UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Tumor versus Matched-Normal Sequencing Analysis and Data Integration

Permalink

https://escholarship.org/uc/item/57f8b691

Author

Sanborn, John Zachary

Publication Date

2012

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

TUMOR VERSUS MATCHED-NORMAL SEQUENCING ANALYSIS AND DATA INTEGRATION

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

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

J. Zachary Sanborn

December 2012

The Dissertation of J. Zachary Sanborn
is approved:

Professor David Haussler, Chair

Professor Joshua Stuart

Professor Nader Pourmand

Dean Tyrus Miller
Vice Provost and Dean of Graduate Studies

Copyright © by

J. Zachary Sanborn

2012

Table of Contents

List of Figures **v**

List of Tables **xiii**

Abstract **xiv**

Dedication **xv**

Acknowledgments **xvi**

1 Introduction **1**

2 Previous Work **7**
2.1 UCSC Cancer Genomics Browser 7
2.2 BamBam v1.0 11

3 Methods to analyze sequencing data of tumors and their matched-normals **15**
3.1 Somatic & germline variant calling 15
3.1.1 Results 19
3.2 Small insertions and deletions 20
3.3 Relative copy number, allele-specific copy number, and allele fraction . . 22
3.4 Inferring regions of structural variation using paired-end clustering . . . 24
3.5 Refinement of structural variation using Bridget 26
3.6 Bridget *de novo* 28
3.6.1 Results 30
3.7 Tumor Genome Browser 34

4 Integrating sequencing results to locally reconstruct the tumor genome **38**
4.1 Breakpoint traversal to infer local tumor genome assemblies 39
4.2 Breakpoints associated with highly amplified regions 42

4.3 Results . . . 42
4.3.1 Homozygous deletion of CDKN2A in glioblastoma multiforme . . 42
4.3.2 Reconstructing highly amplified regions using whole genome sequencing . . . 46
4.3.3 FISH confirmation of double minutes . . . 56
4.3.4 Evidence of tumor-specific reads in blood . . . 60
4.4 Discussion . . . 64

5 Inferring clonality and tumor evolution 66
5.1 Copy number alterations, allele fraction, and the allelic state diagram . 70
5.2 Fitting sequencing data to the ASD . . . 79
5.3 Modeling the clonal mixture of a tumor sample . . . 82
5.4 Linking mutations to clone-specific allelic states . . . 86
5.4.1 Via direct phasing . . . 86
5.4.2 Via amplified allele fraction . . . 87
5.4.3 Comparing allele fractions to infer mutation timing . . . 88
5.5 Results . . . 90
5.5.1 Clonality of glioblastoma multiforme tumors . . . 91
5.5.2 Clonality of squamous cell carcinomas of the lung . . . 100
5.5.3 Phased mutations to allelic states . . . 102
5.6 Discussion . . . 109

6 Conclusion 112
6.1 Clone-directed mutation discovery . . . 113
6.2 Clinical applications . . . 114
6.3 Final thoughts . . . 117

List of Figures

2.1 UCSC Cancer Genomics Browser. (a) Whole-genome view of 119 breast tumor samples measured on a MIP comparative genomic hybridization (CGH) array, sorted by Her2 status (black: Her2-, yellow: Her2+). Amplifications and deletions are shown in red and blue, respectively, whereas normal copy number is shown in white. A Student's t-test was performed comparing Her2+ vs. Her2-, red indicates genomic regions that are highly amplified in the Her2+ group. (b) Zoomed into chromosome 17 for both the CGH data shown in (a) as well as gene expression measured on Affymetrix U133A microarrays for a superset of the breast tumor samples from (a), sorted similarly and with the same t-test applied. Up- and down-regulated genes are shown in red and green, respectively, whereas no change in expression is shown in black. Note the ERBB2 locus is both amplified and highly expressed in the Her2+ samples, as expected. A region co-amplified with the ERBB2 locus is found upstream on chromosome 17. 10

3.1 Normalized mutated triplets for (a) 17 GBM tumors and (b) 11 LUSC tumors, sequenced by TCGA. The scale is mutations per million sites, which corrects for the uneven distribution of base triplets in the human genome. The samples listed on the right provide the order of samples displayed on the bar graphs, from top-to-bottom. 21

3.2 (a) Overview of the Bridget method to refine structural variants detected by BamBam using split reads. Two sets of 16bp tiles (red and blue lines) derived from the putative split read are shown to align to different segments of the reference genome. The two reference segments A & B are each defined by a "spanning set" of contiguous tiles in the reference and unaligned read, and the pair of them define the "dual-spanning set" (DSS) that describe the breakpoint junction. Note that the overlap of the two regions in the DSS defines the junction's homology. (b) An example Bridget result for a tandem duplication identified in a serous ovarian carcinoma sample sequenced by TCGA, denoted OV-0751. A total of 6 unaligned reads contributed to this result (sandwiched by the respective regions in the reference genome), discovering a 3bp microhomology of "AGC" at the junction point. 29

3.3 Bridget *de novo* results for 13 whole genome GBMs and their matched-normals. (a) Histogram of indel size showing distinct peaks for insertions and deletions corresponding to SINE and LINE families of repetitive elements. (b) Junction homology for germline (blue) and somatic (red) breakpoints showing peaks that indicate at least two primary mechanisms for breakpoint induction and repair: microhomology-mediated end-joining (MMEJ), non-homologous end joining (NHEJ) and repeat-mediated, non-allelic homologous recombination (NAHR). Note the enrichment of MMEJ & NHEJ breakpoints among the somatic breakpoints. 33

3.4 Distance between Bridget-refined breakpoints vs. breakpoints available in the Database of Genomic Variants (DGV). Each dot represents a single breakpoint found by Bridget, where its location on the x-axis is defined by $D(b_{bridget}, b_{dgv}) + 1.0$ between the Bridget solution and the closest breakpoint in DGV. A distance of 1.0 is a perfect score to place it on a log scale. The cluster at 2bp is due to a difference in how DGV stored breakpoints belonging to "Sequence mapping" category, utilizing a one-based genome coordinate system instead of the zero-based coordinates that Bridget reports. Also, not that some studies such as those categorized as "Illumina sequencing" have higher differences, with a peak around 12bp. Those particular studies did not incorporate split read methods to refine their breakpoints. Similarly, breakpoints derived from chip-based assays have much higher differences from the closest Bridget solution, as expected. . 35

3.5 UCSC Tumor Genome Browser. The browser shows all of the high-level somatic difference discovered by BamBam in a single image, enabling the synthesis of multiple distinct datasets to give the user an overall picture of the tumor's genome. The browser is able to zoom into and out of genomic regions rapidly, going from the full genome view, as shown above, to a single base resolution in just a few clicks. (a) Linear view of a highly-rearranged and amplified region on chromosome 12 in GBM-06-0648. (b) A circular genome view of the same region showing how the amplified peaks on chromosome 12 are connected to an amplified region of chromosome 22. It also shows a set of clustered rearrangements on chromosome 9 that caused the homozygous deletion of CDKN2A/B. . . 36

4.1 Toy examples of breakpoint graphs and their solutions. (a) One possible two path solution that accounts for all observed breakpoints and tumor copy number. Note that the deletions in both paths could be swapped to create another equally valid solution to the breakpoint graph. (b) A circular solution to the breakpoint graph for a highly-amplified region. The circular nature of the solution may indicate the presence of a double minute chromosome and/or a homogeneously-staining region (HSR). . . 41

4.2 Browser shots demonstrating how CDKN2A was homozygously deleted in: (a) GBM-06-0188, (b) GBM-06-0145, and (c) GBM-06-0648. Inter- and intra-chromosomal rearrangements caused both copies of CDKN2A to be deleted in (a) and (b), while a clustered set of intra-chromosomal rearrangements deleted both copies in (c). 45

4.3 Browser shot of the inter- and intra-chromosomal, highly-supported breakpoints and copy number observed for TCGA sample GBM-06-0648. . . . 49

4.4 (a) Diagram of the circular solution to breakpoint graph for GBM-06-0648 that accounts for all observed, highly-supported breakpoints, as well as the solution for the remaining, un-amplified segments of chromosome 12. (b) Circular assembly of the circular solution, possibly indicating the presence of a double minute chromosome. Note the positions of MDM2 and the CPM-Novel Exon gene. (c) The linear solution of the remaining contiguous segments on chromosome 12. (d) RNA-Seq data for GBM-06-0648 that indicates a novel exon replaced the final exon of CPM in the circular assembly. 50

4.5 Browser shot of the inter- and intra-chromosomal, highly-supported breakpoints and copy number observed for TCGA sample GBM-06-0152 for (a) chromosome 12 and (b) chromosome 7. 52

4.6 (a) Diagram of the two circular solutions to breakpoint graph for GBM-06-0152 that accounts for all observed, highly-supported breakpoints. (b) Circular assemblies of the two circular solutions, possibly indicating the presence of two independent double minute chromosomes. The larger circular solution has two copies of MDM2 and one copy of CDK4, while the smaller solution has EGFR. (c) A zoomed-in browser shot of the amplified segment labeled h, showing the distinct changes in copy number as described in (a). 53
4.7 Browser shot of the intra-chromosomal, highly-supported breakpoints and copy number observed for TCGA sample GBM-06-0145 for chromosome 7. 56
4.8 (a) Diagram of the three paths that together form a potential solution of the breakpoint graph for sample GBM-06-0145 that accounts for all observed, highly-supported breakpoints. The location of the gene EGFR is noted. Note that all paths, A, B and C, can create circular solutions by using the encompassing tandem duplication shown in red to connect the end of each path to the beginning. The tandem duplication may also connect path A to path B, path C to path B, etc. (b) The effects of each path on the EGFR gene, showing that path B creates an oncogenic form of EGFR, EGFRvIII, through specific deletion of exons 2-7. 57
4.9 Fluorescent *in situ* hybridization (FISH) images for samples used to determine the physical state of the amplified regions. (a) FISH image for GBM-06-0648 using probes for MDM2 (red) and the centromere of chromosome 12 (green). Note that the improved focus of the image on the right displays the punctuate pattern in red, characteristic of a DM, confirming that MDM2 is amplified by DM, as predicted. (b) FISH image for GBM-06-0145 using probes for EGFR (red) and centromere of chromosome 7 (green). Note that the red probes in most cells are clustered together, as opposed to the pattern seen in (a), indicative of an HSR. However, there are cells in (b), such as the bottom-left cell in the right image of (b), that display a DM-like pattern similar to that observed in (a). This suggests that, while the dominant pattern indicates EGFR amplification via HSR, the EGFR amplicon may also take the form of double minute chromosomes. 59
4.10 Bar graphs showing breakpoint read support for tumor (black bars) and blood (red bars) for sample: (a) GBM-06-0152, (b) GBM-06-0648, and (c) GBM-06-0145. 62

4.11 Browser shot of raw sequencing reads for GBM-06-0648 surrounding the leftmost position of a highly-supported deletion-type breakpoint that is part of this sample's putative DM. (a) The position of the deletion-type breakpoint and an excerpt of the tumor sequencing reads that support the breakpoint and the associated amplicon. (b) The sequencing reads for the matched-normal (blood) sample for GBM-06-0648 showing a set of 9 sequencing reads that support the breakpoint shown in (a). 63

5.1 A toy example of the evolution of a tumor starting from the germline, to an initial tumor cell, to a population of major and minor clones that are sampled by the tumor biopsy. To avoid complicating this example, mutation and copy number passenger events that occurred prior to the emergence of the initial tumor cell were not included. 69
5.2 An example of an allelic state diagram (ASD) of simulated data for a monoclonal tumor sample with zero normal contamination, $\alpha = 0$. Chromosomal regions exhibiting different copy number alterations are plotted in different colors. This simulated tumor genome exhibits 6 allelic states: normal (black), single-copy amplification (yellow), hemizygous deletion (red), homozygous deletion (dark green), copy-neutral LOH (CN-LOH, blue), and multi-copy biallelic amplification (cyan). 72
5.3 A set of allelic state diagrams for the same simulated monoclonal tumor genome shown in Figure 5.2 with different levels of normal contaminant: (a) 0%, (b) 10%, (c) 50%, and (d) 90%. Notice that, as the level of normal contamination increases, the distance between allelic states is reduced, negatively affecting our ability to distinguish between different states. 73
5.4 Allelic state diagram of the case presented in Figure 5.2 showing some possible bidirectional and unidirectional transitions between its allelic states. Note that unidirectional transitions are those that involve irreversible loss of a parental chromosome. 75
5.5 Allelic state diagram of the case presented in Figure 5.2 showing how transitional allelic states are created when the tumor consists of a mixture of two different clones. In this case, there are two subclones. Clone A is defined by the original allelic states: yellow (2,1), cyan (5,2), and red (1,0). Clone B alters these states through amplifications and a deletion to produce the allelic states: yellow (2,2), cyan (4,2), and red (2,0). They share the same allelic states of blue (2,0), black (1,1), and green (0,0). The percentages denote the percentage of clone B present in the tumor population, where 0% describes a monoclonal population of clone A, and 100% is a monoclonal population of clone B. 76

5.6 Allelic state diagram of the case presented in Figure 5.2 showing the transitional allelic states produced when allelic states are "skipped," which can occur when the tumor consists of two or more unrelated, or distantly-related clones. Notice that the transitional states are not found on the edges connecting allelic states if the allelic states of the two major clones differ in both majority and minority alleles. 78
5.7 Allelic state diagrams for two GBM tumors: (a) GBM-06-0145 and (b) GBM-06-0185. The fitted parameters found normal contamination at 21.5% and 14.6%, respectively. Notice that in (a) there are only clonal allelic states and no evidence of transitional allelic states, indicating that GBM-06-0145 is a monoclonal tumor. However, in (b), there are both clonal states and multiple transitional allelic states. Since the blue and purple transitional allelic states feature three different mixture percentages, this polyclonal tumor must consist of at least three subclones. The black X's plotted in (b) represent "landmark" allelic states that will be used to determine the clonal mixture of GBM-06-0185. 93
5.8 Monoclonal allelic profile for GBM-06-0145. The "Relative Coverage" and "Allele Fraction" displayed at the top of the plot shows both the observed output by BamBam and the imputed coverage and allele fraction generated by rendering the data as the mixture of the single clone and normal contamination. The comparison of real versus modeled data shows very strong agreement. The clone's allelic profile below shows the majority and minority allelic states for the tumor genome, showing amplification of one copy of entire chr7 and chr19, complete loss of one copy of chr10, and arm-level loss of chr9p. 95
5.9 Polyclonal allelic profiles for GBM-06-0185. A total of 4 distinct clones were identified in this tumor, with clone D determined to be the dominant clone of the population, comprising 42.7% of the tumor sample. All clones have single-copy amplifications of chr7, chr19 & chr20, single copy loss of chr10 and chr22, and loss of chr9p in common. Clones B, C, & D all have deletions in chr6, but clone B's deletions are focal while clones C & D display arm-level loss of chr6q. Clone D is further distinguished by amplification of the intact copy of chr9. 96
5.10 Polyclonal allelic profiles for GBM-06-0152. Three clones were identified in this tumor sample that has an estimated 24.1% normal contamination. All clones share amplifications of chr1, chr19 & chr20, deletions of chr10 & chr22, and focal losses of chr12 related to the chromothripsis-like event that created two DMs described in the previous chapter. Clones B & C exhibit amplification of chr7 and deletions of the non-amplified copy of chr1 as well as chr2, chr3, chr4, chr8, chr13, and chr17. Clone C further amplifies the remaining copy of chr8. 98

5.11 Polyclonal allelic profiles for GBM-06-1086. Four clones were identified in this tumor sample that has an estimated 7.5% normal contamination. All clones share amplification of chr21 and deletions of chr9 & chr11p. Clones C & D exhibit significant chromosomal loss, deleting chr1, chr3, chr4, chr5, chr6, chr8, chr10, chr13, chr14, chr15, chr17, chr18, and chr20. The dominant clone D, making up 41.6% of the tumor sample, further deletes the sole remaining copy of chr18 and amplifies chr19. The black arrows indicate the position of CDK2NA in clones A & B, highlighting the arrival of the focal deletion of CDKN2A in the latter clone. 99
5.12 Allelic state diagrams for two LUSC tumors: (a) LUSC-34-2596 and (b) LUSC-66-2756. The fitted parameters found normal contamination at 36.4% and 19.6%, respectively. These tumors are highly polyclonal, with LUSC-66-2756 exhibiting a large number of highly amplified allelic states and zero single copy loss allelic states, which could have resulted from a genome doubling event. 101
5.13 Example of seven phased mutations on the dual ASD. A representative region on the tumor genome is shown with blue and red horizontal lines representing chromosomes in the tumor mapping to this region, consisting of a subregion in the single copy gain allelic state, a subregion in the "normal" allelic state, a subregion exhibiting CN-LOH, and a subregion exhibiting LOH. Mutations 1, 2, and 3 are found in the amplified subregion, where mutations 2 & 3 are majority-phased (red stars) and mutation 1 is minority-phased (blue star). Mutations 4 & 5 are found in the "normal" allelic state, where mutation 4 is majority-phased and mutation 5 is minority-phased. Both regions exhibiting LOH have one mutation each phased to the sole remaining allele, and thus mutations 6 & 7 are majority-phased. The dual ASD below shows where each of these mutations would be found, using each mutation's corrected allele fraction, MAF_c, to determine its location along the x-axis. Note the different placement of mutations 2 & 3 in the single-copy gain allelic state, where mutation 3, the mutant allele that exists on both chromosomes with the majority allele, is found near the single copy gain majority allelic state. Mutation 2 is found near the single copy gain *minority* allelic state, correctly identifying that the mutation exists on only one copy of the majority allele and hence is likely to occur after the duplication event. Finally, note that the minority-phased mutations in blue are all found in the left-half of the dual ASD. 104

5.14 Phased mutations on the dual ASD for tumor GBM-06-0145. 7 regions are encircled on these plots: (a) majority-phased mutations in an amplified allelic state, but presenting with MAF_c suggesting the mutation is only on one of the two copies, (b) majority-phased mutations in an amplified allelic state with MAF_c suggesting mutations are present on both amplified copies, (c) majority-phased mutations with an allele fraction consistent with the LOH allelic state, (d) minority-phased mutations in the majority-amplified allelic state with MAF_c consistent with single copy, (e) unphased mutations with MAF_c consistent with amplified allelic state, and (f) unphased mutations with MAF_c consistent with LOH allelic state. LOH has left this cell with only a mutated copy of the gene DOCK8. 105

5.15 Phased mutations on the dual ASD for tumor LUSC-34-2596. The two encircled regions, (a) and (b), show a number of mutations phased to the majority and minority alleles, respectively, in the balanced amplified allelic state (2,2). One majority-phased mutation in NDRG1 is found in a transitional allelic state with matching MAF_c. The locations of two missense mutations, in BRAF & DNMT3A, and one nonsense mutation in TP53 are shown in the unphased plot, placing BRAF in a highly amplified allelic state, DNMT3A in the "normal" allelic state, and TP53 in the CN-LOH state. 108

List of Tables

4.1 Types of CDKN2A Loss in 17 TCGA GBMs. 46

5.1 Clonality of 12 GBMs. 91
5.2 Phased mutations in GBM and LUSC tumors. 110

Abstract

Tumor versus Matched-Normal Sequencing Analysis and Data Integration

by

J. Zachary Sanborn

Cancer research is a study of tragic complexity. The more we discover about this disease, the more we realize how much we have to learn. It relentlessly unravels new complications, justifying the difficulty we have had in ridding cancer from the human experience. Nevertheless, we strive to conquer it. To accomplish this noble goal, we require tools to understand a patient's specific tumor in as comprehensive a manner as possible. The methods presented here aim to provide a complete assessment of a tumor's genome, discovering the alterations made to the patient's germline genome that may have given rise to cancer. Mutations both large and small are identified and integrated using various techniques to discover the events and elements specific to the tumor genome. The methods will be used on whole genome sequencing data of several tumors, revealing the structural events that occurred in multiple tumors. Double minute chromosomes in GBM tumors will be discovered computationally and confirmed experimentally. Finally, techniques that expose tumor clonality will be described and used to infer whole genome, allele-specific copy number profiles of all major clones comprising a tumor population from a single tumor biopsy.

To my beautiful wife,

Shirley Truong,

for her friendship, support, and, most of all, patience through these many years of graduate school.

Acknowledgments

I would like to thank my advisor David Haussler for all of his advice and encouragement throughout my long graduate school journey. I thank my committee members Josh Stuart and Nader Pourmand for all of their advice. Of course, I thank the many members of the UCSC cancer group, especially Jing Zhu, Sofie Salama, Chris Szeto, Chris Wilks, Amie Radenbaugh, Mia Griffor, Ting Wang, and the late Fan Hsu. Lastly, I would like to thank my friends and Five3 Genomics cofounders Steve Benz and Charlie Vaske who will hopefully help to make the research presented in this thesis actually improve people's lives one day. The work has been partially supported by a NIH Training Grant, the I-SPY INSTINCT project, The Cancer Genome Atlas (TCGA) Pilot and U24 funding sources, the ARCS Foundation, and Stand Up To Cancer (SU2C) project.

Chapter 1

Introduction

According to the American Cancer Society, cancer accounts for nearly a quarter of all deaths in America, nearly 600,000 people in 2012, with only heart disease killing more people per year [1]. Approximately 1 in 2 males and 1 in 3 females will develop cancer at some point in their life, with the leading incidence sites being lung, colon, and prostate (or breast) for males (or females). Fortunately, cancer death rates have been in modest decline since 1990 for both men and women, with the decrease attributed to behavioral changes, such as reduced rates of smoking, earlier detection with improved screening methods, such as the prostate-specific antigen (PSA) blood-based test for prostate cancer, and advances in cancer treatment, but there is great room for improvement. With the advent of precision medicine, significant advances in the treatment of cancer will be possible through improved understanding of both common and unique features of each patient's tumor. For example, the particular treatment giving

to a women diagnosed with breast cancer is influenced by measurements of the biological markers estrogen, progesterone, and HER2 receptors [94, 77]. Also, patients with chronic myelogenous leukemia (CML) are likely to respond well to treatment with the tyrosine kinase inhibitor imatinib (*Gleevec*) if the gene fusion of BCR and the tyrosine kinase ABL is observed [23, 66].

In order to formulate effective precision cancer treatments, it is vital to understand how each cancer fulfills the general "requirements" for tumorigenesis. It is generally accepted that the formation of cancer requires the acquisition of five essential characteristics, the so-called "hallmarks of cancer": self-sufficiency in growth signals, insensitivity to antigrowth signals, avoiding cell death (apoptosis), unlimited replication, and sustained angiogenesis [35]. To develop metastatic cancer, the tumor must further acquire an ability to invade other tissues. Two emerging hallmarks, deregulation of cellular energetics and avoidance of immune destruction, as well as two enabling characteristics, genomic instability and inflammation, were recently added [36]. What makes cancer such a complex disease to both study and treat is that there are a number of different ways that tumors acquire these characteristics. Constitutional activation of BCL2, mutations to TP53, loss of APAF1 function, and alterations to the PI3K-AKT survival signaling pathway are but a few ways that cancer cells inhibit apoptosis [61]. Occasionally, single alterations provide the tumor with more than one of these essential characteristics. For example, in non-small cell lung cancer, angiogenesis is regulated via VEGF (vascular endothelial growth factor) or amplification of EGFR, which also promotes tumor cell growth [95]. Cancer research is focused on discovering and under-

standing the modifications that initiate tumorigenesis and finding treatments that can target cancer cells more effectively, with reduced impact on healthy cells. The goal as always: better treatments with less side-effects.

Identifying perturbations in genes or sets of genes that are shared across many tumors is a powerful method to discover events that may be important to tumorigenesis. A small region of chromosome 17 was discovered to be amplified in a subset of breast cancers, later found to be harboring the growth factor receptor ERBB2 [11]. This subtype later become one of five clinically-relevant breast cancer subtypes (basal-like, luminal A & B, ERBB2-overexpressing, and normal-like) defined by a set of differentially-expressed genes identified in several independent studies using different microarray technologies [62, 80, 81]. Interestingly, these gene expression signatures can occasionally help identify specific genomic events, such as the gene expression signature associated with a recurrent genomic translocation in acute lymphoblastic leukemia (ALL) [26]. Algorithms such GISTIC have been applied to tumor copy number data measured on comparative genomic hybridization (CGH) microarrays to find recurrent amplifications and deletions in various types of cancer [74, 10, 73]. Four recurring mutations in the coding sequence of three genes, IDH1, NRAS, and NPM1, and a potential regulatory region were discovered in a large cohort of acute myeloid leukemia (AML) samples [49]. Probably most famous of all, mutations in TP53 are nearly ubiquitous in many different types of cancer, reflecting its general importance to tumorigenesis [33].

Structural rearrangements of the genome are also known to be recurrent events in cancer, resulting in gene fusions that characterize particular cancers. In many CML

cancers, a genomic translocation between chromosomes 9 and 22 (denoted t(9;22)) results in the fusion of two genes, BCR and ABL, to form the chimeric protein BCR-ABL [8]. When localized to the nucleus, ABL can induce apoptosis in response to DNA damage, but the chimeric protein BCR-ABL is localized exclusively to the cytoplasm and activates anti-apoptotic pathways [91]. By performing a meta-analysis of several gene expression datasets on prostate cancer, researchers discovered chimeric proteins created by fusing TMPRSS2 to members of the ETS family (ERG and ETV1), one of the first recurrent gene fusion events found in cancers of epithelial origin [87]. Although recurrent structural variations appear to be rare in non-hematological cancers, researchers argue that this is due to the difficulty in observing these events in these cancers due to their more complex karyotypes [41]. In fact, it is believed that gene fusions are equally common in blood, soft, or solid cancers, with the bias due to the relative lack of recurrent chromosomal aberrations observed in solid cancers [53].

The identification of recurrent events is essentially a search for events that drive the development and expansion of cancer, so called *driver* mutations that provide a competitive advantage in the cellular environment. Recurrent mutations are thought to be likely drivers because of the low likelihood that similar, or identical, mutations would occur independently in multiple tumors unless they contributed to tumorigenesis. This assumes that mutations arise randomly across the genome (i.e. no mutational hotspots). Cells that acquire these driver mutations are also likely to carry *passenger* mutations that provide neither benefit nor detriment to the tumor cell, but which tend to accumulate in the tumor genome simply because the passenger mutations occurred

in cells undergoing tumorigenesis.

Distinguishing which events act as drivers is difficult, especially since recurrent events are not always necessarily drivers. Recent studies have attempted to distinguish somatic mutations driving the development of cancer from passenger mutations. Large-scale resequencing of the FLT3 coding region in 222 AML patients identified nine non-synonymous mutations, of which only four were gain-of-function mutations found to constitutively activate FLT3's kinase domain [27]. The remaining neutral mutations observed were likely passengers of an unobserved oncogenic mutation. Similarly, another study on 188 AML samples found that, out of 30 novel non-synonymous mutations observed in 22 different kinases, only mutations to FLT3 had any effect on cell transformation [48]. In a study of drivers of liver cancer, researchers used RNAi in a mouse model to knock out orthologous genes recurrently deleted in human hepatocellular carcinomas, identifying 12 tumor suppressors out of the 301 genes tested [96]. Clearly, the evidence supports the notion that a significant fraction of observed mutations, even those occurring in the same genes across many tumors, are likely passenger mutations. To this end, researchers have developed mutational models that better model the non-uniform distribution of mutations across the genome in order to improve our ability to discover significantly mutated genes from genes that carry many passenger mutations [17, 57].

Ultimately, tumorigenesis must be understood from the context of evolution, where tumorigenesis is the process by which cells adapt in order to survive, and unfortunately prosper, inside a human host that actively seeks to destroy misbehaving cells.

Adaptations come about via both loss- and gain-of-function mutations, providing the emerging cancer cell with the ingredients essential for malignant growth. During the cell's evolution towards cancer, an initial tumor cell emerges and rapidly expands to form the majority of the observed tumor mass. During its growth, additional mutations may occur that can further drive or impede the cancer's growth, or are simply benign passenger mutations. From this evolutionary perspective, driver mutations are those that provide the tumor cell with a set of adaptations that optimize its ability to grow within its specific cellular environment. Driver mutations are more likely occur early in the tumor's development. This evolutionary process has been studied directly by investigating the evolution of the Ig heavy chain locus in 22 B-cell chronic lymphocytic leukemias, enabling researchers to divergent subclones in a phylogenetic tree that represents the evolution of the germline genome to that of the dominant clone [13]. The arrival of whole-genome sequences for tumors and matched-normal samples from TCGA provides an opportunity to not only identify genomic events (mutations, structural rearrangements, copy number) at a very high resolution but, as will be shown shortly, chronologically order genome-wide events to understand a single cancer's evolution [56].

Chapter 2

Previous Work

My disseration is based upon my previous work and the work of other members of the Haussler cancer group, which begins with work focused on the visualization and analysis of microarray-based, high-throughput data and ended with my first foray into whole genome sequencing analysis of tumor & matched-normal samples.

2.1 UCSC Cancer Genomics Browser

Given cancer's heterogenous nature, it is important to acquire as much information on each tumor in order to better understand the state of each individual cancer. To this end, cancer studies like TCGA use microarrays to measure gene expression, copy number variation, and methylation on each tumor and, when appropriate, matched-normals derived from blood or solid tissue of the same patient. TCGA has also identified somatic and germline variants through targeted gene re-sequencing using

traditional Sanger sequencing, and has just begun whole-genome sequencing of tumors using next generation sequencing platforms. By combining this data across hundreds of different tumors within the same tumor type, researchers hope to find genomic features that characterize a specific type, or subtype, of tumor while also noting the unique characteristics of each tumor. After collecting this variety of data on thousands of tumors, it becomes very important to have a tool that aids in the exploration and analysis of these data.

We developed the UCSC Cancer Genomics Browser (CGB) as a tool specifically designed for these large cancer studies [98, 67]. It is designed to simultaneously explore many types of data, both genomic and clinical, measured on multiple cancers in a single web-based platform accessible to both bioinformaticians and cancer biologists. The CGB interface features two heatmaps, one displaying genomic data in either genome coordinates or "genesets" and another displaying any available "clinical" data pertaining to each sample displayed. These two heatmaps are coupled vertically, meaning that each row of both heatmaps describes data pertaining to the same tumor sample. Within this vertical coupling lies much of the power of the CGB, because, with a single click, a user can sort samples based on a clinical feature and immediately visualize any patterns in the genomic heatmap that results. For example, amplification/over-expression of receptor ERBB2/Her2 characterizes all Her2+ breast cancer patients, a fact that becomes immediately apparent after sorting the array CGH data of a cohort of breast cancer tumors by their Her2 status, as shown in Figure 2.1. Interestingly, Figure 2.1 shows the presence of another amplicon just upstream in 17q11 that appears amplified

in the same tumors with ERBB2 amplification. These two distinct amplicons have been observed before in many Her2+ breast tumors, and may result by translocation of one region to the other followed by amplification of the newly adjacent regions as a single amplicon [11, 19, 43]. The CGB also features a set of simple statistical tools, such as Student's t and Wilcoxon rank-sum tests, to help discover associations between genomic and clinical data that are not as readily visible.

Another major feature of the UCSC CGB is its ability to group together genomic data into genesets, instead of being restricted to plotting data in genomic coordinates. Since genes often interact with genes that may be located anywhere in the genome, it would be very difficult to analyze genomic data in the proper biological context if constrained to genome coordinates, as the UCSC Genome Browser does. The CGB's "genesets view" does away with genomic coordinates by displaying the genomic data of related genes side-by-side in genesets. Over 7,000 genesets were collected from various databases, such as MSigDB, Reactome, Kyoto Encyclopedia of Genes and Genomes (KEGG), BioCarta, Gene Ontology (GO), and National Cancer Institute Pathway Interaction Database (NCI-PID), and consist of genes associated by cellular processes, genes discovered to be up- or down-regulated in certain cellular conditions, functional gene modules derived from computational scans of high-throughput protein-protein interactions, and other methods [85, 90, 38, 4, 68]. In addition to the built-in genesets, users can create their own genesets. All genesets are searchable via geneset name or by the name of any genes they contain.

The UCSC Cancer Genomics Browser is still under active development by

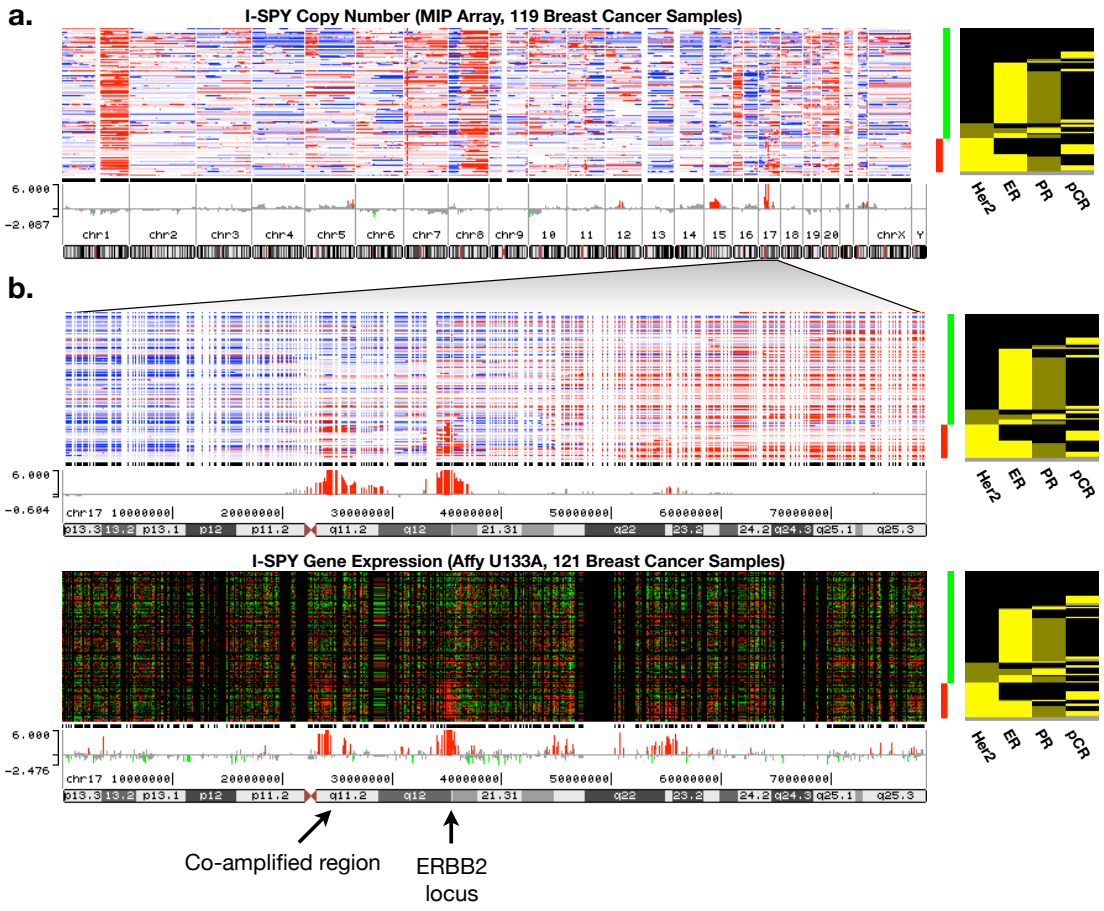

Figure 2.1: UCSC Cancer Genomics Browser. (a) Whole-genome view of 119 breast tumor samples measured on a MIP comparative genomic hybridization (CGH) array, sorted by Her2 status (black: Her2-, yellow: Her2+). Amplifications and deletions are shown in red and blue, respectively, whereas normal copy number is shown in white. A Student's t-test was performed comparing Her2+ vs. Her2-, red indicates genomic regions that are highly amplified in the Her2+ group. (b) Zoomed into chromosome 17 for both the CGH data shown in (a) as well as gene expression measured on Affymetrix U133A microarrays for a superset of the breast tumor samples from (a), sorted similarly and with the same t-test applied. Up- and down-regulated genes are shown in red and green, respectively, whereas no change in expression is shown in black. Note the ERBB2 locus is both amplified and highly expressed in the Her2+ samples, as expected. A region co-amplified with the ERBB2 locus is found upstream on chromosome 17.

the cancer genomics group led by Jing Zhu, with numerous improvements having been made. In addition to the recent upgrades to the web application's user interface, the team have added significant capabilities to the browser, including the ability to dynamically generate sample-specific metrics based on mathematical combinations of gene expression and/or copy number data, enabling gene-based biomarkers to be developed and immediately tested on the many thousands of samples available on the CGB.

2.2 BamBam v1.0

Near the end of August 2009, TCGA released whole-genome sequencing data for GBM and ovarian serous carcinoma tumors and matched-normal data derived from blood of the same patients. The data is stored in the BAM format, the block-compressed, binary equivalent of the Sequence Alignment/Map (SAM) format [46], which has become the de facto standard for storing aligned single- and/or paired-end short-reads due to its ease of use, quick read/write access, and well-designed software libraries. The original GBM tumor and normal samples were sequenced on Illumina GA sequencer to produce approximately 4-5 billion 76bp paired-end reads for each sample. The ovarian tumor and normal samples were sequenced on a later model Illumina GAII sequencer, providing roughly 3 billion paired-end 100bp reads. All reads were aligned to the hg18 reference using MAQ [47].

To determine if a mutation discovered in a tumor sample is somatic or a germline variant requires that we analyze both the tumor and matched-normal data.

This can be done sequentially, by summarizing data at every genomic position for both tumor and matched-normal and later combining the results for analysis. Unfortunately, because whole-genome BAM files are hundreds of gigabytes in their compressed form (1-2 terabytes uncompressed), the intermediate results required for later analysis will be extremely large and slow to merge and analyze. To avoid this and improve the efficiency of analyzing these large datasets, I wrote the first version of my sequencing analysis tool called BamBam.

What made BamBam unique among other sequencing analysis methods of its time was that it performed all of its analyses by simultaneously reading both tumor and matched-normal BAM files, piling up reads that overlap every common genomic location between the two files. This gave BamBam the ability to perform complex analyses, such as paired tumor vs. matched-normal mutation calling, inferring copy number alterations, that required sequencing data from both tumor and normal very quickly and with negligible memory use. Another important benefit to processing these files in sync was that the output BamBam produced could be drastically minimized since only the output could be restricted to just the likely somatic differences, producing what is essentially a whole-genome *diff* between the tumor and its germline.

The initial version of BamBam incorporated rudimentary consensus somatic SNP-calling, germline and somatic inter- and intra-chromosomal rearrangements, estimates of relative copy number, and finding regions exhibiting loss-of-heterozygosity (LOH).

Consensus SNP-calling is performed by a weighted-majority vote of the ac-

ceptable reads in the pileup, weighted by their Phred-scaled base quality scores:

$$MV(x) = \underset{\theta}{\operatorname{argmax}} \sum_{i}^{N} \delta(\theta = b_i, q_i > q_{min}) q_i, \tag{2.1}$$

where $\theta \in \{A, G, C, T\}$, δ is equal to 1 if its arguments are true, and b_i and q_i are the i-th nucleotide base and quality scores of the N reads piling up at the current genomic location, x, respectively. If the consensus base call in the tumor is different than in the matched-normal, then a SNP variant is called for that location. For LOH analysis, if there are enough acceptable reads to call a second allele at the location in the matched-normal, but no reads with that second allele present in tumor, then LOH is called for that location. For example, if the tumor has just 5 reads calling A and germline has 4 reads calling A and 3 reads calling G, then this location may have lost the G allele in the tumor via deletion. By considering only this single location, it could be that the second allele mutated from G→A in the tumor or that the G allele was simply not sampled. However, if we instead consider all of the LOH calls within the same genomic neighborhood and find the density of locations experiencing LOH is high relative to the average region, then it becomes more likely that the set of heterozygous alleles was lost through a deletion event rather than multiple locations all experiencing specific base substitutions or due to sampling bias.

BamBam v1.0 identified putative intra- and inter-chromosomal structural variants by clustering "discordant" paired reads where each read in the pair map to disparate regions of the hg18 reference. The discordant pairs that meet user-defined quality thresholds are aggregated to determine the two genomic regions that may be newly adja-

cent in both tumor and matched-normal. Discordant reads are clustered if they overlap on both sides of the putative rearrangement. To mitigate the effect of low coverage of the breakpoint, such that discordant reads do not directly overlap each other, a 500bp buffer is placed around the start and stop positions of each discordant read to extend its genomic span. When at least two pairs of reads overlap in this way, a structural variant is called and the breakpoint that defines it is recorded in the BamBam output for post-processing and visualization purposes. This version of the rearrangement detection algorithm only allowed a single discordant connection to exist at any given genomic location, so it could miss structural variants that are nearby (within 500bp) or directly overlap other structural variants.

Chapter 3

Methods to analyze sequencing data of tumors and their matched-normals

The methods described below are available in a pair of programs I wrote: an updated version of BamBam & Bridget. Both of these programs are designed to run quickly and use little memory, making them suitable to run on single-core and multi-core machines in addition to high-performance clusters.

3.1 Somatic & germline variant calling

Because BamBam keeps the sequence data in the pair of files in sync across the genome, a complex mutation model that requires sequencing data from both tumor and

matched-normal BAM files as well as the human reference can be implemented easily. This model aims to maximize the joint probability of both the germline genotype (given the matched-normal reads and the reference nucleotide) and the genotype of the tumor (given the germline genotype, a simple mutation model, an estimate of the fraction of contaminating normal tissue in the tumor sample, and the tumor sequence data).

To find the optimal tumor and germline genotype, we aim to maximize the likelihood defined by

$$\begin{aligned} & P(D_g, D_t, G_g, G_t | \alpha, r) \\ & = P(D_g|G_g)P(G_g|r)P(D_t|G_g, G_t, \alpha)P(G_t|G_g) \end{aligned} \tag{3.1}$$

where r is the observed reference allele, α is the fraction of normal contamination, and the tumor and germline genotypes are defined by $G_t = (t_1, t_2)$ and $G_g = (g_1, g_2)$ where $t_1, t_2, g_1, g_2 \in \{A, T, C, G\}$. The tumor and matched-normal sequencing data are defined as a set of reads $D_t = \{d_t^1, d_t^2, ...\}$ and $D_g = \{d_g^1, d_g^2, ...\}$, respectively, with the observed bases $d_t^i, d_g^i \in \{A, T, C, G\}$. All data used in the model must exceed user-defined base and mapping quality thresholds.

The probability of the germline alleles given the germline genotype is modeled as a multinomial over the four nucleotides:

$$P(D_g|G_g) = \frac{n!}{n_A! n_T! n_G! n_C!} \prod_i^n P(d_g^i|G_g), \tag{3.2}$$

where n is the total number of matched-normal reads at this position and n_A, n_T, n_C, n_G are the reads supporting each observed allele. The base probabilities, $P(d_g^i|G_g)$, are

assumed to be independent, coming from either of the two parental alleles represented by the genotype G_g, while also incorporating the approximate base error rate of the sequencer. The prior on the germline genotype is conditioned on the reference base as

$$P(G_g|r = a) = \{\mu_{aa}, \mu_{ab}, \mu_{bb}\}, \tag{3.3}$$

where μ_{aa} is the probability that the position is homozygous reference, μ_{ab} is heterozygous reference, and μ_{bb} homozygous non-reference. The germline prior does not incorporate any information on known SNPs, to avoid unnecessarily biasing the mutation caller at polymorphic sites. SNPs identified by BamBam as mutations can be found and filtered by comparing against recent builds of dbSNP [75].

The probability of the set of observed tumor alleles is again defined as multinomial

$$P(D_t|G_t, G_g, \alpha) = \frac{n!}{n_A!n_T!n_G!n_C!} \prod_i^n P(d_t^i|G_t, G_g, \alpha), \tag{3.4}$$

where the probability of each tumor allele is a mixture of base probabilities derived from both tumor and germline genotypes that is controlled by the fraction of normal contamination, α,

$$P(d_t^i|G_t, G_g, \alpha) = \alpha P(d_t^i|G_t) + (1 - \alpha)P(d_t^i|G_g) \tag{3.5}$$

and the probability of the tumor genotype is defined by a simple mutation model from on the germline genotype

$$P(G_t|G_g) = \max[P(t_1|g_1)P(t_2|g_2), P(t_1|g_2)P(t_2|g_1)], \tag{3.6}$$

where the probability of no mutation (e.g. $t_1 = g_1$) is maximal and the probability of transitions (i.e. $A \to G, T \to C$) are four times more likely than transversions (i.e. $A \to T, T \to G$). All model parameters, $\alpha, \mu_{aa}, \mu_{ab}, \mu_{bb}$, and base probabilities $P(d^i|G)$ for the multinomial distributions can be defined by the user.

The tumor and germline genotypes, G_g^{max}, G_t^{max}, selected are those that maximize Equation 3.1, and the posterior probability defined by

$$\frac{P(D_g, D_t, G_g^{max}, G_t^{max}|\alpha, r)}{\sum_{i,j} P(D_g, D_t, G_g = i, G_t = j|\alpha, r)} \tag{3.7}$$

can be used to score the confidence in the pair of inferred genotypes. If the tumor and germline genotypes differ, the putative somatic mutation(s) will be reported along with its respective confidence.

Maximizing the joint likelihood of both tumor and germline genotypes helps to improve the accuracy of both inferred genotypes, especially in situations where one or both sequence datasets have low coverage of a particular genomic position. Other mutation calling algorithms, such as MAQ and SNVMix, that analyze a single sequencing dataset are more likely to make mistakes when the non-reference or mutant alleles have low support [47, 29]. Since the development of BamBam, new mutation callers that incorporate both tumor and matched-normal sequencing data have been developed, further evidence that this is a powerful way to discover somatic variants from tumor and matched-normal sequencing data [42, 63].

3.1.1 Results

Algorithms used to assess the statistical significance of mutations, such as MutSig [17, 57] and MuSiC [21], normally require 100s of samples. As this thesis focuses on small numbers of whole genomes, assessing BamBam's mutation calls from a significance standpoint is difficult. However, we can instead perform a simple analysis on the genomic context of mutations called in the whole genomes of 17 glioblastoma multiforme (GBM) tumors and 11 squamous cell lung carcinomas (LUSC), both by sequenced by TCGA, to discover if different mutational processes are acting in these two tumor types.

The genomic context of a mutation consists of the bases that precede and follow the mutated base in the reference genome. The mutational context is then fully defined by a mutated triplet, i.e. ACC→ATC, describing a C→T transition within the A-C genomic context. To avoid undercounting mutations, the base that is mutated is allowed to be either C or A. If the base is G or T, then the triplet and mutant base are reverse-complemented. Thus, a total of 32 triplets represent the reference sequence and can be mutated one of 3 ways, for a possible 96 mutations across all possible triplets. The total number of mutations in each triplet category are tallied for a tumor or a set of tumors. These counts are then corrected by the relative frequency of each triplet in the human genome, to determine if certain genomic contexts are preferentially mutated.

Figure 3.1 shows the total number of somatic mutations found in each triplet for 17 GBMs and 11 LUSC tumors, normalized by the triplet count in the human

reference, in a layout inspired by Mike Lawrence of the Broad Institute (Unpublished). GBM tumors almost exclusively exhibit CG→TG mutations, one of the most common mutation types found in the human genome caused by deamination of 5'methyl-cytosine at methylated-CpG sites [3].

LUSC tumors, on the other hand, display a radically different mutational signature, exhibiting a preponderance of mutations in almost all triplet contexts. Notably, C→A (equivalently, G→T) mutations are particularly enriched in all but one LUSC tumor. This agrees with the known mutational signature of smoking-associated lung tumors that are characterized by G→T mutations resulting from the repair of bulky DNA adducts, where polycyclic aromatic hydrocarbons (PAH), the primary carcinogen found in cigarette smoke, covalently bond to DNA [22, 32]. One LUSC tumor, LUSC-21-1078, exhibits the fewest mutations of the 11 LUSC tumors, and very few G→T mutations, and is also the patient that smoked the least out of this set of tumors (9 pack-years vs. median 66 pack-years) [34].

3.2 Small insertions and deletions

BamBam currently employs a simplistic model for calling insertions and deletions (indels). The collection of reads that overlap a position is scanned for any and all evidence of both insertions and/or deletions. For each deletion, its size and genomic location is stored. For each insertion, the position of the insertion and the inserted base sequence is stored. An indel is then called if the identical indel (in location and size

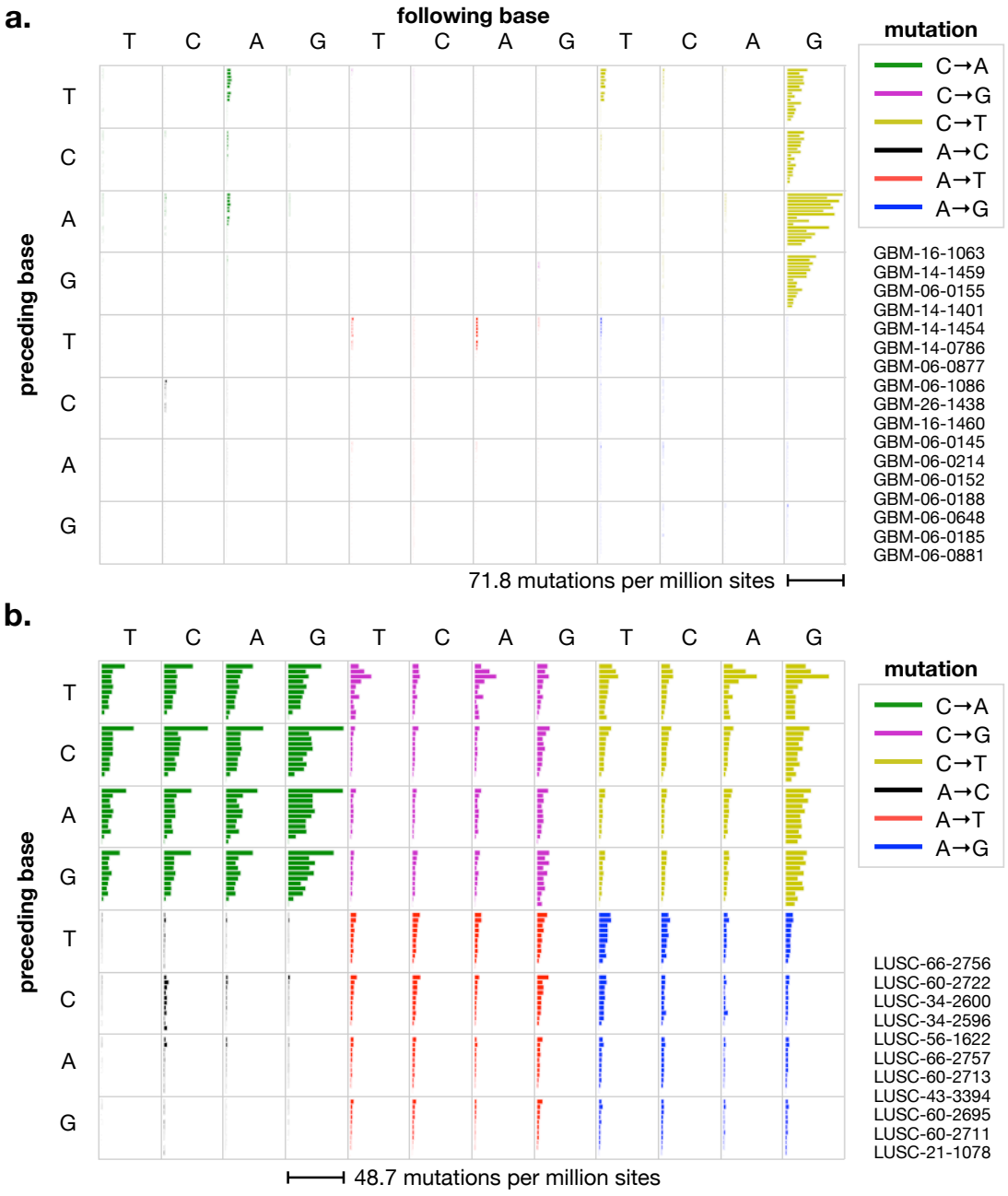

Figure 3.1: Normalized mutated triplets for (a) 17 GBM tumors and (b) 11 LUSC tumors, sequenced by TCGA. The scale is mutations per million sites, which corrects for the uneven distribution of base triplets in the human genome. The samples listed on the right provide the order of samples displayed on the bar graphs, from top-to-bottom.

or sequence) is observed more than a user-defined threshold N_{indel}, where $N_{indel} \geq 2$. If the indel is found only in the tumor and there exist no indels in the same region of the matched-normal, then it's classified as a somatic indel, otherwise it's classified as germline. Base quality of the inserted sequence is not incorporated in this model. The same mapping quality thresholds used for calling point variants are used to filter out poorly aligned reads.

Using the above model, BamBam currently calls a great number of indel variants and, from a brief study of these indel calls, it appears that many are likely false positives produced by systematic errors in the sequencer or aligner. Future versions of BamBam will incorporate an error model for insertions and deletions, but until such a model can be developed and tested, indel variants will only be utilized within a limited scope.

3.3 Relative copy number, allele-specific copy number, and allele fraction

Relative copy number of tumor vs. matched-normal is estimated using a dynamic windowing approach that expands and contracts the window's genomic width according to the read coverage in either the tumor or matched-normal data. The process is initialized with a window of zero width. Each unique read from either the tumor or matched-normal sequencing data will be tallied into tumor counts, N_t, or matched-normal counts, N_n. The start and stop positions of each read define the window's

region, expanding in width as new reads exceed the boundaries of the current window. When either the tumor or matched-normal counts exceed a user-defined threshold, the window's size and location are recorded, as well as the N_t, N_n, and relative coverage $\frac{N_t}{N_n}$. Tailoring the size of the window according to the local read coverage will create wide genomic windows in regions of low coverage (e.g. repetitive regions), or small windows in regions that are amplified in the tumor, thereby increasing the resolution of amplicon boundaries.

Allele-specific copy number is estimated similarly, except that all positions deemed homozygous in the germline are ignored. Germline heterozygosity is defined as a position where the majority allele represents from 50-60% of the total number of unique reads mapping to that position. Majority and minority copy numbers are calculated separately in a manner identical to that described above for overall copy number. All counts ascribed to the majority allele in both tumor and matched-normal data will go towards calculation of the majority copy number, and similarly for the minority allele. Allele-specific copy number is used to identify genomic regions exhibiting loss-of-heterozygosity as well as amplifications or deletions specific to one or both parental chromosomes. This last point is especially important to help distinguish if potentially disease-causing alleles, such as an inherited TP53 mutant, are retained or amplified in the tumor genome. Furthermore, regions that experience hemizygous loss (e.g. one parental chromosome arm) can be used to directly estimate the amount of normal contaminant in the sequenced tumor sample, which can be used to improve the modeling of the germline and tumor genotypes described above. This will be described in further

detail below.

In most cases, it is more useful to calculate allele fraction since its calculation does not depend on sequencing data from the matched-normal. Majority and minority allele fraction is simply defined as:

$$AF_{maj}(t_{maj}, t_{min}) = \frac{t_{maj}}{t_{maj} + t_{min}} \tag{3.8}$$

$$AF_{min}(t_{maj}, t_{min}) = \frac{t_{min}}{t_{maj} + t_{min}}$$

$$= 1 - AF_{maj} \tag{3.9}$$

where t_{maj} and t_{min} are the read depths for the majority and minority alleles in the tumor sample, as defined above. As before, allele fractions are only calculated for positions that are heterozygous in the matched-normal.

3.4 Inferring regions of structural variation using paired-end clustering

Putative intra- and inter-chromosomal rearrangements are discovered using discordant paired reads stored in a pair of coordinate-sorted BAM files, one from the tumor sample and the other from the matched-normal sample, where each read in the discordant pair map to disparate regions of the human reference genome. Intra-chromosomal discordant pairs are those that have an abnormally large insert size (i.e. the genomic distance on the reference separating the paired reads exceeds a user-defined threshold) or those that map in an incorrect orientation (i.e. inversion). Inter-

chromosomal discordant pairs are defined by paired reads that map to different chromosomes.

All discordant paired-end reads from both tumor and matched-normal datasets are clustered according to their genomic locations to define an approximate genomic region where the breakpoint is believed to be. The aggregation process consists of grouping together the reads that overlap other reads on both sides of the putative breakpoint and have the identical orientation, while keeping track of the number of reads that came from tumor and normal datasets. When the number of overlapping discordant pairs in a cluster exceeds a user-defined threshold, the breakpoint that describes the rearrangement is defined and recorded in the output.

Breakpoints are classified as "somatic" if they are significantly supported by reads from the tumor dataset. A minimal amount of read support for "somatic" breakpoints from the matched-normal dataset is allowed to accommodate cases where some minute level of tumor DNA is present in the matched-normal sample. This may occur when highly amplified regions present in a solid tumor's DNA are shed off into the bloodstream at low, but detectable levels. Alternatively, in the case of hematological cancers, it is expected that there will be a high level of tumor introgression in the matched-normal sample (typically skin tissue). The amount of support for somatic breakpoints allowed in the matched-normal dataset can be adjusted according to the expected amount of contaminating tumor tissue present in the matched-normal sample.

"Germline" breakpoints are those that have significant support in both tumor and matched-normal datasets or only find support in the matched-normal dataset.

These breakpoints will not be considered in the analyses presented here.

3.5 Refinement of structural variation using Bridget

The breakpoints discovered from paired-end clustering are approximate such that they utilize fully-mapped reads that, by their nature, cannot substantially overlap the actual junction of the two regions that are believed to rearranged. The base sequence of their junction represents sequence not present in the reference or, in the case of a somatic structural variant, in the matched-normal dataset. To refine the location of the breakpoint, we use a program I developed called Bridget that utilizes the sequence of unmapped reads anchored near the breakpoint to assemble the junction's sequence. An overview of this method and an example result is shown in Figure 3.2.

The regions surrounding each breakpoint are searched for all unaligned reads anchored near the putative breakpoint by a mapped mate. Each of these unmapped reads has the potential to be a "split read" that directly overlaps the rearrangement's breakpoint junction. Localized genomic sequences surrounding both sides of the breakpoint are broken up into a set of unique tiles of length 16bp. A database storing the tile sequences and their location in the reference genome is created. A similar tile database is constructed for each unaligned read, by breaking up the read into overlapping 16bp tiles and noting their position within the read. Comparing the reference tile database and the unaligned tile database, the location of each unaligned tile in the human reference genome is determined. We then look for sets of tiles that are contiguous in both

the reference and the read, defining what is termed a "spanning set" of tiles that span an equal amount of sequence in both reference and read space. The "dual-spanning set" (DSS) is a pair of such spanning sets, with one spanning set for each side of the breakpoint in the reference.

The genomic location of the DSS in reference coordinates precisely determines the breakpoint's location and its orientation (or strandedness) on both sides of the junction. The strandedness is simply determined by looking at each spanning set and determining if the reference tiles are moving in the same direction (same strand) or different directions (opposite strands). With the information describing position and orientation of both left and a right sides of the breakpoint, the junction sequence is fully defined through a precise rearrangement of the reference genome. The DSS also describes the sequence homology present at the breakpoint by noting any overlap of the two spanning sets within the read. Typically, this is short homologous sequence (microhomology), such as "CAA," that is identical in the reference sequences on the both sides of the breakpoint, but is observed only once in the unmapped read where the two junction sequences overlap.

For each unmapped read processed for a given breakpoint, the DSS determines a potential location for the breakpoint. Since every unmapped read may determine slightly different locations for the breakpoint (due to sequence errors near the breakpoint, repetitive reference, etc.), the breakpoint location for every DSS is used to generate a putative junction sequence. All unmapped reads are realigned using a Smith-Waterman alignment algorithm to each putative junction sequences and the over-

all improvement in their alignments is measured relative to the best alignment found in the unaltered reference sequence [79]. The junction sequence that yields the greatest improvement in Smith-Waterman alignment scores is judged as the best candidate for the true junction. If this best junction sequence yields negligible improvement in the alignment scores, then no suitable junction sequence could be found. In this case, the lack of split read confirmation by Bridget could be evidence that the initial breakpoint discovered by BamBam's paired-end clustering method was spurious.

3.6 Bridget *de novo*

Bridget is normally used to simply refine the breakpoints already discovered by BamBam. To reduce the number of structural variants produced, BamBam only reports breakpoints found in the expected orientation if the breakpoint's span (i.e. the distance between the two sides of the breakpoint) is greater than 2,000bp. Unfortunately, this means that deletions smaller than 2,000bp will not be found by BamBam. The need for such a high threshold is due to the fact that the number of called breakpoints goes up dramatically as the threshold is lowered, thereby greatly increasing the number of breakpoints that are sent to Bridget for split read confirmation. This is an inefficient and computationally expensive method for searching for small deletions.

To mitigate this limitation, a secondary mode for running Bridget was developed. This method, called Bridget *de novo*, fuses BamBam's breakpoint caller together with the Bridget's split read aligner. This enables breakpoint searching and split read

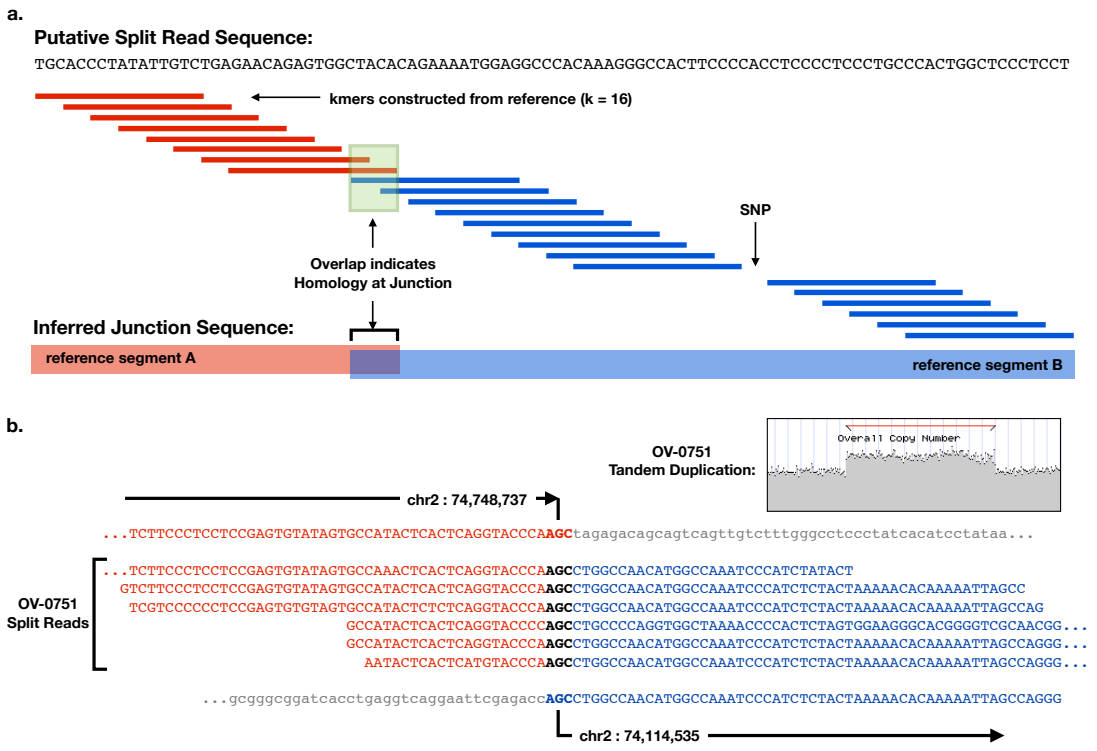

Figure 3.2: (a) Overview of the Bridget method to refine structural variants detected by BamBam using split reads. Two sets of 16bp tiles (red and blue lines) derived from the putative split read are shown to align to different segments of the reference genome. The two reference segments A & B are each defined by a "spanning set" of contiguous tiles in the reference and unaligned read, and the pair of them define the "dual-spanning set" (DSS) that describe the breakpoint junction. Note that the overlap of the two regions in the DSS defines the junction's homology. (b) An example Bridget result for a tandem duplication identified in a serous ovarian carcinoma sample sequenced by TCGA, denoted OV-0751. A total of 6 unaligned reads contributed to this result (sandwiched by the respective regions in the reference genome), discovering a 3bp microhomology of "AGC" at the junction point.

confirmation to be run as a cohesive pair, where immediately upon calling a putative breakpoint, split read evidence for that breakpoint is collected and analyzed. Only breakpoints that are called and have split read support are reported. This enables the threshold for breakpoint span to be significantly lowered, as the number of false positive calls are drastically reduced by the need for all calls to be confirmed by Bridget's inline split read aligner.

The first step of Bridget *de novo* analyzes one million aligned read pairs of the input BAM file to compute statistics on the distribution of insert sizes (i.e. the distance that separates two read pairs, defined during the preparation of the sequencing library). It then sets the threshold for breakpoint span to be one standard deviation away from the computed mean. Thus, the insert size distribution of the sequencing library, instead of a rigid threshold, determines which breakpoints will be analyzed. For example, for typical libraries with an insert size of 500 ± 100bp, the lower limit threshold for reads in the expected orientation will be approximately 600bp. Every breakpoint that exceeds this minimum threshold, or found in a non-reference orientation (e.g. inverted), and is also found to have both split read evidence and exhibits significant improvement in its Smith-Waterman alignment versus reference will be reported by Bridget *de novo*.

3.6.1 Results

Bridget *de novo* was independently run on the whole genome sequencing data of 13 TCGA GBMs and their matched-normals. The two sets of results were merged to identify the somatic status of each breakpoint, where breakpoints found in the matched-

normal exclusively or breakpoints in both matched-normal and tumor were classified as "germline" and those found solely in the tumor were classified "somatic." Two breakpoints are deemed similar if their distance

$$D(b_1, b_2) = \sqrt{(b_{1,start} - b_{2,start})^2 + (b_{1,stop} - b_{2,stop})^2} \tag{3.10}$$

is less than 10bp, where b_1 and b_2 are the two breakpoints being compared and $b_{i,start}$ and $b_{i,stop}$ are the start and stop positions, respectively, of the putative breakpoint b_i.

Figure 3.3(a) shows the distribution of insertion and deletion sizes identified in these 26 whole genome samples (13 tumors and 13 matched-normals). Peaks at indel sizes of 320bp and 6,060bp are observed, which match the approximate size of the SINE and LINE retrotransposons, respectively [64]. Since these mobile elements are still active, we expect the location and existence of these retrotransposons to be different in every person compared to the reference genome. Thus, the observation of these peaks suggest that Bridget *de novo* is identifying true structural variants in these genomes. Note that the peak at 320bp could not be identified by BamBam, and thus Bridget *de novo* has improved the detection of small structural variants, as desired.

Further evidence of the performance of Bridget *de novo* is found in Figure 3.3(b), which shows the distribution of the sizes of the homologous sequenced shared by the two sequences joined together. The results are shown for germline and somatic breakpoints separately. Two peaks, at 0-2bp and 14-16bp, are observed in the set of germline breakpoints. These two peaks suggest that at least two different mechanisms are primarily responsible for producing the observed breakpoints in the germline. There

are three major mechanisms known to induce or repair breakpoints. Microhomology-mediated end-joining (MMEJ) and non-homologous end joining (NHEJ) are two pathways in eukaryotic cells responsible for repairing double-strand breaks (DSBs) [51, 9, 54]. These mechanisms require little to no homology in the sequences to be joined, so they are likely candidates for the mechanisms responsible for breakpoints making up the 0-2bp peak. The third major mechanism, non-allelic homologous recombination (NAHR), requires extensive homology to join two sequences, and is likely the mechanism responsible for producing the indels comprising the 14-16bp peak in Figure 3.3(b) [18].

The distribution of somatic breakpoints in Figure 3.3(b) shows an enrichment for breakpoints created by the MMEJ or NHEJ mechanisms. This suggests that DSB repair is a frequent occurrence in the tumor genome, and supports the notion that DSB induction and repair is an engine of structural change in tumor genomes, as noted by multiple studies [39, 25].

For more evidence that Bridget *de novo* is discovering real structural variants, the breakpoints it identifies are compared to those in the Database of Genomic Variants (DGV) [37]. The DGV is a collection of germline structural variants discovered by multiple groups employing a variety of different wet-lab and computational techniques. Since DGV only stores germline variation, only germline-classified Bridget breakpoints are compared to DGV. For each germline breakpoint, the closest breakpoint in DGV is found using the absolute distance metric given in Equation 3.10. The results are shown in Figure 3.4 for each of the 13 categories of variants available in DGV. This figure demonstrates that Bridget's solutions are very close, and often identical, to the

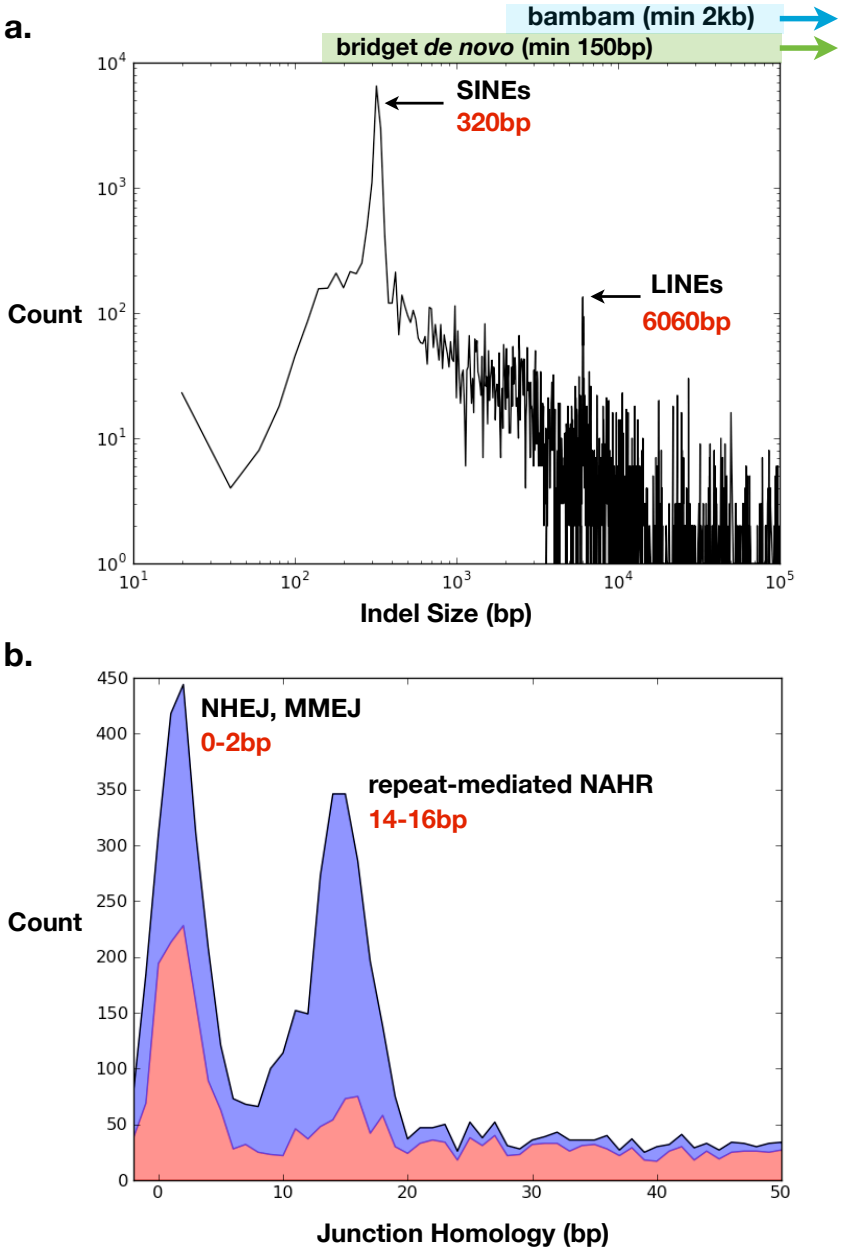

Figure 3.3: Bridget *de novo* results for 13 whole genome GBMs and their matched-normals. (a) Histogram of indel size showing distinct peaks for insertions and deletions corresponding to SINE and LINE families of repetitive elements. (b) Junction homology for germline (blue) and somatic (red) breakpoints showing peaks that indicate at least two primary mechanisms for breakpoint induction and repair: microhomology-mediated end-joining (MMEJ), non-homologous end joining (NHEJ) and repeat-mediated, non-allelic homologous recombination (NAHR). Note the enrichment of MMEJ & NHEJ breakpoints among the somatic breakpoints.

structural variants identified through sequencing-based methods (top 4 categories). The array-based variants tend to be more distant from their nearest Bridget breakpoint (i.e. $D(b_{bridget}, b_{dgv}) >> 10\text{bp}$), but this should be expected as the accuracy of structural variants inferred from an array is limited by the density and separation of its probes on the reference genome.

Similarly, the sequencing-based variants in DGV that are distant ($>> 10\text{bp}$) from the nearest Bridget solution can be explained by the fact that DGV does not require structural variants to have base-level precision. Therefore, DGV accepts structural variants identified by paired-end clustering methods like BamBam's structural variant caller, which are shown to have reduced precision compared to methods, such as Bridget, that incorporate split reads.

In both of the above cases where the two solution are distant, the nearest breakpoints identified by Bridget *de novo* to the sequencing- or array-based counterparts stored in DGV are likely refinements of the variant provided by DGV. In most cases, the Bridget *de novo* solutions appear to improve the knowledge of the size and location of these indels in the human reference to base-level precision.

3.7 Tumor Genome Browser

To visualize all of the results output by BamBam, a tumor genome browser was developed that simultaneously displays all of the genomic variants found in a single tumor sample, versus its matched-normal, as shown in Figure 4. The browser was built

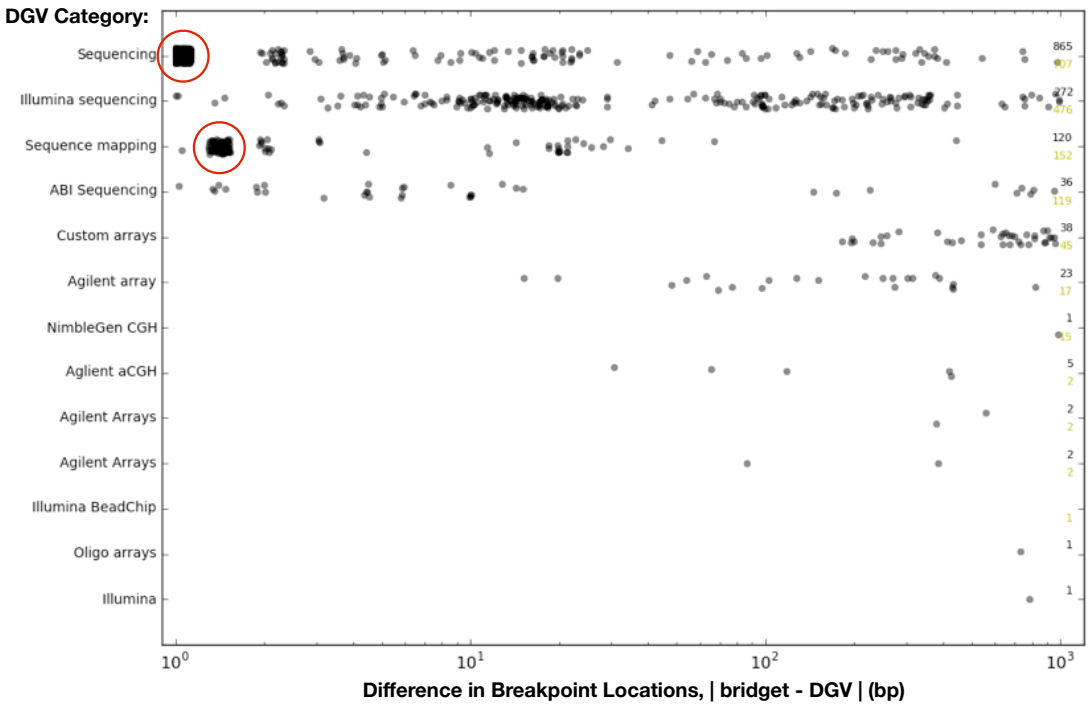

Figure 3.4: Distance between Bridget-refined breakpoints vs. breakpoints available in the Database of Genomic Variants (DGV). Each dot represents a single breakpoint found by Bridget, where its location on the x-axis is defined by $D(b_{bridget}, b_{dgv}) + 1.0$ between the Bridget solution and the closest breakpoint in DGV. A distance of 1.0 is a perfect score to place it on a log scale. The cluster at 2bp is due to a difference in how DGV stored breakpoints belonging to "Sequence mapping" category, utilizing a one-based genome coordinate system instead of the zero-based coordinates that Bridget reports. Also, not that some studies such as those categorized as "Illumina sequencing" have higher differences, with a peak around 12bp. Those particular studies did not incorporate split read methods to refine their breakpoints. Similarly, breakpoints derived from chip-based assays have much higher differences from the closest Bridget solution, as expected.

using the same codebase that powers the UCSC Genome Browser. I wrote a custom drawing engine and MySQL database designed for rapid extraction and display of all BamBam results for a single patient. It is capable of displaying overall & allele specific copy number, intra- and inter-chromosomal rearrangements, and mutations and small indels. It displays data in both linear and circular genome plots, the latter being much better suited for displaying inter-chromosomal rearrangements.

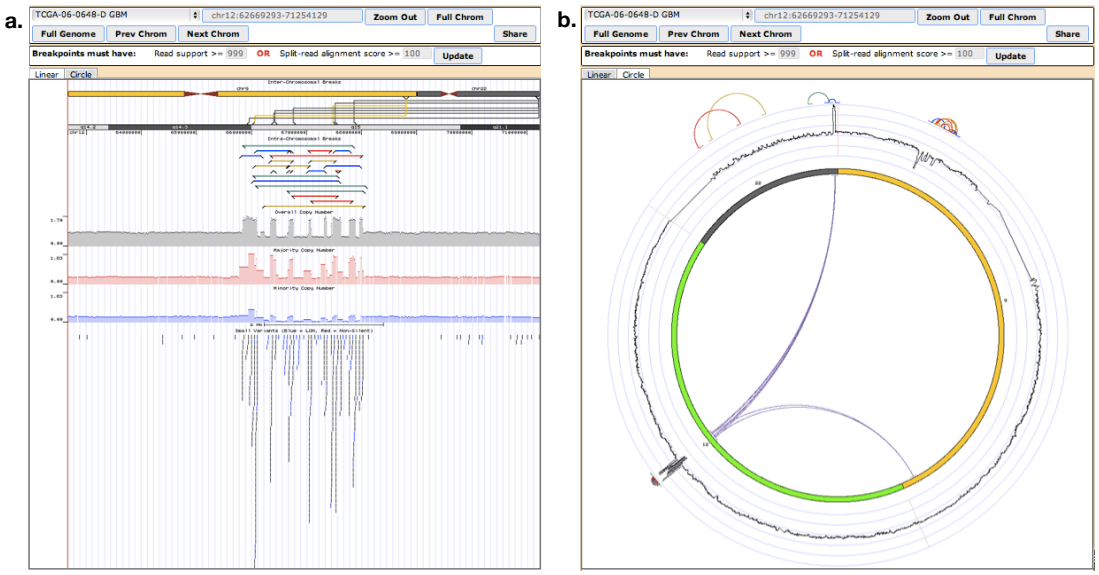

Figure 3.5: UCSC Tumor Genome Browser. The browser shows all of the high-level somatic difference discovered by BamBam in a single image, enabling the synthesis of multiple distinct datasets to give the user an overall picture of the tumor's genome. The browser is able to zoom into and out of genomic regions rapidly, going from the full genome view, as shown above, to a single base resolution in just a few clicks. (a) Linear view of a highly-rearranged and amplified region on chromosome 12 in GBM-06-0648. (b) A circular genome view of the same region showing how the amplified peaks on chromosome 12 are connected to an amplified region of chromosome 22. It also shows a set of clustered rearrangements on chromosome 9 that caused the homozygous deletion of CDKN2A/B.

By displaying the data together in a single image, the user can quickly navigate

a single sample's data and understand the relationship between changes in its copy number and local structural variation. For example, a large intra-chromosomal, deletion-type rearrangement should have a concordant drop in copy number in the region between the breakpoint ends. Also, displaying mutation data with copy number data allows the user to understand if a somatic mutation was subsequently amplified, or if the wild-type allele was deleted in the tumor, both vital datapoints that may suggest the importance of the event in this sample's tumorigenesis.

The ability to visualize and navigate a patient's data within the genome browser has been instrumental to many of the discoveries presented in this thesis. While the browser is still actively used by few members of the cancer group at UCSC, it is such a valuable tool that it was one of the first things we recreated (and drastically improved) at my startup, Five3 Genomics, which is used to produce all browser images presented below.

Chapter 4

Integrating sequencing results to locally reconstruct the tumor genome

As the tumor genome browser demonstrates, the true power of sequencing data comes from when the various data, structural rearrangements, copy number, and mutations, are thoughtfully integrated to form a more complete understanding of the tumor genome. Through this understanding, we can begin to piece together the events that characterize the genesis and subsequent development of individual tumors.

4.1 Breakpoint traversal to infer local tumor genome assemblies

Similar to recently published methods, we construct a breakpoint graph by describing a set of edges that represent the amplified segments of the tumor genome and a set of directional vertices that connect the edges to one another [30, 50]. Here, the edges are defined as the amplified segments of the tumor genome observed in the relative copy number, while the vertices are the highly supported breakpoints found in the manner described above. If an amplified segment is interrupted by a breakpoint, then that segment will be split into two edges at the location of the interrupting breakpoint.

With the segments laid out according to genomic position, we determine the arrangement of edges that represent the rearranged tumor sequence by starting at the leftmost position of the first amplified segment and progress towards the right until a right-oriented vertex is encountered. The path continues by following the vertex to the segment it is connected to and moving in the direction (left or right) specified by the outgoing vertex. A complete solution to the breakpoint graph is made when a path through all edges and vertices have been traversed at least once, satisfying the observed copy number.

Given such loose constraints, it is clear that many satisfactory solutions to the breakpoint graph may be possible. However, the optimal path(s) through the graph will be the one(s) whose inferred copy number most closely agrees with the observed relative copy number produced by BamBam. The number of times a solution traverses a given

segment produces an inference of that segment's copy number. The root mean square deviation (RMSD) of the segment traversal counts to the observed relative copy number for each solution is calculated, and then the solution(s) with the smallest RMSD value are labeled as optimal.

Two toy example breakpoint graphs and their respective solutions are shown in Figure 4.1. Both examples in Figure 4.1 shows a set of breakpoints, the relative copy number between tumor and matched-normal, and respective solutions to the breakpoint graphs. The first example is a simple one, with four breakpoints and three discrete copy number states ranging from homozygous deletion to single-copy gain. The solution presented for this example incorporates two paths that utilize all breakpoints and explain the observed copy number. However, this is not the only solution to this breakpoint graph, as the smaller of the two blue deletions could instead be phased with the inverted breakpoints with no penalty.

The second example in Figure 4.1 describes a slightly more complicated breakpoint graph where a clustered set of five breakpoints lie on the boundaries of regions that are "extremely" amplified, i.e. relative copy number ≥ 10. Here, the simplest solution to this breakpoint graph involves a circular path through the five breakpoints and amplified regions. To account for the extra copies observed in the copy number data, the circular path must be traversed a total of nine times.

In Section 4.3, we describe solutions to breakpoint graphs from real GBM tumors with qualities similar to both of the examples presented in Figure 4.1.

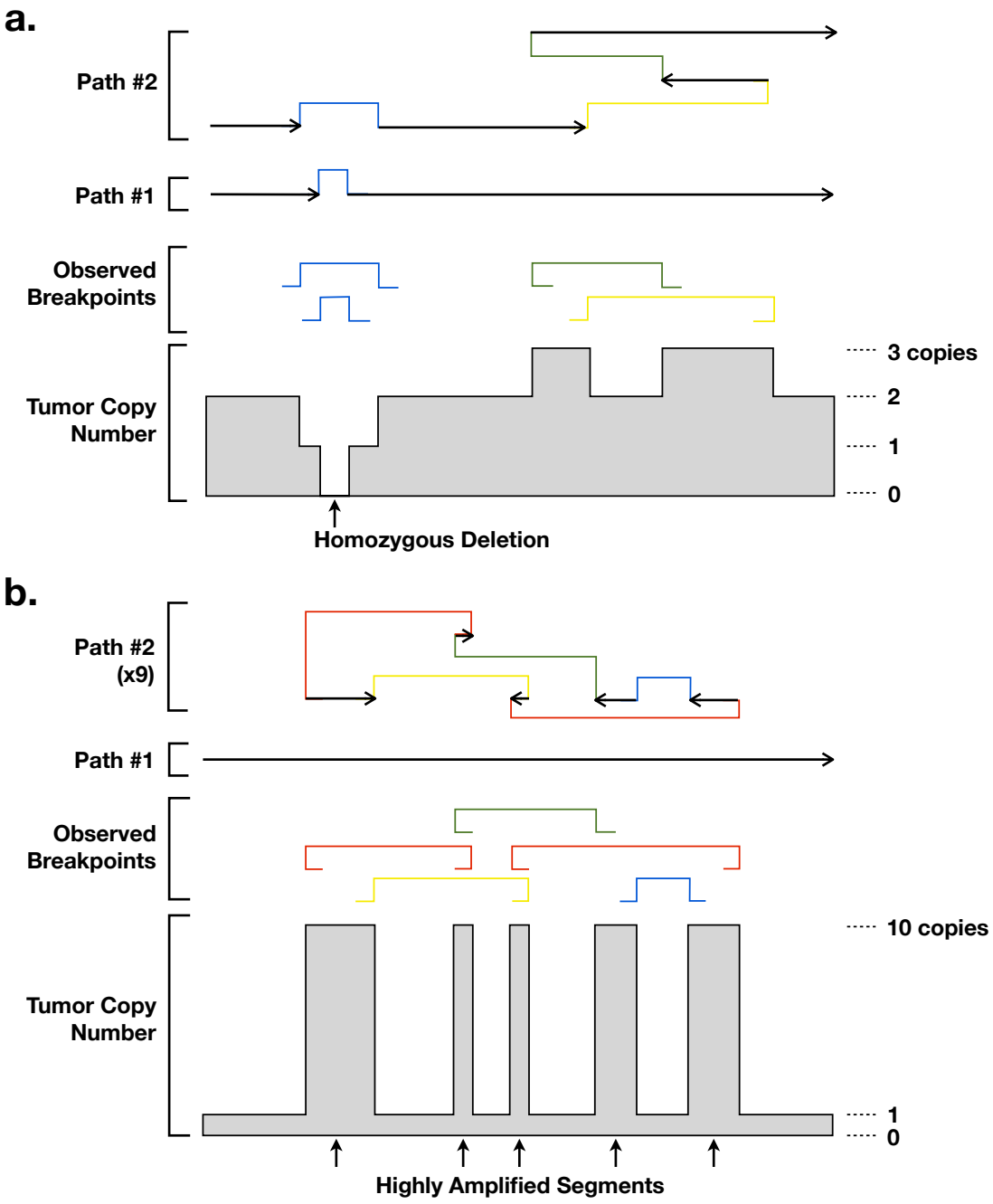

Figure 4.1: Toy examples of breakpoint graphs and their solutions. (a) One possible two path solution that accounts for all observed breakpoints and tumor copy number. Note that the deletions in both paths could be swapped to create another equally valid solution to the breakpoint graph. (b) A circular solution to the breakpoint graph for a highly-amplified region. The circular nature of the solution may indicate the presence of a double minute chromosome and/or a homogeneously-staining region (HSR).

4.2 Breakpoints associated with highly amplified regions

The support of a given breakpoint is directly proportional to the copy number of the regions it connects. Thus, by requiring breakpoints to have a high level of read support, we can filter out breakpoints that are part of copy-neutral rearrangements or lead to low-copy amplifications and deletions, and instead focus on the breakpoints that are likely linked to the highly amplified regions in the tumor. The particular read support threshold is chosen such that breakpoints that have read support expected of copy-neutral regions of the tumor genome are removed.

4.3 Results

The true test of the efficacy of any computational method is how it performs on real data. Here we demonstrate the use of the methods described above on real tumor whole genome sequencing data, creating precise, localized genome assemblies that integrate the observed copy number and breakpoints produced by BamBam. From these assemblies, we can infer the structural changes that occurred and possibly discover the events instrumental to the tumorigenesis of these cancers.

4.3.1 Homozygous deletion of CDKN2A in glioblastoma multiforme

In Figure 4.1, two types of breakpoint graphs, one linear and another circular were presented. This section presents examples of the first type of breakpoint graph in a set of glioblastoma multiforme (GBM) tumors and their matched-normals sequenced by

TCGA. Using the breakpoints and copy number identified by BamBam and Bridget, we focus our study on the homozygous deletions of tumor suppressor CDKN2A, a common occurrence in cancer and notably GBM tumors [60, 56]. CDKN2A inhibits CDK4, arresting cell-cycle progression into G1 and G2 phases. Tumor cells that delete both copies of CDKN2A have vastly increased abilities to proliferate [58, 69].

The chromosomal regions containing CDKN2A are shown for three GBM tumors in Figure 4.2. In each case, the structural events by which the copies of CDKN2A were deleted are different. In Figure 4.2(a) and (b) both show one copy of CDKN2A deleted through a focal intra-chromosomal rearrangement, while the other copy was loss via arm-level deletion that resulted from the translocation of the downstream arm to chromosomes 11 and 22, respectively. The case presented Figure 4.2(c) is complicated by a greater number of intra-chromosomal rearrangements, but it appears that one copy was focally deleted via an intra-chromosomal deletion (upper-left deletion breakpoint colored blue) similar to that presented in (a) and (b). The other copy was deleted via by the complicated rearrangement formed by the remaining breakpoints. The clustered pattern of breakpoints and oscillating nature of copy number and allele fraction is suggestive of chromothripsis, an event where catastrophic genomic disruption creates highly rearranged regions [82, 40].

Table 4.1 lists the status of both copies CDKN2A in a set of 17 TCGA GBM tumors that we have whole genome sequencing data. 12 of the 17 GBMs experienced homozygous loss of CDKN2A, while 5 GBMs retained both copies. The general pattern of loss, shared by all but one GBM sample, is that one copy of CDKN2A was loss via a

small focal deletion and the second copy was deleted via a much more dramatic deletion, i.e. loss of chromosomal arm chr9p, complete loss of chr9, or a set of chr9p deletions caused by a chromothripsis-like event. This pattern shared by multiple tumors allows us to speculate on how these events may have transpired in these tumors.

Similar to how we use recurrent mutations to infer oncogenic importance, the recurrence of focal and large-scale losses of CDKN2A might suggest an importance in the relative timing of the two events. With this assumption, there are two possibilities: (1) focal followed by large-scale loss or (2) large-scale followed by focal loss. In both cases, the loss of the first copy of CDKN2A results in cells with reduced abilities to stop proliferation, helping these cells survive to the next generation. If we assume then that these burgeoning tumor cells are under reduced pressure to maintain their genomic stability, that either loss will be tolerated at first. However, considering that cells become inviable if they lose both copies of "housekeeping" genes, if a cell loses its first copy of CDKN2A via loss of the arm or all of chromosome 9, it cannot lose the second copy by the same means. Only a subsequent focal loss of CDKN2A will produce viable cells that survive to form the tumor. The same argument can be made for the focal losses occurring first, but since we see so few examples of double focal losses suggests that focal losses are less likely to occur first. Of course, all of the above is pure speculation, without a method to directly infer the timing of the loss of CDKN2A, we cannot know which came first. The next chapter of this thesis will describe just such a method.

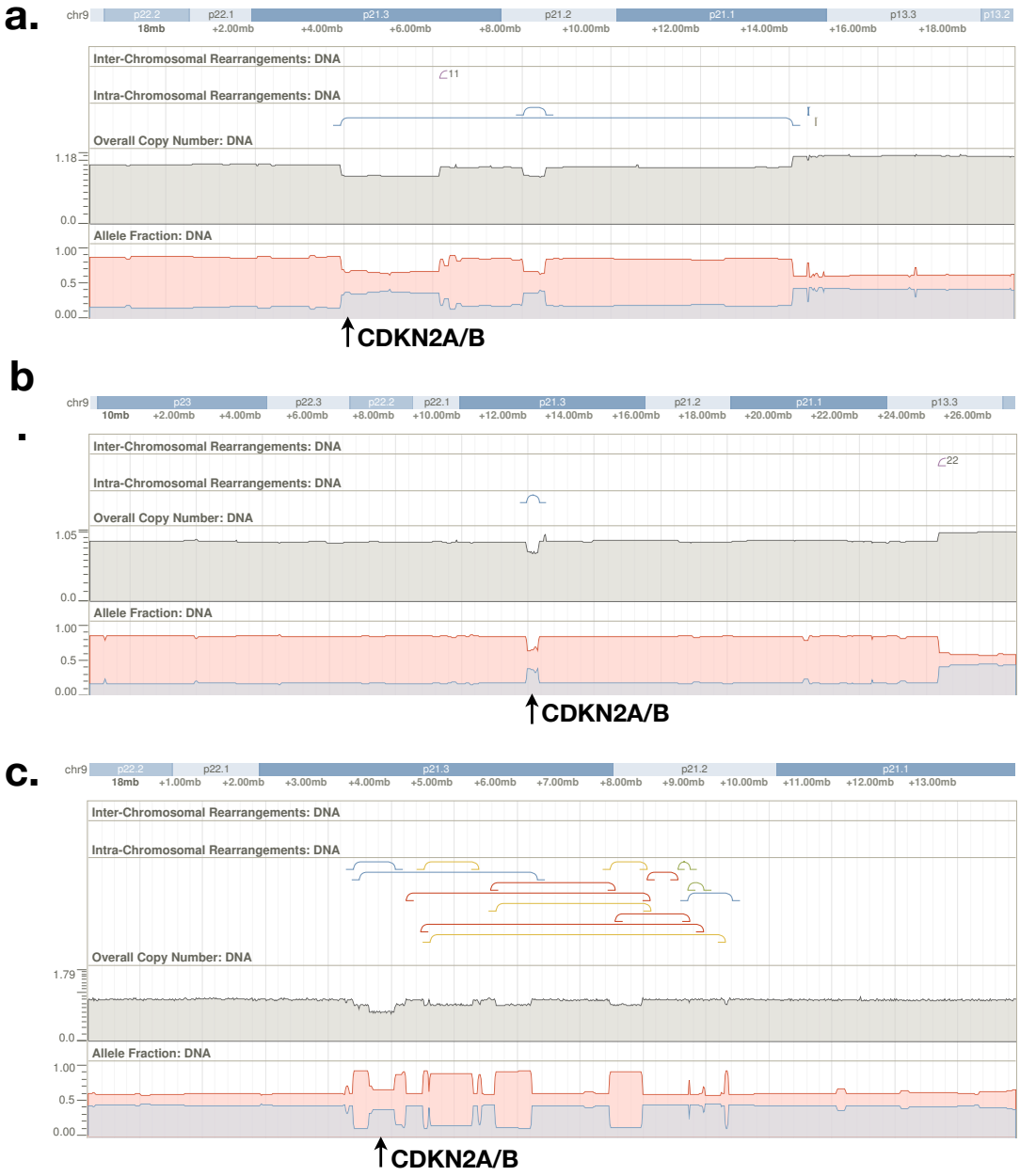

Figure 4.2: Browser shots demonstrating how CDKN2A was homozygously deleted in: (a) GBM-06-0188, (b) GBM-06-0145, and (c) GBM-06-0648. Inter- and intra-chromosomal rearrangements caused both copies of CDKN2A to be deleted in (a) and (b), while a clustered set of intra-chromosomal rearrangements deleted both copies in (c).

Table 4.1: Types of CDKN2A Loss in 17 TCGA GBMs.

Sample	**CDK2NA Copy #1**	**CDKN2A Copy #2**
GBM-06-0145	focal loss	chr9p loss via translocation
GBM-06-0188	focal loss	chr9p loss via translocation
GBM-06-0155	focal loss	chr9p loss via translocation
GBM-06-0185	focal loss	chr9p loss via translocation
GBM-06-0214	focal loss	chr9p loss via translocation
GBM-14-0786	focal loss	chr9p loss
GBM-14-1401	focal loss	chr9p loss
GBM-16-1063	focal loss	chr9p loss
GBM-06-1086	focal loss	complete loss of chr9
GBM-06-0208	focal loss	complete loss of chr9
GBM-06-0648	focal loss	chromothripsis-related focal loss
GBM-06-0877	focal loss	focal loss
GBM-06-0152	*No loss*	*No loss*
GBM-06-0881	*No loss*	*No loss*
GBM-14-1454	*No loss*	*No loss*
GBM-16-1460	*No loss*	*No loss*
GBM-26-1438	*No loss*	*No loss*

4.3.2 Reconstructing highly amplified regions using whole genome sequencing

Next, we will apply the above methods on three GBM tumors from the TCGA project, GBM-06-0648, GBM-06-0152, and GBM-06-0145, and attempt to reconstruct regions in each tumor that are found to be extremely amplified. In each case, we find circular solutions to their breakpoint graphs that could indicate the presence of double minute chromosomes (DMs), self-replicating, circular chromosomes that are often found in high copy number in tumors, conferring drug resistance and harboring oncogenes drive tumorigenesis and growth [7, 65, 70, 7, 6]. By using Bridget-refined breakpoints, the circular assemblies presented in this section are accurate to the base-level.

Double Minute Assembly in GBM-06-0648

A total of 3,696 breakpoints were identified in GBM-06-0648, of which 132 breakpoints were found to have split reads directly spanning the putative breakpoint. Using a minimum read support threshold of 100 filtered all but 16 of the somatic breakpoints with split read support were removed. As shown in Figure 4.3, each of the remaining 16 breakpoints is proximate to the boundaries of highly amplified segments in a clustered region of chromosome 12 and a small 250bp region on chromosome 9. A total of 16 amplified segments are found, total size approximately 891kb, where one segment contains the known oncogene MDM2 implicated in GBM tumorigenesis. In fact, the edges of every amplified segment can be uniquely associated with a single breakpoint oriented such that it enters into or exits from the amplified region. This suggests that the highly supported breakpoints and amplifications represent the rearranged configuration of this region in the tumor's genome.

Figure 4.4(a) represents the single optimal path through the amplified segments discovered by walking the breakpoint graph. The interesting aspect of this solution is that it is circular, Figure 4.4(b), where the final traversal of last segment returns back to the initial position, i.e. the first vertex is also the final vertex. This circular solution passes through every amplified segment exactly once, yet the copy number suggests that the average tumor cell contains multiple copies of each segment. This high-multiplicity circular solution suggests two possibilities: (1) multiple copies of these 16 segments are inserted into the tumor genome in tandem such that each copy maintains the precise

order and orientation described by the circular solution (e.g. homogeneously-staining region, or HSR), or (2) a single 891kb double minute chromosome containing one copy of each of the 16 segments may have been formed of which the average tumor cell has accumulated multiple copies. Note that these two possibilities are not mutually exclusive, as highly amplified regions are known to exist in a state of flux alternating between HSRs and DMs [84].

The assembly of the rearrangements specific to the contiguous, non-DM-associated segments of chromosome 12 is shown in Figure 4.4(c), meaning that the entire structure of chromosome 12 can be explained by the breakpoints and copy number results produced by BamBam and Bridget. The observation that deletions surround some of the amplified peaks suggest that these regions were excised from the genome by some process (e.g. chromothripsis) and circularized to form the double minute, in what is termed as the "episome model" of double minute creation that is confirmed in both hematological and solid tumors [14, 83, 84]. The putative DM contains the oncogene MDM2, which represses the tumor suppressing activity of TP53 [16].

In the circular assembly of GBM-06-0648, we observe that one of the DM-associated breakpoints is located within the final intron of the gene CPM. We may expect that a rearrangement interrupting a gene would severely impact the expression of its transcript. Fortunately, the Broad Institute had sequenced the RNA of this same tumor, providing us the ability to measure the expression of CPM. Here we find something unexpected, not only do we find that CPM is expressed, we find that reads mapped to the penultimate exon are connected to a region that is directly downstream

of CPM *only* in the circular assembly of GBM-06-0648's double minute. This provides yet further support of the veracity of the circular assembly, as this region is 1,470kb distant from the CPM exon in the human reference, but only 13.5kb away in the circular assembly. The truly interesting aspect to this discovery is that the region connected to the penultimate exon of CPM is a completely novel exon, complete with a canonical splice site and terminating stop codon, birthed out of normally unexpressed intergenic DNA. A BLASTp search of this novel exon's 25 amino acid protein sequence did not find any significant hits, so the functional impact of this novel exon to the activity of CPM is currently unknown [2].

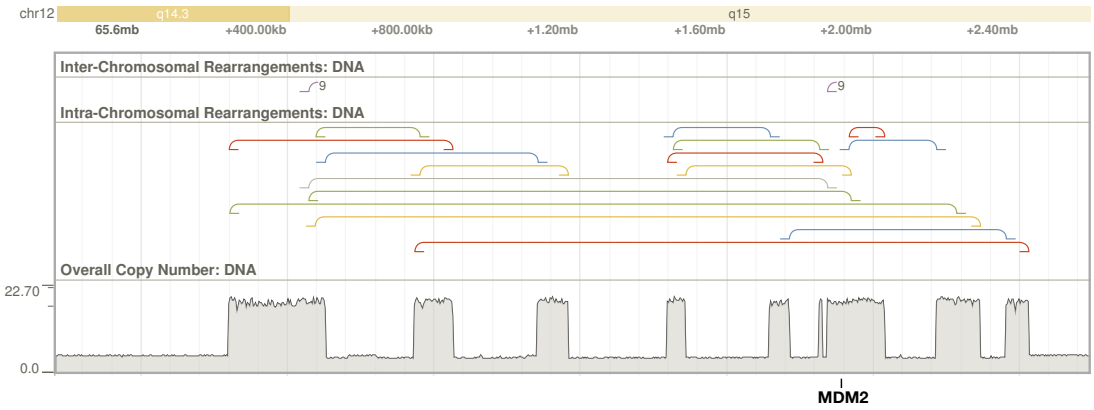

Figure 4.3: Browser shot of the inter- and intra-chromosomal, highly-supported breakpoints and copy number observed for TCGA sample GBM-06-0648.

Assembly of two independent double minutes in GBM-06-0152

A total of 1,669 somatic breakpoints were identified in GBM-06-0152, of which 120 breakpoints were found to have split reads directly spanning the putative breakpoint. 22 breakpoints surpassed the minimum read support threshold of 100 to identify

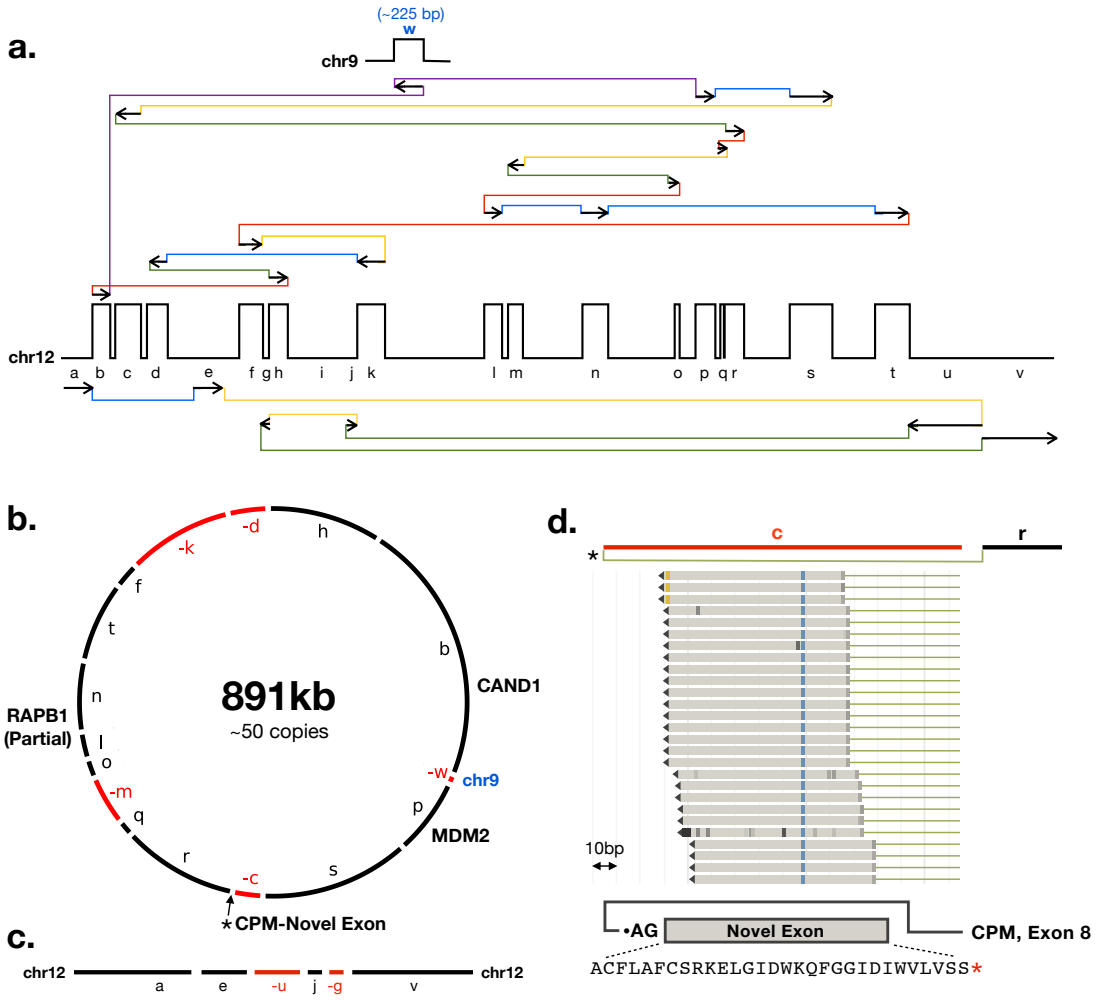

Figure 4.4: (a) Diagram of the circular solution to breakpoint graph for GBM-06-0648 that accounts for all observed, highly-supported breakpoints, as well as the solution for the remaining, un-amplified segments of chromosome 12. (b) Circular assembly of the circular solution, possibly indicating the presence of a double minute chromosome. Note the positions of MDM2 and the CPM-Novel Exon gene. (c) The linear solution of the remaining contiguous segments on chromosome 12. (d) RNA-Seq data for GBM-06-0648 that indicates a novel exon replaced the final exon of CPM in the circular assembly.

breakpoints likely associated with highly-amplified regions, including two breakpoints that did not have split read evidence. All 22 of these highly supported breakpoints are proximate to the boundaries of 29 highly amplified segments in clustered regions on chromosome 7 and chromosome 12, as shown in Figure 4.5, suggesting that the highly supported breakpoints and the amplifications are directly associated and may represent the rearranged configuration of one (or multiple) amplicons in the tumor's genome.

Figure 4.6 shows the solution to the breakpoint graph of GBM-06-0152 that predicts two independent paths that completely account for all observed breakpoints and amplified segments. Path A has a total length of 1,269kb, uses 18 of the 20 intra-chromosomal breakpoints on chromosome 12, and contains one copy of CDK4 and two copies of the oncogene MDM2. CDK4 interacts with CDKN2A to control cell-cycle progression in the G1-S phase, and is known to enhance tumor progression when over-expressed [72, 52]. Path B has a total length of 929kb, incorporating the remaining two intra-chromosomal breakpoints on chromosome 12 and the two inter-chromosomal breakpoints to connect the amplified segments on chromosome 12 to the amplified segment containing EGFR on chromosome 7. Both paths are circular in nature and suggest that two independent oncogenic DMs were created and amplified in this tumor. The observation that the EGFR-DM contains two amplified segments within a cluster of rearrangements specific to the MDM2-CDK4-DM suggests that the creation of these two DMs were coincident.

To incorporate all highly supported breakpoints in the MDM2-CDK4-DM, multiple traversals of some segments were required. 11 of the 29 segments were traversed

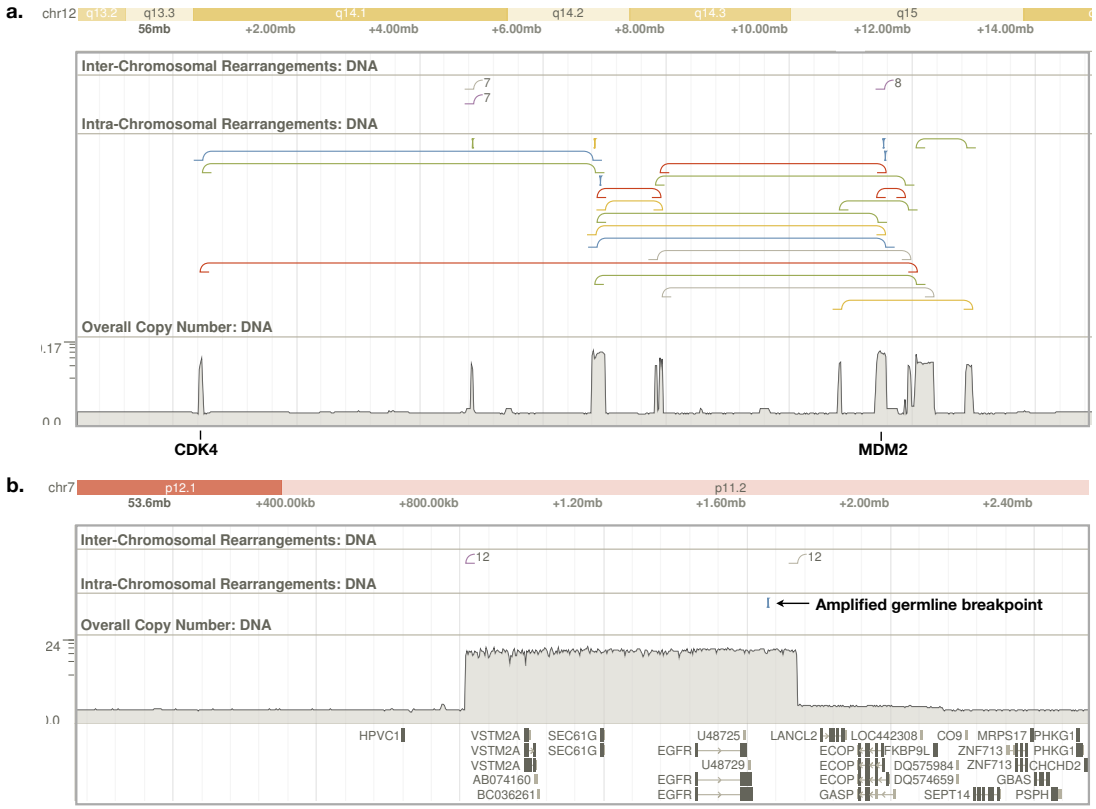

Figure 4.5: Browser shot of the inter- and intra-chromosomal, highly-supported breakpoints and copy number observed for TCGA sample GBM-06-0152 for (a) chromosome 12 and (b) chromosome 7.

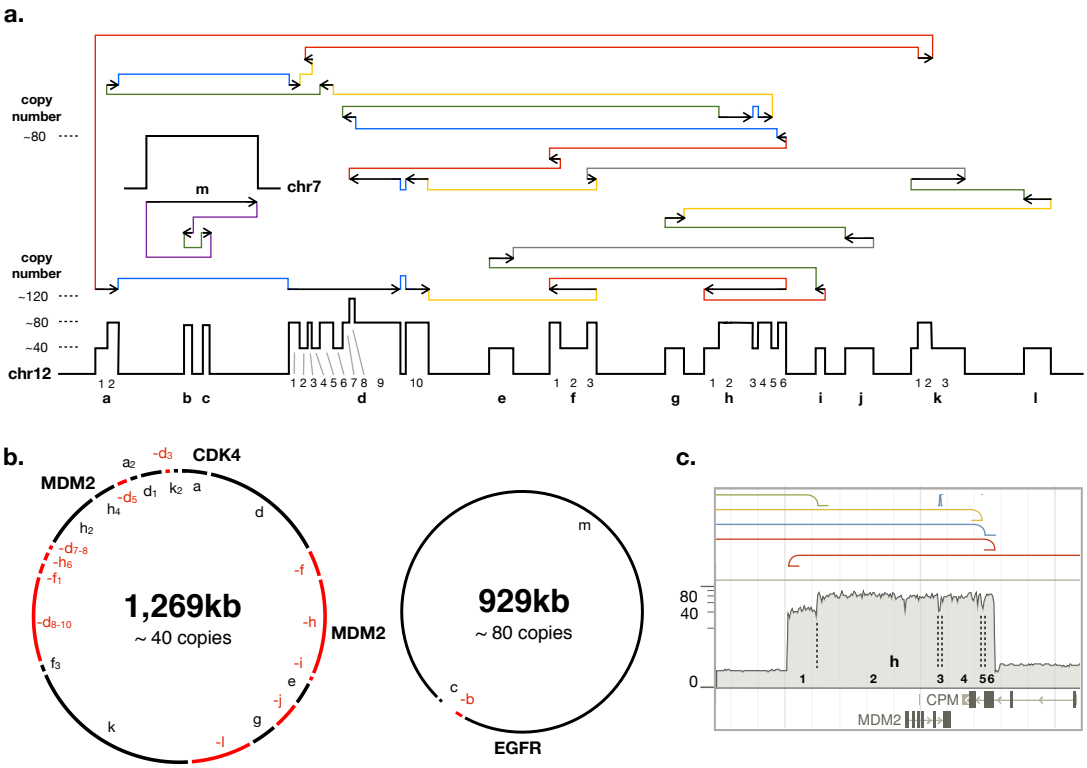

Figure 4.6: (a) Diagram of the two circular solutions to breakpoint graph for GBM-06-0152 that accounts for all observed, highly-supported breakpoints. (b) Circular assemblies of the two circular solutions, possibly indicating the presence of two independent double minute chromosomes. The larger circular solution has two copies of MDM2 and one copy of CDK4, while the smaller solution has EGFR. (c) A zoomed-in browser shot of the amplified segment labeled h, showing the distinct changes in copy number as described in (a).

two times, and one segment was traversed three times. The complex rearrangements featured in the MDM2-CDK4-DM are especially noteworthy in that the multiple-traversed segments feature genomic regions connected in ways that are found only in the putative DM. This suggests that the solution shown here may not represent the initial configuration of this sample's DM, but rather represents the final result of an evolutionary process beginning with the initial (unobserved) DM featuring, at minimum, a single copy of each amplified segment observed in the MDM2-CDK4-DM. Subsequent rearrangements of the initial DM then led to these initial segments being amplified or deleted, where the more "successful" DM configurations, such as the configuration observed with two copies of the MDM2 oncogene, would rapidly sweep through the emerging tumor population. Since the expected increase in copy number was also observed for the multiply traversed segments, the solution presented here likely represents the dominant DM configuration in the GBM-06-0152 tumor, as shown in Figure 4.6(c).

EGFR Amplification in GBM-06-0145

The GBM tumor sample GBM-06-0145 presents both a simple and complicated story. As shown in Figure 4.7, the amplicon containing EGFR has what appears to be uniformly high copy number and we find an extremely well-supported breakpoint (breakpoint support = 2,833 reads) bounding this amplicon. Upon first glance, this data would suggest that this amplicon exists as a simple circular DM as shown in the previous two tumors, or in a tandem array of duplicates.

However, the browser shot shows 6 additional somatic breakpoints with read

support > 100 and the copy number data appears to subtly change near the ends of each breakpoint. A diagram of this same data is shown in Figure 4.8. The solution to the breakpoint graph shown in Figure 4.8(a) shows the possibility of three distinct paths that incorporate all breakpoints and explain the changes in copy number. All paths utilize the bounding tandem-duplication-like breakpoint, where path A is characterized by a simple direct path through the amplicon, path B is a similarly simple path that features a small deletion, and the final complicated path C that incorporates the remaining breakpoints. According to the baseline copy number, path A appears to be the dominant path, found in a ratio of 7 to every 9 copies of the amplicon (77%).

Figure 4.8(b) shows how the structure of the EGFR gene is affected in each path. Since path A features no internal breakpoints, EGFR is intact. On the other hand, paths B and C alter the structure of EGFR. Path C deletes the first exon of EGFR, likely leading to a non-functional version of EGFR. More interesting is path B, whose internal deletion specifically removes exons 2-7 of EGFR, creating a highly oncogenic, constitutively active form of EGFR called EGFRvIII that is found expressed in multiple tumor types [55]. This is interesting since it suggests two scenarios: (1) EGFRvIII emerges after wildtype EGFR is significantly amplified or (2) EGFRvIII is created early but cells with more copies of wildtype EGFR have a selective advantage in the tumor population. The former scenario seems most plausible, as the increased number of copies subsequently improves the chance that the EGFRvIII mutant will arise. Regardless of which scenario is correct, the ratio of EGFRvIII to wildtype EGFR suggests that high copy number of oncogenic EGFRvIII may not be necessary to provide

significant advantage to the growing tumor cell. It would be useful to compare the relative expression of EGFRvIII to EGFR transcripts using RNA-Seq data, if it becomes available for this sample, to determine if EGFRvIII is preferentially expressed over the multiple copies of wildtype EGFR present in this genome.

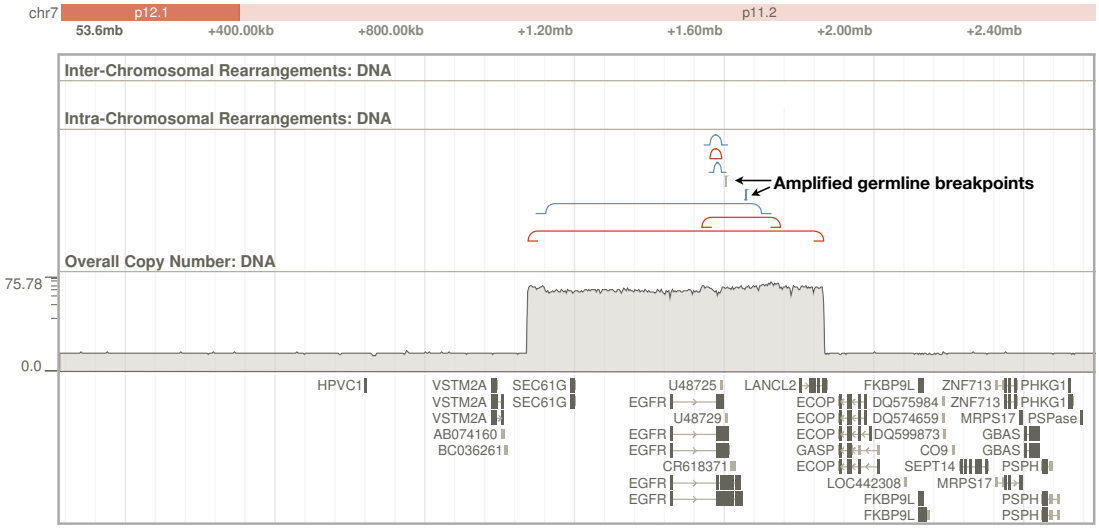

Figure 4.7: Browser shot of the intra-chromosomal, highly-supported breakpoints and copy number observed for TCGA sample GBM-06-0145 for chromosome 7.

4.3.3 FISH confirmation of double minutes

For two tumors, GBM-06-0648 and GBM-06-0145, Fluorescence *in situ* hybridization (FISH) was performed to determine the physical locations of the amplified regions in these tumors. Depending on the pattern of the amplified probes presented by the FISH images, extreme amplification via DMs and/or HSRs can be determined.

Figure 4.9 show the FISH results for GBM-06-0648 and GBM-06-0145, using probes targeting sequences in the amplified and nearby, non-amplified regions of the re-

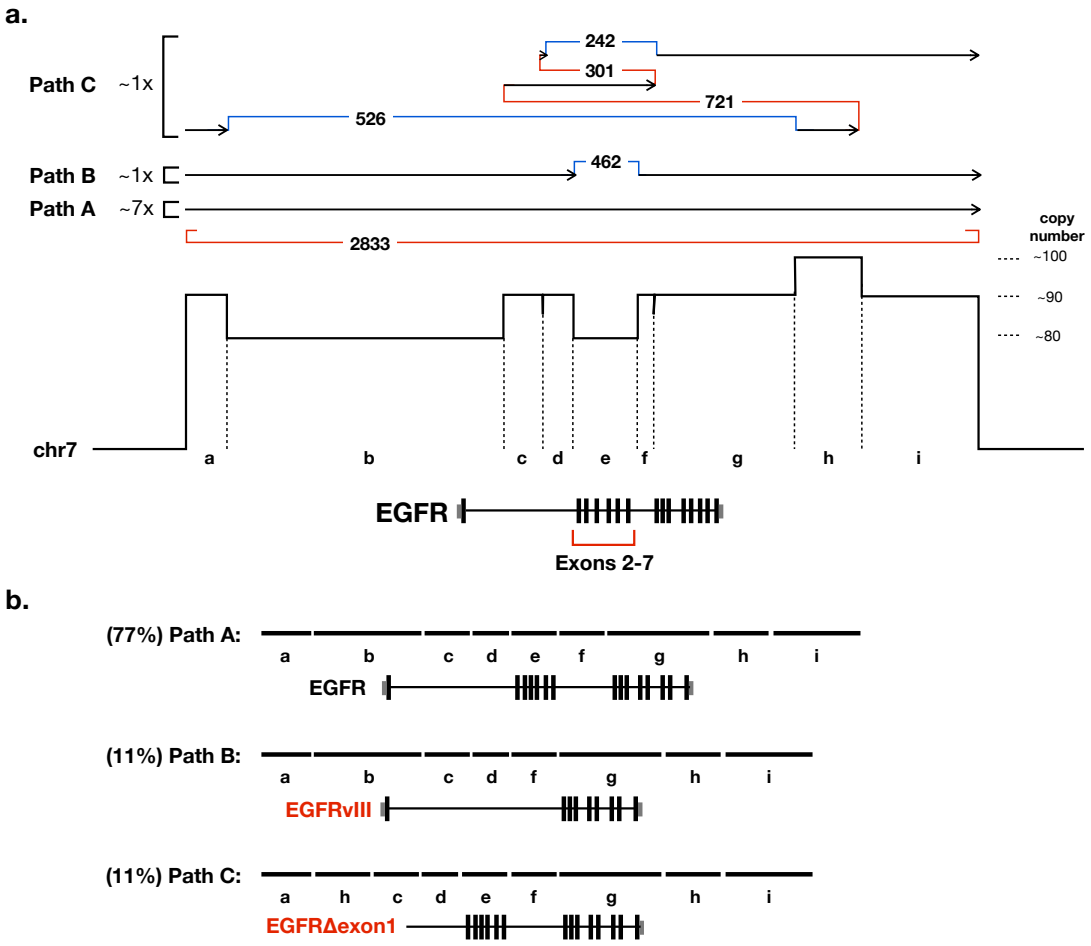

Figure 4.8: (a) Diagram of the three paths that together form a potential solution of the breakpoint graph for sample GBM-06-0145 that accounts for all observed, highly-supported breakpoints. The location of the gene EGFR is noted. Note that all paths, A, B and C, can create circular solutions by using the encompassing tandem duplication shown in red to connect the end of each path to the beginning. The tandem duplication may also connect path A to path B, path C to path B, etc. (b) The effects of each path on the EGFR gene, showing that path B creates an oncogenic form of EGFR, EGFRvIII, through specific deletion of exons 2-7.

spective tumors. For GBM-06-0648, the amplified probe (colored red) targeted MDM2, while the non-amplified probe (colored green) targeted the centromere of chr12. These two probes would show the locations of these two regions in the genome, which for normal cells would show both probes to be in single copy per chromosome and linearly paired with each other (assuming cells are in interphase). From these results, the technicians at the cytogenics laboratory of the Memorial Sloan-Kettering Cancer Center (MSKCC) who produced this FISH data, confirmed the presence of the MDM2 double minute for GBM-06-0648, since the FISH image for this tumor exhibits the characteristically distributed, punctuate pattern of an amplification by double minute.

The story is less clear for the EGFR amplification in GBM-06-0145. For this FISH experiment, the probes targeted the amplified EGFR (red) and the centromere of chr7 (green). Similar to sample GBM-06-0648, the amplified probe is not found in single copy or found proximate to the non-amplified probe on chromosome 7, indicating that the extra copies of EGFR were not caused by simple tandem duplication at the reference location of EGFR. The technicians at MSKCC believe its pattern is primarily HSR-like, as most cells exhibit dense clusters of the amplified probe, but it is clear from this slide that there are cells that also display the punctuate pattern of a DM. It is known that HSRs often spawn DM, and vice versa, so this may indicate that the amplified state of the EGFR is in flux, alternating being present in HSR and DM form from cell to cell [84]. The fluctuating status of this amplicon may also explain why the breakpoint graph for GBM-06-0145, as shown in Figure 4.8, was more complicated than the other examples.

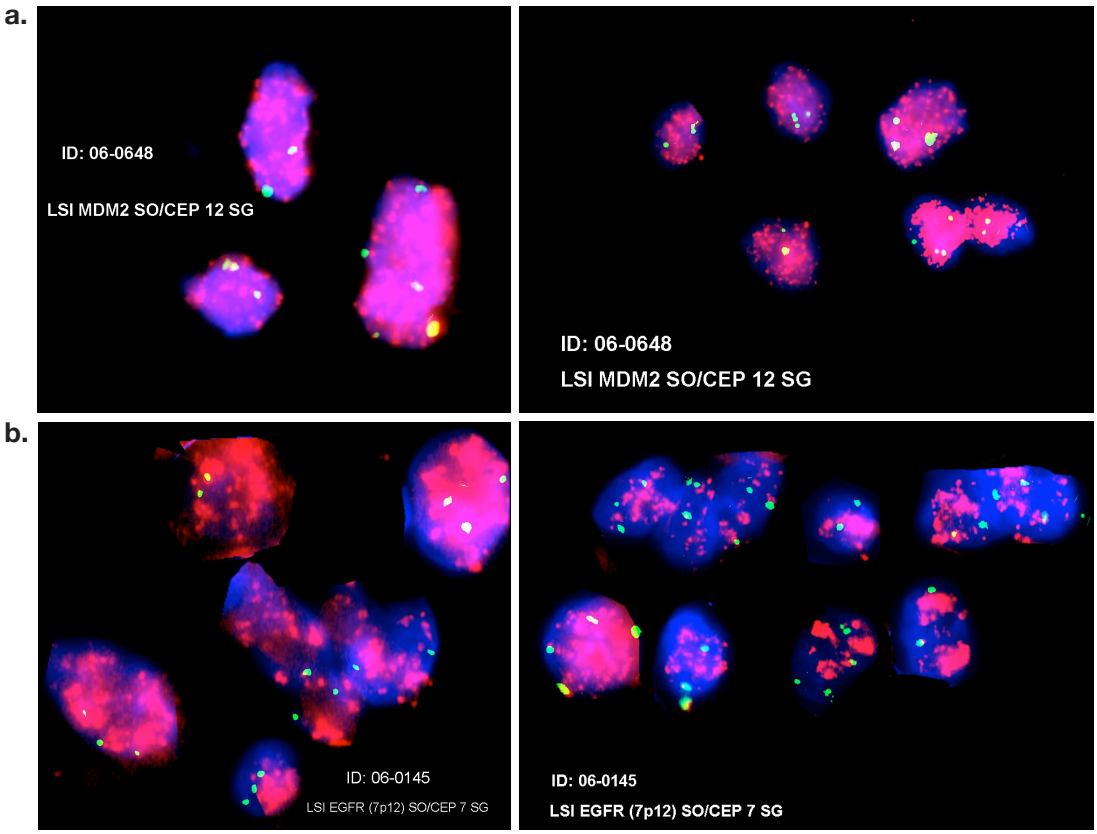

Figure 4.9: Fluorescent *in situ* hybridization (FISH) images for samples used to determine the physical state of the amplified regions. (a) FISH image for GBM-06-0648 using probes for MDM2 (red) and the centromere of chromosome 12 (green). Note that the improved focus of the image on the right displays the punctuate pattern in red, characteristic of a DM, confirming that MDM2 is amplified by DM, as predicted. (b) FISH image for GBM-06-0145 using probes for EGFR (red) and centromere of chromosome 7 (green). Note that the red probes in most cells are clustered together, as opposed to the pattern seen in (a), indicative of an HSR. However, there are cells in (b), such as the bottom-left cell in the right image of (b), that display a DM-like pattern similar to that observed in (a). This suggests that, while the dominant pattern indicates EGFR amplification via HSR, the EGFR amplicon may also take the form of double minute chromosomes.

4.3.4 Evidence of tumor-specific reads in blood

Summarized in Figure 4.10 is the evidence found of tumor-specific breakpoints in the matched-normal sequencing data of the three GBM tumors discussed above. In both GBM-06-0152 and GBM-06-0648, we find significant evidence for the majority of DM-specific breakpoints in the matched-normal (blood) from the same patients, an example of which is shown in Figure 4.11. In GBM-06-0152, 11 of the 20 breakpoints have at least one supporting read in blood, while 9 breakpoints have no read support in the blood. While all DM-specific breakpoints for sample GBM-06-0648 have at least one supporting read in the blood, 13 of its 16 breakpoints are found to have at least 5 supporting blood reads.

Given the propensity of DMs to be lost after successive rounds of mitosis, it seems unlikely that the DM was present in the patient's germline and selectively retained only in the cells that eventually became oncogenic. A more plausible interpretation is that this oncogenic DM is instead somatic in origin, constructed and amplified at some point prior to or during tumorigenesis. The presence of tumor discordant reads in the blood of DM-specific breakpoints then suggests that these GBM-borne DMs cross the blood-brain barrier to enter the patient's bloodstream.

The third GBM tumor shown in Figure 4.10, GBM-06-0145, exhibits an extreme level of EGFR amplification that could indicate the presence of an EGFR-DM. However, FISH analysis presented above shows that amplified EGFR displayed a primarily HSR-like pattern in tumor cells, with a minority of cells with the pattern of

DM-based amplification. Interestingly, only one breakpoint associated with the amplified region was discovered in GBM-06-0145's matched-normal sequencing data.

The discovery of DM-specific DNA in blood begs the question of how DMs could enter the patient's bloodstream. One possible transport mechanism for these oncogenic DMs is via microvesicles (or exosomes), which brain tumors are known to produce possibly as a method for intercellular communication of mRNA transcripts [88]. Microvesicles bud off from tumor cells with lipid bilayers intact, carrying cytoplasmic contents of the tumor cell such as mRNA, miRNA, and angiogenic proteins, which can then be observed in the patient's blood [76]. More recently, studies have discovered isolated microvesicles containing single-stranded DNA (ssDNA) with amplified oncogenic sequences, in particular c-Myc [5]. Tumor nucleic acids leaked into the blood in this manner or via macrophages that engulf necrotic or apoptotic cells have been proposed as possible disease biomarkers in several cancers including glioma [71, 44]. If microvesicles are the primary mechanism that carry the contents of the tumor cytoplasm into the blood, this may help explain why little evidence of GBM-06-0145's primarily HSR-amplified region was detected in blood. Since HSR-amplified regions are integrated in the tumor genome and thus confined to the nucleus, they are far less likely to be included in a budding microvesicle than the cytoplasmic DMs.

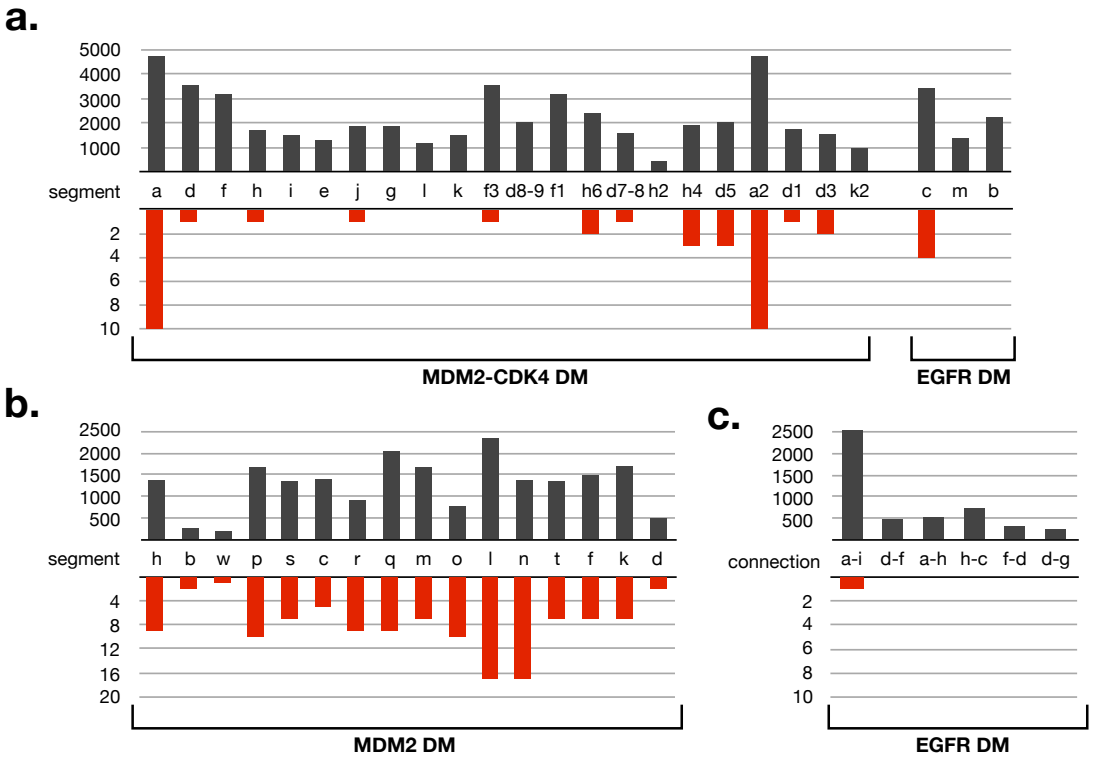

Figure 4.10: Bar graphs showing breakpoint read support for tumor (black bars) and blood (red bars) for sample: (a) GBM-06-0152, (b) GBM-06-0648, and (c) GBM-06-0145.

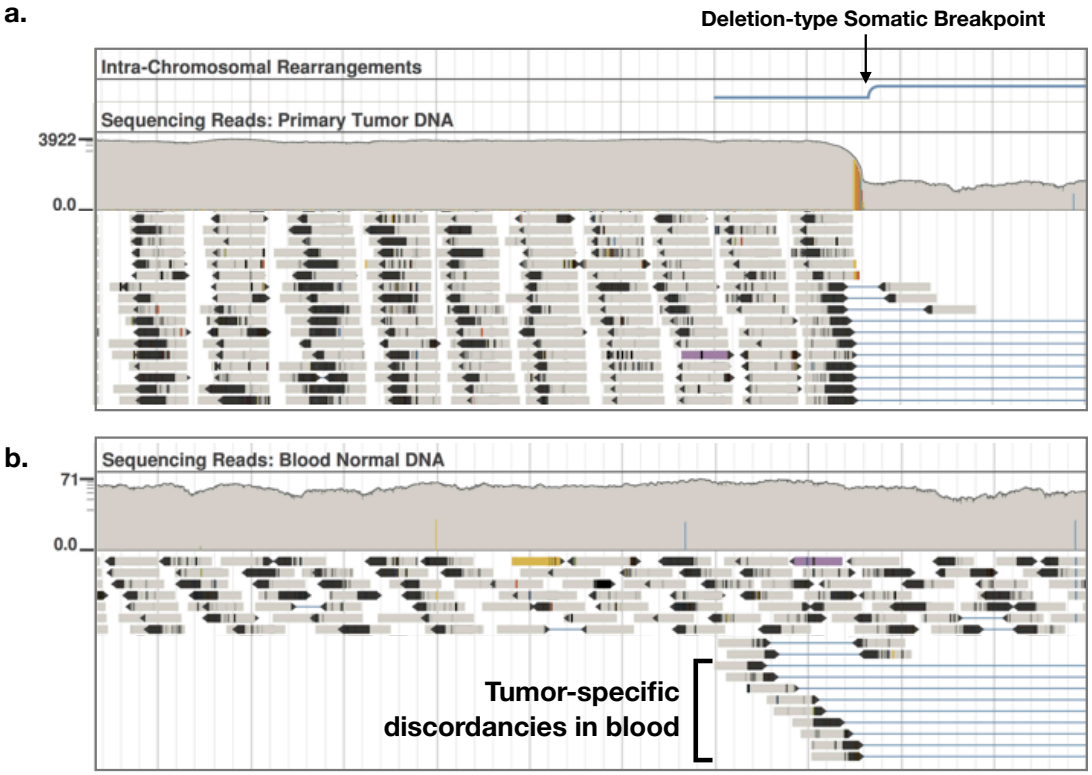

Figure 4.11: Browser shot of raw sequencing reads for GBM-06-0648 surrounding the leftmost position of a highly-supported deletion-type breakpoint that is part of this sample's putative DM. (a) The position of the deletion-type breakpoint and an excerpt of the tumor sequencing reads that support the breakpoint and the associated amplicon. (b) The sequencing reads for the matched-normal (blood) sample for GBM-06-0648 showing a set of 9 sequencing reads that support the breakpoint shown in (a).

4.4 Discussion

The integration of relative copy number with structural variants as described here has demonstrated how such data can be used to improve our understanding of the genomic topology of the cancer genome. We show that, in the case of highly amplified regions of the tumor, we are able to completely explain both the observed copy number and highly supported breakpoints by solving a simple breakpoint graph, which describes the order and orientation of the highly amplified segments in the tumor genome. We can obtain local genome assemblies that are accurate down to the base-level and precisely determine their genetic content, an impressive achievement considering this is done using only a pair of short-read sequencing datasets.

In two of the GBM samples discussed here, the optimal solutions to the breakpoint graphs of amplified segments are circular. These circular solutions suggest that the observed amplified regions are DMs. The presence of oncogenes on each DM and their highly amplified state indicate that the DMs have strong oncogenic potential, confer a selective advantage to the tumor cell, and their formation was likely a key event in the tumorigenesis of these GBM tumors. For the GBM tumor with FISH data, the predicted DM was confirmed to exist. A third sample exhibiting extreme amplification of EGFR yielded a more complicated breakpoint solution, potentially supported by the fluctuating state between HSR- and DM-like patterns observed in its FISH results.

Most interesting, nearly every breakpoint specific to the DMs of two GBMs also has detectable read support in the blood of the patient. In fact, the infiltration of

GBM-borne DMs in the blood for sample GBM-06-0648 is observed at such a high level that DM-specific breakpoints could be reliably inferred using only the whole genome sequencing data derived from the patient's blood. This result suggests some intriguing clinical applications, which will be described further in Section 6.2.

Chapter 5

Inferring clonality and tumor evolution

A tumor growth is a population of cancer cells. This population may homogenous, where all tumor cells share similar characteristics. Such tumors are defined here as "monoclonal" since all tumor cells feature the same variants (copy number aberrations, structural variants, mutations) as each other, all clones of the same progenitor tumor cell. This progenitor tumor cell may be the first cancerous cell that initiated the tumor, or may be a subsequent tumor cell that gained an advantageous mutation that aided a complete sweep of the tumor population.

"Polyclonal" tumor growths are defined as tumors composed of at least two distinct clonal populations of tumor cells. In "polyclonal" tumors, each clonal populations arose from a clone, but each clone differs from the other by some observable

alteration. The multiple clonal populations may be significantly different from each, or, as is more often the case, the clonal populations are related, sharing a set of variants that are found in all or a large subset of tumor cells. For the purposes of this analysis, we will define a polyclonal tumor to consist of multiple "major clones," where "major clone" represents a "detectable" clone, typically representing ∼10% of the tumor population. This distinction is necessary to remind us to remain cognizant of the fact that a growing tumor population is not static and is likely to consist of numerous minor clones that are undetectable with any given method.

In addition to the above definition of tumor clonality, we can further classify individual mutations as either "clonal" or "subclonal." When the dominant clones of a particular tumor is found, "clonal" variants are those shared by all tumor cells of any or all dominant clones, in other words these variants achieved full penetrance in the entire population or polyclonal subpopulation of cells. "Subclonal" variants are those that exist in only a small proportion of the cells belonging to a clonal population.

Let's apply the above definitions to an example of a tumor's evolution, shown in Figure 5.1. Suppose an initial cell acquired a nonsense mutation in a key tumor suppressor (M1) and amplified an oncogene (A1) that supported the initial growth of a tumor. Early on in this tumor's development, another tumor suppressor was deleted (D1) that caused the tumor cell to grow even more rapidly, enabling cells with this deletion to rapidly overtake the entire tumor population. Soon after acquiring deletion D1, a cell also acquires a set of neutral mutations (M2, M3), amplifications (A2, A3), deletions (D2, D3). Since these variants occurred early during the clonal expansion of

this tumor cell variant, but do not provide any selective advantage, the population of tumor cells are split into two "major clones," where 25% of tumor cells have the neutral variants (M2, M3, A2, A3, D2, and D3) and 75% of tumor cells do not. Much further during this tumor's development, additional mutations (M4, M5) appear on one of the two major clones, but do not have a chance to spread through the population prior to the patient's death and/or tissue biopsy.

In the above example, we would define this tumor population as polyclonal, with its two major clones defined such that: clone (1) has variants M1, A1, and D1, and clone (2) shares the variants of clone (1), but in addition has variants M2, M3, A2, A3, D2, and D3. The clonal mixture is determined as 75% clone (1) and 25% clone (2). Mutations M1, M2, and M3 would all be classified as "clonal" since they all achieved full penetrance in their respective clones, whereas M4 and M5 would be classified as "subclonal" mutations.

This chapter presents various methods to extract and synthesize data to reconstruct the clonal evolution of a tumor from whole genome sequencing data of a single tumor biopsy. In a mix of automated and manual processes, these methods form a framework to determine the tumor's clonality, the number and proportion of all major clones, and the variants that distinguish them. Furthermore, a method to phase mutations to parental alleles to time their emergence within the population is described. In addition, the methods will provide an accurate estimate of the amount of contaminating normal tissue was present in the tumor's biopsy.

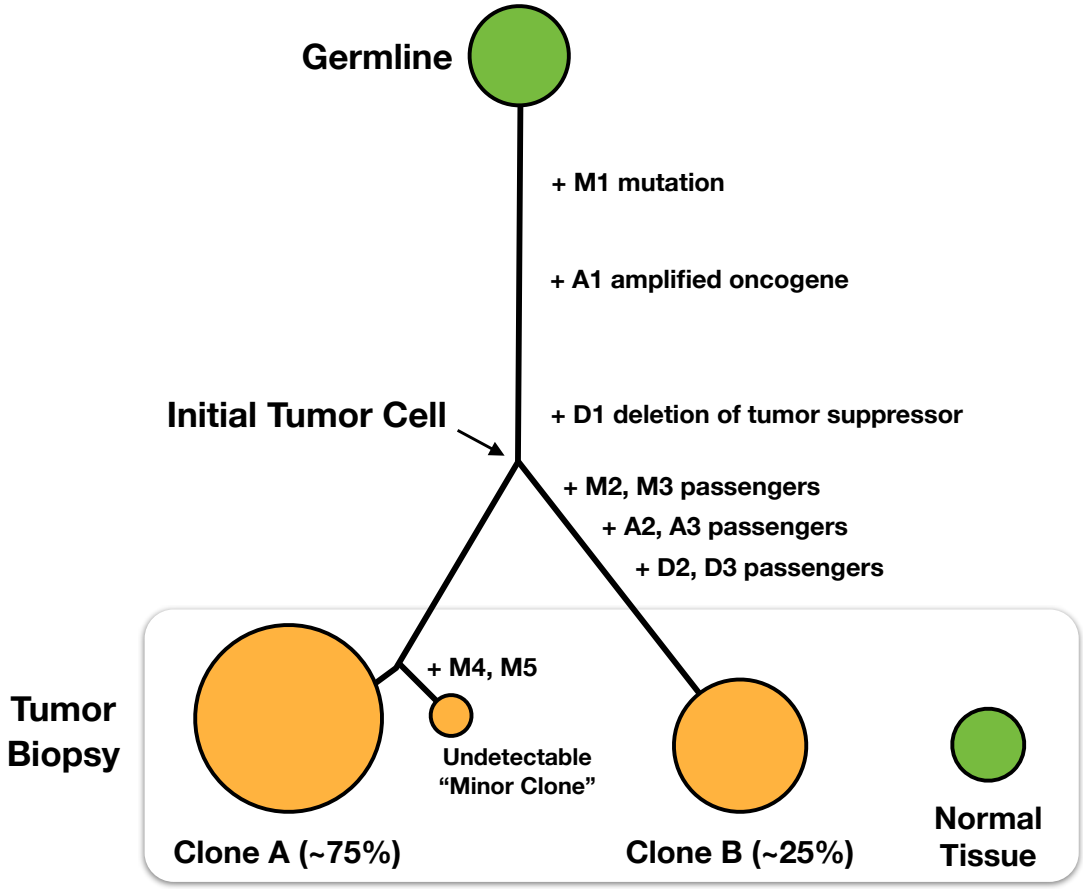

Figure 5.1: A toy example of the evolution of a tumor starting from the germline, to an initial tumor cell, to a population of major and minor clones that are sampled by the tumor biopsy. To avoid complicating this example, mutation and copy number passenger events that occurred prior to the emergence of the initial tumor cell were not included.

5.1 Copy number alterations, allele fraction, and the allelic state diagram

To discover and describe the major clones of a population, relative copy number and allele fraction estimates made by BamBam of copy number alterations are utilized. Other groups have studied tumor clonality using SNP chips or single point variants sequenced at extremely high coverage ($>$ 150x for whole genomes or $>$ 1000x for exomes) [15, 59, 92]. Unfortunately, the precision of allele fraction estimates is poor for point mutations using the average coverage ($30 - 50$x) whole genomes sequenced by TCGA, which is a key measure in determining clonality. By aggregating over large genomic regions, the error in these estimates is drastically reduced, enabling distinct clones to be resolved. However, this does limit the ability of this method. It can resolve major clones only based on copy number alterations. Clones distinguished solely by point mutations are not discoverable by this method, although this limitation is mitigated by the mutation clonality analysis described below in Section 5.4.

Underlying the method to determine both clonality and estimate normal contamination is the "allelic state diagram" (ASD), which is defined here. The ASD describes the positions of clonal allele-specific copy number variants using both relative copy number and allele fraction of copy number alterations, demonstrating the relationship between copy number and allele fraction for all allelic states. The positions of

clonal allelic states in the ASD are determined by the following:

$$CN(t_{maj}, t_{min}, n_{maj}, n_{min}, \alpha) = \frac{(1-\alpha)(t_{maj}+t_{min})+\alpha(n_{maj}+n_{min})}{n_{maj}+n_{min}} \quad (5.1)$$

$$AF(t_{maj}, t_{min}, n_{maj}, n_{min}, \alpha) = \frac{(1-\alpha)t_{maj}+\alpha n_{maj}}{(1-\alpha)(t_{maj}+t_{min})+\alpha(n_{maj}+n_{min})} \quad (5.2)$$

where CN is the relative tumor copy number compared to a matched-normal, AF is allele fraction in the tumor, α is the fraction of normal contamination in the tumor sample, and $t_{maj}, t_{min}, n_{maj}$ & n_{min} are the majority and minority allelic states in the tumor and normal, respectively. Since individual genomes can only have discrete allelic states, such that they have 0, 1, 2, or more copies of a given chromosomal segment, the possible values for t_{maj} and t_{min} are constrained to the set of positive integers, $t_i \in (0, 1, 2, ..., n)$. Furthermore, the majority and minority allelic states for the normal are set to one, $n_i = 1$, which is true for all of the autosomes in a normal human genome. The sex chromosomes, X and Y, are currently ignored in the ASD. Note that since the above math necessarily requires two alleles, only heterozygous sites in the matched-normal genome are considered for the ASD.

The forthcoming figures all share the same set of allelic states: normal copy number, single-copy amplification, single-copy / hemizygous deletion, homozygous deletion, copy-neutral LOH (CN-LOH), and amplification of both parental alleles. Figure 5.2 shows simulated copy number and allele fraction data for these allelic states with no normal contamination, demonstrating how the ASD can be used to determine the allelic state of each cluster of points. Each vertex in the ASD's grid is labeled with its tumor allelic state, (t_{maj}, t_{min}), and the position is determined by the equa-

tions above. Figure 5.3 demonstrates how the location of allelic states are affected by increasing amounts of normal contamination, α. As normal contamination increases, the allelic state positions grow closer together, reducing the ability to resolve different allelic states.

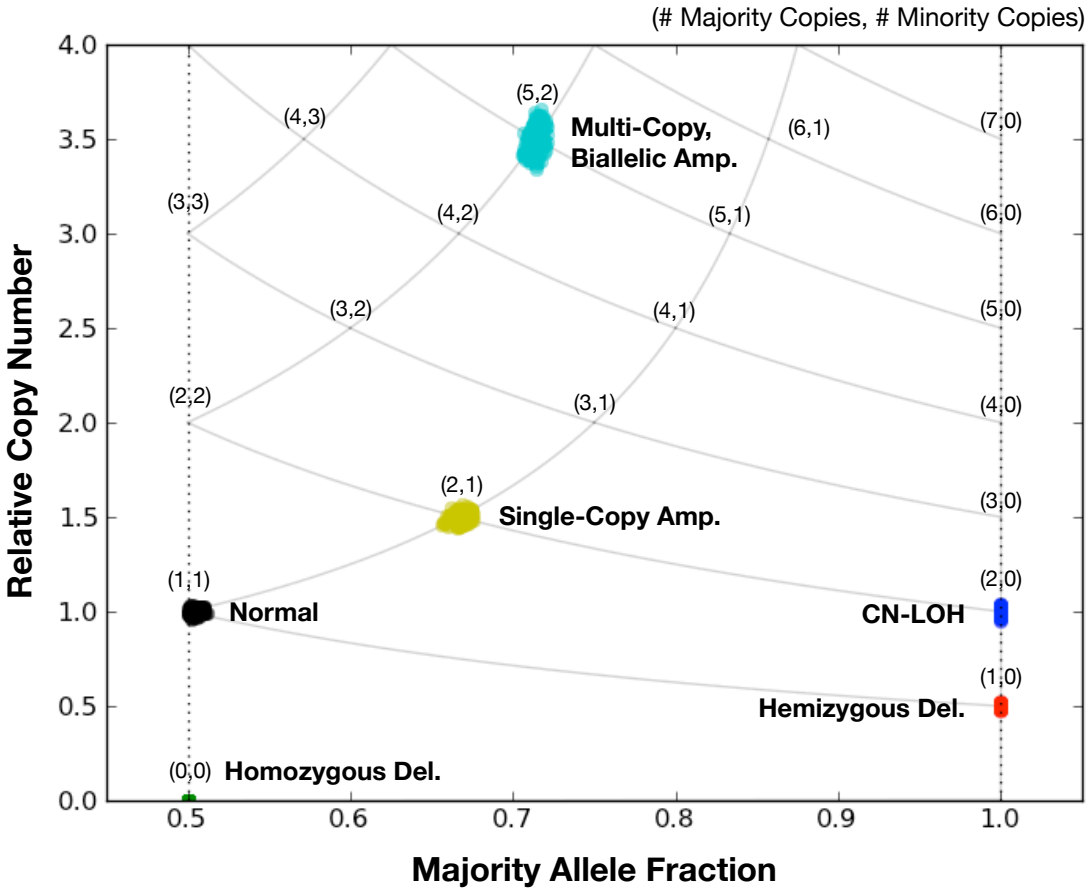

Figure 5.2: An example of an allelic state diagram (ASD) of simulated data for a monoclonal tumor sample with zero normal contamination, $\alpha = 0$. Chromosomal regions exhibiting different copy number alterations are plotted in different colors. This simulated tumor genome exhibits 6 allelic states: normal (black), single-copy amplification (yellow), hemizygous deletion (red), homozygous deletion (dark green), copy-neutral LOH (CN-LOH, blue), and multi-copy biallelic amplification (cyan).

The case presented in Figure 5.3 is a static snapshot of a monoclonal tumor.

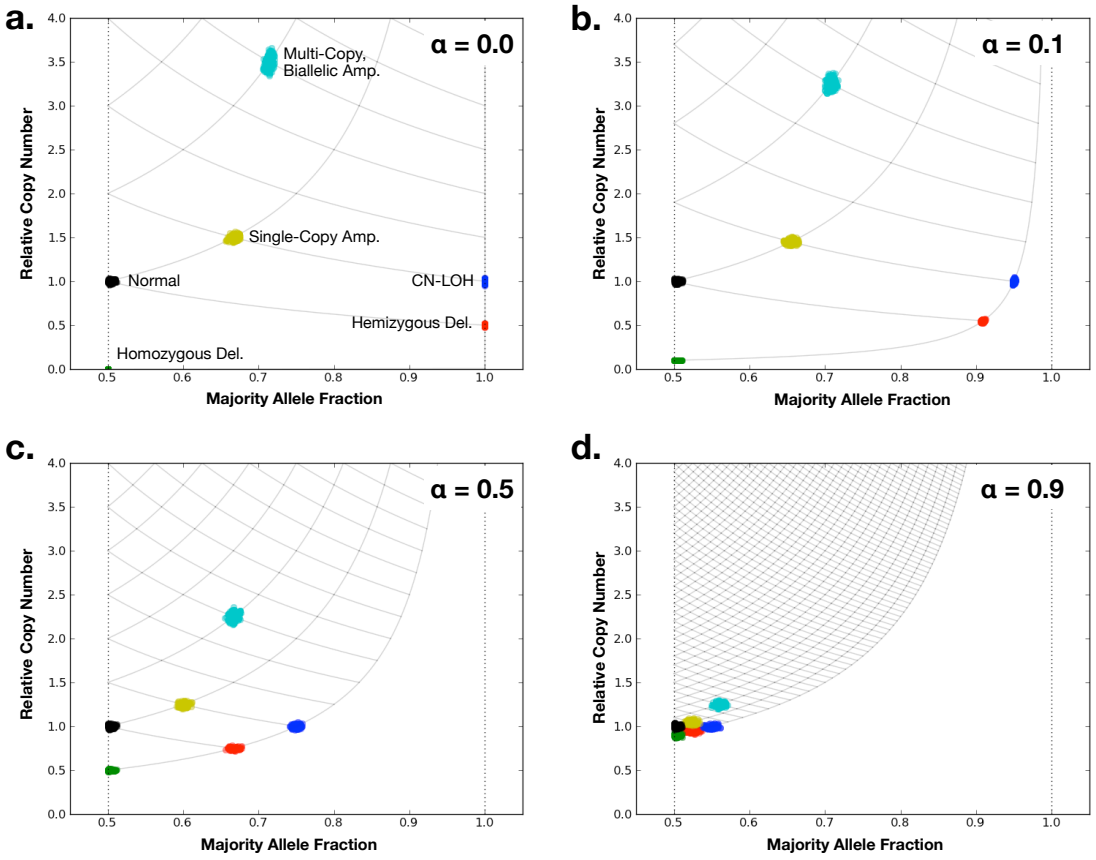

Figure 5.3: A set of allelic state diagrams for the same simulated monoclonal tumor genome shown in Figure 5.2 with different levels of normal contaminant: (a) 0%, (b) 10%, (c) 50%, and (d) 90%. Notice that, as the level of normal contamination increases, the distance between allelic states is reduced, negatively affecting our ability to distinguish between different states.

However, we know that the tumor genome can be very dynamic, with gains and losses of small and large chromosomal segments. Figure 5.4 shows some of the possible transitions between the allelic states described in previous figures. Note that some transitions are deemed "one-way" since they involve the irreversible loss of chromosomal segments. For example, the transition between the normal allelic state (1,1) and the hemizygous deletion state (1,0) is "one-way" because that deleted allele can never be restored. However, the retained allele in this case can be amplified, permitting transitions to the copy-neutral LOH (CN-LOH) state and beyond (2+,0). Notice that the deletions necessary for the transition between the cyan-colored allelic states are not deemed "one-way" because at least one copy remains of each allele remains in the genome.

Figure 5.5 displays the ASD of a tumor genome transitioning from the allelic states presented in the previous figures to new allelic states that differ only by a single copy loss or gain. During such a transition, the population of tumor cells will be a mixture of tumor cells consisting of tumor cells with the original allelic states and tumor cells with the new allelic states. For the example presented in Figure 5.5, we can think of this "transitional" tumor as a population divided between two major clones, A and B, where clone A is defined by the original allelic states, and clone B is defined by the new allelic states. The mixture fractions shown on this figure, M_b, represents the fraction of clone B within the population, such that the tumor population solely consists of clone A when $M_b = 0$, and the population consists only of clone B when $M_b = 1$. It is important to note that the allelic states for both clones, $t_{i,a}$ and $t_{i,b}$, are still constrained to the set of positive integers.

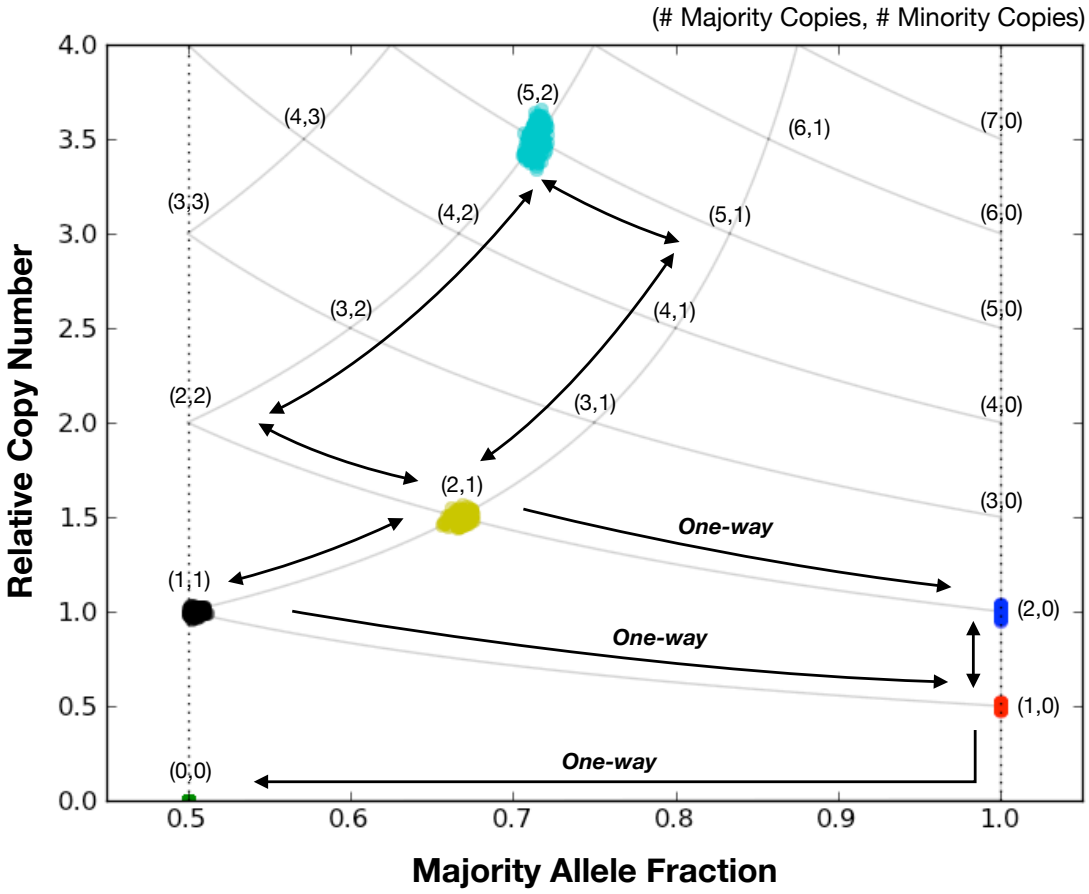

Figure 5.4: Allelic state diagram of the case presented in Figure 5.2 showing some possible bidirectional and unidirectional transitions between its allelic states. Note that unidirectional transitions are those that involve irreversible loss of a parental chromosome.

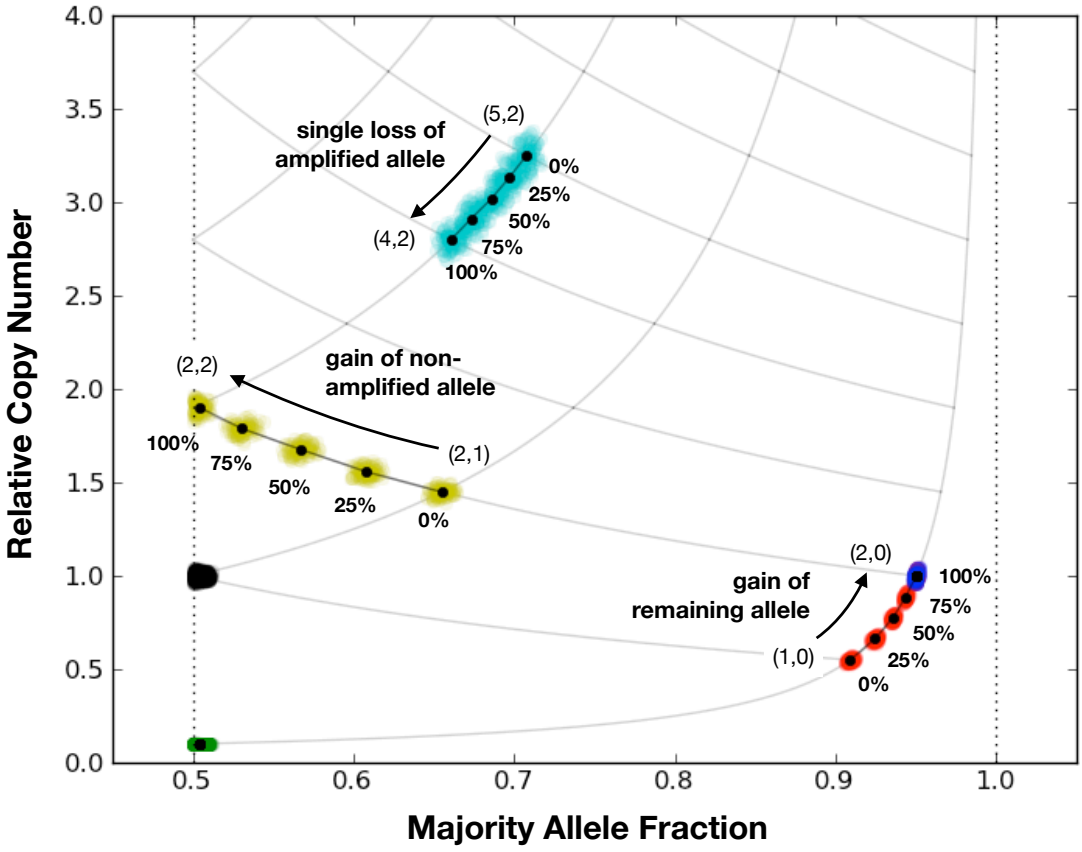

Figure 5.5: Allelic state diagram of the case presented in Figure 5.2 showing how transitional allelic states are created when the tumor consists of a mixture of two different clones. In this case, there are two subclones. Clone A is defined by the original allelic states: yellow (2,1), cyan (5,2), and red (1,0). Clone B alters these states through amplifications and a deletion to produce the allelic states: yellow (2,2), cyan (4,2), and red (2,0). They share the same allelic states of blue (2,0), black (1,1), and green (0,0). The percentages denote the percentage of clone B present in the tumor population, where 0% describes a monoclonal population of clone A, and 100% is a monoclonal population of clone B.

From Figure 5.5, when the mixture fraction M_b is such that the tumor is a heterogenous population of cells, $M_b = 0.25, 0.5, 0.75$, the allelic states do not lie on the vertices of the ASD, but rather on the edge connecting two vertices. A tumor population in such a state would be classified as polyclonal. Take, for example, the cluster of red points in Figure 5.5. In this region of the genome, clone A has the allelic state of hemizygous deletion, or $(1, 0)$, whereas clone B has amplified clone A's retained allele, altering its allelic state in this region to copy-neutral LOH, or $(2, 0)$. When $M_b = 0$, the allelic state of the red points are found clustered on the ASD vertex representing the hemizygous deletion allelic states. As M increases, meaning increasing amounts of clone B in the population, the cluster of points progresses along the edge towards the CN-LOH state. At $M_b = 0.5$, where there are equal amounts of clone A and B in the population, the cluster of points is found precisely in the middle of the edge between the LOH and CN-LOH allelic states.

If the tumor population consists of non-derivative clones, or clones that are distantly related to one another such that their allelic states do not differ by single copy gains or losses, the position of the mixture of allelic states will not lie along the edges of the ASD, as shown in Figure 5.6. We will see later that such abnormal allelic states can also occur when more than 2 major clones exist in a polyclonal tumor.

This demonstrates how the ASD can, at a glance, indicate the presence of one or more major clone in a tumor sample, determine the allelic states of the major clones, and provide a visual estimate of the proportion of each major clone in the tumor population. Thus, the ASD can, at a minimum, be used as a powerful visual diagnostic

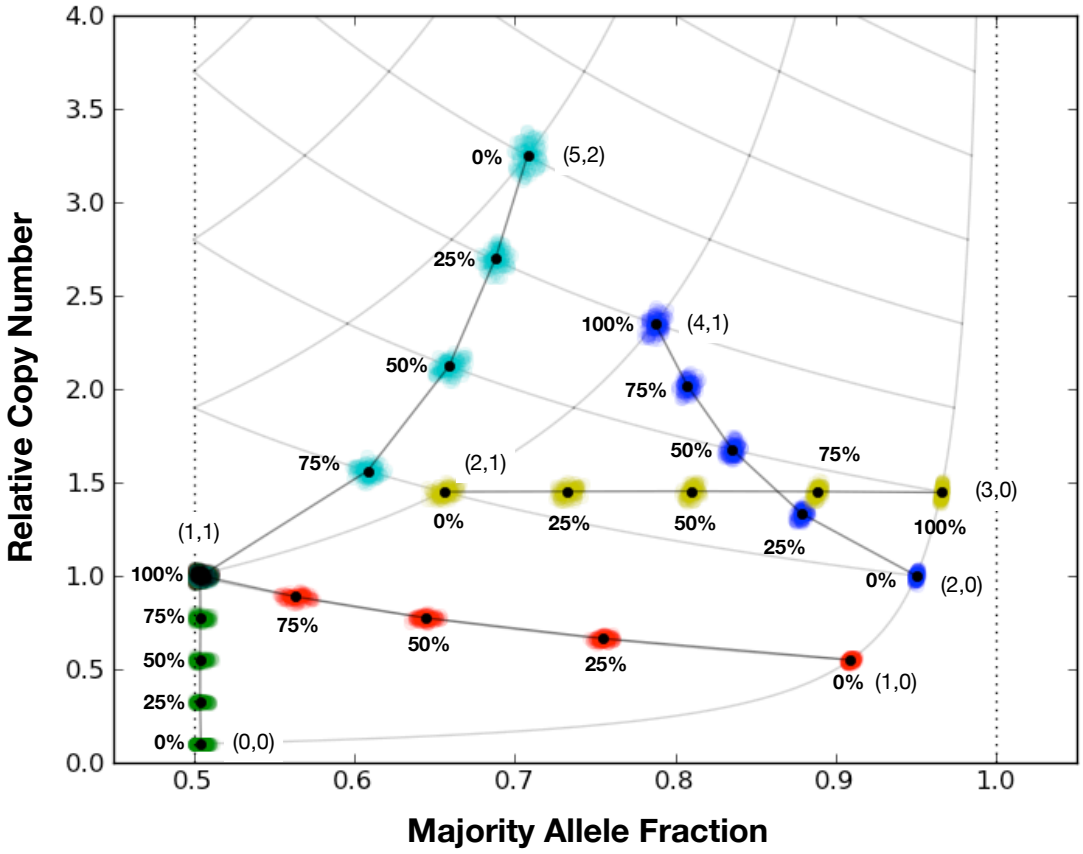

Figure 5.6: Allelic state diagram of the case presented in Figure 5.2 showing the transitional allelic states produced when allelic states are "skipped," which can occur when the tumor consists of two or more unrelated, or distantly-related clones. Notice that the transitional states are not found on the edges connecting allelic states if the allelic states of the two major clones differ in both majority and minority alleles.

tool for determining the clonality of a tumor sample.

5.2 Fitting sequencing data to the ASD

The mathematics behind the ASD, Equation 5.1 and Equation 5.2, is modeling the ideal case where the relative copy number is 1.0 and the majority allele fraction is 0.5 for "normal" $(1, 1)$ allelic states. However, the results produced by BamBam on real world data do not precisely fit this idealized case. To estimate relative copy number, BamBam calculates the relative *coverage* between tumor and normal. If the tumor and normal samples are sequenced at the same coverage level, relative coverage is an accurate measure of relative copy number. However, this will not be the case if the tumor sample is sequenced at a much higher coverage than its matched-normal, in an attempt to improve detection of mutations, particularly subclonal mutations, in the tumor sample.

For example, assuming no normal contamination, if a tumor is sequenced at twice the coverage of its matched-normal, then a region with a "normal" allelic state will have twice as many reads in the tumor as it has in the normal. Thus, this region has a relative coverage of 2.0 and a relative copy number of 1.0, and BamBam's use of relative coverage will not fit the ASD. Unfortunately, the precise coverage level of a given sequencing dataset is unknown, as the sequencing centers merely target the desired coverage level, but have no guarantee of achieving it. Using the raw number of reads found in the tumor and matched-normal datasets as an estimate of overall

coverage level can help correct the imbalance, but is complicated by the ploidy of the tumor sample. If a tetraploid tumor (ploidy = 4.0) and its matched-normal (ploidy = 2.0) are sequenced at the same physical coverage, the tumor will have two times the number of raw reads than the matched-normal. So, using the ratio of their raw numbers of reads to scale local relative coverage estimates would this tetraploid tumor to appear to have normal copy number.

The error in the estimate of allele fraction that BamBam produces is caused by a limitation in how the majority allele is selected in regions of allelic balance, such as the "normal" allelic state. Ideally, the allele fraction for such regions should be approximately 0.5, but this only occurs when both alleles have equal read depths. More often, due to the stochastic nature of how heterozygous alleles are sampled from the pool of genomic DNA, one of the two alleles will likely have a slightly higher read depth than the other, causing a slight increase in the majority allele fraction that is estimated.

For example, assuming no normal contamination, a whole genome with 30x coverage would ideally produce 15 reads of each allele at heterozygous "normal" allelic states. However, if one allele's read depth was shifted by just a single read, such that allele A has read support of 16, BamBam would estimate the majority allele fraction to be $\frac{16}{30} = .53$, a deviation of 0.03 from the actual allele fraction. Usually averaging over multiple positions can reduce the effect of such errors, however the error in majority allele fraction for these balanced allelic states cannot be averaged out because majority allele fraction, by its definition, can never dip below 0.5. Fortunately, sampling error has a much less pronounced effect on amplified and deleted allelic states. In these cases,

the majority allele is readily identifiable and the sampling error can be averaged out over multiple positions.

In order to fit BamBam results onto the idealized ASD, we must then model and correct out the above errors from the data. The model has four parameters: normal contamination α, allele fraction delta AF_d, coverage delta COV_d, and coverage scaling factor COV_s. The α parameter only affects grid layout of the ASD, as shown in Figure 5.3. The latter three parameters transform the BamBam results. The paramaters COV_d and COV_s affect the y-axis shift of the copy number data and scale of copy number data from the "normal" allelic state according to the following equation:

$$CN_{corr}(CN, COV_d, COV_s) = COV_s(CN - COV_d) + 1.0 \tag{5.3}$$

where CN is the relative copy number estimate produced by BamBam and CN_{corr} is the corrected copy number used to compare against the ASD. The final parameter, AF_d has its strongest effect on the the allele fraction estimates of the allelic balanced states. It does this with the following equation:

$$AF_{corr}(AF, AF_d, CN_{corr}, x) = AF - \frac{AF_d}{CN_{corr}}(\frac{1.0 - (AF - AF_d)}{0.5})^x \tag{5.4}$$

where x is set to a large integer (e.g. $x = 20$) to rapidly reduce the degree to which allele fraction estimates are corrected as they diverge from balanced allelic states. This parameter helps reduce the error in allele fraction estimates introduced when the "majority" allele achieved a higher read support from unbalanced sampling of the heterozygous site, such as the normal allelic state, $(1, 1)$, and any amplified state where both alleles are amplified equally, (N, N). It is very important to not appreciably alter the allele

fraction estimates at deleted states, as they are the determining factor for estimating normal contamination.

The optimal values for these four parameters are discovered using gradient steepest descent search, optimizing the RMSD of the corrected copy number and allele fraction estimates, CN_{corr} and AF_{corr}, to the ASD defined by the normal contamination parameter, α. The search begins with a set of initial values for each parameter, and a set of increments for each parameter, COV_d^i, COV_s^i, AF_d^i and α^i. For each parameter, p, and parameter increment, p^i, the RMSD from the ASD is calculated for $p, p + p^i$ and $p - p^i$. The parameter value that yields the greatest reduction in RMSD among all four parameters is chosen as the new current value for that parameter, and the cycle repeats. If no reduction in RMSD is possible with the current parameter increments, the increments are divided by half and the search resumes. Once three rounds of divisions have occurred, the search is concluded and the best fit parameters are reported.

Since gradient descent can often get stuck in a local minimum, gradient search is performed with a number of different initial parameters until a consistent set of fit parameters is found.

5.3 Modeling the clonal mixture of a tumor sample

The ASD can be used to determine a set of allelic state "landmarks" that help define the number of clones and their proportions within the tumor population, $L_i = (CN_{corr,i}, AF_{corr,i})$. The landmarks used in this analysis will be defined by the

large clusters of points on the ASD, as they indicate major portions of the tumor that have undergone a copy number change in a significant fraction of the overall tumor population. See Figure 5.7(b) for the landmark allelic states used to analyze GBM-06-0185. For each landmark on the ASD, all plausible clonal mixtures that would result in its observed copy number and allele fraction are considered, then the optimal clonal mixture is chosen such that it can account for all ASD landmarks most parsimoniously.

As observed in Figure 5.5, for monoclonal tumor populations, we expect the landmarks to all lie on ASD vertices. In polyclonal tumors consisting of two major clones, where clone B inherits all of clone A's allelic states and has additional allelic states distinct from clone A, we will find landmarks on both the vertices and edges of the ASD. The landmarks that lie on vertices are those that represent the allelic states shared by both clone A & B, while the landmarks on ASD edges represent the mixture of the different allelic states. The position along this connecting edge determines the proportions of clones A and B in the mixture. If multiple landmarks are found on edges and not on vertices, then the variety of positions along their respective edges will determine the number of clones.

For example, if all landmarks are found halfway between two allelic states, that is most simply explained by two major clones in equivalent proportion within the population. If, however, one landmark is located at the halfway mark and another is found 25% along the way towards an allelic state, there must be more than two clones in the population. One simple explanation for this is that there are three clones, A, B, & C, where A makes up 50% of the tumor population and clones B & C make up 25%

each. Suppose clones B & C both exhibit a single-copy allelic state change from clone A, explaining the halfway landmark. The 25% landmark is then explained if, in that chromosomal segment, clone B (or C) experienced a single-copy allelic state change not found in clones A and C (or B).

Thus, the problem at hand is determining the least number of major clones that explain n observed landmarks:

$$L^{obs} \in (L_0^{obs}, L_1^{obs}, ..., L_n^{obs}), \tag{5.5}$$

where $L_i^{obs} = (CN_i^{obs}, AF_i^{obs})$. We assume a mixture of m clones, each with k integral majority and minority allelic states $C_i = [(t_{maj,i}^0, t_{min,i}^0), (t_{maj,i}^1, t_{min,i}^1), ..., (t_{maj,i}^k, t_{min,i}^k)]$, and mixture proportions, M_i, such that $\sum M_i = 1.0 - \alpha$. The relative copy number and allele fraction of each landmark, L_i^{mix}, is a linear combination of the allelic states indexed by i across the clonal mixture:

$$CN_i^{mix} = \frac{2\alpha + \sum_k^m M_k(t_{maj,i}^k + t_{min,i}^k)}{2} \tag{5.6}$$

$$AF_i^{mix} = \frac{\alpha + \sum_k^m M_k(t_{maj,i}^k)}{2\alpha + \sum_k^m M_k(t_{maj,i}^k + t_{min,i}^k)} \tag{5.7}$$

where the normal allelic states for all clones are assumed to be $n_{maj,i}^k = n_{min,i}^k = 1$. The optimal solution is the one that most closely approximates the observed landmarks with a simplest mixture of major clones, or optimizing the objective function:

$$O(L^{obs}, L^{mix}) = \frac{1}{n}\sum_i^n (CN_i^{obs} - CN_i^{mix})^2 + (AF_i^{obs} - AF_i^{mix})^2 + m^x \tag{5.8}$$

which is the squared distance from the observed data plus a penalty, m, equal to the number of clones in the population, that is controlled by the strength parameter x. This

is a complicated objective minimization problem that due to the inclusion of integral allelic states. Shortcuts based on modeling the evolutionary process of the tumor's development are possible.

The optimization of Equation 5.8 is performed after finding the best fit parameters for α, AF_d, COV_d and COV_s as described in Section 5.2. It begins by identifying all "shared" landmark allelic states, which every clone in the mixture must exhibit. If we assume a tumor is step-wise evolving, these shared allelic states represent the "root" of the tumors evolutionary tree. If there are no landmarks on ASD edges, the procedure is complete and the tumor population is classified as monoclonal.

If landmarks exist along ASD connecting edges, between two bounding allelic states, then additional clones are necessary. The procedure adds one additional "daughter" clone to the mixture, which inherits all of the shared allelic states and gains an allelic state and mixture proportion necessary to explain the edge-bound landmark. If more than one edge-bound landmark can be explained with the same mixture proportion, then those new allelic states are added to the new clone. This process is repeated until all non-vertex landmarks are explained by the clonal mixture, wherein each additional "daughter" clone can derive from any current clone in the mixture that bounds one side of an unexplained landmark. Once all landmarks can be reasonably explained, the clones' allelic states and mixture proportions are reported.

Note from Equation 5.6 and Equation 5.7 that the combination of allelic states that uniquely determine each landmark's position on the ASD can also determines the phased set allelic states for all positions in the genome that correspond to the landmark.

This can only work when the mixture proportions are unique for each clone, i.e. the major clones must unevenly split the tumor population. In such cases, this enables whole genome, allele-specific copy number profiles to be inferred for each clone in the tumor population.

5.4 Linking mutations to clone-specific allelic states

To achieve an even greater understanding of the evolution of a tumor, we should not constrain ourselves to only looking at copy number changes. By integrating somatic mutations into this framework, we can attempt to determine when a mutation arose during the tumor's development. To do this, we will first attempt to link mutations directly to the majority or minority allele in the encompassing chromosomal region on the ASD and then use its mutational allele fraction to determine whether the mutation occurred prior to the change in allelic state, soon after the allelic state change, or much later.

5.4.1 Via direct phasing

For every mutation discovered by BamBam, we will find all nearby germline heterozygous variants and look for paired reads that physically connect, or "phase," the mutation allele to a specific germline allele. "Nearby" is defined as being separated by no more than double the insert size of the paired read library, typically ∼1,000bp for these whole genome libraries, as that is well outside the expected distance that would

separate two paired reads.

All read pairs that overlap the positions of both the mutation and the germline variant are collected and the number of times the mutation is phased to either of the two germline variant alleles is recorded. If the mutation is found linked to the same germline variant allele more than once, and is not found also phased to the other allele of that germline variant, it is considered to be directly phased to that germline variant allele. Phasing can be made either within a single read if the mutation and germline variant are separated by less than a read's length, or can occur across mates of a read pair. Mutations can also be phased to multiple germline variant positions.

For every mutation that can be directly phased to a germline variant, the germline variant's allele fraction is used to determine if the mutation is phased to the majority or minority allele. If the germline variant's allele fraction is determined to be greater than or equal to 0.5, then the mutation is deemed "majority-phased," otherwise it is phased to the minority allele, or "minority-phased." Note that in the cases when the two allelic states are equal, such as normal $(1, 1)$ or bi-allelic, balanced amplifications $(2, 2)$, the mutation's assignment to "majority" or "minority" allele depends on whichever allele was sampled slightly deeper in the sequencing data. Thus, classifying mutations as "majority-phased" or "minority-phased" in such cases is not as meaningful.

5.4.2 Via amplified allele fraction

When direct phasing cannot be made, our ability to determine which allele the mutation is linked to is severely limited. However, when mutations are found within an

amplified chromosomal segment, we can use the mutation's allele fraction to determine to which allele the mutation may be linked. When the mutation's allele fraction is approximately equal to the majority allele fraction, this can only have occurred if the mutation was present in the amplified allele prior to the amplification. If the mutation were instead on the un-amplified allele, the mutation's allele fraction would necessarily be much lower.

However, low mutation allele fractions do not necessarily indicate that the allele is not "majority-phased," since mutations can occur post-amplification. For example, if a region was amplified by a single copy, allelic state $(2, 1)$, a post-amplification mutation could be, at most, present on one copy of the majority allele, with a maximal allele fraction of $\frac{1}{2+1} = 0.33$, compared to the expected allele fraction of a pre-amplification mutation, $\frac{2}{2+1} = 0.67$.

Thus, we are limited to linking un-phased mutations to amplified segments when the mutation occurs prior to the amplification. This can be still be useful, as we expect oncogenic mutations to occur early in the tumor's development, as they are likely to drive tumor growth. If multiple copies of these oncogenic mutations are selectively advantageous for the tumor cell, then we'd expect the requisite increase in mutation copy number and allele fraction to enable us to use this method.

5.4.3 Comparing allele fractions to infer mutation timing

After assigning mutations to the majority or minority alleles, we compare the allele fraction of the mutation to the allele fraction of the majority or minority allele

fraction of the chromosomal segment that encompasses the germline variant allele. We use the allele fraction of the chromosomal segment instead of the germline variant allele because the estimate of the chromosomal segment's allele fraction is more accurate due to averaging over all germline heterozygous positions within the segment.

To accurately compare a mutation's allele fraction to the majority or minority allele fraction of heterozygous positions, we must add in some "normal" contamination to the mutated allele. Recall from Equation 5.2 that majority allele fraction, AF, features normal contamination in both its numerator and denominator. This is due to the fact that the positions considered in these equations are heterozygous in the normal, and thus you expect to get normal contamination from both alleles. However, for a somatic mutation, there is no normal contamination of the mutant allele, as the mutation does not exist in the normal:

$$MAF = \frac{(1-\alpha)mt_{maj}}{(1-\alpha)(mt_{maj} + (1-m)t_{maj} + t_{min}) + \alpha(n_{maj} + n_{min})} \tag{5.9}$$

where MAF is mutant allele fraction, m is the fraction of copies of the tumor allele t_{maj} that are mutated, and t_{maj}, t_{min}, n_{maj} & n_{min} represent the same homozygous allele. To fairly compare MAF to allele fractions estimated at heterozygous positions, we must add the following correction:

$$\begin{aligned} MAF_c &= MAF + \frac{\alpha n_{maj}}{(1-\alpha)(t_{maj} + t_{min}) + \alpha(n_{maj} + n_{min})} \\ &= \frac{(1-\alpha)mt_{maj} + \alpha n_{maj}}{(1-\alpha)(t_{maj} + t_{min}) + \alpha(n_{maj} + n_{min})} \end{aligned} \tag{5.10}$$

where MAF_c is the corrected mutant allele fraction. Note that while m is allowed to be any fraction less than or equal to zero in the above equations, there are some values of

m that have special meaning. If $m = 1$, then all of the t_{maj} alleles were mutated, and in the cases where t_{maj} represents an amplified allele, when $m = 1$ the mutation must have occurred prior to the amplification. When $m = \frac{1}{t_{maj}}$, where t_{maj} represents the number of copies of the amplified allele, then we know the mutation must have occurred soon after the amplification, since it exists on a single copy of the amplified allele but is found in this state in the majority of tumor cells. If, however, $m << \frac{1}{t_{maj}}$, then the mutation must have occurred after the amplification, likely very late during the tumor's growth, as its very low allele fraction indicates it's only found in a small fraction of tumor cells.

If the mutation is phased to the minority allele, t_{min}, we expect to find the maximum mutation fraction to be $m = \frac{t_{min}}{t_{maj}}$ as that indicates that all copies of the minority allele were converted. So, when the minority allele state exists in single copy and all copies of it were mutated, $m = \frac{1}{t_{maj}}$, precisely the same mutant fraction one would calculate for a "majority-phased" mutation present at single copy. Thus, only with the direct phasing can you distinguish between an early "minority-phased" mutation and mutations that occur post-amplification.

5.5 Results

We will test the above methods on whole genome sequencing data from two tumor types studied by TCGA: glioblastoma multiforme (GBM) and lung squamous cell carcinoma (LUSC). These two types of tumors were chosen as they demonstrate

Table 5.1: Clonality of 12 GBMs.

Sample	**Clonality**	**Normal Cont.** (α)	**# Major Clones**
GBM-06-0145	monoclonal	22.5%	1
GBM-06-0155	monoclonal	24.5%	1
GBM-06-0877	monoclonal	29.8%	1
GBM-06-0648	polyclonal	12.5%	2
GBM-06-0152	polyclonal	24.1%	3
GBM-06-1086	polyclonal	7.5%	4
GBM-06-0185	polyclonal	14.6%	4
GBM-14-1454	polyclonal	5.9%	> 1
GBM-14-0786	polyclonal	13.9%	> 1
GBM-06-0188	polyclonal	20.6%	> 1
GBM-06-0214	polyclonal	33.6%	> 1
GBM-26-1438	polyclonal	50.0%	> 1

the full range of clonality, from monoclonal to highly polyclonal, possible in different cancers.

5.5.1 Clonality of glioblastoma multiforme tumors

12 whole genome GBM samples were processed with the above methods to determine the level of normal contamination and the clonality present in each tumor biopsy. The relative coverage and allele fraction produced by BamBam for the other 5 whole genome GBM samples discussed in previous sections possessed too much variability to be analyzed by these methods. The results of the clonality analysis are summarized in Table 5.1. Surprisingly, only 3 GBM tumor samples were found to be monoclonal, while the other 9 samples consisted of at least two major clones. For 7 GBM tumors, the precise mixture of clones were determined, while the remaining 5 tumor were inspected visually to determine their clonality.

Results of two tumors, GBM-06-0145 and GBM-06-0185, are shown in Figure 5.7. The relative coverage and allele fraction data of these two samples were transformed using the best fit paramaters as described above, demonstrating close fit to the ASD with estimated normal contamination levels of 21.5% and 14.6%, respectively. By inspecting where the data cluster, whether on vertices or edges, we can visually determine the clonality of these tumors. Since all of GBM-06-0145's data cluster around ASD vertices, it is likely this tumor is monoclonal. On the other hand, GBM-06-0185 is clearly polyclonal, since the several large clusters along the ASD edges indicate the presence of at least two major clones in this tumor. In fact, since the edge-bound clusters are found at different positions along their edges (e.g. the blue clusters are at the halfway mark, while the two purple clusters are approximately .75 and .80 of the way towards the single-copy deletion state, respectively), this can only occur from a mixture of at least three major clones.

To determine precisely the number of clones in these samples, we used the methods described above to determine the number of clones and their allelic states. For each inferred clonal mixture, we computationally determine relative copy number and allele fraction for every position in the genome given the derived clonal mixture and compare it against the results produced by BamBam. This provides a metric to determine how well the clonal mixture models the observed data.

As shown in Figure 5.8, we find a single clone for GBM-06-0145, as expected. The computationally-derived relative copy number and allele fraction data shows a very good fit to the observed data. We find a total of four major clones for GBM-06-0185,

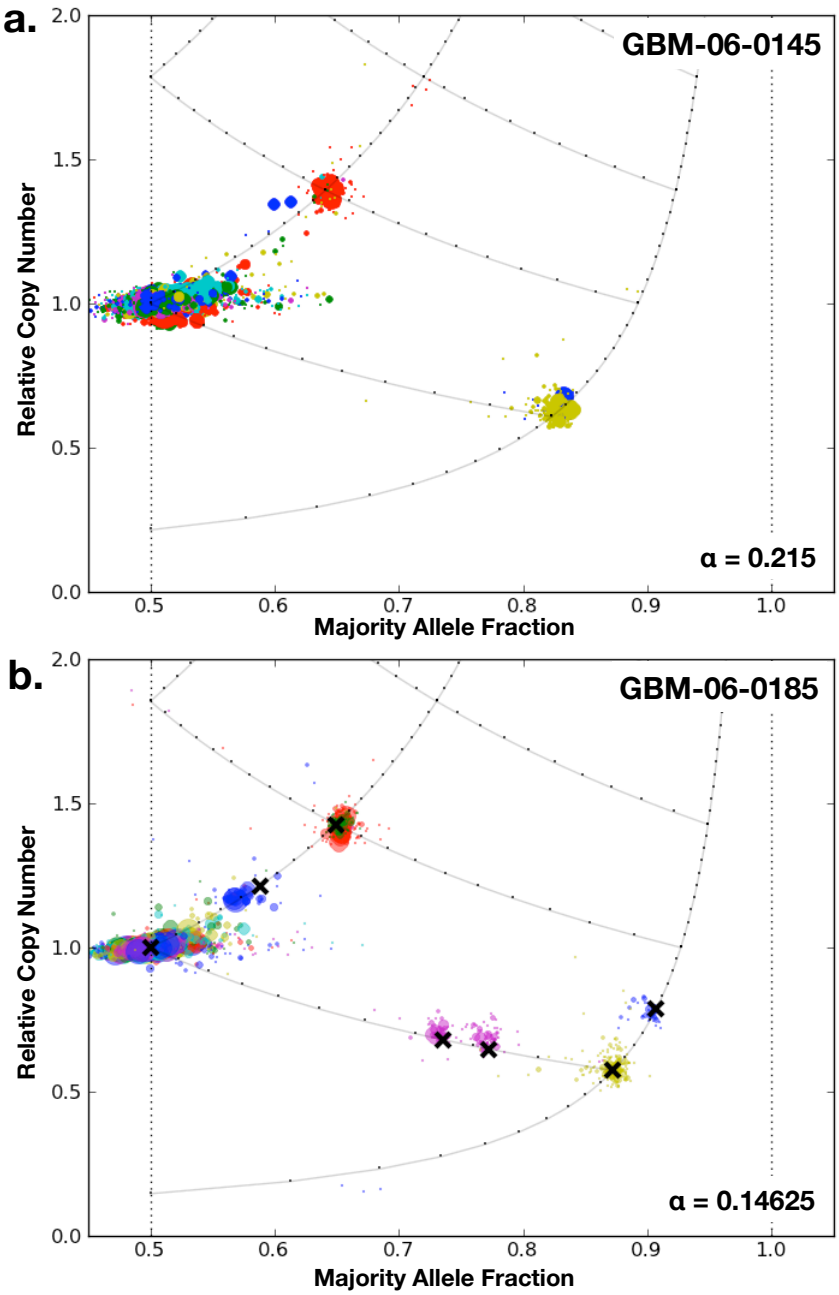

Figure 5.7: Allelic state diagrams for two GBM tumors: (a) GBM-06-0145 and (b) GBM-06-0185. The fitted parameters found normal contamination at 21.5% and 14.6%, respectively. Notice that in (a) there are only clonal allelic states and no evidence of transitional allelic states, indicating that GBM-06-0145 is a monoclonal tumor. However, in (b), there are both clonal states and multiple transitional allelic states. Since the blue and purple transitional allelic states feature three different mixture percentages, this polyclonal tumor must consist of at least three subclones. The black X's plotted in (b) represent "landmark" allelic states that will be used to determine the clonal mixture of GBM-06-0185.

whose clonal allelic profiles (CAPs) are presented in Figure 5.9. There are two important things to note from the four clones presented here. Firstly, as described before, the fact that each clone's mixture proportions is different from all others helps us to phase the allelic states across the whole genome into allele-specific copy number profiles for each clone.

Secondly, all clones appear to have derived from clone A. Each derivative clone shares all of the events found in clone A, suggesting that clone A is the progenitor of clones B, C, and D. It's unclear if this set of clones evolved linearly, in a step-wise progression, or if clone B and clones C & D represent different lineages. These latter three clones differ by the deletions in chr6q, where clone B features a set of focal deletions while clones C & D have lost all of chr6q. These are not mutually exclusive events, so it's possible that clone C had derived from B, inheriting its focal deletions and subsequently deleting the remainder of chr6q. However, nothing precludes clones B and C from deriving directly from clone A and independently deleting parts of chr6q. It is interesting that it is clone D, the last clone of the tree, that becomes the dominant clone in the tumor population according to mixture proportion, suggesting that the events unique to this clone (e.g. amplification of chr9) may have provided a growth advantage to this clone.

The clonal allelic profiles for GBM-06-0152 are shown in Figure 5.10. This tumor is interesting because the amplification of chr7, a characteristic of approximately 40% of GBM tumors [31, 56], does not occur until clone B. Recall that this sample was shown in the previous chapter to have two double minute chromosomes, one with

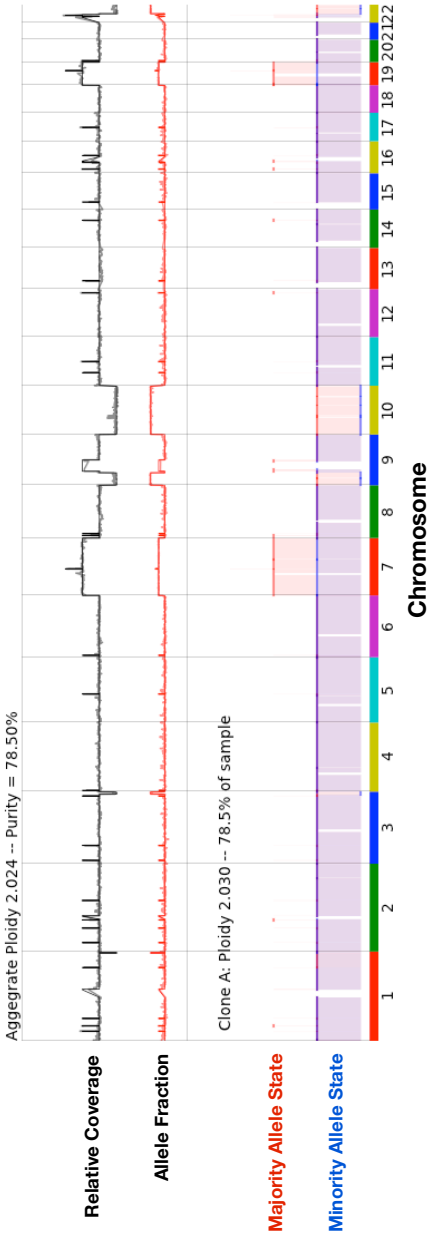

Figure 5.8: Monoclonal allelic profile for GBM-06-0145. The "Relative Coverage" and "Allele Fraction" displayed at the top of the plot shows both the observed output by BamBam and the imputed coverage and allele fraction generated by rendering the data as the mixture of the single clone and normal contamination. The comparison of real versus modeled data shows very strong agreement. The clone's allelic profile below shows the majority and minority allelic states for the tumor genome, showing amplification of one copy of entire chr7 and chr19, complete loss of one copy of chr10, and arm-level loss of chr9p.

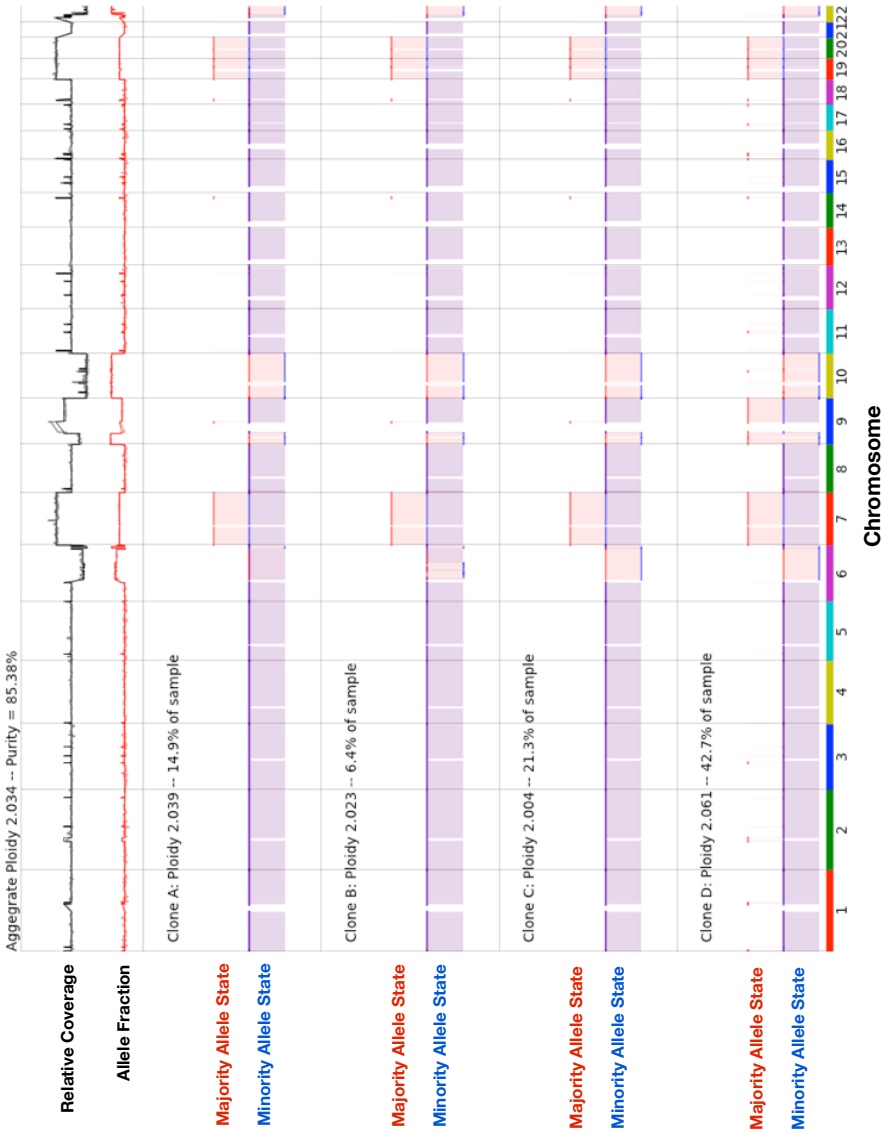

Figure 5.9: Polyclonal allelic profiles for GBM-06-0185. A total of 4 distinct clones were identified in this tumor, with clone D determined to be the dominant clone of the population, comprising 42.7% of the tumor sample. All clones have single-copy amplifications of chr7, chr19 & chr20, single copy loss of chr10 and chr22, and loss of chr9p in common. Clones B, C, & D all have deletions in chr6, but clone B's deletions are focal while clones C & D display arm-level loss of chr6q. Clone D is further distinguished by amplification of the intact copy of chr9.

MDM2 & CDK4 and another containing EGFR, that were borne out of a chromothripsis-like event. While extremely amplified genomic regions are difficult to render in these clonal allelic profiles, you can see evidence of the deletions related to these events on chr12 in clone A, suggesting that these double minutes occurred early in this tumor's development. It's possible that the early focal amplification of EGFR may have played a role in the later emergence of chr7 amplification.

The clonal evolution described by the allelic states for sample GBM-06-1086 has a few interesting aspects. The first subtle thing to notice in its allelic profile, shown in Figure 5.11, is that the focal deletion of CDKN2A occurs does not occur until clone B, suggesting it occurred after the complete loss of chr9 observed first in clone A. This is strong evidence supporting the hypothesis presented in Section 4.3.1 that focal losses of CDKN2A likely occur after arm-level or entire chromosomal losses of chr9. The second interesting aspect is that clones C & D exhibit hemizygous losses of 13 different *entire* chromosomes. Clone D takes this one step further by deleting its last copy of chr18, as well as amplifying chr19. This reduces the ploidy of both clones C & D to 1.31, from the approximately normal ploidy shared by the other two clones (ploidy = 1.95). It's fascinating how cells that have lost almost 30% of their genomic content, becoming haploid for most chromosomes and losing both copies of one chromosome, can not only survive but, given the 41.8% mixture proportion of clone D, apparently thrive in a population of tumor cells.

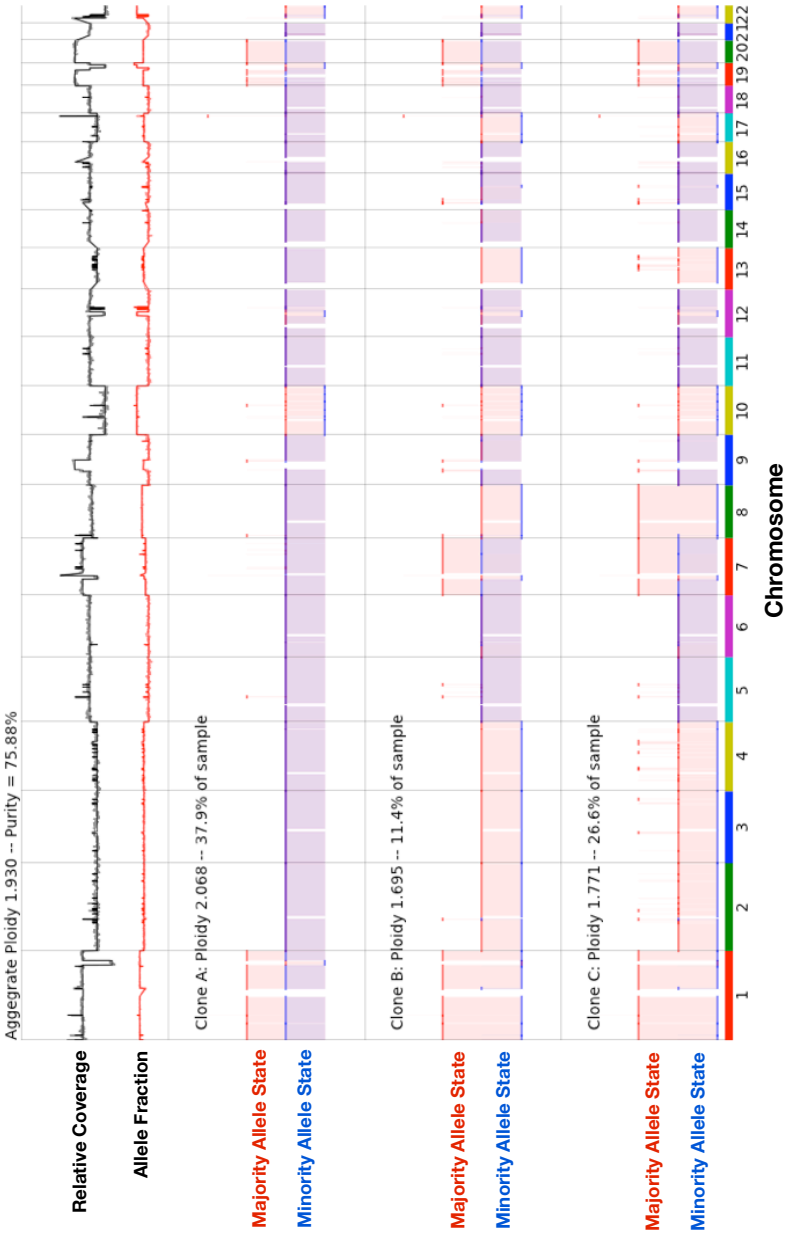

Figure 5.10: Polyclonal allelic profiles for GBM-06-0152. Three clones were identified in this tumor sample that has an estimated 24.1% normal contamination. All clones share amplifications of chr1, chr19 & chr20, deletions of chr10 & chr22, and focal losses of chr12 related to the chromothripsis-like event that created two DMs described in the previous chapter. Clones B & C exhibit amplification of chr7 and deletions of the non-amplified copy of chr1 as well as chr2, chr3, chr4, chr8, chr13, and chr17. Clone C further amplifies the remaining copy of chr8.

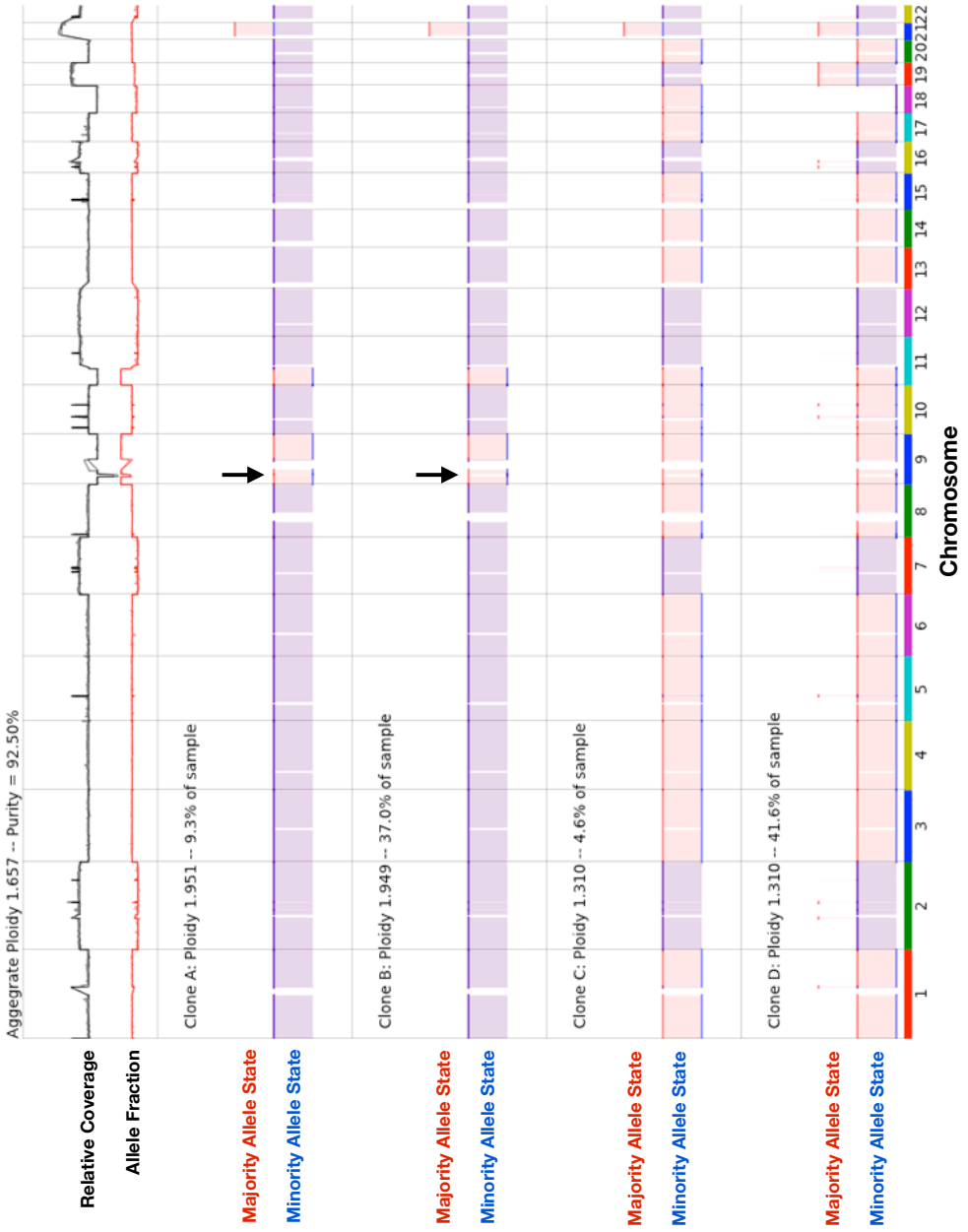

Figure 5.11: Polyclonal allelic profiles for GBM-06-1086. Four clones were identified in this tumor sample that has an estimated 7.5% normal contamination. All clones share amplification of chr21 and deletions of chr9 & chr11p. Clones C & D exhibit significant chromosomal loss, deleting chr1, chr3, chr4, chr5, chr6, chr8, chr10, chr13, chr14, chr15, chr17, chr18, and chr20. The dominant clone D, making up 41.6% of the tumor sample, further deletes the sole remaining copy of chr18 and amplifies chr19. The black arrows indicate the position of CDK2NA in clones A & B, highlighting the arrival of the focal deletion of CDKN2A in the latter clone.

5.5.2 Clonality of squamous cell carcinomas of the lung

Whole genome data for nine squamous cell carcinomas of the lung (LUSC) sequenced by TCGA were analyzed by these methods to infer clonality. The allelic state diagrams of two tumors are shown in Figure 5.12. From the greater number of transitional allelic states evident in these two samples, it appears that these LUSC tumors exhibit a much higher degree of clonality compared to the GBM tumors described above.

The tumor sample LUSC-66-2756 shown in Figure 5.12(b) exhibits numerous highly amplified states at ASD vertices (states common among all major clones) and ASD edges (states shared by only a subset of major clones). A wide variety of mixture proportions is evident from the almost continuous set of different positions of point clusters along, and in between, ASD edges, suggesting that this sample is highly polyclonal. Another interesting feature of this sample is that none of its genome is found in the single copy loss allelic state (1,0). This may have occurred via a genome doubling event where the tumor genome was briefly tetraploid (N=4), then a series of chromosomal deletions led to either single copy gain, "normal," or CN-LOH allelic states. Genome doubling events are believed to often occur in serous ovarian carcinomas to explain how large portions of their genomes are observed in the CN-LOH allelic state [15].

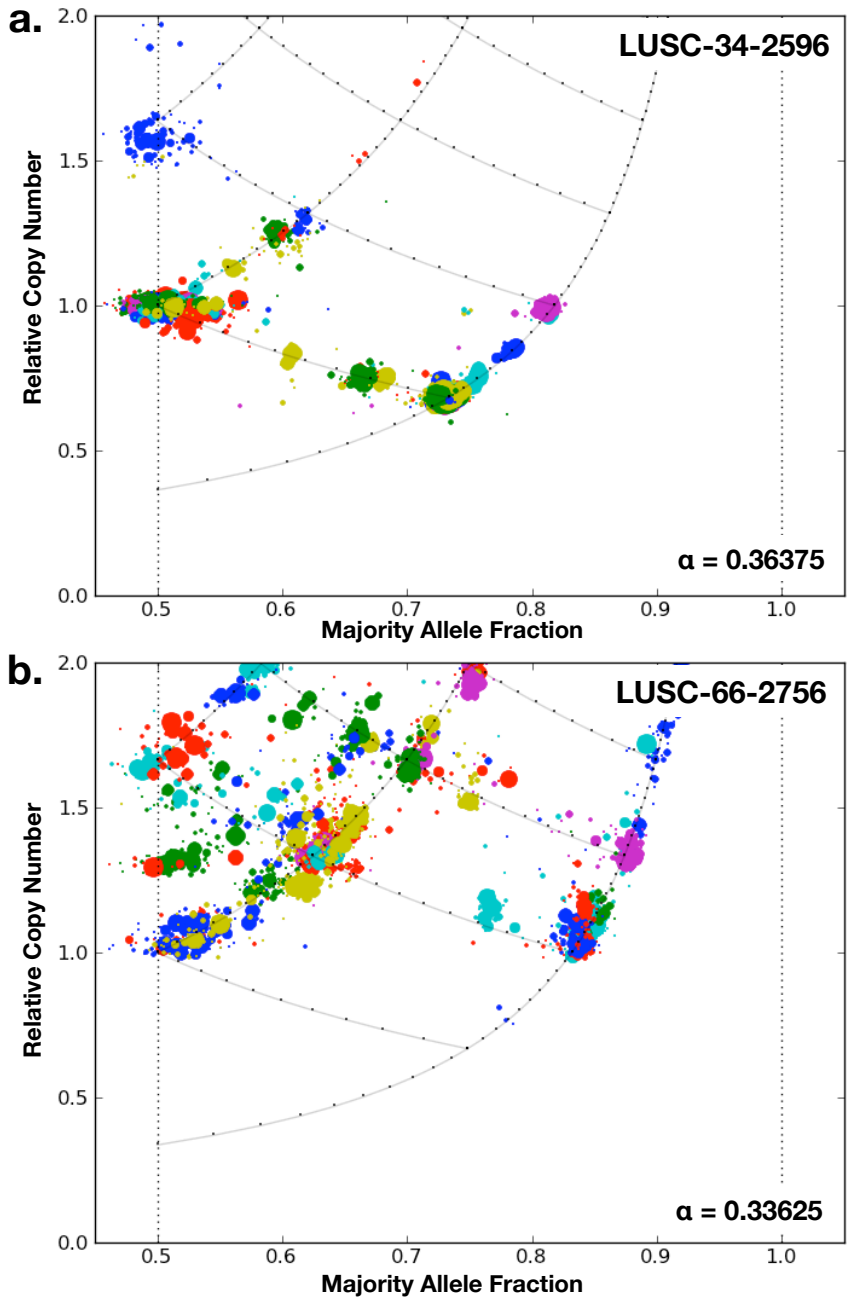

Figure 5.12: Allelic state diagrams for two LUSC tumors: (a) LUSC-34-2596 and (b) LUSC-66-2756. The fitted parameters found normal contamination at 36.4% and 19.6%, respectively. These tumors are highly polyclonal, with LUSC-66-2756 exhibiting a large number of highly amplified allelic states and zero single copy loss allelic states, which could have resulted from a genome doubling event.

5.5.3 Phased mutations to allelic states

To visualize the phased mutations onto allelic states, we will use a slightly modified allelic state diagram, called the dual allelic state diagram (dual ASD). Noting from Equation 3.9 that minority allele fraction is the complement of majority allele fraction ($AF_{min} = 1.0 - AF_{maj}$), we can construct a dual ASD by placing a mirror image of the ASD in order to display the location of the minority allelic states. Mutations phased to germline variants corresponding to the majority allele, minority allele, or neither, are plotted on the dual ASD. By determining which allelic state (majority or minority) the mutations are nearest and using their phased status (if any), we can infer the the timing of the mutations.

An example dual ASD is shown in Figure 5.13 to help explain how to interpret it. This dual ASD presents a series of mutations that are phased to germline variants belonging to either the majority or minority alleles in two different allelic states. Each mutation's allele fraction is corrected using the equations presented in Section 5.4.3 and placed onto the dual ASD. Based on their phase and mutant allele fraction, the dual ASD helps you identify on how many copies of the majority (or minority) alleles the mutation is present. In the case of the amplification presented in Figure 5.13, mutations present on both alleles versus those on just one of the amplified copies are readily distinguished, allowing you to determine visually whether a mutation occurred before or after the amplification. Similarly, for mutations phased to the minority allele, we should see them having $MAF_c \leq 0.5$ for all but the "normal" allelic state where

their phased assignment is not meaningful.

Not shown on Figure 5.13 are unphased mutations, where as discussed above, any time an unphased mutation has $MAF_c >> 0.5$ in unbalanced allelic states, the mutation must be phased to the majority allele. Otherwise, the mutation exists in single copy or less, and thus we cannot use its allele fraction to determine whether it's phased to the majority or minority alleles.

The dual ASDs for tumor GBM-06-0145 are shown in Figure 5.14. 6 regions on these diagrams are highlighted to help with interpretation of these diagrams on real data. Regions (a) and (b) show mutations that were directly phased via nearby germline variants to the majority allele. Only the 2 mutations in (b) are found to have the MAF_c corresponding to an amplified mutation and hence are expected to have been present prior to the duplication event (both mutations are located in intergenic regions). Most majority-phased mutations are found in region (a), corresponding to mutations at single copy number, suggesting that these mutations occurred post-duplication. An unphased, missense mutation in DOCK8 is found in the single copy loss allelic state, meaning that the only copy of DOCK8 remaining in this tumor is in the mutated state. Inactivation of DOCK8 through homozygous deletion has been linked to progression of lung cancers, so the lack of wildtype DOCK8 in this GBM tumor may have played a role in its tumorigenesis [86]. Figure 5.14 also demonstrates the high degree of variation in estimates of MAF_c from these average coverage whole genomes. As TCGA increases tumor sequencing depth from 30x to 60x, these estimates will improve.8

The most striking thing about the dual ASD for tumor LUSC-34-2596, shown

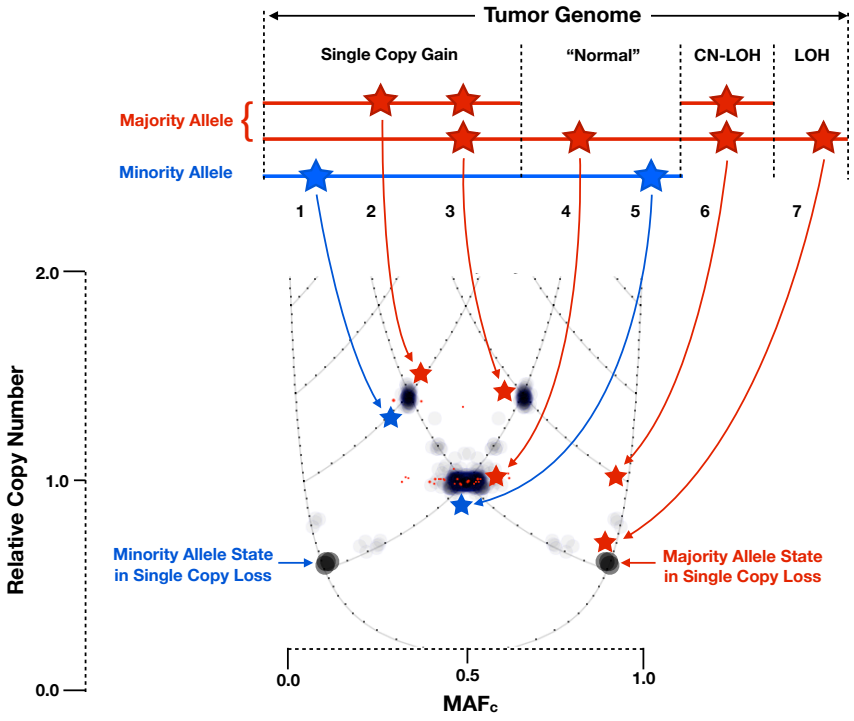

Figure 5.13: Example of seven phased mutations on the dual ASD. A representative region on the tumor genome is shown with blue and red horizontal lines representing chromosomes in the tumor mapping to this region, consisting of a subregion in the single copy gain allelic state, a subregion in the "normal" allelic state, a subregion exhibiting CN-LOH, and a subregion exhibiting LOH. Mutations 1, 2, and 3 are found in the amplified subregion, where mutations 2 & 3 are majority-phased (red stars) and mutation 1 is minority-phased (blue star). Mutations 4 & 5 are found in the "normal" allelic state, where mutation 4 is majority-phased and mutation 5 is minority-phased. Both regions exhibiting LOH have one mutation each phased to the sole remaining allele, and thus mutations 6 & 7 are majority-phased. The dual ASD below shows where each of these mutations would be found, using each mutation's corrected allele fraction, MAF_c, to determine its location along the x-axis. Note the different placement of mutations 2 & 3 in the single-copy gain allelic state, where mutation 3, the mutant allele that exists on both chromosomes with the majority allele, is found near the single copy gain majority allelic state. Mutation 2 is found near the single copy gain *minority* allelic state, correctly identifying that the mutation exists on only one copy of the majority allele and hence is likely to occur after the duplication event. Finally, note that the minority-phased mutations in blue are all found in the left-half of the dual ASD.

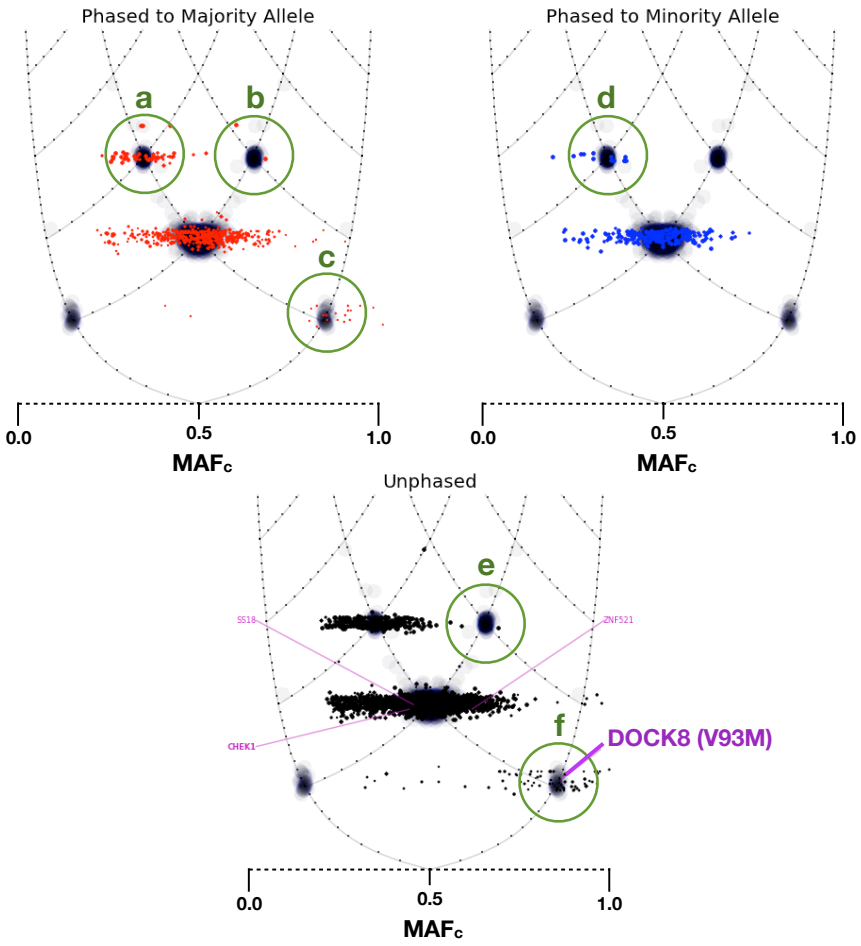

Figure 5.14: Phased mutations on the dual ASD for tumor GBM-06-0145. 7 regions are encircled on these plots: (a) majority-phased mutations in an amplified allelic state, but presenting with MAF_c suggesting the mutation is only on one of the two copies, (b) majority-phased mutations in an amplified allelic state with MAF_c suggesting mutations are present on both amplified copies, (c) majority-phased mutations with an allele fraction consistent with the LOH allelic state, (d) minority-phased mutations in the majority-amplified allelic state with MAF_c consistent with single copy, (e) unphased mutations with MAF_c consistent with amplified allelic state, and (f) unphased mutations with MAF_c consistent with LOH allelic state. LOH has left this cell with only a mutated copy of the gene DOCK8.

in Figure 5.15, is the sheer number of mutations, phased or unphased, across all expected allelic states. Compared to the previous GBM tumor, it is clear that the mutation rate of LUSC-34-2596 is significantly higher. This is expected since lung tumors exhibit some of the highest mutation rates among the cancers studied thus far by TCGA [34]. We observe a great number of both majority- and minority-phased mutations in the balanced-amplified allelic state (2,2) at the expected $MAF_c \approx 0.5$, labeled (a) and (b) in Figure 5.15. We also observe a large cluster of mutations to the left of the $(2, 2)$ balanced-amplified state that correspond to mutations in the $(3, 1)$ amplified state at single copy number. Moving down to the $(2, 1)$ amplified state, as in the GBM sample, we observe many single copy mutations on the majority allele and hence likely occurring post-duplication. However, unlike the GBM case, we observe a large cluster of mutations to the right (present in two copies), representing mutations that were likely there before the duplication event.

The location of a majority-phased missense mutation in NDRG1, a gene recently discovered to be up-regulated in squamous cell lung cancer [93], is found in a genomic region in between the "normal" and single copy loss allelic states. Its MAF_c is approximately equal to the allele fraction of the genomic region, suggesting that the mutation exists on both clones (i.e. the clone with "normal" allelic state and the clone with single copy loss allelic state). This is evidence that the mutation occurred early, prior to the emergence of the second clone featuring the new deletion, and that the deletion contained the wildtype version of NDRG1.

The location of three unphased mutations, BRAF, DNMT3A, and TP53, are

also highlighted in Figure 5.15. The nonsense mutation in TP53 is found in a CN-LOH state and has a mutant allele fraction that precisely corresponds to the CN-LOH allele fraction, meaning that this tumor has deleted one copy of TP53, knocked out the remaining copy via mutation, and then amplified the mutant allele. The region encompassing BRAF was highly amplified, and it is clear from BRAF's MAF_c that the mutation occurred prior to or early in its amplification. BRAF mutations occur frequently in melanomas but have been recently discovered in a small percentage of non-small cell lung carcinomas [20, 12]. Since more than half of the copies are mutated, the mutation could not have occurred after the amplification process had finished unless BRAF was independently and identically mutated on multiple copies, a highly improbable event. A mutation in DNMT3A, a gene whose loss is implicated in lung cancer and other tumor types [28, 97, 45], is found in the "normal" allelic state and has the expected $MAF_c \approx 0.5$. In all of these cases, mutations to these genes must have occurred early during tumorigenesis as they are present in all (or, in the case of BRAF, at least the majority) of the tumor's major clones. Since these are genes known to be drivers in multiple tumor types, this suggests that one or more of these mutations may be drivers of this particular tumor.

Table 5.2 summarizes the phaseable mutations for 12 GBM and 8 LUSC tumors. Again, note the higher overall rate of mutations in the LUSC tumors relative to the GBM tumors. Also, it's clear that significantly more mutations are found at single copy within the amplified regions of the GBM tumors, whereas in LUSC tumors we see no consistent bias, with some tumors having more mutations in single copy and others

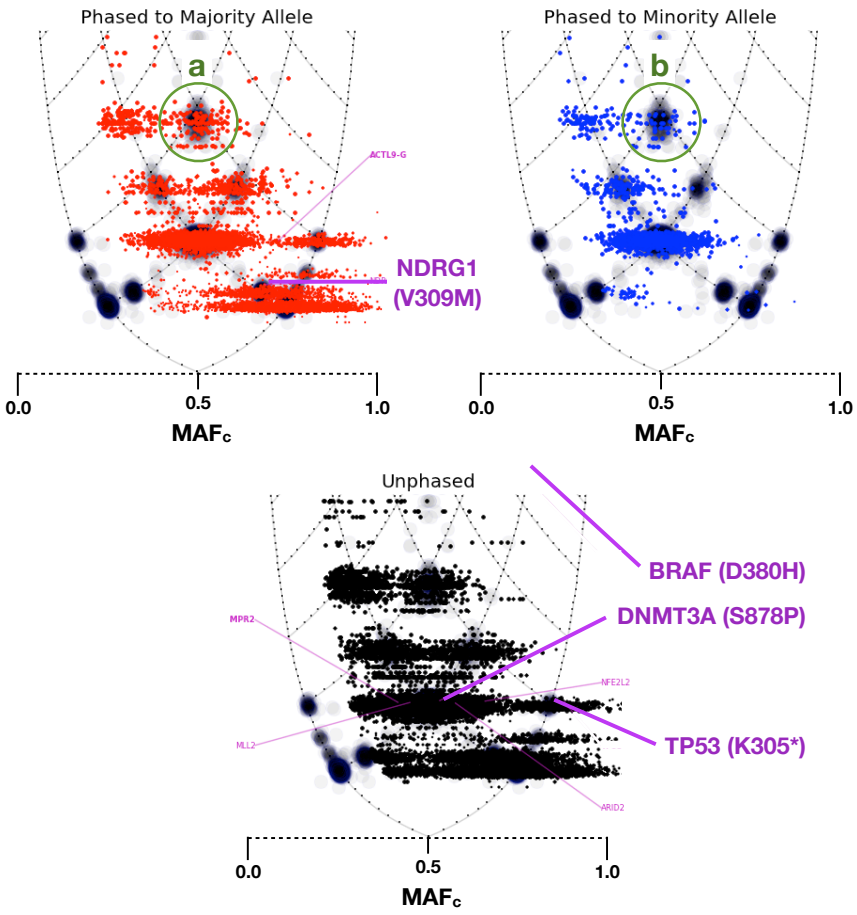

Figure 5.15: Phased mutations on the dual ASD for tumor LUSC-34-2596. The two encircled regions, (a) and (b), show a number of mutations phased to the majority and minority alleles, respectively, in the balanced amplified allelic state (2,2). One majority-phased mutation in NDRG1 is found in a transitional allelic state with matching MAF_c. The locations of two missense mutations, in BRAF & DNMT3A, and one nonsense mutation in TP53 are shown in the unphased plot, placing BRAF in a highly amplified allelic state, DNMT3A in the "normal" allelic state, and TP53 in the CN-LOH state.

having more mutations in double copy (or more). Assuming the mutation rate remains constant throughout the development of these tumors, then this suggest that amplification usually occurs early in the development of the GBM tumors, before most other mutations occur, whereas in LUSC tumor development, amplification may occur at any stage. Mutations and copy number alterations are frequent occurrences of LUSC tumor development, with large numbers of mutations occurring prior to and after amplification events.

Another possibility is that the mutation rate does not remain constant during development of GBM tumors. Suppose that amplification of the growth factor EGFR, a common event in these GBM tumors, increases the cell's rate of growth and subsequently reduces the cell's ability to correct mistakes made during genome replication, thereby increasing the mutation rate per cellular division. This would also explain the enrichment of mutations present in single copy within amplified allelic states. However, without knowledge of the number of generations that occur before and after EGFR amplification, we cannot determine if the effect we observe is due to timing bias or mutation rate change.

5.6 Discussion

The methods presented here demonstrate our ability to computationally dissect a tumor's cell population using whole genome sequencing data. We present an effective way to visually assess a tumor sample's clonality using an allelic state diagram.

Table 5.2: Phased mutations in GBM and LUSC tumors.

Sample	Total	Maj-Phased	In Amp.	Single Copy	Min-Phased
GBM-06-0188	1,595	67	2	1 (50%)	19
GBM-06-0648	1,600	68	9	5 (56%)	34
GBM-06-0877	4,792	367	130	84 (65%)	145
GBM-14-1454	2,539	291	55	38 (69%)	80
GBM-06-0152	2,896	269	47	35 (74%)	80
GBM-26-1438	2,820	239	36	32 (89%)	86
GBM-06-0155	7,167	1,108	99	93 (94%)	440
GBM-06-0214	2,827	237	31	29 (94%)	83
GBM-06-0145	3,800	511	56	53 (95%)	225
GBM-14-0786	5,863	783	117	114 (97%)	241
GBM-06-1086	3,993	412	10	10 (100%)	129
GBM-06-0185	1,195	20	2	2 (100%)	14
Total GBM	**41,087**	**4,372**	**594**	**496 (84%)**	**1,576**
LUSC-66-2756	45,995	7,336	6,869	2,107 (31%)	2,104
LUSC-56-1622	19,343	3,159	286	99 (35%)	1,016
LUSC-60-2695	6,265	830	408	148 (36%)	302
LUSC-43-3394	18,905	4,162	694	267 (38%)	1,184
LUSC-66-2757	11,400	1,477	374	177 (47%)	498
LUSC-60-2713	21,337	4,290	503	262 (52%)	1,207
LUSC-34-2596	30,794	7,166	663	349 (53%)	1,410
LUSC-60-2722	32,485	5,101	504	319 (63%)	1,682
Total LUSC	**186,524**	**33,521**	**10,301**	**3,728 (36%)**	**9,403**
Total LUSC*	**140,529**	**26,185**	**3,432**	**1,621 (47%)**	**7,299**

Total = Number of phased and unphased mutations called, **Maj-Phased** = Number of majority-phased mutations, **In Amp.** = Number of majority-phased mutations in regions of amplified allelic state, **Single Copy** = Number of majority-phased mutations in regions of amplified allelic state, but having MAF_c corresponding to single copy, **Min-Phased** = Number of minority-phased mutations. **Total LUSC*** is the total for all LUSC tumors excluding the outlier case of LUSC-66-2756 that has almost double the number of mutations of any other LUSC tumor.

We show how the clonal mixture of a tumor can be determined, decomposing the tumor population into all of the major clones of the tumor cell population and estimate normal contamination in order to fully account for the copy number and allele fraction data measured by BamBam. We further show that whole genome allelic profiles of all major clones can be fully determined and phased, allowing us to infer the phylogenetic tree of polyclonal tumor genomes to time the emergence of clone-specific copy number alterations. Finally, by using phasing and mutant allele fraction, we can time the emergence of mutations with respect to copy number alterations of the chromosomal region that carry them.

The clonality and timing information helps us gain a better appreciation for the dynamic nature of individual tumors. We observe tumors belonging to the same tumor type exhibiting a wide range of clonality, from monoclonal to highly polyclonal. We also find that certain tumor types reflect an overall higher degree of clonality than others, possibly suggesting that capacity to nurture multiple clones is in some way tissue-dependent. Most importantly, all of this information can be discovered from just a single tumor biopsy, making such methods feasible for future applications in the clinical environment, as will be further discussed below in Section 6.2.

Chapter 6

Conclusion

In this thesis, numerous whole genome sequencing analysis methods were described and used on sequencing data from real tumors to prove their individual and collective efficacies. Starting with BamBam and Bridget, we discover the full complement of germline and somatic variants from the very small alterations (e.g. point mutations) to large structural events (e.g. deletions of entire chromosomes), and everything in between. While this knowledge is powerful in isolation, through clever integration of these data we gain a much truer understanding of the tumor.

In this concluding chapter, I will briefly describe some of the directions I foresee this work taking in both research and clinical arenas.

6.1 Clone-directed mutation discovery

By determining the clonal mixture, we could determine the expected mutant allele fraction for mutations specific to each clone in the tumor. This information could be used to adjust the mutation model used by BamBam, enabling confident calls to be made at lower mutation allele frequencies than they would otherwise given the prior knowledge that clones exist in the sample that should produce mutations with certain mutant allele frequencies.

There is a lower limit to our ability to detect variants that is dependent on the coverage of the tumor genome. For example, from a 30x tumor whole genome with 20% normal contamination, we expect only 5 reads on average to come from a clone that comprises 20% of the tumor population (assuming a "normal" allelic state). Since most mutations are found on just one of the two parental alleles, only 2-3 of those reads are likely to exhibit the mutation. Therefore, increased coverage, possibly by integrating high coverage exomes (200-1000x), would be necessary to yield sufficient coverage to identify these clone-specific mutations.

The tweak to the current algorithm suggested above would be a good first step to extracting clone-specific mutations, but a mutation model that incorporates the phylogenetic relationships between clones may be much more powerful. This method would help find mutations that are specific to each clone, but would also be able to infer the node in the tree where the mutation likely emerged, to better model the situation where multiple clones may share the mutation. A phylogenetic-based mutation model

could also prove useful to analyze the mutations of related samples, such as a primary tumor and its metastases, to reconstruct the mutational history of a cancer as it spreads.

6.2 Clinical applications

Presently, there is much hype surrounding the promise of genome sequencing in improving the care and treatment of cancer. The hope is that with improved knowledge about the unique features of each tumor genome, physicians will be able to tailor treatment to more precisely target the cancer. Most efforts focus on specific oncogenic mutations, or genes commonly amplified in tumors, and therefore employ targeted sequencing (e.g. exome capture, or even more restrictive kits that capture ∼200 cancer genes targeted by drugs that either FDA-approved or in late-phase clinical trials) [24]. Less effort is applied to whole genome sequencing data since it currently generates a huge amount of data, most of which is likely to go unused as its relevance to cancer treatment is unknown or non-existent. However, the methods described in this thesis clearly suggest some novel clinical applications of whole genome sequencing data.

By integrating copy number and structural variants in regions exhibiting extreme amplification, double minute chromosomes (DM) were identified and reconstructed to determine their genetic content, accurately down to the base level. DNA reads from these DMs, born in a brain tumor, were readily detectable in the bloodstream of the same patients. This has obvious clinical relevance, as it suggests that brain tumors could be detected or monitored by a simple blood draw. If the creation of a DM is an early or

initiating event of tumorigenesis, then it's plausible that a sensitive blood-based assay could be developed to detect certain brain tumors even before they could be detected by more expensive means, such as an MRI.

One can envision a new class of sequencing-based assays that monitor DNA from blood samples to determine, in the *de novo* sense, if a region surrounding any known oncogene, such as EGFR, c-Myc, and MDM2, exhibit a clustered pattern of breakpoints indicative of an amplified DM. Collecting evidence of discordant reads across regions that have oncogenic potential should improve our ability to identify these regions even when the concentration of DMs in the bloodstream is low.

The ability to detect DMs in the bloodstream should extend to other cancers that commonly feature highly amplified DMs containing known oncogenes, such as EGFR in non-small cell lung cancer and c-Myc in acute myelogenous leukemia [89, 78]. In fact, our ability to detect DMs with this method may improve for tumor types where the bloodstream is more accessible to the tumor. Furthermore, drugs that specifically target genes commonly amplified in DMs may be prescribed based on evidence collected solely from the bloodstream, avoiding painful and dangerous tumor biopsies. Future studies on a wider panel of tumors and tumor types will help us learn how well this method can be extended to other cancers.

In my opinion, tumor clonality has even more potential in the clinic. The ability to determine a tumor's clonality and identify all of the major clones that comprise the growing tumor mass, all from the whole genome sequencing data of a single biopsy, opens up a wide variety of potential clinical applications. For example, suppose a

newly-diagnosed patient's tumor is biopsied and all of the major clones are identified via clonal analysis. The clinician could use this clonality analysis to tailor the patient's treatment according to the alterations specific to the clone furthest up the evolutionary tree, the progenitor tumor cell, with the hope that by treating the initial tumor mass, all derivative clones will also be targeted.

In a related twist, suppose a patient is diagnosed with a slow-growing tumor that can be safely monitored for a longer period of time before surgery or the beginning of chemotherapy. Clonal analysis of a series of biopsies could be performed and, by tracking the clonal composition of the tumor through time, the clinician could identify the clones that are growing most rapidly. By designing a treatment that targets not what is currently the dominant clone, but the clone that is becoming the dominant clone, we might more effectively treat the cancer.

Clonality analysis may also prove useful in better understanding the metastatic spread of cancers. A clonal analysis of a primary tumor and a series of metastases could determine all of the major clones present in the spreading tumor. By inspecting the clonal composition of the primary tumor and each metastasis, we could determine how each clone spreads and discover if there are particular clones that exhibit superior metastatic potential. If we can determine the characteristics unique to the metastatic clones, it may be possible to identify the emergence of these characteristics in the minor clones of another patient's primary tumor, providing an "early warning" for the likelihood of imminent metastasis.

6.3 Final thoughts

While this thesis presents a comprehensive range of techniques to explore and understand the cancer genome, I know the work presented here merely represents the "tip of the iceberg" of what is to be learned about cancer genome. The paradoxical nature of cancer research is such that, for every bit of information discovered, we both add knowledge and yet reduce our overall understanding by revealing further complexities of this disease. One day soon the balance of knowledge will tip toward our favor, and the days when cancer is considered a death sentence will be behind us.

Bibliography

[1] ACS. Cancer Facts & Figures 2012. *American Cancer Society*, 2012.

[2] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

[3] Stylianos E Antonarakis, Michael Krawczak, and David N Cooper. Disease-causing mutations in the human genome. *European Journal of Pediatrics*, 159(15):S173–S178, January 2000.

[4] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29, May 2000.

[5] Leonora Balaj, Ryan Lessard, Lixin Dai, Yoon-Jae Cho, Scott L Pomeroy, Xandra O Breakefield, and Johan Skog. Tumour microvesicles contain retrotransposon

elements and amplified oncogene sequences. *Nature Communications*, 2:180, February 2011.

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

[7] F Baskin, R N Rosenberg, and V Dev. Correlation of double-minute chromosomes with unstable multidrug cross-resistance in uptake mutants of neuroblastoma cells. *Proceedings of the National Academy of Sciences*, 78(6):3654–3658, January 1981.

[8] Y Ben-Neriah, G Q Daley, A M Mes-Masson, O N Witte, and D Baltimore. The chronic myelogenous leukemia-specific P210 protein is the product of the bcr/abl hybrid gene. *Science*, 233(4760):212–214, July 1986.

[9] J Bentley. DNA double strand break repair in human bladder cancer is error prone and involves microhomology-associated end-joining. *Nucleic Acids Research*, 32(17):5249–5259, September 2004.

[10] 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. *Proc Natl Acad Sci USA*, 104(50):20007–20012, December 2007.

[11] I Bièche, C Tomasetto, C H Régnier, C Moog-Lutz, M C Rio, and R Lidereau. Two distinct amplified regions at 17q11-q21 involved in human primary breast cancer. *Cancer Res*, 56(17):3886–3890, September 1996.

[12] Marcia S Brose, Patricia Volpe, Michael Feldman, Madhu Kumar, Irum Rishi, Renee Gerrero, Eugene Einhorn, Meenhard Herlyn, John Minna, Andrew Nicholson, Jack A Roth, Steven M Albelda, Helen Davies, Charles Cox, Graham Brignell, Philip Stephens, P Andrew Futreal, Richard Wooster, Michael R Stratton, and Barbara L Weber. BRAF and RAS mutations in human lung cancer and melanoma. *Cancer Research*, 62(23):6997–7000, December 2002.

[13] 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. *Proc Natl Acad Sci USA*, 105(35):13081–13086, September 2008.

[14] S M Carroll, M L DeRose, P Gaudray, C M Moore, D R Needham-Vandevanter, D D Von Hoff, and G M Wahl. Double minute chromosomes can be produced from precursors derived from a chromosomal deletion. *Molecular and Cellular Biology*, 8(4):1525–1533, April 1988.

[15] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, Rameen Beroukhim, David Pellman, Douglas A Levine, Eric S Lander, Matthew Meyerson, and Gad Getz. Absolute quantification of somatic DNA alterations in human cancer. *Nature*, 30(5):413–421, April 2012.

[16] James A Goodrich Robert Tjian Catherine J Thut. Repression of p53-mediated transcription by MDM2: adualmechanism. *Genes & Development*, 11(15):1974, August 1997.

[17] Michael A Chapman, Michael S Lawrence, Jonathan J Keats, Kristian Cibulskis, Carrie Sougnez, Anna C Schinzel, Christina L Harview, Jean-Philippe Brunet, Gregory J Ahmann, Mazhar Adli, Kenneth C Anderson, Kristin G Ardlie, Daniel Auclair, Angela Baker, P Leif Bergsagel, Bradley E Bernstein, Yotam Drier, Rafael Fonseca, Stacey B Gabriel, Craig C Hofmeister, Sundar Jagannath, Andrzej J Jakubowiak, Amrita Krishnan, Joan Levy, Ted Liefeld, Sagar Lonial, Scott Mahan, Bunmi Mfuko, Stefano Monti, Louise M Perkins, Robb Onofrio, Trevor J Pugh, S Vincent Rajkumar, Alex H Ramos, David S Siegel, Andrey Sivachenko, A Keith Stewart, Suzanne Trudel, Ravi Vij, Douglas Voet, Wendy Winckler, Todd Zimmerman, John Carpten, Jeff Trent, William C Hahn, Levi A Garraway, Matthew Meyerson, Eric S Lander, Gad Getz, and Todd R Golub. Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339):467–472, March 2011.

[18] S H Chen and C R Scott. Recombination between two 14-bp homologous sequences

as the mechanism for gene deletion in factor IX Seattle 1. *Am J Hum Genet*, 47(6):1020–1022, December 1990.

[19] E D Coene, V Schelfhout, R A Winkler, A M Schelfhout, N Van Roy, M Grootecla es, F Speleman, and C R De Potter. Amplification units and translocation at chromosome 17q and c-erbB-2 overexpression in the pathogenesis of breast cancer. *Virchows Arch*, 430(5):365–372, May 1997.

[20] Helen Davies, Graham R Bignell, Charles Cox, Philip Stephens, Sarah Edkins, Sheila Clegg, Jon Teague, Hayley Woffendin, Mathew J Garnett, William Bottomley, Neil Davis, Ed Dicks, Rebecca Ewing, Yvonne Floyd, Kristian Gray, Sarah Hall, Rachel Hawes, Jaime Hughes, Vivian Kosmidou, Andrew Menzies, Catherine Mould, Adrian Parker, Claire Stevens, Stephen Watt, Steven Hooper, Rebecca Wilson, Hiran Jayatilake, Barry A Gusterson, Colin Cooper, Janet Shipley, Darren Hargrave, Katherine Pritchard-Jones, Norman Maitland, Georgia Chenevix-Trench, Gregory J Riggins, Darell D Bigner, Giuseppe Palmieri, Antonio Cossu, Adrienne Flanagan, Andrew Nicholson, Judy W C Ho, Suet Y Leung, Siu T Yuen, Barbara L Weber, Hilliard F Seigler, Timothy L Darrow, Hugh Paterson, Richard Marais, Christopher J Marshall, Richard Wooster, Michael R Stratton, and P Andrew Futreal. Mutations of the BRAF gene in human cancer. *Nature*, 417(6892):949–954, June 2002.

[21] N D Dees, Q Zhang, C Kandoth, M C Wendl, W Schierding, D C Koboldt, T B Mooney, M B Callaway, D Dooling, E R Mardis, R K Wilson, and L Ding. Mu-

SiC: Identifying mutational significance in cancer genomes. *Genome Research*, 22(8):1589–1598, August 2012.

[22] M F Denissenko, A Pao, G P Pfeifer, and M Tang. Slow repair of bulky DNA adducts along the nontranscribed strand of the human p53 gene may explain the strand bias of transversion mutations in cancers. *Oncogene*, 16(10):1241–1247, March 1998.

[23] B J Druker, M Talpaz, D J Resta, B Peng, E Buchdunger, J M Ford, N B Lydon, H Kantarjian, R Capdeville, S Ohno-Jones, and C L Sawyers. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N Engl J Med*, 344(14):1031–1037, April 2001.

[24] Michael Eisenstein. Foundation Medicine. *Nature*, 30(1):14–14, January 2012.

[25] B Elliott and M Jasin. Double-strand breaks and translocations in cancer. *Cellular and molecular life sciences : CMLS*, 59(2):373–385, February 2002.

[26] Bernard M Fine, Martin Stanulla, Martin Schrappe, Minh Ho, Susanne Viehmann, Jochen Harbott, and Linda M Boxer. Gene expression patterns associated with recurrent chromosomal translocations in acute lymphoblastic leukemia. *Blood*, 103(3):1043–1049, February 2004.

[27] Stefan Fröhling, Claudia Scholl, Ross L Levine, Marc Loriaux, Titus J Boggon, Olivier A Bernard, Roland Berger, Hartmut Döhner, Konstanze Döhner, Benjamin L Ebert, Sewit Teckie, Todd R Golub, Jingrui Jiang, Marcus M Schitten-

helm, Benjamin H Lee, James D Griffin, Richard M Stone, Michael C Heinrich, Michael W Deininger, Brian J Druker, and D Gary Gilliland. Identification of driver and passenger mutations of FLT3 by high-throughput DNA sequence analysis and functional assessment of candidate alleles. *Cancer Cell*, 12(6):501–513, December 2007.

[28] Qing Gao, Eveline J Steine, M Inmaculada Barrasa, Dirk Hockemeyer, Mathias Pawlak, Dongdong Fu, Seshamma Reddy, George W Bell, and Rudolf Jaenisch. Deletion of the de novo DNA methyltransferase Dnmt3a promotes lung tumor progression. *Proceedings of the National Academy of Sciences of the United States of America*, 108(44):18061–18066, November 2011.

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

[30] C D Greenman, E D Pleasance, S Newman, F Yang, B Fu, S Nik-Zainal, D Jones, K W Lau, N Carter, P A W Edwards, P A Futreal, M R Stratton, and P J Campbell. Estimation of rearrangement phylogeny for cancer genomes. *Genome Research*, 22(2):346–361, February 2012.

[31] W J Gullick. Prevalence of aberrant expression of the epidermal growth factor receptor in human cancers. *British Medical Bulletin*, pages 87–98, January 1991.

[32] P Hainaut. Patterns of p53 G-¿T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. *Carcinogenesis*, 22(3):367–374, March 2001.

[33] P Hainaut, T Hernandez, A Robinson, P Rodriguez-Tome, T Flores, M Hollstein, C C Harris, and R Montesano. IARC Database of p53 gene mutations in human tumors and cell lines: updated compilation, revised formats and new visualisation tools. *Nucleic Acids Res*, 26(1):205–213, January 1998.

[34] Peter S Hammerman, D Neil Hayes, Matthew D Wilkerson, Nikolaus Schultz, Ron Bose, Andy Chu, Eric A Collisson, Leslie Cope, Chad J Creighton, Gad Getz, James G Herman, Bruce E Johnson, Raju Kucherlapati, Marc Ladanyi, Christopher A Maher, Gordon Robertson, Chris Sander, Ronglai Shen, Rileen Sinha, Andrey Sivachenko, Roman K Thomas, William D Travis, Ming-Sound Tsao, John N Weinstein, Dennis A Wigle, Stephen B Baylin, Ramaswamy Govindan, and Matthew Meyerson. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, September 2012.

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

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

[37] A John Iafrate, Lars Feuk, Miguel N Rivera, Marc L Listewnik, Patricia K Donahoe,

Ying Qi, Stephen W Scherer, and Charles Lee. Detection of large-scale variation in the human genome. *Nature genetics*, 36(9):949–951, September 2004.

[38] Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, and Yoshihiro Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480–4, January 2008.

[39] K K Khanna and S P Jackson. DNA double-strand breaks: signaling, repair and the cancer connection. *Nature genetics*, 27(3):247–254, March 2001.

[40] Wigard P Kloosterman, Marlous Hoogstraat, Oscar Paling, Masoumeh Tavakoli-Yaraki, Ivo Renkens, Joost S Vermaat, Markus J van Roosmalen, Stef van Lieshout, Isaac J Nijman, Wijnand Roessingh, Ruben van 't Slot, Jose van de Belt, Victor Guryev, Marco Koudijs, Emile Voest, and Edwin Cuppen. Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. *Genome Biol*, 12(10):R103, 2011.

[41] Chandan Kumar-Sinha, Scott A Tomlins, and Arul M Chinnaiyan. Evidence of recurrent gene fusions in common epithelial tumors. *Trends in molecular medicine*, 12(11):529–536, November 2006.

[42] D E Larson, C C Harris, K Chen, D C Koboldt, T E Abbott, D J Dooling, T J Ley, E R Mardis, R K Wilson, and L Ding. SomaticSniper: identification of somatic

point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, January 2012.

[43] C Latham, A Zhang, A Nalbanti, S Månér, P Zickert, H Blegen, and A Zetterberg. Frequent co-amplification of two different regions on 17q in aneuploid breast carcinomas. *Cancer Genet Cytogenet*, 127(1):16–23, May 2001.

[44] I Lavon, D Zrihan, B Zelikovitch, Y Fellig, D Fuchs, D Soffer, and T Siegal. Longitudinal Assessment of Genetic and Epigenetic Markers in Oligodendrogliomas. *Clinical Cancer Research*, 13(5):1429–1437, March 2007.

[45] Timothy J Ley, Li Ding, Matthew J Walter, Michael D McLellan, Tamara Lamprecht, David E Larson, Cyriac Kandoth, Jacqueline E Payton, Jack Baty, John Welch, Christopher C Harris, Cheryl F Lichti, R Reid Townsend, Robert S Fulton, David J Dooling, Daniel C Koboldt, Heather Schmidt, Qunyuan Zhang, John R Osborne, Ling Lin, Michelle O'Laughlin, Joshua F McMichael, Kim D Delehaunty, Sean D McGrath, Lucinda A Fulton, Vincent J Magrini, Tammi L Vickery, Jasreet Hundal, Lisa L Cook, Joshua J Conyers, Gary W Swift, Jerry P Reed, Patricia A Alldredge, Todd Wylie, Jason Walker, Joelle Kalicki, Mark A Watson, Sharon Heath, William D Shannon, Nobish Varghese, Rakesh Nagarajan, Peter Westervelt, Michael H Tomasson, Daniel C Link, Timothy A Graubert, John F DiPersio, Elaine R Mardis, and Richard K Wilson. DNMT3AMutations in Acute Myeloid Leukemia. *N Engl J Med*, 363(25):2424–2433, December 2010.

[46] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor

Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[47] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858, November 2008.

[48] Marc M Loriaux, Ross L Levine, Jeffrey W Tyner, Stefan Fröhling, Claudia Scholl, Eric P Stoffregen, Gerlinde Wernig, Heidi Erickson, Christopher A Eide, Roland Berger, Olivier A Bernard, James D Griffin, Richard M Stone, Benjamin Lee, Matthew Meyerson, Michael C Heinrich, Michael W Deininger, D Gary Gilliland, and Brian J Druker. High-throughput sequence analysis of the tyrosine kinome in acute myeloid leukemia. *Blood*, 111(9):4788–4796, May 2008.

[49] Elaine R Mardis, Li Ding, David J Dooling, David E Larson, Michael D McLellan, Ken Chen, Daniel C Koboldt, Robert S Fulton, Kim D Delehaunty, Sean D McGrath, Lucinda A Fulton, Devin P Locke, Vincent J Magrini, Rachel M Abbott, Tammi L Vickery, Jerry S Reed, Jody S Robinson, Todd Wylie, Scott M Smith, Lynn Carmichael, James M Eldred, Christopher C Harris, Jason Walker, Joshua B Peck, Feiyu Du, Adam F Dukes, Gabriel E Sanderson, Anthony M Brummett, Eric Clark, Joshua F McMichael, Rick J Meyer, Jonathan K Schindler, Craig S Pohl, John W Wallis, Xiaoqi Shi, Ling Lin, Heather Schmidt, Yuzhu Tang, Carrie Haipek, Madeline E Wiechert, Jolynda V Ivy, Joelle Kalicki, Glendoria Elliott, Rhonda E

Ries, Jacqueline E Payton, Peter Westervelt, Michael H Tomasson, Mark A Watson, Jack Baty, Sharon Heath, William D Shannon, Rakesh Nagarajan, Daniel C Link, Matthew J Walter, Timothy A Graubert, John F DiPersio, Richard K Wilson, and Timothy J Ley. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med*, 361(11):1058–1066, September 2009.

[50] A W McPherson, C Wu, A Wyatt, S P Shah, C Collins, and S C Sahinalp. nFuse: Discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Research*, June 2012.

[51] Mitch McVey and Sang Eun Lee. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Cell*, 24(11):529–538, November 2008.

[52] Paula L Miliani de Marval, Everardo Macias, Claudio J Conti, and Marcelo L Rodriguez-Puebla. Enhanced malignant tumorigenesis in Cdk4 transgenic mice. *Nature*, 23(10):1863–1873, December 2003.

[53] Felix Mitelman, Bertil Johansson, and Fredrik Mertens. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nature genetics*, 36(4):331–334, April 2004.

[54] J K Moore and J E Haber. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, 16(5):2164–2173, May 1996.

[55] David K Moscatello, Marina Holgado-Madruga, Andrew K Godwin, Gloria Ramirez, George Gunn, Philip W Zoltick, Jaclyn A Biegel, Roberta L Hayes, and Albert J Wong. Frequent Expression of a Mutant Epidermal Growth Factor Receptor in Multiple Human Tumors. *cancerres.aacrjournals.org*.

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

[57] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, June 2011.

[58] G P Nielsen, A O Stemmer-Rachamimov, Y Ino, M B Moller, A E Rosenberg, and D N Louis. Malignant transformation of neurofibromas in neurofibromatosis 1 is associated with CDKN2A/p16 inactivation. *The American Journal of Pathology*, 155(6):1879–1884, December 1999.

[59] 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 Fa-

tima, 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.

[60] Tsutomu Nobori, Kaoru Miura, David J Wu, Augusto Lois, Kenji Takabayashi, and Dennis A Carson. Deletions of the cyclin-dependent kinase-4 inhibitor gene in multiple human cancers. *Nature*, 368(6473):753–756, April 1994.

[61] Hitoshi Okada and Tak W Mak. Pathways of apoptotic and non-apoptotic death in tumour cells. *Nat Rev Cancer*, 4(8):592–603, August 2004.

[62] C M Perou, T Sørlie, M B Eisen, M van de Rijn, S S Jeffrey, C A Rees, J R Pollack, D T Ross, H Johnsen, L A Akslen, O Fluge, A Pergamenschikov, C Williams, S X Zhu, P E Lønning, A L Børresen-Dale, P O Brown, and D Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, August 2000.

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

[64] Astrid M Roy-Engel. LINEs, SINEs and other retroelements: do birds of a feather flock together? *Frontiers in Bioscience*, 17:1345.

[65] V Ling C P Stanners S M Robertson. Co-amplification of double minute chromosomes, multiple drug resistance, and cell surface P-glycoprotein in DNA-mediated transformants of mouse cells. *Molecular and Cellular Biology*, 4(3):500, March 1984.

[66] Giuseppe Saglio, Alessandro Morotti, Giovanna Mattioli, Emanuela Messa, Emilia Giugliano, Gisella Volpe, Giovanna Rege-Cambrin, and Daniela Cilloni. Rational approaches to the design of therapeutics targeting molecular markers: the case of chronic myelogenous leukemia. *Ann N Y Acad Sci*, 1028:423–431, December 2004.

[67] J Z Sanborn, S C Benz, B Craft, C Szeto, K M Kober, L Meyer, C J Vaske, M Goldman, K E Smith, R M Kuhn, D Karolchik, W J Kent, J M Stuart, D Haussler, and J Zhu. The UCSC cancer genomics browser: update 2011. *Nucleic Acids Research*, 39(Database):D951–D959, December 2010.

[68] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. PID: the Pathway Interaction Database. *Nucleic Acids Res*, 37(Database issue):D674–9, January 2009.

[69] T Schepeler, P Lamy, J R Laurberg, N Fristrup, T Reinert, J Bartkova, L Tropia, J Bartek, T D Halazonetis, C-C Pan, M Borre, L Dyrskjøt, and T F Ørntoft. A high resolution genomic portrait of bladder cancer: correlation between genomic aberrations and the DNA damage response. *Nature*, August 2012.

[70] P V Schoenlein, D W Shen, J T Barrett, I Pastan, and M M Gottesman. Double

minute chromosomes carrying the human multidrug resistance 1 and 2 genes are generated from the dimerization of submicroscopic circular DNAs in colchicine-selected KB carcinoma cells. *Molecular Biology of the Cell*, 3(5):507–520, May 1992.

[71] Heidi Schwarzenbach, Dave S B Hoon, and Klaus Pantel. Cell-free nucleic acids as biomarkers in cancer patients. *Nature*, 11(6):426–437, May 2011.

[72] M Serrano, G J Hannon, and D Beach. A new regulatory motif in cell-cycle control causing specific inhibition of cyclin D/CDK4. *Nature*, 366(6456):704–707, December 1993.

[73] S P Shah. Computational methods for identification of recurrent copy number alteration patterns by array CGH. *Cytogenet Genome Res*, 123(1-4):343–351, January 2008.

[74] Jonathan S T Sham, Terence C-M Tang, Yan Fang, Li Sun, Lun-Xiu Qin, Qiu-Liang Wu, Dan Xie, and Xin-Yuan Guan. Recurrent chromosome alterations in primary ovarian carcinoma in Chinese women. *Cancer Genet Cytogenet*, 133(1):39–44, February 2002.

[75] S T Sherry, M H Ward, M Kholodov, J Baker, L Phan, E M Smigielski, and K Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29(1):308–311, January 2001.

[76] Johan Skog, Tom Würdinger, Sjoerd van Rijn, Dimphna H Meijer, Laura Gainche,

William T Curry, Bob S Carter, Anna M Krichevsky, and Xandra O Breakefield. Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. *Nature*, 10(12):1470–1476, November 2008.

[77] D J Slamon, B Leyland-Jones, S Shak, H Fuchs, V Paton, A Bajamonde, T Fleming, W Eiermann, J Wolter, M Pegram, J Baselga, and L Norton. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N Engl J Med*, 344(11):783–792, March 2001.

[78] M L Slovak, J P Ho, M J Pettenati, A Khan, D Douer, S Lal, and S T Traweek. Localization of amplified MYC gene sequences to double minute chromosomes in acute myelogenous leukemia. *Genes Chromosomes Cancer*, 9(1):62–67, January 1994.

[79] T F Smith and M S Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, March 1981.

[80] T Sørlie, C M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, M B Eisen, M van de Rijn, S S Jeffrey, T Thorsen, H Quist, J C Matese, P O Brown, D Botstein, P Eystein Lønning, and A L Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA*, 98(19):10869–10874, September 2001.

[81] Therese Sorlie, Robert Tibshirani, Joel Parker, Trevor Hastie, J S Marron, An-

drew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, Stephanie Geisler, Janos Demeter, Charles M Perou, Per E Lønning, Patrick O Brown, Anne-Lise Børresen-Dale, and David Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA*, 100(14):8418–8423, July 2003.

[82] 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.

[83] Clelia Tiziana Storlazzi, Thoas Fioretos, Cecilia Surace, Angelo Lonoce, Angela Mastrorilli, Bodil Strombeck, Pietro D'Addabbo, Francesco Iacovelli, Crescenzio Minervini, Anna Aventin, Nicole Dastugue, Christa Fonatsch, Anne Hagemeijer, Martine Jotterand, Dominique Muhlematter, Marina Lafage-Pochitaloff, Florence Nguyen-Khac, Claudia Schoch, Marilyn L Slovak, Arabella Smith, Francesc Sole, Nadine Van Roy, Bertil Johansson, and Mariano Rocchi. MYC-containing double minutes in hematologic malignancies: evidence in favor of the episome model and

exclusion of MYC as the target gene. *Hum Mol Genet*, 15(6):933–942, March 2006.

[84] Clelia Tiziana Storlazzi, Angelo Lonoce, Maria C Guastadisegni, Domenico Trombetta, Pietro D'Addabbo, Giulia Daniele, Alberto L'Abbate, Gemma Macchia, Cecilia Surace, Klaas Kok, Reinhard Ullmann, Stefania Purgato, Orazio Palumbo, Massimo Carella, Peter F Ambros, and Mariano Rocchi. Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. *Genome Research*, 20(9):1198–1206, September 2010.

[85] 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. *Proc Natl Acad Sci USA*, 102(43):15545–15550, October 2005.

[86] Kenji Takahashi, Takashi Kohno, Rieko Ajima, Hiroki Sasaki, John D Minna, Toshiyoshi Fujiwara, Noriaki Tanaka, and Jun Yokota. Homozygous deletion and reduced expression of the DOCK8 gene in human lung cancer. *International Journal of Oncology*, 28(2):321.

[87] Scott A Tomlins, Daniel R Rhodes, Sven Perner, Saravana M Dhanasekaran, Rohit Mehra, Xiao-Wei Sun, Sooryanarayana Varambally, Xuhong Cao, Joelle Tchinda, Rainer Kuefer, Charles Lee, James E Montie, Rajal B Shah, Kenneth J Pienta, Mark A Rubin, and Arul M Chinnaiyan. Recurrent fusion of TMPRSS2 and ETS

transcription factor genes in prostate cancer. *Science*, 310(5748):644–648, October 2005.

[88] Kristan E van der Vos, Leonora Balaj, Johan Skog, and Xandra O Breakefield. Brain tumor microvesicles: insights into intercellular communication in the nervous system. *Cellular and molecular neurobiology*, 31(6):949–959, August 2011.

[89] Marileila Varella-Garcia. Stratification of non-small cell lung cancer patients for therapy with epidermal growth factor receptor inhibitors: the EGFR fluorescence in situ hybridization assay. *Diagnostic Pathology*, 1(1):19, 2006.

[90] Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Geeta Joshi-Tope, Gopal Gopinath, David Croft, Bernard de Bono, Marc Gillespie, Bijay Jassal, Suzanna Lewis, Lisa Matthews, Guanming Wu, Ewan Birney, and Lincoln Stein. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol*, 8(3):R39, January 2007.

[91] P Vigneri and J Y Wang. Induction of apoptosis in chronic myelogenous leukemia cells through nuclear entrapment of BCR-ABL tyrosine kinase. *Nat Med*, 7(2):228–234, February 2001.

[92] Matthew J Walter, Dong Shen, Li Ding, Jin Shao, Daniel C Koboldt, Ken Chen, David E Larson, Michael D McLellan, David Dooling, Rachel Abbott, Robert Fulton, Vincent Magrini, Heather Schmidt, Joelle Kalicki-Veizer, Michelle O'Laughlin, Xian Fan, Marcus Grillot, Sarah Witowski, Sharon Heath, John L Frater, William

Eades, Michael Tomasson, Peter Westervelt, John F DiPersio, Daniel C Link, Elaine R Mardis, Timothy J Ley, Richard K Wilson, and Timothy A Graubert. Clonal Architecture of Secondary Acute Myeloid Leukemia. *N Engl J Med*, 366(12):1090–1098, March 2012.

[93] D Wang, X Tian, and Y Jiang. NDRG1/Cap43 overexpression in tumor tissues and serum from lung cancer patients. *Journal of Cancer Research and Clinical Oncology*, June 2012.

[94] J L Wittliff. Steroid-hormone receptors in breast cancer. *Cancer*, 53(3 Suppl):630–643, February 1984.

[95] Seiji Yano, Yasuhiko Nishioka, Hisatsugu Goto, and Saburo Sone. Molecular mechanisms of angiogenesis in non-small cell lung cancer, and therapeutics targeting related molecules. *Cancer Sci*, 94(6):479–485, June 2003.

[96] Lars Zender, Wen Xue, Johannes Zuber, Camile P Semighini, Alexander Krasnitz, Beicong Ma, Peggy Zender, Stefan Kubicka, John M Luk, Peter Schirmacher, W Richard McCombie, Michael Wigler, James Hicks, Gregory J Hannon, Scott Powers, and Scott W Lowe. An oncogenomics-based in vivo RNAi screen identifies tumor suppressors in liver cancer. *Cell*, 135(5):852–864, November 2008.

[97] Zhujiang Zhao, Qingxiang Wu, Jian Cheng, Xuemei Qiu, Jianqiong Zhang, and Hong Fan. Depletion of DNMT3A Suppressed Cell Proliferation and Restored

PTEN in Hepatocellular Carcinoma Cell. *Journal of Biomedicine and Biotechnology*, 2010, May 2010.

[98] Jingchun Zhu, J Zachary Sanborn, Stephen Benz, Christopher Szeto, Fan Hsu, Robert M Kuhn, Donna Karolchik, John Archie, Marc E Lenburg, Laura J Esserman, W James Kent, David Haussler, and Ting Wang. The UCSC Cancer Genomics Browser. *Nat Methods*, 6(4):239–240, April 2009.