snapshot of a tumor genome should improve the ability of CN-AVG to reconstruct its history.

This work describes a way to predict the chronological order of specific types of mutations, genomic rearrangements, but understanding a tumor's genetic evolutionary history will only be a small piece of the enormous cancer puzzle. Although, as stated in Chapter 1, cancer is currently understood as a disease of the genome, to understand it only as a disease of the genome would be an over-simplification and a detriment to finding a cure. We already know there are numerous environmental factors known to cause the disease, including infectious agents and chemical carcinogens [162], and probably just as many factors remain to be discovered. We also know that epigenetics can play a role in the disease [33], and metabolites have recently been proposed to play a role in prostate cancer [143] with metabolic changes in general hypothesized to lead to cancer [20]. Therefore, the complete mutational landscape and history of a tumor and the precise ordering of every single SNV and structural rearrangement, will not necessarily explain why the tumor arose in the first place or even how to treat it in the long run. We may determine which mutations occurred first and are therefore probably driving the cancer, but this will not explain why or how those first mutations arose. It will be important in future efforts to cure cancer to recognize the limitations of genetic studies in a time when cancer is frequently described as a genetic disease.

Nevertheless, while genetic mutations are not the complete picture of a diseased cell, they play a key role in disease and can provide clues to things happening outside of the genome or outside the cell as well. With genome sequencing becoming common place, future disease treatment will monitor a tumor genome in real time, allowing its current unique mutational combination to be precisely targeted by drug combinations.

Finding evolutionary patterns will give further advantage by allowing us to anticipate what the next mutational step will be, giving doctors a chance to cut off cancer regrowth at the outset. The CN-AVG and discord-retro methods will assist in finding these evolutionary patterns, but it will be important in future evolutionary studies of cancer to remember that cancer is not simply a genetic disease, but is heavily influenced by the microenvironment within and without the cell. Cancer development should be viewed as the evolution of ecological environments as well as of genomes [103].

Bibliography

[1] B N Ames, W E Durston, E Yamasaki, and F D Lee. Carcinogens are mutagens: a simple test system combining liver homogenates for activation and bacteria for detection. *Proceedings of the National Academy of Sciences*, 70(8):2281–2285, August 1973.

[2] Alexei A Aravin, Gregory J Hannon, and Julius Brennecke. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science (New York, N.Y.)*, 318(5851):761–4, 2007.

[3] 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):666–677, April 2013.

[4] Martin Bader. Sorting by reversals, block interchanges, tandem duplications, and deletions. *BMC Bioinformatics*, 10(Suppl 1):S9, January 2009.

[5] J Kenneth Baillie, Mark W Barnett, Kyle R Upton, Daniel J Gerhardt, Todd A Richmond, Fioravante De Sapio, Paul Brennan, Patrizia Rizzu, Sarah Smith, Mark Fell, Richard T Talbot, Stefano Gustincich, Thomas C Freeman, John S Mattick, David A Hume, Peter Heutink, Piero Carninci, Jeffrey A Jeddeloh, and Geoffrey J Faulkner. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, pages 1–4, October 2011.

[6] P E Barker. Double minutes in human tumor cells. *Cancer genetics and cytogenetics*, 5(1):81–94, February 1982.

[7] M T Barrett, C A Sanchez, L J Prevo, D J Wong, P C Galipeau, T G Paulson, P S Rabinovitch, and B J Reid. Evolution of neoplastic cell lineages in Barrett oesophagus. *Nature Genetics*, 22(1):106–109, May 1999.

[8] C R Bartram, A de Klein, A Hagemeijer, T van Agthoven, A Geurts van Kessel, D Bootsma, G Grosveld, M A Ferguson-Smith, T Davies, and M Stone. Transloca-

tion of c-ab1 oncogene correlates with the presence of a Philadelphia chromosome in chronic myelocytic leukaemia. *Nature*, 306(5940):277–280, November 1983.

[9] Christine R Beck, Pamela Collier, Catriona Macfarlane, Maika Malig, Jeffrey M Kidd, Evan E Eichler, Richard M Badge, and John V Moran. Line-1 retrotransposition activity in human genomes. *Cell*, 141(7):1159–1170, Jun 2010.

[10] V P Belancio, A M Roy-Engel, and P Deininger. The impact of multiple splice sites in human L1 elements. *Gene*, 411(1-2):38–45, March 2008.

[11] Rameen Beroukhim, Gad Getz, Leia Nghiemphu, Jordi Barretina, Teli Hsueh, David Linhart, Igor Vivanco, Jeffrey C Lee, Julie H Huang, Sethu Alexander, Jinyan Du, Tweeny Kau, Roman K Thomas, Kinjal Shah, Horacio Soto, Sven Perner, John Prensner, Ralph M Debiasi, Francesca Demichelis, Charlie Hatton, Mark A Rubin, Levi A Garraway, Stan F Nelson, Linda Liau, Paul S Mischel, Tim F Cloughesy, Matthew Meyerson, Todd A Golub, Eric S Lander, Ingo K Mellinghoff, and William R Sellers. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences*, 104(50):20007–20012, December 2007.

[12] Graham R Bignell, Thomas Santarius, Jessica C M Pole, Adam P Butler, Janet Perry, Erin Pleasance, Chris Greenman, Andrew Menzies, Sheila Taylor, Sarah Edkins, Peter Campbell, Michael Quail, Bob Plumb, Lucy Matthews, Kirsten McLay, Paul A W Edwards, Jane Rogers, Richard Wooster, P Andrew Futreal, and Michael R Stratton. Architectures of somatic genomic rearrangement in hu-

man cancer amplicons at sequence-level resolution. *Genome Research*, 17(9):1296–1303, September 2007.

[13] J D Boeke, D J Garfinkel, C A Styles, and G R Fink. Ty elements transpose through an rna intermediate. *Cell*, 40(3):491–500, Mar 1985.

[14] Déborah Bourc'his and Timothy H Bestor. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking dnmt3l. *Nature*, 431(7004):96–9, Sep 2004.

[15] Guillaume Bourque, Bernard Leong, Vinsensius B Vega, Xi Chen, Yen Ling Lee, Kandhadayar G Srinivasan, Joon-Lin Chew, Yijun Ruan, Chia-Lin Wei, Huck Hui Ng, and Edison T Liu. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res*, 18(11):1752–62, Nov 2008.

[16] Marília D V Braga, Eyla Willing, and Jens Stoye. Double Cut and Join with Insertions and Deletions. *Journal of Computational Biology*, 18(9):1167–1184, September 2011.

[17] Cameron W Brennan, Roel G W Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie R Salama, Siyuan Zheng, Debyani Chakravarty, J Zachary Sanborn, Samuel H Berman, Rameen Beroukhim, Brady Bernard, Chang-Jiun Wu, Giannicola Genovese, Ilya Shmulevich, Jill Barnholtz-Sloan, Lihua Zou, Rahulsimham Vegesna, Sachet A Shukla, Giovanni Ciriello, W K Yung, Wei Zhang, Carrie Sougnez, Tom Mikkelsen, Kenneth Aldape, Darell D

Bigner, Erwin G Van Meir, Michael Prados, Andrew Sloan, Keith L Black, Jennifer Eschbacher, Gaetano Finocchiaro, William Friedman, David W Andrews, Abhijit Guha, Mary Iacocca, Brian P O'Neill, Greg Foltz, Jerome Myers, Daniel J Weisenberger, Robert Penny, Raju Kucherlapati, Charles M Perou, D Neil Hayes, Richard Gibbs, Marco Marra, Gordon B Mills, Eric Lander, Paul Spellman, Richard Wilson, Chris Sander, John Weinstein, Matthew Meyerson, Stacey Gabriel, Peter W Laird, David Haussler, Gad Getz, and Lynda Chin. The Somatic Genomic Landscape of Glioblastoma. *Cell*, 155(2):462–477, October 2013.

[18] Brook Brouha, Joshua Schustak, Richard M Badge, Sheila Lutz-Prigge, Alexander H Farley, John V Moran, and Haig H Kazazian. Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America*, 100(9):5280–5, April 2003.

[19] Talha Khan Burki. Stand Up To Cancer. *The Lancet Oncology*, 13(12):1197–1198, December 2012.

[20] Rob A Cairns, Isaac S Harris, and Tak W Mak. Regulation of cancer cell metabolism. *Nature Reviews Cancer*, 11(2):85–95, February 2011.

[21] Peter J Campbell, Erin D Pleasance, Philip J Stephens, Ed Dicks, Richard Rance, Ian Goodhead, George A Follows, Anthony R Green, P Andy Futreal, and Michael R Stratton. Subclonal phylogenetic structures in cancer revealed by

ultra-deep sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 105(35):13081–13086, September 2008.

[22] Peter J Campbell, Shinichi Yachida, Laura J Mudie, Philip J Stephens, Erin D Pleasance, Lucy A Stebbings, Laura A Morsberger, Calli Latimer, Stuart McLaren, Meng-Lay Lin, David J McBride, Ignacio Varela, Serena A Nik-Zainal, Catherine Leroy, Mingming Jia, Andrew Menzies, Adam P Butler, Jon W Teague, Constance A Griffin, John Burton, Harold Swerdlow, Michael A Quail, Michael R Stratton, Christine Iacobuzio-Donahue, and P Andrew Futreal. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319):1109–1113, October 2010.

[23] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, October 2008.

[24] Hui Chen, Caroline E Lilley, Qin Yu, Darwin V Lee, Jody Chou, Iñigo Narvaiza, Nathaniel R Landau, and Matthew D Weitzman. APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. *Current Biology : CB*, 16(5):480–5, March 2006.

[25] 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 Mey-

erson, Scott L Carter, and Gad Getz. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421, April 2012.

[26] Lara S Collier, Corey M Carlson, Shruthi Ravimohan, Adam J Dupuy, and David A Largaespada. Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature*, 436(7048):272–6, Jul 2005.

[27] Francis S Collins and Anna D Barker. Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Scientific American*, 296(3):50–57, March 2007.

[28] Neal G Copeland and Nancy A Jenkins. Deciphering the genetic landscape of cancer – from genes to pathways. *Trends in Genetics*, 25(10):455–462, October 2009.

[29] Nicole G Coufal, José L Garcia-Perez, Grace E Peng, Gene W Yeo, Yangling Mu, Michael T Lovci, Maria Morell, K Sue O'Shea, John V Moran, and Fred H Gage. L1 retrotransposition in human neural progenitor cells. *Nature*, 460(7259):1127–31, Aug 2009.

[30] Steven W Criscione, Yue Zhang, William Thompson, John M Sedivy, and Nicola Neretti. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics*, 15(1):583, 2014.

[31] F T Cutts, S Franceschi, S Goldie, X Castellsague, S de Sanjose, G Garnett, W J Edmunds, P Claeys, K L Goldenthal, D M Harper, and L Markowitz. Hu-

man papillomavirus and HPV vaccines: a review. *Bulletin of the World Health Organization*, 85(9):719–726, September 2007.

[32] Frederik Damm, Birgit Markus, Felicitas Thol, Michael Morgan, Gudrun Göhring, Brigitte Schlegelberger, Jürgen Krauter, Michael Heuser, Olivier A Bernard, and Arnold Ganser. TET2 mutations in cytogenetically normal acute myeloid leukemia: Clinical implications and evolutionary patterns. *Genes, Chromosomes and Cancer*, pages n/a–n/a, June 2014.

[33] Mark A Dawson and Tony Kouzarides. Cancer epigenetics: from mechanism to therapy. *Cell*, 150(1):12–27, July 2012.

[34] A P Jason de Koning, Wanjun Gu, Todd A Castoe, Mark A Batzer, and David D Pollock. Repetitive elements may comprise over two-thirds of the human genome. *PLoS genetics*, 7(12):e1002384, December 2011.

[35] Marie Dewannieux, Cécile Esnault, and Thierry Heidmann. LINE-mediated retrotransposition of marked Alu sequences. *Nature genetics*, 35(1):41–8, October 2003.

[36] D Dickson. Wellcome funds cancer database. *Nature*, 401(6755):729–729, October 1999.

[37] L Ding, T J Ley, D E Larson, C A Miller, and D C Koboldt. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481:506–510, January 2012.

[38] Li Ding, Matthew J Ellis, Shunqiang Li, David E Larson, Ken Chen, John W

Wallis, Christopher C Harris, Michael D McLellan, Robert S Fulton, Lucinda L Fulton, Rachel M Abbott, Jeremy Hoog, David J Dooling, Daniel C Koboldt, Heather Schmidt, Joelle Kalicki, Qunyuan Zhang, Lei Chen, Ling Lin, Michael C Wendl, Joshua F McMichael, Vincent J Magrini, Lisa Cook, Sean D McGrath, Tammi L Vickery, Elizabeth Appelbaum, Katherine DeSchryver, Sherri Davies, Therese Guintoli, Li Lin, Robert Crowder, Yu Tao, Jacqueline E Snider, Scott M Smith, Adam F Dukes, Gabriel E Sanderson, Craig S Pohl, Kim D Delehaunty, Catrina C Fronick, Kimberley A Pape, Jerry S Reed, Jody S Robinson, Jennifer S Hodges, William Schierding, Nathan D Dees, Dong Shen, Devin P Locke, Madeline E Wiechert, James M Eldred, Josh B Peck, Benjamin J Oberkfell, Justin T Lolofie, Feiyu Du, Amy E Hawkins, Michelle D O'Laughlin, Kelly E Bernard, Mark Cunningham, Glendoria Elliott, Mark D Mason, Dominic M Thompson Jr, Jennifer L Ivanovich, Paul J Goodfellow, Charles M Perou, George M Weinstock, Rebecca Aft, Mark Watson, Timothy J Ley, Richard K Wilson, and Elaine R Mardis. Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, 464(7291):999–1005, April 2010.

[39] Brian J Druker. Imatinib and chronic myeloid leukemia: validating the promise of molecularly targeted therapy. *European journal of cancer (Oxford, England : 1990)*, 38 Suppl 5:S70–6, September 2002.

[40] Adam J Dupuy, Keiko Akagi, David A Largaespada, Neal G Copeland, and Nancy A Jenkins. Mammalian mutagenesis using a highly mobile somatic sleeping

beauty transposon system. *Nature*, 436(7048):221–6, Jul 2005.

[41] Steffen Durinck, Christine Ho, Nicholas J Wang, Wilson Liao, Lakshmi R Jakkula, Eric A Collisson, Jennifer Pons, Sai-Wing Chan, Ernest T Lam, Catherine Chu, Kyunghee Park, Sung-woo Hong, Joe S Hur, Nam Huh, Isaac M Neuhaus, Siegrid S Yu, Roy C Grekin, Theodora M Mauro, James E Cleaver, Pui-Yan Kwok, Philip E LeBoit, Gad Getz, Kristian Cibulskis, Jon C Aster, Haiyan Huang, Elizabeth Purdom, Jian Li, Lars Bolund, Sarah T Arron, Joe W Gray, Paul T Spellman, and Raymond J Cho. Temporal Dissection of Tumorigenesis in Primary Cancers. *Cancer Discovery*, 1(2):137–143, July 2011.

[42] Michael T Eickbush and Thomas H Eickbush. Retrotransposition of R2 elements in somatic nuclei during the early development of Drosophila. *Mobile DNA*, 2(1):11, September 2011.

[43] C Esnault, J Maestre, and T Heidmann. Human LINE retrotransposons generate processed pseudogenes. *Nature Genetics*, 24(4):363–7, May 2000.

[44] M Esteller, J Garcia-Foncillas, E Andion, S N Goodman, O F Hidalgo, V Vanaclocha, S B Baylin, and J G Herman. Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *The New England journal of medicine*, 343(19):1350–1354, November 2000.

[45] A. D Ewing and H. H Kazazian. High-throughput sequencing reveals extensive

variation in human-specific l1 content in individual human genomes. *Genome Res*, 20(9):1262–1270, Sep 2010.

[46] A D Ewing and H H Kazazian. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Research*, 20(9):1262–1270, September 2010.

[47] A D Ewing and H H Kazazian. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Research*, 21(6):985–990, June 2011.

[48] Anna-Sophie Fiston-Lavier, Matthew Carrigan, Dmitri A Petrov, and Josefa González. T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. *Nucleic Acids Research*, 39(6):e36, March 2011.

[49] Simon A Forbes, Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Kok, David Beare, Mingming Jia, Rebecca Shepherd, Kenric Leung, Andrew Menzies, Jon W Teague, Peter J Campbell, Michael R Stratton, and P Andrew Futreal. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 39(Database issue):D945–50, Jan 2011.

[50] Josep V Forment, Abderrahmane Kaidi, and Stephen P Jackson. Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nature Reviews Cancer*, 12(10):663–670, October 2012.

[51] Veronique Frattini, Vladimir Trifonov, Joseph Minhow Chan, Angelica Castano, Marie Lia, Francesco Abate, Stephen T Keir, Alan X Ji, Pietro Zoppoli, Francesco Niola, Carla Danussi, Igor Dolgalev, Paola Porrati, Serena Pellegatta, Adriana Heguy, Gaurav Gupta, David J Pisapia, Peter Canoll, Jeffrey N Bruce, Roger E McLendon, Hai Yan, Ken Aldape, Gaetano Finocchiaro, Tom Mikkelsen, Gilbert G Privé, Darell D Bigner, Anna Lasorella, Raul Rabadan, and Antonio Iavarone. The integrated landscape of driver genomic alterations in glioblastoma. *Nature Genetics*, 45(10):1141–1149, October 2013.

[52] Frank B Furnari, Tim Fenton, Robert M Bachoo, Akitake Mukasa, Jayne M Stommel, Alexander Stegh, William C Hahn, Keith L Ligon, David N Louis, Cameron Brennan, Lynda Chin, Ronald A DePinho, and Webster K Cavenee. Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes and Development*, 21(21):2683–2710, November 2007.

[53] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–83, Mar 2004.

[54] Dale W Garsed, Owen J Marshall, Vincent D A Corbin, Arthur Hsu, Leon Di Stefano, Jan Schröder, Jason Li, Zhi-Ping Feng, Bo W Kim, Mark Kowarsky, Ben Lansdell, Ross Brookwell, Ola Myklebost, Leonardo Meza-Zepeda, Andrew J Holloway, Florence Pedeutour, K H Andy Choo, Michael A Damore, Andrew J Deans,

Anthony T Papenfuss, and David M Thomas. The Architecture and Evolution of Cancer Neochromosomes. *Cancer Cell*, 26(5):653–667, November 2014.

[55] D Gisselsson, L Pettersson, M Höglund, M Heidenblad, L Gorunova, J Wiegant, F Mertens, P Dal Cin, F Mitelman, and N Mandahl. Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity. *Proceedings of the National Academy of Sciences*, 97(10):5357–5362, May 2000.

[56] Mary Goldman, Brian Craft, Teresa Swatloski, Melissa Cline, Olena Morozova, Mark Diekhans, David Haussler, and Jingchun Zhu. The UCSC Cancer Genomics Browser: update 2015. *Nucleic Acids Research*, 43(Database issue):D812–7, January 2015.

[57] 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 Research*, 22(2):346–361, February 2012.

[58] Steven I Hajdu. A note from history: landmarks in history of cancer, part 1. *Cancer*, 117(5):1097–1102, March 2011.

[59] Steven I Hajdu. A note from history: landmarks in history of cancer, part 2. *Cancer*, 117(12):2811–2820, June 2011.

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

[61] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, March 2011.

[62] D. C. Hancks, J. L. Goodier, P. K. Mandal, L. E. Cheung, and H. H. Kazazian. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Human Molecular Genetics*, 20(17):3386–3400, June 2011.

[63] Dustin C Hancks, Adam D Ewing, Jesse E Chen, Katsushi Tokunaga, and Haig H Kazazian. Exon-trapping mediated by the human retrotransposon SVA. *Genome Research*, 19(11):1983–91, November 2009.

[64] Dustin C Hancks and Haig H Kazazian. Active human retrotransposons: variation and disease. *Current Opinion in Genetics and Development*, 22(3):191–203, March 2012.

[65] Dustin C Hancks and Haig H Kazazian Jr. Active human retrotransposons: variation and disease. *Current Opinion in Genetics & Development*, pages 1–13, March 2012.

[66] S Hannenhalli and P A Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *IEEE 36th Annual Foundations of Computer Science*, pages 581–592, Milwaukee, WI, October 1995. IEEE Comput. Soc. Press.

[67] Sridhar Hannenhalli and Pavel A Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM*, 46(1):1–27, January 1999.

[68] Kasper Daniel Hansen, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, Eirikur Briem, Kun Zhang, Rafael A Irizarry, and Andrew P Feinberg. Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8):768–75, Aug 2011.

[69] C R Harris, A Dewan, A Zupnick, R Normart, A Gabriel, C Prives, A J Levine, and J Hoh. p53 responsive elements in human retrotransposons. *Oncogene*, 28(44):3857–65, November 2009.

[70] Monika E Hegi, Annie-Claire Diserens, Thierry Gorlia, Marie-France Hamou, Nicolas de Tribolet, Michael Weller, Johan M Kros, Johannes A Hainfellner, Warren Mason, Luigi Mariani, Jacoline E C Bromberg, Peter Hau, René O Mirimanoff, J Gregory Cairncross, Robert C Janzer, and Roger Stupp. MGMT gene silencing and benefit from temozolomide in glioblastoma. *The New England Journal of Medicine*, 352(10):997–1003, March 2005.

[71] Elena Helman, Michael S Lawrence, Chip Stewart, Carrie Sougnez, Gad Getz, and Matthew Meyerson. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Research*, 24(7):1053–1063, July 2014.

[72] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max D M Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, Jiashan Zhang, Cyriac Kandoth, Rehan Akbani, Hui Shen, Larsson Omberg, Andy Chu, Adam A Margolin, Laura J Van't Veer, Nuria López-Bigas, Peter W Laird, Benjamin J Raphael, Li Ding, A Gordon Robertson, Lauren A Byers, Gordon B Mills, John N Weinstein, Carter Van Waes, Zhong Chen, Eric A Collisson, Cancer Genome Atlas Research Network, Christopher C Benz, Charles M Perou, and Joshua M Stuart. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, August 2014.

[73] F Hormozdiari, C Alkan, M Ventura, I Hajirasouliha, M Malig, F Hach, D Yorukoglu, P Dao, M Bakhshi, S. C Sahinalp, and E. E Eichler. Alu repeat discovery and characterization within human genomes. *Genome Research*, 21(6):1–36, Dec 2010.

[74] F Hormozdiari, I Hajirasouliha, P Dao, F Hach, D Yorukoglu, C Alkan, E. E Eichler, and S. C Sahinalp. Next-generation variationhunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, 26(12):i350–i357, Jun 2010.

[75] 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):873–885, March 2012.

[76] Cheng Ran Lisa Huang, Anna M Schneider, Yunqi Lu, Tejasvi Niranjan, Peilin Shen, Matoya A Robinson, Jared P Steranka, David Valle, Curt I Civin, Tao Wang, Sarah J Wheelan, Hongkai Ji, Jef D Boeke, and Kathleen H Burns. Mobile interspersed repeats are major structural variants in the human genome. *Cell*, 141(7):1171–1182, Jun 2010.

[77] 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, 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, 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, Mark Lathrop, Nuria López-Bigas, B F Francis Ouellette, Gilles Thomas, Alfonso Valencia, Teruhiko Yoshida, Karen L Kennedy, Myles Axton, Stephanie O M Dyke, Daniela S Gerhard, Chris Gunter, Mark Guyer, Thomas J Hudson, John D McPherson, Linda J Miller, Brad Ozenberger, Kenna M Shaw, 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, 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, Ronald A DePinho, Sarah Thayer, Kamran Shazand, Timothy Beck, Michelle Sam, Lee Timms, Vanessa Ballin, Youyong Lu, Jiafu Ji, Xiuqing Zhang, Feng Chen, Xueda Hu, Qi Yang, Geng Tian, Lianhai Zhang, Xiaofang Xing, Xianghong Li, Zhenggang Zhu, Yingyan Yu, Jun Yu, 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, Ewan Birney, Åke Borg, Anne-Lise Børresen-Dale, and ... Caldas. International network of cancer genome projects. *Nature*, 464(7291):993–998, April 2010.

[78] Rebecca C Iskow, Michael T Mccabe, Ryan E Mills, Spencer Torene, W Stephen Pittard, Andrew F Neuwald, Erwin G Van Meir, Paula M Vertino, and Scott E

Devine. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*, 141(7):1253–1261, Jun 2010.

[79] George F Jenks and Fred C. Caspall. Error on Choroplethic Maps: Definition, Measurement, Reduction. *Annals of the Association of American Geographers*, 61(2):217–244, March 2010.

[80] Hiroki Kano, Irene Godoy, Christine Courtney, Melissa R Vetter, George L Gerton, Eric M Ostertag, and Haig H Kazazian. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes & development*, 23(11):1303–12, June 2009.

[81] R J Kaufman, P C Brown, and R T Schimke. Amplified dihydrofolate reductase genes in unstably methotrexate-resistant cells are associated with double minute chromosomes. *Proceedings of the National Academy of Sciences*, 76(11):5669–5673, November 1979.

[82] H H Kazazian, C Wong, H Youssoufian, A F Scott, D G Phillips, and S E Antonarakis. Haemophilia a resulting from de novo insertion of l1 sequences represents a novel mechanism for mutation in man. *Nature*, 332(6160):164–6, Mar 1988.

[83] Hameed Khan, Arian Smit, and Stéphane Boissinot. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome research*, 16(1):78–87, January 2006.

[84] Masanobu Kinomoto, Takayuki Kanno, Mari Shimura, Yukihito Ishizaka, Asato Kojima, Takeshi Kurata, Tetsutaro Sata, and Kenzo Tokunaga. All APOBEC3 family proteins differentially inhibit LINE-1 retrotransposition. *Nucleic acids research*, 35(9):2955–64, January 2007.

[85] A G Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, April 1971.

[86] Tomoko Kuwabara, Jenny Hsieh, Alysson Muotri, Gene Yeo, Masaki Warashina, Dieter Chichung Lie, Lynne Moore, Kinichi Nakashima, Makoto Asashima, and Fred H Gage. Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. *Nature neuroscience*, 12(9):1097–105, 2009.

[87] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T

Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowski, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[88] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, January 2014.

[89] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, Adam Kiezun, Peter S Hammerman, Aaron McKenna, Yotam Drier, Lihua Zou, Alex H Ramos, Trevor J Pugh, Nicolas Stransky, Elena Helman, Jaegil Kim, Carrie Sougnez, Lauren Ambrogio, Elizabeth Nickerson, Erica Shefler, Maria L Cortés, Daniel Auclair, Gordon Saksena, Douglas Voet, Michael Noble, Daniel DiCara, Pei Lin, Lee Lichtenstein, David I Heiman, Timothy Fennell, Marcin Imielinski, Bryan Hernandez, Eran Hodis, Sylvan Baca, Austin M Dulak, Jens Lohr, Dan-Avi Landau, Catherine J Wu, Jorge Melendez-Zajgla, Al-

fredo Hidalgo-Miranda, Amnon Koren, Steven A McCarroll, Jaume Mora, Ryan S Lee, Brian Crompton, Robert Onofrio, Melissa Parkin, Wendy Winckler, Kristin Ardlie, Stacey B Gabriel, Charles W M Roberts, Jaclyn A Biegel, Kimberly Stegmaier, Adam J Bass, Levi A Garraway, Matthew Meyerson, Todd R Golub, Dmitry A Gordenin, Shamil Sunyaev, Eric S Lander, and Gad Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, July 2013.

[90] E Lee, R Iskow, L Yang, O Gokcumen, P Haseley, L J Luquette, J G Lohr, C C Harris, L Ding, R K Wilson, D A Wheeler, R A Gibbs, R Kucherlapati, C Lee, P V Kharchenko, P J Park, and The Cancer Genome Atlas Research Network. Landscape of Somatic Retrotransposition in Human Cancers. *Science*, 337(6097):967–971, August 2012.

[91] Eunjung Lee, Rebecca Iskow, Lixing Yang, Omer Gokcumen, Psalm Haseley, Lovelace J Luquette, Jens G Lohr, Christopher C Harris, Li Ding, Richard K Wilson, David A Wheeler, Richard A Gibbs, Raju Kucherlapati, Charles Lee, Peter V Kharchenko, and Peter J Park. Landscape of Somatic Retrotransposition in Human Cancers. *Science (New York, N.Y.)*, June 2012.

[92] Mark D M Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A Ryslik, Nuria López-Bigas, Gad Getz, Li Ding, and

Benjamin J Raphael. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2):106–114, February 2015.

[93] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, Jul 2009.

[94] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, August 2009.

[95] T Liehr, K Mrasek, A Weise, A Dufke, L Rodríguez, N Martínez Guardia, A Sanchís, J R Vermeesch, C Ramel, A Polityko, O A Haas, J Anderson, U Claussen, F von Eggeling, and H Starke. Small supernumerary marker chromosomes–progress towards a genotype-phenotype correlation. *Cytogenetic and Genome Research*, 112(1-2):23–34, 2006.

[96] Anthony W I Lo, Laure Sabatier, Bijan Fouladi, Géraldine Pottier, Michelle Ricoul, and John P Murnane. DNA amplification by breakage/fusion/bridge cycles initiated by spontaneous telomere loss in a human cancer cell line. *Neoplasia*, 4(6):531–538, November 2002.

[97] D D Luan, M H Korman, J L Jakubczak, and T H Eickbush. Reverse transcription

of r2bm rna is primed by a nick at the chromosomal target site: a mechanism for non-ltr retrotransposition. *Cell*, 72(4):595–605, Feb 1993.

[98] Patrícia M Machado, Rita D Brandão, Branca M Cavaco, Joana Eugénio, Sandra Bento, Mónica Nave, Paula Rodrigues, Aires Fernandes, and Fátima Vaz. Screening for a brca2 rearrangement in high-risk breast/ovarian cancer families: evidence for a founder effect and analysis of the associated phenotypes. *Journal of Clinical Oncology*, 25(15):2027–34, May 2007.

[99] Norra MacReady. Stand Up 2 Cancer. *The Lancet Oncology*, 11(11):1027–1028, November 2010.

[100] Carlo C Maley and Brian J Reid. Natural selection in neoplastic progression of Barrett's esophagus. *Seminars in Cancer Biology*, 15(6):474–483, December 2005.

[101] B McClintock. The Stability of Broken Ends of Chromosomes in *Zea Mays*. *Genetics*, 26(2):234–282, March 1941.

[102] Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G Van Meir, Daniel J Brat, Gena M Mastrogianakis, Jeffrey J Olson, Tom Mikkelsen, Norman Lehman, Ken Aldape, W K Alfred Yung, Oliver Bogler, Scott VandenBerg, Mitchel Berger, Michael Prados, Donna Muzny, Margaret Morgan, Steve Scherer, Aniko Sabo, Lynn Nazareth, Lora Lewis, Otis Hall, Yiming Zhu, Yanru Ren, Omar Alvi, Jiqiang Yao, Alicia Hawes, Shalini Jhangiani, Gerald Fowler, Anthony San Lucas, Christie Kovar, Andrew Cree, Huyen Dinh, Jireh Santibanez, Vandita

Joshi, Manuel L Gonzalez-Garay, Christopher A Miller, Aleksandar Milosavljevic, Larry Donehower, David A Wheeler, Richard A Gibbs, Kristian Cibulskis, Carrie Sougnez, Tim Fennell, Scott Mahan, Jane Wilkinson, Liuda Ziaugra, Robert Onofrio, Toby Bloom, Rob Nicol, Kristin Ardlie, Jennifer Baldwin, Stacey Gabriel, Eric S Lander, Li Ding, Robert S Fulton, Michael D McLellan, John Wallis, David E Larson, Xiaoqi Shi, Rachel Abbott, Lucinda Fulton, Ken Chen, Daniel C Koboldt, Michael C Wendl, Rick Meyer, Yuzhu Tang, Ling Lin, John R Osborne, Brian H Dunford-Shore, Tracie L Miner, Kim Delehaunty, Chris Markovic, Gary Swift, William Courtney, Craig Pohl, Scott Abbott, Amy Hawkins, Shin Leong, Carrie Haipek, Heather Schmidt, Maddy Wiechert, Tammi Vickery, Sacha Scott, David J Dooling, Asif Chinwalla, George M Weinstock, Elaine R Mardis, Richard K Wilson, Gad Getz, Wendy Winckler, Roel G W Verhaak, Michael S Lawrence, Michael O'Kelly, Jim Robinson, Gabriele Alexe, Rameen Beroukhim, Scott Carter, Derek Chiang, Josh Gould, Supriya Gupta, Josh Korn, Craig Mermel, Jill Mesirov, Stefano Monti, Huy Nguyen, Melissa Parkin, Michael Reich, Nicolas Stransky, Barbara A Weir, Levi Garraway, Todd Golub, Matthew Meyerson, Lynda Chin, Alexei Protopopov, Jianhua Zhang, Ilana Perna, Sandy Aronson, Narayan Sathiamoorthy, Georgia Ren, Jun Yao, W Ruprecht Wiedemeyer, Hyunsoo Kim, Sek Won Kong, Yonghong Xiao, Isaac S Kohane, Jon Seidman, Peter J Park, Raju Kucherlapati, Peter W Laird, Leslie Cope, James G Herman, Daniel J Weisenberger, Fei Pan, David Van Den Berg, Leander Van Neste, Joo Mi Yi, Kornel E Schuebel, Stephen B Baylin, Devin M Absher, Jun Z Li, Audrey Southwick, Shannon Brady, Amita Aggarwal, Tisha Chung, Gavin Sherlock, James D Brooks, Richard M Myers, Paul T Spellman, Elizabeth Purdom, Lakshmi R Jakkula, Anna V Lapuk, Henry Marr, Shannon Dorton, Yoon Gi Choi, Ju Han, Amrita Ray, Victoria Wang, Steffen Durinck, Mark Robinson, Nicholas J Wang, Karen Vranizan, Vivian Peng, Eric Van Name, Gerald V Fontenay, John Ngai, John G Conboy, Bahram Parvin, Heidi S Feiler, Terence P Speed, Joe W Gray, Cameron Brennan, Nicholas D Socci, Adam Olshen, Barry S Taylor, Alex Lash, Nikolaus Schultz, Boris Reva, Yevgeniy Antipin, Alexey Stukalov, Benjamin Gross, Ethan Cerami, Wei Qing Wang, Li-Xuan Qin, Venkatraman E Seshan, Liliana Villafania, Magali Cavatore, Laetitia Borsu, Agnes Viale, William Gerald, Chris Sander, Marc Ladanyi, Charles M Perou, D Neil Hayes, Michael D Topal, Katherine A Hoadley, Yuan Qi, Sai Balu, Yan Shi, Junyuan Wu, Robert Penny, Michael Bittner, Troy Shelton, Elizabeth Lenkiewicz, Scott Morris, Debbie Beasley, Sheri Sanders, Ari Kahn, Robert Sfeir, Jessica Chen, David Nassau, Larry Feng, Erin Hickey, Jinghui Zhang, John N Weinstein, Anna Barker, Daniela S Gerhard, Joseph Vockley, Carolyn Compton, Jim Vaught, Peter Fielding, Martin L Ferguson, Carl Schaefer, Subhashree Madhavan, Kenneth H Buetow, Francis Collins, Peter Good, Mark Guyer, Brad Ozenberger, Jane Peterson, and Elizabeth Thomson. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, October 2008.

[103] Lauren M F Merlo, John W Pepper, Brian J Reid, and Carlo C Maley. Cancer

as an evolutionary and ecological process. *Nature Reviews Cancer*, 6(12):924–935, December 2006.

[104] Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12(4):R41, 2011.

[105] Nicholas Metropolis and S Ulam. The Monte Carlo Method. *Journal of the American Statistical Association*, 44(247):335–341, April 2012.

[106] Isamu Nishisho Akira Horii Yasuo Miyoshi Joji Utsunomiya Kenneth W Kinzler Bert Vogelstein Yusuke Nakamura Yoshio Miki. Disruption of the apc gene by a retrotransposal insertion of l1 sequence in a colon cancer. *Cancer Research*, 52:643–645, Feb 1992.

[107] 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 Computational Biology*, 10(8):e1003665, August 2014.

[108] Heide Muckenfuss, Matthias Hamdorf, Ulrike Held, Mario Perkovic, Johannes Löwer, Klaus Cichutek, Egbert Flory, Gerald G Schumann, and Carsten Münk.

APOBEC3 proteins inhibit human LINE-1 retrotransposition. *The Journal of biological chemistry*, 281(31):22161–72, August 2006.

[109] Siddhartha Mukherjee. *The Emperor of All Maladies. A Biography of Cancer.* Simon and Schuster, New York NY, 2010.

[110] Alysson R Muotri, Vi T Chu, Maria C N Marchetto, Wei Deng, John V Moran, and Fred H Gage. Somatic mosaicism in neuronal precursor cells mediated by l1 retrotransposition. *Nature*, 435(7044):903–10, Jun 2005.

[111] Donna M. Muzny, Matthew N. Bainbridge, Kyle Chang, Huyen H. Dinh, Jennifer A. Drummond, Gerald Fowler, Christie L. Kovar, Lora R. Lewis, Margaret B. Morgan, Irene F. Newsham, Jeffrey G. Reid, Jireh Santibanez, Eve Shinbrot, Lisa R. Trevino, Yuan-Qing Wu, Min Wang, Preethi Gunaratne, Lawrence A. Donehower, Chad J. Creighton, David A. Wheeler, Richard A. Gibbs, Michael S. Lawrence, Douglas Voet, Rui Jing, Kristian Cibulskis, Andrey Sivachenko, Petar Stojanov, Aaron McKenna, Eric S. Lander, Stacey Gabriel, Gad Getz, Li Ding, Robert S. Fulton, Daniel C. Koboldt, Todd Wylie, Jason Walker, David J. Dooling, Lucinda Fulton, Kim D. Delehaunty, Catrina C. Fronick, Ryan Demeter, Elaine R. Mardis, Richard K. Wilson, Andy Chu, Hye-Jung E. Chun, Andrew J. Mungall, Erin Pleasance, A. Gordon Robertson, Dominik Stoll, Miruna Balasundaram, Inanc Birol, Yaron S. N. Butterfield, Eric Chuah, Robin J. N. Coope, Noreen Dhalla, Ranabir Guin, Carrie Hirst, Martin Hirst, Robert A. Holt, Darlene Lee, Haiyan I. Li, Michael Mayo, Richard A. Moore, Jacqueline E. Schein, Jared R.

Slobodan, Angela Tam, Nina Thiessen, Richard Varhol, Thomas Zeng, Yongjun Zhao, Steven J. M. Jones, Marco A. Marra, Adam J. Bass, Alex H. Ramos, Gordon Saksena, Andrew D. Cherniack, Stephen E. Schumacher, Barbara Tabak, Scott L. Carter, Nam H. Pho, Huy Nguyen, Robert C. Onofrio, Andrew Crenshaw, Kristin Ardlie, Rameen Beroukhim, Wendy Winckler, Matthew Meyerson, Alexei Protopopov, Juinhua Zhang, Angela Hadjipanayis, Eunjung Lee, Ruibin Xi, Lixing Yang, Xiaojia Ren, Hailei Zhang, Narayanan Sathiamoorthy, Sachet Shukla, Peng-Chieh Chen, Psalm Haseley, Yonghong Xiao, Semin Lee, Jonathan Seidman, Lynda Chin, Peter J. Park, Raju Kucherlapati, J. Todd Auman, Katherine A. Hoadley, Ying Du, Matthew D. Wilkerson, Yan Shi, Christina Liquori, Shaowu Meng, Ling Li, Yidi J. Turman, Michael D. Topal, Donghui Tan, Scot Waring, Elizabeth Buda, Jesse Walsh, Corbin D. Jones, Piotr A. Mieczkowski, Darshan Singh, Junyuan Wu, Anisha Gulabani, Peter Dolina, Tom Bodenheimer, Alan P. Hoyle, Janae V. Simons, Matthew Soloway, Lisle E. Mose, Stuart R. Jefferys, Saianand Balu, Brian D. O'Connor, Jan F. Prins, Derek Y. Chiang, D. Neil Hayes, Charles M. Perou, Toshinori Hinoue, Daniel J. Weisenberger, Dennis T. Maglinte, Fei Pan, Benjamin P. Berman, David J. Van Den Berg, Hui Shen, Timothy Triche Jr, Stephen B. Baylin, Peter W. Laird, Michael Noble, Doug Voet, Nils Gehlenborg, Daniel DiCara, Chang-Jiun Wu, Spring Yingchun Liu, Lihua Zhou, Pei Lin, Richard W. Park, Marc-Danie Nazaire, Jim Robinson, Helga Thorvaldsdottir, Jill Mesirov, Vesteinn Thorsson, Sheila M. Reynolds, Brady Bernard, Richard Kreisberg, Jake Lin, Lisa Iype, Ryan Bressler, Timo Erkkilä, Madhumati Gundapuneni, Yuexin Liu, Adam Norberg, Tom Robinson, Da Yang, Wei Zhang, Ilya Shmulevich, Jorma J. de Ronde, Nikolaus Schultz, Ethan Cerami, Giovanni Ciriello, Arthur P. Goldberg, Benjamin Gross, Anders Jacobsen, Jianjiong Gao, Bogumil Kaczkowski, Rileen Sinha, B. Arman Aksoy, Yevgeniy Antipin, Boris Reva, Ronglai Shen, Barry S. Taylor, Timothy A. Chan, Marc Ladanyi, Chris Sander, Rehan Akbani, Nianxiang Zhang, Bradley M. Broom, Tod Casasent, Anna Unruh, Chris Wakefield, Stanley R. Hamilton, R. Craig Cason, Keith A. Baggerly, John N. Weinstein, David Haussler, Christopher C. Benz, Joshua M. Stuart, Stephen C. Benz, J. Zachary Sanborn, Charles J. Vaske, Jingchun Zhu, Christopher Szeto, Gary K. Scott, Christina Yau, Sam Ng, Ted Goldstein, Kyle Ellrott, Eric Collisson, Aaron E. Cozen, Daniel Zerbino, Christopher Wilks, Brian Craft, Paul Spellman, Robert Penny, Troy Shelton, Martha Hatfield, Scott Morris, Peggy Yena, Candace Shelton, Mark Sherman, Joseph Paulauskis, Julie M. Gastier-Foster, Jay Bowen, Nilsa C. Ramirez, Aaron Black, Robert Pyatt, Lisa Wise, Peter White, Monica Bertagnolli, Jen Brown, Gerald C. Chu, Christine Czerwinski, Fred Denstman, Rajiv Dhir, Arnulf Dörner, Charles S. Fuchs, Jose G. Guillem, Mary Iacocca, Hartmut Juhl, Andrew Kaufman, Bernard Kohl III, Xuan Van Le, Maria C. Mariano, Elizabeth N. Medina, Michael Meyers, Garrett M. Nash, Phillip B. Paty, Nicholas Petrelli, Brenda Rabeno, William G. Richards, David Solit, Pat Swanson, Larissa Temple, Joel E. Tepper, Richard Thorp, Efsevia Vakiani, Martin R. Weiser, Joseph E. Willis, Gary Witkin, Zhaoshi Zeng, Michael J. Zinner, Carsten Zornig, Mark A. Jensen, Robert Sfeir, Ari B. Kahn, Anna L. Chu, Prachi

Kothiyal, Zhining Wang, Eric E. Snyder, Joan Pontius, Todd D. Pihl, Brenda Ayala, Mark Backus, Jessica Walton, Jon Whitmore, Julien Baboud, Dominique L. Berton, Matthew C. Nicholls, Deepak Srinivasan, Rohini Raman, Stanley Girshik, Peter A. Kigonya, Shelley Alonso, Rashmi N. Sanbhadti, Sean P. Barletta, John M. Greene, David A. Pot, Kenna R. Mills Shaw, Laura A. L. Dillon, Ken Buetow, Tanja Davidsen, John A. Demchok, Greg Eley, Martin Ferguson, Peter Fielding, Carl Schaefer, Margi Sheth, Liming Yang, Mark S. Guyer, Bradley A. Ozenberger, Jacqueline D. Palchik, Jane Peterson, Heidi J. Sofia, and Elizabeth Thomson. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, July 2012.

[112] 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):90–94, April 2011.

[113] 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, and Peter J Campbell. The Life History of 21 Breast Cancers. *Cell*, 149(5):994–1007, May 2012.

[114] C O Nordling. A new theory on cancer-inducing mechanism. *British Journal of Cancer*, 7(1):68–72, March 1953.

[115] P C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, October 1976.

[116] Layla Oesper, Ahmad Mahmoody, and Benjamin J Raphael. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology*, 14(7):R80, 2013.

[117] Shuji Ogino, Jérôme Galon, Charles S Fuchs, and Glenn Dranoff. Cancer immunology–analysis of host and tumor factors for personalized medicine. *Nature reviews. Clinical oncology*, 8(12):711–9, January 2011.

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

[119] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, David N Louis, Orit Rozenblatt-Rosen, Mario L Suvà, Aviv Regev, and Bradley E Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, June 2014.

[120] Benedict Paten, Mark Diekhans, Dent Earl, John St John, Jian Ma, Bernard Suh, and David Haussler. Cactus Graphs for Genome Comparisons. *Journal of Computational Biology*, 18(3):469–481, March 2011.

[121] Benedict Paten, Daniel R Zerbino, Glenn Hickey, and David Haussler. A unifying model of genome evolution under parsimony. *BMC Bioinformatics*, 15(1):206, 2014.

[122] A-M Pearse and K Swift. Allograft theory: transmission of devil facial-tumour disease. *Nature*, 439(7076):549–549, February 2006.

[123] P A Pevzner. DNA physical mapping and alternating Eulerian cycles in colored graphs. *Algorithmica*, 13(1-2):77–105, February 1995.

[124] George Poulogiannis, Ian M Frayling, and Mark J Arends. DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. *Histopathology*, 56(2):167–179, January 2010.

[125] Elizabeth Purdom, Christine Ho, Catherine S Grasso, Michael J Quist, Raymond J Cho, and Paul Spellman. Methods and Challenges in Timing Chromosomal Ab-

normalities within Cancer Samples. *Bioinformatics*, 29(24):btt546–3120, September 2013.

[126] A. R Quinlan, R. A Clark, S Sokolova, M. L Leibowitz, Y Zhang, M. E Hurles, J. C Mell, and I. M Hall. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Research*, 20(5):623–635, May 2010.

[127] 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, January 2012.

[128] Steven A Rosenberg and Nicholas P Restifo. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science*, 348(6230):62–68, April 2015.

[129] 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. *Nature Methods*, 11(4):396–398, April 2014.

[130] J Zachary Sanborn, Sofie R Salama, Mia Grifford, Cameron W Brennan, Tom Mikkelsen, Suresh Jhanwar, Sol Katzman, Lynda Chin, and David Haussler. Double Minute Chromosomes in Glioblastoma Multiforme Are Revealed by Precise Reconstruction of Oncogenic Amplicons. *Cancer Research*, 73(19):6036–6045, October 2013.

[131] David Sankoff. Edit distance for genome comparison based on non-local operations. In *Combinatorial Pattern Matching*, pages 121–135. Springer Berlin Heidelberg, 1992.

[132] Charles L Sawyers, Andreas Hochhaus, Eric Feldman, John M Goldman, Carole B Miller, Oliver G Ottmann, Charles A Schiffer, Moshe Talpaz, Francois Guilhot, Michael W N Deininger, Thomas Fischer, Steve G O'Brien, Richard M Stone, Carlo B Gambacorti-Passerini, Nigel H Russell, Jose J Reiffers, Thomas C Shea, Bernard Chapuis, Steven Coutre, Sante Tura, Enrica Morra, Richard A Larson, Alan Saven, Christian Peschel, Alois Gratwohl, Franco Mandelli, Monique Ben-Am, Insa Gathmann, Renaud Capdeville, Ronald L Paquette, and Brian J Druker. Imatinib induces hematologic and cytogenetic responses in patients with chronic myelogenous leukemia in myeloid blast crisis: results of a phase II study. *Blood*, 99(10):3530–3539, May 2002.

[133] Ton N Schumacher and Robert D Schreiber. Neoantigens in cancer immunotherapy. *Science*, 348(6230):69–74, April 2015.

[134] Manfred Schwab. Oncogene amplification in solid tumors. *Seminars in Cancer Biology*, 9(4):319–325, August 1999.

[135] Judith A Schwartzbaum, James L Fisher, Kenneth D Aldape, and Margaret Wrensch. Epidemiology and molecular pathology of glioma. *Nature Clinical Practice Neurology*, 2(9):494–503, September 2006.

[136] A F Scott, B J Schmeckpeper, M Abdelrazik, C T Comey, B O'Hara, J P Rossiter, T Cooley, P Heath, K D Smith, and L Margolet. Origin of the human l1 elements: proposed progenitor genes deduced from a consensus dna sequence. *Genomics*, 1(2):113–25, Oct 1987.

[137] Maria del Carmen Seleme, Melissa R Vetter, Richard Cordaux, Laurel Bastone, Mark A Batzer, and Haig H Kazazian. Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 103(17):6611–6, April 2006.

[138] S P Shah, A Roth, R Goya, A Oloumi, G Ha, and Y Zhao. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486:395–399, June 2012.

[139] Sohrab P Shah, Ryan D Morin, Jaswinder Khattra, Leah Prentice, Trevor Pugh,

Angela Burleigh, Allen Delaney, Karen Gelmon, Ryan Guliany, Janine Senz, Christian Steidl, Robert A Holt, Steven Jones, Mark Sun, Gillian Leung, Richard Moore, Tesa Severson, Greg A Taylor, Andrew E Teschendorff, Kane Tse, Gulisa Turashvili, Richard Varhol, Ren eacute L Warren, Peter Watson, Yongjun Zhao, Carlos Caldas, David Huntsman, Martin Hirst, Marco A Marra, and Samuel Aparicio. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, 461(7265):809–813, October 2009.

[140] Mingfu Shao and Yu Lin. Approximating the edit distance for genomes with duplicate genes under DCJ, insertion and deletion. *BMC Bioinformatics*, 13(Suppl 19):S13, December 2012.

[141] S Solyom, A D Ewing, E P Rahrmann, T Doucet, H H Nelson, M B Burns, R S Harris, D F Sigmon, A Casella, B Erlanger, S Wheelan, K R Upton, R Shukla, G J Faulkner, D A Largaespada, and H H Kazazian. Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Research*, 22(12):2328–2338, December 2012.

[142] Andrea Sottoriva, Inmaculada Spiteri, Sara G M Piccirillo, Anestis Touloumis, V Peter Collins, John C Marioni, Christina Curtis, Colin Watts, and Simon Tavaré. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 110(10):4009–4014, March 2013.

[143] Arun Sreekumar, Laila M Poisson, Thekkelnaycke M Rajendiran, Amjad P Khan,

Qi Cao, Jindan Yu, Bharathi Laxman, Rohit Mehra, Robert J Lonigro, Yong Li, Mukesh K Nyati, Aarif Ahsan, Shanker Kalyana-Sundaram, Bo Han, Xuhong Cao, Jaeman Byun, Gilbert S Omenn, Debashis Ghosh, Subramaniam Pennathur, Danny C Alexander, Alvin Berger, Jeffrey R Shuster, John T Wei, Sooryanarayana Varambally, Christopher Beecher, and Arul M Chinnaiyan. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, 457(7231):910–914, February 2009.

[144] Mark D Stenglein and Reuben S Harris. APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. *The Journal of Biological Chemistry*, 281(25):16837–41, June 2006.

[145] 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, January 2011.

[146] Chip Stewart, Deniz Kural, Michael P Strömberg, Jerilyn A Walker, Miriam K

Konkel, Adrian M Stütz, Alexander E Urban, Fabian Grubert, Hugo Y K Lam, Wan-Ping Lee, Michele Busby, Amit R Indap, Erik Garrison, Chad Huff, Jinchuan Xing, Michael P Snyder, Lynn B Jorde, Mark A Batzer, Jan O Korbel, Gabor T Marth, and 1000 Genomes Project. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genetics*, 7(8):e1002236, Aug 2011.

[147] Roger Stupp, Warren P Mason, Martin J van den Bent, Michael Weller, Barbara Fisher, Martin J B Taphoorn, Karl Belanger, Alba A Brandes, Christine Marosi, Ulrich Bogdahn, Jürgen Curschmann, Robert C Janzer, Samuel K Ludwin, Thierry Gorlia, Anouk Allgeier, Denis Lacombe, J Gregory Cairncross, Elizabeth Eisenhauer, René O Mirimanoff, European Organisation for Research and Treatment of Cancer Brain Tumor and Radiotherapy Groups, and National Cancer Institute of Canada Clinical Trials Group. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *The New England journal of medicine*, 352(10):987–996, March 2005.

[148] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005.

[149] Sebastian Szpakowski, Xueguang Sun, José M Lage, Andrew Dyer, Jill Rubinstein, Diane Kowalski, Clarence Sasaki, Jose Costa, and Paul M Lizardi. Loss of

epigenetic silencing in tumors preferentially affects primate-specific retroelements. *Gene*, 448(2):151–67, December 2009.

[150] Mariko Taniguchi-Ikeda, Kazuhiro Kobayashi, Motoi Kanagawa, Chih-chieh Yu, Kouhei Mori, Tetsuya Oda, Atsushi Kuga, Hiroki Kurahashi, Hasan O Akman, Salvatore DiMauro, Ryuji Kaji, Toshifumi Yokota, Shin'ichi Takeda, and Tatsushi Toda. Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. *Nature*, 478(7367):127–31, October 2011.

[151] TCGA Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–8, October 2008.

[152] TCGA Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–15, June 2011.

[153] H M Temin. Homology between RNA from Rouse Sarcoma Virous and DNA from Rouse Sarcoma virus-infected cells. *Proceedings of the National Academy of Sciences*, 52:323–329, August 1964.

[154] Erik Teugels, Sylvia De Brakeleer, Guido Goelen, Willy Lissens, Erica Sermijn, and Jacques De Grève. De novo alu element insertions targeted to a sequence common to the brca1 and brca2 genes. *Hum Mutat*, 26(3):284, Sep 2005.

[155] J L Tsao, Y Yatabe, R Salovaara, H J Järvinen, J P Mecklin, L A Aaltonen, S Tavaré, and D Shibata. Genetic reconstruction of individual colorectal tu-

mor histories. *Proceedings of the National Academy of Sciences*, 97(3):1236–1241, February 2000.

[156] Kaja Urbańska, Justyna Sokołowska, Maciej Szmidt, and Paweł Sysa. Glioblastoma multiforme - an overview. *Contemporary oncology (Poznań, Poland)*, 18(5):307–312, 2014.

[157] B Vogelstein, E R Fearon, S R Hamilton, S E Kern, A C Preisinger, M Leppert, Y Nakamura, R White, A M Smits, and J L Bos. Genetic alterations during colorectal-tumor development. *The New England Journal of Medicine*, 319(9):525–532, September 1988.

[158] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, March 2013.

[159] Fei Wang, Wenbo Meng, Bingyuan Wang, and Liang Qiao. Helicobacter pylori-induced gastric inflammation and gastric cancer. *Cancer letters*, 345(2):196–202, April 2014.

[160] Jianxin Wang, Lei Song, Deepak Grover, Sami Azrak, Mark A Batzer, and Ping Liang. dbrip: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Human Mutation*, 27(4):323–9, Apr 2006.

[161] Ting Wang, Jue Zeng, Craig B Lowe, Robert G Sellers, Sofie R Salama, Min Yang, Shawn M Burgess, Rainer K Brachmann, and David Haussler. Species-

specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences*, 104(47):18613–8, Nov 2007.

[162] Robert A Weinberg. *The Biology of Cancer, Second Edition.* Garland Science, Taylor & Francis Group, LLC, New York, 2nd edition edition, 2014.

[163] John S Welch, Timothy J Ley, Daniel C Link, Christopher A Miller, David E Larson, Daniel C Koboldt, Lukas D Wartman, Tamara L Lamprecht, Fulu Liu, Jun Xia, Cyriac Kandoth, Robert S Fulton, Michael D McLellan, David J Dooling, John W Wallis, Ken Chen, Christopher C Harris, Heather K Schmidt, Joelle M Kalicki-Veizer, Charles Lu, Qunyuan Zhang, Ling Lin, Michelle D O'Laughlin, Joshua F McMichael, Kim D Delehaunty, Lucinda A Fulton, Vincent J Magrini, Sean D McGrath, Ryan T Demeter, Tammi L Vickery, Jasreet Hundal, Lisa L Cook, Gary W Swift, Jerry P Reed, Patricia A Alldredge, Todd N Wylie, Jason R Walker, Mark A Watson, Sharon E Heath, William D Shannon, Nobish Varghese, Rakesh Nagarajan, Jacqueline E Payton, Jack D Baty, Shashikant Kulkarni, Jeffery M Klco, Michael H Tomasson, Peter Westervelt, Matthew J Walter, Timothy A Graubert, John F DiPersio, Li Ding, Elaine R Mardis, and Richard K Wilson. The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell*, 150(2):264–278, July 2012.

[164] John S Welch, Timothy J Ley, Daniel C Link, Christopher A Miller, David E Larson, Daniel C Koboldt, Lukas D Wartman, Tamara L Lamprecht, Fulu Liu, Jun

Xia, Cyriac Kandoth, Robert S Fulton, Michael D McLellan, David J Dooling, John W Wallis, Ken Chen, Christopher C Harris, Heather K Schmidt, Joelle M Kalicki-Veizer, Charles Lu, Qunyuan Zhang, Ling Lin, Michelle D O'Laughlin, Joshua F McMichael, Kim D Delehaunty, Lucinda A Fulton, Vincent J Magrini, Sean D McGrath, Ryan T Demeter, Tammi L Vickery, Jasreet Hundal, Lisa L Cook, Gary W Swift, Jerry P Reed, Patricia A Alldredge, Todd N Wylie, Jason R Walker, Mark A Watson, Sharon E Heath, William D Shannon, Nobish Varghese, Rakesh Nagarajan, Jacqueline E Payton, Jack D Baty, Shashikant Kulkarni, Jeffery M Klco, Michael H Tomasson, Peter Westervelt, Matthew J Walter, Timothy A Graubert, John F DiPersio, Li Ding, Elaine R Mardis, and Richard K Wilson. The origin and evolution of mutations in acute myeloid leukemia. *Cell*, 150(2):264–278, July 2012.

[165] David J Witherspoon, Jinchuan Xing, Yuhua Zhang, W Scott Watkins, Mark A Batzer, and Lynn B Jorde. Mobile element scanning (me-scan) by targeted high-throughput sequencing. *BMC Genomics*, 11:410, Jan 2010.

[166] D M Woodcock, C B Lawler, M E Linsenmeyer, J P Doherty, and W D Warren. Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. *The Journal of biological chemistry*, 272(12):7810–6, March 1997.

[167] J Xing, Y Zhang, K Han, A. H Salem, S. K Sen, C. D Huff, Q Zhou, E. F Kirkness,

S Levy, M. A Batzer, and L. B Jorde. Mobile elements create structural variation: Analysis of a complete human genome. *Genome Res*, 19(9):1516–1526, Sep 2009.

[168] Shinichi Yachida, Siân Jones, Ivana Bozic, Tibor Antal, Rebecca Leary, Baojin Fu, Mihoko Kamiyama, Ralph H Hruban, James R Eshleman, Martin A Nowak, Victor E Velculescu, Kenneth W Kinzler, Bert Vogelstein, and Christine A Iacobuzio-Donahue. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, 467(7319):1114–1117, October 2010.

[169] S Yancopoulos, O Attie, and R Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, August 2005.

[170] Sophia Yancopoulos and Richard Friedberg. DCJ Path Formulation for Genome Transformations which Include Insertions, Deletions, and Duplications. *Journal of Computational Biology*, 16(10):1311–1338, October 2009.

[171] J A Yoder, C P Walsh, and T H Bestor. Cytosine methylation and the ecology of intragenomic parasites. *Trends in genetics : TIG*, 13(8):335–40, August 1997.

[172] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18(5):821–9, May 2008.

[173] Daniel R Zerbino, Benedict Paten, Glenn Hickey, and David Haussler. An algebraic framework to sample the rearrangement histories of a cancer metagenome with double cut and join, duplication and deletion events. *arXiv.org*, March 2013.

[174] Yuan Zhu, Frantz Guignard, Dawen Zhao, Li Liu, Dennis K Burns, Ralph P Mason, Albee Messing, and Luis F Parada. Early inactivation of p53 tumor suppressor gene cooperating with NF1 loss induces malignant astrocytoma. *Cancer Cell*, 8(2):119–130, August 2005.

Appendix A

Additional CN-AVG results for TCGA GBM patients

The following are plots for checking the quality of CN-AVG sampling as in Figure 4.2, and for viewing the overall consensus evolutionary history as in Figure 4.3 for three other TCGA GBM patients.

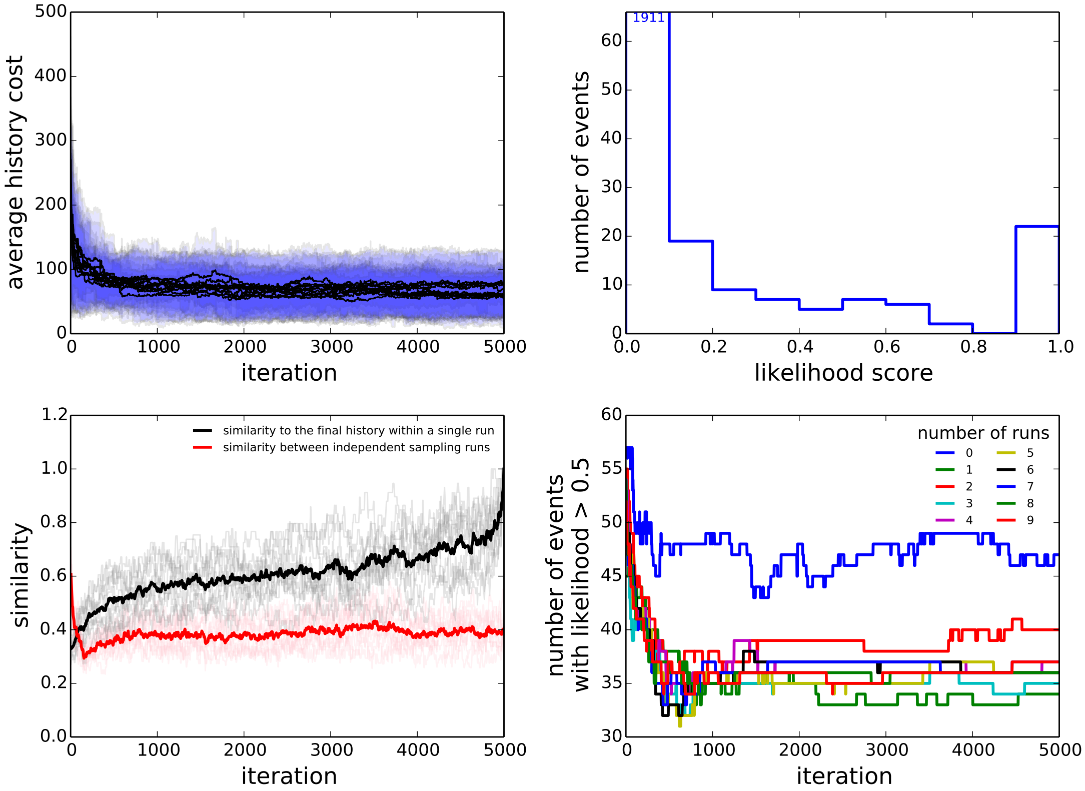


Figure A.1: Plots for checking the quality of CN-AVG sampling for patient 0152. The flat cost curves in the top left show that we have reached a cost minimum in the sampling. The bottom left plot shows that about 40 of the histories are the same between independent runs. The histogram in the top right shows that the likelihood scores have a bimodal distribution, so the sample is behaving as the simulations did. Lastly, the bottom right plot shows that there are 34 to 36 consistent events, and that more sampling will not change this set.

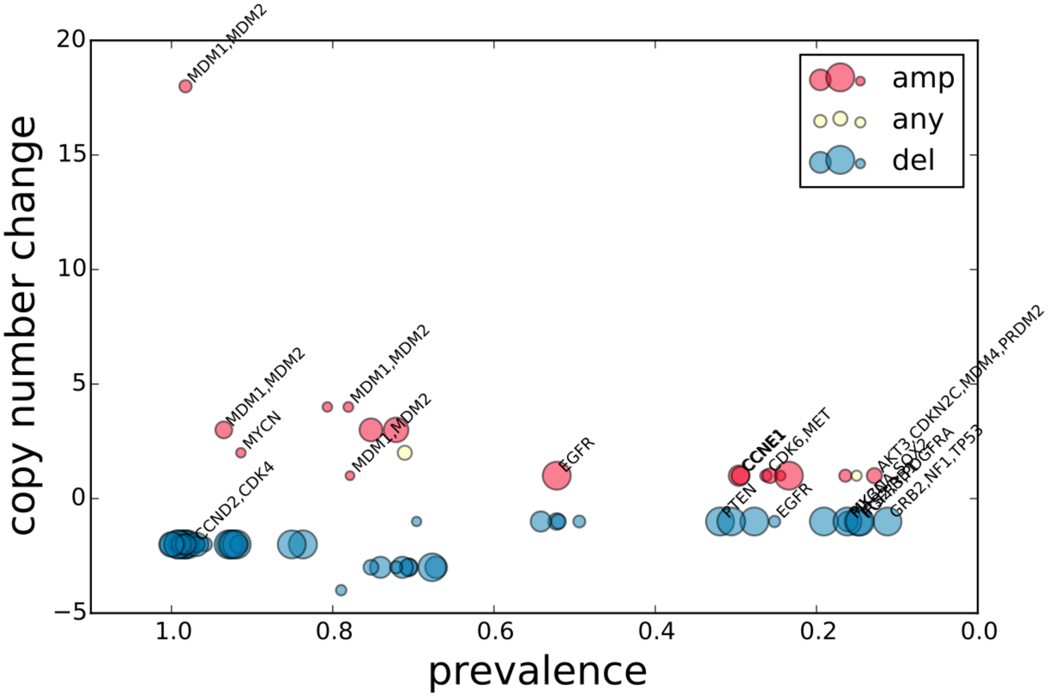


Figure A.2: Overview of the events in patient 0152 with likelihood of greater than 0.1 in the CN-AVG sampling, labeled with commonly mutated genes in GBM. Here we see a large amplification of MDM1 and MDM2 early in the tumor, followed by several other amplifications throughout its history. CCND2 and CDK4, both cyclins involved in tumor suppressor pathways, are deleted early. Of note, this patient has a double minute chromosome containing MDM1 and MDM2, which would explain the extremely high copy numbers of those genes.

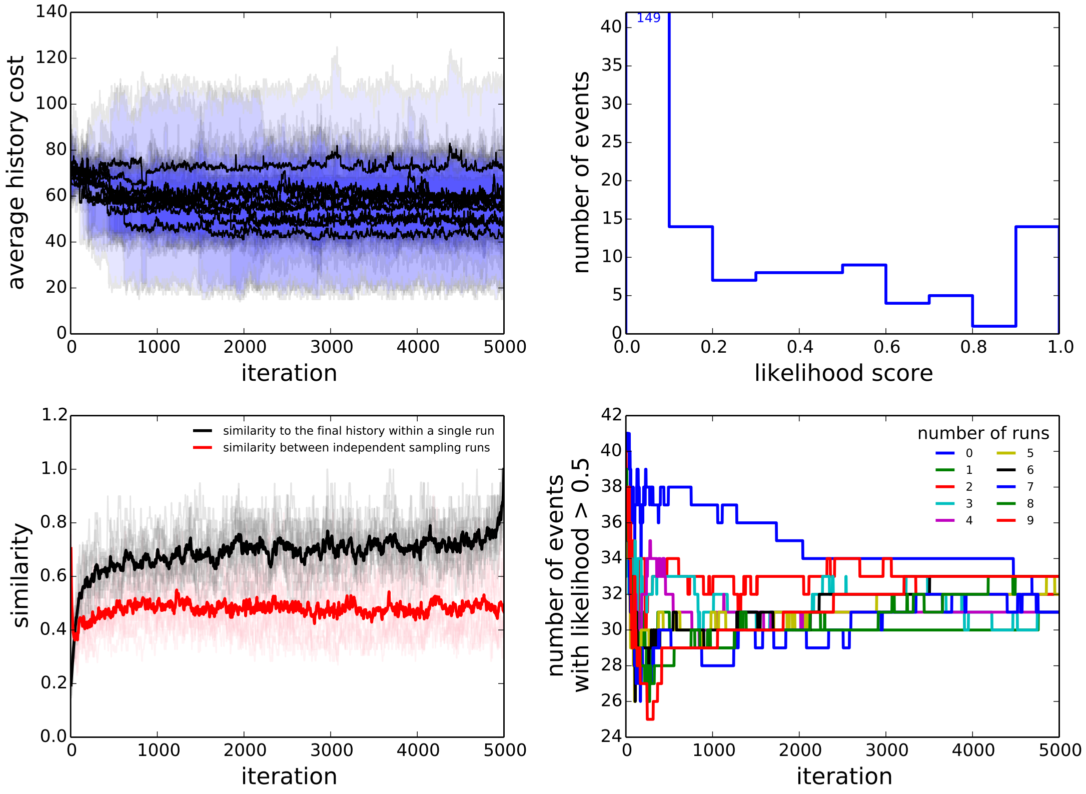


Figure A.3: Plots for checking the quality of CN-AVG sampling for patient 0155. The cost curves in the top left show how the random starting point in the sampling affects the results, with some sampling runs achieving much lower costs than others. This illustrates the importance of multiple random starts in MCMC. The bottom left plot shows that a little over 40% of the histories are the same between independent runs, and that the sampling within a single run never reaches this level of diversity, as it lies at around 60%. The histogram in the top right shows that the likelihood scores have a bimodal distribution, although the modes are not as strong as in the simulations. For this sample, we may want to chood a more stringent cutoff (0.8) for a set of accurate events. Lastly, the bottom right plot shows that there are 32 consistent events, and that more sampling will not change this set.

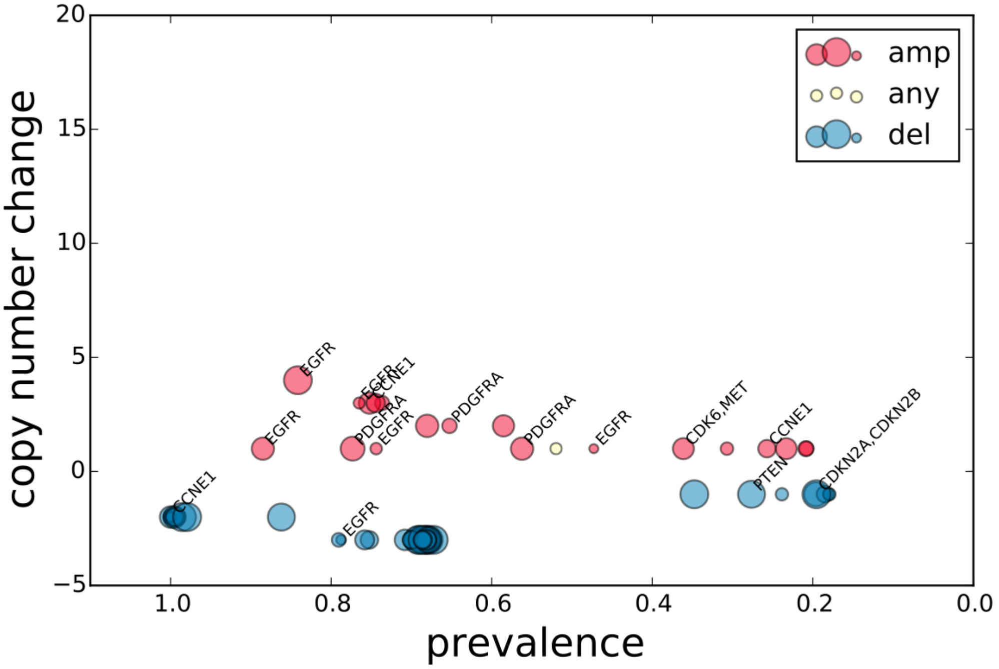


Figure A.4: Overview of the events in patient 0155 with likelihood of greater than 0.1 in the CN-AVG sampling, labeled with commonly mutated genes in GBM. In this patient, there is an early complete loss of CCNE1 (at -2 copy number change and near 1.0 prevalence) followed by multiple EGFR amplifications, although at a low amplitude. PDGFRA, another growth factor receptor, is amplified after EGFR.

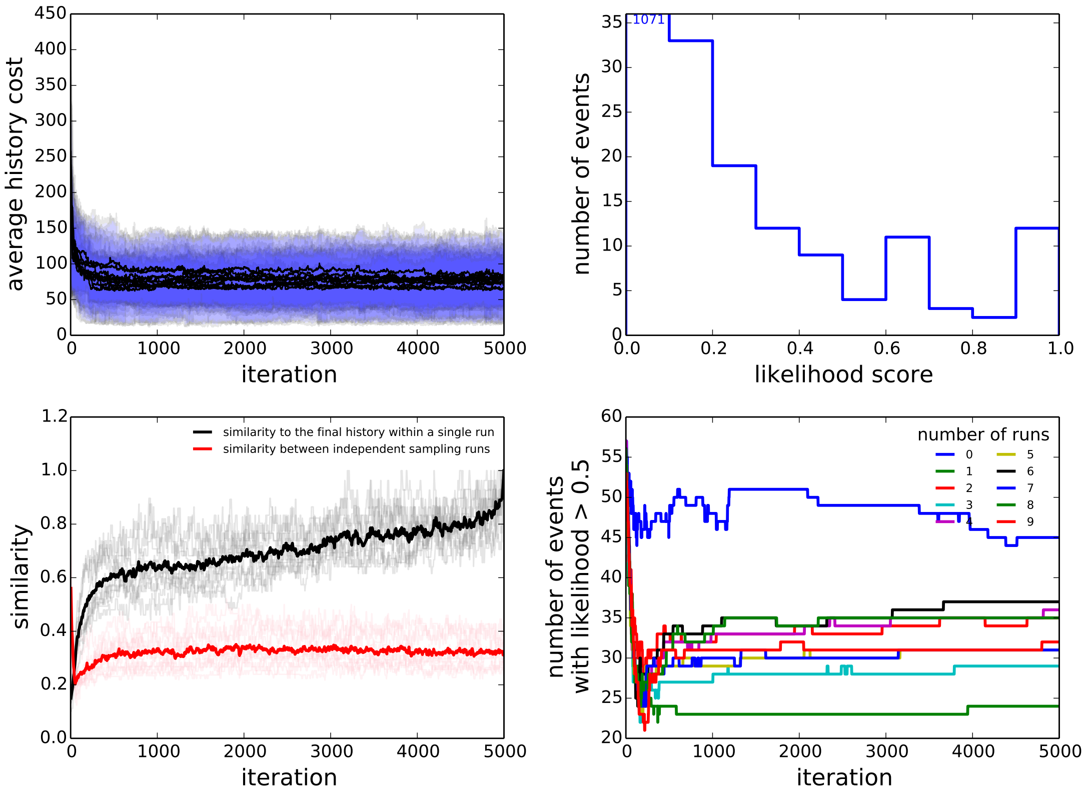


Figure A.5: Plots for checking the quality of CN-AVG sampling for patient 0185. The cost curves in the top left show that all sampling runs converge to similar costs. The bottom left plot shows that around 30% of the histories are the same between independent runs, and that the sampling within a single run never reaches this level of diversity, as it lies at around 60%. The histogram in the top right shows that the likelihood scores have a bimodal distribution, although the modes are not as strong as in the simulations. For this sample, we may want to chood a more stringent cutoff (0.8) for a set of accurate events. Lastly, the bottom right plot shows that there are around 30 consistent events, and that more sampling will not change this set.

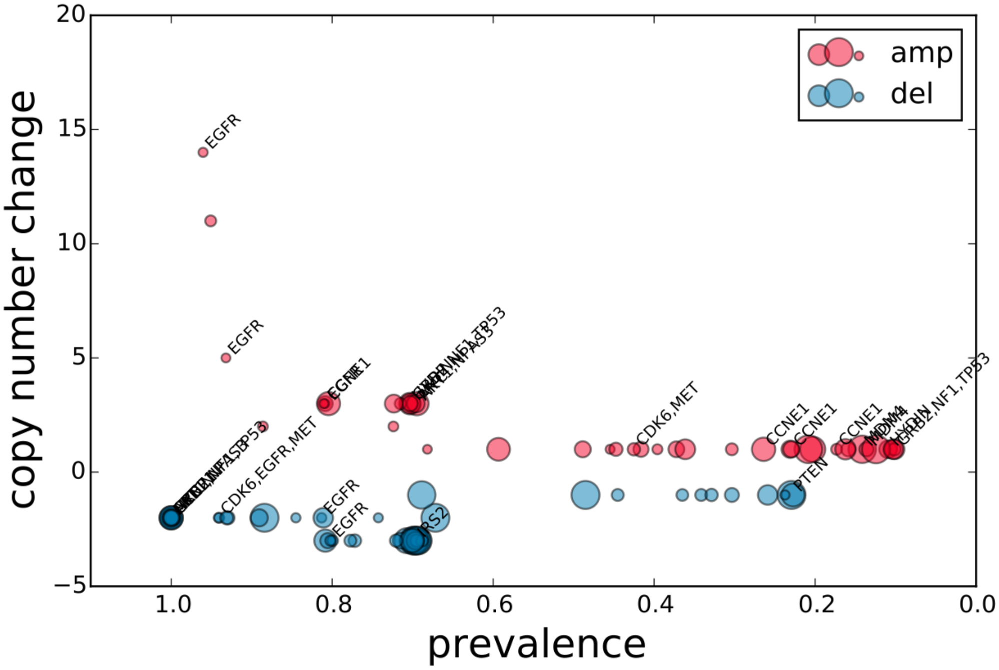


Figure A.6: Overview of the events in patient 0185 with likelihood of greater than 0.1 in the CN-AVG sampling, labeled with commonly mutated genes in GBM. Similar to patient 0145, we see multiple amplifications of EGFR early on in tumor formation, and may indicate the presence of a double minute. We also see some deletions of EGFR, and this type of balancing effect may also indicate double minute formation, as it results from complex genomic rearrangements occurring over multiple genomic regions. CN-AVG may be constructing a history in which EGFR was amplified as part of a large chromosomal amplification and subsequently deleted in a subset of cells.