Structural variation detection in the human genome

Author: Jiantao Wu

Boston College

The Graduate School of Arts and Sciences

Department of Biology

Structural Variation Detection in the Human Genome

a dissertation

by

Jiantao Wu

submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

December 2013

Structural Variation Detection in the Human Genome

Abstract

Jiantao Wu
Dissertation advisor: Gabor T. Marth

Structural variations (SVs), like single nucleotide polymorphisms (SNPs) and short insertion-deletion polymorphisms (INDELs), are a ubiquitous feature of genomic sequences and are major contributors to human genetic diversity and disease. Due to technical difficulties, *i.e.* the high data-acquisition cost and/or low detection resolution of previous genome-scanning technologies, this source of genetic variation has not been well studied until the completion of the Human Genome Project and the emergence of next-generation sequencing (NGS) technologies. The assembly of the human genome and economical high-throughput sequencing technologies enable the development of numerous new SV detection algorithms with unprecedented accuracy, sensitivity and precision.

Although a number of SV detection programs have been developed for various SV types, such as copy number variations, deletions, tandem duplications, inversions and translocations, some types of SVs, *e.g.* copy number variations (CNVs) in capture sequencing data and mobile element insertions (MEIs) have undergone limited study. This is a result of the lack of suitable statistical models and computational approaches, *e.g.* efficient mapping method to handle multiple aligned reads from mobile element (ME) sequences.

The focus of my dissertation was to identify and characterize CNVs in capture sequencing data and MEI from large-scale whole-genome sequencing data. This was achieved by building sophisticated statistical models and developing efficient algorithms and analysis methods for NGS data. In Chapter 2, I present a novel algorithm that uses the read depth (RD) signal to detect CNVs in deep-coverage exon capture sequencing data that are originally designed for SNPs discovery. We were one of the early pioneers to tackle this problem. In Chapter 3, I present a fast, convenient and memory-efficient program, Tangram, that integrates read-pair (RP) and split-read (SR) signals to detect and genotype MEI events. Based on the results from both simulated and

experimental data, Tangram has superior sensitivity, specificity, breakpoint resolution and genotyping accuracy, when compared to other recently published MEI detection methods. Lastly, Chapter 4 summarizes my work for SV detection in human genomes during my PhD study and describes the future direction of genetic variant researches.

Contents

1 Introduction 1

1.1 Structural variation in human diseases and phenotypes 2

1.2 Technologies for SV detection 4

1.2.1 Cytogenetic methods 4

1.2.2 Microarray 6

1.2.3 Next-generation sequencing 8

1.3 Algorithms for SV detection with NGS data 10

1.3.1 Read depth algorithm 11

1.3.2 Read pair algorithm 14

1.3.3 Split read algorithm 18

2 CNV detection from exon capture sequencing data 21

2.1 Introduction 22

2.2 Results 24

2.2.1 Brief algorithmic overview 24

2.2.2 Dataset 26

2.2.3 Sample coverage and gene affinities 28

2.2.4 CNV candidates detected in the data 30

2.2.5 Call-set accuracy assessment 34

2.2.6 Sensitivity 38

2.2.7 The number of CNV events in the samples 39

2.2.8 Detection efficiency as a function of data quantity and data quality 39

2.2.9 Functions of affected genes 43

2.3 Discussion 45

2.4 Methods 46

2.4.1 Observed read depth 46

2.4.2 Data filtering 47

2.4.3 Expected read depth based on uni-linear fit and tri-linear fit 47
2.4.4 Copy number probabilities 48
2.4.5 Insight from empirical data and accounting for over-dispersion 49
2.4.6 Neighboring gene deletions 52
2.4.7 Sensitivity estimation 52
2.4.8 Common CNVs 52
2.4.9 Validation experiments 56

3 Tangram: An inclusive toolbox for MEI detection **57**
3.1 Introduction 58
3.2 Results 60
3.2.1 Performance evaluation on simulated datasets 60
3.2.2 Performance comparisons using 1000 Genomes Project data 63
3.2.3 Running Tangram on population data 66
3.2.4 Experimental validation 68
3.2.5 Resource requirements and software availability 73
3.3 Discussion 74
3.4 Methods 75
3.4.1 The algorithmic overview of Tangram 75
3.4.2 Sequence alignment to genome and mobile element reference 76
3.4.3 MEI detection based on read-pair mapping positions 77
3.4.4 MEI detection based on split-read mapping positions 77
3.4.5 Candidate MEI event filtering and post-processing 78
3.4.6 Sample genotype calling and genotype likelihood calculation 78
3.4.7 Simulation data generation 79
3.4.8 Genotype mixing 80
3.4.9 Alignments for RetroSeq 80
3.4.10 Identification of events across MEI callsets 80
3.4.11 PCR validation 81
3.5 Software availability 82

4 Concluding Remarks **83**
4.1 Summary 84
4.1.1 CNV detection from exon capture sequencing data 84
4.1.2 Tangram: An inclusive toolbox for MEI detection 85
4.1.3 Discussion 86
4.2 Future directions 88

4.2.1 Current challenges 88
4.2.2 Prospect of new sequencing technologies 89
4.2.3 Prospect of new algorithms for SV detection 90
4.2.4 Prospect of function studies 91

References **126**

List of figures

1.2.1 Cytogenetic detection of structural variations 5
1.2.2 Idetification of CNVs with microarray 7
1.2.3 DNA amplification methods for NGS 9
1.3.1 SV detection with the RD and the SR signal 13
1.3.2 Breakpoint resolution of the RD algorithm 14
1.3.3 SV detection with the RP signal 16
1.3.4 Breakpoint resolution of the RP algorithm 17
1.3.5 Fragment length distribution 18
1.3.6 Breakpoint resolution of the SR algorithm 20

2.2.1 Workflow of the CNV detection method 25
2.2.2 Data characteristics for the 1000GP exon sequencing datasets 29
2.2.3 Detected CNV events 36
2.2.4 Exon length distribution 40
2.2.5 CNV detection efficiency 41
2.4.1 Simulation for the over-dispersion factor 50
2.4.2 Analysis of genes that failed simple linear fit 54
2.4.3 Simulated Common CNVs 55

3.2.1 Breakpoint resolution 64
3.2.2 Allele frequency spectrum of AFR population 69
3.2.3 ALU L1 recognition motif 70
3.4.1 MEI detection with RP and SR signal 76

List of tables

1.1.1 CNV and human diseases 3

1.2.1 Sequencing cost 9

2.2.1 Properties of datasets from different sequencing centers 27

2.2.2 Data characterized by sequencing center and population 28

2.2.3 Gene duplication calls in the SC dataset 31

2.2.4 Gene deletion calls in the BI dataset 32

2.2.5 Gene deletion calls in the SC dataset 35

2.2.6 Gene deletion calls in the BCM dataset 35

2.2.7 Validation results 37

2.2.8 Prior probabilities 42

3.2.1 MEI detection sensitivity on simulated data 62

3.2.2 Genotype accuracy of MEI detection on simulated data 63

3.2.3 Sequnce coverage of CEU Trio 65

3.2.4 Sensitivity and genotype accuracy on deep sequencing data 66

3.2.5 Contingency table of genotyp accuracy 67

3.2.6 Genomic distribution of MEI events in AFR population 68

3.2.7 Samples for validation 71

3.2.8 Validation result for Tangram 72

3.2.9 Comparison of validation results 72

I dedicate my dissertation to my family and friends.

Acknowledgments

I am deeply indebted to my mentor, Dr. Gabor T. Marth, for his critcal mind, constant support and invaluable guidance during my PhD study. After joining his lab, he guided me to the palace of bioinformatics step by step with his intellegence, enthusiasm to the scientific research and influence in the community. I would never have finished my doctoral studies successfully without his enormous help.

I am eternally grateful to my families, especially my wife, Xining Yan, who has been supportive, loving and patient during my long term studies for the past five years. I would be totally lost without her. Although I was thousands of miles away from home, my parents, Xiuling Wang and Bingshu Wu, still continously provided me their understanding, care and encouragement. I am forever appreciated for their love.

I greatly thank all my colleagues in my lab for their cooperations and insightful suggestions. Their creative work greatly faciliated my researches in the computational biology.

I would like to express my special thanks to Dr. Chip Stewart for his patience, willingness to guide me and rich knowledge about mathematics and statistics. My discussion with him concerning data analysis and statistical modeling significantly accelerate my understanding in bioinformatics.

Finally I would like to acknowlege all the people who helped me during the slow development of my research in Boston College. I could not be more grateful for all your support.

The greatest Oaks have been little Acorns.

Geoffrey Chaucer

1 Introduction

THE ENORMOUS DIVERSITIES in the human population can be explained by a difference of only 0.1% in the genomic sequence between any two individuals [1, 2]. Thus, identification and characterization of these genetic variants is a crucial step in understanding the link between the genomic information and phenotype. Genetic variations between human genomes could range from a single nucleotide up to several million base pairs. Despite this large size range, in the last 15 years of the 20th century the study of these variants was limited to large-scale events that can be observed under the microscope, such as aneuploidies [3–5], rearrangements [6–8], heteromorphisms [9–11], chromosomal fragile sites [12], and single nucleotide polymorphisms (SNPs) that can be detected using traditional

PCR-based DNA sequencing methods [13]. Typically, variants under 50 bp are considered to be short polymorphisms, including short insertion and deletion events. Those variants with sizes ranging from 50 bp to millions of base pairs are typically termed as structural variations (SVs). Due to the limitations of available technologies, these variants were not deeply studied until the emergence and popularization of array-based comparative genome hybridization (array-CGH) and next-generation sequencing (NGS) technologies. With these higher resolution technologies, the whole-genome SV detection at the population scale became practicable. In the last ten years, various types of SVs, including copy number variations (CNVs, such as deletions and duplications) that alter the net amount of DNA and copy neutral variations (such as inversions and translocations) that do not alter the net amount of DNA, have been discovered at a rapid rate. By the end of June 2013, 2,888,526 CNVs and 3,380 inversions have been reported to the Database of Genomic Variants (DGV) [14]. Recent large international genome study projects, *e.g.* the 1000 Genomes Projects [15] and International Cancer Genome Consortium [16], have started to generate the map of almost all types of SVs at the single nucleotide resolution, which further accelerates the SV research. This map will set a solid stage for understanding the relationship between genetic variants and phenotypic diversities and many common and rare human diseases.

1.1 Structural variation in human diseases and phenotypes

Like SNPs, SVs are ubiquitous in the human genome and are a major source of genomic and phenotypic diversities [17]. Recent studies suggest an unexpected result that SVs actually affect more heritable DNA sequences than SNPs between individuals (0.1% for SNPs and 0.5% – 1% for SVs) [18, 19]. Also the rate of novel SVs formed at a specific genomic location is relatively high. A new locus-specific SV (*de novo* variant introduced at the same genomic location among individuals) may occur in every 7,000 newborns [20], which is at least 1,000 to 10,000 times more frequent than locus-specific SNPs [21]. Although most SVs have a neutral phenotypic effect, mounting evidences show that some SVs play an important role in many phenotypic traits

Table 1.1.1: The phenotypic impact of copy number variation (CNV) in human genome. The copy number change of genes may lead to various types of genetic disorders. CNVs have been associated with many human diseases [30].

Affected gene	Copy number change	Phenotype
GSTT1	Deletion	Halothane/epoxide sensitivity
GSTM1	Deletion	Toxin resistance, cancer susceptibility
CYP2D6	Amplification	Antidepressant sensitivity
CYP21A2	Amplification	Congenital andrenal hyperplasia
OPN1LW, OPN1MW	Deletion	X-linked color blindness
LPA	Deletion	Coronary heart disease risk
RHD	Deletion	Rhesus blood group sensitivity
C4A/C4B	Deletion	Systemic lupus erythematosus
DEFB4,103	Deletion	Crohn's disease, IBD
DEFB4,103	Amplification	Psoriasis
CCL3L1	Deletion	HIV susceptibility
FCGR3B	Deletion	SLE and glomerulonephritis
IRGM	Deletion	Crohn's disease
GPRC5B	Upstream Deletion	Obesity
C4	Amplification	Lupus
SMN2	Amplification	Severity of spinal muscular atrophy
AZF region	Deletion	Spermatogenetic failure
UGT2B17	Deletion	Graft-versus-host disease
NEGR1	Upstream deletion	Obesity
NBPF23	Deletion	Neuroblastoma
TSPAN8	Amplification	Type 2 diabetes
HLA	Multiple CNVs	Crohn's disease, reheumatoid arthritis
LCE3B, LCE3C	Deletion	Psoriasis
CRIPAK	Deletion	Breast cancer

and genetic disorders, such as Mendelian disease [22, 23], sporadic chromosomal microdeletion syndrome [21], autism [24, 25], schizophrenia [26] and different types of cancers [27–29]. Table 1.1.1 summarizes some human diseases that are correlated with SVs [30].

In general, SVs can affect the phenotype through two well-recognized mechanisms: dosage effect [31, 32] and position effect [33]. Deletion and duplication (CNV) of genes and regulatory elements may cause significant dosage changes in the expression level (mRNAs) and the translation level (proteins). If affected genes are dosage-sensitive, these rearrangements can cause genetic abnormalities. Results from many studies carried out in model organisms like mice [34–36] and transformed human cells [37, 38] have demonstrated the close relationship between gene copy numbers and their expression levels. The position effect mechanism is dominated by duplications and translocations. These rearrangements can affect the causative gene even from a long distance (~1 Mbp). For example, a ~2 Mbp duplication has been found in the regulatory region upstream of the SOX9 gene to be associated with brachydactyly-anonychia disease [39]. Also, in the study of chronic myelogenous leukemia (CML), a recurrent translocation between chromosome 9 and 22 has been reported. This rearrangement forms a fusion gene between BCR and ABL genes that has been implicated in the development of this type of cancer [40]. A number of other mechanisms linking copy number changes with diseases have also been proposed, including the coding sequence disruption [41] and unmasking of recessive mutations [42].

1.2 Technologies for SV detection

1.2.1 Cytogenetic methods

As previously mentioned, the SV detection is generally limited by the development of technologies. Back in 1920s, long before the establishment of modern molecular biology and genomics, SVs could only be detected at a microscopic level (variants are so large that they can be observed under the microscope). Mega-base-pair CNVs, inversions and chromosomal rearrangements could be detected through cytogenetic methods such as chromosome banding (Figure 1.2.1A), spectral karyotyping (SKY) (Figure 1.2.1B) and fluorescent in situ hybridization (FISH) (Figure 1.2.1C, D, E, F, G and H) [17]. These large-scale genome abnormalities and heteromorphisms are usually associated with severe genetic diseases like Down and Turner syndrome [43]. However, these types of SVs are rarely implicated in common complex diseases and non-disease traits.

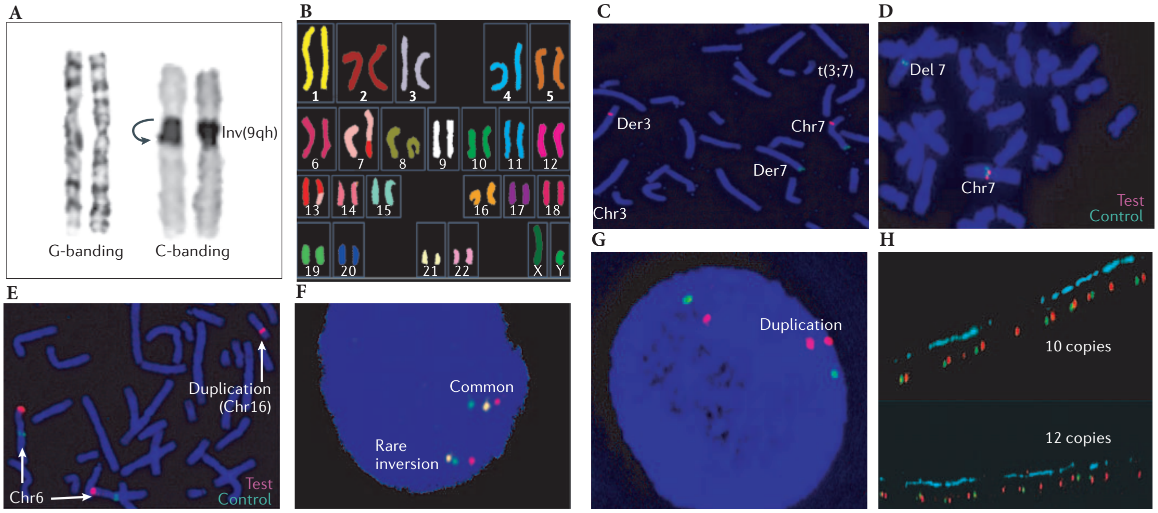


Figure 1.2.1: Structure variation detection with cytogenetic technology. **A.** An inversion event detected using the centromere (C)-banding method. **B.** A translocation event between chromosome 7 and 13 detected using the spectral karyotyping (SKY) method. **C.** a translocation between chromosome 3 and 7 detected using fluorescence in situ hybridization (FISH) carried out using metaphase chromosomes. **D, E.** copy number decrease and increase events detected using the FISH method. In panel **D**, two copies of control probes (green) in chromosome 7 have been detected whereas the test probe (red) only presents on one of the homologous chromosome 7. In panel **E**, an amplification signal is observed on chromosome 16 in additional to the signal of two copies on chromosome 6. **F.** A micro inversion event of length 700kbp detected using a two-color FISH method. Reprinted from [44] with permission. **G.** Two-color FISH has revealed a large genomic rearrangement (duplication). **H.** Copy number differences can be detected with FISH. Reprinted from [14] with permission.

1.2.2 Microarray

The first wave of systematic studies of SVs at the whole-genome level began in the late 1990s and early 2000s when the full assembly of the human genome [45] and microarray technologies (aCGH) [46, 47] became available. Figure 1.2.2 is a flow chart that demonstrates how a microarray is used to detect SVs. The sample and reference DNA are first fragmented and then labeled with different fluorescent dyes, for example Cy5 and Cy3. Both sample and reference DNA are then treated with COT-1 DNA that is primarily composed of repetitive sequences, to block genomic regions with repeats. These DNA sequences are then hybridized to arrays that are covered with oligonucleotides (60 – 100bp) derived from the reference genome. Finally, SVs (deletions and duplications) can be detected by measuring the ratio of fluorescent signal between the sample and reference DNA. To reduce the noise and false positive detection rate, array-CGH usually includes an assay format called "dye-swap". In this format, an extra hybridization is carried out with sample and reference DNA swapping their fluorescent tag (say sample-Cy5 and reference-Cy3 for the first hybridization and reference-Cy5 and sample-Cy3 for the second hybridization). The ratio will be measured twice. These two ratios are almost the reciprocal of each other for real events. Any spurious calls can be excluded if only one ratio is off from the neutral ratio (1.0), which might be caused by the random fluctuation of the fluorescent signal instead of a real CNV event.

The strength of this technology is its effectiveness of both cost and time. In 2000, the whole-genome shotgun sequencing (Sanger sequencing [48]) was already being used in the Human Genome Project. However, it is prohibitively expensive for the routine SV detection at the population scale. Compared with the first generation sequencing technology, microarrays are vastly cheaper. Additionally, microarrays are very high-throughput: hundreds of thousands of genomic regions can be probed for SV detections simultaneously on a single array, making it an ideal method for large-scale projects. Microarrays can also be used to detect submicroscopic SV events. In fact, a resolution on the order of tens kbp [17] can be achieved. This would be

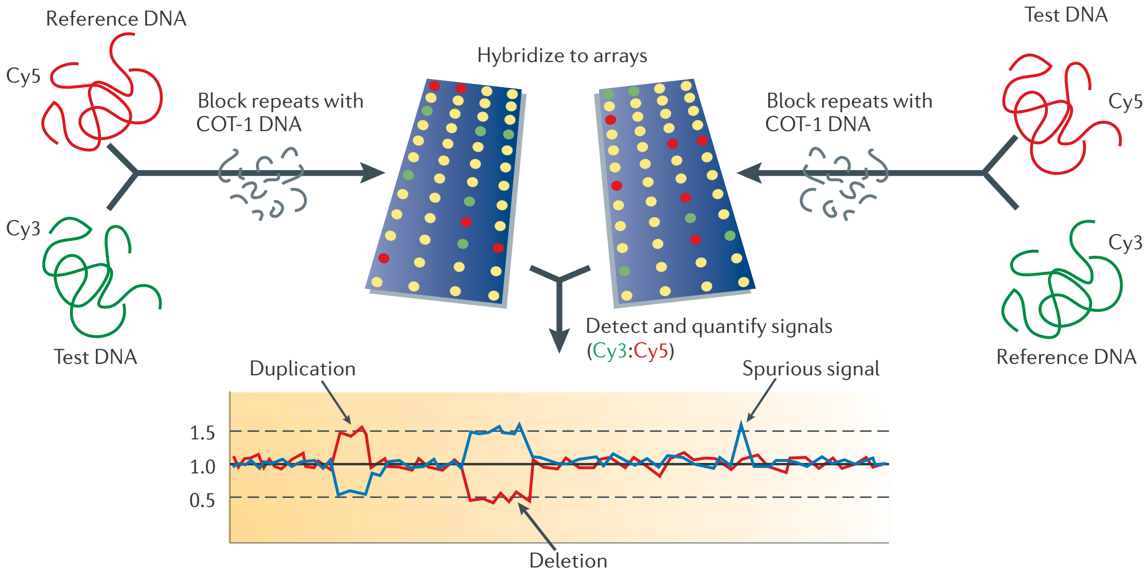


Figure 1.2.2: Array based, genome-wide methods for SV detection. Test and reference DNA sequences are fragmented, labeled with different fluorescent tags and hybridized to arrays covered with oligonucleotide probes derived from the reference sequence. The copy number variants can be detected as those regions in which the ratio between sample and reference data deviates significantly from 1.0. To reduce noise, the sample and reference DNA sequences have their fluorescent tags swapped for an extra round of measurement. Reprinted from [17] with permission.

practically impossible to observe using cytogenetic methods.

Although the development of microarrays was a significant advance in SV detection technology, it is not without its limitations. First of all, aCGH can be only used for detecting CNVs (deletions and duplications), not copy neutral variations (inversions and translocations) since it only measures the copy number difference between the sample and reference DNA. Secondly, although microarray is a much improved genome-scanning technology, the fluorescent signal is typically very noisy. The signal can be affected by many factors such as the base composition, the proportion of repetitive sequences and the amount of "hybridizable" DNA in the array element. The fluorescent intensities can fluctuate by a factor of 30 even if there are no CNVs [49]. Because of these reasons, microarray data normally require a sophisticated computational process to decode. This limits the sensitivity and breakpoint resolution to smaller (under 1 kbp) SVs of algorithms designed for microarray data. Last but not least, microarray based methods are intentionally designed to avoid genomic regions embedded in repeat sequences, making it insensitive to breakpoints located in repetitive elements, which compose 66 – 69% of the human genome [50].

1.2.3 Next-generation sequencing

The recent success in building up high-resolution SV map within human populations is largely attributable to the rapid development of the high throughput NGS technology. The NGS technology was first introduced by Roche company with its 454 sequencing machine in 2005 [51]. Soon, many other companies like Illumina [52], Applied Biosystem (ABI) [53] and Complete Genomics [54] joined this market with their own NGS technologies. The widespread adoption of these sequencing technologies greatly facilitated the discovery of SVs. The number of reported SVs grew dramatically since the late 2000s. Compared to the first generation sequencing technology, Sanger sequencing, NGS technology replaces the time-consuming bacterial cloning with much more efficient PCR techniques to amplify DNA samples (Figure 1.2.3), which significantly reduces the sequencing cost (Table 1.2.1 [54, 55]). The length of output reads (25 – 100 bp) from NGS machines is usually shorter than that of the Sanger sequencing (~1 kbp). However, NGS is able to generate much more data per run: the latest Illumina HiSeq sequencing machine can produce up to 200 Gb high quality reads per run in about eight days whereas the most recent Sanger capillary machine introduced in 1999 can only produce 1.6 Mb data per run. Also, most current NGS technologies apply the paired-end sequencing technique to increase the effective sequencing length. DNA samples are digested into long fragments with a length ranging from several hundred base pairs to thousands base pairs, depending on the sequencing technology and the final read length. Then sequencing machines read the nucleotides from both sides of these fragments and leave an unsequenced region in the middle. The width of the distribution of these fragments, or inserts, is usually very tight. The fragment length of most sequencing reads is within a very narrow region. So the mapping distance of a given pair of reads from this technology can be easily estimated from this distribution if there are not any SVs occurring in the

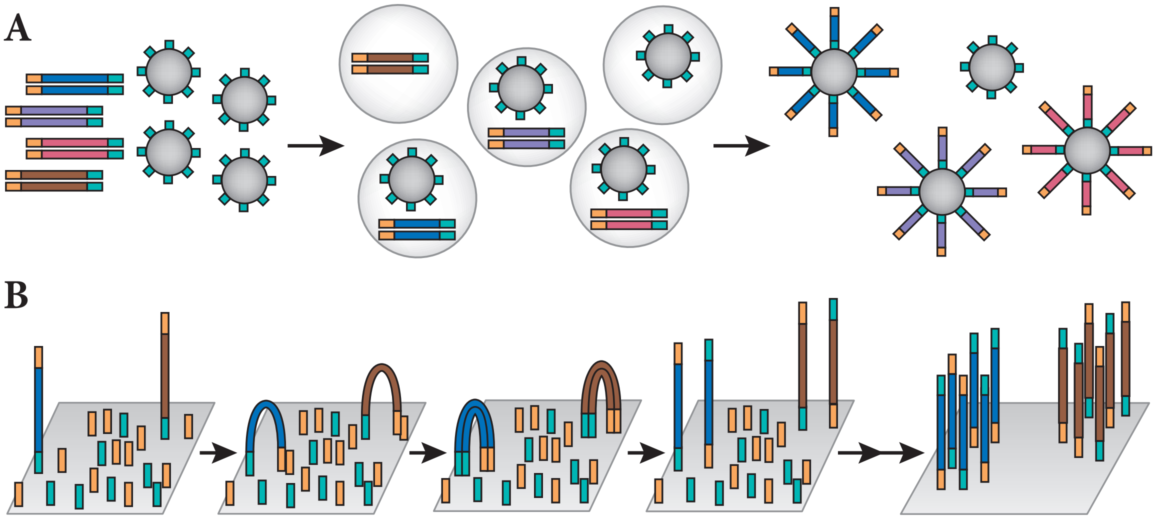


Figure 1.2.3: DNA amplification methods used in next-generation sequencing technologies. **A.** Emulsion PCR. This method is mainly used in 454 and Solid sequencing machines. DNA fragments with adapters (gold and turquoise) are PCR amplified within a water-in-oil emulsion. **B.** Bridge PCR. Illumina invents this technique. One end of the DNA fragments for amplification is first ligated to adaptors that attached to a membrane. The other end of these fragments is then flanked with another adapter. The bridge-shape fragment will then be amplified iteratively as shown in the figure. Reprinted from [56] with permission.

Table 1.2.1: Approximate cost of generating reads with 1× coverage of human genome by using different sequencing technologies [54, 55].

Technology	**Cost per** 1×
Sanger capillary	$1.4M
Roche 454	$93k
Illumina	$123
ABI SOLiD	$8k
Complete Genomics	$110
PacBio	$6k
Ion Torrent	$3k

unsequenced region. This constraint provides valuable information to detect SVs with NGS data.

By utilizing the NGS technology, researchers can now identify a certain types of SVs in the whole-genome and population scale, like deletions and duplications, at the single nucleotide resolution with high accuracy. However, due to the limitations of the NGS technology, especially the read length, and biological complexities of the human genome, some other types of SVs, *e.g.* inversions (usually buried in repetitive regions) and mobile element insertions (MEIs, inserted elements themselves are repetitive sequences), are hard to detect. The detection of full-spectrum SV types will require further advances in the sequencing technology (with read length at tens of kbp) and the development of more sophisticated algorithms.

1.3 Algorithms for SV detection with NGS data

NGS data opened many possibilities for bioinformaticians to develop different types of computational methods to comprehensively identify and characterize SVs in human genomes. To handle the huge amount of data generated from NGS machines, many efficient algorithms that take advantage of different aspects of sequencing data have been proposed. Most of these approaches are based on the resequencing strategy — sequencing reads have to be first mapped to the reference genome with aligners, such as MOSAIK [57], BWA [58] and BFAST [59], and then the SVs can be detected as the differences between alignment reads and the human genome reference, the major achievement of the Human Genome Project. Due to the limitations of current sequencing technologies (short read length and fragment length) and biological features of the human genome (full of repetitive elements), many reads cannot be aligned uniquely to the reference genome. Reads that can be mapped to multiple positions are usually assigned only to a random location by most of sequencing alignment programs with a low mapping quality (0, in most cases) that is dominantly affected by the number of locations a read can be aligned in addition to some other factors such as the number of mismatches in the alignment and base qualities of the sequencing read and excluded from the analysis by most of SV detection programs for the sake of lower false discovery rate (FDR). However, in order to detect some complicated types of SV, such as MEI, these ambiguous reads have to be taken into account with special handling at both the primary aligning level and SV detection level — the aligner must provide the extra information about these reads, such as the type of repetitive elements where these reads are sampled, for the downstream analysis.

This section will review three most frequently applied algorithms in current available SV detectors for NGS data: read depth (RD), read pair (RP) and split read (SR).

1.3.1 Read depth algorithm

The depth of coverage is one of the well-known statistics to describe the quality of NGS alignment data — usually the higher the coverage the better the data. In most cases, the depth of coverage refers to the base coverage: the number of reads that contain a certain nucleotide in the reference sequence: $c = \frac{NL}{G}$, where c is the base coverage, N is the number of sequencing reads, L is the average length of reads and G is the length of the reference sequence. There is another expression of the depth of coverage that is often used in the CNV detection: read depth, the number of alignments (DNA fragments) that fall into a given size of window at a particular genome location. By analyzing the read depth (RD) signal with NGS alignment data, CNVs can be detected with the similar computational method that is applied on microarray data. Instead of measuring the difference of the fluorescent intensity between the sample and reference DNA, the RD method measures the difference between the observed read depth and the expected or control read depth. For example, the observed read depth at a given genome region should be about half of the expected or control read depth if the genomic region harbors a heterozygous deletion or about zero if the genomic region harbors a homozygous deletion (Figure 1.3.1). In this method, the whole genome region is first segmented into numerous non-overlap windows with fixed size around 50 bp – 100 bp (depending on the quality of the data) and then the algorithm will count how many alignments (the start of each alignment) are within each of these windows. Each of these counts is the observed read depth. To detect CNVs, it is also necessary to estimate the number of read counts in the same window if there is no CNV at all (null hypothesis). One efficient way of estimating expected read counts is to generate the same amount of simulated reads with the same read length as the real sequencing data from the reference genome with a practical error model (sequencing error) similar to the sequencing technology used for generating the real data. Many toolboxes, such as WgSim [60] and MASON [61], can be used for this task. The simulated reads will be aligned with the same aligner and the same parameters as the real sequencing data and the count of simulated alignments at the corresponding window will be served as the expected read depth. In cancer sequencing data, there is usually no need to generate simulated data since the number of alignments from the normal tissue in the same patient can be served as the control read depth. If we assume that sequencing reads are sampled uniformly from the genome, the number of observed read depth at a given window should follow the *Poisson* distribution with the median of $\mathrm{RD}_{\mathrm{expected}}$ and the standard deviation of $\sqrt{\mathrm{RD}_{\mathrm{expected}}}$. The candidate CNV events then can be detected with a pre-defined p-value threshold. In practice, detectors using the RD signal usually call a CNV event only if at least two or three consecutive windows all have the significant difference between the observed read depth and the expected read depth for specificity consideration.

The advantage of this algorithm is that it is computationally lightweight since only the alignment position of each read is used for calculation. After calculating the read count for each window, the rest of computational work can be easily performed even with a personal computer. Moreover, the RD algorithm can be applied to both whole-genome sequencing data and the capture sequencing data where sequencing reads are only from selected genomics regions, such as exons. The major problem of CNV detection in capture sequencing is that breakpoints may not be included in sequencing regions, which is a requirement for RP and SR algorithms. Since the RD algorithm only measures the read depth change breakpoint positions being outside the sequencing region does not affect the detection of CNVs.

Like microarray technology, the major limitation of the RD algorithm is the relatively low breakpoint resolution (approximately several hundred base pairs, Figure 1.3.2) and sensitivity to smaller events. Although NGS is technologically better than aCGH, the RD signal is still

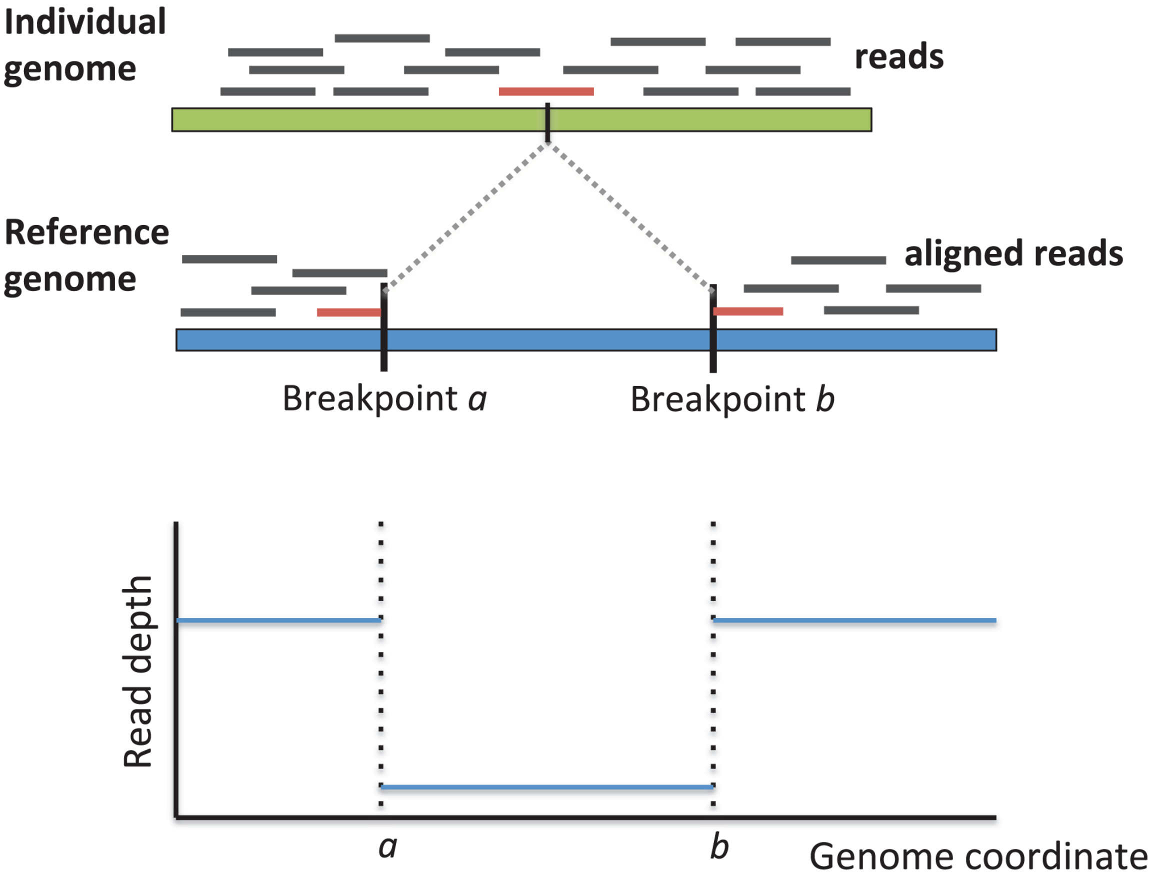


Figure 1.3.1: Detection of a homozygous deletion event with split read (red read in the middle) and read depth (bottom panel) signal. Reprinted from [62] with permission.

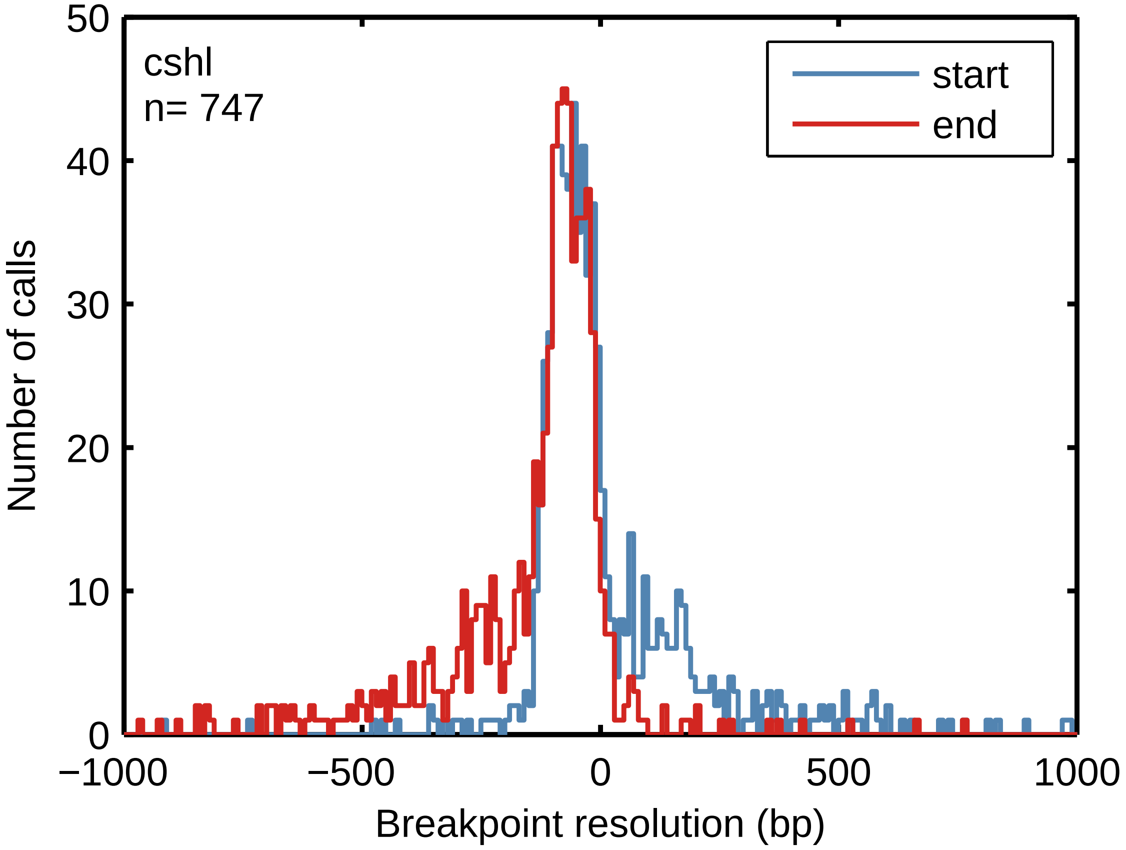


Figure 1.3.2: Breakpoint resolution (blue for start position and red for end position) of deletion events detected by the read depth method with WGS data in the 1000 Genomes Project Pilot studies [63].

sometimes too noisy to precisely locate CNVs and sensitively detect those small events. Moreover, the RD algorithm is totally blind to copy neutral variations like inversions and balanced translocations since it only measures read count changes.

1.3.2 Read pair algorithm

The RP algorithm takes advantage of a special feature of the NGS technology, paired-end mapping. In the protocol of current available sequencing technologies, the input sample DNA is usually sheared into small fragments, ranging from several hundred to several thousand base pairs. The sequencing machine will read nucleotides from both ends of each fragment and leave an unsequenced region in the middle. The length of fragments from the same batch of sequencing jobs should be within a very narrow range. If the unsequenced region of the fragment does not harbor any SVs then the mapping distance between the two mates of a read pair should be slightly deviated from the expected fragment length. These pairs are called concordant pairs. If there is an SV between the two mates then the mapping distance of them should be much different from the expected fragment. For example, if a read pair span a deletion breakpoint, the mapping length of this read pair should be significantly larger than its fragment length due to the absence of the deleted region in the sample DNA and the existence of it in the reference genome. These pairs are called discordant pairs (Figure 1.3.3 left panel). So the first step in the RP algorithm is to calculate the fragment length distribution from those read pairs with high mapping qualities (both mates are uniquely aligned with few mismatches, Figure 1.3.3 right panel). SV candidates then can be identified as those read pairs on both edges of the fragment length distribution with a pre-defined p-value cutoff. These read pairs will then be clustered with a particular clustering algorithm to increase the detection specificity. Most SV detectors equipped with the RP algorithm required a minimum number of candidate fragments in a cluster to make an event call to reduce the possibility of false detections. Since the exact length of DNA fragments input into the sequencing machine is unknown, the breakpoint position and the length of the detected event can be only estimated approximately from the mapping positions of alignment reads in the cluster and the fragment length distribution. For example, a cluster with two discordant reads identifies a deletion event. The read length of these two pairs is fixed: 50 bp. The mapping start and end positions of the first mates in the first pair are: 1000 bp and 1049 bp. The mapping start and end positions of the second mate are: 2000 bp and 2049 bp. The corresponding mapping start and end position of the two mates in the second pair are: 1100 bp, 1149 bp, 2050 bp and 2099 bp. The median fragment length of this sequencing library is 500 bp. It represents the expected fragment length without any SV events. Based on the information given above, we can estimate the breakpoint position of this deletion event to be at 1149 bp, the rightmost position of the mapping end position of the first mate in these two pairs, and the event length to be 525bp, the average difference between mapping distances of these two pairs and the median fragment length,

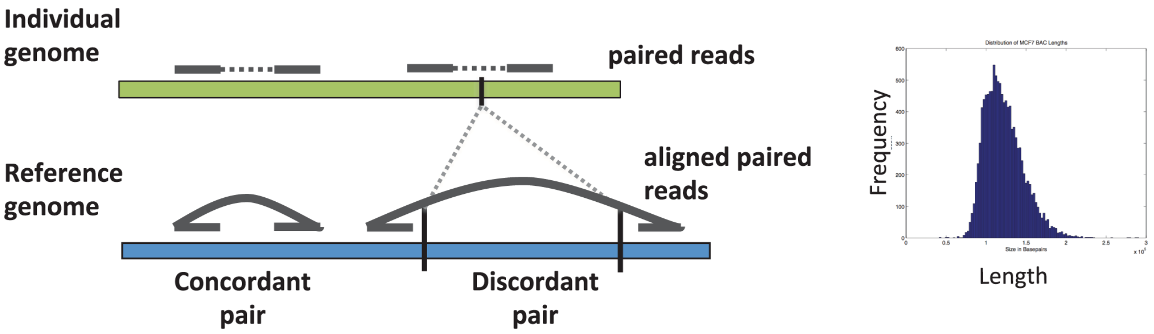


Figure 1.3.3: Illustrations of concordant, discordant pairs (left panel) and fragment length distribution (right panel). The discordant pairs can be identified as those alignments whose mapping distance between two mates does not agree with the fragment length distribution or mapping orientation does not agree with the expected read orientation. The left panel demonstrates an instance of using discordant pair to detect a deletion event. Reprinted from [62] with permission.

(1050bp – 500bp + 1000bp – 500bp) / 2. Besides the fragment length, the orientation of a read pair can also provide useful information for SVs detection. For a given sequencing technology, the orientation of two mates in a read pair should follow a predictable pattern if they are sampled from a genomic region without any SVs. For example, the orientation pattern of read pairs from Illumina sequencing machines is that the mate with smaller genomic position should be on the positive strand and the mate with larger genomic position should be on the minus strand. If one mate of a read pair hits a inversion then its orientation will be different from the expected orientation. Inversion events can be identified through grouping these mis-oriented read pairs.

The advantage of the RP algorithm is that it provides much higher breakpoint resolution. The uncertainty of reported events by the RP algorithm is usually around 50 – 100 bp (Figure 1.3.4), depending on the coverage and the shape of the fragment length distribution. Also, the RP signal is generally very strong and clear. It usually requires a few RP supporting fragments to identify a SV event. So the RP method can be applied to low coverage data (~5×). Moreover, the RP algorithm cannot only identify CNV events such as deletions and duplications but also can detect copy neutral variation like inversions and translocations. Almost all types of SVs have their corresponding RP signatures. For example, the RP signature for deletions is that the mapping length of a read pair is larger than the expected fragment length; the RP signature for insertions is

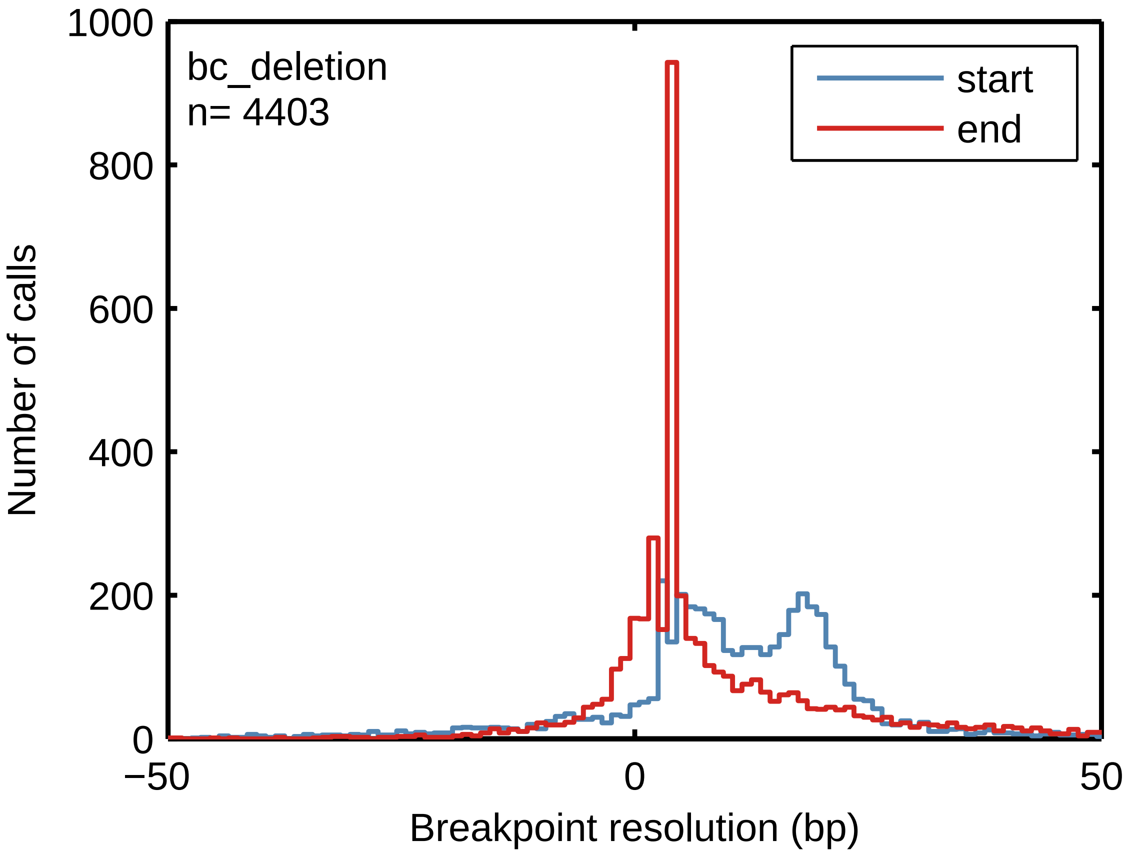
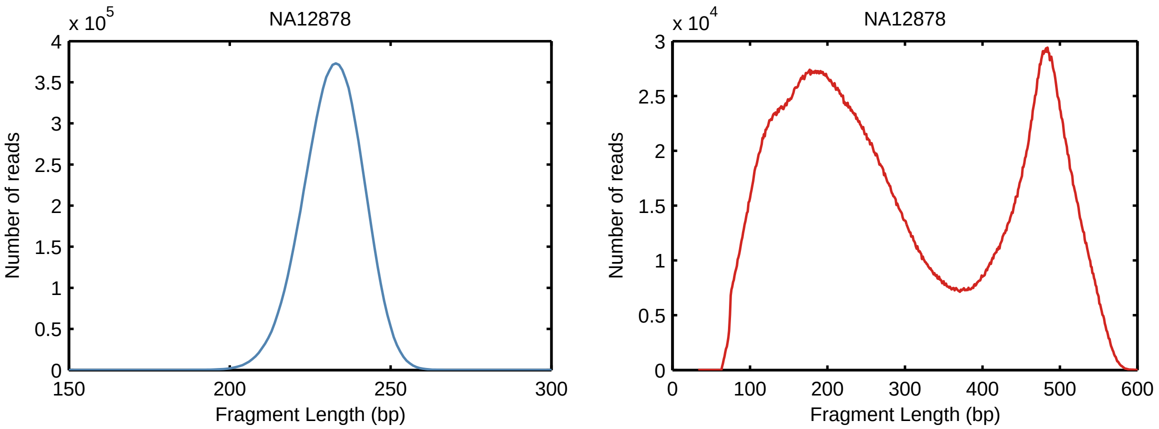


Figure 1.3.4: Breakpoint resolution (blue for start position and red for end position) of deletion events detected by the read pair method with WGS data in the 1000 Genomes Project Pilot studies [63].

that the mapping length of a read pair is smaller than the expected fragment length; the RP signature for inversions is that the mapping orientation of a read pair is discordant with the expected orientation (the mapping length might be discordant too); the RP signature for translocations is that two mates of a read pair will be aligned to two different chromosomes.

The major limitation of this approach is that the detection sensitivity to SV events highly depends on the quality of the fragment length distribution. If the fragment length distribution of sequencing data is in regular shape (bell-shaped) and tight (Figure 1.3.5A), the RP algorithm can achieve high detection efficiency. However, if the fragment length distribution is wide and in irregular shape(Figure 1.3.5B), it might limit the sensitivity of the RP algorithm. Also, although the RP algorithm can provide much better breakpoint resolution than the RD algorithm, it still



A: Tight fragment length distribution with well-formed bell shape — suitable for the RP algorithm.

B: Wide fragment length distribution with irregular shape — may cause lower sensitivity with the RP detection.

Figure 1.3.5: Two different fragment length distributions from two different sequencing libraries of a 1000GP sample (WGS), NA12878.

can only provide the approximate position of a reported SV instead of the exact breakpoint location.

1.3.3 Split read algorithm

The SR algorithm is the latest player for SV discovery. The first SV detection program based on the SR algorithm, Pindel [64], is not published until 2009. Before that, few detectors take those unaligned reads and soft clipped reads (only part of these reads can be aligned to the reference genome) into account for SV detection since they are hard to handle. These reads are usually sampled from genome regions that cross SV breakpoints. The basic idea of the SR algorithm is to split these unaligned and soft clipped reads into several partial reads so that they can be aligned separately to different genome positions, before breakpoints, within SVs and/or after breakpoints (Figure 1.3.1). For example, one mate of a read pair with 100 bp length crosses a deletion (500 bp) breakpoint in the middle. This mate is actually a fusion read with the first 50 bp before the deletion region and the second 50 bp after the deletion region. This read usually cannot be aligned back to the reference genome or it will be aligned with 50 bp soft clipped (either the first or the second 50 bp). With the SR algorithm, the first 50 bp partial alignment can be found by searching a local region, about 2 times of the median fragment length, around the other anchor mate (usually required to be aligned uniquely to the genome). The second 50 bp partial alignment can be then found in a region after the mapping end position of the first partial alignment. For running time consideration, the size of the search region for the second partial alignment is usually limited to several kbp to 1 mbp since large-size SVs are generally very rare. After both partial alignments are found the position, the length and type of the detected variation can be determined. To avoid high FDR, most SV detectors based on the SR algorithm require at least two SR alignments for a given call.

The advantage of the SR algorithm is that it can locate SV at the single nucleotide resolution (Figure 1.3.6), which is a huge improvement from RD and RP algorithms. The mapping position and orientation of partial alignments can provide the precise information about the location, length and type of reported SVs. Like the RP algorithm, the SR method can detect almost all types of simple SVs as well as some complex events.

Although powerful, the SR algorithm requires an additional mapping effort after the primary alignment. Depending on the size of the search region for the second partial alignment, the length of sequencing reads and the base coverage of the alignment, the split mapping step may become time-consuming. Moreover, a long read length (>50 bp) is usually required for reliable split mapping results.

Three algorithms utilize reads sampled from three different regions associated with SVs: candidates for the RD algorithm are those reads inside SV events; candidates for the RP algorithm are those reads whose two mates span SV breakpoints; and candidates for the SR algorithm are those reads that one mate is uniquely aligned to the normal reference region and the other mate hits the breakpoint of a SV event. These three sources of signal for SV detection are generally independent of each other. Several recently published SV detectors, such as DELLY [65] and Tangram (described in Chapter 3), utilize two or more algorithms together for higher detection efficiency and specificity. As the read length becomes longer, we could anticipate that toolboxes

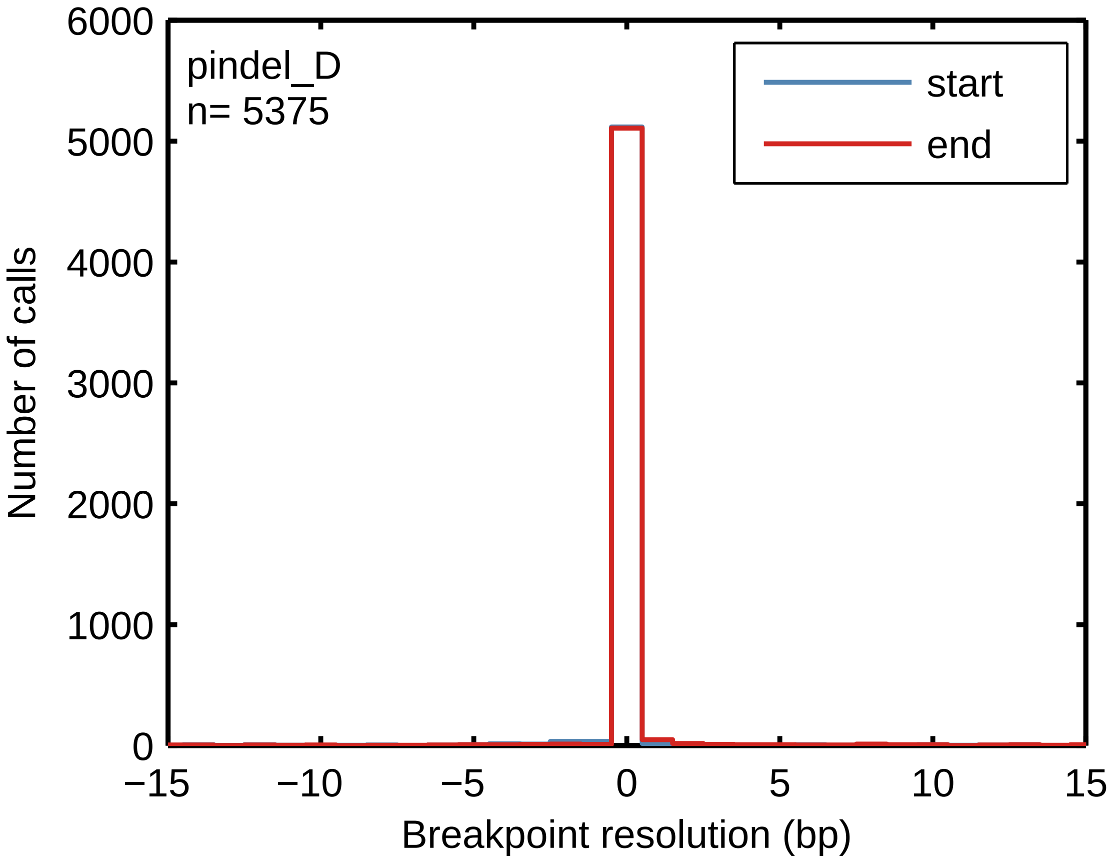


Figure 1.3.6: Breakpoint resolution (blue for start position and red for end position) of deletion events detected by the split read method with WGS data in the 1000 Genomes Project Pilot studies [63].

that integrate *de novo* or local assembly algorithms will soon become available in the near future.

If you would be a real seeker after truth, it is necessary that at least once in your life you doubt, as far as possible, all things.

Rene Descartes

2

CNV detection from exon capture sequencing data

DNA capture technologies combined with high-throughput sequencing now enable cost-effective, deep-coverage, targeted sequencing of complete exomes. This is well suited for SNP discovery and genotyping. However, there has been little attention devoted to Copy Number Variation (CNV) detection from exome capture datasets despite the potentially high impact of CNVs in exonic regions on protein function.

As members of the 1000 Genomes Project analysis effort, we investigated 697 samples in which 931 genes were targeted and sampled with 454 or Illumina paired-end sequencing. We developed a rigorous Bayesian method to detect CNVs in the genes, based on read depth within target regions. Despite substantial variability in read coverage across samples and targeted exons, we were able to identify 107 heterozygous deletions in the dataset. The experimentally determined false discovery rate (FDR) of the cleanest dataset from the Wellcome Trust Sanger Institute is 12.5%. We were able to substantially improve the FDR in a subset of gene deletion candidates that were adjacent to another gene deletion call (17 calls with 0% FDR). From the simulation experiment and our calculation, the estimated sensitivity of our call-set was 45%.

This study demonstrates that exonic sequencing datasets, collected both in population based and medical sequencing projects, will be a useful substrate for detecting genic CNV events, particularly deletions. Based on the number of events we found and the sensitivity of the methods in the present dataset, we estimate on average 16 genic heterozygous deletions per individual genome. Our power analysis informs ongoing and future projects about sequencing depth and uniformity of read coverage required for efficient detection.

2.1 Introduction

Copy Number Variations (CNVs) *i.e.* deletions and amplifications, are an essential part of normal human variability [66]. Specific CNV events have also been associated with various human diseases [67], including cancer [68] autism [69, 70] and schizophrenia [71]. Historically, large CNV events can be observed using FISH [14] but systematic, genome-wide discovery of CNVs started with microarray-based methods [72–74] which can detect events down to tens of kbp. As with all hybridization based approaches, these methods are blind in repetitive and low complexity regions of the genome where probes cannot be designed. High throughput sequencing with next-generation technologies have enabled CNV detection at higher resolution (*i.e.* down to smaller event size), in whole-genome shotgun datasets [63, 75, 76]. However, despite decreasing costs, deep-coverage ($\geq 25\times$) whole-genome data is still prohibitively expensive for routine sequencing of hundreds of samples, and in low-coverage (2-6× base coverage) datasets detection sensitivity and resolution is limited to long genomic events [66].

Targeted DNA capture technologies combined with high-throughput sequencing now provide a reasonable balance between coverage and sequencing cost in a substantial portion of the genome, and full-exome sequencing projects are presently collecting $\geq 25\times$ average sequence coverage in thousands of samples. CNV events in exonic regions are important because the deletions of one or both copies, or amplifications affecting exons, are likely to incur phenotypic consequences.

Current algorithms for detecting CNVs in whole-genome shotgun sequencing data use one of four types of signal as evidence for an event: (1) aberrantly mapped mate-pair reads (RP or read pair methods); (2) split-read mapping positions (SR); (3) *de novo* assembly (AS); and (4) a significant decrease or increase of mapped read depth (RD methods). Unfortunately, these methods are not generally applicable for CNV detection in capture sequence data without substantial modifications. SR, RP, and AS based methods are sensitive only to CNVs in which mapped reads or fragments span the event breakpoint(s). In the case of exon capture data, this restricts detection to CNV events where at least one breakpoint falls in a targeted exon. RD based methods suffer from large fluctuations of sequence coverage stemming from variability in probe-specific hybridization affinities across different capture targets (in this case: exons) and sets of such targets (in our case: genes), and from the over-dispersion of the read coverage distribution in the same target across different samples. Presumably because of the technical challenges, and despite the importance of deletion or amplification events within exons, there are currently no reported CNV detection algorithms for targeted DNA capture based exon-sequencing data (with the exception of methods for tumor-normal datasets [77] where the read depth measured in the normal sample can be used for normalization, which is not available in the case of population sequencing).

In this study, we set out to develop a CNV detection algorithm for capture sequencing data. This algorithm is based on RD measurement, and detects samples with non-normal copy number in the capture target regions. As participants of the 1000 Genomes Project, we took part in the data analysis of the "Exon Sequencing Pilot" dataset [15], where 12,475 exons from over 900 genes (representing about 10% of the whole exome) were targeted and sequenced with a variety of DNA capture sequencing technologies.

2.2 Results

2.2.1 Brief algorithmic overview

Our algorithm is an extended version of RD-based CNV detection that aims to mitigate the vast target-to-target (and consequently gene-to-gene) heterogeneity of read coverage by normalization procedures roughly corresponding to those employed in CNV detection methods from microarray hybridization intensity data. The overall workflow of our method is shown in Figure 2.2.1 and described in greater detail in the Methods 2.4 section. For a given gene in a given sample (we will use the abbreviation GSS: Gene-Sample Site throughout the paper), we define the read depth as the number of uniquely mapped reads whose 5' end falls within any of the targeted exons within that gene. We compare this measurement with an expected read depth (Eq. 2.2, Methods 2.4.3), based on a "gene affinity" calculated from measured read depth for that gene across all samples (to account for across-target read coverage variance due to target-specific hybridization), and the overall read depth for the sample (to account for the variance of read coverage due to the overall sequence quantity collected for the sample under examination). We then use a Bayesian scheme, calculating the posterior probability for each copy number with prior probablities estimated from previous study [18] and the data likelihood computed based on the data (See Methods 2.4.4), to determine whether the measured coverage is consistent with normal copy number (*e.g.* CN = 2 for autosomes), or aberrant copy number (*i.e.* homozygous deletion: CN = 0, heterozygous deletion: CN = 1, or amplification: CN > 2). We have included two algorithmic variants: One is suitable for CNV events that occur at a low allele frequency (*i.e.* in a small fraction of the samples), and the other for capturing higher-frequency deletion events (see Methods 2.4.8).

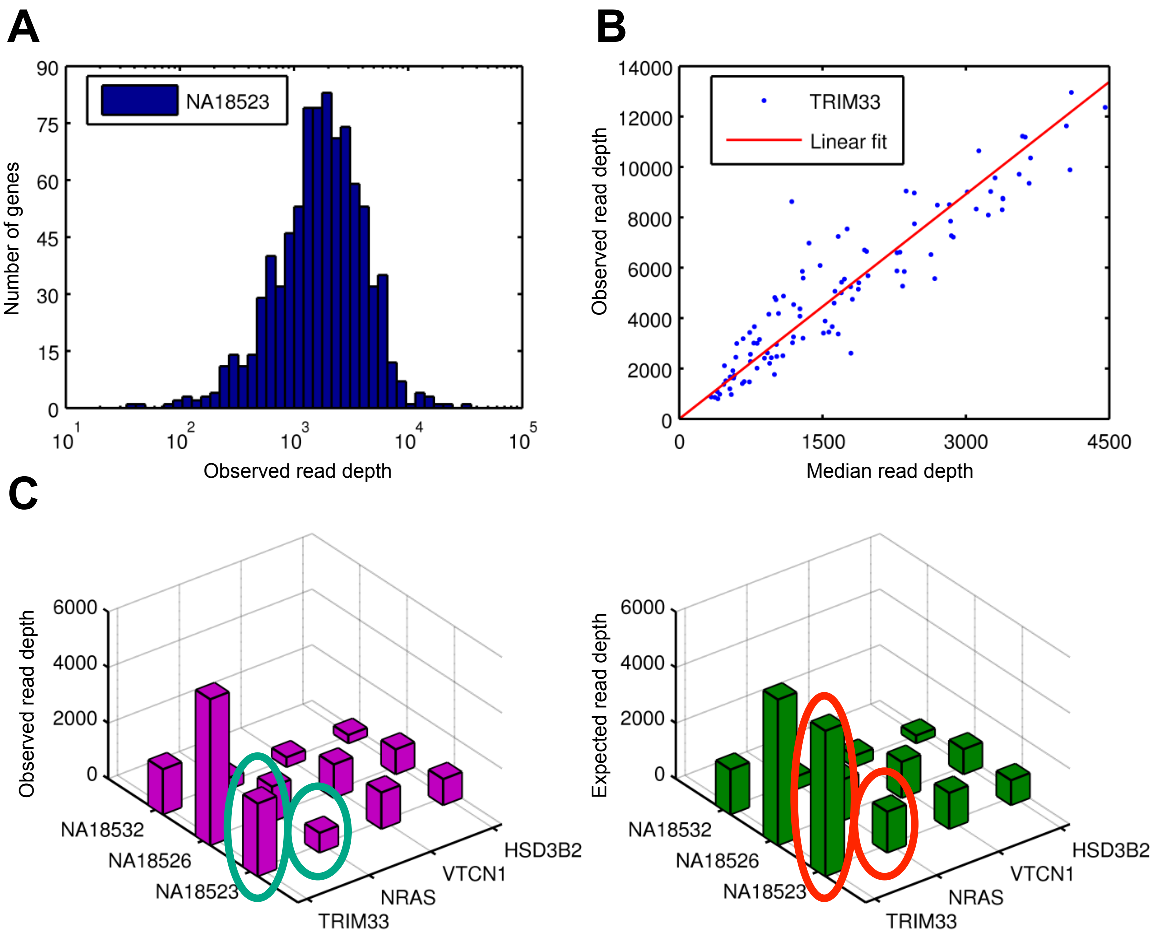


Figure 2.2.1: A. Median Read Depth (MRD) is calculated for each sample, as a measure of sample coverage (NA18523 shown). **B.** The gene affinity is estimated for each gene as the slope of the least-square-error linear fit between MRD and RD for that gene (TRIM33 shown). **C.** Example of observed (magenta) and expected (green) read depth for three samples and four genes. The observed read depths were roughly half of the expected values for genes TRIM33 and NRAS, in sample NA18523, and detected as deletions.

2.2.2 Dataset

In this study we analyzed the exon capture sequencing dataset collected by the 1000 Genomes Project Exon Sequencing Pilot, including 931 genes (about 4.6% of the protein-coding genes in the human genome) processed with Agilent liquid-phase and Nimblegen solid-phase capture methods, and sequenced from 697 individuals with Illumina paired-end and/or 454 technologies. The samples in the dataset have been sequenced by four different data collection centers (Washington University, WU; Wellcome Trust Sanger Institute, SC; Broad Institute, BI; and Baylor College of Medicine, BCM) using different pairings of capture and sequencing technologies (Table 2.2.1 and Table 2.2.2). Initially 1,000 genes were randomly selected by the Exon Piolt Project from the CCDS [78, 79] database as targeted sequences. However, the capture target designs used in the four production centers were significantly different. To eliminate the inconsistency, the Pilot Project defined a set of consensus exon target sequences by intersecting the intial designs. The consensus targets, 931 genes used in this study, has approximately 1.43 Mbp in length, covering 86.1% coding regions in the initial 1,000 genes [80]. As our method relies on an estimate of the gene-specific hybridization affinity, it requires that such affinities are consistent across all samples analyzed simultaneously. According to the principal component analysis (PCA) of the observed read depths, (Figure 2.2.2A, see Methods 2.4.1), target and genes affinities are inconsistent across data from different centers, and therefore we analyzed each dataset separately. We only considered datasets with at least 100 samples (SC, BI, BCM) so we can obtain sufficient sample statistics across genes. After filtering out genes and samples that did not meet our minimum read depth requirements (see Methods 2.4.2), we were left with the following datasets: SC (862 genes in 106 individuals sequenced with Illumina), BI (739 genes in 110 samples sequenced with Illumina), and BCM (439 genes in 349 samples sequenced with 454) (Table 2.2.1). The number of genes that passed our filters was substantially lower in the BCM dataset both due to lower overall 454 coverage (see below), and because the longer 454 reads result in lower RD (fewer reads) when compared to shorter Illumina reads, even at

Table 2.2.1: Properties of datasets from different sequencing centers

	SC	BCM	BI	WU
Total sample count	117	352	161	93
Sample count after quality control	106	349	110	82
Technology	Illumina	454	Illumina	Illumina
Duplicate rate	0.21	0.3	0.5	0.72
Mapping quality (mean)	50	33	45	51
Base coverage (mean ± standard deviation)	56 ± 34	23 ± 12	70 ± 61	29 ± 9
Read depth per gene (mean ± standard deviation)	2309 ± 3166	106 ± 171	1329 ± 2053	977 ± 1382
MRD (mean ± standard deviation)	1710 ± 1073	97 ± 52	1070 ± 803	599 ± 164
Number of exons	8174	8174	8174	8174
Exons overlapped with segmental duplication regions	458 (5.6%)	458 (5.6%)	458 (5.6%)	458 (5.6%)
Number of genes (passing QC)	862	439	739	1
Genes overlapped with segmental duplication regions	29 (3.3%)	11(2.5%)	23(3.1%)	0(0.0%)
Over-dispersion factor (mean ± standard deviation)	7.9 ± 8.2	2.1 ± 1.1	6.4 ± 5.5	N/A
Quality index (mean ± standard deviation)	9.4 ± 8.8	5.5 ± 2.3	7.6 ± 5.6	N/A
Expected detection sensitivity based on quality index	0.46	0.2	0.41	N/A
Number of calls $h = 0.65$ either with or without a neighboring call	36	4	56	N/A
Number of calls $h = 0.1$ either with a neighboring call	17	0	11	N/A

Table 2.2.2: Data characterized by sequencing center and population

	SC				
	CEU	CHB	JPT	TSI	YRI
Number of samples	18	14	9	51	14
Male/Female	9/9	5/9	5/4	24/27	2/12
Average read depth per gene	1679	1701	1597	1617	1865
Read Length	36	36	36	36	36

	BCM					
	CEU	CHB	CHD	JPT	LWK	YRI
Number of samples	40	62	78	16	108	45
Male / Female	20/20	15/47	38/40	5/11	51/57	22/23
Average read depth per gene	178	131	171	243	128	165
Read length	258	323	339	300	336	322

	BI				
	CEU	CHB	CHD	JPT	YRI
Number of samples	16	13	28	34	19
Male / Female	9/7	11/2	12/16	16/18	12/7
Average read depth per gene	1623	1631	1675	1104	1612
Read length	73	75	74	75	76

Population abbreviations:
CEU — Utah residents with Northern and Western European ancestry
CHB — Han Chinese in Beijing
CHD — Chinese in Denver, Colorado
JPT — Japanese in Tokyo, Japan
LWK — Luhya in Webuye, Kenya
TSI — Tuscans in Italy
YRI — Yoruba in Ibadan, Nigeria

equivalent base coverage.

2.2.3 Sample coverage and gene affinities

As a metric of coverage for each sample, we calculated the sample-specific median gene RD, referred to as "Median Read Depth" (MRD); see Figure 2.2.1A and Methods 2.4.3. MRD was highest for the SC samples (1,710 ± 1,073, median 1,491 reads/gene; data presented as mean ± standard deviation), see Figure 2.2.2B. MRD was somewhat lower for the BI samples (1,070 ±

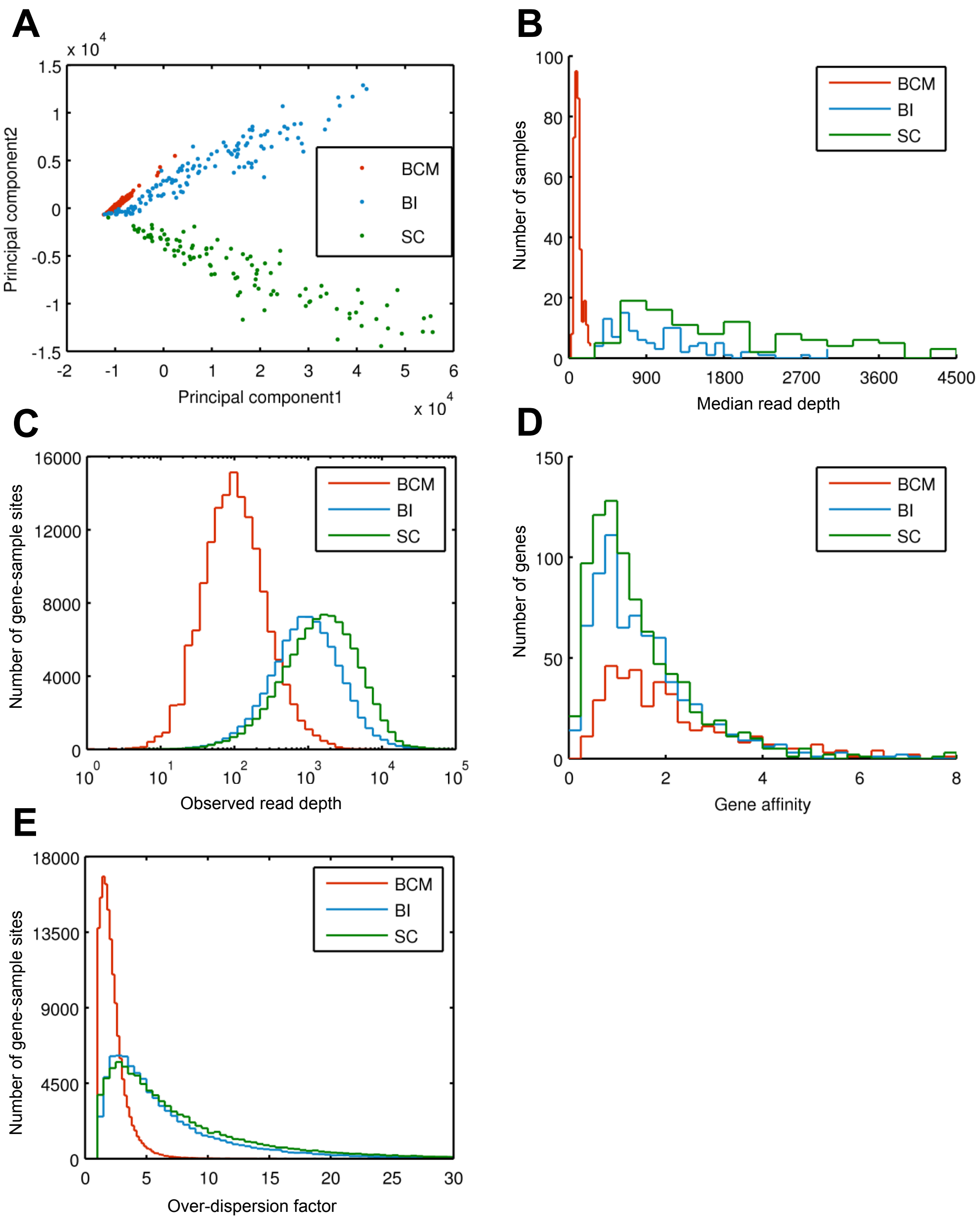


Figure 2.2.2: A. Principal component analysis of a "mixed" read depth matrix built with data from 3 different sequencing centers, SC (Wellcome Trust Sanger Institute), BI (Broad Institute) and BCM (Baylor College of Medicine). Each sample is represented as a point in the plot, with the first principal component plotted vs. the second principal component. Samples from different sequencing centers cluster separately from each other within this space, suggesting significant differences in the gene affinities among these three datasets. **B.** Distributions of MRD for each of the BCM, BI and SC samples **C.** Histogram of RD across all GSSs in the three datasets. **D.** Histogram of gene affinities across genes within each of the three datasets. **E.** Distributions of the RD over-dispersion factor (ODF) in our data.

803, median 860 reads/gene), and much lower in the BCM dataset (97 ± 52, median 87 reads/gene). As mentioned above, RD (distributed as in Figure 2.2.2C) is not determined by base coverage alone. Base coverage was highest in the BI data (70 ± 61, median 56 reads/base), followed by SC (56 ± 34, median 50 reads/base). The much lower RD in the 454 reads from BCM corresponds to only somewhat lower base coverage (23 ± 12, median 21 reads/base).

For each target we define a quantity, the "target affinity", intended to describe the number of reads (RD) being mapped to a given target, relative to the sample-specific MRD over all capture targets. Analogously, we define the gene-specific affinity as the ratio of the number of reads (RD) mapped to the targets (exons) belonging to that gene and the gene-specific MRD for that same sample (see Methods 2.4.3, Figure 2.2.2D). In general, tighter distributions of affinities, with mean and median as close to 1 as possible, are desirable because these correspond to more even target coverage. The observed gene affinities for our datasets (Figure 2.2.2D) were as follows: SC (1.40 ± 1.43, median 1.04), BI (1.58 ± 1.59, median 1.20), and BCM (2.63 ± 3.03, median 1.73). Because of the more favorable gene affinities, we used the SC data as our primary dataset for method development and experimental validations.

2.2.4 CNV candidates detected in the data

According to our Bayesian detection scheme, we call a heterozygous deletion event in a gene if the posterior probability value of CN = 1, *i.e.* P(CN=1 | RD) ≥ *h* where *h* is a pre-defined probability cutoff value. Similarly, a homozygous deletion is where P(CN=0 | RD) ≥ *h*. Although we detected both deletions and amplifications in the analyzed datasets, deletion events (even when in a heterozygous state) provide easier detectable signal than amplifications. For this reason we only discuss deletion events here and report candidate amplifications in Table 2.2.3.

Using a cutoff value *h* = 0.65, we detected 96 deletion events in the three datasets (36 in SC, 56 in BI, and 4 in BCM), all heterozygous deletions (Table 2.2.4, Table 2.2.5 and Table 2.2.6). The top ranked deletions are shown in Figure 2.2.3A. Most of the events were found in the Tuscan population, which constituted about half of the sample set. A subset of 10 of 36 gene deletions in

Table 2.2.3: Gene duplication calls in the SC dataset (PP: posterior probability)

Population	Sample	Gene name	Chr	Start [bp]	End [bp]	PP	$\mathrm{RD}_{\mathrm{obs}}$	$\mathrm{RD}_{\mathrm{exp}}$
CEU	NA12348	CD300LB	17	70030472	70039195	1	638	420
TSI	NA20533	CLDN10	13	95003009	95028269	1	2108	1582
CHB	NA18526	SNRNP27	2	69974621	69977184	1	530	383
CHB	NA18532	CES1	16	54401930	54424468	1	501	337
TSI	NA20752	NOM1	7	156435193	156455158	1	1335	966
TSI	NA20796	AHNAK	11	62040792	62059238	1	7330	5169
TSI	NA20796	ZNF264	19	62408577	62416161	0.999	1276	888
TSI	NA20801	GPR128	3	101811391	101896535	0.998	14747	8265
TSI	NA20772	STX16	20	56660469	56684753	0.998	2101	1605
TSI	NA20769	MRPS6	21	34419511	34436770	0.998	1585	1203
TSI	NA20774	ELAVL4	1	50383216	50439437	0.998	782	567
TSI	NA20804	CYP2A13	19	46291375	46293686	0.997	1289	984
TSI	NA20774	CREB5	7	28494318	28825421	0.996	1435	954
TSI	NA20796	ZNF32	10	43459504	43461587	0.996	911	646
TSI	NA20520	C6orf145	6	3668852	3683381	0.995	2015	1601
CEU	NA12348	GDNF	5	37851510	37870647	0.994	306	217
CHB	NA18561	PSMB4	1	149638688	149640730	0.986	3461	2216
CEU	NA12546	DAZAP2	12	49920394	49922509	0.985	2265	1651
TSI	NA20752	AATF	17	32380539	32488077	0.976	1157	843
CEU	NA12749	PAQR5	15	67439474	67483215	0.976	1684	1239
TSI	NA20769	BCL2L11	2	111597794	111638279	0.965	1813	1435
TSI	NA20804	PILRA	7	99809603	99835466	0.909	962	752
TSI	NA20589	C8orf85	8	118019664	118024121	0.903	147	91
TSI	NA20752	CCKAR	4	26092358	26100987	0.902	712	532
JPT	NA18973	HBG2	11	5278820	5523329	0.901	4151	3094
TSI	NA20774	HIPK1	1	114298778	114317657	0.9	2374	1626
TSI	NA20774	ODC1	2	10498301	10502609	0.897	1489	935
TSI	NA20796	STBD1	4	77446947	77450177	0.885	978	664
TSI	NA20589	CRIPAK	4	1378300	1379640	0.877	76	38
YRI	NA19189	PSMB4	1	149638688	149640730	0.853	2622	2090
TSI	NA20774	STX16	20	56660469	56684753	0.811	949	704
JPT	NA18980	CES1	16	54401930	54424468	0.788	1679	1036
TSI	NA20774	PAQR5	15	67439474	67483215	0.788	1048	676
CHB	NA18561	CRNN	1	150648694	150651333	0.778	4845	3172
TSI	NA20774	DKK4	8	42350775	42353720	0.76	493	362
TSI	NA20589	NOM1	7	156435193	156455158	0.74	1052	801
TSI	NA20769	RNF122	8	33525813	33535831	0.734	2574	2004
TSI	NA20796	ZNF521	18	20896674	21184908	0.721	3536	2738
TSI	NA20769	VLDLR	9	2625453	2631499	0.676	2092	1624

the SC dataset were found in two samples (NA18523 and NA20533), clustered in a contiguous string of deleted genes extending approximately 3 Mbp on chromosome 1 and 17, respectively, a genomic deletion event that we were also able to find in the 1000 Genomes Project whole-genome Low Coverage Pilot data [15] from the same samples.

Table 2.2.4: Gene deletion calls in the BI dataset (PP: posterior probability)

Population	Sample	Gene name	Chr	Start[bp]	End[bp]	PP	$\mathrm{RD}_{\mathrm{obs}}$	$\mathrm{RD}_{\mathrm{exp}}$
CHD	NA18695	TPM3	1	152396739	152422219	1	166	337
JPT	NA19066	TPM3	1	152396739	152422219	1	169	288
CHD	NA18687	RPL27A	11	8661325	8663929	1	93	182
JPT	NA18983	POU5F1	6	31240357	31241803	1	122	256
JPT	NA19066	POU5F1	6	31240357	31241803	1	166	318
JPT	NA19066	RPL27A	11	8661325	8663929	1	106	203
CHD	NA18687	TPM3	1	152396739	152422219	1	155	258
CHD	NA18687	POU5F1	6	31240357	31241803	1	156	285
JPT	NA19054	TPM3	1	152396739	152422219	1	135	230
CHD	NA18695	POU5F1	6	31240357	31241803	1	194	371
JPT	NA18960	SETD8	12	122441130	122455574	1	221	347
CHD	NA18164	RPL27A	11	8661325	8663929	1	129	223
JPT	NA19054	POU5F1	6	31240357	31241803	1	130	254
CHD	NA18695	SETD8	12	122441130	122455574	1	142	309
CHD	NA18695	RPL27A	11	8661325	8663929	1	128	238
CHD	NA18695	AKR1B1	7	133778020	133787045	1	310	554
CHD	NA18164	HAX1	1	152512874	152514801	1	214	339
CHD	NA18687	SETD8	12	122441130	122455574	1	125	237
JPT	NA19054	HFE	6	26201326	26202433	1	56	122
JPT	NA18983	RPL27A	11	8661325	8663929	0.99	95	164
JPT	NA18983	TPM3	1	152396739	152422219	0.99	147	232
JPT	NA19561	TRIM55	8	67202058	67209944	0.99	119	193
CHD	NA18687	RBMS1	2	160840394	160932124	0.99	334	575
CHB	NA18757	CRIPAK	4	1378300	1379640	0.99	327	669
JPT	NA19054	PSAT1	9	80109471	80113319	0.98	140	253

Table 2.2.4: Gene deletion calls in the BI dataset — continuation from previous page

Population	Sample	Gene name	Chr	Start[bp]	End[bp]	PP	RD_{obs}	RD_{exp}
JPT	NA19066	PSAT1	9	80109471	80113319	0.98	190	317
CHD	NA18164	TPM3	1	152396739	152422219	0.98	209	317
JPT	NA19568	OR8A1	11	123945175	123946141	0.98	471	764
JPT	NA19066	RAN	12	129923334	129926424	0.98	229	462
CHD	NA18695	KLHL12	1	201128284	201160913	0.97	767	1358
JPT	NA19066	SETD8	12	122441130	122455574	0.97	154	265
JPT	NA19066	RPS15A	16	18706886	18707936	0.96	83	161
CHD	NA18695	RPS15A	16	18706886	18707936	0.96	88	188
CHD	NA18687	KLHL12	1	201128284	201160913	0.96	621	1041
JPT	NA18983	SETD8	12	122441130	122455574	0.96	120	213
JPT	NA18983	DCTN5	16	23560365	23585966	0.96	177	298
JPT	NA18983	EIF2B5	3	185500333	185509372	0.94	856	1482
CHD	NA18687	ARG2	14	67187855	67187951	0.94	28	62
CHD	NA18695	PSAT1	9	80109471	80113319	0.93	221	371
CHD	NA18695	RBMS1	2	160840394	160932124	0.9	442	750
JPT	NA19561	OR8A1	11	123945175	123946141	0.89	254	466
YRI	NA19247	TIMM8B	11	111461229	111462657	0.88	40	89
CHD	NA18164	POU5F1	6	31240357	31241803	0.85	226	349
CHD	NA18164	KLHL12	1	201128284	201160913	0.8	803	1276
CHD	NA18164	SETD8	12	122441130	122455574	0.79	181	291
CHD	NA18687	RPS15A	16	18706886	18707936	0.79	81	144
JPT	NA19066	EIF2B5	3	185500333	185509372	0.78	1137	1840
JPT	NA19568	GABARAPL2	1	157676173	157676631	0.76	254	476
JPT	NA19560	OR8A1	11	123945175	123946141	0.75	614	1119
JPT	NA19058	RPL27	17	38404294	38408463	0.73	356	518
CHD	NA18699	SDPR	2	192408894	192419896	0.72	524	1033
JPT	NA18983	SPRR2G	1	151388989	151389210	0.67	81	147
JPT	NA19066	SPRR2G	1	151388989	151389210	0.67	105	182
JPT	NA19066	RBMS1	2	160840394	160932124	0.67	404	642
JPT	NA19054	EIF2B5	3	185500333	185509372	0.67	869	1470

Table 2.2.4: Gene deletion calls in the BI dataset — continuation from previous page

Population	**Sample**	**Gene name**	**Chr**	**Start[bp]**	**End[bp]**	**PP**	**$\mathrm{RD}_{\mathrm{obs}}$**	**$\mathrm{RD}_{\mathrm{exp}}$**
CHD	NA18695	RAN	12	129923334	129926424	0.66	290	539

When two or more gene deletions are detected in close proximity, it is likely that these events are part of a single, longer genomic deletion spanning the genes. With this in mind, we searched the sequenced genes for deletion events at a lower probability cutoff value ($h = 0.1$), but required that an immediate neighbor of a candidate gene be located within 3 Mbp and also show evidence for a deletion at the same probability cutoff. This procedure produced 17 heterozygous deletion calls in the SC dataset, 11 calls in the BI dataset (but no such calls were made in the BCM dataset). The union of both callsets (*i.e.* those made with and without use of neighboring information) resulted in a total of 107 unique deletion events (41 in SC dataset, 62 in BI, and 4 in BCM). We note that none of the events we detected in our data were at high allele frequency. In fact, even the most "common" events were only present in two samples, as heterozygotes.

2.2.5 Call-set accuracy assessment

To assess the accuracy of deletion calls made in the SC dataset, researchers from Stanford University (Dr. Fabian Grubert and Dr. Alexander Urban) helped me perform experimental validations on calls made with posterior probability ≥0.65 without neighbor information, using quantitative PCR (qPCR) (see Methods 2.4). The validation results are summarized in Figure 2.2.3B. Many of the CNV calls submitted for qPCR validation are not given a conclusive results. This is gernerally caused by some limitations of this technologies such as the high similarity between the test DNA fragments and the target template and the defective design of the primers [81] Of the 36 calls made, we evaluated 26. All 22 calls with posterior probability ≥0.95 and 4 out of 12 calls (randomly selected) with posterior probability between 0.65 and 0.95 were

Table 2.2.5: Gene deletion calls in the SC dataset (PP: posterior probability)

Population	Sample	Gene name	Chr	Start [bp]	End [bp]	PP	RD_{obs}	RD_{exp}
YRI	NA18523	BCL2L15	1	114225268	114231520	1	533	1158
YRI	NA18523	HIPK1	1	114298778	114317657	1	2539	5272
TSI	NA20533	GLOD4	17	610163	632245	1	1322	2295
TSI	NA20533	C1QBP	17	5277059	5282317	1	793	1416
TSI	NA20533	C17orf91	17	1562414	1563890	1	369	574
YRI	NA18523	NRAS	1	115052679	115060304	1	702	1462
YRI	NA18523	TRIM33	1	114741793	114808533	1	2610	5225
TSI	NA20533	TRPV3	17	3363961	3404894	1	3365	5275
TSI	NA20774	PTMAP1	6	30725671	30728671	1	132	260
TSI	NA20796	SNRNP27	2	69974621	69977184	0.998	105	194
TSI	NA20807	HIST1H2BC	6	26231731	26232111	0.998	42	90
TSI	NA20772	ULBP1	6	150331436	150332954	0.997	104	205
TSI	NA20807	CYP2A13	19	46291375	46293686	0.996	126	204
YRI	NA18508	PTMAP1	6	30725671	30728671	0.992	145	230
CEU	NA07000	PSG8	19	47950287	47960273	0.99	29	70
CEU	NA11893	PSG8	19	47950287	47960273	0.985	43	86
TSI	NA20771	PTMAP1	6	30725671	30728671	0.98	533	862
TSI	NA20773	CCK	3	42274594	42280126	0.971	282	474
CEU	NA07000	HMGN4	6	26653414	26653686	0.966	68	132
CEU	NA12749	HMGN4	6	26653414	26653686	0.966	156	286
TSI	NA20772	AIF1	6	31692086	31692262	0.964	51	124
CEU	NA12348	DUSP10	1	219942377	219946216	0.962	155	242
YRI	NA18508	ULBP1	6	150331436	150332954	0.941	40	79
YRI	NA18523	PPM1J	1	113056116	113057756	0.891	560	924
TSI	NA20807	POU5F1	6	31240884	31241803	0.891	124	193
TSI	NA20772	SERPINA11	14	93978696	93984864	0.889	786	1243
CEU	NA07000	KRT18P19	12	51630379	51632393	0.887	85	174
CEU	NA12348	ULBP1	6	150331436	150332954	0.879	49	88
YRI	NA18523	RHOC	1	113054308	113055529	0.867	557	955
CEU	NA12348	STBD1	4	77446947	77450177	0.839	246	395
CEU	NA07000	POU5F1	6	31240884	31241803	0.823	106	169
CEU	NA12749	SNRNP27	2	69974621	69977184	0.775	142	216
TSI	NA20752	POU5F1	6	31240884	31241803	0.723	76	142
TSI	NA20807	HIST1H2BO	6	27969220	27969600	0.723	48	88
TSI	NA20589	POU5F1	6	31240884	31241803	0.697	61	117
TSI	NA20786	NPSR1	7	34884213	34884321	0.678	51	88

Table 2.2.6: Gene deletion calls in the BCM dataset (PP: posterior probability)

Population	Sample	Gene name	Chr	Start [bp]	End [bp]	PP	RD_{obs}	RD_{exp}
LWK	NA19355	MBD5	2	148932798	148986980	0.999	618	973
CHD	NA17970	MTERFD2	2	241684086	241687982	0.996	255	393
CHB	NA18618	GABARAPL2	16	74159436	74168768	0.8	58	99
CHD	NA18135	PSMB4	1	149638688	149640929	0.729	390	605

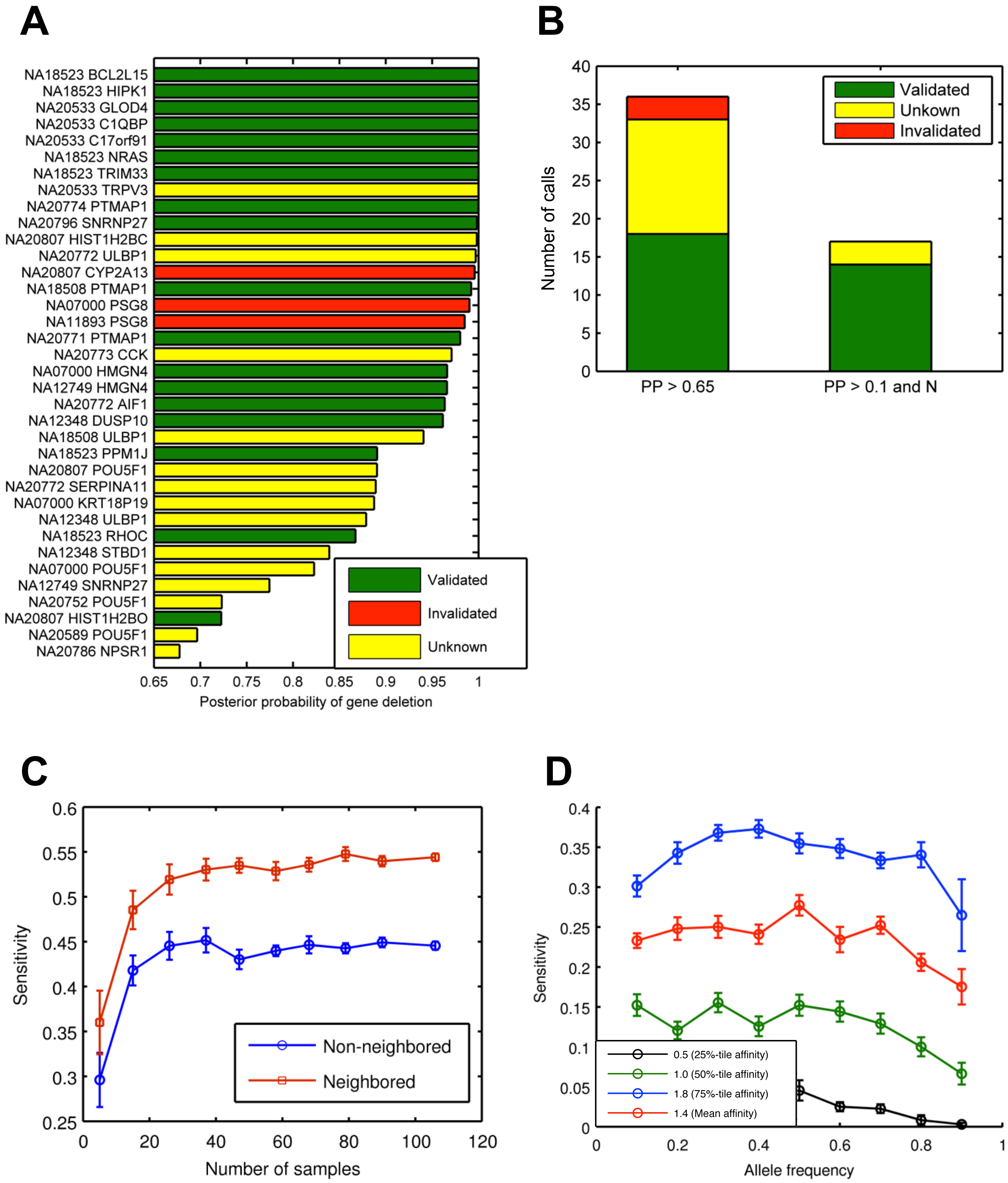


Figure 2.2.3: A. Top-ranked (by posterior probability) deletion events in the SC dataset. **B.** Validation results for different callsets (left — without neighboring information, right — with use of neighboring information). Green denotes events positively validated either in our experiments or as known events [18]; red — calls validated negatively in our experiments; yellow — calls without validation status (not submitted for validation or validation experiments without conclusive outcomes). **C.** Detection sensitivity as a function of number of samples. **D.** Sensitivity of detecting common CNV as a function of the deleted allele frequency.

Table 2.2.7: Validation results

	Posterior>=0.95 without neighbor information	0.65<=Posterior<0.95 without neighbor information	Posterior>=0.1 with neighbor information
Validated per previous publication	4	2	7
Validated positively *de novo*	11	1	7
Validated inconclusively *de novo*	4	1	0
Validated negatively *de novo*	3	0	0
Submitted for validation but without result	0	10	3
Total calls	22	14	17

submitted for validation. A set of 6 were considered positively validated as they appeared in an earlier publication [18] and 20 were validated *de novo* using qPCR. The qPCR validations produced positive results for 12 calls (measured fold change <0.7) and negative results for 3 calls (measured fold change >0.8). The validation results for the remaining 5 were inconclusive. All the 17 neighbored calls with posterior probability ≥0.1 were selected for validation. A set of 7 were considered valid per previous publication [18], 7 were positively validated *de novo* and none was found invalid; validation was not obtained for the remaining 3. The union of those two callsets counted 41 calls and 32 of them were evaluated. Among these 32 calls 7 were considered positively validated per previous publication [18], 14 were positively validated *de novo*, 3 were invalidated, 5 were inconclusive and 3 did not obtain the validation results. The numbers of validated calls are presented in Table 2.2.7. The selection procedure for site validation was as follows: (1) We selected sites for validation (in some categories, all candidates, in others, a random selection); (2) we searched the literature [18], and removed from the validation list events that we found as validated in one of the publications we consulted; (3) events that remained on the list were sent for experimental validation. The overall FDR for the union of calls made with and without neighboring information can be estimated as 12.5% (3/24).

2.2.6 Sensitivity

We performed simulations to assess the detection efficiency of our method, both for individual gene and for pairs of neighboring genes deletions. Specifically, in each sample we randomly selected (1) 5 out of 862 genes in one simulation and (2) 5 pairs of neighboring genes in another simulation. In the selected genes we down-sampled the actual read depth seen in the experimental data by a factor of 2 to simulate a heterozygous deletion. The results of those simulations are presented in Figure 2.2.3C. Of the 530 gene deletions, we detected 237 (45%). Of the 530 gene-pair deletions we detected 287 (54%). We also performed simulations on smaller subsets of the original 106 samples to assess the impact of sample size on detection sensitivity. Reduction of sample size did not substantially degrade detection sensitivity as long as the number of samples was >20. Therefore, our detection efficiency is around 45% without using neighboring information and approximately 50-55% with the use of neighboring information, in the SC dataset.

In addition to simulations, we compared our dataset to a published study [18]. This study reported 12 heterozygous deletion events in samples and genes (in our terminology, GSS) that were part of our analyzed dataset. We detected 6 of these 12 events, which is broadly consistent with our overall sensitivity estimate.

Finally, we investigated our sensitivity to common events (see Methods 2.4.8) using simulations. Figure 2.2.3D shows detection sensitivity as a function of gene-level affinity: for a gene affinity value of 1.8 (representing the 75th percentile of our data), sensitivity to common events (allele frequency between 10% and 90%) approaches 40%. Note that the detection efficiency starts to decrease at high allele frequency (>70%) due to a reduction of the overall read depth because more samples have a deletion and a corresponding depleted read depth signal. The estimated gene affinity will be dominated by these deleted events. Instead of detecting these deletion events, the samples with normal copy numbers will be detected as amplifications. We can also see that the median gene affinity is substantially lower than the mean because the distribution of gene affinity has a long tail at the high end (Figure 2.2.2D). Since sensitivity is directly related to the gene affinity, the simulated data with the substantially higher mean gene affinity (red) has better sensitivity than with the substantially lower median gene affinity (green).

2.2.7 The number of CNV events in the samples

We estimated the total number of gene deletions in the SC dataset from the number of detected events (41), the FDR (12.5%) and the detection efficiency (45%), as ~66 in total 106 samples, or a nominal 0.62 deletions per sample . By projecting the per-sample number, corresponding to 3.9% of the exome (862 genes of 21,999), onto the whole exome, our estimate for the average number of genic deletion events is 16 ± 4 per sample. This estimation is very close to that from a large-scale whole-genome scanning CNV study with high-resolution CGH technology published in 2011 [18]. In that study, 6187 heterozygous deletions were found in exon regions from 450 samples (on average, it is ~14 heterozygous deletions per exome). This estimation is representative for the whole-exome sequencing data since the 1000 Genomes Exon Pilot Project randomly selected all the exon targets from the CCDS collection. Our gene set is therefore a quasi-random sampling of known human genes, with no intentional enrichment for any given gene family. Figure 2.2.4A and 2.2.4B show the distributions of exon length in the gene list used for our analysis and the full human exome. There is no significant difference between these two distributions: the median and the standard deviation of the exon length for our study are 125 bp and 236 bp, whereas the corresponding values for the whole exome are 127 bp and 264 bp. The similarity of these two distributions suggests that our estimation of the number of events per sample is unbiased and is representative for a whole-exome analysis.

2.2.8 Detection efficiency as a function of data quantity and data quality

As discussed earlier, our algorithm's sensitivity was 45% at 87.5% accuracy. Both sensitivity and accuracy are considerably lower than achievable for SNP detection in the same datasets [15]. This poses the more general question of how detection efficiency is influenced by sample size, data quantity, and data quality. Our simulations show that sensitivity only modestly depends on

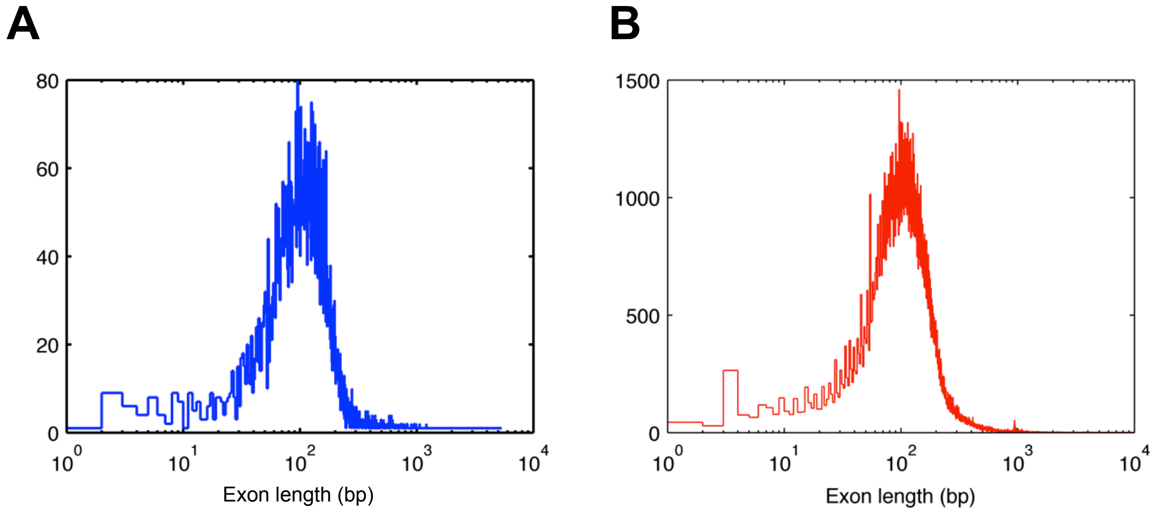


Figure 2.2.4: A. Exon length distribution in the gene list used for our analysis (median: 125 bp, standard deviation: 236 bp). **B.** Exon length distribution of the whole exome (median: 127 bp, standard deviation: 264 bp). These two distributions are very similar to each other, suggesting our estimation of the number of events per sample is unbiased and is representative for a whole-exome study.

sample size, above approximately 20 samples (Figure 2.2.3C).

We found that the primary factors that determine detection efficiency are (1) sequence coverage, or more precisely, RD (higher RD supplies more statistical power to detect systematic changes in coverage); (2) the level of over-dispersion of the RD distribution for individual genes (the more the RD distribution departs from an expected *Poisson* distribution, the less one can rely on the statistics); and (3) the shape of the distribution of RD across all genes in the dataset, determined by the gene affinities (uneven distribution means that detection power is low in a high fraction of the genes, but this effect is not compensated by the extra coverage in other, "over-sequenced" genes where detection efficiency is already high, see Figure 2.2.5A. Favorable scenarios therefore involve distributions in which all or most genes have sufficient RD for detection).

For each gene, we compute a quality index (QI) taking into account the variance of the expected read depth for that gene (assuming the ideal, *Poisson* distribution), $\mathrm{RD}_{\mathrm{expected}}$, and a over-dispersion factor, ODF (see Method 2.4.5), that quantifies the over-dispersion of RD

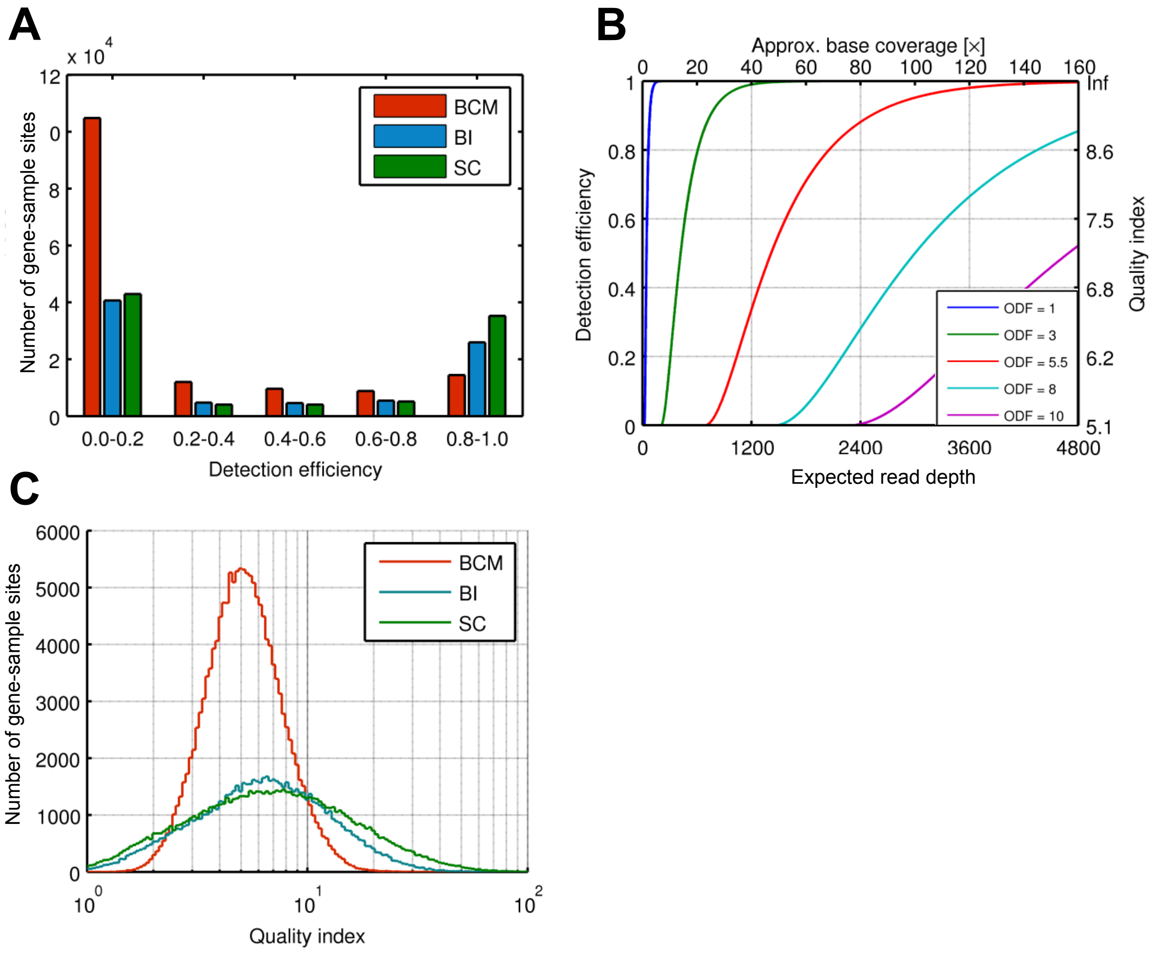


Figure 2.2.5: A. Distributions of the detection efficiency estimated from the quality index for each gene-sample site. **B.** Theoretical detection efficiency (at posterior probability cutoff $h = 0.65$) as a function of expected read depth, plotted for various values of the over-dispersion factor. **C.** Histograms of the quality index (QI) distribution in the three datasets. Overall, QI was highest in SC: 9.4±8.8 (median 6.6); second highest in BI: QI $=$ 7.6 $\pm$ 5.6 (median 6.2); and lowest in BCM: QI $=$ 5.5 $\pm$ 2.3 (median 5.0).

Table 2.2.8: Nominal prior probabilities corresponding to the range of gene region copy numbers derived from Conrad *et al.* 2010 [18]

Copy number	Prior probability per gene
0	$6.34 \cdot 10^{-4}$
1	$2.11 \cdot 10^{-3}$
2	$9.96 \cdot 10^{-1}$
3	$5.38 \cdot 10^{-4}$
4	$6.68 \cdot 10^{-4}$
5	$3.57 \cdot 10^{-5}$
6	$7.52 \cdot 10^{-6}$
7	$1.39 \cdot 10^{-6}$
8	$3.61 \cdot 10^{-7}$
9	$4.37 \cdot 10^{-8}$

relative to the *Poisson* expectation:

$$QI = \frac{\sqrt{RD_{expected}}}{ODF} \tag{2.1}$$

QI is directly related to detection sensitivity, as shown in Figure 2.2.5B. According to our power calculations, for the posterior detection threshold value we used in this study ($h = 0.65$), sensitivity is completely diminished for genes with QI < 5.1. QI ≥ 7.2 is required to achieve 50% sensitivity, and QI ≥ 9.5 to achieve 90% sensitivity. This estimated sensitivity from QI is made only for heterozygous deletions. To achieve the same sensitivity for detecting higher copy number variation (CN ≥ 3), higher QI value will be required since the difference of prior probability between higher copy and normal copy (CN = 2) is greater than that between heterozygous deletion and normal copy (Table 2.2.8).

The distributions of QI values in our three datasets are shown in Figure 2.2.5C. Overall, QI was highest in SC: 9.4 ± 8.8 (median 6.6); second highest in BI: QI = 7.6 ± 5.6 (median 6.2); and lowest in BCM: QI = 5.5 ± 2.3 (median 5.0). The corresponding distributions of detection efficiency values are shown in Figure 2.2.5A. Because detection efficiency increases abruptly from 0 to almost 1 over a narrow range of QI values (note the mapping between the vertical axes in Figure 2.2.5B), the distribution of detection sensitivity (Figure 2.2.5A) is strongly bimodal, with the vast majority of GSS having either close to zero or close to 100% sensitivity. Even in the SC dataset with the highest overall QI values, in less than half of the GSS does the quantity and quality of the data support >80% detection efficiency. There was also very substantial variation across samples: only 15 of the 106 SC samples had sufficiently high coverage to support ≥ 90% overall sensitivity, and in 22 samples overall sensitivity was below 10%.

Given that QI improves only with the square root of RD, over-dispersion can profoundly influence detection performance, as shown in Figure 2.2.5B. The ODF values we chose for this figure correspond to the 25th, 50th and 75th percentile, and the mean values (ODF = 3, 5.5, 10, and 8, respectively) in the SC dataset. Using the observed distribution of QI in the SC dataset, we predict 46% sensitivity, in good agreement with our estimate based on simulations. The QI formulation permits one to estimate CNV (or specifically in our case, heterozygous deletion) detection power in any given exon capture dataset, based on the read mappings. One can also use the formulation to calculate the amount of base coverage required for a given level of desired power, to guide data collection. For example, using the distributions of QI values in the SC dataset, one would need to collect an overall 110× coverage, assuming 36 bp reads, to achieve 60% detection power, and 320× coverage to achieve 80% detection power. However, if DNA capture methods improved to support a median ODF = 3, assuming an accordingly scaled version of the observed distribution of QI in the SC dataset, one would only need to collect 33× coverage for 60% power, and 96× for 80% power. It is important to also point out that, in the case of whole-exome data, sensitivity would also improve just by virtue of the higher density of targeted genes, if one were to integrate in one's pipeline neighbor-gene based detection.

2.2.9 Functions of affected genes

Although function study is not our major goal for this research work, we still found some genes affected by CNVs in the callset that are correlated with human diseases. For example, heterozygous deletions are detected at POU5F1, a gene that is responsible for the self-renewal activity and pluripotency of embryonic stem cells and germ cells [82], in many Asian samples from both BI and SC datasets. The mutations of this gene and EWSR1 together are reported to play an important role in sarcomagenesis and tumor cell maintenance [83]. Two genes from BCL2 familiy, BCL2L11 and BCL2L15 are detected as duplications and heterozygous deletions recpectively in the SC dataset. BCL2 family is well known as one of the regulators for programmed cell death. When it dominants, the programmed cell death will be suppressed and the cell can therefore survive [84]. The dysfunction of this gene is associated with many types of cancers such as breast cancer [85] and prostate cancer [86]. Many other cancer-related genes are discovered as CNVs in the callset as well, such as NRAS [87], ODC1 [88] and CRIPAK [89, 90]. Besides cancers, genes associated with neurodegenerative genetic disorders are also seen. SETD2, also known as HYPB (huntingtin yeast partner B), is involved in the modulation of chromatin structure and may also bind to DNA promoters and interact with Pol II, thereby promoting transcription [91]. The mutation of SETD2 is associated with the pathogenesis of Huntington's disease [92], which is characterized by a loss of striatal neurons, leading to brain deterioration and, ultimately, death. Another gene in the detected in our callset, GDNF, a highly conserved neurotrophic factor. The major function of the protein production of this gene is to promote the survival and differentiation of dopaminergic neurons in culture and to prevent apoptosis of motor neurons induced by axotomy [93]. The dysfuction of this gene may lead to Parkinson's disease, a degenerative disorder of the central nervous system. HFE, a gene that econdes a membrane protein that is responsible for regulating iron absorption, is invloved in the devlopment of Alzheimer's disease [94] since the iron imbalance may have impact on plaque formation, amyloid processing, and expression of and response to inflammatory agents. Many other disease-correlated genes, such as TPM3 (muscle weakness [95]), DAZAP2 (male infertility [96]) and HAX1 (neutropenia [97]) are also seen in our callset. Due to the design of 1000GP exon capture sequencing study, the phenotype data of all the samples are not available so it is very hard for us to do any further functional studies of these detected CNVs. However, for other large whole-exome sequencing projects that focus on functional studies, our method could be potentially used for detecting events with significant biological impact.

2.3 Discussion

We have developed a novel, Bayesian method to identify CNVs in exon-capture data. We applied this method (and a simple extension using neighbor-gene information) to the 1000 Genomes Project Exon Sequencing Pilot dataset. We were able to achieve reasonable sensitivity (which is limited by the quality of the dataset instead of our methodology) and specificity in a dataset that was optimized for SNP discovery and, as discussed above, is far from ideal for CNV detection. As new whole-exome sequencing data become easily available nowadays with higher coverage and low or even none (single molecule sequencing) PCR bias, the detection efficiency of our method should be significantly improved based on our statistical analysis (quality index).

Krumm and his colleagues recently published a method, CoNIFER [98], that also used read-depth signal to detect CNV in the exome capturing sequencing data. Like our method, CoNIFER normalizes the read depth signal in order to discover the CNV. However, it is quite different for these two algorithms in the approach of calling samples copy number variants on the basis that they present aberrant read depth. As we mentioned previously, our method deploys specific models for copy numbers 0, 1, 2, and is capable of detecting both rare, intermediate frequency, and common CNV events. On the other hand, CoNIFER deploys singular value decomposition (SVD) to remove noise from the read depth data, and interprets the first "k" singular values as noise in the data. This approach may identify systematic variance in the data caused by a high-frequency CNV event as noise and removes it. Therefore CoNIFER has limited power for detecting common CNV events. On the other hand, our method is capable of detecting CNV events on the entire frequency spectrum, and is therefore more generally applicable.

The main accomplishment of this work is that we provide a statistically rigorous algorithm for CNV detection in exon capture data, backed by experimental validations, that can be applied to the thousands of exomes sequenced to date in various medical projects, and to nascent and on-going projects targeting increasingly higher numbers of samples. Our formulation allows investigators to assess detection power in existing datasets and to take into account CNV detection power during experimental design for future datasets. We also uncovered >100 heterozygous deletion events in the 1000 Genomes samples we examined, allowing us to estimate the average number of heterozygous deletions per exome (as ~16 events per exome for a diploid genome. See Results 2.2.7). Because we focused on algorithm we only did some brief functional assessment of these sites is beyond in this study. Nevertheless, these and other gene deletions that will be found using our methods are very likely to uncover events with strong functional significance.

2.4 Methods

The overall detection workflow (shown in Figure 2.2.1) consists of five main steps: (1) We tabulate the observed read depth for every GSS. (2) We determine whether the distribution of read depth for a specific gene distribute across samples should be modeled using simple uni-linear fit or using a more sophisticated tri-linear fit. (3) If the simple uni-linear fit is found suitable, we determine an expected read depth for every GSS under a null hypothesis of a normal copy number, using a simple linear fit model. (4) Subsequently, we compare the observed read depth for a GSS to the corresponding expectation , calculate a Bayesian posterior probability for each copy number considered (CN = 0-9) and report events that pass the pre-defined posterior probability threshold with a non-normal CN. (5) If data do not allow for modeling using a simple uni-linear fit model, we perform a more sophisticated tri-linear fit. The tri-linear fit directly assigns copy number to every sample.

2.4.1 Observed read depth

Capture sequencing reads from the 1000 Genomes Project Exon Sequencing Pilot Project were downloaded, in FASTQ format, from the 1000 Genomes Project DCC site: http://1000genomes.org. The reads were mapped using the MOSAIK read mapping program [57], to the NCBI build 36.3 human reference genome. The resulting read alignments (in BAM format) were further processed to remove duplicate reads, and reads with low mapping quality (<20) [57].

Gene target regions were also downloaded from the 1000 Genomes Project site. For each GSS, we determined RD as the number of distinct reads that had their first (5') base uniquely mapped within an exon of that gene. This resulted in a matrix of RD observations (illustrated in Figure 2.2.1C left).

2.4.2 Data filtering

We discarded all duplicate reads and all reads with mapping quality less than 20. We also discarded all the targets with median RD less than 30. Similarly, we discarded all the samples with median RD less than 30. In 454-sequenced data, this led to discarding almost all targets and samples; therefore we relaxed those criteria to 5 and 1, respectively. Additionally, we discarded all the genes that failed to exhibit correlation between observed RD and MRD at $r^2 \geq 0.7$.

2.4.3 Expected read depth based on uni-linear fit and tri-linear fit

In the first attempt, we use the simple uni-linear fit; we calculate the expected read depth for normal copy number (CN = 2) as the product of a gene-specific capture affinity value, α_g, and a sample-specific measure of read coverage, the median of read depths, MRD_s, across all genes for that sample:

$$RD_{gs} = \alpha_g \cdot MRD_s \tag{2.2}$$

The gene-specific capture affinity (α_g) is determined as the slope of a least-squares zero-intercept linear fit between the gene-specific read depth (RD_{gs}) and the median read depth (MRD_s) for all samples (illustrated in Figure 2.2.1B). This procedure resulted in a matrix of RD expectations (Figure 2.2.1C right).

The afore-mentioned procedure requires a single-line linear fit between RD_{gs} and MRD_s. The quality of such a fit is evaluated by comparing r^2 against a predetermined threshold (≥0.7 as described before). When this indicates poor quality of the single-line linear fit, we attempt to perform a tri-linear fit.

Briefly, we attempted to minimize error function:

$$error_g = \sum_s min\{(RD_{g,s} - \alpha_g \cdot MRD_s), (RD_{g,s} - \frac{\alpha_g}{2} \cdot MRD_s), (RD_{g,s} - 0 \cdot MRD_s)\} \quad (2.3)$$

where s iterates over samples and g indicates the gene in question. Note that the tri-linear fit directly assigns copy number to each GSS. Please see Common CNVs (Methods 2.4.8) for more detail.

2.4.4 Copy number probabilities

We used a Bayesian scheme to calculate the probability $\mathrm{P}(\mathrm{CN}_{gs}|\mathrm{RD}_{gs})$ of a given copy number at a given GSS, based on the observed read depth. We only considered CN = 0-9 *i.e.* homozygous deletion (CN = 0), heterozygous deletion (CN = 1), normal copy number (CN = 2), and amplifications of various magnitudes (CN > 2). We assigned prior probabilities $\mathrm{P}(\mathrm{CN}_{gs})$ to each copy number based on CNV events reported in an earlier study [18] (Table 2.2.8). We assumed that, for each distinct CN, the observed RD obeys an over-dispersed *Poisson* distribution. Its mean value for normal copy number (CN = 2) is calculated according to (Eq. 2.2) and for other copy numbers it is proportionally scaled. The standard deviation of the distribution includes an over-dispersion factor (ODF) in the range of 1 to 20 to account for over-dispersion (variance beyond the level of *Poisson* fluctuations, see Method 2.4.5).

Briefly, to account for over-*Poisson* dispersion, we used observed RD_{gs} and calculated corresponding *z-score* under an assumption of an ideal *Poisson* distribution at every GSS. Subsequently, we calculated a sample-specific standard deviation of that *z-score* for every sample and annotated it as sample over-dispersion factor. Similarly, we calculated a gene-specific standard deviation of *z-score* for every gene and annotated it as the gene-specific over-dispersion factor. If the assumption of an ideal *Poisson* distribution were true, those sample- and gene-specific standard deviations should equal 1. Subsequently, we calculated the over-dispersion factor for every GSS as a product of respective sample- and gene-specific ODFs. The ODF was then normalized and assigned to 1 if less than 1.

We used the over-dispersed *Poisson* distributions to calculate the data likelihoods $P(RD_{gs}|CN)$ for all considered CN values. Finally, we used Bayesian method to estimate the posteriors for each considered CN (Eq. 2.4).

$$P(CN_{gs}|RD_{gs}) = \frac{P(CN) \cdot P(RD_{gs}|CN)}{\sum_{CN'} P(CN') \cdot P(RD_{gs}|CN')} \tag{2.4}$$

A CNV event is reported the posterior probability of a non-normal copy number is above a pre-defined threshold value, *h*.

2.4.5 Insight from empirical data and accounting for over-dispersion

We performed a simulation to assess potential variability in the gene affinities on the over-dispersion. Using this data, we calculated expected read depth $RD_{expected}$ for every GSS as product of respective gene affinity and MDR. Subsequently, we calculated read depth using *Poisson* distribution with $RD_{expected}$ as parameter. The *z-score* calculated from that distribution followed a normal distribution $N(0, 1)$, as expected for an ideal case.

Subsequently, we randomly distorted the vector of gene affinities; *i.e.* we drew a random number from a normal distribution $N(a_g, 0.15 \cdot a_g)$to be used instead of the exact affinity a_g. With increased variability in gene affinities, the distribution becomes progressively wider; at a 15% increase in variability the results are comparable to the distribution of the empirically calculated *z-score* (Figure 2.4.1). This result indicates that as little as 15% variability in gene affinities is enough to reproduce the distribution over-dispersion observed in the experimental data.

If we knew ODF for every GSS in our data, we could correct for it, so that

$$\frac{RD_{observed} - RD_{expected}}{c \cdot \sqrt{RD_{expected}}} \sim N(0, 1) \tag{2.5}$$

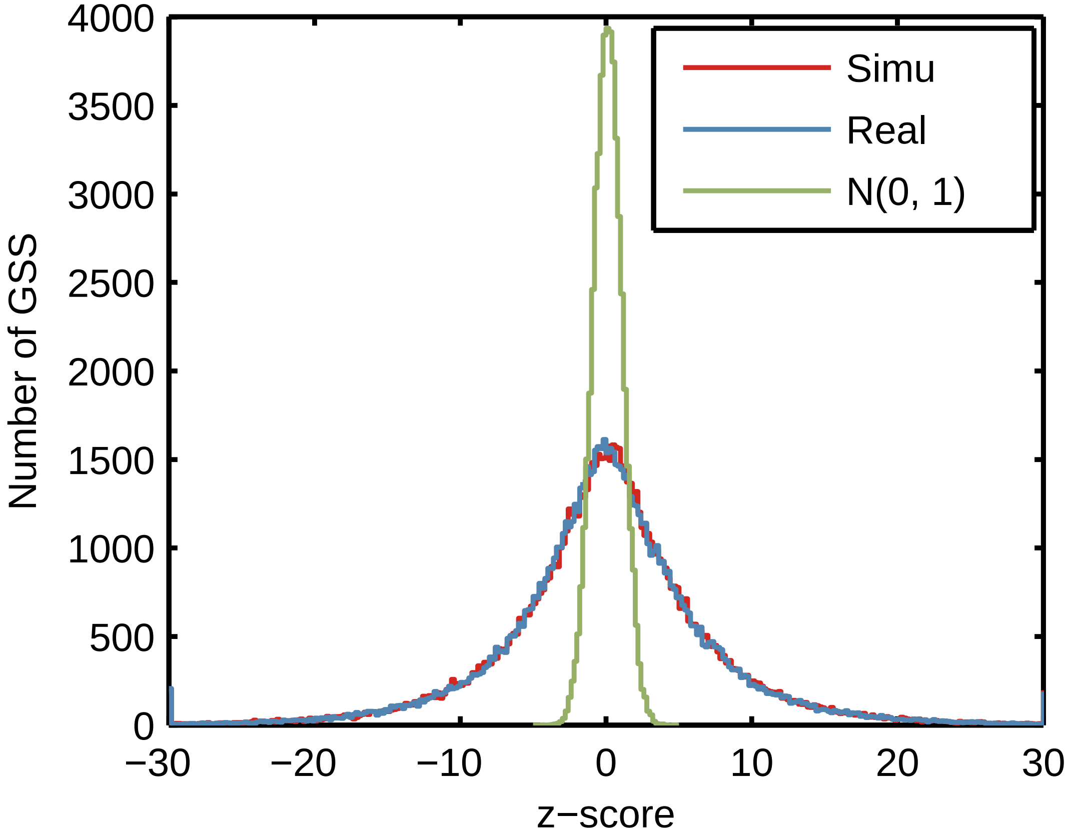


Figure 2.4.1: To generate the simulated data, we introduced a normal random noise to each target affinity calculated from the real data with 15% of the value of the target affinity, N(α, 0.15$\cdot\alpha$). The distribution of the *z-score* ($\frac{\mathrm{RD_{obs}} - \mathrm{RD_{exp}}}{\sqrt{\mathrm{RD_{exp}}}}$) from the simulated data (red) is very similar to that of the real data (blue). Note that both *z-score* distributions from simulated and real data are much wider (dispersed) than the ideal normal distribution (green) due to the over-dispersion effect.

where c is the sample-gene-specific correction factor for the over-dispersed *Poisson* effect (over-dispersion factor, ODF).

As indicated above, ODF remains constant over a range of coverage only under assumption of mutual independence of subsequent runs. When the entire *z-score* matrix is considered, that assumption is obviously violated (*i.e.* RDs in different genes in a sample are correlated by sharing the same MDR and RDs in a gene in different samples are correlated by sharing the same gene affinity).

In the absence of a fundamental model describing interplay between gene affinities varying across genes, samples and machine runs, we developed an empirical procedure to account and correct for over-dispersion.

We estimated the over-dispersion factor for each site according to the following steps. First we calculated a *z-score* matrix $[z_{s,g}]$,

$$z_{s,g} = \frac{observed_{s,g} - expected_{s,g}}{\sqrt{expected_{s,g}}} \tag{2.6}$$

from the observed read depth matrix $[observed_{s,g}]$ and expected read depth matrix $[expected_{s,g}]$.

Then for every row and for every column in the "z-score" matrix, we calculated their respective standard deviations. This procedure generated a column vector $[c_{s,*}]$ of row (sample-specific) standard deviations and a row vector $[c_{*,g}]$ of column (gene-specific) standard deviations. Subsequently, the over-dispersion factor matrix $[c_{s,g}]$ was calculated as:

$$c_{s,g} = \frac{c_{s,*} \cdot c_{*,g}}{mean(c_{*,g})} \tag{2.7}$$

If any over-dispersion factor was to fall below 1, it was assigned 1 since no counting experiment of independent trials should have a variance less than that of a *Poisson* distribution.

Once the over-dispersion factor was calculated, we could model data likelihood using a normal distribution $N(RD_{exptected}, c \cdot \sqrt{RD_{expected}})$.

2.4.6 Neighboring gene deletions

A simple extension of the algorithm used neighboring gene deletion events as part of the detection method. For the purpose of our algorithm, the genes were deemed "neighboring" if they were located on the same chromosome, the segment between those genes was no longer than 3 Mbp and no gene was sequenced in between. In principle, when a gene has a deleted neighbor, we should assume a higher prior probability of a deletion in the gene in question. Since the posterior probability usually scales monotonically with the prior, for practical reasons we assumed a lower Bayesian posterior probability threshold ($h = 0.1$) to produce a preliminary list of candidate events. Events on this list for which at least one of the two immediate neighbor genes was also on the list were retained.

2.4.7 Sensitivity estimation

We carried out sensitivity estimation in the SC dataset, using simple simulations. In each simulation cycle, we drew 5 genes randomly from every sample, and downscaled the observed RD for those genes by a factor of 2, to emulate heterozygous deletions. We then applied our standard detection procedure to this "spiked" dataset, and tabulated the fraction of simulated events that were detected by the algorithm.

2.4.8 Common CNVs

We evaluated all genes that failed to achieve $r^2 \geq 0.7$ using the linear fit model from Figure 2.2.1B. The results of that evaluation are shown in Figure 2.4.2. The last row describes result for gene RNF150 that achieved the worst r^2 of 0.48. The histogram shown in the left columns demonstrates distribution of observed RD to MRD (taken as from Figure 2.2.1B), In case of a rare CNV (or lack of CNVs at all), one would expect a unimodal distribution centered around that gene affinity. For a common CNV, one additional peak corresponding to CN = 1 centered around half of that gene affinity, and another peak corresponding to homozygous deletion (CN = 0) around 0, should be visible. However, the data shown do not allow identifying such a pattern of either bi- or tri-modal distribution.

Additionally, the histogram of quality index calculated for that gene is presented in the right column. The low values of quality index further corroborate the conclusion that the absence of a call in that locus is due to lack of high quality data rather than due to a hypothetical common CNV event. Careful inspection of the graphs calculated for all 69 genes the failed simple linear fit reveals lack of evidence for a common CNV in any of them. Notably, in the SC dataset only 28% of GSS in genes with $r^2 < 0.7$ were potentially detectable vs. 62% in genes with $r^2 \geq 0.7$.

With no common CNV present in the experimental data, we tested the sensitivity of our algorithm using simulated deletions. We used realistic gene affinities (mean and three quartiles from Figure 2.2.2B) and the empirical MRD_s for 106 samples. We assumed frequency of the deleted allele among 106 samples varying from 0 to 100% in 10% increments; we allowed for random segregation, so that both homo- and heterozygous deletions were introduced. Then for each sample we calculated the expected read depth as a product of MRD and affinity; however in the samples drawn for a heterozygous deletion we used halves of the nominal affinities and in the samples drawn for a homozygous deletion, we multiplied the MRD by 0.01 to account for reads erroneously mapped into that region. Having an expected read depth m for each sample, we drew a random read depth using a normal distribution, $N(m, ODF\sqrt{m})$, where ODF was assumed as 8. In Figure 2.4.3B and 2.4.3C we show the results of analysis performed on simulated common CNV events. Panel B shows r^2 values obtained from the simple linear fit (as in Figure 2.2.1B) and panel C shows the r^2 values obtained from the tri-linear fit (as in Figure 2.4.3C). The uni-linear r^2 values deteriorate with the increase of the deleted allele frequency. To the contrary, the tri-linear r^2 values stay relatively high over wide range of the allele frequency. Finally, Figure 2.2.3D demonstrates that the sensitivity of the algorithm to the common CNVs remains relatively stable over wide range of the deleted allele frequency (up to 90%).

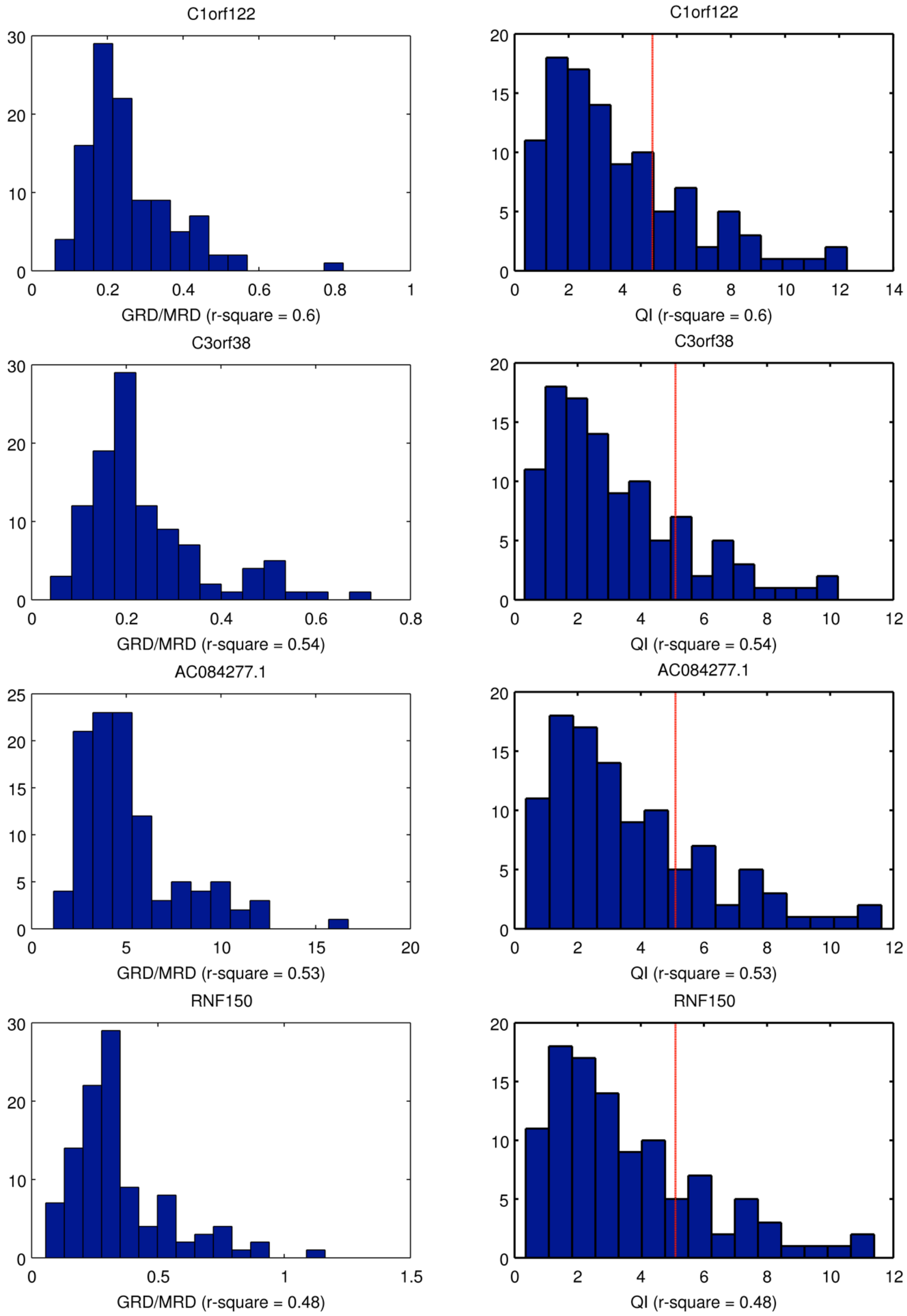


Figure 2.4.2: Analysis of genes that failed simple linear fit. Each row describes a different gene. Left panels — distribution of the ratio of RD at the GSS sites to the sample MRD. Right panels — distribution of the quality index for that gene. The non-multimodal distributions and the low quality-index values of these genes suggest that there are no common CNV events on these loci.

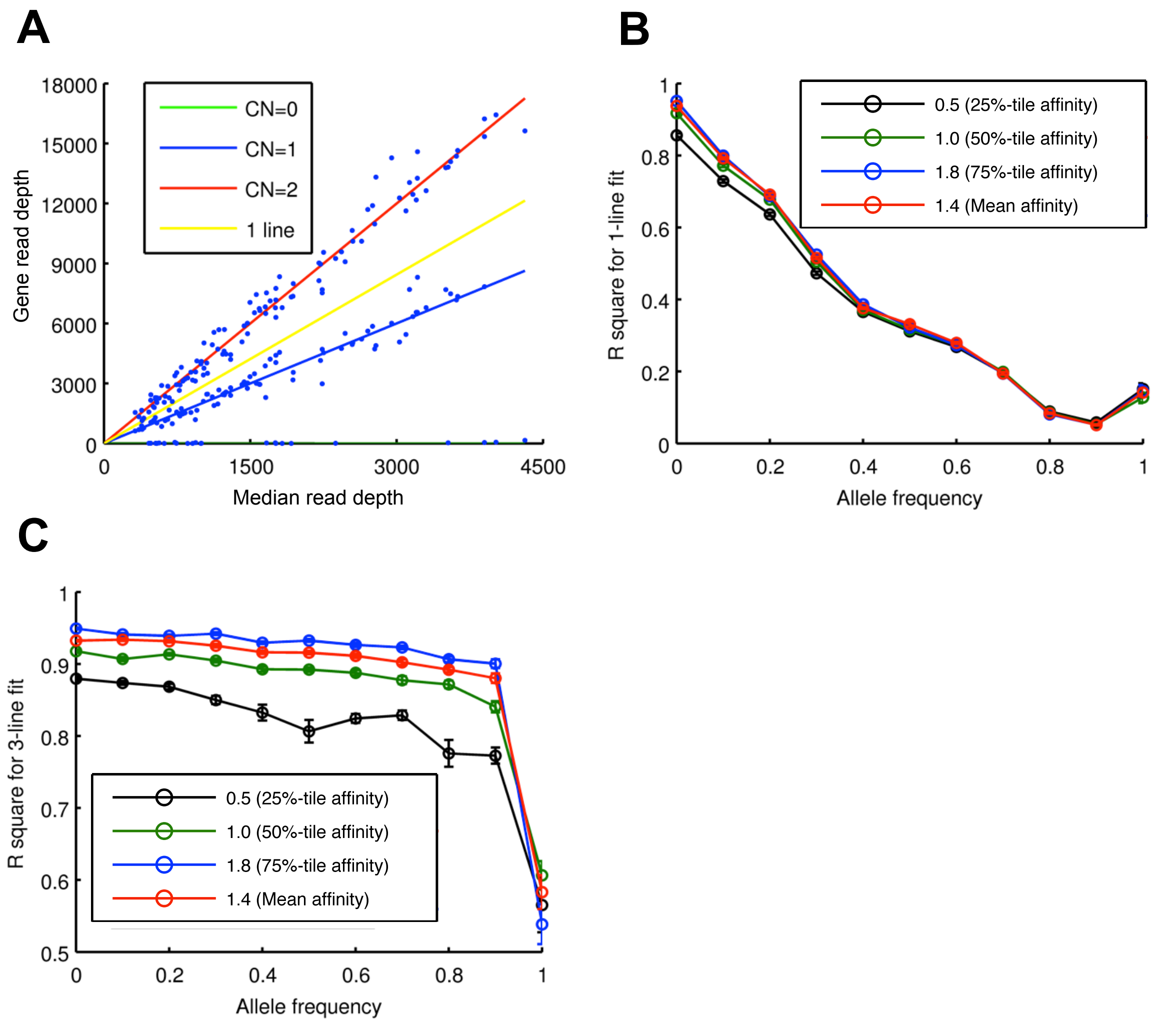


Figure 2.4.3: A. If a simple linear fit fails, the gene affinity is estimated for each gene as the slope of the least-square-error tri-linear fit between MRD and RD for that gene. **B and C.** r^2 values of a simple linear fit (**B**) and a tri-linear fit (**C**) as a function of the deleted allele frequency.

2.4.9 Validation experiments

All primers were designed using Primer3 [99, 100] with default settings to obtain a desired PCR amplicon size between 200 bp and 250 bp. All primers were checked with BLAT [101] to avoid known SNPs that could influence primer hybridization. PCR products were run on an agarose gel to make sure they gave no additional bands besides the expected amplicon.

Primer efficiencies were determined by calculating the standard curve of a serial dilution (4 times, 10-fold) of pooled genomic DNA (Promega, Madison, WI). All experiments were performed in triplicates on the Roche LightCycler 480 platform with LightCycler 480 SYBR Green I Master (cat# 04707516001). The volume of each reaction was 20 μl with final primer concentrations of 400 nM. The PCR was performed according to the following protocol: 5 min at 95 °C, and 45 cycles of 5s at 95 °C, 10s at 60 °C, 30s at 72 °C. To determine the copy number state of an event locus, we used the Delta-Delta-Ct-Method (2-ΔΔCt) for each event locus compared to a reference locus in the sample and a control pool of seven individuals (Promega, Madison, WI), respectively. This reference locus was not previously known to show any copy number variation.

Among the calls made without neighboring information, we exhaustively validated all the calls with posterior probability of 0.95 or more (4 coincided with known events [18]; we experimentally validated the remaining 18 events). Additionally, we performed qPCR validations for 4 events randomly selected from those with posterior probability between 0.65 and 0.95 (2 coincided with known events [18]; we experimentally validated the remaining 2 events).

Of the calls made with the neighboring information, we deemed 7 calls coincided with known events [18]; 7 out of 10 remaining calls were submitted for qPCR validation. For the purpose of validation, the fold change for a given gene <0.7 was classified as a positive validation, >0.8 as a negative validation and in the intermediate range as inconclusive.

If the facts don't fit the theory, change the facts.

Albert Einstein

3

Tangram: An inclusive toolbox for MEI detection

Mobile elements (MEs) constitute greater than 45% of the human genome as a result of repeated insertion events during human genome evolution. Although most of these elements are now fixed in the population, some MEs, including ALU, L1, SVA and HERV-K, are still actively duplicating. Mobile element insertions (MEIs) have been associated with human genetic disorders, including Crohn's disease [102], hemophilia [103], and various types of cancers [104, 105], motivating the need for accurate MEI detection methods. To comprehensively identify and accurately characterize these variants in whole genome next-generation sequencing (NGS) data, a computationally efficient detection and genotyping method is required. Current computational tools [64, 65, 76, 106] are unable to call MEI polymorphisms with sufficiently high sensitivity and specificity, or call individual genotypes with sufficiently high accuracy.

Here we report Tangram, a computationally efficient MEI detector program that integrates read-pair (RP) and split-read (SR) mapping to detect MEI events. By utilizing SR mapping in its primary detection module, Tangram is able to pinpoint MEI breakpoints with single-nucleotide precision. To understand the role of MEI events in disease, it is essential to produce accurate individual genotypes in clinical samples. Tangram is able to predict sample genotypes with very high accuracy. Using simulations and experimental datasets, we demonstrate that Tangram has superior sensitivity, specificity, breakpoint resolution and genotyping accuracy, when compared to other, recently developed MEI detection methods. Tangram serves as the primary MEI detection tool in the 1000 Genomes Project, and is implemented as a highly portable, memory-efficient, easy-to-use C/C++ computer program, built under an open-source development model.

3.1 Introduction

Structural variations (SVs), like single nucleotide polymorphisms (SNPs), are a ubiquitous feature of genomic sequences and are major contributors to human genetic diversity and disease [107–109]. With the advent of next-generation sequencing (NGS) technologies providing vast throughput for individual resequencing, a number of new algorithms have been developed for various SV types, including copy number variations (CNVs) [64–66, 110, 111], and large deletion events [112]. These algorithms take advantage of various signals provided by NGS mapping algorithms primarily read-depth (RD), and read-pair (RP) mapping positions. However, the computational identification of mobile element insertions (MEIs) with NGS data is less well established because mobile elements (MEs) are highly repetitive DNA sequences that are difficult to align against a reference genome with commonly used mapping strategies. MEs have propagated in the human genome through a copy-and-paste mechanism [113–115] and undergone continuous amplification in early primate evolution. Through more than 40 million years of accumulation, MEs account for nearly half of the human genome sequence [116]. Although the current insertion/duplication rate of these elements is substantially reduced, many genetic disorders, such as Crohn's disease [102], hemophilia [103] and cancers [104, 105], have been reported to be associated with their continuing transposition into new genomic locations.

To address effective detection of MEI events we developed an MEI detection pipeline around our SPANNER SV discovery tool [117], and deployed it on the Pilot data of the 1000 Genomes Project (1000GP) [15]. Using this pipeline we compiled the most comprehensive catalog of MEI events in the human genome to date [118]. Although an effective SV detector used extensively in the 1000GP [63], SPANNER only uses RP signal, limiting the precision of breakpoint prediction, detection sensitivity as well as the genotype accuracy that can be achieved.

More recently, three NGS-based MEI detectors, RetroSeq [119], TEA [105] and VariationHunter [120], have been published, each with specific limitations. For example, TEA and VariationHunter do not provide sample genotypes, limiting their use for single-sample detection pipelines *e.g.* in personal genome sequencing projects; or genotype data likelihoods that are essential for phasing structural variants together with SNPs and short INDELs. Also, none of these detectors efficiently integrate the SR and RP signals: VariationHunter detects MEIs using RP signal alone; RetroSeq and TEA only trigger SR analysis when RP signal suggests a potential MEI, and therefore misses events for which only SR evidence is available from the reads (See Table 3.2.1). Because of the steady increase in the read lengths generated by today's sequencing technologies, SR methods are becoming more powerful because these longer reads support confident mapping across SV event breakpoints. Therefore, it is reasonable to expect that using both SR signal and RP signal on an equal footing, as primary observations for "nucleating" SV event calls, will be more sensitive than RP signal alone, or RP signal in combination with a secondary SR search. As a more practical point, the TEA and VariationHunter programs produce reports in non-standard formats, rather than the well established standard VCF format [121], an issue for data communication and downstream analysis. Finally, all the above tools focus on the detection of NON-LTR events, such as ALUs, L1s and SVA, and they do not address the detection of LTRs, such as HERV-K, in the human genome.

3.2 Results

Here we report a fast and convenient MEI detection toolbox, Tangram, which effectively integrates signals provided by both RP and SR mapping. What sets our approach apart from existing methods is the "global" use of SR mapping: we perform a SR mapping step for all orphaned or substantially soft-clipped reads before the detection begins, and therefore both RP and SR mappings are available at the outset, and can nucleate SV event calls. We target both NON-LTR and LTR mobile element types. The global use of SR mapping substantially improves the accuracy of identifying SV event boundaries (breakpoints). Our method produces sample genotypes as well as genotype likelihoods. Unlike other SV detection tools, Tangram is able to detect MEIs for a single individual genome and simultaneously process multiple sequence alignment (BAM) [122] files to call MEI events on population-scale data, and can deal with multiple fragment length libraries and a mixture of read lengths within a single detection step. Tangram is memory and CPU efficient, as analysis is carried out locally *i.e.* event detection in any given region only requires reading the alignment within that region. To our knowledge, there are currently no other detectors that can provide such a comprehensive set of features required for the full characterization of MEIs within a single sample, or a large collection of samples.

3.2.1 Performance evaluation on simulated datasets

We evaluated the detection and genotyping performance of Tangram with a series of in silico experiments involving the insertion of 1,000 full-length AluY elements into the sequence of human chromosome 20 (to closely reflect the real insertion, each inserted AluY element was attached with 15 bp poly-A tails and 15 bp target-site duplication sequence), and generating simulated paired-end sequencing reads of various lengths with realistic base error properties (See Methods 3.4.7). After aligning these reads to the human reference genome sequence using our MOSAIK read mapping program [57], we applied Tangram detect MEI events and to generate sample genotype calls (see Table 3.2.1 and 3.2.2). For comparison, we also ran the RetroSeq program on the same dataset (aligned with the BWA mapping program [58], using default parameters, as instructed by the RetroSeq paper [119]), and compared detection sensitivity and genotyping accuracy, for various read lengths and levels of sequence coverage, considering both heterozygous and homozygous events *i.e.* case where the MEI event is present in one or both chromosome copies within the cell. TEA and VariationHunter do not report sample genotypes, and therefore we did not use these two programs in the comparisons.

As Table 3.2.1 shows, Tangram's sensitivity exceeds 97% both for heterozygous and homozygous events in 10× sequence coverage or greater. Even in low-coverage sequence (5× is the approximate average sequence coverage in the low-coverage 1000GP datasets), Tangram maintains >80% sensitivity. Tangram's sensitivity substantially exceeds that of the RetroSeq program, especially when detecting heterozygous events in low-coverage (5×) data.

We also tabulated genotype calling accuracy *i.e.* the rate at which a given algorithm provides the correct genotype for a given simulated sample (*i.e.* no MEI, heterozygous MEI, homozygous MEI). As Table 3.2.2 indicates, Tangram is able to call sample genotypes with >90% accuracy for all coverage levels and event ploidy we considered. Accuracy in our simulated data is nearly perfect for heterozygous events over 10× coverage, and for homozygous events over 20× coverage. These accuracy values compare very favorably with those obtained for RetroSeq, which appears to heavily favor homozygous calls in low-coverage data, and heterozygous calls in deeper sequence coverage, and has a very high error rate in the non-favored category. The overall accuracy of the Tangram genotypes, obtained by a judicious mixing of heterozygous and homozygous events, is high, over 96%, in every category, again, substantially higher than what was obtained with RetroSeq.

Determining the exact location of SV event boundaries is notoriously difficult. In the simulation experiments performed here, Tangram was able to assign MEI breakpoints at or near single nucleotide resolution using the SR signal. For 106 bp reads, greater than 65% of the

Table 3.2.1: Results are shown for the Tangram and RetroSeq programs applied to simulated data (1,000 ALUY insertions introduced at random positions on human chromosome 20).Simulated reads were generated under: different ploidy values (homozygous or heterozygous), read length (76bp and 106bp) and read coverage (5×, 10×, 20×). The two columns "Sen (RP\SR)" and "Sen (SR\RP)" indicate the sensitivity of the RP and SR methods respectively, when considered in isolation. The best result in each row is indicated in boldface text.(Pldy: Ploidy, RL: Read Length, Cov: Coverage)

Parameters			**Tangram**					**RetroSeq**
Pldy	RL	Cov	Sen(RP)	Sen(SR)	Sen(RP\SR)	Sen(SR\RP)	Sen	Sen
Het	76bp	5×	67.6%	60.0%	25.4%	17.8%	**85.4%**	43.7%
		10×	83.4%	88.9%	8.8%	14.3%	**97.7%**	93.6%
		20×	84.2%	97.8%	1.2%	14.8%	**99.0%**	98.9%
	106bp	5×	45.1%	67.3%	13.9%	36.1%	**81.2%**	12.0%
		10×	77.0%	93.0%	4.5%	20.5%	**97.5%**	68.9%
		20×	83.4%	98.9%	0.4%	15.9%	**99.3%**	97.7%
Homo	76bp	5×	83.4%	88.9%	8.8%	14.3%	**97.7%**	95.2%
		10×	84.2%	97.8%	1.2%	14.8%	**99.0%**	98.8%
		20×	84.6%	99.1%	0.4%	14.9%	**99.5%**	99.2%
	106bp	5×	77.0%	93.0%	4.5%	20.5%	**97.5%**	68.9%
		10×	83.4%	98.9%	0.4%	15.9%	**99.3%**	97.7%
		20×	83.8%	99.3%	0.4%	15.9%	**99.7%**	98.9%

Table 3.2.2: For each simulated dataset corresponding to a specific read length and coverage, we randomly chose 500 MEI loci. 400 were designated as heterozygous sites, and 100 as homozygous sites. The genotype accuracy was then calculated for these loci. The random selection and genotype accuracy experiment was then repeated five times (to give a sample of 2,500 MEI loci) and the overall genotype accuracy was determined by averaging the results of the five experiments. The best result in each row is indicated in boldface text. (RL: Read Length, Cov: Coverage)

Parameters		**Tangram**			**RetroSeq**		
RL	Cov	Het	Homo	Total	Het	Homo	Total
76bp	5×	99.3%	90.8%	**97.6%**	2.3%	92.8%	20.4%
	10×	100.0%	94.2%	**98.8%**	40.6%	63.6%	45.2%
	20×	100.0%	98.4%	**99.7%**	96.5%	8.8%	78.9%
106bp	5×	96.6%	93.4%	**96.0%**	0.0%	91.6%	18.3%
	10×	99.6%	92.6%	**98.2%**	38.8%	64.4%	43.9%
	20×	100.0%	95.6%	**99.1%**	95.1%	10.8%	19.6%

reported breakpoints co-locate exactly with, and over 99% are within 15 bp of the true breakpoints (see Figure 3.2.1). This performance is attributable to SR-mapped reads identifying the breakpoints at a resolution that RP-only methods are unable to match.

3.2.2 Performance comparisons using 1000 Genomes Project data

We ran Tangram and two other MEI detection algorithms, RetroSeq and TEA, to analyze deep-coverage sequencing data from a CEU trio consisting of samples NA12878 (89×), NA12891 (78×) and NA12892 (78×), obtained from the public 1000GP ftp site. The DNA of these individuals were collected from fresh blood cells. All people who contributed their DNA to this project are anonymous and have no phenotype data available. Trio data were sampled from mother-father-adult child families. The detailed data collection guideline can be found from the supplemental information of [123]. The data consists of 101 bp paired-end reads generated by Illumina HiSeq sequencing machines; insert size was 465 ± 50 bp (median ± standard deviation). We mapped the reads with MOSAIK 2.0 [57] for Tangram and BWA [58] for RetroSeq and TEA,

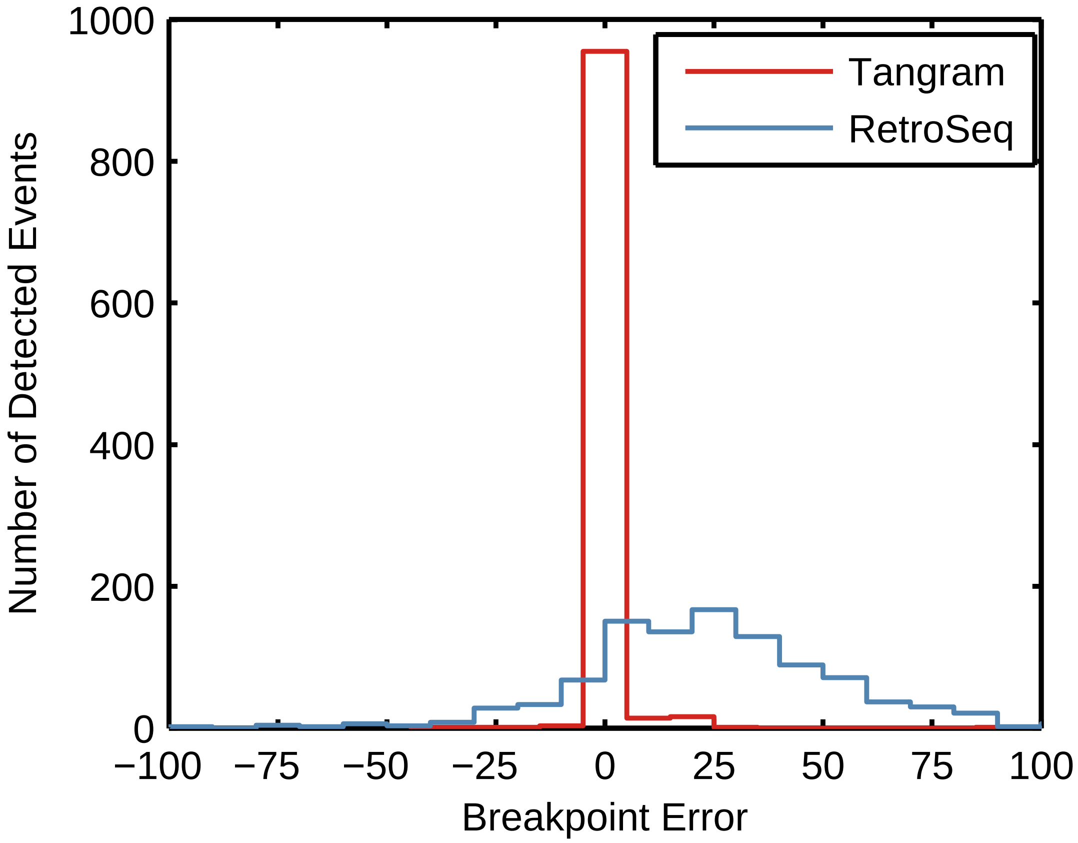


Figure 3.2.1: Breakpoint resolution of Tangram and Retroseq. The difference between reported and true breakpoint position in simulated data is shown for the Tangram and the RetroSeq MEI detection algorithms (homozygous events in 76 bp paired-end reads, 20× sequence coverage). The majority of breakpoints reported by Tangram exactly match the true breakpoint.

according to author instructions. To assess sensitivity and genotype accuracy, we compared the MEI loci (ALU and L1) reported by the three detectors to the events reported and experimentally characterized in a previous large-scale study [118] using an earlier set of 1000GP data for the same samples (characteristics of this dataset from the 1000GP Pilot 2 trio data are reported in Table 3.2.3). The Stewart *et al.* 2011 [118] callset consisted of 1,208 Alu and 180 L1 calls, including 486 Alu and 48 L1 insertions that were experimentally confirmed with a PCR-based validation technique. As shown in Table 3.2.4, Tangram recovered >98% of PCR validated events and > 93% of all reported events. RetroSeq provided comparable results, but TEA was unable to achieve this level of sensitivity to ALU events. Tangram's genotype accuracy for ALU events was > 91% for all three samples. Tangram detected approximately 87% of PCR validated L1 insertion events, outperforming the two competing algorithms. Tangram's sensitivity to L1 events reported in the Stewart *et al.* 2011 data set drops markedly in comparison to the PCR-validated events. This is likely the result of the high false discovery rate (FDR) for L1 events (18.8%) in the Stewart *et al.* 2011 data set. Notably, our algorithms called none of the events reported in the Stewart *et al.* 2011 dataset that failed PCR validation. It is notable that sample NA12878 had the highest number of MEI calls using either of the calling methods. This is likely the result of the substantially higher read coverage in this sample, as well as longer reads from 454 sequencing machines, not available for the other two samples (Table 3.2.3).

Table 3.2.3: Sequence coverage (base coverage) for two sequencing technologies (454 and Illumina) of CEU trio (NA12878, NA12891 and NA12892) used in 1000GP Pilot MEI paper [118].

Samples	454	Illumina
NA12878	11.0×	15.9×
NA12891	0.0×	14.9×
NA12892	0.0×	9.2×

Our experiments here demonstrate that Tangram provides accurate MEI genotypes across all MEI types (see Table 3.2.5). The TEA program does not provide sample genotypes, and

Table 3.2.4: Comparisons are shown for a CEU trio (NA12878, NA12891 and NA12892) processed with Tangram, RetroSeq and TEA.Sensitivity and genotype accuracy was measured by comparing the reported events with those in Stewart *et al.*, 2011 [118]. The total number of validated and reported MEI loci are shown under the "Stewart *et al.* 2011" column. The two sub columns under each detector, "Validated" and "Reported", show the sensitivity to PCR validated loci and all reported loci in Stewart *et al.* 2011, respectively. The TEA program does not provide genotype calls, and therefore could not be used for genotype accuracy comparisons. The best result in each row is indicated in boldface text.

	Stewart *et al.* 2011			**Tangram**			**RetroSeq**			**TEA**	
	Sample	Loci		Sensitivity		Genotype	Sensitivity		Genotype	Sensitivity	
		Validated	Reported	Validated	Reported		Validated	Reported		Validated	Reported
ALU	NA12878	408	965	**98.8%**	**93.0%**	**95.0%**	94.10%	87.70%	76.40%	89.50%	82.20%
	NA12891	309	675	98.1%	96.3%	**91.2%**	**98.40%**	**96.40%**	67.90%	96.10%	93.80%
	NA12892	312	650	98.1%	96.9%	**92.6%**	**99.00%**	**97.40%**	71.20%	94.20%	92.50%
L1	NA12878	38	157	**86.8%**	**52.2%**	**87.5%**	78.90%	45.80%	83.30%	84.20%	49.70%
	NA12891	26	64	**92.3%**	**75.0%**	**100.0%**	76.90%	64.10%	66.70%	84.60%	70.30%
	NA12892	34	76	**94.1%**	**76.3%**	**85.7%**	79.40%	65.80%	50.00%	76.50%	64.50%

therefore we were not able to include it in this comparison. RetroSeq appears to suffer from a systematic bias when applied to deep-coverage data: it called almost all MEI loci as heterozygous. In comparison, Tangram can effectively distinguish between homozygous and heterozygous loci.

3.2.3 Running Tangram on population data

We deployed Tangram on 218 samples from the 1000GP Phase 1 release [123]. Three populations were included in this dataset: ASW (people with African ancestry in Southwest United States, 50 individuals), LWK (Luhya in Webuye, Kenya, 83 individuals) and YRI (Yoruba in Ibadan, Nigeria, 85 individuals). These data were collected with same strategy as the deep-coverage trio data mentioned above. However, the sequencing coverage for these samples is much lower. On average, each sample had 5× sequence coverage so the overall coverage of this dataset is ~1,000×. The allele frequency spectrum (AFS) of all MEIs for each of the three populations (4,085 ALU, 1,548 L1, 88 SVA and 44 HERV-K insertions) is shown in Figure 3.2.2. The expectation is that the AFS of MEIs is similar to AFS observed from SNP data [118]. This is indeed the case, except at very low allele frequency, where detection sensitivity drops off in the low-coverage 1000GP datasets (as there may be too few RP and/or SR mapped reads supporting

Table 3.2.5: A contingency table is shown for MEI genotypes reported by Tangram and RetroSeq on deep coverage sequencing data from a CEU trio (NA12878, NA12891 and NA12892).The “Genotype from validation” column shows the genotype that was validated in Stewart *et al.* 2011 [118]. The “Genotype call” column shows the genotype predicted by Tangram and RetroSeq at the same loci. The “Genotype” column in Table 3.2.4 was calculated based on the results in this table.

			Tangram		**RetroSeq**	
		Genotype from validation	Genotype call		Genotype call	
			Het	Homo	Het	Homo
ALU	NA12878	Het	120	8	119	0
		Homo	1	26	37	1
	NA12891	Het	95	13	93	0
		Homo	0	40	44	0
	NA12892	Het	106	11	104	0
		Homo	0	32	42	0
L1	NA12878	Het	5	1	4	0
		Homo	0	2	1	1
	NA12891	Het	4	0	2	0
		Homo	0	2	1	0
	NA12892	Het	3	1	3	0
		Homo	0	3	3	0

Table 3.2.6: Genomic distribution of MEI events detected from the AFR dataset.

Genomic Region	Number of MEIs
Intergenic	3,249
Intron	2,439
3' UTR	54
5' UTR	23
Exon	0

a MEI event). The genomic distribution of these 5,765 MEI events is shown in Table 3.2.6. Most of the detected MEI events (98.7%) fall into the intergenic and intronic regions whereas none of the events are found in the exon regions. This observation is very similar to the results from Stewart et al. 2011 [118]. The absence of MEI events in exonic regions could be attributed to the selection pressure since such long insertion events could substantially interrupt the transcription process (See Discussion 4.1.3).

3.2.4 Experimental validation

To assess the specificity of Tangram, researchers (Dr. Miriam Konkel and Dr. Mark Batzer) from Louisiana State University helped us perform the PCR validation experiment on 23 1000GP Phase 1 [123] samples (Table 3.2.7), including a CEU trio (NA12878, NA12891 and NA12892) with deep coverage (~20×) and 20 low-coverage (~5×) samples from the CHS and LWK populations. Tangram detected 2,874 ALU, 256 L1, 53 SVA and 22 HERV-K insertions in these samples. Of the 3,205 loci, 357 were novel, *i.e.* not reported in previous studies [118, 124–130], and absent from the dbRIP database [131]. Two random subsets, 160 sites in all, were randomly selected for PCR validation: (1) 80 loci (66 known + 14 novel) were randomly selected from the entire callset of 3,205 MEIs; and (2) additional 80 loci were randomly selected only from the novel 357 novel calls. PCR validation results for Tangram and VariationHunter are shown in Table 3.2.8 and Table 3.2.9. Tangram achieved very low FDR for all three non-LTR MEI types (<6%). Although the numbers are low, no false positive L1 and SVA calls were reported. The

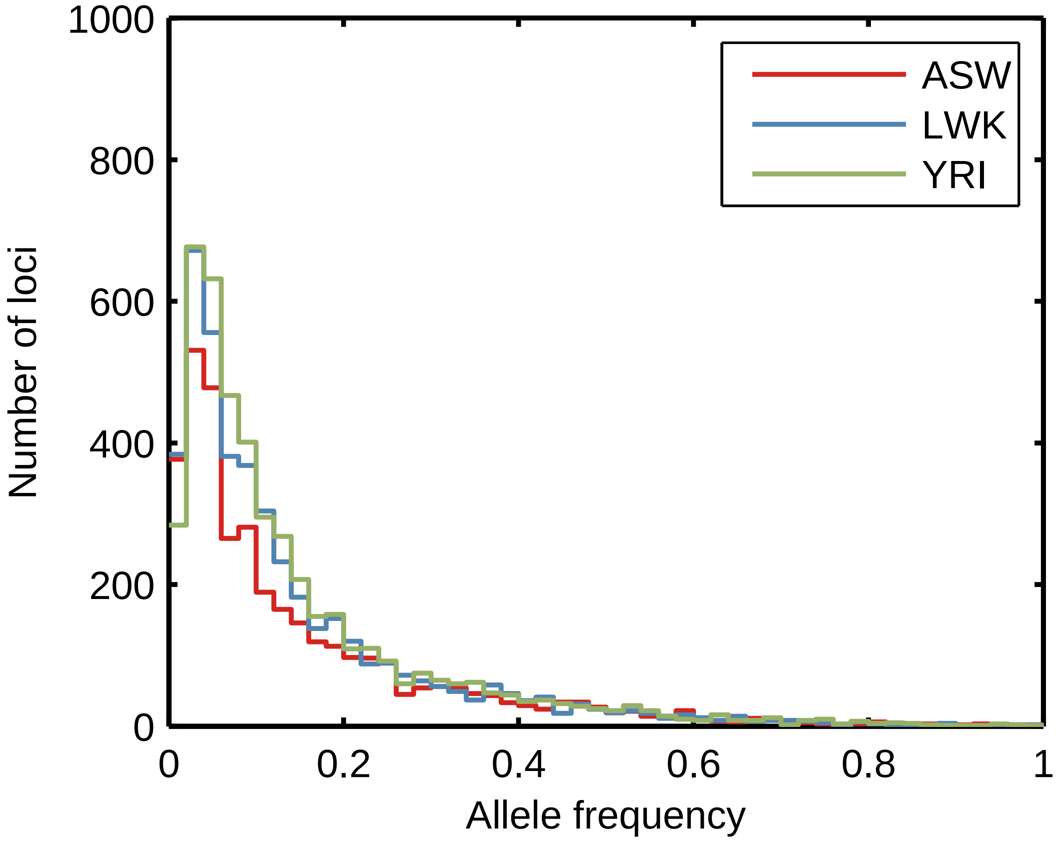


Figure 3.2.2: Allele frequency spectrum for MEI variants detected in 3 African populations. Results for samples designated as ASW, LWK and YRI are shown, for 4 types of MEIs: ALU, L1, SVA and HERV-K. There is limited sensitivity to low frequency events because of sparse or absent supporting reads in low-coverage data.

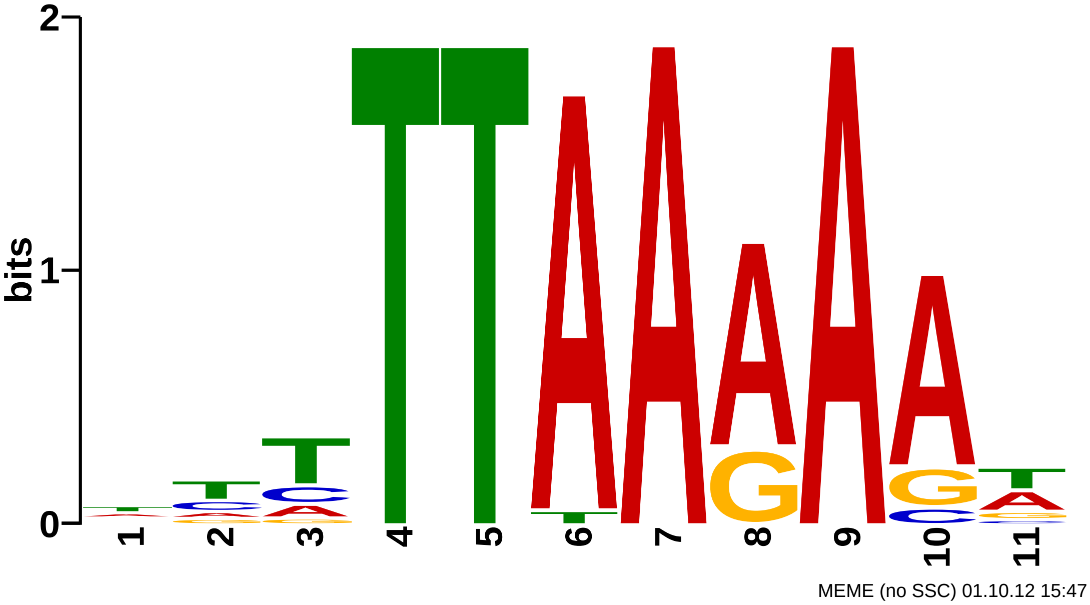


Figure 3.2.3: Motifs reported by MEME software [132] by using sequences (±25bp) around the ALU and L1 breakpoints detected by Tangram in 23 1000 Genome Project Phase 1 samples. They are highly consistent with the canonical ALU and L1 recognition motifs.

overall estimated FDR for the first and second validation sets were 2.53% and 9.21%, respectively. This result is consistent with expectations that newly detected, previously unknown events have higher FDR. In Table 3.2.9, we compared experimental validation results for three algorithms: Tangram, RetroSeq, and VariationHunter, for event types detected by each calling algorithm. Tangram achieves substantially higher specificity than the two competing algorithms. In fact, this level of accuracy is comparable to or better than the FDR of SNP calls from current state-of-the-art variant callers [123].

Consistently with the validation results, a copy of the canonical ALU and L1 recognition motif, `5'-TTAAAAA-3'`, was found within a 25 bp window of all reported breakpoints (Figure 3.2.3), further confirming the high specificity of our detection method.

Table 3.2.7: Samples and sequence coverage of CEU trio and 20 1000GP phase I samples used for PCR validation

Sample	Population	Platform
NA19397	LWK	ILLUMINA
NA19398	LWK	ILLUMINA
NA19399	LWK	ILLUMINA
NA19404	LWK	ILLUMINA
NA19428	LWK	ILLUMINA
NA19429	LWK	ILLUMINA
NA19434	LWK	ILLUMINA
NA19435	LWK	ILLUMINA
NA19440	LWK	ILLUMINA
NA19443	LWK	ILLUMINA
HG00662	CHS	ILLUMINAHiSEQ
HG00663	CHS	ILLUMINAHiSEQ
HG00671	CHS	ILLUMINAHiSEQ
HG00672	CHS	ILLUMINAHiSEQ
HG00683	CHS	ILLUMINAHiSEQ
HG00684	CHS	ILLUMINAHiSEQ
HG00689	CHS	ILLUMINAHiSEQ
HG00690	CHS	ILLUMINAHiSEQ
HG00464	CHS	ILLUMINAHiSEQ
HG00614	CHS	ILLUMINAHiSEQ
NA12878	CEU	Multiple
NA12892	CEU	Multiple
NA12891	CEU	Multiple

Table 3.2.8: PCR validation results for the Tangram MEI detector. Validation results and estimated false discovery rates are shown for MEI calls from 23 1000 Genomes Project Phase 1 samples.

	ALU		**L1**		**SVA**		**HERV-K**		**Total**	
	Random	Novel	Random	Novel	Random	Novel	Random	Novel	Random	Novel
Analyzed by PCR	68	64	7	3	3	6	1	3	80	78
Validated Loci	66	58	7	3	3	6	1	2	77	69
Invalidated Loci	2	6	0	0	0	0	0	1	2	7
FDR	2.94%	9.38%	0.00%	0.00%	0.00%	0.00%	0.00%	33.33%	2.53%	9.21%

Table 3.2.9: Comparison of PCR validation results across three MEI detection algorithms. Calls were made in 23 1000 Genomes Project Phase 1 samples by Tangram, RetroSeq and VariationHunter.The best result is indicated in boldface text.

	Tangram			**RetroSeq**			**VariationHunter**		
	Random	Novel	Combined	Random	Novel	Combined	Random	Novel	Combined
Analyzed by PCR	80	78	158	80	80	159	83	51	134
Validated Loci	77	69	142	73	58	131	69	29	98
Invalidated Loci	2	7	9	7	21	28	14	22	36
FDR	**2.53%**	**9.21%**	**5.96%**	8.75%	26.58%	17.61%	16.87%	43.14%	26.86%

3.2.5 Resource requirements and software availability

The primary motivation behind developing Tangram was to provide highly accurate MEI calls. To be a useful software tool, however, it must be easy to install, easy to run, and generate results in a timely fashion, using reasonable computational resources. We characterized resource usage and analysis time on our analysis of the 218 1000GP low-coverage samples described [123]. When using other MEI detection software, it is a common requirement that only a single BAM file can be processed at a time, necessitating all input BAM files to be merged into a single file (a lengthy task), or to process each BAM file individually (reducing sensitivity to low-frequency events). Tangram, in contrast, can process all input BAM files simultaneously. Most currently available structural variant callers employ multiple passes through the entire input file, requiring substantial memory and computation time. To reduce the memory footprint and increase the throughput, Tangram was designed to call MEI events regionally, *i.e.* within shorter windows of the sequence alignment. Single-pass analysis is made possible by annotation tags produced by our MOSAIK read mapper software [57], marking reads whose fragment-end paired mate maps into ME reference sequence. Additional parallelization was accomplished by multi-threaded implementation of the software. In this test, we submitted one Tangram detection job for each chromosome (Chr1-ChrX). Each job used one AMD Opteron 6134 CPU (8 cores at 2.3GHz). The detection process finished within 58 hours (wall time) or 96 hours (CPU time). Repeating the detection process in 1 Mbp detection windows on the same cluster resource requires 0.24 hours (wall time) or 0.40 hours (CPU time).

Tangram is easy to install and run. Users can download it from its main github repository (https://github.com/jiantao/Tangram). We have also integrated it into our pipeline and tool launcher system, GKNO, available at http://gkno.me.

3.3 Discussion

Many MEI events have strong impact on gene function and they are therefore essential to accurately detect and genotype within individuals. Mobile elements are, by nature, repetitive sequences and are therefore difficult to detect. To our knowledge, our Tangram software is the only robust software capable of detecting all classes of MEIs, providing accurate individual genotype information, and accurate, near base-perfect breakpoint localization. We believe that Tangram can achieve higher sensitivity, specificity, genotyping accuracy, and breakpoint calling accuracy than competing MEI detection methods because of the global use of split-read mapping information into the detection process. Competing algorithms either only use RP mapping information to call events, or perform SR mapping in regions where RP mappings indicate a possible MEI events. In contrast, Tangram analyses both RP and SR mapped reads from the start, and can therefore detect events for which only SR mapping evidence exists.

Table 3.2.1 illustrates detection sensitivity when RP or SR signal is used in isolation, or in combination with each other. At almost all read length and coverage values, the SR method on its own is more sensitive that the RP method (except for low, 5 × coverage in 76 bp reads). Importantly, RP detection sensitivity does not exceed 85%, even in deep-coverage data. This is because RP-mapped reads localize the ME insertion point to a window. If the reference sequence already contains a ME within this window, one must filter out the candidate event because of the high likelihood of spurious detection. SR mapping localizes the insertion site with much greater resolution, making it possible to distinguish between ME elements in the reference, and polymorphic insertions not present in the reference.

Table 3.2.1 also illustrates that RP based methods that use a secondary SR mapping step can perform very well in deep sequencing data because in such high-coverage datasets there are likely read pairs mapping across the breakpoints, and then additional reads that can be SR-mapped across the breakpoint for fine localization. In low-coverage data however, there are many events without read pairs mapping across the breakpoints. When using shorter reads, reliable SR mapping becomes difficult. In both cases, sensitivity suffers. As through technology development read lengths increase, the same sequence coverage will be accomplished with fewer, but longer, reads. Moving forward, this trend clearly favors SR mapping methods, and in particular, methods that use SR mapping as part of their primary detection approach. As we demonstrate in this study, such methods are more sensitive and specific, have higher genotype accuracy, and are able to localize event boundaries more accurately.

Our MEI detector program, Tangram is a fast, accurate tool that has been extensively tested and benchmarked in the analysis of the 1000GP sequencing datasets. It is easy to install, easy to use, and is available as a stand-alone package or as part of our tool and pipeline launching system, making it especially useful for medical or population sequencing projects.

3.4 Methods

3.4.1 The Tangram detector — algorithmic overview

As input, Tangram uses reads aligned to the genome reference sequence as well as to mobile element reference sequences, available in BAM format alignment file(s). Currently, alignments to ME reference sequences can be produced by the MOSAIK mapping software (version 2.0 or above) [57]. Tangram's RP detection module first scans the alignment for read pairs where one mate uniquely aligns to the genome reference, and the other mate maps to a ME reference sequence (Figure 3.4.1A). Second, read pairs where one mate is aligned to the genome reference uniquely (*i.e.* with high read mapping quality value, or MQ), but the other mate either soft-clipped or entirely unaligned, are collected as the starting material for SR mapping (Figure 3.4.1B). The SR module attempts to align these soft-clipped or unaligned mates both the genome reference and to the ME reference sequences in a split fashion (*i.e.* aligning one section of the read to the genome reference and another section to the ME reference). Loci in the genome with either RP or SR evidence for a candidate MEI event are then extracted. Candidate events are filtered on the number and type of supporting fragments. A genotyping module produces

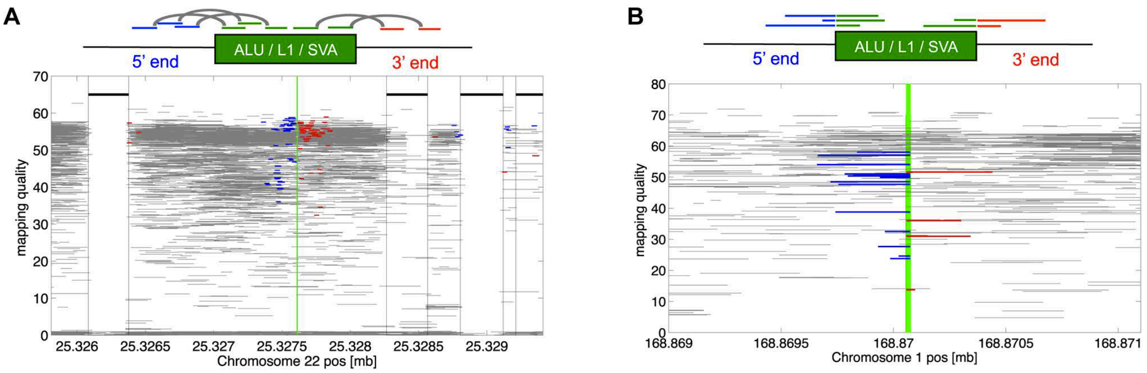


Figure 3.4.1: Illustration of MEI detection algorithms in Tangram. **A.** MEI detection with RP signal: RP algorithm will cluster those read pairs with one mate uniquely aligned to the normal reference (5' – blue and 3' – red) and the other mate aligned to the MEI special references (green). **B.** MEI detection with SR signal: SR algorithm will search for unaligned or soft-clipped reads (crossing the breakpoint from 5' – blue or 3' – red) and align these reads to both the normal reference and the special MEI reference (green) after splitting them into two subsections. Reprinted from [118] with permission.

individual genotype likelihoods and calls sample genotypes. A reporting module produces a VCF format variant report including the location and type of the events, as well as individual sample genotype information.

3.4.2 Sequence alignment to genome and mobile element reference

Alignments were created with the MOSAIK program, a hash-based read mapper that is aware of user-specified insertion sequences, *e.g.* MEIs. When the insertion sequences are provided, the reference hashes are prioritized such that alignment to the MEI sequences are attempted prior to alignment to the genome reference. Since MEIs are repetitive elements, a read from an MEI can be mapped to several locations within the genome (potentially hundreds of locations). An additional tag in the BAM file (the ZA tag) is then populated with information about the reads mate, including location, mapping quality and number of mapping locations for the mate. This information ensures that BAM search operations (which can be lengthy for large alignment files) can be avoided.

3.4.3 MEI detection based on read-pair mapping positions

Tangram first establishes the fragment length distribution for each library in the input BAM files using “normal” read pairs (*i.e.* those read pairs where both mates are uniquely aligned to the same chromosome with expected orientation). Tangram then searches the BAM files for MEI-candidate read pairs that have one mate uniquely aligned to the reference genome and the other aligned to a ME reference. Such read pairs must also satisfy one of the following three requirements: (1) they do not have the expected orientation; (2) they are not aligned to the same chromosome or (3) the fragment length is not consistent with the fragment length distribution (*p-value* ≤ 0.005). For each type of ME (ALU, L1, SVA and HERV-K), Tangram clusters these candidate read pairs with a customized nearest-neighbor algorithm [133, 134] according to their fragment center position (aligned position of the uniquely aligned mate plus one half of the median of the fragment length distribution). During this process read pairs cluster with other read pairs within a range determined by the fragment length distribution. This algorithm can handle candidate read pairs from different libraries and samples effectively, which can significantly improve the sensitivity for multiple low-coverage samples. Also, the complexity of this algorithm is linear in the number of candidate read pairs, making it suitable for large-scale sequencing data. Read pairs that span into MEs from the 5’ end will be clustered separately from those spanning in from the 3’ end. Tangram will identify an MEI event if a pair of clusters in the MEI neighborhood range span into the insertion from both the 5’ and 3’ ends (Figure 3.4.1A). The true breakpoint should locate somewhere between the end of the 5’ cluster or the beginning of the 3’ cluster. Tangram reports the estimated breakpoint following a leftmost convention (smallest genomic coordinate of the two positions).

3.4.4 MEI detection based on split-read mapping positions

We used the Scissors software [135], both a stand-alone split-read mapping program, and a library providing an application programming interface (API) to its functions. Scissors uses a uniquely aligned mate and the fragment length distribution to identify a candidate genomic region for aligning an unaligned/soft-clipped mate (Figure 3.4.1B). The alignment is performed using a sensitive and fast algorithm, single instruction multiple data Smith-Waterman (SIMD SW) [136]. Several candidate alignments may be obtained in this step, each of which may have a different segment of the read successfully aligned. The unaligned/soft-clipped read is then aligned to the MEI reference sequences, using the SIMD SW algorithm (Figure 3.4.1B). This step may again yield several candidate alignments. After obtaining the candidate alignments, Scissors calculates a score for each, based on the number of mapped bases and the number of mismatches. In our application, we use the best SR alignment *i.e.* the alignment with the highest score.

3.4.5 Candidate MEI event filtering and post-processing

The MEI candidates are first filtered using the number of supporting fragments. An MEI candidate with at least two RP supporting fragments from both 5' and 3' or at least two SR supporting fragments were retained. Candidates that are supported by RP signal alone undergo additional filtering. If the candidate MEI falls within a predefined distance of a locus annotated in RepeatMasker [137] downloaded from UCSC Genome Browser [138] they are removed from the candidate list. The distance used is the approximate maximum expected fragment length ($p\text{-}value \approx 0.005$) in the clusters of supporting RP fragments. For ALU and HERV-K events, the candidate call is only filtered out if the MEI in RepeatMasker is also an ALU or HERV-K event. L1 and SVA elements are filtered out if they also co-locate with an L1, SVA or ALU event in RepeatMasker. For MEI events supported by SR signal, no further filtering steps will be applied. All remaining MEI candidates will be reported in the final VCF file. These filtering steps can be performed using the PERL program (`tangram_filter.pl`) that is included in the toolbox.

3.4.6 Sample genotype calling and genotype likelihood calculation

Tangram uses a Bayesian framework to predict the genotype of MEI events [118]. We calculate the posterior probability of a given sample MEI genotype g (*i.e.* monomphic: REF/REF;

heterozygous MEI: REF/MEI: or homozygous MEI: MEI/MEI) as follows:

$$P(g|D) = \frac{P(g)P(D|g)}{\sum_{g'} P(g')P(D|g')} \tag{3.1}$$

where D is the observed read evidence at the site; and $P(g)$ is the prior probability of the genotype. By default, Tangram will set a flat prior probability (1/3) for all three possible genotypes. The data likelihood, $P(D|g)$, is calculated as a binomial probability with the following parameters:

$$P(D|g) = p_{bin}(N_{alt}, N_{alt} + N_{ref}, p_g) \tag{3.2}$$

where p_g is the expected ratio of MEI alleles to the total number of fragments (~0 for homozygous reference, 0.5 for heterozygous MEI and ~1 for homozygous MEI); N_{ref} and N_{alt} are the numbers of read pair fragments that support reference and MEI (alternate) alleles, respectively. Reference and MEI alleles are defined as follows: any uniquely mapped read pairs spanning the predicted breakpoint with a consistent insert size and orientation will be counted as a fragment supporting the reference allele. Fragments supporting an alternate allele (insertion) are those inconsistent with the conditions for a reference allele collected during the detection step (both RP and SR signal). The meaning of the data likelihood is the binominal probability that $N_{ref} + N_{alt}$ will fluctuate to N_{alt}, given the expected p_g.

The genotype reported by Tangram is that with the highest posterior probability and the output VCF file is populated with the corresponding data likelihoods.

3.4.7 Simulation data generation

1,000 full-length ALUY elements with a 15 bp poly-A tail and a 15 bp target-site duplication (TSD) sequence were randomly introduced into chromosome 20. No elements were allowed to insert within a 100 bp window of the reference MEs or other simulated elements. Simulated Illumina paired-end reads were generated for both heterozygous and homozygous insertions, with two different read lengths (76 bp and 106 bp) and three different coverages (5×, 10× and 20×) using the MASON read simulator [61] with the default error model. This led to 12 different different sets of simulated data. All of the simulated reads had a 500 bp ± 100 bp (median ± standard deviation) insert size. MOSAIK 2.0 [57] with default parameters was used to align these simulated reads against a customized human reference that combined hg19 and 23 ME sequences (4 ALU, 17 L1, 1 SVA and 1 HERV) downloaded from RepBase [139]. The output BAM files from MOSAIK were sorted by genomic coordinates using BamTools [140]. The final BAM files served as the input to Tangram for MEI discovery and genotyping.

3.4.8 Genotype mixing

For each dataset corresponding to a specific read length and coverage, we randomly chose 500 MEI loci. 400 were designated as heterozygous sites, and 100 as homozygous sites (the 4:1 ration was based on experimentally validated genotypes from our earlier study, Stewart *et al.* 2011 [118]). The genotype accuracy was then calculated for these loci. The random selection and genotype accuracy experiment was then repeated five times (to give a sample of 2,500 MEI loci) and the overall genotype accuracy was determined by averaging the results of the five experiments.

3.4.9 Alignments for RetroSeq

RetroSeq calls were based on BWA [58] alignments with default parameters as suggested in the RetroSeq publication.

3.4.10 Identification of events across MEI callsets

In this experiment, we report a detected MEI event as a match to the locus in Stewart *et al.* 2011 [118], if the two events are within 500 bp of each other. This criterion is a result of the large breakpoint uncertainty in Stewart *et al.* 2011.

3.4.11 PCR validation

Two sets of 80 loci each were selected for PCR validations from the whole dataset of candidate loci containing ALU, L1, SVA, and LTR elements. The first set contained loci from the whole dataset while the second one included only loci identified as novel based on previous studies [118, 124–130] and the dbRIP database [131]. Due to the nature of paired-end reads and low coverage data, breakpoint coordinates for MEIs were commonly not available. Thus, an insertion range was provided for each locus within which the MEI was predicted. For primer design, 600 bp of flanking sequence were added upstream and downstream of the insertion coordinates. The sequence was extracted from the human reference genome (hg19) using Galaxy [141–143].

ALU elements were masked using RepeatMasker [137]. After adding a safety margin of 50 nucleotides up- and downstream of the insertion coordinates, primers were selected using BatchPrimer3 v2.0 [144]. The uniqueness of each primer was determined using BLAT [101]. An in silico PCR was performed for each locus when at least one primer had more than one match. If several matches were identified or the in silico PCR provided evidence for more than one PCR product primers were manually redesigned. In these cases the repeat content of the flanking sequence was determined using RepeatMasker. Moreover, the flanking sequence was "Blatted" against the human reference genome (hg19) to determine if the flanking sequence matched to highly homologous loci. In cases with high sequence homology, the other orthologous sequences were retrieved using the UCSC genome browser [138]. Following an alignment of the candidate locus with the other orthologous loci using BioEdit [145] primers design was attempted in regions with sequence divergence between the different loci. All manually designed primers were tested with Primer3 [146]. For loci with ambiguous PCR results, no amplification, or amplification of only the empty insertions site, a second primer pair was designed using the same primer design criteria described above.

Due to the size and high GC-content of SVA elements we used previously designed internal PCR primers [118]. The internal primers were designed within the 3' end of the SVA sequence matching the consensus sequences of the youngest SVA subfamily (SVA_F) which is human-specific. All PCR primers were ordered from Sigma Aldrich, Inc. (St. Louis, MO). The PCR primer sequences used in this validation study are available at http://batzerlab.lsu.edu.

3.5 Software availability

The source code and instruction are available at https://github.com/jiantao/Tangram. Our pipeline and tool launcher system, GKNO, available at https://github.com/gkno.

Reviewing what you have learned and learning anew, you are fit to be a teacher.

Confucius

4

Concluding Remarks

STRUCTURAL VARIATIONS are now recognized as one of the major contributors to human diseases and phenotypic variants. In order to enable downstream functional studies about these variants, it is first necessary to establish reliable methods to detect them. Current excitement surrounding the SV discoveries mainly stem from the advent of NGS sequencing technologies. The focus of my PhD study in the Marth lab is to develop efficient and lightweight computational methods for SV detection in the human genome based on NGS data.

4.1 Summary

4.1.1 CNV detection from exon capture sequencing data

DNA capture technologies combined with high-throughput sequencing now enable cost-effective, deep-coverage and targeted sequencing of complete exomes. This is well suited for SNP discovery and genotyping. However, there has been little attention devoted to CNV detection from exome capture datasets despite the potential impact on the protein function for CNVs in exonic regions.

To fill this gap, I developed a computational method based on the RD signal to identify CNVs in exon capture sequencing data. I first established a mathematical model to calculate the expected number of reads for each target region (gene), which is one of the most difficult problems in the CNV detection from capture sequencing data. This model does not only normalize the read depth signal from sample to sample (sample specific median read depth) but also from gene to gene (gene affinity). With the expected read depth, I can calculate the data likelihood of each gene-sample site (GSS) and each possible genotype based on the *Poisson* (Normal) distribution with a correction factor (ODF) accounting for the random noise and PCR bias. I plugged these data likelihoods to a Bayesian framework to calculate the posterior probability for each possible copy number. CNVs can be detected as those GSS whose largest posterior probability is not from copy number 2.

I evaluated this algorithm on 1000GP exon capture sequencing data generated from four sequencing centers. Totally my program detected 96 heterozygous deletions and 39 duplications from about 4.6% of the human exome (Table 2.2.3, 2.2.4, 2.2.5 and 2.2.6). Due to the limitation of the data quality, the estimated detection efficiency from both mathematical derivation and simulation experiments is about 50%. I derived a statistical measurement, quality index (QI), to describe the relationship between the quality of sequencing data (coverage and ODF) and the estimated detection efficiency. From the calculation, I found the detection efficiency of my program could be significantly improved if better data are available (high coverage and/or low

ODF) (Figure 2.2.5B). Based on the number of CNV calls in this study and the estimated sensitivity, I gave the approximate number of genes affected by CNV, 0.62, in each individual genome on average. Finally the result of PCR validation experiments performed on 24 random selected heterozygous deletion events indicated an FDR of 12.5%, which is comparable to or lower than the FDR of CNV detectors based on the RD signal in 1000GP Pilot 1 low coverage data (Table 2.2.7).

4.1.2 Tangram: An inclusive toolbox for MEI detection

Although it is possible today to detect large deletions and duplications with high accuracy, effective methods still need to be developed for several other structural variation (SV) types. MEI was still one of the most difficult SV types to detect and genotype, although a few methods have been published to tackle this problem [105, 118–120].

To address this difficult SV type, I developed a novel variant calling program, Tangram, designed to provide a flexible and efficient SV detection tool for genomics researchers to identify and characterize MEI accurately and sensitively in the human genome. This new tool relied heavily on split-read mappings performed on all problematic mates (*i.e.* read pairs where one end-mate is aligned with high mapping quality, but the other mate is either unmapped or mapped with many unaligned or "clipped-off" bases). This approach is different from other SV detection methods employing the SR mapping, which only attempt SR mappings in regions where the RP signal indicates the possibility of a candidate event. I found that a significant fraction of SV events were supported only by SR mapped reads but not RP mappings (Table 3.2.1). I also developed a genotyping module to assign genotype data likelihoods based on the number of RP and SR mappings, as well as the mapping quality values associated with sequencing reads.

I evaluated Tangram on simulated data, applied it to 1000GP data, and compared its performance to competing methods. The analysis of simulated data indicates a high-degree of sensitivity, specificity and genotype accuracy, across a wide range of sequence coverage values, both for heterozygous and for homozygous MEI events (Table 3.2.1 and 3.2.2). This experiment also demonstrates that the global SR method makes a key contribution for the sensitive MEI discovery (*e.g.* at 20× coverage nearly 15% of events are only detected from SR mappings). It is also able to report SV events with very accurate breakpoint locations (Figure 3.2.1). I ran Tangram on deep CEU trio data, and compared our detection performance with two competing methods, RetroSeq and TEA. Tangram had higher sensitivity to both known MEI events from the literature and experimentally validated events found in the 1000GP Pilot dataset, especially for L1 elements. The genotyping accuracy of our program as compared to experimentally determined genotypes was far better than those two competing methods (Table 3.2.4). Finally, PCR based validation experiments performed by our collaborators in the Batzer laboratory on 160 randomly selected events indicated an FDR of 5.93%, an accuracy that equals or exceeds the SNP calling specificity from the best variant callers (Table 3.2.8 and 3.2.9).

4.1.3 Discussion

During my PhD study, I developed two variant callers based on two different detection strategies, read-depth and read-pair plus split-read approaches for two different types of sequencing data, exon capture and whole genome sequence. Because each type of sequencing data has its own unique characteristic, it is necessary to adopt different SV detection algorithms. As mentioned in Chapter 1, compared to the RD algorithm, RP and SR are more superior methods in both breakpoint resolution and sensitivity to smaller events. However, they are not suitable for the SV detection in exon capture sequencing data since breakpoints of SV events might be outside the sequencing regions (breakpoints could locate at intronic or intergenic regions). Due to this special characteristic of capture sequencing data, candidate read pairs for RP (read pairs span across the breakpoint) and SR (reads pairs are sampled from the breakpoint) will not be obtained for the analysis. On the other hand, the RD method does not have this limitation. No presence of breakpoints in sequencing data will not keep it from detecting CNVs properly, since it only measures the change of read depth coverage in a given genomic region. Moreover, since the RD algorithm is computationally light-weight it is a good fit for analyzing large-scale data, *e.g.*

sequencing data from 1000GP. In the second research work, MEI detection from WGS data, RP + SR methods instead of the RD algorithm were applied because of their high detection efficiency and breakpoint resolution. Also, although MEI belongs to CNVs (changes in the net amount of DNA), the RD approach is basically blind to this type of SV since MEs are highly repetitive DNA elements. To accurately measure the read depth of a given genomic region, only those uniquely aligned sequencing reads will be taken into consideration for the statistical analysis and those reads aligned to multiple genomic positions will be excluded. So for the MEI detection, the RD method can hardly collect any signal. Moreover, due to the repetitiveness of MEs, traditional RP and SR methods also have to be customized enough for the special need of the detection: the postdoctoral research associate in our lab, Wan-Ping Lee, modified our sequencing read aligner, MOSAIK, in order to provide the extra MEI information (an optional BAM file tag, called "ZA") in the alignment file, which makes it possible for Tangram to detect MEIs with the RP method; I implemented a customized split alignment module in Tangram that can align soft-clipped or unaligned reads to both normal and ME references.

One interesting observation in these two research works is that although based on our study results from the exon capture sequencing data, we estimated that there should be many CNV events occur in exonic regions for a given individual, no MEI events were found in exonic regions when we looked at the detection result for 218 1000GP phase 1 samples (Table 3.2.6). This seemingly contradiction actually has several reasonable explanations: (1) MEI events in exonic regions are so destructive to genes that the individual carried these variations can not survive under the selection pressure. Even the shortest ME, Alu, has a length of about 300 bp. L1, SVA and HERV are all thousands bp long. Such a long DNA element inserted in the exon region will definitely has a great impact on the transcription process of a gene. Moreover, non-LTR MEs, Alu, L1 and SVA, carry their own insertion recognition motif, `5'-TTAAAA-3'`. One insertion of this kind of MEs will introduce more insertions at the same area, which will create a MEI "hotspot". This is further unfavorable under the selection pressure. (2) As shown in Figure 3.2.2, Tangram has relatively low sensitivity to those low allele frequency events due to the absence or sparseness of supporting fragments. It is possible that some low allele frequency MEIs occur in exonic regions but Tangram might not be able to detect them due to the detection efficiency issue.

Both of my research works were aiming at developing efficient algorithms for those SVs that were not addressed by any previous studies or very difficult to detect accurately in the past. My first research work opens a new door for the exploration of exon capture sequencing data as they are originally generated only for SNP and INDEL detections. In my second research work, I developed the state-of-the-art MEI detector that is capable of analyzing large-scale NGS data for the routine use. By properly introducing new modules and integrating new algorithms in the future, my current detector could be expanded to a comprehensive detection toolbox for more other SV types, such as inversions, translocations and *de novo* insertions (See Future directions 4.2).

4.2 Future directions

4.2.1 Current challenges

The future of the SV detection largely depends on the development of sequencing technologies and new computational methods that can take advantage of them. As most simple SVs, like deletions and duplications, are already well characterized by current available SV detection programs, the researching focus has moved to those much difficult SVs, *e.g.* inversions, translocations and complex events. Although some SV toolboxes, such DELLY [65], Pindel [64] and BreakDancer [106], have already provided the function to detect these types of events, their performance is less than satisfactory. For example, recent validation results in the 1000GP indicated a 70 – 100% FDR for current methods attempting to detect inversion events. Current challenges of the SV detection come from technology restrictions, algorithm limitations and biological complexities. From the aspect of technology, the current generation of sequencing technology can only provide short length reads (36bp – 250bp) due to the restrictions of chemical agents and image processing. The length of the sequencing read greatly limits the possibilities of the exploration of those SV events buried in the complex genomic context, like inversions which are usually surrounded by repeat sequences [147, 148]. In terms of current SV detection algorithms, most of them rely on the alignment of sequencing reads to the human reference assembly. This single-reference detection model could cause systematic biases. For example, most false detections of translocation events are caused by the mis-assembly in the human genome reference. Also sequencing reads from those highly mutated human genomes, like those from solid tumor tissues, might be difficult to align to the normal reference. As to the biological complexities of the human genome, many recent studies have found that SV events tend to aggregate at some certain genomic locations. For example, a paper published in 2011 [118] for MEI studies reported many "hot spots" for MEI events in the human genome. The early MEI events set stage for later events. Some newly inserted MEs are very close or even inside previous MEs. Such complex genomic regions create tremendous difficulties for current SV detection methods.

4.2.2 Prospect of new sequencing technologies

The fast and continuous advance in both sequencing technologies and computational methods may offer solutions to all the mentioned issues in the near future.

Many sequencing companies have already announced their third-generation products, such as Ion Torrent from Life Technologies and PacBio from Pacific Biosciences. Unlike the second-generation sequencing (NGS) technology that DNA molecules need to be amplified through PCR step before sequencing, the third-generation sequencing machine applied a brand new technique — Single-Molecule Real-Time (SMRT) sequencing technology [149]. Through this technique, the sequencing machine can directly observe the synthesis process of a single DNA polymerase, which significantly increases the sequencing speed and addresses many shortcomings of the second-generation sequencing technology, such as the PCR bias (not all the genomic regions can be amplified at the same rate due to the GC content difference) and short read length. The length of output reads from the third-generation sequencing machine could range from 1,000bp to 10,000bp, which is much longer than that from the NGS technology. Although this new technology is still not mature yet due to the relative high sequencing error (currently about 15%), it is not hard to imagine the bright future and wide use of it for the high-quality *de novo* assembly algorithm, direct identification of haplotypes and the SV detection in complex and repetitive genomic regions.

4.2.3 Prospect of new algorithms for SV detection

As new sequencing technologies become available, there is little doubt that new companion computational methods will also be developed rapidly. The much longer read length from the third generation sequencing machine opens many opportunities for multi-reference or even reference-free SV detection approaches. The multi-reference system is gradually formed these years as more and more genomic variants are detected and submitted to public variant databases such as dbSNP [150] and DGV [14]. It is highly possible that variants between newly sequenced genomes and the reference are already existed in these databases. Thus detection of these existed variants will become a simple task if a well-designed aligner can map sequencing reads not only to the normal reference but also to those alternative alleles. Several attempts have already been carried out based on NGS short reads [151, 152]. As the continuous expansion and improvement of variant databases, such as the removal of duplicated entries and the refinement of breakpoint positions, this approach could be applied routinely in the future for the detection of common SVs in large-scale sequencing projects. Another direction of the future SV studies is *de novo* assembly method. The performance of current *de novo* assemblers are greatly restricted by the read length of the NGS technology. According to recent study results, tens of thousands of errors could be generated with short sequencing reads by the-state-of-art *de novo* assemblers [153] for human genomes. Moreover, the memory and time cost is prohibitively expensive for current *de novo* assemblers for routine uses due to the huge number of reads generated by NGS machines. The future development of *de novo* assemblers will greatly benefit from the longer length and less number of reads from the third-generation sequencing technology. Also, the memory usage of *de* *novo* assemblers could be significantly reduced by using the compressed data structure during the assembly process [154]. With high-quality assembly data, almost all SV types should be easily identified and characterized.

4.2.4 Prospect of function studies

The ultimate goal of genomics studies is the continuous improvement of the human health. The last ten years since the completeness of the Human Genome Project has witnessed the huge advance in understanding genetic variations that distinguish different people and are responsible for specific traits and diseases. Based on the results of numerous genomic variants studies, genome-wide association studies in humans have been carried out to identify the relationship between inherited mutations and various common human diseases, such as heart disease [155, 156], diabetes [157–159], Alzheimer's disease [160, 161] and Crohn's disease [162–164]. Although more than 13,000 GWAS papers have been published in the last 5 years, germline variants discovered in these researches only address a small fraction of the heritability of traits and diseases [165] (less than 500 types). Until recently most GWAS studies only take SNP variants into account as the SNP database and detection methods are pretty mature. However, in the past few years it has been clear that SV is also a major contributor to human genomic variations and can actually affect more genomic regions than SNPs [14, 73]. Moreover, since there were no cost-effective methods to call all genetic variants in a large number of human genomes, currently many GWAS studies only focus their attentions on common variants whose allele frequencies are higher than 5%. The "missing heritability" gap due to these two limitations mentioned above is the major bottleneck for GWAS studies [166]. The further development and improvement of both sequencing technologies and SV detection algorithms in the next ten years will enable the systematical discoveries and characterizations of all types of germline SVs in the human genome and create a complete list of genomic variants that will greatly facilitate association studies that can translate the genetic information into phenotypic diversity or pathogenesis. Here the "complete" does not mean we will sequence the DNA sample from every individual in the world. Instead, it is more desirable to have a comprehensive SV database with the accurate information, such as the type, position and length of a given SV event, at a satisfactory population allele frequency deepness (say <0.5%). For example, to catch variants down to 0.5% AFS in a population with 90% sensitivity, only 230 individuals need to be sampled ($log(1-0.9)/log(1-0.005)/2$). As the rapid development of technologies and measuring algorithms, soon this database could be set up for the downstream functional study and serve as the major resource to fill the "missing heritability" gap for future GWAS studies.

Besides those population-scale genomics studies, another branch of human genomics, personal genomics and medical, is also under fast development. The preliminary results from variant researches in the human genome have already attracted the attetions of the public. More and more people are willing to explore their own genomic information to identify variants that may threat their future health. This useful information could help them to take some preventive actions or appropriate treatments to avoid their future health risks. Many personal genomics projects, e.g. Personal Genome Project (PGP) [167–169], have already started to collect and sequencing DNA samples from a broader space than that of normal large-scale genomics projects, such as 1000GP, in a long-term run. Also many companies have already sensed the commercial interest of delivering the genomic analysis to individual customers. For example, 23andMe sells mail order of SNP genotyping kits for people who want to assess their risks of 178 diseases and estimate their ancestry origins. Other firms, such as HelloGenome and deCODEme.com, all offer similar services to the public. As the cost of WGS rapidly and continuously drops, sequencing-based services, instead of SNP genotyping kits, may become the mainstream. However, currently SV studies did not play an important role in these analyses. As the reason mentioned above, compared to the current knowledge of SNPs, our understanding about SVs is still not comprehensive enough. Methods that can be used to accurately characterize all types of SVs are still under developing. Until then personal genomic studies could extend to broader areas that have never been explored before due to the lack of associations between phenotypes and genotypes and we should be able to understand more clearly of the pathogenesis of most common and rare diseases. The personal genomic information at that time may become much valuable to us for the purpose of personalized medicine and therapies that could substantially improve our health quality.

Another possible high-impact direction of SV researches in the future is the identification and characterization of somatic mutations for different types of cancers in different tissues (organs). Unlike germline mutations that are inherited from parents, somatic mutations are accumulated during the lifetime of an individual. These mutations are tissue specific or even single-cell specific (the mutations you got on your skin due to the sun burn could be much different from those in your stomach due to the alcohol damage) and they are driven factors for various types of cancers [170]. These somatic mutations inherited by daughter cells in tumors are under continuous selections, which make the cancer a "microevolutionary process" [171–174]. More and more "passenger" mutations are introduced during this whole evolutionary process as a result of the increasing instability of the DNA repair machinery. Cancer genomes, especially those in solid tumors, are extensively rearranged compared to the normal healthy genome [175–177]. Although somatic mutations have been recognized as the "top criminal" that is responsible for the cancer formation for decades, it is still very difficult to detect driver variations (in most cases SVs) since they are usually buried in a background of germline (could be filtered out with normal control genome from the same patient but it still depends on the sensitive SV detection on both DNA samples) and "passenger" mutations. The signal to noise ratio (SNR) in the cancer genome is generally very low. With the help of the current high-resolution genomics technology, several recurrent fusion genes are discovered in solid tumors, such as prostate [178] and lung cancers [179] but we are still far away from accurately and systematically detecting these driver mutations from various types of cancers. As the read length from the future sequencing technology becomes longer and longer, one possible breakthrough of somatic mutations detection could be *de novo* assembly method. Using the reference-free method to detect the SVs in the cancer genome could overcome some limitations of resequencing-based detection methods, such as the mapping accuracy for those highly mutated genomes and the sensitivity to insertion events, and could reconstruct the organization of the cancer genome at the single nucleotide resolution. Although there are no publications for this type of study some genomics scientists have already started to explore this promising research direction [180]. With the rapid development of technologies and SV detection algorithms and the broad corporation of international institutions in large cancer genome projects, such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC), searching driven somatic mutations at the genome scale will become practical and very cost-effective, which could significantly facilitate the downstream pathogenesis and medicine targeting study.

It will be a long journey to decode all the secrets in the human genome and we are just passing the start line by studying variants and some of their functional impacts. The full picture of the human genome will become more and more clear as we collect more and more variations like jigsaw puzzles from different sources, population-scale, personal-scale and tissue- and disease-specific data. With sufficient data preparation, bold hypothesis proposal and prudent experiment design from the entire biology community, we are gradually approaching the comprehensive understanding of the relationship between the genetic information and its complicated functions. Of course, studying the variants on the DNA sequence level is just a beginning. Many other inheritable factors, such as epigenetic variants, also plays a significant role in affecting our phenotypic traits [181–183] or susceptibility to different diseases, including Angelman syndrome [184], Prader-Willi syndrome [185], Beckwith-Wiedemann syndrome [186, 187] and various types of cancers [188–196]. Some epigenetics problems, *e.g.* methylation variation detection, are very similar to those in the SV detection (CNV detection). Many methods used for SV discovery in high throughput sequencing data could also be transplanted easily on large-scale epigenetic data [197–199]. So the future achievement of SV studies could also greatly benefit the development of epigenetics. The progress of variant studies, including SNP, INDEL, SV and epigenetic variations, will accelerate the process of finding "missing heritability" in the human genome and facilitate downstream GWAS studies, which could potentially bring revolutionary improvements to the human health.

References

[1] M. Przeworski, R. R. Hudson, and A. D. Rienzo, "Adjusting the focus on human variation," *Trends in genetics*, vol. 16, pp. 296–302, Jul 2000.

[2] D. E. Reich, S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin, J. M. Higgins, D. J. Richter, E. S. Lander, and D. Altshuler, "Human genome sequence variation and the influence of gene history, mutation and recombination," *Nature genetics*, vol. 32, pp. 135–142, Sep 2002.

[3] P. A. Jacobs, A. G. Baikie, W. M. C. Brown, and J. A. Strong, "The somatic chromosomes in mongolism," *Lancet*, vol. 1, p. 710, Apr 4 1959.

[4] J. H. Edwards, D. G. Harnden, A. H. Cameron, V. M. Crosse, and O. H. Wolff, "A new trisomic syndrome," *Lancet*, vol. 1, pp. 787–790, Apr 9 1960.

[5] K. Patau, D. W. Smith, E. Therman, S. L. Inhorn, and H. P. Wagner, "Multiple congenital anomaly caused by an extra autosome," *Lancet*, vol. 1, pp. 790–793, Apr 9 1960.

[6] M. Bobrow, L. F. Joness, and G. Clarke, "A complex chromosomal rearrangement with formation of a ring 4," *Journal of medical genetics*, vol. 8, pp. 235–239, Jun 1971.

[7] P. A. Jacobs, J. S. Matsuura, M. Mayer, and I. M. Newlands, "A cytogenetic survey of an

institution for the mentally retarded: I. Chromosome abnormalities," *Clinical genetics*, vol. 13, pp. 37–60, Jan 1978.

[8] R. Coco and V. B. Penchaszadeh, "Cytogenetic findings in 200 children with mental retardation and multiple congenital anomalies of unknown cause," *American Journal of Medical Genetics*, vol. 12, pp. 155–173, Jun 1982.

[9] R. S. Verma, J. Rodriguez, and H. Dosik, "The clinical significance of pericentric inversion of the human Y chromosome: a rare 'third' type of heteromorphism," *The Journal of heredity*, vol. 73, pp. 236–238, May-Jun 1982.

[10] L. Y. Hsu, P. A. Benn, H. L. Tannenbaum, T. E. Perlis, and A. D. Carlson, "Chromosomal polymorphisms of 1, 9, 16, and Y in 4 major ethnic groups: a large prenatal study," *American Journal of Medical Genetics*, vol. 26, pp. 95–101, Jan 1987.

[11] R. S. Verma, H. Dosik, and H. A. Lubs, "Size and pericentric inversion heteromorphisms of secondary constriction regions (h) of chromosomes 1, 9, and 16 as detected by CBG technique in Caucasians: classification, frequencies, and incidence," *American Journal of Medical Genetics*, vol. 2, no. 4, pp. 331–339, 1978.

[12] H. A. Lubs, "A marker X chromosome," *American Journal of Human Genetics*, vol. 21, pp. 231–244, May 1969.

[13] P. Stankiewicz and J. R. Lupski, "Structural variation in the human genome and its role in disease," *Annual Review of Medicine*, vol. 61, pp. 437–455, 2010.

[14] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee, "Detection of large-scale variation in the human genome," *Nature genetics*, vol. 36, pp. 949–951, Sep 2004.

[15] 1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean, "A map of human genome

variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, Oct 28 2010.

[16] International Cancer Genome Consortium, T. J. Hudson, W. Anderson, A. Artez, A. D. Barker, C. Bell, R. R. Bernabe, M. K. Bhan, F. Calvo, I. Eerola, D. S. Gerhard, A. Guttmacher, M. Guyer, *et al.*, "International network of cancer genome projects," *Nature*, vol. 464, pp. 993–998, Apr 15 2010.

[17] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nature reviews: Genetics*, vol. 7, pp. 85–97, Feb 2006.

[18] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. Macarthur, J. R. Macdonald, I. Onyiah, A. W. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, Wellcome Trust Case Control Consortium, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles, "Origins and functional impact of copy number variation in the human genome," *Nature*, vol. 464, pp. 704–712, Apr 1 2010.

[19] A. W. Pang, J. R. MacDonald, D. Pinto, J. Wei, M. A. Rafiq, D. F. Conrad, H. Park, M. E. Hurles, C. Lee, J. C. Venter, E. F. Kirkness, S. Levy, L. Feuk, and S. W. Scherer, "Towards a comprehensive structural variation map of an individual human genome," *Genome biology*, vol. 11, no. 5, pp. R52–2010–11–5–r52. Epub 2010 May 19, 2010.

[20] D. J. Turner, M. Miretti, D. Rajan, H. Fiegler, N. P. Carter, M. L. Blayney, S. Beck, and M. E. Hurles, "Germline rates of *de novo* meiotic deletions and duplications causing several genomic disorders," *Nature genetics*, vol. 40, pp. 90–95, Jan 2008.

[21] J. R. Lupski, "Genomic rearrangements and sporadic disease," *Nature genetics*, vol. 39, pp. S43–7, Jul 2007.

[22] J. R. Lupski, "Genomic disorders: structural features of the genome can lead to DNA

rearrangements and human disease traits," *Trends in genetics*, vol. 14, pp. 417–422, Oct 1998.

[23] J. R. Lupski, "Genomic disorders ten years on," *Genome medicine*, vol. 1, p. 42, Apr 24 2009.

[24] C. R. Marshall, A. Noor, J. B. Vincent, A. C. Lionel, L. Feuk, J. Skaug, M. Shago, R. Moessner, D. Pinto, Y. Ren, B. Thiruvahindrapduram, A. Fiebig, S. Schreiber, J. Friedman, C. E. Ketelaars, Y. J. Vos, C. Ficicioglu, S. Kirkpatrick, R. Nicolson, L. Sloman, A. Summers, C. A. Gibbons, A. Teebi, D. Chitayat, R. Weksberg, A. Thompson, C. Vardy, V. Crosbie, S. Luscombe, R. Baatjes, L. Zwaigenbaum, W. Roberts, B. Fernandez, P. Szatmari, and S. W. Scherer, "Structural variation of chromosomes in autism spectrum disorder," *American Journal of Human Genetics*, vol. 82, pp. 477–488, Feb 2008.

[25] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y. H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimaki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M. C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye, and M. Wigler, "Strong association of de novo copy number mutations with autism," *Science (New York, N.Y.)*, vol. 316, pp. 445–449, Apr 20 2007.

[26] A. J. Sharp, S. Hansen, R. R. Selzer, Z. Cheng, R. Regan, J. A. Hurst, H. Stewart, S. M. Price, E. Blair, R. C. Hennekam, C. A. Fitzpatrick, R. Segraves, T. A. Richmond, C. Guiver, D. G. Albertson, D. Pinkel, P. S. Eis, S. Schwartz, S. J. Knight, and E. E. Eichler, "Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome," *Nature genetics*, vol. 38, pp. 1038–1042, Sep 2006.

[27] D. G. Albertson, C. Collins, F. McCormick, and J. W. Gray, "Chromosome aberrations in solid tumors," *Nature genetics*, vol. 34, pp. 369–376, Aug 2003.

[28] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X. W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta,

M. A. Rubin, and A. M. Chinnaiyan, “Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer,” *Science (New York, N.Y.)*, vol. 310, pp. 644–648, Oct 28 2005.

[29] M. Soda, Y. L. Choi, M. Enomoto, S. Takada, Y. Yamashita, S. Ishikawa, S. Fujiwara, H. Watanabe, K. Kurashina, H. Hatanaka, M. Bando, S. Ohno, Y. Ishikawa, H. Aburatani, T. Niki, Y. Sohara, Y. Sugiyama, and H. Mano, “Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer,” *Nature*, vol. 448, pp. 561–566, Aug 2 2007.

[30] S. Girirajan, C. D. Campbell, and E. E. Eichler, “Human copy number variation and complex genetic disease,” *Annual Review of Genetics*, vol. 45, pp. 203–226, 2011.

[31] J. R. Lupski, “An inherited DNA rearrangement and gene dosage effect are responsible for the most common autosomal dominant peripheral neuropathy: Charcot-Marie-Tooth disease type 1A,” *Clinical research*, vol. 40, pp. 645–652, Dec 1992.

[32] J. R. Lupski, C. A. Wise, A. Kuwano, L. Pentao, J. T. Parke, D. G. Glaze, D. H. Ledbetter, F. Greenberg, and P. I. Patel, “Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A,” *Nature genetics*, vol. 1, pp. 29–33, Apr 1992.

[33] D. A. Kleinjan and V. van Heyningen, “Long-range control of gene expression: emerging mechanisms and disruption in disease,” *American Journal of Human Genetics*, vol. 76, pp. 8–32, Jan 2005.

[34] E. Chaignat, E. A. Yahya-Graison, C. N. Henrichsen, J. Chrast, F. Schutz, S. Pradervand, and A. Reymond, “Copy number variation modifies expression time courses,” *Genome research*, vol. 21, pp. 106–113, Jan 2011.

[35] C. N. Henrichsen, N. Vinckenbosch, S. Zollner, E. Chaignat, S. Pradervand, F. Schutz, M. Ruedi, H. Kaessmann, and A. Reymond, “Segmental copy number variation shapes tissue transcriptomes,” *Nature genetics*, vol. 41, pp. 424–429, Apr 2009.

[36] L. D. Orozco, S. J. Cokus, A. Ghazalpour, L. Ingram-Drake, S. Wang, A. van Nas, N. Che, J. A. Araujo, M. Pellegrini, and A. J. Lusis, "Copy number variation influences gene expression and metabolic traits in mice," *Human molecular genetics*, vol. 18, pp. 4118–4129, Nov 1 2009.

[37] A. Schlattl, S. Anders, S. M. Waszak, W. Huber, and J. O. Korbel, "Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions," *Genome research*, vol. 21, pp. 2004–2013, Dec 2011.

[38] O. Vazquez-Mena, I. Medina-Martinez, E. Juarez-Torres, V. Barron, A. Espinosa, N. Villegas-Sepulveda, L. Gomez-Laguna, K. Nieto-Martinez, L. Orozco, E. Roman-Basaure, S. M. Cortez, M. B. Ibanez, C. Venegas-Vega, M. Guardado-Estrada, A. Rangel-Lopez, S. Kofman, and J. Berumen, "Amplified genes may be overexpressed, unchanged, or downregulated in cervical cancer cell lines," *PloS one*, vol. 7, no. 3, p. e32667, 2012.

[39] I. Kurth, E. Klopocki, S. Stricker, J. van Oosterwijk, S. Vanek, J. Altmann, H. G. Santos, J. J. van Harssel, T. de Ravel, A. O. Wilkie, A. Gal, and S. Mundlos, "Duplications of noncoding elements 5' of SOX9 are associated with brachydactyly-anonychia," *Nature genetics*, vol. 41, pp. 862–863, Aug 2009.

[40] L. M. Merlo, J. W. Pepper, B. J. Reid, and C. C. Maley, "Cancer as an evolutionary and ecological process," *Nature reviews: Cancer*, vol. 6, pp. 924–935, Dec 2006.

[41] C. M. Carvalho and J. R. Lupski, "Copy number variation at the breakpoint region of isochromosome 17q," *Genome research*, vol. 18, pp. 1724–1732, Nov 2008.

[42] N. Kurotaki, J. J. Shen, M. Touyama, T. Kondoh, R. Visser, T. Ozaki, J. Nishimoto, T. Shiihara, K. Uetake, Y. Makita, N. Harada, S. Raskin, C. W. Brown, P. Hoglund, N. Okamoto, and J. R. Lupski, "Phenotypic consequences of genetic variation at hemizygous alleles: Sotos syndrome is a contiguous gene syndrome incorporating

coagulation factor twelve (FXII) deficiency," *Genetics in medicine*, vol. 7, pp. 479–483, Sep 2005.

[43] R. D. Schmickel, "Contiguous gene syndromes: a component of recognizable syndromes," *The Journal of pediatrics*, vol. 109, pp. 231–241, Aug 1986.

[44] L. Feuk, J. R. MacDonald, T. Tang, A. R. Carson, M. Li, G. Rao, R. Khaja, and S. W. Scherer, "Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies," *PLoS genetics*, vol. 1, p. e56, Oct 2005.

[45] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, Feb 15 2001.

[46] S. Solinas-Toldo, S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Dohner, T. Cremer, and P. Lichter, "Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances," *Genes, chromosomes & cancer*, vol. 20, pp. 399–407, Dec 1997.

[47] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson, "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays," *Nature genetics*, vol. 20, pp. 207–211, Oct 1998.

[48] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, pp. 5463–5467, Dec 1977.

[49] D. Pinkel and D. G. Albertson, "Array comparative genomic hybridization and its applications in cancer," *Nature genetics*, vol. 37 Suppl, pp. S11–7, Jun 2005.

[50] A. P. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock, "Repetitive elements may comprise over two-thirds of the human genome," *PLoS genetics*, vol. 7, p. e1002384, Dec 2011.

[51] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, pp. 376–380, Sep 15 2005.

[52] D. R. Bentley, "Whole-genome re-sequencing," *Current opinion in genetics & development*, vol. 16, pp. 545–552, Dec 2006.

[53] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church, "Accurate multiplex polony sequencing of an evolved bacterial genome," *Science (New York, N.Y.)*, vol. 309, pp. 1728–1732, Sep 9 2005.

[54] R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, F. Dahl, A. Fernandez, B. Staker, K. P. Pant, J. Baccash, A. P. Borcherding, A. Brownley, R. Cedeno, L. Chen, D. Chernikoff, A. Cheung, R. Chirita, B. Curson, J. C. Ebert, C. R. Hacker, R. Hartlage, B. Hauser, S. Huang, Y. Jiang, V. Karpinchyk, M. Koenig, C. Kong, T. Landers, C. Le, J. Liu, C. E. McBride, M. Morenzoni, R. E. Morey, K. Mutch, H. Perazich, K. Perry, B. A. Peters, J. Peterson, C. L. Pethiyagoda, K. Pothuraju, C. Richter, A. M. Rosenbaum, S. Roy,

J. Shafto, U. Sharanhovich, K. W. Shannon, C. G. Sheppy, M. Sun, J. V. Thakuria, A. Tran, D. Vu, A. W. Zaranek, X. Wu, S. Drmanac, A. R. Oliphant, W. C. Banyai, B. Martin, D. G. Ballinger, G. M. Church, and C. A. Reid, "Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays," *Science (New York, N.Y.)*, vol. 327, pp. 78–81, Jan 1 2010.

[55] M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu, "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers," *BMC genomics*, vol. 13, pp. 341–2164–13–341, Jul 24 2012.

[56] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nature biotechnology*, vol. 26, pp. 1135–1145, Oct 2008.

[57] W.-P. Lee, M. Stromberg, A. Ward, C. Stewart, E. Garrison, and G. Marth, "MOSAIK: A next-generation reference-guided aligner." In preparation.

[58] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1754–1760, Jul 15 2009.

[59] N. Homer, B. Merriman, and S. F. Nelson, "BFAST: an alignment tool for large scale genome resequencing," *PloS one*, vol. 4, p. e7767, Nov 11 2009.

[60] L. Heng, "wgsim - Read simulator for next generation sequencing." http://github.com/lh3/wgsim.

[61] M. Holtgrewe, "Mason — a read simulator for second generation sequencing data," tech. rep., Freie Universität Berlin, Jun 2011.

[62] B. J. Raphael, "Chapter 6: Structural variation and medical genomics," *PLoS computational biology*, vol. 8, no. 12, p. e1002821, 2012.

[63] R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. Lam, J. Leng, R. Li, Y. Li, C. Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stutz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, J. O. Korbel, and 1000 Genomes Project, "Mapping copy number variation by population-scale genome sequencing," *Nature*, vol. 470, pp. 59–65, Feb 3 2011.

[64] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads," *Bioinformatics (Oxford, England)*, vol. 25, pp. 2865–2871, Nov 1 2009.

[65] T. Rausch, T. Zichner, A. Schlattl, A. M. Stutz, V. Benes, and J. O. Korbel, "DELLY: structural variant discovery by integrated paired-end and split-read analysis," *Bioinformatics (Oxford, England)*, vol. 28, pp. i333–i339, Sep 15 2012.

[66] P. H. Sudmant, J. O. Kitzman, F. Antonacci, C. Alkan, M. Malig, A. Tsalenko, N. Sampas, L. Bruhn, J. Shendure, 1000 Genomes Project, and E. E. Eichler, "Diversity of human copy number variation and multicopy genes," *Science (New York, N.Y.)*, vol. 330, pp. 641–646, Oct 29 2010.

[67] M. Fanciulli, E. Petretto, and T. J. Aitman, "Gene copy number variation and common human disease," *Clinical genetics*, vol. 77, pp. 201–213, Mar 2010.

[68] B. Frank, J. L. Bermejo, K. Hemminki, C. Sutter, B. Wappenschmidt, A. Meindl, M. Kiechle-Bahat, P. Bugert, R. K. Schmutzler, C. R. Bartram, and B. Burwinkel, "Copy

number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk," *Carcinogenesis*, vol. 28, pp. 1442–1445, Jul 2007.

[69] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y. H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimaki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M. C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye, and M. Wigler, "Strong association of de novo copy number mutations with autism," *Science (New York, N.Y.)*, vol. 316, pp. 445–449, Apr 20 2007.

[70] M. Kusenda and J. Sebat, "The role of rare structural variants in the genetics of autism spectrum disorders," *Cytogenetic and genome research*, vol. 123, no. 1-4, pp. 36–43, 2008.

[71] J. G. Mulle, A. F. Dodd, J. A. McGrath, P. S. Wolyniec, A. A. Mitchell, A. C. Shetty, N. L. Sobreira, D. Valle, M. K. Rudd, G. Satten, D. J. Cutler, A. E. Pulver, and S. T. Warren, "Microdeletions of 3q29 confer high risk for schizophrenia," *American Journal of Human Genetics*, vol. 87, pp. 229–236, Aug 13 2010.

[72] A. Baross, A. D. Delaney, H. I. Li, T. Nayar, S. Flibotte, H. Qian, S. Y. Chan, J. Asano, A. Ally, M. Cao, P. Birch, M. Brown-John, N. Fernandes, A. Go, G. Kennedy, S. Langlois, P. Eydoux, J. M. Friedman, and M. A. Marra, "Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data," *BMC bioinformatics*, vol. 8, p. 368, Oct 2 2007.

[73] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler, "Large-scale copy number polymorphism in the human genome," *Science (New York, N.Y.)*, vol. 305, pp. 525–528, Jul 23 2004.

[74] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H.

Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles, "Global variation in copy number in the human genome," *Nature*, vol. 444, pp. 444–454, Nov 23 2006.

[75] G. M. Cooper, D. A. Nickerson, and E. E. Eichler, "Mutational and selective effects on copy-number variants in the human genome," *Nature genetics*, vol. 39, pp. S22–9, Jul 2007. LR: 20101118; GR: HL066682/HL/NHLBI NIH HHS/United States; GR: T32 HG00035/HG/NHGRI NIH HHS/United States; JID: 9216904; RF: 75; ppublish.

[76] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat, "Sensitive and accurate detection of copy number variants using read depth of coverage," *Genome research*, vol. 19, pp. 1586–1592, Sep 2009.

[77] J. F. Sathirapongsasuti, H. Lee, B. A. Horst, G. Brunner, A. J. Cochran, S. Binder, J. Quackenbush, and S. F. Nelson, "Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV," *Bioinformatics (Oxford, England)*, vol. 27, pp. 2648–2654, Oct 1 2011.

[78] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M. M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. Dicuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B. L. Maidak, J. Mudge, M. R. Murphy, T. Murphy, J. Rajan, B. Rajput, L. D. Riddick, C. Snow, C. Steward, D. Webb, J. A. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin, and D. Lipman, "The consensus coding sequence

(CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes," *Genome research*, vol. 19, pp. 1316–1323, Jul 2009.

[79] R. A. Harte, C. M. Farrell, J. E. Loveland, M. M. Suner, L. Wilming, B. Aken, D. Barrell, A. Frankish, C. Wallin, S. Searle, M. Diekhans, J. Harrow, and K. D. Pruitt, "Tracking and coordinating an international curation effort for the CCDS Project," *Database : the journal of biological databases and curation*, vol. 2012, p. bas008, Mar 20 2012.

[80] G. T. Marth, F. Yu, A. R. Indap, K. Garimella, S. Gravel, W. F. Leong, C. Tyler-Smith, M. Bainbridge, T. Blackwell, X. Zheng-Bradley, Y. Chen, D. Challis, L. Clarke, E. V. Ball, K. Cibulskis, D. N. Cooper, B. Fulton, C. Hartl, D. Koboldt, D. Muzny, R. Smith, C. Sougnez, C. Stewart, A. Ward, J. Yu, Y. Xue, D. Altshuler, C. D. Bustamante, A. G. Clark, M. Daly, M. DePristo, P. Flicek, S. Gabriel, E. Mardis, A. Palotie, R. Gibbs, and . G. Project, "The functional spectrum of low-frequency coding variation," *Genome biology*, vol. 12, pp. R84–2011–12–9–r84, Sep 14 2011.

[81] H. Vikalo, B. Hassibi, and A. Hassibi, "Limits of performance of quantitative polymerase chain reaction systems," *IEEE Transactions on Information Theory*, vol. 56, no. 2, pp. 688–695, 2010.

[82] C. Hansis, J. A. Grifo, and L. C. Krey, "Oct-4 expression in inner cell mass and trophectoderm of human blastocysts," *Molecular human reproduction*, vol. 6, pp. 999–1004, Nov 2000.

[83] T. Fujino, K. Nomura, Y. Ishikawa, H. Makino, A. Umezawa, H. Aburatani, K. Nagasaki, and T. Nakamura, "Function of EWS-POU5F1 in sarcomagenesis and tumor cell maintenance," *The American journal of pathology*, vol. 176, pp. 1973–1982, Apr 2010.

[84] S. J. Korsmeyer, "BCL-2 gene family and the regulation of programmed cell death," *Cancer research*, vol. 59, pp. 1693s–1700s, Apr 1 1999.

[85] S. J. Dawson, N. Makretsov, F. M. Blows, K. E. Driver, E. Provenzano, J. L. Quesne, L. Baglietto, G. Severi, G. G. Giles, C. A. McLean, G. Callagy, A. R. Green, I. Ellis, K. Gelmon, G. Turashvili, S. Leung, S. Aparicio, D. Huntsman, C. Caldas, and P. Pharoah, "BCL2 in breast cancer: a favourable prognostic marker across molecular subtypes and independent of adjuvant therapy received," *British journal of cancer*, vol. 103, pp. 668–675, Aug 24 2010.

[86] S. D. Catz and J. L. Johnson, "BCL-2 in prostate cancer: a minireview," *Apoptosis : An International Journal on Programmed Cell Death*, vol. 8, pp. 29–37, Jan 2003.

[87] W. D. Roock, B. Claes, D. Bernasconi, J. D. Schutter, B. Biesmans, G. Fountzilas, K. T. Kalogeras, V. Kotoula, D. Papamichael, P. Laurent-Puig, F. Penault-Llorca, P. Rougier, B. Vincenzi, D. Santini, G. Tonini, F. Cappuzzo, M. Frattini, F. Molinari, P. Saletti, S. D. Dosso, M. Martini, A. Bardelli, S. Siena, A. Sartore-Bianchi, J. Tabernero, T. Macarulla, F. D. Fiore, A. O. Gangloff, F. Ciardiello, P. Pfeiffer, C. Qvortrup, T. P. Hansen, E. V. Cutsem, H. Piessevaux, D. Lambrechts, M. Delorenzi, and S. Tejpar, "Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis," *The lancet oncology*, vol. 11, pp. 753–762, Aug 2010.

[88] M. D. Hogarty, M. D. Norris, K. Davis, X. Liu, N. F. Evageliou, C. S. Hayes, B. Pawel, R. Guo, H. Zhao, E. Sekyere, J. Keating, W. Thomas, N. C. Cheng, J. Murray, J. Smith, R. Sutton, N. Venn, W. B. London, A. Buxton, S. K. Gilmour, G. M. Marshall, and M. Haber, "ODC1 is a critical determinant of MYCN oncogenesis and a therapeutic target in neuroblastoma," *Cancer research*, vol. 68, pp. 9735–9745, Dec 1 2008.

[89] P. Castro, C. J. Creighton, M. Ozen, D. Berel, M. P. Mims, and M. Ittmann, "Genomic profiling of prostate cancers from African American men," *Neoplasia (New York, N.Y.)*, vol. 11, pp. 305–312, Mar 2009.

[90] A. H. Talukder, Q. Meng, and R. Kumar, "CRIPak, a novel endogenous Pak1 inhibitor," *Oncogene*, vol. 25, pp. 1311–1319, Mar 2 2006.

[91] S. Rega, T. Stiewe, D. I. Chang, B. Pollmeier, H. Esche, W. Bardenheuer, G. Marquitan, and B. M. Putzer, "Identification of the full-length huntingtin-interacting protein p231HBP/HYPB as a DNA-binding factor," *Molecular and cellular neurosciences*, vol. 18, pp. 68–79, Jul 2001.

[92] P. W. Faber, G. T. Barnes, J. Srinidhi, J. Chen, J. F. Gusella, and M. E. MacDonald, "Huntingtin interacts with a family of WW domain proteins," *Human molecular genetics*, vol. 7, pp. 1463–1474, Sep 1998.

[93] N. K. Patel and S. S. Gill, "GDNF delivery for Parkinson's disease," *Acta neurochirurgica.Supplement*, vol. 97, no. Pt 2, pp. 135–154, 2007.

[94] J. R. Connor and S. Y. Lee, "HFE mutations and Alzheimer's disease," *Journal of Alzheimer's disease : JAD*, vol. 10, pp. 267–276, Nov 2006.

[95] B. Ilkovski, N. Mokbel, R. A. Lewis, K. Walker, K. J. Nowak, A. Domazetovska, N. G. Laing, V. M. Fowler, K. N. North, and S. T. Cooper, "Disease severity and thin filament regulation in M9R TPM3 nemaline myopathy," *Journal of neuropathology and experimental neurology*, vol. 67, pp. 867–877, Sep 2008.

[96] S. Tsui, T. Dai, S. Roettger, W. Schempp, E. C. Salido, and P. H. Yen, "Identification of two novel proteins that interact with germ-cell-specific RNA-binding proteins DAZ and DAZL1," *Genomics*, vol. 65, pp. 266–273, May 1 2000.

[97] C. Klein, M. Grudzien, G. Appaswamy, M. Germeshausen, I. Sandrock, A. A. Schaffer, C. Rathinam, K. Boztug, B. Schwinzer, N. Rezaei, G. Bohn, M. Melin, G. Carlsson, B. Fadeel, N. Dahl, J. Palmblad, J. I. Henter, C. Zeidler, B. Grimbacher, and K. Welte, "HAX1 deficiency causes autosomal recessive severe congenital neutropenia (Kostmann disease)," *Nature genetics*, vol. 39, pp. 86–92, Jan 2007.

[98] N. Krumm, P. H. Sudmant, A. Ko, B. J. O'Roak, M. Malig, B. P. Coe, N. E. S. Project, A. R. Quinlan, D. A. Nickerson, and E. E. Eichler, "Copy number variation detection and genotyping from exome sequence data," *Genome research*, vol. 22, pp. 1525–1532, Aug 2012.

[99] T. Koressaar and M. Remm, "Enhancements and modifications of primer design program Primer3," *Bioinformatics (Oxford, England)*, vol. 23, pp. 1289–1291, May 15 2007.

[100] A. Untergasser, I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, and S. G. Rozen, "Primer3–new capabilities and interfaces," *Nucleic acids research*, vol. 40, p. e115, Aug 2012.

[101] W. J. Kent, "BLAT–the BLAST-like alignment tool," *Genome research*, vol. 12, pp. 656–664, Apr 2002.

[102] S. A. McCarroll, A. Huett, P. Kuballa, S. D. Chilewski, A. Landry, P. Goyette, M. C. Zody, J. L. Hall, S. R. Brant, J. H. Cho, R. H. Duerr, M. S. Silverberg, K. D. Taylor, J. D. Rioux, D. Altshuler, M. J. Daly, and R. J. Xavier, "Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease," *Nature genetics*, vol. 40, pp. 1107–1112, Sep 2008.

[103] H. H. K. Jr, C. Wong, H. Youssoufian, A. F. Scott, D. G. Phillips, and S. E. Antonarakis, "Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man," *Nature*, vol. 332, pp. 164–166, Mar 10 1988.

[104] Y. Miki, T. Katagiri, F. Kasumi, T. Yoshimoto, and Y. Nakamura, "Mutation analysis in the BRCA2 gene in primary breast cancers," *Nature genetics*, vol. 13, pp. 245–247, Jun 1996.

[105] E. Lee, R. Iskow, L. Yang, O. Gokcumen, P. Haseley, L. J. L. 3rd, J. G. Lohr, C. C. Harris, L. Ding, R. K. Wilson, D. A. Wheeler, R. A. Gibbs, R. Kucherlapati, C. Lee, P. V. Kharchenko, P. J. Park, and C. G. A. R. Network, "Landscape of somatic retrotransposition in human cancers," *Science (New York, N.Y.)*, vol. 337, pp. 967–971, Aug 24 2012.

[106] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang, D. P. Locke, X. Shi, R. S. Fulton, T. J. Ley, R. K. Wilson, L. Ding, and E. R. Mardis, "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation," *Nature methods*, vol. 6, pp. 677–681, Sep 2009.

[107] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavare, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis, "Relative impact of nucleotide and copy number variation on gene expression phenotypes," *Science (New York, N.Y.)*, vol. 315, pp. 848–853, Feb 9 2007.

[108] J. O. Korbel, T. Tirosh-Wagner, A. E. Urban, X. N. Chen, M. Kasowski, L. Dai, F. Grubert, C. Erdman, M. C. Gao, K. Lange, E. M. Sobel, G. M. Barlow, A. S. Aylsworth, N. J. Carpenter, R. D. Clark, M. Y. Cohen, E. Doran, T. Falik-Zaccai, S. O. Lewin, I. T. Lott, B. C. McGillivray, J. B. Moeschler, M. J. Pettenati, S. M. Pueschel, K. W. Rao, L. G. Shaffer, M. Shohat, A. J. V. Riper, D. Warburton, S. Weissman, M. B. Gerstein, M. Snyder, and J. R. Korenberg, "The genetic architecture of Down syndrome phenotypes revealed by high-resolution analysis of human segmental trisomies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 12031–12036, Jul 21 2009.

[109] F. Zhang, W. Gu, M. E. Hurles, and J. R. Lupski, "Copy number variation in human health, disease, and evolution," *Annual review of genomics and human genetics*, vol. 10, pp. 451–481, 2009.

[110] P. J. Campbell, P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li, T. Santarius, L. A. Stebbings, C. Leroy, S. Edkins, C. Hardy, J. W. Teague, A. Menzies, I. Goodhead, D. J. Turner, C. M. Clee, M. A. Quail, A. Cox, C. Brown, R. Durbin, M. E. Hurles, P. A. Edwards, G. R. Bignell, M. R. Stratton, and P. A. Futreal, "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing," *Nature genetics*, vol. 40, pp. 722–729, Jun 2008.

[111] J. Wu, K. R. Grzeda, C. Stewart, F. Grubert, A. E. Urban, M. P. Snyder, and G. T. Marth, "Copy Number Variation detection from 1000 Genomes Project exon capture sequencing data," *BMC bioinformatics*, vol. 13, pp. 305–2105–13–305, Nov 17 2012.

[112] R. E. Handsaker, J. M. Korn, J. Nemesh, and S. A. McCarroll, "Discovery and genotyping of genome structural polymorphism by sequencing on a population scale," *Nature genetics*, vol. 43, pp. 269–276, Mar 2011.

[113] P. L. Deininger, M. A. Batzer, C. A. H. 3rd, and M. H. Edgell, "Master genes in mammalian repetitive DNA amplification," *Trends in genetics*, vol. 8, pp. 307–311, Sep 1992.

[114] R. Cordaux, D. J. Hedges, and M. A. Batzer, "Retrotransposition of Alu elements: how many sources?," *Trends in genetics*, vol. 20, pp. 464–467, Oct 2004.

[115] R. Cordaux and M. A. Batzer, "The impact of retrotransposons on human genome evolution," *Nature reviews: Genetics*, vol. 10, pp. 691–703, Oct 2009.

[116] J. Xing, D. J. Witherspoon, D. A. Ray, M. A. Batzer, and L. B. Jorde, "Mobile DNA elements in primate and human evolution," *American Journal of Physical Anthropology*, vol. Suppl 45, pp. 2–19, 2007.

[117] C. Stewart, "SPANNER: a structural variation detection tool." In preparation.

[118] C. Stewart, D. Kural, M. P. Stromberg, J. A. Walker, M. K. Konkel, A. M. Stutz, A. E. Urban, F. Grubert, H. Y. Lam, W. P. Lee, M. Busby, A. R. Indap, E. Garrison, C. Huff, J. Xing, M. P. Snyder, L. B. Jorde, M. A. Batzer, J. O. Korbel, G. T. Marth, and 1000 Genomes Project, "A comprehensive map of mobile element insertion polymorphisms in humans," *PLoS genetics*, vol. 7, p. e1002236, Aug 2011.

[119] T. M. Keane, K. Wong, and D. J. Adams, "RetroSeq: transposable element discovery from next-generation sequencing data," *Bioinformatics (Oxford, England)*, vol. 29, pp. 389–390, Feb 1 2013.

[120] F. Hormozdiari, I. Hajirasouliha, P. Dao, F. Hach, D. Yorukoglu, C. Alkan, E. E. Eichler, and S. C. Sahinalp, "Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery," *Bioinformatics (Oxford, England)*, vol. 26, pp. i350–7, Jun 15 2010.

[121] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and . G. P. A. Group, "The variant call format and VCFtools," *Bioinformatics (Oxford, England)*, vol. 27, pp. 2156–2158, Aug 1 2011.

[122] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics (Oxford, England)*, vol. 25, pp. 2078–2079, Aug 15 2009.

[123] 1000 Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, pp. 56–65, Nov 1 2012.

[124] J. Xing, Y. Zhang, K. Han, A. H. Salem, S. K. Sen, C. D. Huff, Q. Zhou, E. F. Kirkness, S. Levy, M. A. Batzer, and L. B. Jorde, "Mobile elements create structural variation: analysis of a complete human genome," *Genome research*, vol. 19, pp. 1516–1526, Sep 2009.

[125] C. R. Huang, A. M. Schneider, Y. Lu, T. Niranjan, P. Shen, M. A. Robinson, J. P. Steranka, D. Valle, C. I. Civin, T. Wang, S. J. Wheelan, H. Ji, J. D. Boeke, and K. H. Burns, "Mobile interspersed repeats are major structural variants in the human genome," *Cell*, vol. 141, pp. 1171–1182, Jun 25 2010.

[126] R. C. Iskow, M. T. McCabe, R. E. Mills, S. Torene, W. S. Pittard, A. F. Neuwald, E. G. V. Meir, P. M. Vertino, and S. E. Devine, "Natural mutagenesis of human genomes by endogenous retrotransposons," *Cell*, vol. 141, pp. 1253–1261, Jun 25 2010.

[127] D. J. Witherspoon, J. Xing, Y. Zhang, W. S. Watkins, M. A. Batzer, and L. B. Jorde, "Mobile element scanning (ME-Scan) by targeted high-throughput sequencing," *BMC genomics*, vol. 11, pp. 410–2164–11–410, Jun 30 2010.

[128] C. R. Beck, J. L. Garcia-Perez, R. M. Badge, and J. V. Moran, "LINE-1 elements in structural variation and disease," *Annual review of genomics and human genetics*, vol. 12, pp. 187–215, Sep 22 2011.

[129] A. D. Ewing and H. H. K. Jr, "Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans," *Genome research*, vol. 21, pp. 985–990, Jun 2011.

[130] F. Hormozdiari, C. Alkan, M. Ventura, I. Hajirasouliha, M. Malig, F. Hach, D. Yorukoglu, P. Dao, M. Bakhshi, S. C. Sahinalp, and E. E. Eichler, "Alu repeat discovery and characterization within human genomes," *Genome research*, vol. 21, pp. 840–849, Jun 2011.

[131] J. Wang, L. Song, D. Grover, S. Azrak, M. A. Batzer, and P. Liang, "dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans," *Human mutation*, vol. 27, pp. 323–329, Apr 2006.

[132] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "MEME SUITE: tools for motif discovery and searching," *Nucleic acids research*, vol. 37, pp. W202–8, Jul 2009.

[133] D. E. Knuth, *The art of computer programming*. Reading, Mass.: Addison-Wesley Pub. Co., 1968.

[134] S. Youssef, "Clustering with local equivalence relations," *Computer Physics Communications*, vol. 45(1-3), pp. 423–426, 1987.

[135] W.-P. Lee and J. Wu, "SCISSORS: A Split-Read Mapper for Structural Variants on Next-Generation Sequencing Data." https://github.com/wanpinglee/scissors. In preparation.

[136] M. Zhao, W.-P. Lee, E. Garrison, and G. T. Marth, "SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications," Apr. 2013. www.github.com/mengyao/Complete-Striped-Smith-Waterman-Library.

[137] A. Smit, R. Hubley, and P. Green, "RepeatMasker Open-3.0." http://www.repeatmasker.org, 1996-2010.

[138] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, "The human genome browser at UCSC," *Genome research*, vol. 12, pp. 996–1006, Jun 2002.

[139] O. Kohany, A. J. Gentles, L. Hankus, and J. Jurka, "Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor," *BMC bioinformatics*, vol. 7, p. 474, Oct 25 2006.

[140] D. W. Barnett, E. K. Garrison, A. R. Quinlan, M. P. Stromberg, and G. T. Marth, "BamTools: a C++ API and toolkit for analyzing and managing BAM files," *Bioinformatics (Oxford, England)*, vol. 27, pp. 1691–1692, Jun 15 2011.

[141] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko, "Galaxy: a platform for interactive large-scale genome analysis," *Genome research*, vol. 15, pp. 1451–1455, Oct 2005.

[142] D. Blankenberg, G. V. Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor, "Galaxy: a web-based genome analysis tool for experimentalists," *Current protocols in molecular biology / edited by Frederick M.Ausubel ...[et al.]*, vol. Chapter 19, pp. Unit 19.10.1–21, Jan 2010.

[143] J. Goecks, A. Nekrutenko, J. Taylor, and G. Team, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life

sciences," *Genome biology*, vol. 11, no. 8, pp. R86–2010–11–8–r86. Epub 2010 Aug 25, 2010.

[144] F. M. You, N. Huo, Y. Q. Gu, M. C. Luo, Y. Ma, D. Hane, G. R. Lazo, J. Dvorak, and O. D. Anderson, "BatchPrimer3: a high throughput web application for PCR and sequencing primer design," *BMC bioinformatics*, vol. 9, pp. 253–2105–9–253, May 29 2008.

[145] T. Hall, "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT," *Nucleic Acids Symposium Series*, vol. 41, pp. 95–98, 1999.

[146] S. Rozen and H. Skaletsky, "Primer3 on the WWW for general users and for biologist programmers," *Methods in molecular biology (Clifton, N.J.)*, vol. 132, pp. 365–386, 2000.

[147] N. Tayebi and H. Khodaei, "A Rare Case of Pericentric Inversion, Inv (21) (p12;q22) in Repeated Pregnancy Loss: A Case Report," *Oman medical journal*, vol. 26, pp. 441–443, Nov 2011.

[148] M. R. Mehan, N. B. Freimer, and R. A. Ophoff, "A genome-wide survey of segmental duplications that mediate common human genetic variation of chromosomal architecture," *Human genomics*, vol. 1, pp. 335–344, Aug 2004.

[149] E. E. Schadt, S. Turner, and A. Kasarskis, "A window into third-generation sequencing," *Human molecular genetics*, vol. 19, pp. R227–40, Oct 15 2010.

[150] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation," *Nucleic acids research*, vol. 29, pp. 308–311, Jan 1 2001.

[151] L. Huang, V. Popic, and S. Batzoglou, "Short read alignment with populations of genomes," *Bioinformatics (Oxford, England)*, vol. 29, pp. i361–i370, Jul 1 2013.

[152] K. Schneeberger, J. Hagmann, S. Ossowski, N. Warthmann, S. Gesing, O. Kohlbacher, and D. Weigel, "Simultaneous alignment of short reads against multiple genomes," *Genome biology*, vol. 10, no. 9, pp. R98–2009–10–9–r98. Epub 2009 Sep 17, 2009.

[153] S. L. Salzberg, A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marcais, M. Pop, and J. A. Yorke, "GAGE: A critical evaluation of genome assemblies and assembly algorithms," *Genome research*, vol. 22, pp. 557–567, Mar 2012.

[154] J. T. Simpson and R. Durbin, "Efficient de novo assembly of large genomes using compressed data structures," *Genome research*, vol. 22, pp. 549–556, Mar 2012.

[155] J. Y. Lee, B. S. Lee, D. J. Shin, K. W. Park, Y. A. Shin, K. J. Kim, L. Heo, J. Y. Lee, Y. K. Kim, Y. J. Kim, C. B. Hong, S. H. Lee, D. Yoon, H. J. Ku, I. Y. Oh, B. J. Kim, J. Lee, S. J. Park, J. Kim, H. K. Kawk, J. E. Lee, H. K. Park, J. E. Lee, H. Y. Nam, H. Y. Park, C. Shin, M. Yokota, H. Asano, M. Nakatochi, T. Matsubara, H. Kitajima, K. Yamamoto, H. L. Kim, B. G. Han, M. C. Cho, Y. Jang, H. S. Kim, J. E. Park, and J. Y. Lee, "A genome-wide association study of a coronary artery disease risk variant," *Journal of human genetics*, vol. 58, pp. 120–126, Mar 2013.

[156] X. Lu, L. Wang, S. Chen, L. He, X. Yang, Y. Shi, J. Cheng, L. Zhang, C. C. Gu, J. Huang, T. Wu, Y. Ma, J. Li, J. Cao, J. Chen, D. Ge, Z. Fan, Y. Li, L. Zhao, H. Li, X. Zhou, L. Chen, D. Liu, J. Chen, X. Duan, Y. Hao, L. Wang, F. Lu, Z. Liu, C. Yao, C. Shen, X. Pu, L. Yu, X. Fang, L. Xu, J. Mu, X. Wu, R. Zheng, N. Wu, Q. Zhao, Y. Li, X. Liu, M. Wang, D. Yu, D. Hu, X. Ji, D. Guo, D. Sun, Q. Wang, Y. Yang, F. Liu, Q. Mao, X. Liang, J. Ji, P. Chen, X. Mo, D. Li, G. Chai, Y. Tang, X. Li, Z. Du, X. Liu, C. Dou, Z. Yang, Q. Meng, D. Wang, R. Wang, J. Yang, H. Schunkert, N. J. Samani, S. Kathiresan, M. P. Reilly, J. Erdmann, C. A. D. G.-W. Replication, M.-A. C. Consortium, X. Peng, X. Wu, D. Liu, Y. Yang, R. Chen, B. Qiang, and D. Gu, "Genome-wide association study in Han Chinese identifies four new

susceptibility loci for coronary artery disease," *Nature genetics*, vol. 44, pp. 890–894, Jul 1 2012.

[157] N. Sandholm, R. M. Salem, A. J. McKnight, E. P. Brennan, C. Forsblom, T. Isakova, G. J. McKay, W. W. Williams, D. M. Sadlier, V. P. Makinen, E. J. Swan, C. Palmer, A. P. Boright, E. Ahlqvist, H. A. Deshmukh, B. J. Keller, H. Huang, A. J. Ahola, E. Fagerholm, D. Gordin, V. Harjutsalo, B. He, O. Heikkila, K. Hietala, J. Kyto, P. Lahermo, M. Lehto, R. Lithovius, A. M. Osterholm, M. Parkkonen, J. Pitkaniemi, M. Rosengard-Barlund, M. Saraheimo, C. Sarti, J. Soderlund, A. Soro-Paavonen, A. Syreeni, L. M. Thorn, H. Tikkanen, N. Tolonen, K. Tryggvason, J. Tuomilehto, J. Waden, G. V. Gill, S. Prior, C. Guiducci, D. B. Mirel, A. Taylor, S. M. Hosseini, D. R. Group, H. H. Parving, P. Rossing, L. Tarnow, C. Ladenvall, F. Alhenc-Gelas, P. Lefebvre, V. Rigalleau, R. Roussel, D. A. Tregouet, A. Maestroni, S. Maestroni, H. Falhammar, T. Gu, A. Mollsten, D. Cimponeriu, M. Ioana, M. Mota, E. Mota, C. Serafinceanu, M. Stavarachi, R. L. Hanson, R. G. Nelson, M. Kretzler, H. M. Colhoun, N. M. Panduru, H. F. Gu, K. Brismar, G. Zerbini, S. Hadjadj, M. Marre, L. Groop, M. Lajer, S. B. Bull, D. Waggott, A. D. Paterson, D. A. Savage, S. C. Bain, F. Martin, J. N. Hirschhorn, C. Godson, J. C. Florez, P. H. Groop, and A. P. Maxwell, "New susceptibility loci associated with kidney disease in type 1 diabetes," *PLoS genetics*, vol. 8, p. e1002921, Sep 2012.

[158] H. Li, W. Gan, L. Lu, X. Dong, X. Han, C. Hu, Z. Yang, L. Sun, W. Bao, P. Li, M. He, L. Sun, Y. Wang, J. Zhu, Q. Ning, Y. Tang, R. Zhang, J. Wen, D. Wang, X. Zhu, K. Guo, X. Zuo, X. Guo, H. Yang, X. Zhou, D. Consortium, A.-T. Consortium, X. Zhang, L. Qi, R. J. Loos, F. B. Hu, T. Wu, Y. Liu, L. Liu, Z. Yang, R. Hu, W. Jia, L. Ji, Y. Li, and X. Lin, "A genome-wide association study identifies GRK5 and RASGRP1 as type 2 diabetes loci in Chinese Hans," *Diabetes*, vol. 62, pp. 291–298, Jan 2013.

[159] J. M. Hotaling, D. R. Waggott, J. Goldberg, G. Jarvik, A. D. Paterson, P. A. Cleary, J. Lachin, A. Sarma, H. Wessells, and D. R. Group, "Pilot genome-wide association search

identifies potential loci for risk of erectile dysfunction in type 1 diabetes using the DCCT/EDIC study cohort," *The Journal of urology*, vol. 188, pp. 514–520, Aug 2012.

[160] M. I. Kamboh, F. Y. Demirci, X. Wang, R. L. Minster, M. M. Carrasquillo, V. S. Pankratz, S. G. Younkin, A. J. Saykin, A. D. N. Initiative, G. Jun, C. Baldwin, M. W. Logue, J. Buros, L. Farrer, M. A. Pericak-Vance, J. L. Haines, R. A. Sweet, M. Ganguli, E. Feingold, S. T. Dekosky, O. L. Lopez, and M. M. Barmada, "Genome-wide association study of Alzheimer's disease," *Translational psychiatry*, vol. 2, p. e117, May 15 2012.

[161] J. C. Lambert, B. Grenier-Boley, D. Harold, D. Zelenika, V. Chouraki, Y. Kamatani, K. Sleegers, M. A. Ikram, M. Hiltunen, C. Reitz, I. Mateo, T. Feulner, M. Bullido, D. Galimberti, L. Concari, V. Alvarez, R. Sims, A. Gerrish, J. Chapman, C. Deniz-Naranjo, V. Solfrizzi, S. Sorbi, B. Arosio, G. Spalletta, G. Siciliano, J. Epelbaum, D. Hannequin, J. F. Dartigues, C. Tzourio, C. Berr, E. M. Schrijvers, R. Rogers, G. Tosto, F. Pasquier, K. Bettens, C. V. Cauwenberghe, L. Fratiglioni, C. Graff, M. Delepine, R. Ferri, C. A. Reynolds, L. Lannfelt, M. Ingelsson, J. A. Prince, C. Chillotti, A. Pilotto, D. Seripa, A. Boland, M. Mancuso, P. Bossu, G. Annoni, B. Nacmias, P. Bosco, F. Panza, F. Sanchez-Garcia, M. D. Zompo, E. Coto, M. Owen, M. O'Donovan, F. Valdivieso, P. Caffarra, E. Scarpini, O. Combarros, L. Buee, D. Campion, H. Soininen, M. Breteler, M. Riemenschneider, C. V. Broeckhoven, A. Alperovitch, M. Lathrop, D. A. Tregouet, J. Williams, and P. Amouyel, "Genome-wide haplotype association study identifies the FRMD4A gene as a risk locus for Alzheimer's disease," *Molecular psychiatry*, vol. 18, pp. 461–470, Apr 2013.

[162] M. C. Dubinsky, S. Kugathasan, S. Kwon, T. Haritunians, I. Wrobel, G. Wahbeh, A. Quiros, R. Bahar, G. Silber, S. Farrior, M. Stephens, N. Teleten, D. Panikkath, A. Ippoliti, E. Vasiliauskas, P. Fleshner, C. Williams, C. Landers, J. I. Rotter, S. R. Targan, K. D. Taylor, and D. P. McGovern, "Multidimensional prognostic risk assessment identifies association

between IL12B variation and surgery in Crohn's disease," *Inflammatory bowel diseases*, vol. 19, pp. 1662–1670, Jul 2013.

[163] K. Yamazaki, J. Umeno, A. Takahashi, A. Hirano, T. A. Johnson, N. Kumasaka, T. Morizono, N. Hosono, T. Kawaguchi, M. Takazoe, T. Yamada, Y. Suzuki, H. Tanaka, S. Motoya, M. Hosokawa, Y. Arimura, Y. Shinomura, T. Matsui, T. Matsumoto, M. Iida, T. Tsunoda, Y. Nakamura, N. Kamatani, and M. Kubo, "A genome-wide association study identifies 2 susceptibility Loci for Crohn's disease in a Japanese population," *Gastroenterology*, vol. 144, pp. 781–788, Apr 2013.

[164] L. Jostins, S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, J. Essers, M. Mitrovic, K. Ning, I. Cleynen, E. Theatre, S. L. Spain, S. Raychaudhuri, P. Goyette, Z. Wei, C. Abraham, J. P. Achkar, T. Ahmad, L. Amininejad, A. N. Ananthakrishnan, V. Andersen, J. M. Andrews, L. Baidoo, T. Balschun, P. A. Bampton, A. Bitton, G. Boucher, S. Brand, C. Buning, A. Cohain, S. Cichon, M. D'Amato, D. D. Jong, K. L. Devaney, M. Dubinsky, C. Edwards, D. Ellinghaus, L. R. Ferguson, D. Franchimont, K. Fransen, R. Gearry, M. Georges, C. Gieger, J. Glas, T. Haritunians, A. Hart, C. Hawkey, M. Hedl, X. Hu, T. H. Karlsen, L. Kupcinskas, S. Kugathasan, A. Latiano, D. Laukens, I. C. Lawrance, C. W. Lees, E. Louis, G. Mahy, J. Mansfield, A. R. Morgan, C. Mowat, W. Newman, O. Palmieri, C. Y. Ponsioen, U. Potocnik, N. J. Prescott, M. Regueiro, J. I. Rotter, R. K. Russell, J. D. Sanderson, M. Sans, J. Satsangi, S. Schreiber, L. A. Simms, J. Sventoraityte, S. R. Targan, K. D. Taylor, M. Tremelling, H. W. Verspaget, M. D. Vos, C. Wijmenga, D. C. Wilson, J. Winkelmann, R. J. Xavier, S. Zeissig, B. Zhang, C. K. Zhang, H. Zhao, I. I. G. C. (IIBDGC), M. S. Silverberg, V. Annese, H. Hakonarson, S. R. Brant, G. Radford-Smith, C. G. Mathew, J. D. Rioux, E. E. Schadt, M. J. Daly, A. Franke, M. Parkes, S. Vermeire, J. C. Barrett, and J. H. Cho, "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease," *Nature*, vol. 491, pp. 119–124, Nov 1 2012.

[165] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 9362–9367, Jun 9 2009.

[166] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, pp. 747–753, Oct 8 2009.

[167] G. M. Church, "The personal genome project," *Molecular systems biology*, vol. 1, p. 2005.0030, 2005.

[168] E. A. Ashley, A. J. Butte, M. T. Wheeler, R. Chen, T. E. Klein, F. E. Dewey, J. T. Dudley, K. E. Ormond, A. Pavlovic, A. A. Morgan, D. Pushkarev, N. F. Neff, L. Hudgins, L. Gong, L. M. Hodges, D. S. Berlin, C. F. Thorn, K. Sangkuhl, J. M. Hebert, M. Woon, H. Sagreiya, R. Whaley, J. W. Knowles, M. F. Chou, J. V. Thakuria, A. M. Rosenbaum, A. W. Zaranek, G. M. Church, H. T. Greely, S. R. Quake, and R. B. Altman, "Clinical assessment incorporating a personal genome," *Lancet*, vol. 375, pp. 1525–1535, May 1 2010.

[169] J. E. Lunshof, J. Bobe, J. Aach, M. Angrist, J. V. Thakuria, D. B. Vorhaus, M. R. Hoehe, and G. M. Church, "Personal genomes in progress: from the human genome project to the personal genome project," *Dialogues in clinical neuroscience*, vol. 12, no. 1, pp. 47–60, 2010.

[170] A. M. Soto and C. Sonnenschein, "The somatic mutation theory of cancer: growing problems with the paradigm?," *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 26, pp. 1097–1107, Oct 2004.

[171] P. C. Nowell, "The clonal evolution of tumor cell populations," *Science (New York, N.Y.)*, vol. 194, pp. 23–28, Oct 1 1976.

[172] M. Casas-Selves and J. Degregori, "How cancer shapes evolution, and how evolution shapes cancer," *Evolution*, vol. 4, pp. 624–634, Dec 2011.

[173] T. Tian, S. Olson, J. M. Whitacre, and A. Harding, "The origins of cancer robustness and evolvability," *Integrative biology : quantitative biosciences from nano to macro*, vol. 3, pp. 17–30, Jan 2011.

[174] M. Shackleton, E. Quintana, E. R. Fearon, and S. J. Morrison, "Heterogeneity in cancer: cancer stem cells versus clonal evolution," *Cell*, vol. 138, pp. 822–829, Sep 4 2009.

[175] L. X. Qin, "Chromosomal aberrations related to metastasis of human solid tumors," *World journal of gastroenterology : WJG*, vol. 8, pp. 769–776, Oct 2002.

[176] G. Bardi, B. Johansson, N. Pandis, N. Mandahl, E. Bak-Jensen, A. Andren-Sandberg, F. Mitelman, and S. Heim, "Karyotypic abnormalities in tumours of the pancreas," *British journal of cancer*, vol. 67, pp. 1106–1112, May 1993.

[177] M. Ozery-Flato, C. Linhart, L. Trakhtenbrot, S. Izraeli, and R. Shamir, "Large-scale analysis of chromosomal aberrations in cancer karyotypes reveals two distinct paths to aneuploidy," *Genome biology*, vol. 12, pp. R61–2011–12–6–r61, Jun 29 2011.

[178] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X. W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan, "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer," *Science (New York, N.Y.)*, vol. 310, pp. 644–648, Oct 28 2005.

[179] M. Soda, Y. L. Choi, M. Enomoto, S. Takada, Y. Yamashita, S. Ishikawa, S. Fujiwara, H. Watanabe, K. Kurashina, H. Hatanaka, M. Bando, S. Ohno, Y. Ishikawa, H. Aburatani,

T. Niki, Y. Sohara, Y. Sugiyama, and H. Mano, "Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer," *Nature*, vol. 448, pp. 561–566, Aug 2 2007.

[180] A. R. Carson, W. Pfeiffer, T. Schwartz, G. Oliveira, T. Nicholas, G. Zhang, M. A. Miller, E. J. Topol, and S. Levy, "Detection of somatic mutations in tumor genomes using de novo assembly with assembly to assembly mapping," American Society of Human Genetics Meeting, 2012.

[181] M. F. Fraga, E. Ballestar, M. F. Paz, S. Ropero, F. Setien, M. L. Ballestar, D. Heine-Suner, J. C. Cigudosa, M. Urioste, J. Benitez, M. Boix-Chornet, A. Sanchez-Aguilera, C. Ling, E. Carlsson, P. Poulsen, A. Vaag, Z. Stephan, T. D. Spector, Y. Z. Wu, C. Plass, and M. Esteller, "Epigenetic differences arise during the lifetime of monozygotic twins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 10604–10609, Jul 26 2005.

[182] S. Costa and P. Shaw, "'Open minded' cells: how cells can change fate," *Trends in cell biology*, vol. 17, pp. 101–106, Mar 2007.

[183] Z. A. Kaminsky, T. Tang, S. C. Wang, C. Ptak, G. H. Oh, A. H. Wong, L. A. Feldcamp, C. Virtanen, J. Halfvarson, C. Tysk, A. F. McRae, P. M. Visscher, G. W. Montgomery, I. I. Gottesman, N. G. Martin, and A. Petronis, "DNA methylation profiles in monozygotic and dizygotic twins," *Nature genetics*, vol. 41, pp. 240–245, Feb 2009.

[184] M. Lalande and M. A. Calciano, "Molecular epigenetics of Angelman syndrome," *Cellular and molecular life sciences : CMLS*, vol. 64, pp. 947–960, Apr 2007.

[185] T. Ohta, T. A. Gray, P. K. Rogan, K. Buiting, J. M. Gabriel, S. Saitoh, B. Muralidhar, B. Bilienska, M. Krajewska-Walasek, D. J. Driscoll, B. Horsthemke, M. G. Butler, and R. D. Nicholls, "Imprinting-mutation mechanisms in Prader-Willi syndrome," *American Journal of Human Genetics*, vol. 64, pp. 397–413, Feb 1999.

[186] D. Viljoen and R. Ramesar, “Evidence for paternal imprinting in familial Beckwith-Wiedemann syndrome,” *Journal of medical genetics*, vol. 29, pp. 221–225, Apr 1992.

[187] R. Weksberg, A. C. Smith, J. Squire, and P. Sadowski, “Beckwith-Wiedemann syndrome demonstrates a role for epigenetic control of normal development,” *Human molecular genetics*, vol. 12 Spec No 1, pp. R61–8, Apr 1 2003.

[188] M. Esteller, J. M. Silva, G. Dominguez, F. Bonilla, X. Matias-Guiu, E. Lerma, E. Bussaglia, J. Prat, I. C. Harkes, E. A. Repasky, E. Gabrielson, M. Schutte, S. B. Baylin, and J. G. Herman, “Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors,” *Journal of the National Cancer Institute*, vol. 92, pp. 564–569, Apr 5 2000.

[189] M. Sanchez-Cespedes, M. Esteller, L. Wu, H. Nawroz-Danish, G. H. Yoo, W. M. Koch, J. Jen, J. G. Herman, and D. Sidransky, “Gene promoter hypermethylation in tumors and serum of head and neck cancer patients,” *Cancer research*, vol. 60, pp. 892–895, Feb 15 2000.

[190] M. Esteller, E. Avizienyte, P. G. Corn, R. A. Lothe, S. B. Baylin, L. A. Aaltonen, and J. G. Herman, “Epigenetic inactivation of LKB1 in primary tumors associated with the Peutz-Jeghers syndrome,” *Oncogene*, vol. 19, pp. 164–168, Jan 6 2000.

[191] R. Agrelo, W. H. Cheng, F. Setien, S. Ropero, J. Espada, M. F. Fraga, M. Herranz, M. F. Paz, M. Sanchez-Cespedes, M. J. Artiga, D. Guerrero, A. Castells, C. von Kobbe, V. A. Bohr, and M. Esteller, “Epigenetic inactivation of the premature aging Werner syndrome gene in human cancer,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 8822–8827, Jun 6 2006.

[192] L. Shen, Y. Kondo, G. L. Rosner, L. Xiao, N. S. Hernandez, J. Vilaythong, P. S. Houlihan, R. S. Krouse, A. R. Prasad, J. G. Einspahr, J. Buckmeier, D. S. Alberts, S. R. Hamilton, and

J. P. Issa, "MGMT promoter methylation and field defect in sporadic colorectal cancer," *Journal of the National Cancer Institute*, vol. 97, pp. 1330–1338, Sep 21 2005.

[193] K. H. Lee, J. S. Lee, J. H. Nam, C. Choi, M. C. Lee, C. S. Park, S. W. Juhng, and J. H. Lee, "Promoter methylation status of hMLH1, hMSH2, and MGMT genes in colorectal cancer associated with adenoma-carcinoma sequence," *Langenbeck's archives of surgery / Deutsche Gesellschaft fur Chirurgie*, vol. 396, pp. 1017–1026, Oct 2011.

[194] A. Facista, H. Nguyen, C. Lewis, A. R. Prasad, L. Ramsey, B. Zaitlin, V. Nfonsam, R. S. Krouse, H. Bernstein, C. M. Payne, S. Stern, N. Oatman, B. Banerjee, and C. Bernstein, "Deficient expression of DNA repair enzymes in early progression to sporadic colon cancer," *Genome integrity*, vol. 3, pp. 3–9414–3–3, Apr 11 2012.

[195] D. Koutsimpelas, W. Pongsapich, U. Heinrich, S. Mann, W. J. Mann, and J. Brieger, "Promoter methylation of MGMT, MLH1 and RASSF1A tumor suppressor genes in head and neck squamous cell carcinoma: pharmacological genome demethylation reduces proliferation of head and neck squamous carcinoma cells," *Oncology reports*, vol. 27, pp. 1135–1141, Apr 2012.

[196] J. Brieger, S. A. Mann, W. Pongsapich, D. Koutsimpelas, K. Fruth, and W. J. Mann, "Pharmacological genome demethylation increases radiosensitivity of head and neck squamous carcinoma cells," *International journal of molecular medicine*, vol. 29, pp. 505–509, Mar 2012.

[197] C. A. Miller, O. Hampton, C. Coarfa, and A. Milosavljevic, "ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads," *PloS one*, vol. 6, p. e16327, Jan 31 2011.

[198] N. Li, M. Ye, Y. Li, Z. Yan, L. M. Butcher, J. Sun, X. Han, Q. Chen, X. Zhang, and J. Wang, "Whole genome DNA methylation analysis based on high throughput sequencing technology," *Methods (San Diego, Calif.)*, vol. 52, pp. 203–212, Nov 2010.

[199] E. Meaburn and R. Schulz, “Next generation sequencing in epigenetics: insights and challenges,” *Seminars in cell & developmental biology*, vol. 23, pp. 192–199, Apr 2012.