RICE UNIVERSITY

Calculating Variant Allele Fraction of Structural Variation in Next Generation Sequencing by Maximum Likelihood

by

Xian Fan

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

Master of Science

APPROVED. THESIS COMMITTEE:

[Signatures]

Dr. Luay K. Nakhleh
Associate Professor, Chair,
Computer Science

Dr. Ken Chen
Assistant Professor,
Dept. of Bioinformatics & Comp. Biology,
UT MD Anderson Cancer Center

Dr. Lydia E. Kavraki
Professor,
Computer Science

Dr. Chris Jermaine
Associate Professor,
Computer Science

HOUSTON, TEXAS

DECEMBER 2014
Abstract

Calculating Variant Allele Fraction of Structural Variation in Next Generation Sequencing by Maximum Likelihood

by

Xian Fan

Cancer cells are intrinsically heterogeneous. Multiple clones with their unique variants co-exist in tumor tissues. The variants include point mutations and structural variations. Point mutations, or single nucleotide variants are those variants on one base; structural variations are variations involving sequence with length not smaller than 50 bases. Approaches to estimate the number of clones and their respective percentages from point mutations have been recently proposed. However, structural variations, although involving more reads than point mutations, have not been quantitatively studied in characterizing cancer heterogeneity. I describe in this thesis a maximum likelihood approach to estimate variant allele fraction of a putative structural variation, as a step towards the characterization of tumor heterogeneity. A software tool, BreakDown, implemented in Perl realizing this statistical model is publicly available. I studied the performance of BreakDown through both simulated and real data, and found BreakDown outperformed other methods such as THetA in estimating variant allele fractions.
Acknowledgements

First of all, I thank my committee members, Dr. Luay Nakhleh, Dr. Ken Chen, Dr. Lydia Kavraki and Dr. Chris Jermaine for their help and support on this project. I thank Dr. Kavraki and Dr. Jermaine for their time attending the thesis defense and their valuable comments on the direction of the project. Dr. Ken Chen, as my advisor in MD Anderson, patiently guided my research in the past two years. Dr. Luay Nakhleh oversaw the direction of my research and offered his generous help whenever it was needed. I thank you for what you have done for me. Without your help I cannot finish this project. Thank you.

Second, I thank my colleagues for their help on this project. They are Wanding Zhou, Zechen Chong, Tenghui Chen, Hamim Safar and Yun Yu.

Third, I thank my parents for their daily prayers and emotional support. I thank them for their encouragement when I was upset, and their valuable advices when I almost lost the direction.

Fourth, I thank my dear sisters for their selfless care and love. They are Cindy Tan, Ying Hu, Yu Shen, Veivei Zhu, Becky Shaw, and sisters in Baltimore. I am grateful for having you in my life.

Fifth, I thank my brother Yong Xia who brought me support and help whenever they were needed. In particular, I thank him for curing my Carpal Tunnel Syndrome so that I can type now.

Sixth, I thank my dear former colleagues/advisors for their encouragement and generous support. They are Charles Lu, Li Ding, and my mentor Arthur Delcher. I
am so grateful knowing you, and I will continue remembering you and your help.

Last but not the least, I thank my Lord Jesus Christ for leading me through the past two years. I did not give up the original plan because Jesus guides me, comforts me, encourages me, and loves me. Thank you Lord.
Contents

Abstract ............................................................... ii

1 Introduction ......................................................... 1

2 Structural Variations ............................................... 5
   2.1 Overview ....................................................... 5
   2.2 Signals in NGS ................................................ 8
   2.3 Literature of SV Detection ................................. 15
      2.3.1 ReadDepth ............................................... 16
      2.3.2 BreakDancer ............................................. 17
      2.3.3 CREST .................................................... 20

3 Tumor Heterogeneity ............................................... 21
   3.1 ABSOLUTE ................................................... 23
   3.2 THetA ........................................................ 25

4 Methods ................................................................... 27
   4.1 A maximum likelihood VAF estimator ..................... 27
   4.2 The Maximum Likelihood Formulation ..................... 29
4.3 Modelling normal read pairs ........................................... 31
4.4 Modeling discordant read pairs ................................. 32
4.5 Modelling soft-clipped read ................................. 32
4.6 VAF Estimation .................................................... 33
4.7 Confidence scoring ............................................ 34
4.8 Genome-wide parameter initialization ............................... 35

5 Results ........................................ 36
5.1 Characterizing the estimation accuracy ........................... 36
5.2 Comparison with published tools ................................. 38
5.3 Identifying somatic SVs from tumor-normal matched WGS data . . . 40
5.4 Identify sub-clonal SVs in breast cancer samples ............... 45

6 BreakDown ........................................ 50
6.1 Input ..................................................... 50
6.2 Output ..................................................... 52
6.3 Steps ..................................................... 54
6.4 Examples .................................................. 56
6.5 Dependence on Other Perl Modules .............................. 57

7 Conclusions and Future Research .................................. 58

8 Conclusions ........................................ 61
List of Figures

2.1 Illustration of four SV types. For each type, a reference and a target genome is shown. On the reference genome, the two breakpoints bp1 and bp2 are indicated with vertical blue bars. The blacks lines show the correspondence between the breakpoints on the target genome and those on the reference. Novel adjacencies occur at these breakpoints. (a) Deletion (DEL). The red segment on the reference is deleted on the target. (b) Inversion (INV). The red segment is inverted and replaced at the same location. (c) Tandem duplication (DUP). The red segment is duplicated and placed next to the original copy. (d) Insertion (INS). The yellow segment is inserted in between the blue and green ones. When the yellow segment can be found on the reference, the insertion is templated; otherwise non-templated.
2.2 Illustration of three possible translocation orientations. For each case, shown are the adjacencies at the left for reference and right for target. On the reference side, solid lines are connected by the blue curves, representing novel adjacency. (a) 5’ to 3’ connection; (b) 5’ to 5’ connection; (c) 3’ to 3’ connection. Note that 3’ to 5’ connection is the same as (a) except the order of the two chromosomes are reversed, and thus eliminated.

2.3 Illustration of the process of massively parallel paired-end sequencing. Each blue horizontal line denotes an insert, sequenced from the target genome in black. Each insert is composed of the left and right read, with the orientation pointing towards each other. Elements composing a read are A, C, G and T.

2.4 Number of sequenced reads corresponding to GC content for two Illumina libraries of sample COLO-829. It can be seen a peak stands at around 40, with the two end dropping, causing the GC bias problem. The number of read pairs is estimated by randomly selecting 10 regions of size 10Mbp in the whole genome, each region further divided into 10K non-overlapping small regions. The number of read pairs per bp and corresponding GC content are calculated from each 10K region. Plotted are the median numbers of read pairs per bp (Y axis) at different GC content (X axis) in two DNA libraries: COLO829IL (left) and COLO829_v2_74 (right). The straight lines on the left and the right sides of the curve are linearly extrapolated from the real values in the middle.
2.5 Shown are the illustrations of five SV types from DRP signal. Each DRP is illustrated as a left read in red and right read in green, connected with the blue dotted line. The black horizontal lines are the reference. The breakpoints, shown as bp1 and bp2 at the vertical blue lines on the references are where the novel adjacencies occur. (a) Deletion (DEL) is characterized by an increased insert size (distance between the left and right reads). Inversion (INV) has two groups of DRP, one with both left and right reads pointing to the right, or 3' end; the other to the left or 5' end. Duplication (DUP) has the paired reads' positions reversed, i.e., the right read is on the 5' end of the left read. Note that this applies only to tandem duplications. (b) Translocation (TRA) has the two reads sequenced from two chromosomes, or two distinct locations on one chromosome. Shown is the case of a 5' connection on chr1 with 3' of chr2. (c) Illustrated is an insertion (INS). The sequence on chr2 highlighted in orange is inserted into the position with the vertical blue bar on chr1. The discordant read pair has one read sequenced from the flanking region of the inserted position, and the other from the inserted sequence. DRP has two patterns: one has a read on the left side of the inserted position, and the other on the right.
2.6 Shown are the illustrations of five SV types from SR signal. Each SR is illustrated as a left read in red and right read in green, connected with the blue dotted line. When a read is sequenced from two genomic loci, the adjacency in the read is shown as the red or green dotted lines, corresponding to the left and right read respectively. The black horizontal lines are the reference. The breakpoints, shown as bp1 and bp2 at the vertical blue lines on the references are where the novel adjacencies occur. All types have two groups of SRs: either the left or the right read is split. (a) Deletion (DEL) is characterized by the split of either the left or the right read. Inversion (INV) has either the left or the right read split and the resulting two segments are facing towards each other. Duplication (DUP) has either the left or the right read split and the resulting segments are in a reversed position. Note that this applies only to tandem duplications. (b) Translocation (TRA) has either the left or the right read split and the resulting segments come from two chromosomes, or two distinct locations on one chromosome. Shown is the case of a 5’ connection on chr1 with 3’ of chr2. (c) Illustrated is an insertion (INS). The sequence on chr2 highlighted in orange is inserted into the position with the vertical blue bar on chr1. Both the left (5’) and the right (3’) end of the inserted place has split reads, with the other segments mapped to the inserted sequence.
4.1 Schematic overview of BreakDown. a) In the input are the bam files and a set of SV calls. Reads encompassing each SV are extracted for analysis. Genome-wide parameters such as average read count per bp per GC content are initialized. b) The encompassed reads are divided into three groups: normal, discordant and soft-clipped. Reads in each group are counted. c) Normal reads for large CNVs are counted in a series of consecutive non-overlapping bins. d) Read counts are normalized by GC contents. e) Expected read counts are defined as functions of SV, VAF and sequencing data. f) Genotype and VAF that maximize the likelihood function are derived from the expected and observed read counts. g) Variant scores are estimated that quantify the confidence of the results. Detailed explanations of the mathematics are available in Section 4.2.

4.2 Comparison between the observed and the expected read counts. Data in box plots are the number of discordant read pairs and soft-clipped reads from the known deletions from the 1000 Genomes Project [MWS+11] sample NA12878 [CPR+10] with 67x sequence coverage and 100 bp long reads. The boxes represent the 75% and 25% percentiles on the high and low ends. The red horizontal lines indicate the medium, i.e. 50% percentile. The highest and lowest horizontal lines outside of the box indicate the maximum and minimum counts. Expected read counts for the heterozygous variants (left) and homozygous variants (right), corresponding to VAFs of 0.5 and 1, respectively are plotted as black horizontal lines above the boxes.
5.1 Accuracy of estimated VAF in simulation. Plotted are the chances (Y axis) of the estimated VAF falling within 10% of the true VAF (X axis). Each data point is estimated from 1000 random samples. Each subplot in the figure contains 4 curves representing the accuracy of VAFs estimated from SNVs (red plus), 1 Kb deletions (green cross), 1 Mb deletions (blue triangle), and inversions/reciprocal-translocations (purple circle). Various types of sequencing data are simulated and results compared: a) short-insert (500 bp) short-read (100 bp) at 5x sequence coverage, b) short insert short read at 30x coverage, c) long insert (3000 bp) short read at 30x coverage and d) short insert short read at 500x coverage.

5.2 A mock phylogeny tree of a polyclonal tumor mass. The root is C₀, representing the founding germline clone. Each node, C₀ to C₅, corresponds to a clone that existed in the phylogeny. C₁ and C₂ evolved from C₁₂, C₃ and C₄ evolved from C₃₄ and C₅ evolved directly from C₀. The ratios in the parentheses (ranging from 0.05 to 1.0), represents the relative abundance of a clone at the time of branching. The SVs in different clones are labeled in different shapes based on their lineages. SV types are differentiated by colors: blue for deletion, orange for tandem duplication. To simulate progressive tumor clone development, both C₁ and C₂ inherited the two SVs (diamond) in C₁₂, whereas C₂ had two novel SVs (square). The similar labeling was applied to C₃₄, C₃, C₄, and C₅.
5.3 Comparison of VAF estimation errors between BreakDown and THetA.

Plotted are the mean and the maximum errors (Y axis) estimated from 10 lineage specific SVs by BreakDown and THetA under six normal contamination rates from 0 to 0.5 (X axis). Details are described in the text and Table 5.4.

5.4 Plots of estimated VAF of validated deletions. Diamond, triangle and circles represent validated germline, putative germline and somatic events. For tumor ploidy analysis, those events on the chromosomes which are mostly diploid are in green, and mostly polyploid in red. Those that are partly diploid are in black.

5.5 A loss-of-heterozygosity deletion in NRXN3 in COLO-829. Plotted is an image from the integrative genome viewer (IGV) that represents read alignment over this deletion (chr14:79175898?79184805) in the tumor sample (top panel) and in the normal sample (bottom panel). The discordant read pairs (dark-red bars connected by long grey lines) and drop in coverage (white space in the center) correspond to the start and the end of the deletion (red horizontal bar). VAFs are estimated by BreakDown to be 1.0 and 0.49, respectively in the tumor and the normal.
5.6 Sub-clonal SVs in PD4120. Plotted are the integrative genome viewer (IGV) screenshots that display reads in the tumor sample (top panel) and in the normal sample (bottom panel) of a) A 1751 bp deletion between Chr10:117351105 and Chr10:117352856 with an estimated tumor VAF 0.12, and b) an intra-chromosomal translocation between Chr22:19372768 and Chr22:21122617 with an estimated tumor VAF 0.055. Reads are displayed from the top to the bottom in the following order: split reads (partially clipped bars), discordant read pairs (brown bars and lines), and normal read pairs.
List of Tables

5.1 Comparison of the VAF estimation errors between BreakDown and THetA based on simulation. For columns 3 to 8, listed are the estimation errors (estimated - true value) for BreakDown (left of slash) and THetA (right of slash) for normal contamination rate = 0, 0.1, ..., 0.5. The second column is the true VAF without normal contamination rate.  

5.2 A list of 32 previously reported deletions with BreakDown estimated VAFs in COLO-829. The Germline_Tag column indicates the structural variation is estimated by BreakDown to be somatic if 0, putative germline if 1 and validated germline if 2. Validation is reported by CREST. The Polyploid_Tag column indicates the ploidy of the genomic loci of the structural variation: partly diploid for 0, polyploid for 1 and mostly diploid for 2. 

5.3 A list of 41 LOH deletions detected by BreakDown in COLO-829. The VAF_Diff column lists the absolute difference of estimated tumor and normal VAF by BreakDown; the Hit_Pleasance_LOH column lists the LOH estimated by [PCS+10] overlapping with this entry; the Hit_Gene lists the hit gene if there is any.
5.4 Novel sub-clonal somatic structural variants detected by BreakDown. Listed are the novel somatic structural variants in the breast cancer sample PD4120, PD4115 and PD4088 along with BreakDown estimated VAFs (column 9). Also shown (column 8) are the genes overlapping with either deletion loci or translocation breakpoints.
Chapter 1

Introduction

Cancer is one of the main causes of death in the world. It is estimated that 590,000 people will die of cancer in 2015. The overall cancer survival rate was less than 70% between 2004 and 2010 (http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2015/index). Cancer cells come into existence when there are genetic changes that enable them to avoid programmed cell death or evade the immune system, in which case they will continue to divide, avoiding death, and invade normal cells. Thus the detection of genetic variations in tumor cells is significant in cancer diagnosis and treatment. The advent of the sequencing technology such as next generation sequencing [Mar12] makes this detection possible.

Cancer cells evolve. The evolution causes heterogeneity. A bulk tissue taken from the solid tumor, like a snapshot taken at a particular moment in the long cancer evolutionary history, contains multiple sub-clones [NKT+11]. Each sub-clone is characterized by its unique signature: mutations. While two sub-clones contain the same mutation, they might evolve from the same parent clone. We use variant allele frac-
tion (VAF), the fraction of the alleles carrying the variant as a measurement of the extent of the mutation.

Variations or mutations are the genomic changes on the target genome with respect to the reference, a nucleotide acid sequence database which was assembled from long segments as a representative of human genomes. Next Generation Sequencing (NGS) technology sequences genomic segments (called an insert) after shearing amplified DNA. If the sequencing technology is paired-end, the left and the right ends (each called a read) of the sheared genomic segment are sequenced. A standard Illumina short-read library has 200 base pair (bp) to 400bp long insert, and 100bp long read. Amplification makes multiple copies of sequences covering the same nucleotide. The number of such copies is called coverage, represented as an integer followed by “X”. For instance, if on average there are 5 copies of a genome sequenced, the coverage is 5X. A standard coverage varies from low at 1-10X to high at >100X. Since the human genome is about 3 billion bps long, a 30X paired-end sequencing library contains 0.3 billion inserts and 0.6 billion reads with 300bp insert size. Computational methods have been developed to detect variations in a human genome by comparing these inserts/reads with the reference [CLC+13, YSL+09, CWM+09].

Inserts lose genomic positions after sheering. Alignment algorithms were developed [LYL+09, LD10, LTPS09] to align reads back to the reference. The output contains the chromosome and start position of the read on the reference, along with a CIGAR string indicating the alignment status of each nucleotide of the read on the reference [LHW+09]. The status includes “matching”, “insertion”, “deletion” and “clipped”. This enables a detailed look at the variation.
Variations are divided into three categories in terms of scale: single nucleotide polymorphism or variation (SNP or SNV), small insertions and deletions (INDEL) and large structural variations (SV). SNP is a single nucleotide change in the pool of A, C, G, T. In cancer genomic research area, it is usually called SNV. Methods for the detection of SNP/SNV have been well developed [CLC+13] and the accuracy and consistency among methods are relatively high compared with that of SVs [XDS+14]. INDEL refers to insertions and deletions with size ranging from a few bps to a few hundred bps. SVs are those variations with size varying from 50 to millions of bps. SVs can be categorized into different types, including deletion, insertion, duplication, inversion and translocation. Combination of two or more types forms complex SVs. While the SNP and INDEL calling has been looked into in a deep and thorough way [Con10], SV detection lags far behind. First, a high discordance has been observed among different methods [MWS+11]. Accuracy varies from 15 percent to 90 percent. Each method focuses on one or more signals, but few methods utilizes all SV signals. Second, the existence of complex SVs makes it more challenging [QH12, LEN+11]. The discovery of new SV types, such as chromothripsis [SGF+11] and chromoplexy [BPL+14] implies that we are far away from having a comprehensive knowledge of SV types. Third, due to the abundance of repetitive regions in the human genome and the tendency of SVs to occur in such regions [SKS+11], the detection of SVs becomes even more challenging. Short inserts further limit the scale of detection. Not spanning the whole repetitive regions makes SV detection methods prone to generate a myriad of false positives. All of these make SV detection a challenging task. Yet SV is a significant type of variation since it involves the highest amount of nucleotides, and it
has been implicated in various forms of cancer [Mar12, LEN+11, MLF+13]. Therefore, a deeper and more thorough understanding of SV is important.

While methods have been proposed for SV detection, few have been developed to statistically model all the signals of an SV. Moreover, since cancer genomes are heterogeneous, none of the existing methods is capable of characterizing such heterogeneity through a single SV [FZC+14]. Some of them [CCH+12, OMR13] can merely utilize a group of SVs, yet not all SV types are addressable. Other methods use SNVs for heterogeneity characterization [NZVLW+12], yet they are mostly limited to high coverage data. In this thesis, I introduce a statistical model utilizing all the signals of an SV to estimate its VAF [FZC+14]. This serves as an intermediate step towards the full characterization of tumor heterogeneity. Our proposed method, BreakDown [?], is not limited to coverage and SV types, and is applicable to a single SV. The tool can be freely downloaded from https://bitbucket.org/xianfan/breakdown.
Chapter 2

Structural Variations

2.1 Overview

Structural variations involve a large number of nucleotides. They have been reported to cause genetic diseases such as cancer [Mar12]. Structural variations include deletion, insertion, inversion, translocation and duplication, as is shown in Fig. 2.1 and Fig. 2.2. Deletion is a loss of a genomic locus, with the two ends concatenated. Insertions can be divided into templated and non-templated ones. For the templated insertions, the inserted sequence is from the genome. This is also called duplication when the original copy remains, otherwise it is a translocation. If the inserted sequence is next to the original copy, it is called tandem duplication. Non-templated insertions have novel inserted sequence. Inversion replaces the reverse complement of the original copy of the sequence in the same position. Translocation has the adjacency between two genomic loci that were not originally adjacent to each other. Translocations can be categorized into intra-chromosomal and inter-
chromosomal translocations. The former results in the connection of two distinct loci that are far away from each other in the same chromosome. The latter results in the connection of two loci on distinct chromosomes. The orientations of their connection can be 5’ to 3’, 5’ to 5’, 3’ to 3’ and 3’ to 5’, as shown in Fig. 2.2. If a novel adjacency forms between two genomic loci, we call the nucleotide at which the adjacency occurs as the breakpoint. With respect to the reference genome, breakpoint refers to the nucleotide position at which the normal adjacency is missing. For brevity, we will refer to the breakpoints as relative to the reference genome in the rest of the text unless explicitly stated otherwise. An insertion has only one breakpoint, which indicates the inserted locus. All other SV types mentioned above have two breakpoints.
Figure 2.1: Illustration of four SV types. For each type, a reference and a target genome is shown. On the reference genome, the two breakpoints bp1 and bp2 are indicated with vertical blue bars. The blacks lines show the correspondence between the breakpoints on the target genome and those on the reference. Novel adjacencies occur at these breakpoints. (a) Deletion (DEL). The red segment on the reference is deleted on the target. (b) Inversion (INV). The red segment is inverted and replaced at the same location. (c) Tandem duplication (DUP). The red segment is duplicated and placed next to the original copy. (d) Insertion (INS). The yellow segment is inserted in between the blue and green ones. When the yellow segment can be found on the reference, the insertion is templated; otherwise non-templated.
Figure 2.2: Illustration of three possible translocation orientations. For each case, shown are the adjacencies at the left for reference and right for target. On the reference side, solid lines are connected by the blue curves, representing novel adjacency. (a) 5’ to 3’ connection; (b) 5’ to 5’ connection; (c) 3’ to 3’ connection. Note that 3’ to 5’ connection is the same as (a) except the order of the two chromosomes are reversed, and thus eliminated.

2.2 Signals in NGS

Next Generation Sequencing technology takes pieces of subsequence from the target genome by random shotgun, amplify them via polymerase chain reaction (PCR),
and sequences them in a massively parallel way [Mar12]. Each newly identified string of bases composed of A, T, C and G is called a read. If the sequencing technology supports paired-end (PE) protocols, the sequenced string is in a form of a pair of two reads. Both of them have an orientation, and they point towards each other, that is, the left read is sequenced from left to right (or 5’ to 3’ end) and the right one from right to left (or 3’ to 5’ end), as illustrated in Fig. 2.3. The insert size is defined as the length of the whole pair, from the 5’ end of the left read to the 3’ end of the right one. The insert size of a pair corresponds to a certain distribution based upon the sequencing technology.

![Diagram](image)

**Figure 2.3:** Illustration of the process of massively parallel paired-end sequencing. Each blue horizontal line denotes an insert, sequenced from the target genome in black. Each insert is composed of the left and right read, with the orientation pointing towards each other. Elements composing a read are A, C, G and T.

When a structural variation is present, there are three types of signals in paired-end reads: read depth, discordant read pair and split reads.

Read depth characterizes the variation of the number of reads aligned to a genomic locus. It can be used to detect deletions and duplications. Deletions have a lower read
depth in the deleted area as compared with the flanking region. If a deletion occurs at
30% of the alleles, one would expect a 30% drop in read depth in the deleted region.
This is because these 30% of the alleles do not generate any reads on the deleted
sequence. Assuming a uniform sampling, 30% fewer reads will be observed inside the
deleted region as compared to the flanking region.

In fact, the read sampling after PCR-amplification is not uniform. Low sampling
rate has been observed at the two extremes of GC-content, a measurement of the
fraction of the sum of G and C versus all four nucleotides A, C, G and T. One can
use a fixed size window to calculate GC content. As shown in Fig. 2.4, we observed
a peak in 40% GC and a drop at the low and high GC for the two libraries of a
sequencing data from a healthy human individual. Methods have been proposed to
correct read depth according to GC-content [YXM+09] so that the signals of read
depth reflect the true loss of sequence in the variant alleles.
Figure 2.4: Number of sequenced reads corresponding to GC content for two Illumina libraries of sample COLO-829. It can be seen a peak stands at around 40, with the two end dropping, causing the GC bias problem. The number of read pairs is estimated by randomly selecting 10 regions of size 10Mbp in the whole genome, each region further divided into 10K non-overlapping small regions. The number of read pairs per bp and corresponding GC content are calculated from each 10K region. Plotted are the median numbers of read pairs per bp (Y axis) at different GC content (X axis) in two DNA libraries: COLO829IL (left) and COLO829_v2_74 (right). The straight lines on the left and the right sides of the curve are linearly extrapolated from the real values in the middle.

When a pair of reads is sequenced from the target genome in such a way that the segment in between the two reads intersects the breakpoint on the target genome, discordant read pair (DRP) occurs. While read depth is a signal for deletion and duplication, DRP signifies all SV types. Actually, by studying the orientation, position and insert size of DRP, one can infer the SV type. Fig. 2.3.2 shows different signatures of each SV type for DRPs. A deletion has larger-than-normal insert size, and an insertion the opposite. It is noted that due to the variation of insert size in the absence of SVs, short deletions may not have an obvious signature of DRP. Large insertions (larger than average insert size) usually do not have any DRP signal. Small
insertions may display very weak DRP signal, and thus it is not recommended to use abnormal insert size of DRP to detect insertions. On the other hand, if the insertion is a tandem duplication, the abnormal positions of the left and the right read can be used to infer the SV. An inversion has both abnormal insert size and abnormal orientation to one of the reads. The latter is more obvious and straightforward to infer an inversion. The insert size minus the read length can be used to infer the inversion size. Finally, the translocation has a pair of reads coming from distinct chromosomes if it is an inter-chromosomal one. Intra-chromosomal translocations (ITX) mostly can be explained by duplication, deletion, insertion, inversion or inverted duplication.
Figure 2.5: Shown are the illustrations of five SV types from DRP signal. Each DRP is illustrated as a left read in red and right read in green, connected with the blue dotted line. The black horizontal lines are the reference. The breakpoints, shown as bp1 and bp2 at the vertical blue lines on the references are where the novel adjacencies occur. (a) Deletion (DEL) is characterized by an increased insert size (distance between the left and right reads). Inversion (INV) has two groups of DRP, one with both left and right reads pointing to the right, or 3’ end; the other to the left or 5’ end. Duplication (DUP) has the paired reads’ positions reversed, i.e., the right read is on the 5’ end of the left read. Note that this applies only to tandem duplications. (b) Translocation (TRA) has the two reads sequenced from two chromosomes, or two distinct locations on one chromosome. Shown is the case of a 5’ connection on chr1 with 3’ of chr2. (c) Illustrated is an insertion (INS). The sequence on chr2 highlighted in orange is inserted into the position with the vertical blue bar on chr1. The discordant read pair has one read sequenced from the flanking region of the inserted position, and the other from the inserted sequence. DRP has two patterns: one has a read on the left side of the inserted position, and the other on the right.

When a read sequenced from the target genome intersects with the breakpoint on the target genome, split read (SR) comes into existence. Similar to DRP, SR can be used to identify all types of SVs as shown in Fig.2.6. Instead of using a pair of reads, SR mostly relies on the two alignments of one read, i.e., the read is split into two parts during alignment. In NGS where read length is limited (typically 36bp
- 250bp), it is possible that only one alignment of the two exists; the other end of
the read is too short to have a unique anchor on the reference genome. In this case,
only one breakpoint on the reference genome can be inferred, leaving the other one
(excluding INS) and the SV type unknown. In insertions, split reads with only one
alignment is quite common. This is because no matter whether the inserted sequence
is templated which results in multiple alignments of the other end, or non-templated
that leads to no alignment on the reference, there is only one breakpoint reported.
A potential INS is indicated if the following conditions are met: 1. A combination
of two split reads with their inferred breakpoints located close to each other (<20bp
typically); 2. These two split reads face towards each other. Of all SV types, when
both alignments are present, and the aligned sequences compose the whole read, it is
quite straightforward to infer SV type.
Figure 2.6: Shown are the illustrations of five SV types from SR signal. Each SR is illustrated as a left read in red and right read in green, connected with the blue dotted line. When a read is sequenced from two genomic loci, the adjacency in the read is shown as the red or green dotted lines, corresponding to the left and right read respectively. The black horizontal lines are the reference. The breakpoints, shown as bp1 and bp2 at the vertical blue lines on the references are where the novel adjacencies occur. All types have two groups of SRs: either the left or the right read is split. (a) Deletion (DEL) is characterized by the split of either the left or the right read. Inversion (INV) has either the left or the right read split and the resulting two segments are facing towards each other. Duplication (DUP) has either the left or the right read split and the resulting segments are in a reversed position. Note that this applies only to tandem duplications. (b) Translocation (TRA) has either the left or the right read split and the resulting segments come from two chromosomes, or two distinct locations on one chromosome. Shown is the case of a 5’ connection on chr1 with 3’ of chr2. (c) Illustrated is an insertion (INS). The sequence on chr2 highlighted in orange is inserted into the position with the vertical blue bar on chr1. Both the left (5’) and the right (3’) end of the inserted place has split reads, with the other segments mapped to the inserted sequence.

2.3 Literature of SV Detection

The methods for SV detection can be categorized by the signals they use. While some methods proposed earlier [CWM+09, WME+11, KAM+09, MHCM11] use only
one signal, recently proposed ones tend to combine two or more signals [HKNM11, RZS+12]. The combination of the signals yields higher sensitivity and specificity. The strategies of combination, however, differ from one method to another. In this section, three methods using a single signal, one from each described above are introduced.

2.3.1 ReadDepth

The signal of read depth can be only used to detect deletions and duplications. In the context of copy number variation (CNV, the variation of the number of allele copies in a genome), they are usually referring to large deletions and duplications (>1Mbp). A popular method for CNV detection is circular binary segmentation (CBS) proposed by Olshen [OVLW04]. Although it was initially proposed for array-based data, CBS can be easily applied to NGS data, and a successful application has been proposed in an R package called ReadDepth [MHCM11]. Dividing the whole chromosome into small contiguous non-overlapping windows, CBS counts the number of reads aligned to each window. Then it divides the whole chromosome into two halves with equal size. A statistical test to measure the homogeneity of the read depth on the small windows is done to each half. If the test reports a high variance of read depth among windows, it further divides that half into two halves. This process continues until no more division can be made. Finally each segment can be taken as containing a homogeneous copy number. Taking the most common one as diploid, the rest can be classified into copy number gain (duplication) and loss (deletions).

Two key aspects lead to a high sensitivity and specificity of CNV detection. The
first is the handling of GC-content bias [YXM+09]. According to [BBS+08] and as mentioned above, the number of reads sequenced corresponds to the G+C percentage in the window, instead of being uniform across the genome. It requires additional pre-processing by G+C percentage before CNV detection, either by read depth correction from read depth distribution corresponding to G+C percentage, or adjusting the expectation. In our proposed model, we used the latter strategy.

The second key aspect is the window size. First of all, the window size selection is highly correlated with the overall coverage of the genome. The higher the coverage, the smaller the window size. A large window size limits the resolution of breakpoint identification, whereas a small one may lead to over-segmentation and thus high false positive rate [YXM+09]. [YXM+09] used a 100bp window at 30X coverage. CNV-seq ([XT09]), with the assumption of Poisson distribution of the number of reads falling into a fixed window size, selected window size such that the number of reads falling into each window is > 10. In our proposed method, we use a window size that is adapted to the coverage and the minimum number of reads required in each window.

### 2.3.2 BreakDancer

BreakDancer utilizes discordant read pairs to detect SVs. Given a binary alignment/map (BAM) file [LHW+09], BreakDancer reports all putative SV calls including deletion, insertion, inversion, intra-chromosomal translocations and inter-chromosomal translocations. Originally proposed as two tools (BreakDancerMax and BreakDancerMini) targeting different scales of SVs, only BreakDancerMax is being
actively maintained and used. Also, originally implemented in Perl and slow to run, only BreakDancerMax was refactored into C++ [FALC14] incorporating Samtools API for speedup. Thus BreakDancer represents BreakDancerMax in most cases in recent years in the community. Similarly, I will describe BreakDancerMax only and use BreakDancer to represent the tool BreakDancerMax in the following text.

BreakDancer scans through all reads in the BAM file, collects discordant ones and cluster them by both genomic distance and the types of inherent SVs. A quality score was proposed to rank the clustering in the aspects of both width and depth. In particular, a statistical model assuming Poisson distribution of random occurrence of discordant read pairs was proposed, and the p-value characterizing that the cluster of discordant reads comes from this distribution can be calculated by the following equation:

\[
P(k_i) = P(n_i > k_i),
\]

(2.1)

in which \(k_i\) is the total number of discordant reads of a cluster with type \(i\), and \(n_i\) represents all possible number of discordant reads of a cluster with type \(i\). Assuming Poisson distribution, the mean of a cluster with size \(l\) (in bp) can be pre-calculated as

\[
\lambda = \frac{N \times l}{G},
\]

(2.2)

in which \(N\) is the total number of discordant reads and \(G\) (in bp) is the total length of the whole genome. An initialization pre-calculates \(N\) for each SV type, and
estimates the corresponding $\lambda$.

When there are multiple libraries in a single genome, BreakDancer combines them naturally by assuming the independence across libraries, and uses a Chi-square distribution to combine the probabilities. Finally, it converts the probability into a phred quality score by the following equation:

$$S = -10 \log(P)$$  \hspace{1cm} (2.3)

Note that the higher the score, the higher confidence of the SV being true. If the score is 30, the probability that the SV is a false positive is 0.001, taking 10 as the logarithm base. When running BreakDancer, users can set up their own threshold to select high confidence call set.

BreakDancer has been actively used by the community including the 1000 Genomes Project [MWS+11] and compared with newly proposed methods. The advantage of BreakDancer is that compared with other SV detection methods, it is easy to install and run, and the total runtime after refactoring is limited within a day if users parallelize jobs on the high performance clusters by chromosomes. Also, it is able to report all SV types. The drawbacks are the high false discovery rate and the ambiguity of breakpoints. The false positive calls are mainly due to mis-alignment from repetitive regions, the sequencing errors, and the large range of insert size. It cannot report the nucleotide resolution breakpoint due to the intrinsic property of using discordant read pairs. The range of putative breakpoint resolution varies by the insert size distribution, usually predicted to be from hundreds to a thousand base pair for Illumina
short reads.

2.3.3 CREST

CREST [WME11] (Clipping REveals STructure) utilizes soft-clipped reads to detect the same set of types of structural variation as BreakDancer. The motivation is to collect soft-clipped reads on one breakpoint end, and use a cohort of them to find the other end. In particular, with a BAM file as the input, CREST extracts soft-clipped reads, which is indicated in the CIGAR string in BAM. At a particular putative region with the abundance of soft-clipped reads, CREST assembles them into a long contig: a continuous sequence of DNA. The contig extends a read with the cohort of all extracted reads. CREST then aligns the contig to the reference, allocating the other end that the soft clipped parts span. The long contig, when compared with the short reads, makes it more likely to anchor the other end on the reference. To confirm the paired breakpoints, the soft-clipped reads from the other end allocated in the previous process are extracted and the similar approach is applied to reversely allocate the end that the algorithm starts from.

CREST is the first approach to utilize soft-clipped reads to discover SV breakpoints. It has a nucleotide resolution, and the specificity is relatively high, with the validation rate > 80% from PCR. The disadvantage is that it takes a long time to run and that it can miss calls when alignment fails to report soft-clipped reads if the reads are short or the soft-clipped part comes from non-templated sequence.
Tumor Heterogeneity

Tumor cells may differ in terms of the variants they carry, even in the same bulk tissue of the tumor. Such intra-tumor heterogeneity makes it challenging for accurate diagnosis and treatment. For example, several tumor clones may co-exist, each with its own amount of cells. Those with a larger percentage is easier to detect and thus more likely to be targeted in the treatment. However, if the small clones are not taken good care of, they may grow and become a dominant clone due to the drug resistance and selective advantage after the extinction of large clones, leading to even more severe cancer to the patient [DLL+12]. Therefore, it is significant to characterize the intra-tumor heterogeneity. Such characterization includes listing the number of clones, their respective percentage, and the mutations they carry. Due to that tumor cells might be contaminated by normal ones with unknown percentage or merely a rough estimate, the intra-tumor heterogeneity sometimes includes the step of estimating tumor purity or normal contamination rate. The advent of NGS makes such comprehensive characterization possible. However, except single cell sequencing,
the sequencing data from NGS are mostly an admixture of the cells from all tumor clones and contaminating normal ones. It requires computational methods to infer the most likely proportion and normal contamination rate given the combined snapshot. Numerous methods have hence been proposed to achieve this goal. The information they utilize are mostly copy number change and SNV.

Copy number is the number of allelic copies. A copy number change (gain or loss, or can be understood as duplication and deletion in a larger scope) at a particular clone results in the corresponding read depth change. For example, if a one-copy-gain happens to the clone with 40% cells in tumor bulk tissue, the resulting read depth is 20% larger than copy number neutral regions, taking into account human chromosomes are diploid except sex chromosomes. Reverse engineering makes it possible to answer the question of which percentage of clones has such copy number gain. To finally characterize intra-tumor heterogeneity, the percentages of cells containing copy number changes are clustered in a statistical modeling. Certain assumptions are made to make the models within a manageable scope. The assumptions include: 1. The maximum number of clones; 2. A genomic region may not have copy number gain in one clone and loss in the other.

The other information, SNV, has much more abundance of events than copy number change. Existing methods utilize SNVs for the characterization of intra-tumor heterogeneity by calculating its variant allele fraction (VAF), which is the fraction of variant alleles over all. In the case of a homogeneous bulk tissue with one diploid clone, if the SNV happens to one of the two alleles, VAF is 0.5. We call it to be a heterozygous event because it increases the heterozygosity between the two alleles. If the
SNV happens to both alleles, VAF is 1. Such event is called a homozygous event. In NGS, the VAF of a SNV is computed by counting the number of reads supporting the variant and dividing it by all reads covering the SNV locus. An accurate estimation of VAF is critical in characterizing intra-tumor heterogeneity because it determines the quality of clustering SNVs with similar VAFs, a step for the preparation of statistical modeling.

In the following section I will describe two state-of-the-art methods in characterizing intra-tumor heterogeneity, and finally state the motivation of my proposed model: BreakDown.

### 3.1 ABSOLUTE

ABSOLUTE [CCH+12] is an algorithm to infer the most likely tumor purity and ploidy given segmented and smoothed copy number from either microarray or parallel sequencing. The VAF of each SNV is an optional input. Methods for CNV detection (or deletion and duplication detection) have been described in the last chapter. Some post-processing such as HAPSEG (https://www.broadinstitute.org/cancer/cga/hapseg) is used to normalize average read depth of each segmented copy number region so that they can be taken as relative copy number. In the case of heterogeneous and impure tumor sample, the copy number is usually not integer, as a result of combining those from multiple sub-clones and normal contamination with their respective percentage. ABSOLUTE assumes $\tau$ as the average tumor ploidy, and $\alpha$ as the tumor contamination rate. Ploidy, the number of chromosomes in a genome,
can be understood as the copy number in a genome-wide range. The average ploidy of the contaminated tumor is \( D = \tau \alpha + 2(1 - \alpha) \). For a certain segmented locus \( x \) in the preprocessing step, suppose the underlying integer copy number in pure tumor’s major sub-clone is \( q(x) \), the following equation holds if there is only one clone in tumor:

\[
R(x) = (\alpha q(x) + 2(1 - \alpha))/D = \alpha q(x)/D + 2(1 - \alpha)/D, \tag{3.1}
\]

in which \( R(x) \) is the relative copy number ratio after copy number normalization as mentioned above for locus \( x \).

Similarly, if given somatic point mutations’ VAF \( F(x) \) for a specific locus \( x \), the following equation holds if there is only one clone in tumor:

\[
F(x) = (\alpha s_q(x))/D_s, \tag{3.2}
\]

in which \( s_q(x) \) is the number of copies supporting the SNV, and \( D_s = \alpha q(x) + 2(1 - \alpha) \). \( D_s \) represents the ploidy as a combination of both that of tumor and contaminating normal. Note that \( s_q(x) \) is smaller than \( q(x) \) and takes the integer values. With the combination of Eq.3.1 and Eq.3.2, ABSOLUTE uses accumulated \( R(x) \) and \( F(x) \) for each \( x \) to infer \( \tau \) and \( \alpha \) by adapting the data in a mixture Gaussian model. It outputs the mostly likely combinations of \( \tau \) and \( \alpha \), and when there are more than one solution, it uses recurrent cancer-karyotype models based on existing large data sets to infer the most common one.

ABSOLUTE has been widely used for SNParray data. Although the paper claimed
that it can deal with NGS data, it does not give a detailed explanation of the method such as the scoring system for NGS. Moreover, it can only deal with moderate sub-clones, as can be seen from its assumptions in equations 3.1 and 3.2. Lastly, it relies on the existing recurrent cancer-karyotypes to infer the most likely one. In situations of unknown type of sub-type of cancer, or incomplete database for a certain cancer type, ABSOLUTE cannot be suitably applied to infer ploidy and purity. The next method, THetA [OMR13] is designed to infer clonality and purity for NGS data and can deal with two tumor clones instead of one in ABSOLUTE.

3.2 THetA

THetA (Tumor Heterogeneity Analysis) [OMR13] takes the copy number aberrations as the input, models the tumor purity and clonal ploidy as a maximum likelihood mixture decomposition problem, and solves it by convex optimization. In particular, given the number of sub-clones in the tumor \( m \), THetA represents the copy number at each genomic locus as an entry in the interval count matrix \( C \). \( C \) is of the size \( m \times n \), in which \( n \) is the number of genomic loci. The first column of \( C \) is occupied by the copy number of contaminating normal tissue. To limit the solution space, the assumption \( m < n \) is held. On the other hand, the array \( \mu \) is used to represent the tumor and normal fraction in the total alleles. Given the array of read depth \( r \) at each genomic locus, THetA maximizes the probably \( P(r|C, \mu) \). Since the entries of \( C \) are integers, whereas those of \( \mu \) are real values, THetA solves the convex optimization problem to obtain best \( \mu \) for each possible \( C \). To further lower the search space in
$C$, the possible value of which increases exponentially with the increase of $m$ and $n$, two additional assumptions are held.

- The first column of $C$ corresponding to those from normal has all entries as 2.
- The entries of $C$ are integers varying between 0 and $k$, where $k > 2$.

Suppose all genomic loci are with the same length. The probability of a read aligning to locus $j$ for a pure clone $i$ is $\frac{C_{ij}}{\sum_i C_{ij}}$. When there are normal admixture and other tumor sub-clones, the probability is that of drawing a read from a multinomial distribution because the observed read depth $r_j$ for genomic loci $j$ is the sum of reads contributed from all clones at this particular loci.

The major limitation of THetA is the high computational demand. A polynomial algorithm has been derived when $m = 2$. Some extra derivation has been done to reduce computational complexity. However, the algorithm (beta version 0.6) can handle $m$ at maximum 3, including the normal admixture. Thus any tumor containing more than two sub-clones cannot be handled by THetA, as shown in the following chapters. Also, the assumptions THetA holds are not necessarily true. It is possible for the normal to have copy number other than 2. In the next chapter, I will describe my proposed model BreakDown, an algorithm to infer VAF from structural variations, followed by experiments including comparisons with THetA.
Chapter 4

Methods

4.1 A maximum likelihood VAF estimator

I have developed a maximum likelihood approach \cite{?}, which estimates the VAF of an SV that best explains the associated alignment data. Briefly, my approach includes the following five steps (Fig. 4.1). First, parameters such as average coverage and insert size are initialized based on measures from randomly selected regions of the genome. Read pairs encompassing an SV are extracted from the data (usually in BAM format [LHW+09]) (Fig. 4.1a). Second, these read pairs are classified into three groups: A) normal, B) discordant and C) soft-clipped, based on their alignment patterns. The numbers of read pairs \(n\), \(d\) and \(s\) are counted respectively in these 3 groups (Fig. 4.1b-c). These counts are normalized with respect to (w.r.t.) GC content and mapping quality (Fig. 4.1d). Third, given the expected numbers of counts (Fig. 4.1e) and the observed counts \(n\), \(d\) and \(s\), the maximum likelihood estimation of VAF and genotype can be obtained. VAF is a continuous variable ranging between
0 and 1 and is suitable to represent allelic structure in heterogeneous samples, while genotype is a discrete variable (e.g., AA, AB, and BB) that is suitable in homogeneous samples. I keep both variables in one formulation to achieve unified maximal likelihood inference in either heterogeneous or homogeneous samples (Fig. 4.1f). In this thesis, I focus on measuring VAFs in heterogeneous tumor samples. A Bayesian variant score is computed to quantify the confidence of the results (Fig. 4.1g).

**Figure 4.1:** Schematic overview of BreakDown. a) In the input are the bam files and a set of SV calls. Reads encompassing each SV are extracted for analysis. Genome-wide parameters such as average read count per bp per GC content are initialized. b) The encompassed reads are divided into three groups: normal, discordant and soft-clipped. Reads in each group are counted. c) Normal reads for large CNVs are counted in a series of consecutive non-overlapping bins. d) Read counts are normalized by GC contents. e) Expected read counts are defined as functions of SV, VAF and sequencing data. f) Genotype and VAF that maximize the likelihood function are derived from the expected and observed read counts. g) Variant scores are estimated that quantify the confidence of the results. Detailed explanations of the mathematics are available in Section 4.2.
Based on the above method, I implemented a software tool called BreakDown [FZC+14], which can be used in conjunction with SV discovery tools such as BreakDancer [CWM+09], GenomeSTRiP [HKNM11], Delly [RZS+12], CREST [WME+11], Pindel [YSL+09], etc. to measure the genotype or VAF of candidate SVs in a BAM file.

4.2 The Maximum Likelihood Formulation

BreakDown starts from an NGS paired-end BAM file produced by BWA [42] and a set of SV calls produced by BreakDancer, Delly, Pindel and other SV discovery tools. All these tools predict a start and an end coordinates and the variant types such as deletion, duplication, translocation and so on (Fig. 4.1a). The VAF or the genotype of a variant is determined by maximizing the following likelihood function:

\[
[g^*, p^*] = \arg \max_{[g, p]} L(D | g, p)
\] (4.1)

where \( D \) denotes the alignment data in a window \( w \) that encompasses the SV, and \( p \) the variant allele fraction (VAF) ranging from 0 to 1. In a monoclonal diploid genome, we use \( g \) to represent genotypes (AA, AB, BB), which are equivalent to \( p \) equals to 0, 0.5, and 1, respectively. Including both \( p \) and \( g \) makes it convenient to apply BreakDown to both heterogeneous and homogeneous samples. The analytic form of the likelihood function \( L \) is parameterized by read length \( r \), insert size \( t \), average number of inserts (read pairs) per bp \( c \), which are estimated from normal diploid regions of the genome.
BreakDown classifies read pairs into three groups based on their alignment to the reference and counts the numbers of: 1) normal read pair $n$, 2) discordant read pair (DRP) $d$, and 3) soft-clipped reads (SR) $s$ in $w$ (Fig. 4.1b). The definitions of these read groups are similar to previous work [MSB09, ACE11]. Soft-clipped reads are recognized from the CIGAR strings in the BWA alignment (Fig. 4.1b and http://bio-bwa.sourceforge.net/bwa.shtml). A read pair is defined as normal if its two reads align with reference in expected orientation and distance, or otherwise as discordant.

An SV typically associates with three counts $D = \{n, d, s\}$. However, for a balanced SV such as an inversion or a reciprocal translocation, the normal counts are irrelevant: $D = \{d, s\}$. If an SV contains multiple breakpoints, $d$ and $s$ each becomes an array of counts from constituent breakpoints. Unlike other SV detection methods, which sequentially analyze these different groups of reads [HKNM11, RZS+12, ZNH+12], I jointly analyze all the reads in $w$. Because read pairs in these 3 groups are sequenced and aligned independently, the likelihood function can be expanded into the following product:

$$L = P(n|p)P(d|p)P(s|p)$$  \hspace{1cm} (4.2)

which probabilistically integrates different types of counts. In an unbiased shotgun sequencing experiment, these counts should follow Poisson distributions with parameter $\lambda$ being defined in the following sections [PSMW95].
4.3 Modelling normal read pairs

An SV such as a copy number variant (CNV) can span a very large genomic region with excessive GC content variation. It is known that GC content can introduce substantial sequencing bias and need to be normalized against (as shown in Fig. 2.4) [DMW+11]. BreakDown split $w$ into smaller non-overlapping bins $w_i$ and count in each bin the number of normal read pairs $n_i$ instead of having one normal count $n$ from the entire length of the encompassing window $w$. I assume that $n_i$ follows a Poisson distribution with

$$\lambda_{ni} = c_i(\theta)w_i(1 - p)$$

(4.3)

where $c_i$ is the average number of normal read pairs per bp, normalized by the GC content in $w_i$. $c_i(\theta)$ can be pre-estimated as a function of GC-content $\theta$ from randomly selected regions in the genome (Fig. 4.1d). $w_i$ has a fixed width of 100 bp under our default setting. Splitting a large window into small bins allows normalization being performed at fine resolution and is particularly effective for large SVs that span GC-content variable region. As another corrective measure, I exclude bins that contain more than 50% of zero mapping quality reads, an indicator of potentially unreliable data that are introduced by mapping errors in repetitive regions.
4.4 Modeling discordant read pairs

The expected count of DRP $d$ should be linear to the span coverage [KAM+09], i.e., insert size $t$ minus twice read length $r$ (Fig. 4.1b):

$$\lambda_d = v_d c(\theta) (t - 2r)p,$$  \hspace{1cm} (4.4)

where $c(\theta)$ denotes the mean number of inserts per bp, a function of the GC-content. The observed numbers of DRPs often turn to be smaller than what is expected from the span coverage due to simplification in the above definition and peculiarities of alignment algorithms. I used $v_d$, a trainable parameter ranging between 0 and 1 to compensate such offset. Empirical $v_d$ is around 0.8 in typical WGS data as shown in Fig. 4.2.

4.5 Modelling soft-clipped read

The expected number of soft-clipped read $s$ should be proportional to the summation of read length (sequence coverage) in a read pair (Fig. 4.1b), assuming a read would become soft-clipped if it has any overlap with the breakpoint:

$$\lambda_s = v_s c(\theta) 2rp$$ \hspace{1cm} (4.5)

This formulation is approximate because an aligner may choose not to soft-clip a read when it only slightly overlaps the breakpoint. Sometimes, an aligner may incorrectly soft-clip a non-breakpoint containing read. However, such aligner-specific
behaviors can hardly be modeled post-alignment. To alleviate this bias, I use a trainable parameter to compensate for such offset. Empirical $v_s$ is around 0.7 in typical WGS data as shown in Fig. 4.2.

### 4.6 VAF Estimation

Taken together, the likelihood function can be expressed as

$$L = f_{\text{Pois}}(d; \lambda_d) \cdot f_{\text{Pois}}(s; \lambda_s) \cdot \prod_{i=1}^{m} f_{\text{Pois}}(n_i; \lambda_{ni}),$$

where $m$ denotes the number of bins for counting normal read pairs (Fig. 4.1c). For genomes sequenced with multiple DNA libraries, the quantities estimated from
each library are combined through multiplication, assuming that the libraries are independently constructed. Without loss of generality, I present the derivation of VAF from a single library. Solving equation \( dL/dp = 0 \) yields the variant allele fraction that maximizes the likelihood function in a close-form quadratic solution,

\[
p = \frac{-B - \sqrt{B^2 - 4AC}}{2A}
\]

(4.7)

where

\[
A = c(\theta)(t - 2r)v_d + 2rc(\theta)v_s - \sum_{i=1}^{m} c_i(\theta)w_i
\]

\[
B = -(A + d + s + \sum_{i=1}^{m} n_i)
\]

\[
C = d + s
\]

4.7 Confidence scoring

I use variant score (VarScore) to quantify error probability, i.e., the chance that there is no SV at the input site (Fig. 4.1g):

\[
S_v = -10\log_{10}(P(p = 0|D)),
\]

(4.8)

where \( P(p = 0|D) \) represents the posterior probability that VAF equals to 0 given the data. For practical implementation, I used discretized genotype to estimate the
error probabilities:

\[ S_v = -10 \log_{10}(P(g = AA|D)) \]  

(4.9)

where \( P(g = AA|D) \) is the posterior probability that the genotype is homozygous reference. This quantity can be calculated based on Bayesian Theorem:

\[ S_v = -10 \log_{10}\left( \frac{L(D|G_0)P_V(G_0)}{\sum_{l=0}^{2} L(D|G_l)P_V(G_l)} \right) \]  

(4.10)

where \( P_V(G) \) is the prior variant probability of genotype \( G \) and \( G_l \), \( l = 0, 1, \) or 2 representing homozygous reference, heterozygous variant and homozygous variant genotypes, respectively. For a heterogeneous tumor sample, uniform genotype prior is assumed \( P_V(G) = 1 \). For a homogeneous normal sample, the genotype prior can be defined based on population genetics [HKNM11].

### 4.8 Genome-wide parameter initialization

Parameters that are needed by BreakDown are initialized from the data before they are applied to VAF estimation. I randomly choose \( N \) (\( N=10 \) by default) 10 Mb regions from the BAM file (excluding centromere and telomere regions). The median read length and insert size from the data are estimated. A lookup table is created to store average read pair per bp as a function of GC content (an integer ranging from 0 to 100) (see Fig.2.4).
Chapter 5

Results

5.1 Characterizing the estimation accuracy

BreakDown involves sophisticated numerical calculations. Does it always return the correct results? To answer this question, I simulate read counts under various combinations of parameters including VAFs, variant types, variant sizes, sequence coverage, and insert sizes. In particular, I simulated a set of read counts for SNPs, deletions (with size 1K and 1Mbp), inversions or reciprocal translocations at coverages of 5X, 30X and 500X based on short insert size (500bp) and short read length (100bp). I also simulated read counts from long insert size (3000bp) and short read length (100bp) at 30X coverage (Fig. 5.1c). For each parameterization, I randomly sampled 1000 data points from the Poisson distributions. For SNVs, I assumed that the number of variant supporting reads follows a binomial distribution parameterized by the given coverage and VAF. For an inversion or a reciprocal translocation that have two breakpoints, counts at each breakpoint were simulated independently.
BreakDown was then run to estimate VAFs from simulated counts. To measure accuracy, the chance of an estimated VAF falling near (±10%) the true value in 1000 random trials was recorded.

From short-insert (500 bp) short-read (100 bp) low-coverage (5x) data (Fig. 5.1a), it is very challenging to estimate VAF accurately from SNPs, medium-size deletions (1 Kb), inversions (INV) and reciprocal translocations (TRA). However, it is possible to accurately (≥ 90% chance) estimate relatively high (≥ 0.1) VAF from large (1 Mb) deletions. When coverage increases to 30X (Fig. 5.1b), which is typically for WGS data, VAF as low as 0.05 can be accurately estimated from large deletions. Notably, VAF estimated from INVs and TRAs are more accurate than those estimated from SNVs, thanks to larger physical coverage than sequence coverage. Medium-size deletions perform always better than SNVs but worse than INVs and TRAs in low (< 0.2) VAFs. However it outperforms INVs and TRAs at high (> 0.3) VAFs. The accuracy of SVs over SNVs becomes even more striking as the insert size becomes longer (3 Kb) (Fig. 5.1c), which indicates that our method has successfully leveraged physical coverage. Even at ultra-high (500x) coverage (Fig. 5.1d), the SNVs still have limited accuracy (< 0.6) in estimating small (< 0.1) VAFs from short insert data. This result indicates that current methods that measure VAFs from only SNVs are suffering from great challenges in delineating low-abundance sub-clones, whereas when SVs are included, low-abundance sub-clones would have much higher chance to be identified.

In summary, this simulation results indicate that BreakDown can accurately estimate VAFs for various types of SVs and can enhance the heterogeneity analysis from
either short or long insert data at any coverage.

Figure 5.1: Accuracy of estimated VAF in simulation. Plotted are the chances (Y axis) of the estimated VAF falling within 10% of the true VAF (X axis). Each data point is estimated from 1000 random samples. Each subplot in the figure contains 4 curves representing the accuracy of VAFs estimated from SNVs (red plus), 1 Kb deletions (green cross), 1 Mb deletions (blue triangle), and inversions/reciprocal-translocations (purple circle). Various types of sequencing data are simulated and results compared: a) short-insert (500 bp) short-read (100 bp) at 5x sequence coverage, b) short insert short read at 30x coverage, c) long insert (3000 bp) short read at 30x coverage and d) short insert short read at 500x coverage.

5.2 Comparison with published tools

The genomes of a tumor often evolve from a complex history that spans multiple years. To understand if BreakDown is more accurate than other tools in inferring
complex history, I created a 5 clone mixture tumor sample from chr20, based on a mock phylogeny tree (Fig. 5.2). Each branch in the tree represents the birth of a new clone that contains two novel SVs, which are randomly placed non-overlapping 1.5 Mb heterozygous deletions or one-copy tandem duplications. I generated synthetic reads from this bulk tumor genome using wgsim. Additional reads were generated from the wild-type genome to simulate normal contamination that are frequently observed in real tumor samples. I created 6 data sets by varying the normal contamination rate from 0 to 0.5 with an incremental of 0.1. The total coverage was kept at a constant 50X across all conditions. All the deletions and the duplications were simulated as single copy alterations, and therefore the true VAF ranged from 0.05 to 0.3 when the normal contamination rate is 0. When the normal contamination rate is 0.5, the true VAFs ranged from 0.025 to 0.15. The synthetic reads were mapped to the wide-type chr20 reference using bwa-mem [LD10].

THetA beta version 0.60 was run under default parameters on these 6 data sets. The whole chr20 was segmented into 19 regions, corresponding to 10 non-overlapping copy number alterations with copy number neutral regions in between. The interval count file, serving as input to THetA, was generated by counting reads aligned into each of the 19 regions, for both the tumor and the normal samples. This version of THetA supported the inference of up to 3 clones. However, it reported that $n = 3$ was not a good model for this data. Therefore, all the results reported in this thesis from THetA are based on $n = 2$, i.e., one tumor clone plus one contaminating normal clone. Since the maximum copy number THetA estimated was 3, I converted copy number of those 10 intervals into VAF by $\frac{|C_i - 2|}{2} \times (1 - \mu)$, in which $C_i$ represents the copy
Table 5.1: Comparison of the VAF estimation errors between BreakDown and THetA based on simulation. For columns 3 to 8, listed are the estimation errors (estimated - true value) for BreakDown (left of slash) and THetA (right of slash) for normal contamination rate = 0, 0.1, ..., 0.5. The second column is the true VAF without normal contamination rate.

<table>
<thead>
<tr>
<th>SV</th>
<th>True VAF w/o Normal</th>
<th>0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV1_DUP</td>
<td>0.15</td>
<td>0.0180/-0.1047</td>
<td>0.0080/-0.1350</td>
<td>-0.0150/-0.1200</td>
<td>0.0170/-0.1050</td>
<td>0.0060/-0.0900</td>
<td>-0.0050/-0.0750</td>
</tr>
<tr>
<td>SV2_DEL</td>
<td>0.15</td>
<td>0.0040/-0.1500</td>
<td>-0.0160/-0.1350</td>
<td>0.0090/-0.1200</td>
<td>-0.0030/-0.1050</td>
<td>0.0120/-0.0900</td>
<td>-0.0050/-0.0750</td>
</tr>
<tr>
<td>SV3_DEL</td>
<td>0.125</td>
<td>0.0130/-0.1250</td>
<td>-0.0025/-0.1125</td>
<td>-0.0140/-0.1000</td>
<td>0.0125/-0.0875</td>
<td>-0.0030/-0.0750</td>
<td>-0.0185/-0.0625</td>
</tr>
<tr>
<td>SV4_DUP</td>
<td>0.125</td>
<td>-0.0100/-0.1250</td>
<td>-0.0155/-0.1125</td>
<td>0.0090/-0.1000</td>
<td>-0.0025/-0.0875</td>
<td>-0.0080/-0.0750</td>
<td>0.0225/-0.0625</td>
</tr>
<tr>
<td>SV5_DUP</td>
<td>0.1</td>
<td>-0.0150/-0.0453</td>
<td>0.0020/-0.0184</td>
<td>0.0220/-0.2400</td>
<td>-0.0020/-0.2100</td>
<td>0.0150/-0.1800</td>
<td>0.0280/-0.1500</td>
</tr>
<tr>
<td>SV6_DEL</td>
<td>0.3</td>
<td>-0.0120/-0.0453</td>
<td>-0.0070/-0.0184</td>
<td>0.0010/-0.2400</td>
<td>0.0100/-0.2100</td>
<td>-0.0180/-0.1800</td>
<td>-0.0110/-0.1500</td>
</tr>
<tr>
<td>SV7_DEL</td>
<td>0.25</td>
<td>0.0180/0.0047</td>
<td>-0.0110/0.0266</td>
<td>0.0040/0.2000</td>
<td>0.0210/-0.1750</td>
<td>-0.0080/-0.1500</td>
<td>0.0020/-0.1250</td>
</tr>
<tr>
<td>SV8_DUP</td>
<td>0.25</td>
<td>-0.0090/0.0047</td>
<td>0.0150/0.0266</td>
<td>-0.0090/-0.2000</td>
<td>0.0020/-0.1750</td>
<td>-0.0130/-0.1500</td>
<td>0.0070/-0.1250</td>
</tr>
<tr>
<td>SV9_DUP</td>
<td>0.05</td>
<td>-0.0040/-0.0500</td>
<td>-0.0030/-0.0450</td>
<td>-0.0050/-0.0400</td>
<td>-0.0140/-0.0350</td>
<td>-0.0120/-0.0300</td>
<td>0.0210/-0.0250</td>
</tr>
<tr>
<td>SV10_DEL</td>
<td>0.05</td>
<td>0.0180/-0.0500</td>
<td>0.0050/-0.0450</td>
<td>0.0040/-0.0400</td>
<td>0.0000/-0.0350</td>
<td>-0.0070/-0.0300</td>
<td>-0.0150/-0.0250</td>
</tr>
</tbody>
</table>

number estimated for the $i$th interval, and $\mu$ the estimated normal contamination rate.

I ran BreakDown with default parameters and compared the accuracy of estimated VAFs with that of THetA. In all 6 sets, BreakDown achieved lower VAF estimation errors than THetA (Fig. 5.3). It was noticed that the accuracy of THetA starts to deteriorate when the normal contamination rate increases, whereas BreakDown achieved consistently small errors ($<0.028$). This result indicated that BreakDown is likely more accurate in modeling real tumor sample that often contain multiple tumor clones with variable rates of normal contamination.

### 5.3 Identifying somatic SVs from tumor-normal matched WGS data

To assess BreakDown’s capability in accurately estimating VAF and detecting somatic events in real data, I downloaded the WGS data of a metastatic melanoma
cancer cell-line COLO-829 and its matched normal cell-line [PCS+10]. For each putative SV, I computed its VAF in the tumor and the normal samples independently. I first ran BreakDown on 32 previously validated deletions in the tumor [WME+11, PCS+10]. Except for 5 events, which are suspected to be germline (3 out of these 5 are also reported as germline by CREST), the estimated VAFs in the normal sample for the other 27 events (see Table 5.2) had a mean of 0.024 and a standard deviation of 0.032, as expected for somatic deletions.
Table 5.2: A list of 32 previously reported deletions with BreakDown estimated VAFs in COLO-829. The Germline_Tag column indicates the structural variation is estimated by BreakDown to be somatic if 0, putative germline if 1 and validated germline if 2. Validation is reported by CREST. The Polyploid_Tag column indicates the ploidy of the genomic loci of the structural variation: partly diploid for 0, polyploid for 1 and mostly diploid for 2.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>DEL_Start</th>
<th>DEL_End</th>
<th>Tumor_VAF</th>
<th>Normal_VAF</th>
<th>Germline_Tag</th>
<th>Polyploidy_Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>89690279</td>
<td>89702321</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>X</td>
<td>33969699</td>
<td>33972626</td>
<td>1</td>
<td>0.08</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>X</td>
<td>32008454</td>
<td>32111172</td>
<td>0.92</td>
<td>0.05</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>57182046</td>
<td>57220832</td>
<td>1</td>
<td>0.13</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>X</td>
<td>31987776</td>
<td>32203018</td>
<td>0.69</td>
<td>0.05</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>14948624</td>
<td>14961844</td>
<td>0.57</td>
<td>0.04</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>28024299</td>
<td>28147692</td>
<td>0.53</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>77486140</td>
<td>77652508</td>
<td>0.56</td>
<td>0.04</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>X</td>
<td>31211124</td>
<td>31948671</td>
<td>0.52</td>
<td>0.01</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>48401118</td>
<td>48403688</td>
<td>1</td>
<td>0.5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>X</td>
<td>31106864</td>
<td>3126132</td>
<td>0.51</td>
<td>0.01</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>85512695</td>
<td>85518865</td>
<td>1</td>
<td>0.52</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>28823753</td>
<td>28998853</td>
<td>0.54</td>
<td>0.07</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>13108730</td>
<td>13112103</td>
<td>0.45</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>125533359</td>
<td>125954137</td>
<td>0.43</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>80463250</td>
<td>80771428</td>
<td>0.43</td>
<td>0.01</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>110180568</td>
<td>110181697</td>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>14910958</td>
<td>14961951</td>
<td>0.37</td>
<td>0.05</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>32089249</td>
<td>32168399</td>
<td>0.28</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>64530442</td>
<td>64538080</td>
<td>0.29</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>60107794</td>
<td>60179583</td>
<td>0.27</td>
<td>0.02</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>77819775</td>
<td>77920185</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>78028143</td>
<td>78095623</td>
<td>0.24</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>6047165</td>
<td>60988255</td>
<td>0.26</td>
<td>0.03</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>61955582</td>
<td>62319120</td>
<td>0.19</td>
<td>0.02</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>14936825</td>
<td>15056818</td>
<td>0.18</td>
<td>0.04</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>7277162</td>
<td>7278457</td>
<td>0.66</td>
<td>0.57</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>222849418</td>
<td>222852657</td>
<td>0.07</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>131181342</td>
<td>131224000</td>
<td>0.5</td>
<td>0.53</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>206047856</td>
<td>206081445</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>65803522</td>
<td>65814988</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>143390165</td>
<td>143719727</td>
<td>0.69</td>
<td>0.69</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 5.3: Comparison of VAF estimation errors between BreakDown and THetA. Plotted are the mean and the maximum errors (Y axis) estimated from 10 lineage specific SVs by BreakDown and THetA under six normal contamination rates from 0 to 0.5 (X axis). Details are described in the text and Table 5.4.

In the tumor sample, I observed that those events on chromosome 2, 3, 7, 17, 18, 20, 22 and X have VAFs diverging from 0, 0.5 or 1, which were unexpected from a homogeneous diploid sample. This implied aneuploidy on these chromosomes, assuming the tumor sample is pure. This speculation was confirmed by a previous independent study that characterized the genome-wide copy number profile in this sample [GWP+12]. On chromosomes 5, 10, 15 and 16 that were indicated as mostly diploid [GWP+12], the estimated VAFs were within 0.06 of either 0.5 or 1.0 (dot plot shown in Fig. 5.4). Thus, the VAFs estimated by Breakdown in both the tumor and
the normal samples appeared to be valid.

Figure 5.4: Plots of estimated VAF of validated deletions. Diamond, triangle and circles represent validated germline, putative germline and somatic events. For tumor ploidy analysis, those events on the chromosomes which are mostly diploid are in green, and mostly polyploid in red. Those that are partly diploid are in black.

I ran BreakDancer on the tumor and the normal samples independently, and ran Breakdown on the resulting deletion calls. Since previous studies have been fairly comprehensive at identifying somatic heterozygous deletions (gain-of-heterozygosity
or GOH) I focused on identifying loss-of-heterozygosity (LOH) events. 41 candidate LOH events were found that have VAFs between 0.45 and 0.55 in the normal and between 0 and 0.05 or 0.95 and 1 in the tumor (see Table 5.3). Among them, 40 (97.6%) overlapped previously reported LOH regions deriving from segmenting the B-allele frequency of single nucleotide variations [PCS+10]. Two LOHs were of potential functional impact. One at chr10q22.3 hit C10orf11, a melanocyte-differentiation gene that has been reported to relate with autosomal-recessive albinism in humans [GnDS+13]. This homozygous deletion overlapped the only homozygous region found in four affected individuals but not in any unaffected ones. The other LOH at chr14q31.1 hit gene NRXN3 (Fig. 5.5), which has been reported to relate to malignant melanoma http://www.ncbi.nlm.nih.gov/clinvar?term=NRXN3.

5.4 Identify sub-clonal SVs in breast cancer samples

I analyzed three pairs of matched tumor-normal breast cancer samples to show that BreakDown can accurately estimate VAFs of medium-size deletions, inversions and translocations. The first pair consists of an estrogen-receptor (ER)-positive primary breast cancer sample PD4120, sequenced by Illumina Hi-seq at 188x, and a matched normal sample at 38x. The clonal structure of this tumor has been previously inferred based on SNVs and large CNVs [NZWL+12, OMR13, FVGI14]. The other two tumor samples PD4115 and PD4088 were sequenced at around 40x
Table 5.3: A list of 41 LOH deletions detected by BreakDown in COLO-829. The VAF_Diff column lists the absolute difference of estimated tumor and normal VAF by BreakDown; the Hit_Pleasance_LOH column lists the LOH estimated by [PCS+10] overlapping with this entry; the Hit_Gene lists the hit gene if there is any.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>DEL_Start</th>
<th>DEL_End</th>
<th>VAF_Diff</th>
<th>Hit_Pleasance_LOH</th>
<th>Hit_Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>56282888</td>
<td>56285348</td>
<td>0.55</td>
<td>16:33325823-88690776</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>98372949</td>
<td>98375378</td>
<td>0.55</td>
<td>5:81949-180629495</td>
<td>SBF2</td>
</tr>
<tr>
<td>11</td>
<td>10216431</td>
<td>10217898</td>
<td>0.51</td>
<td>11:188510-134449982</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>65179249</td>
<td>65181406</td>
<td>0.45</td>
<td>10:33409859-135284541</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>97427311</td>
<td>97428510</td>
<td>0.45</td>
<td>5:81949-180629495</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>82099212</td>
<td>82101905</td>
<td>0.44</td>
<td>5:81949-180629495</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>77925584</td>
<td>77931020</td>
<td>0.54</td>
<td>10:33409859-135284541</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>8300680</td>
<td>8301820</td>
<td>0.54</td>
<td>3:41866-2538574</td>
<td>LOC100288428</td>
</tr>
<tr>
<td>14</td>
<td>4840822</td>
<td>48403687</td>
<td>0.53</td>
<td>14:19272965-106356482</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>61031662</td>
<td>61035008</td>
<td>0.52</td>
<td>10:33409859-135284541</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>54641943</td>
<td>54648825</td>
<td>0.52</td>
<td>11:188510-134449982</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>46179972</td>
<td>46181746</td>
<td>0.53</td>
<td>5:81949-180629495</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>127435327</td>
<td>127438813</td>
<td>0.53</td>
<td>5:81949-180629495</td>
<td>FLJ33630</td>
</tr>
<tr>
<td>5</td>
<td>65063854</td>
<td>65066134</td>
<td>0.52</td>
<td>5:81949-180629495</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>67086658</td>
<td>67089030</td>
<td>0.42</td>
<td>11:188510-134449982</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>157227910</td>
<td>157232829</td>
<td>0.52</td>
<td>1:86818284-274165315</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>39679221</td>
<td>39687427</td>
<td>0.51</td>
<td>14:19272965-106356482</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>56352780</td>
<td>56354363</td>
<td>0.48</td>
<td>10:33409859-135284541</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>21465015</td>
<td>21495342</td>
<td>0.47</td>
<td>NA</td>
<td>MIR650</td>
</tr>
<tr>
<td>5</td>
<td>50270142</td>
<td>50275225</td>
<td>0.47</td>
<td>5:81949-180629495</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>7746850</td>
<td>7748577</td>
<td>0.51</td>
<td>11:188510-134449982</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>88604135</td>
<td>88607580</td>
<td>0.51</td>
<td>11:188510-134449982</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>10377958</td>
<td>10377845</td>
<td>0.51</td>
<td>11:188510-134449982</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>79175898</td>
<td>79184805</td>
<td>0.51</td>
<td>14:19272965-106356482</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>66068841</td>
<td>66070741</td>
<td>0.49</td>
<td>10:33409859-135284541</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>45040578</td>
<td>45050283</td>
<td>0.49</td>
<td>5:81949-180629495</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>31350332</td>
<td>31354240</td>
<td>0.48</td>
<td>11:188510-134449982</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>112493349</td>
<td>112506227</td>
<td>0.5</td>
<td>1:86818284-274165315</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>69668624</td>
<td>69670027</td>
<td>0.49</td>
<td>15:21520763-82648782</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>54453533</td>
<td>54458848</td>
<td>0.49</td>
<td>10:33409859-135284541</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>84117796</td>
<td>84120351</td>
<td>0.49</td>
<td>10:33409859-135284541</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>176190593</td>
<td>176192015</td>
<td>0.5</td>
<td>5:81949-180629495</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>60493110</td>
<td>60495082</td>
<td>0.48</td>
<td>15:21520763-82648782</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>227879686</td>
<td>227887613</td>
<td>0.48</td>
<td>1:86818284-274165315</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>241194400</td>
<td>241195846</td>
<td>0.48</td>
<td>1:86818284-274165315</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>119460219</td>
<td>119461278</td>
<td>0.48</td>
<td>11:188510-134449982</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>51987074</td>
<td>51990577</td>
<td>0.46</td>
<td>15:21520763-82648782</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>6690271</td>
<td>6700672</td>
<td>0.47</td>
<td>9:30910-30341247</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>132798754</td>
<td>132802774</td>
<td>0.45</td>
<td>10:33409859-135284541</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>58350768</td>
<td>58352594</td>
<td>0.46</td>
<td>11:188510-134449982</td>
<td>TCERG1L</td>
</tr>
<tr>
<td>7</td>
<td>125832879</td>
<td>125838697</td>
<td>0.45</td>
<td>7:125536046-125880817</td>
<td></td>
</tr>
</tbody>
</table>
and their clonal structure have been characterized by THetA [OMR13]. For these 3 cases, I called structural variation using BreakDancer on the paired tumor and normal samples. Then BreakDown was run on each of the candidate SV calls. Through this process I discovered several subclonal deletions (all of which are shorter than 10 Kb), inversion and translocation (Table 5.4) that have not been previously reported [NZAW+12].

For the novel somatic SVs detected in PD4120, two deletions (at chr5q14.3 and chr14q22.1) have BreakDown predicted VAFs of 0.38 and 0.39, respectively (see...
Table 5.4: Novel sub-clonal somatic structural variants detected by BreakDown. Listed are the novel somatic structural variants in the breast cancer sample PD4120, PD4115 and PD4088 along with BreakDown estimated VAFs (column 9). Also shown (column 8) are the genes overlapping with either deletion loci or translocation breakpoints.

<table>
<thead>
<tr>
<th>SAMPLE</th>
<th>CHR1</th>
<th>POS1</th>
<th>CHR2</th>
<th>POS2</th>
<th>TYPE</th>
<th>SIZE (bp)</th>
<th>GENES</th>
<th>EST_VAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD4120</td>
<td>10</td>
<td>117351105</td>
<td>10</td>
<td>117352856</td>
<td>Deletion</td>
<td>1751</td>
<td>ATRNL1</td>
<td>0.12</td>
</tr>
<tr>
<td>PD4120</td>
<td>5</td>
<td>86223184</td>
<td>5</td>
<td>86259088</td>
<td>Deletion</td>
<td>2804</td>
<td></td>
<td>0.38</td>
</tr>
<tr>
<td>PD4120</td>
<td>14</td>
<td>5189637</td>
<td>14</td>
<td>51896099</td>
<td>Deletion</td>
<td>6462</td>
<td></td>
<td>0.39</td>
</tr>
<tr>
<td>PD4120</td>
<td>22</td>
<td>19372768</td>
<td>22</td>
<td>21122617</td>
<td>Translocation</td>
<td>N/A</td>
<td>HIRA, PI4KA</td>
<td>0.055</td>
</tr>
<tr>
<td>PD4115</td>
<td>1</td>
<td>10151402</td>
<td>1</td>
<td>10152851</td>
<td>Deletion</td>
<td>1449</td>
<td>UBE4B</td>
<td>0.296</td>
</tr>
<tr>
<td>PD4115</td>
<td>9</td>
<td>108351356</td>
<td>9</td>
<td>108352767</td>
<td>Deletion</td>
<td>1411</td>
<td>FKTN</td>
<td>0.285</td>
</tr>
<tr>
<td>PD4115</td>
<td>9</td>
<td>140773612</td>
<td>9</td>
<td>140777195</td>
<td>Deletion</td>
<td>3583</td>
<td>CACNA1B</td>
<td>0.400</td>
</tr>
<tr>
<td>PD4115</td>
<td>X</td>
<td>73961886</td>
<td>X</td>
<td>73967477</td>
<td>Deletion</td>
<td>5591</td>
<td>XIST</td>
<td>0.394</td>
</tr>
<tr>
<td>PD4088</td>
<td>10</td>
<td>60908347</td>
<td>10</td>
<td>60920370</td>
<td>Inversion</td>
<td>11823</td>
<td></td>
<td>0.594</td>
</tr>
</tbody>
</table>

Fig. 5.6). Because the tumor purity of PD4120 is around 70% [NZVLW+12, OMR13], these two deletions are thereby likely heterozygous deletions in the primary clone (cluster D in [NZVLW+12]). The third deletion (at chr10q25.3) (Fig. 5.6a) has a predicted VAF of 0.12, which may originate from one of the sub-clones (cluster B in [NZVLW+12]) that has a SNV VAF peak at 0.11. It can be seen Fig. 5.6a that although the read depth signal is not apparent, there are 15 discordant read pairs and 4 split reads, all of which are important for estimating VAF but have been ignored in the previous studies. The translocation Fig. 5.6b has a BreakDown estimated VAF of 0.055, which matches another sub-clone (cluster A in [NZVLW+12]) with a SNV VAF peak at 0.05. For the four deletions discovered in PD4115 (76% tumor purity [OMR13]), two (at chr9q34.3 and chrXq13.2) have VAFs estimated at around 0.4, which implies that they are likely heterozygous variants in the founding clone. The other two (at chr1p36.22 and chr9q31.2) had estimated VAFs at around 0.3, and are likely homozygous deletions in the subclone of 32.7% abundance [OMR13]. One novel somatic inversion (at chr10q21.1) found in PD4088 (59% tumor purity [OMR13]) has
an estimated VAF 0.594, which may be a homozygous event in the founding clone. In
summary, all of the novel somatic SVs I have identified were consistent with previously
inferred clonal architecture, which demonstrated the accuracy of our method.

**Figure 5.6:** Sub-clonal SVs in PD4120. Plotted are the integrative genome viewer (IGV)
screenshots that display reads in the tumor sample (top panel) and in the normal sample
(bottom panel) of a) A 1751 bp deletion between Chr10:117351105 and Chr10:117352856
with an estimated tumor VAF 0.12, and b) an intra-chromosomal translocation between
Chr22:19372768 and Chr22:21122617 with an estimated tumor VAF 0.055. Reads are dis-
played from the top to the bottom in the following order: split reads (partially clipped
bars), discordant read pairs (brown bars and lines), and normal read pairs.
Chapter 6

BreakDown

In this chapter, I will introduce the software tool BreakDown, including its usage, examples etc. BreakDown [FZC+14] (https://bitbucket.org/xianfan/breakdown) is a perl package that estimates VAF and infers genotypes of structural variants from the next generation paired-end sequencing reads. There are two steps in BreakDown. First, \texttt{bam2cfg.pl} is used to generate a configuration file of the sequence data. Next, \texttt{BreakDown.pl} uses this configuration file and the putative SV files to generate a VAF/genotype summary file. The followings are the usages and program dependency.

6.1 Input

- A configuration file describing meta-data of bam files. This is the output of Step 1.

- Reference file from which the alignment algorithm was used for generating
the bam file.

• An SV file that is in either user-defined format or a BreakDancer file format. For a user-defined format, a header line beginning with # with the following columns are required in the SV file.

    chr1 start chr2 end type size

They are respectively chromosome 1, position1, chromosome2, position 2, type of the SV (one of DEL, CTX, DUP, INS, INV and ITX, see explanation in the text below for BreakDancer format), and size of the SV. The rows in the SV file other than the header line should have the columns corresponding to the header, but not with # as the first character. If your SV file contains more columns or extra columns in between the above required ones, BreakDown will not be affected. The only thing that BreakDown requires to find the entries is the correct header strings, as indicated in the example above. Please add option -z to indicate that this is a user-defined SV format.

On another hand, if you feed BreakDown with BreakDancer output format, which should be with the following columns, option -z should not be used.

1. Chromosome 1
2. Position 1
3. Orientation 1
4. Chromosome 2
5. Position 2

6. Orientation 2

7. Type of a SV

8. Size of a SV

9. Confidence Score

Columns 1-3 and 4-6 are used to specify the coordinates of the two SV breakpoints. The orientation is a string that records the number of reads mapped to the plus (+) or the minus (-) strand in the anchoring regions.

Column 7 is the type of SV detected: DEL (deletions), INS (insertion), INV (inversion), ITX (intra-chromosomal translocation), CTX (inter-chromosomal translocation), and Unknown. DUP (tandem duplication) is supported by BreakDown1.1.

Column 8 is the size of the SV in bp. It is meaningless for inter-chromosomal translocations.

Column 9 is the confidence score associated with the prediction.

If BreakDancer is used to generate putative SV calls, the output of BreakDancer is in the qualified format already.

6.2 Output

BreakDown’s output file consists of the following columns:
1. Chromosome 1
2. Position 1
3. Chromosome 2
4. Position 2
5. Type of an SV
6. Size of an SV
7. Variant Score
8. Inferred Genotype
9. Inferred VAF
10. Genotype Score

Columns 1-6 are given in the input SV file, describing the SV coordinates, type and size.

Columns 7-10 are inferred from BreakDown algorithm. Specifically, column 9 is the inferred VAF, ranging from 0 to 1, with the corresponding variant score in column 7.

Column 8 and 10 are genotype estimation (categorized into homo-ref, het-variant and homo-variant) and its genotype score. These are mainly for normal genomes without genetic diseases, and its beyond the scope of this thesis since cancer genomics is the focus here.
6.3 Steps

Step 1. *bam2cfg.pl* to generate configuration files. Usage:

```perl
perl bam2cfg.pl bam_files
```

Inputs are all the bam files separated by space. The usage can be found in [CWM+09].

Step 2. *BreakDown.pl* to estimate VAF and infer genotype for structural variations. A few options useful to users are listed below. Usage:

```bash
BreakDown.pl <BreakDancer configuration file> <SV file>
```

Options:

- `-g` Turn on GC correction, default: off
- `-d` DIR Directory where all intermediate files are saved
- `-c` STR Select chromosome STR
- `-q` INT Only analyze reads with mapping quality ≥ [0]
- `-r` FLOAT Prior of SVs in the genome [0.0001]
- `-x` STR Type of evidences included in VAF and genotype estimation: n,d,v,nd,rv,dr,ndv (n: normal, d: discordant, v: variant)
- `-F` STR Path to the samtools faidx-ed reference sequence (useful for GC estimation)
- `-S` Skip sites with size bigger than [1000000]
- **-b**  Bin size. Should be larger when genome coverage is low. [100]

- **-P**  The mean mapping quality of a bin below which the bin will be not participated in the genotyping decision and AF estimation. [30]

- **-y**  The libraries to be ignored, with colon to separate.

Most of these options are self-explanatory. If the bam file contains only of one chromosome, **-c** option needs to be specified to avoid accessing other chromosomes not existing in the bam during initialization. If only one or two of the evidences are included for VAF/genotype estimation, **-x** option is used for the specification. **n** stands for normal reads, **d** for discordant and **v** for soft-clipped reads. **-g** turns on GC correction, which is possible only when a whole chromosome or a whole genome exists in the bam. Otherwise the algorithm will make a homogeneous assumption and takes the estimates from bam2cfg script. When GC correction is on, the reference sequence is required, and could be specified by **-F** option. When putative SV has extremely large size, BreakDown has the option to skip it to avoid stucking on one SV by option **-S**. Bin size can be tuned by option **-b** in unbalanced SVs. 100 for small SVs (size < 1Kbp) or 200 for large SVs (size ≥ 1Kbp) is reasonable. When GC correction is off, this option is deactivated. Skipping low quality bins in VAF estimation of genotyping is made through **-P** option, so that the specified mean mapping quality threshold disqualifies those highly repetitive bins. If a specific set of libraries is to be ignored, use **-y** option, with colon separating each of the library.
6.4 Examples

Three examples of running BreakDown are listed (bam2cfg step is skipped and please refer to [CWM+09] for details).

- **Example 1** Estimate VAF and genotype on a whole chromosome/genome bam file. Since the bam file contains the whole chromosome/genome, it is possible to do GC correction. Please refer to example 3 if this is not preferred.

  \[
  \text{perl BreakDown.pl -g -F reference_file config_file SV_file}
  \]

  If the bam file does not contain all 22 autosomes, please specify the chromosome from which the GC correction is learned by `-c`. For example, if only chromosome 1 exists in the bam file, the following is the command.

  \[
  \text{perl BreakDown.pl -c 1 -g -F reference_file config_file SV_file}
  \]

  It takes around 15-20 minutes for GC correction initialization on a standard 30-50X NGS file. Once the initialization is done, the process of each SV goes much faster, < 1 second for a medium sized SV.

  A few temporary files from GC correction are written and stored for potential rerun to remove randomness. They are gc.tmp.[library_names] and cfg.tmp. Please store them in a separate folder in case they are overwritten if you will run the program from the same location again. To rerun with the stored temporary files, please refer to Example 2.

- **Example 2** Estimate VAF and genotype with GC files.
perl BreakDown.pl -g -F reference_file config_file SV_file
"gc.tmp.*"

If the third input argument is not empty, the program will look for the temporary files with given names. Once the contents of the files are loaded, it will directly do VAF and genotype estimation, skipping the initialization step.

• **Example 3** Estimate VAF and genotype without GC correction. When GC correction is not needed (such as in simple simulation data) or not possible (a bam covering only a small part of the genome surrounding the SV), it is possible to turn GC correction off. The command is the same as that in Example 1 except without -g option. Homogeneous insert/bp versus GC content is assumed in this case.

### 6.5 Dependence on Other Perl Modules

CPAN module Statistics::Descriptive is used by both `bam2cfg.pl` and `BreakDown.pl`, which can be downloaded at

http://search.cpan.org/~colink/Statistics-Descriptive-2.6/Descriptive.pm
Chapter 7

Conclusions and Future Research

In this thesis, I presented a novel sequence analysis method that can estimate the VAFs of structural variations from a heterogeneous tumor sample.

I showed that VAFs estimated from structural variations by BreakDown are at least several times more accurate than those estimated from single nucleotide variations, thanks to the integration of diverse alignment signals (or coverage) from multiple groups of reads. I also clearly demonstrated that different types of structural variations are associated with different accuracy in VAF estimation and BreakDown can fully harness the structural nature of these structural variations.

This work has extended current clonal inference from single nucleotide variations and large copy number variations to medium-sized structural variations and balanced structural variations such as inversions and translocations. It is possible to further extend it to account for complex structural variations such as chromothripsis [SGF+11, ZLP13, KC13] and chromoplexy [She13, BPL+14].

In terms of accurately estimating VAFs, BreakDown compared favorably to exist-
ing tools. In the simulation, it could more reliably estimate VAFs than THetA from tumor samples that have multiple clones and a high level of normal contamination. Other approaches such as ABSOLUTE were not directly comparable to BreakDown, because they were designed to infer tumor purity and ploidy without further characterizing clonal structure or subclonal mutations [CCH\textsuperscript{+12}]. In the analysis of sample COLO-829, I identified 40 putative germline deletions in somatic LOH regions. These events, although potentially important, have not been systematically reported in previous studies that examined the same set of data. Therefore, the results obtained by BreakDown can facilitate more accurate characterization than the previous studies that utilized only single nucleotide variations and indels [PCS\textsuperscript{+10}]. These findings also indicated that previous studies might have narrowly focused on detecting gain of heterozygosities (GOHs) but somewhat ignored LOHs. As medical genomics research continues, it is expected that BreakDown will have a tremendous opportunity to improve the comprehensiveness of mutational profiling from an unprecedented amount of WGS data that are accumulated by the tumor genome atlas, the international cancer genome consortium [HAA\textsuperscript{+10}] and broader biomedical research communities.

The analysis of breast cancer samples PD4120, PD4115 and PD4088 demonstrated the identification of subclone SVs using BreakDown. The estimated VAF, together with those from single nucleotide variations and large copy number variations, will serve as input to further infer clonality, ploidy and purity. While these three perspectives are intermingled, I believe that incorporating structural variations into the picture [CCH\textsuperscript{+12}, RKY\textsuperscript{+14}, OMR13, FVGIM14] will greatly enhance the quality of the inference.
Future research involves using estimated VAFs from SVs to help with tumor heterogeneity characterization. As previous methods [CCH+12, OMR13] incorporate only copy number and SNVs to estimate tumor clonality, ploidy and purity, now that the VAFs of each SV is available, more data at a different layer can be incorporated for a more complete analysis. Also, the combined study of VAFs on SV and SNV will yield a more complete understanding of the mechanism of mutations.
Bibliography


[BPL+14] Sylvan C. Baca, Davide Prandi, Michael S. Lawrence, Juan M. Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, Naoki Kitabayashi, Theresa Y. MacDonald, Mahmoud Ghandi, Eliezer Van Allen, Gregory V. Kryukov, Andrea Sboner, Jean-Philippe Theurillat, T. David Soong, Elizabeth Nickerson, Daniel Auclair,


[Con10] The 1000 Genomes Project Consortium, *A map of human genome variation from population-scale sequencing*, Nature **467** (2010), no. 7319,


Li Ding, Timothy J. Ley, David E. Larson, Christopher A. Miller, Daniel C. Koboldt, John S. Welch, Julie K. Ritchey, Margaret A. Young, Tamara Lamprecht, Michael D. McLellan, Joshua F. McMichael, John W. Wallis, Charles Lu, Dong Shen, Christopher C. Harris, David J. Dooling, Robert S. Fulton, Lucinda L. Fulton, Ken Chen, Heather Schmidt, Joelle Kalicki-Veizer, Vincent J. Magrini, Lisa Cook, Sean D. McGrath, Tammi L. Vickery, Michael C. Wendl,


[LD10] Heng Li and Richard Durbin, *Fast and accurate long-read alignment


[MWS+11] Ryan E. Mills, Klaudia Walter, Chip Stewart, Robert E. Handsaker, Ken Chen, Can Alkan, Alexej Abyzov, Seungtae C. Yoon, Kai Ye, R. Keira Cheetham, Asif Chinwalla, Donald F. Conrad, Yutao Fu,


[XDS+14] Huilei Xu, John DiCarlo, Ravi V. Satya, Quan Peng, and Yexun Wang, Comparison of somatic mutation calling methods in amplicon and whole exome sequence data, BMC Genomics 15 (2014), no. 1, 244+.

[XT09] Chao Xie and Martti Tammi, CNV-seq, a new method to detect copy number variation using high-throughput sequencing, BMC Bioinformatics 10 (2009), no. 1, 80+.


[YXM+09] Seungtai Yoon, Zhenyu Xuan, Vladimir Makarov, Kenny Ye, and Jonathan Sebat, Sensitive and accurate detection of copy number variants using read depth of coverage., Genome research 19 (2009), no. 9, 1586–1592.