ALGORITHMS FOR STRUCTURAL VARIATION DISCOVERY USING MULTIPLE SEQUENCE SIGNATURES

A DISSERTATION SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

COMPUTER ENGINEERING

By

Arda Söylev

September 2018

Algorithms for Structural Variation Discovery Using Multiple Sequence Signatures
By Arda Söylev
September 2018

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Can Alkan(Advisor)

Mehmet Somel

A. Ercüment Çiçek

Çiğdem Gündüz Demir

Elif Sürer

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

ALGORITHMS FOR STRUCTURAL VARIATION DISCOVERY USING MULTIPLE SEQUENCE SIGNATURES

Arda Söylev
Ph.D. in Computer Engineering
Advisor: Can Alkan
September 2018

Genomic variations including single nucleotide polymorphisms (SNPs), small INDELs and structural variations (SVs) are known to have significant phenotypic effects on individuals. Among them, SVs, that alter more than 50 nucleotides of DNA, are the major source of complex genetic diseases such as Crohn's, schizophrenia and autism. Additionally, the total number of nucleotides affected by SVs are substantially higher than SNPs (3.5 Mbp SNP, 15-20 Mbp SV). Today, we are able to perform whole genome sequencing (WGS) by utilizing high throughput sequencing technology (HTS) to discover these modifications unimaginably faster, cheaper and more accurate than before. However, as demonstrated in the 1000 Genomes Project, HTS technology still has significant limitations. The major problem lies in the short read lengths (<250 bp) produced by the current sequencing platforms and the fact that most genomes include large amounts of repeats make it very challenging to unambiguously map and accurately characterize genomic variants. Thus, most of the existing SV discovery tools focus on detecting relatively simple types of SVs such as insertions, deletions, and short inversions. In fact, other types of SVs including the complex ones are of crucial importance and several have been associated with genomic disorders. To better understand the contribution of these SVs to human genome, we need new approaches to accurately discover and genotype such variants. Therefore, there is still a need for accurate algorithms to fully characterize a broader spectrum of SVs and thus improve calling accuracy of more simple variants.

Here we introduce TARDIS that harbors novel algorithms to accurately characterize various types of SVs including deletions, novel sequence insertions, inversions, transposon insertions, nuclear mitochondria insertions, tandem duplications and interspersed segmental duplications in direct or inverted orientations using short read whole genome sequencing datasets. Within our framework, we make use of multiple sequence signatures including read pair, read depth and

split read in order to capture different sequence signatures and increase our SV prediction accuracy. Additionally, we are able to analyze more than one possible mapping location of each read to overcome the problems associated with repeated nature of genomes. Recently, due to the limitations of short-read sequencing technology, newer library preparation techniques emerged and 10x Genomics is one of these initiatives. This technique is regarded as a cost-effective alternative to long read sequencing, which can obtain long range contiguity information. We extended TARDIS to be able to utilize Linked-Read information of 10x Genomics to overcome some of the constraints of short-read sequencing technology.

We evaluated the prediction performance of our algorithms through several experiments using both simulated and real data sets. In the simulation experiments, TARDIS achieved 97.67% sensitivity with only 1.12% false discovery rate. For experiments that involve real data, we used two haploid genomes (CHM1 and CHM13) and one human genome (NA12878) from the Illumina Platinum Genomes set. Comparison of our results with orthogonal PacBio call sets from the same genomes revealed higher accuracy for TARDIS than state of the art methods. Furthermore, we showed a surprisingly low false discovery rate of our approach for discovery of tandem, direct and inverted interspersed segmental duplications prediction on CHM1 (less than 5% for the top 50 predictions). The algorithms we describe here are the first to predict insertion location and the various types of new segmental duplications using HTS data.

Keywords: Structural variation, high throughput sequencing, combinatorial algorithms.

ÖZET

ÇOKLU DİZİ SİNYALLERİ KULLANARAK YAPISAL VARYASYON KEŞFİ İÇİN ALGORİTMALAR

Arda Söylev
Bilgisayar Mühendisliği, Doktora
Tez Danışmanı: Can Alkan
Eylül 2018

Tek nükleotid polimorfizmi (TNP), baz çifti ekleme/çıkarma (Indel) ve yapısal varyasyon (YV) gibi genetik varyasyonların canlılar üzerinde önemli fenotipik etkileri vardır. Bunların içinde 50'den fazla baz çiftini etkileyen YV'ler, Crohn Hastalığı, şizofreni ve otizm gibi çeşitli kalıtsal hastalıkların da temel sebebidir. Ayrıca YV'lerin etkilediği baz çifti sayısı TNP'lere göre çok daha fazladır (3,5 Mbp TNP, 15-20 Mbp YV). Bugün, yeni nesil dizileme (YND) teknolojisini kullanarak tam genom hizalama (WGS) yapabiliyor ve bu tip varyasyonları çok daha hızlı, ucuz ve yüksek doğrulukla keşfedebiliyoruz. Ancak 1000 Genom Projesi'nde de gördüğümüz gibi, YND teknolojisinin bazı yetersizlikleri vardır. En önemli sorun şu an kullanılan YND platformlarının ürettiği kısa okuma (<250 bp) boyutları ve genomların çok tekrarlı bölgeler barındırması sebebiyle bu kısa okumaların yüksek doğrulukla hizalanmasını zorlaştırmasıdır. Bu durum, keşfedilen genomik varyasyonların doğruluk oranını da etkilemektedir. Bu sebeple, bugüne kadar geliştirilmiş algoritmalar ekleme, silinme ve kısa inversiyonlar gibi görece olarak daha basit YV'leri karakterize edebilmesine rağmen birçok genetik hastalıkla bağdaştırılan daha karmaşık varyasyonları göz ardı etmiştir. Bu tip YV'lerin insan genomuna etkilerini gözlemlemek için daha farklı yaklaşımlar kullanan, yüksek doğruluk oranına sahip yeni algoritmalar gerekmektedir.

Bu tezde, YND teknolojisiyle kısa okumaları kullanarak bir canlının genomundaki YV'leri bulan TARDIS algoritmasını tanıtıyoruz. TARDIS; silinme, yeni dizi ekleme, inversiyon, transpozon ekleme, mitokondriyal ekleme, ardışık kopya ve ters/düz ayrışık kopya gibi birçok YV'yi karakterize edebilmektedir. Bu varyasyonların yüksek doğrulukta keşfi için okuma çiftleri, okuma derinliği ve ayrık okumalar gibi farklı sinyalleri birarada kullanmaktadır. Ayrıca TARDIS, genomun tekrarlı yapısı sebebiyle aynı okumanın birden çok yere benzer doğrulukta hizalanmasından dolayı oluşan hataları göz önünde bulundurarak, tüm hizalanma lokasyonlarını da kullanabilme özelliğine sahiptir. Son zamanlarda

kısa okumaların barındırdığı kısıtlamalar sebebiyle yeni kütüphane hazırlama protokolleri geliştirilmiştir. 10x Genomics de bunlardan biridir. Bu teknik, düşük maliyetle uzun mesafeli bitişiklik bilgisi (Long range contiguity) sağlayan, yüksek maliyetli uzun okumalara alternatif bir yöntemdir. TARDIS, kısa okumaların sebep olduğu kısıtlamaların önüne geçebilmek için 10x Genomics'in bağlantılı okumalarını da kullanabilmektedir.

Geliştirdiğimiz algoritmaların doğruluk oranlarını simülasyon ve gerçek veriler kullanarak değerlendirdik. Simülasyonlarda TARDIS %97,67 hassasiyet ve %1,12 hatalı tahmin oranını yakaladı. Gerçek veri deneyleri için de iki haploid (CHM1 ve CHM13) ve bir diploid (NA12878) insan genomu kullandık. Sonuçları PacBio veri setleriyle karşılaştırdığımızda TARDIS'in literatürdeki en başarılı metotlara göre daha yüksek doğruluğa sahip olduğunu gördük. Ayrıca CHM1 genomu için TARDIS'in ardışık ve ayrışık kopya varyasyonlarında çok düşük hata oranına sahip olduğunu gösterdik (En iyi 50 tahmininde hata oranı %5'den azdır). Son olarak belirtmeliyiz ki burada tanıttığımız algoritmalar YND teknolojisini kullanarak ayrışık yapısal varyasyonları karakterize edebilen ilk algoritmalardır.

Anahtar sözcükler: Yapısal varyasyon, yeni nesil dizileme, kombinatorik algoritmalar.

Acknowledgement

First and foremost I wish to express my sincere gratitude to my advisor Dr. Can Alkan. He gave me the opportunity to work in a field where molecular biology and computer science meets. I wouldn't have imagined such an excellent research field that fits my intentions. He was always sensible, patient and through his immense knowledge and vision, he guided me to great ideas and problems to work on. I want to thank him once more.

I would like to give my appreciation to Dr. Fereydoun Hormozdiari. With his suggestions and contributions, we introduced novel approaches in the field. My visit to his lab in UC Davis was a great experience and it influenced the direction of my research. His motivation and hard work was a model for me. I owe him much and hope to carry on our collaboration.

I would also like to thank the members of my dissertation committee; Dr. Mehmet Somel, Dr. A. Ercüment Çiçek, Dr. Çiğdem Gündüz Demir and Dr. Elif Sürer for their insightful comments and encouragements, which allowed me to widen my research from various perspectives.

Besides, I would like to thank our collaborators Iman Hajirasouliha, Camir Ricketts, Thong Minh Le, Baraa Orabi, Ezgi Ebren, Fatih Karaoğlanoğlu and all the members of Alkan Lab.

I also acknowledge the support given by TUBITAK through a 1001 grant (215E172).

I would also like to express my deepest gratitude to my family; my wife Zeynep and my little son Hasan Emre, who have given me the strength and motivation throughout my PhD education.

Last, but not least, I would like to thank my parents. My mother and father always supported me unconditionally from the day I born until today. I wouldn't have done any of these without their presence. This thesis is dedicated to them.

Contents

1 Introduction **1**

1.1 DNA and Computational Genomics 1

1.2 Genomic Variation: Changes in DNA Sequence 2

1.2.1 SNPs, INDELs and STRs 4

1.2.2 Structural Variations 4

1.3 High Throughput Sequencing 8

1.3.1 Short-read sequencing 8

1.3.2 Long-read sequencing 10

1.3.3 Linked-Read sequencing 11

1.4 Reference Based Analysis 12

1.4.1 Read mapping 12

1.5 *De novo* Assembly 13

1.6 Structural Variation Discovery Signatures 16

1.6.1 Read-pair 17

1.6.2 Read-depth 20

1.6.3 Split read 21

1.6.4 Assembly 22

1.7 Contributions 22

2 Overview of TARDIS: Toolkit for Automated and Rapid Discovery of Structural Variants **26**

2.1 Introduction 26

2.1.1 Motivation 26

2.1.2 Our approach 27

2.2 Read Mapping 29

2.2.1 Quick Mode . . . 30
2.2.2 Sensitive Mode . . . 31
2.3 SV Discovery via Maximum Parsimony . . . 32
2.3.1 Building clusters . . . 33
2.3.2 Set-Cover approximation to find putative SVs . . . 34
2.4 Using Linked-Read Information . . . 39

3 Structural Variation Discovery with TARDIS 43
3.1 Introduction . . . 43
3.2 Characterizing Various Types of SV . . . 44
3.2.1 Discovering deletions and insertions . . . 44
3.2.2 Characterizing inversions . . . 47
3.2.3 Transposon insertions . . . 49
3.2.4 Nuclear mitochondria (NUMT) insertions . . . 52
3.2.5 Duplications . . . 54
3.3 Incorporating Split Read Information To Improve SV Calls . . . 58
3.3.1 Detection and clustering of split reads . . . 59
3.3.2 Runtime and memory usage of split reads . . . 62

4 Results 67
4.1 Simulation . . . 68
4.2 Real Data Experiments . . . 69
4.2.1 Deletions . . . 70
4.2.2 Inversions . . . 72
4.2.3 Duplications . . . 73
4.2.4 Insertions . . . 74
4.3 Sensitive Mode . . . 76
4.4 Linked-Reads . . . 78
4.5 Time and Memory Consumption . . . 80

5 Conclusion and Discussion 82
5.1 Future Work . . . 84

A Data and Code Availability 120

List of Figures

1.1 Structural Variations (SV) types of deletion, insertion, inversion, mobile element insertion (MEI), interspersed segmental duplication with direct and inverted orientations, tandem duplication and translocation are depicted. 5
1.2 Sequencing approach employed by the first generation sequencers, i.e., Sanger Sequencing. Briefly, DNA is cut into multiple fragments randomly and each fragment is sequenced using clones. . . 8
1.3 The change of cost to sequence a genome between years 2001 - 2017. Rapid decrease in 2007-2008 demonstrates the transition to HTS platforms. Data is retrieved from [1]. 9
1.4 Whole Genome Shotgun (WGS) sequencing. Firstly, DNA is fragmented randomly and then each fragment is sequenced from both left and right ends using a sequencer. These are called paired-end reads. The number of base-pairs that the sequencer reads called read length depends on the sequencing machine used. Finally, the reads; (a) can be assembled into contigs when a reference genome is not available or (2) can be mapped to a reference genome . . . 14
1.5 Read-pair signatures of each SV event is shown. Reads are mapped to the reference genome and the distance between the mate-pairs or the orientation of the mapped reads determine whether a potential SV is implied or not. 19

1.6 Span size distribution of read-pairs sampled from NA11930 genome. When the read-pairs are mapped to the reference genome, distance between them is a possible indication of a genomic variation. If the distance is below or above the expected cut-off values; $min = mean - (4 \times stdev) = 170$, $max = mean + (4 \times stdev) = 500$ in this case; then these read-pairs are called "discordant" and they are the ones considered as potential SVs [2]. 20

1.7 Read-depth signatures of SV events are shown. When reads are mapped a region, divergence from the distribution implies a deletion (decrease in read-depth) or a duplication (increase in read-depth) event. 21

1.8 Split read signatures of SV events are displayed. Here, unmapped reads are split and each fragment is remapped to the reference genome to observe potential SVs. 25

2.1 Read-pair sequence signatures of some SV events are the same such as inversions, interspersed inverted duplications and gene conversions. Similarly deletions, interspersed direct duplications and gene conversions also show the same signature. 35

2.2 Details of 10x technology. Each step of 10x pipeline is briefly depicted. Firstly, 1 ng of high molecular weight DNA is extracted from the sample and distributed to approximately 10^6 pools, where they are barcoded and subjected to priming and polymerase amplification. After the library preparation process, they undergo Illumina sequencing process. 40

3.1 Figure shows the IGV [3, 4] visualization of a deletion event predicted by TARDIS within 19:8,231,867-8,256,118 for CHM1 genome [5, 6]. Absence or decrease of read-depth within the breakpoint is an indication of a deletion. 45

3.2 Read pairs mapped to the reference genome A) insertion signature, B) deletion signature. 46

3.3 Figure shows the IGV output of an inversion event predicted by TARDIS within 3:44,740,482-44,743,019 for CHM1 genome. . . . 48

3.4 Inversion signature of the read pairs mapped to the reference genome. 49
3.5 An example of a false SV prediction is depicted in the figure. There is a deletion event in the left mapping when the duplication in the genome is not considered. We need to check wheter any of the pairs hit the annotated transposon interval in order to make a correct prediction since the MEI insertions can be underestimated. 50
3.6 The figure depicts the overview of MEI clustering approach we utilize in TARDIS. (A) We first check the paired-end reads where one end maps to an annotated transposon and the other to elsewhere within the genome. (B) For such cases, we cluster the pairs that map to elsewhere in the the genome based on their orientations within an interval. Then we bring forward and reverse pairs together inside the same cluster and treat them as paired-end reads in order to detect the insertion breakpoints. 51
3.7 There are four different cases for mobile elements (copy events) in direct orientation. The cases are based on the position of P_{Br}, and orientation of the mappings. 52
3.8 There are 4 different cases for mobile elements (copy events) in inverted orientation. The cases are based on the position of P_{Br}, and orientation of the mappings. 53
3.9 For NUMT insertion, we check whether any of the pair maps to mitochondria. Such cases is an indication of NUMT insertions within the genome. 54
3.10 Relative abundance of complex SVs among the inversion calls reported in the 1000 Genomes Project [7]. 54% of predicted inversions are in fact inverted duplications and only 20% are correctly predicted as simple inversions. 55
3.11 Figure shows the IGV output of a tandem duplication event predicted by TARDIS within 3:1,216,580-1,217,848 for CHM1 genome. 56
3.12 Tandem duplication signature of the read pairs mapped to the reference genome. 57
3.17 Mapping soft-clips to the reference genome. 61

3.13 IGV visualization of interspersed SD in A) direct orientation and B) inverted orientation. It should be clear that the signature in (A) is +/− and −/+, in (B) −/− and +/+. The first one is exactly the same as the signature of deletion and tandem duplication, the second one as inversion. 63
3.14 The sequence signatures for interspersed SDs in (A) inverted (B) direct orientations. 64
3.15 Figure shows some reads mapped to the reference genome with multiple mappings allowed. We also show how the reads align with some mismatches allowed. The nucleotides in red color are the mismatches. 65
3.16 When aligning a read to a reference genome, some bases match, some don't. SAM/BAM specification outputs this information in a CIGAR string. The position of the read aligned to the reference is 0-based starting position of the alignment. 65
3.18 Split read signatures used by TARDIS to characterize SV types of deletion, novel sequence insertion, transposon insertion, inversion, tandem duplication and interspersed segmental duplication in direct and inverted orientations. Briefly, when a read is mapped to the reference genome, the SV is hidden inside the read and this is resolved by splitting the read into two segments. 66

4.1 Comparison of deletion accuracy (>100 bp) between TARDIS and LUMPY using NA12878 genome (a). We also provide a deletion length histogram (b) exhibiting the expected peaks at 300 bp and 5,900 bp for ALU and L1 deletions 72
4.2 Receiver operator characteristic (ROC) curve for the comparison of inversion predictions on CHM1 and CHM13 datasets. Overall, TARDIS achieves better area under the curve (AUC) statistics compared to the other tools. (a), (b) comparison of CHM1 and CHM13 predicted inversions using PacBio reads based on BLASR mappings. (c) validation of top predicted inversion of different tools using local assembly of the PacBio reads of CHM1. 73

4.3 Comparison of top inversion prediction on NA12878 sample against predicted and validated set of inversion of the same samples using PacBio data from [8] 74
4.4 a) Illumina signature for an inverted duplication, b) PacBio validation. 76
4.5 Alu insertion predictions in CHM1 and CHM13 datasets, compared against PacBio calls [9]. 77

List of Tables

1.1 SV discovery algorithms that use short reads. 18

2.1 Mandatory fields of mrFAST output and how TARDIS handles them. 31
2.2 Formulation for λ_d and λ_p for maximum valid cluster S_i 38

4.1 Summary of simulation predictions by TARDIS, LUMPY and DELLY. 69
4.2 Characterization of different types of segmental duplications using TARDIS on simulated data. 70
4.3 Properties of the datasets we utilized in our eperiments. 71
4.4 Comparison of deletion accuracy between TARDIS, LUMPY and DELLY using CHM1, CHM13 and NA12878 data sets 71
4.5 50 highest scoring segmental duplications predicted by TARDIS in the CHM1 genome. 75
4.6 Comparison of simulation predictions for Sensitive and Quick Mode of TARDIS. 77
4.7 Comparison of real data (CHM1 genome) predictions for Sensitive and Quick Modes of TARDIS. 78
4.8 Evaluation of Linked-Read performance of TARDIS. 79
4.9 Performance comparison in terms of time and memory. 81

Chapter 1

Introduction

To expose the mysteries of genome, various biological studies had been conducted in the past by researchers. With computational approaches today, we are able to make progress much more rapidly than before. Human Genome Project and 1000 Genomes Project are two milestones in biological research executed in the last 20 years, although they are not the only ones. With the introduction of high-throughput sequencing, the results of genomics became more valuable as this platform opened tremendous research opportunities to researchers.

1.1 DNA and Computational Genomics

The first clues of heredity in living organisms became evident in 1865 when Gregor Mendel explained how traits passed down from parent to child. A few years later, Friedrich Miescher identified "nuclein" in white blood cells of human, which is what we know as deoxyribonucleic acid today. Although the importance of Miescher's findings was not realized for many years, Franklin, Watson and Crick's description of the double helix structure of DNA [10] opened a new era in the field of genetics.

Today, we know that DNA (deoxyribonucleic acid) is found within nearly every

cell of living organisms. It has two long sequences of nucleic acids made up of four bases; Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). These sequences are attached together by chemical bonds and are called base pairs; A pairs with T and C pairs with G on complementary strands.

DNA of an organism is packaged into structures called chromosomes inside the nuclei of cells. Human genome includes 46 chromosomes; 22 autosome pairs and two sex chromosomes that can be either XX or XY, for females and males respectively. The total length of these chromosomes is around 3 km in length and contains nearly 3.2 billion base-pairs (bp) with the addition of mitochondrial DNA (mtDNA) present inside the mitochondria that represents only a small fraction of the total DNA. These chromosomes and mtDNA contain the genes that code for proteins and human genome contains 20,000 to 25,000 of them. The combination of all the genes or the genetic makeup of an organism is what we call genome.

Human Genome Project (HGP) was launched in 1990 with the aim to sequence whole human genome and released the first results in 2001 [11]. This is a near complete human genome sequence created from the genomes of a few individuals and is called the "reference genome". The project was completed in 2004 [12], but is still being updated. HGP attempt led to various new projects in the field of genomics [13, 14, 15, 16, 17, 18, 19, 20, 21], therefore the amount of data increased tremendously. With this increase, researchers were forced to rely on computational methods and subsequently new techniques and tools emerged. This can be regarded as the second birth of computational genomics area in the intersection of genomics and computer science.

1.2 Genomic Variation: Changes in DNA Sequence

Genomic variation is defined as the genomic differences between individuals. It has been shown that 99.9% of any two copies of human genomes are identical

(approximately 1 variant per 1,000 bases) [22, 23]. This minor variation causes biological differences between individuals and is what makes each unique. On the other hand, some of these variations are the causes of genetic diseases such as psoriasis [24], Crohn's disease [25], renal disease [26], diabetes [26], AIDS susceptibility [27], neurodevelopmental diseases (e.g., epilepsy, intellectual disability, autism, and schizophrenia) [28, 26, 29, 30] and many more. Thus, studying genomic variations is crucial not only for most of the branches in molecular biology and genetics but also for medical sciences.

Genomic variations can be broadly classified into four groups based on their sizes; (1) Single Nucleotide Polymorphisms (SNPs) are the point mutations i.e., changes in one nucleotide of the genome; (2) INDELs are small insertions and deletions up to 50 bps and short tandem repeats (STR) are repeated small segments up to approximately 171 bps; (3) Structural Variations are the genomic changes that affect more than 50 bps to several megabases; (4) Chromosomal changes that affect the whole chromosome i.e., trisomy or monosomy. The frequency of these variations are inversely proportional to their sizes.

In 2008, 1000 Genomes Project was launched to sequence the genomes of at least one thousand humans to create a catalog of human genetic variations using newer sequencing technologies. They launched the initial results in 2010 [14] and 2012 [18]. With the completion of the project in 2015 [20, 21], genomes of 2,504 individuals from 26 different populations were sequenced using predominantly Illumina technology, an integrated map of 84.7 million SNPs, 3.6 million INDELs, and $> 65,000$ SVs were publicly reported. The project also revealed that a typical genome differs from the reference at around 4 million sites where more than 99.9% of these are SNPs or small INDELs and 2,100 to 2,500 of them are SVs.

There are also other genome sequencing projects [31, 32, 33, 34, 7] and among these Turkey has started a new project where the aim is to sequence and analyze 100,000 Turkish genomes in three years.

1.2.1 SNPs, INDELs and STRs

In 2001, The International SNP Map working group revealed 1.42 million SNPs in human genome [23] and in 2002, Intenational HapMap Project was initiated with the goal of determining the common patterns of DNA sequence variations in human genome by characterizing sequence variants, their frequencies and correlations among them [13]. In phase 1 of the project, 1.3 million SNPs and in phase 2, a further 2.1 million SNPs were genotyped and phased using 270 individuals from diverse populations [35, 36]. Thus, having these associations led to more accurate, faster and cheaper Genome Wide Association (GWAS) studies.

On the other hand, INDELs have not been studied as broadly as SNPs but they comprise 16% to 25% of all sequence polymorphisms in human genomes [37]. INDELs are known to cause phenotypic changes and diseases like cystic fibrosis [38] and fragile X syndrome [39]. There are some methods that discover and genotype INDELs using high throughput sequencing datasets such as SPLITREAD [40] and Scalpel [41].

Finally, Short Tandem Repeats (STRs) are repetitive DNA motifs that consist of micro, mini, beta and alpha satellites that are utilized frequently in forensics, population genetics, and genetic genealogy [42]. They are also known to play an important role in genetic diseases such as various types of neurological disorders including Huntington's disease [43]. Although detection of these events with computational approaches is very challenging, there are still some methods such as lobSTR [44] and hipSTR [45] that discover STRs.

1.2.2 Structural Variations

Structural Variations (SVs) are genomic rearrangements that affect >50 bps in the sequence of a genome including insertions, deletions, duplications, inversions,

mobile element transpositions and translocations [46, 47, 48, 49, 2, 50] (Figure 1.1). Among these, copy number variations (CNVs) are referred to as unbalanced structural variants that change the number of base pairs in the genome including insertions, deletions and duplications, whereas balanced rearrangements include inversions and translocations [51].

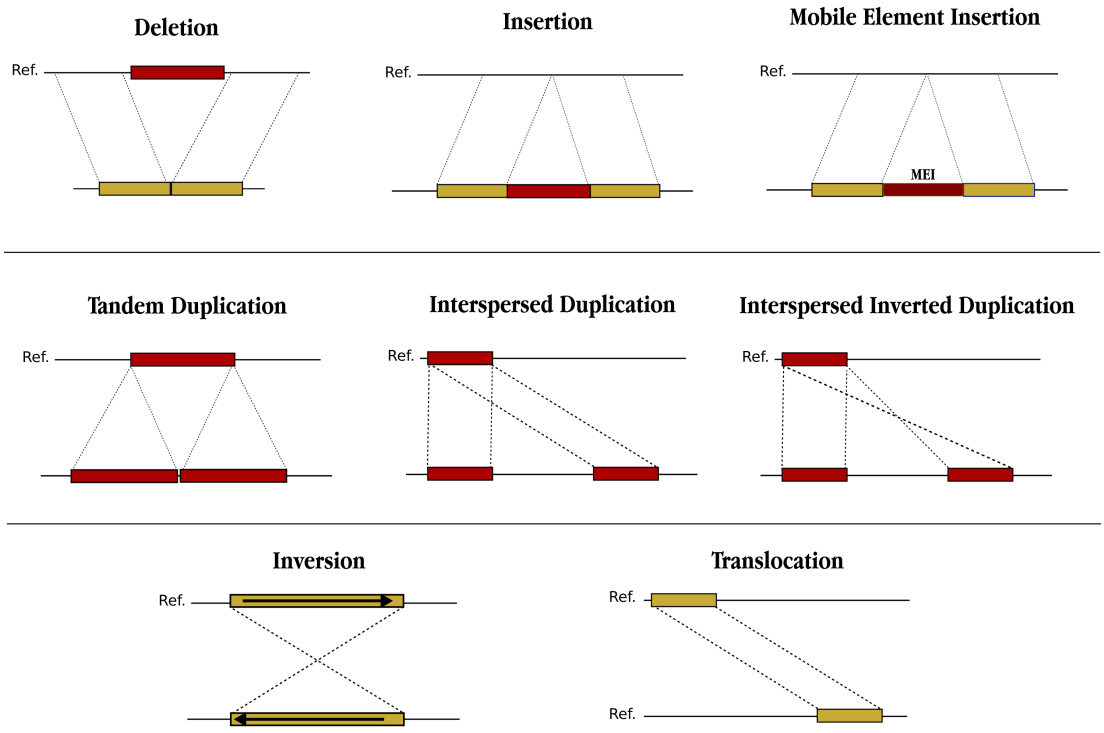


Figure 1.1: Structural Variations (SV) types of deletion, insertion, inversion, mobile element insertion (MEI), interspersed segmental duplication with direct and inverted orientations, tandem duplication and translocation are depicted.

There are various studies that associate SVs with genetic diseases ranging from neurological and neurocognitive disorders to autism, obesity, bipolar disorder, schizophrenia [52, 53, 54, 55, 56, 57] and cancer [58, 59]. Therefore the discovery and genotyping of SVs are of crucial importance in understanding their affects on human health.

The approaches to detect SVs can be broadly categorized into two groups; hybridization-based microarrays and sequencing-based computational approaches.

1.2.2.1 Hybridization-based microarrays

Microarrays have traditionally been used for multiple purposes in molecular biology; gene expression, fusion gene profiling, alternative splicing, etc. Before high throughput sequencing, they were the main instruments for SNP, INDEL and CNV detection and genotyping [46, 47, 60, 61]. There are mainly two different methods; SNP microarrays and array comparative genomic hybridization arrays (arrayCGH) and they are both based on hybridization.

In arrayCGH, reference and test DNA samples are labeled with fluorescent tags and are hybridized to target genomic arrays (long oligonucleotides, bacterial artificial chromosome (BAC) clones are used for this purpose). After hybridization, the signal ratio reveals copy-number differences between the DNA samples. [62, 51].

On the other hand, SNP microarrays are used to find CNVs and single nucleotide polymorphisms in spite of the fact that their probe designs are specific to SNPs. They were mainly used in HapMap project to find millions of SNPs. Unlike arrayCGH, hybridization is performed on a single sample per microarray in SNP microarrays. Hybridization intensities and allele frequencies are compared with average values, which indicate a change in copy number.

In general, microarrays are cheap and fast, however, they have drawbacks; poor breakpoint resolution, always specific to a reference individual, not able to detect transposon insertions, novel sequence insertions and balanced rearrangements, i.e., inversions and translocations. They are also unreliable within segmental duplications.

There are additional approaches like polymerase chain reaction (PCR) used for SNPs, quantitative real-time PCR (qRT-PCR) for CNVs and fluorescent *in situ* hybridization (FISH) used for larger events [63, 64].

1.2.2.2 Sequencing-based computational approaches

DNA sequencing is the process of determining the order of nucleotides in a DNA molecule. This is a challenging task since there is currently no machine that takes a genome as input, reads it from start to end, and outputs the entire sequence.

First attempts to sequence a genome involved breaking the DNA into many small pieces, sequencing them and assembling them again. As Figure 1.2 shows, DNA is cut into multiple fragments or inserts randomly where each fragment is sequenced using clones (Plasmids carry 3-7 Kbps, Fosmids carry ∼40 Kbps and BACs carry ∼150 - 200 Kbps). However, as entire clone cannot be sequenced, only some number of bases (∼1000) are sequenced and these are called "reads". Sequencing can be done from rightmost, leftmost or both ends. Paired-end sequencing is the process when the sequencer sequences from both left and right ends. It should be noted that the number of reads are redundant in order to reconstruct the original genome. As the redundancy increases, accuracy also increases and restoring the original genome becomes easier. This is called depth of coverage indicating the average number of reads that cover each base pair.

The history of DNA sequencing goes back to Gilbert [65] and Sanger [66] methods in 1977, where the first one is based on chemical sequencing and the latter on chain termination sequencing and both generate labeled fragments of varying lengths that are further electrophoresed. However, Sanger method gained more popularity and was used as the main sequencing tool in Human Genome Project.

Briefly, Sanger Sequencing relies on ddNTP, which is a modified nucleotide that can stop replication. As the DNA polymerase copies a DNA strand, when one of four dideoxy nucleotides (ddATP, ddGTP, ddCTP, or ddTTP), which lack a 3' hydroxyl group, became incorporated instead of a dNTP, synthesis terminates. So four different test tubes containing a template DNA strand and a primer attached to it, a DNA polymerase, dNTPs and a few ddNTPs of the same type are used. Then, by using capillary gel electrophoresis, the molecules ending with ddNTPs with various different lengths are separated by size and then fluorescent tag on

each ddNTP are read in order to determine the nucleotides.

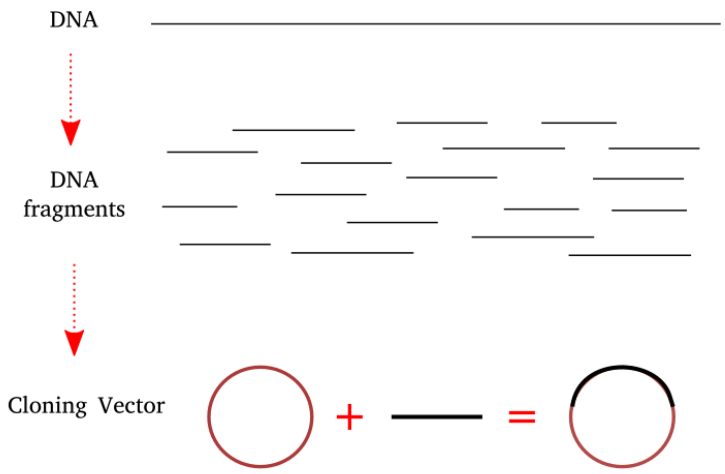


Figure 1.2: Sequencing approach employed by the first generation sequencers, i.e., Sanger Sequencing. Briefly, DNA is cut into multiple fragments randomly and each fragment is sequenced using clones.

Thus, Sanger sequencing allows long stretches of DNA fragments to be sequenced ($\sim$ 1000) with high accuracy using clone libraries, which can be used in further processing. However, this technology is very expensive and slow. Also, building and storing clone libraries is difficult and time consuming. With the introduction of next-generation sequencing (NGS), or currently called high-throughput sequencing (HTS), the field of genomics has been revolutionized. However, we should also note that the methods such as read-depth [67], read-pair [2] and split read [37] that we utilize with HTS technology today were first developed and used with Sanger sequencing technology.

1.3 High Throughput Sequencing

1.3.1 Short-read sequencing

Before 2005, Sanger sequencing, considered as the first generation sequencing platform, was the most feasible approach to sequence a genome harboring long

read length and high accuracy. However, newer sequencing platforms have emerged and changed genomics entirely as they are much faster and cheaper. Figure 1.3 shows the change of costs to sequence a genome between 2001 - 2017. The sudden decrease in 2008 displays the transition from Sanger sequencing to HTS [1]. It is noteworthy that sequencing a genome was around $\sim$ \$100 million in 2001 and is only $\sim$ \$1,200 in 2017, a decrease of five orders of magnitude.

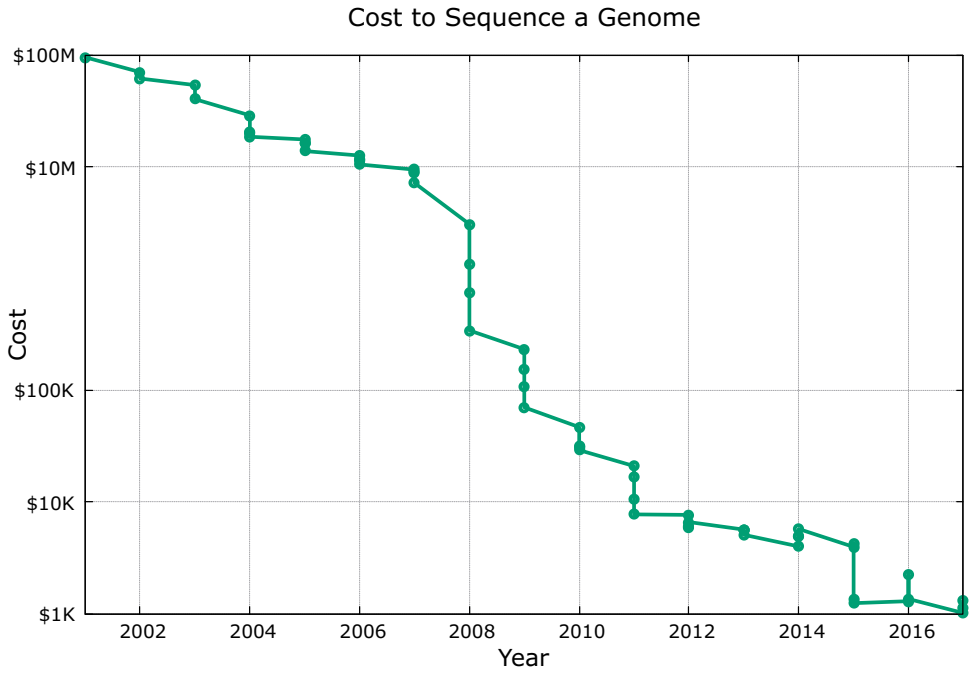


Figure 1.3: The change of cost to sequence a genome between years 2001 - 2017. Rapid decrease in 2007-2008 demonstrates the transition to HTS platforms. Data is retrieved from [1].

Main differences between the short-read sequencing and traditional Sanger sequencing is that it produces up to billions of reads in parallel, which are much shorter ($35 - 250$ bps depending on the platform). Nevertheless, it has higher error rate and possesses bias against high and low GC contents. HTS platforms commonly utilize three main steps but differ in how they handle them; (1) template preparation; (2) sequencing and imaging; (3) data analysis [68]. Indeed, the fundamental difference between the commercial HTS sequencers is in terms of sequencing technology they utilize; 454 Life Sciences is based on pyrosequencing, Illumina uses sequencing by synthesis and AB SOLiD employs sequencing by ligation.

There are two main approaches to analyze the reads of short-read sequencers: read mapping and *de novo* assembly. As both of these methods are highly complex and difficult to achieve, coupled with short nature of the reads, the problem gets even more difficult. Consequently, a combination of Sanger sequencing (longer read lengths) and HTS platform (faster and cheaper) was needed. 3rd generation sequencing (also known as long-read sequencing, single molecule sequencing or next-next generation sequencing) can be regarded as the result of this demand.

1.3.2 Long-read sequencing

Single molecule DNA sequencing was launched in 2008 with Helicos Biosciences [69], then Pacific Biosystems (PacBio) with Single Molecule Real-Time (SMRT) sequencing [70] and Oxford Nanopore Techniques (ONT) with nanopore sequencing [71]. In contrast to 2nd generation sequencing, there is no clonal amplification step in library preparation as they are able to detect single molecule in real time, i.e., the optics of these systems are very sensitive such that they can detect incorporation of one fluorescently labeled nucleotide [72].

Thus, 3rd generation sequencers have reduced PCR based errors, they have much longer read length (10-15 Kbps for PacBio and 6 Kbps for Oxford Nanopore) and they do not suffer from GC bias. Additionally, they are not slow either [73, 74, 75]. However, compared to short-read sequencers, their error rates are higher (0.1% for Illumina, $5-20\%$ for Oxford Nanopore and $10-15\%$ for PacBio) and they are much more expensive [76, 77, 78].

Currently, these sequencers are mostly utilized in *de novo* assembly algorithms such as FALCON [79], HGAP [80] and MHAP [81] and there are relatively few algorithms such as PBHoney [82], SMRT-SV [83], Sniffles [84] and HySA [85] for reference based SV detection.

1.3.3 Linked-Read sequencing

Recently, to overcome the limitations of short and long read sequencers, a new approach called Linked-Read sequencing developed by the 10x Genomics (10xG) company has been introduced. According to this approach, short reads are generated with additional long range information producing Linked-Reads of tens of Kbs originating from the same haplotype [86, 87], obtaining high sequence coverages with the cost of generating moderate coverage data.

This new technology works by first partitioning large DNA molecules (typically 10-100 Kbps) into partitions called GEMs or pools, that contain $\leq 0.3\times$ copies of the genome (2-30 large molecules) with unique barcodes, which are then sequenced using Illumina sequencer. Looking at the barcode information of each read-pair, long range information can be deduced; sequences derived from the same molecule shares the same barcode, thus they are linked [88].

There are currently a few algorithms that use Linked-Reads to identify SVs. LongRanger [89] is one of the pioneering approaches developed by 10x Genomics, which is a comprehensive package capable of doing both sequence alignment and SV detection using barcoded reads. On the other hand GROC-SVs [90] focus on cancer genomes using Linked-Reads and is also able to detect complex SV events employing local assembly. ZoomX [91], another algorithm that uses Linked-Read sequencing, is able to identify complex genomic rearrangements (>200 Kbp) in somatic and germline cells. Finally VALOR [92] characterizes large (>500 Kbp) inversions using "split molecule" signature, which is a similar approach to traditional split reads, but having the additional advantage of spanning over repetitive regions.

1.4 Reference Based Analysis

1.4.1 Read mapping

Read mapping or read alignment is the process of aligning reads onto the reference genome, only if available, in order to detect which part of the genome they likely originated from and expose genomic variations. However this problem is very challenging. First, $\sim$ 50% of the human genome is repetitive [93, 94] and it is impossible to know which copy of the repeat the read should belong to (ambiguity). Second, alignments may contain mismatches, which may be due to sequencing errors, genuine differences (SNPs, INDELs) between reference and query organisms, or both [95]. Indeed, in order to achieve higher accuracy, a confidence score (mapping quality) is needed for the alignments [96]. Finally, due to the huge number of reads, memory consumption will be very high and the speed of the algorithm will decrease. Therefore an optimal alignment (i.e., Smith-Waterman local alignment [97] algorithm) is not possible, so read mapping algorithms apply heuristics. There are two main approaches; (1) hash based seed-and-extend aligners such as mrFAST [98], BFAST [99], mrsFAST [100], SHRIMP 2 [101], FASTHash [102], NovoAlign [103]; (2) Burrows-Wheeler Transform & Ferragina-Manzini Index based aligners PatternHunter 2 [104], Bowtie [95], BWA [105], Bowtie 2 [106].

Hash based aligners initially partition the reference genome into overlapping, equal sized $k-mers$ and index them in a lookup table (i.e., hash table). When searching for the position of the reads, each read is also cut into $k-mers$ similarly and these k-mers are used as keys to look up the matching positions in the reference. Once a match is found, it is extended to align the entire read. Although this approach is sensitive, it is very costly in terms of memory and it is computationally slow. Indeed, most of the time, more than 90% [107, 108], is spent in the verification step that is based on edit distance. Algorithms such as Levenshtein's edit distance [109], Smith-Waterman [97] or Needleman-Wunsch [110] are used for this purpose [111].

On the other hand, Burrows-Wheeler Transformation (BWT) is a data compression method that is used to compress the genome, which can be used to reduce memory load. Utilizing this strategy, second type of aligners store a memory-efficient representation of the reference genome and use Ferragina-Manzini index data structure that retains the suffix array's ability for rapid subsequence search [112]. Then, each read is aligned character by character against the transformed string [113]. By this way, hits can be found very quickly in a memory efficient manner with reduced sensitivity.

There are also other approaches that apply different strategies to map long reads such as LAST [114], BLASR [115], BWA-MEM [116], DALIGNER [117], GraphMap [118], MECAT [119], LAMSA [120], Minimap 2 [121] and NGMLR [84].

1.5 *De novo* Assembly

De novo assembly involves assembling the reads to reconstruct the original genome. However, this is currently an expensive task, comprising many challenges. Shown in Figure 1.4, fragments are randomly sheared and expected to be overlapping with each other. Theoretically, entire genome can be assembled using the similarities of the overlapping parts inside the fragments and larger contiguous sequences called contigs, can be obtained [122]. The first challenge is the repetitive nature of the genome ($\sim 50\%$ of the human genome is repetitive). This becomes even more difficult with shorter read sizes; when the repeat element is larger than the read length, the algorithm cannot distinguish between the two copies. This results in gaps or missing sequence information.

Second, due to the heterozygosity of the genomes (human genome is diploid; 2 alternates of each locus), two inherited copies will have differences and both copies should ideally be constructed by the assembler. Third, because of the double helix structure of the DNA, reverse and complemented forms of the strings should be considered and sequencing errors need to be handled properly. Finally, similar

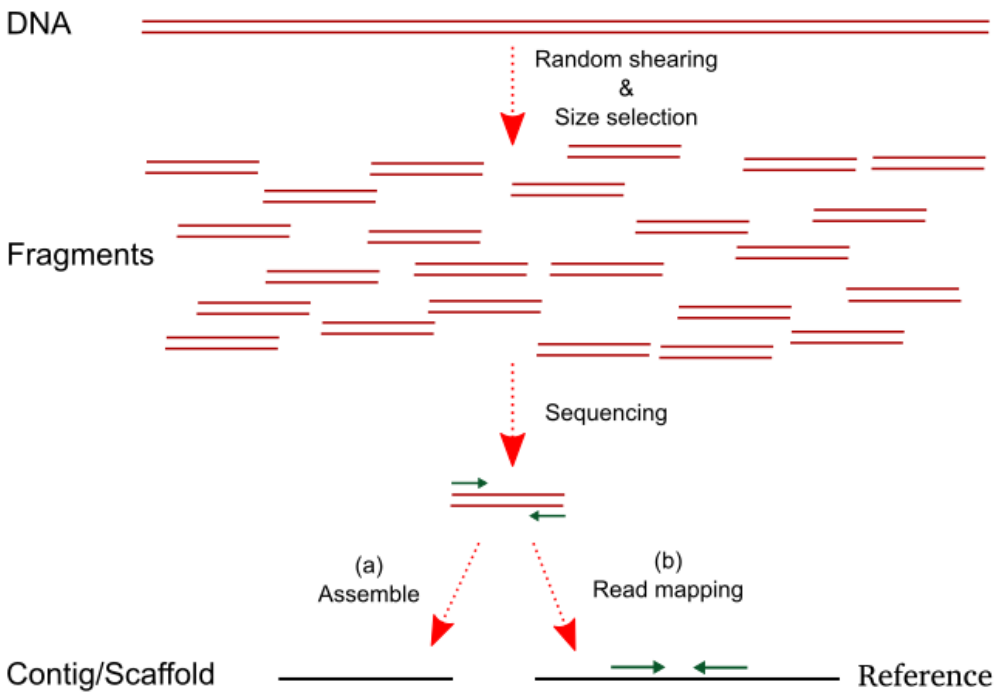


Figure 1.4: Whole Genome Shotgun (WGS) sequencing. Firstly, DNA is fragmented randomly and then each fragment is sequenced from both left and right ends using a sequencer. These are called paired-end reads. The number of base-pairs that the sequencer reads called read length depends on the sequencing machine used. Finally, the reads; (a) can be assembled into contigs when a reference genome is not available or (2) can be mapped to a reference genome

to most of the bioinformatics problems, size of the data is huge and there are billions of reads, therefore proper memory management is crucial.

First attempts to solve *de novo* assembly problem were formulized as shortest common superstring problem (SCS), which is known to be NP-Complete [123]. Given a collection of strings $S = \{s_1, s_2, ..., s_n\}$, SCS asks to find the shortest string s that contains all the substrings in S. Various heuristics or approximation algorithms for either SCS or its reductions to other NP-Complete problems such as Traveling Salesman and Hamiltonian path have been devised. Today three main approaches are being used: (1) Construction of contigs greedily (TIGR [124], PHRAP [125]); (2) Overlap-Layout-Consensus (OLC) (ARACHNE [126],

Phusion [127], PCAP [128], BOA [129]); (3) De-Bruijn Graphs (DBG) (EULER-SR [130], Velvet [131], EULER-USR [132], ABySS [133], Ray [134], SOAP-denovo [135], ALLPATHS-LG [136], [137], Meraculous [138], Cortex [139], SPAdes [140], HipMer [141]).

In the greedy approach, best matching prefix-suffix pairs are merged into longer sequences iteratively in a greedy manner. Most algorithms use overlap graphs or lists to keep the overlaps but generally these algorithms does not scale well with repeats (not appropriate for eukaryotic genomes, can be used for some bacterial or viral genomes).

OLC is composed of three main steps; in the "overlap" step, an overlap graph with prefix-suffix matches of all pairs of reads is created. The second step, called "layout", consists of building contigs by passing over each node exactly once using Hamiltonian path or Traveling Salesman Path, which is NP-Hard. Finally, in the "consensus" step, consensus sequence is determined using multiple sequence alignment of overlapping contigs created in the second step. The main drawback of OLC is the huge and slow-to-build overlap graphs, which is inappropriate for 2nd generation sequencers that have billions of reads. They were mostly used for Sanger sequencing and currently have applications to 3rd generation sequencers [142].

De Bruijn graph (DBG) assembly is considered as the most appropriate approach for short-read sequencing. Formally, given a set of reads $S = \{s_1, s_2, \ldots, s_n\}$ and an integer k, we define De Bruijn graph as $G(S, k)$, where the verticies are substrings of length k. There exists a directed edge between any two vertices u and v if and only if the last $k-1$ characters of u is equal to the first $k-1$ characters of v. Thus, the paths in the graph are the reads and a solution to the assembly problem is an Eulerian path that includes all reads as subpaths [143]. The assembler constructs the contigs using Eulerian walks in $O(|E|)$ time where E is the number of edges. However, DBG also has drawbacks; sequencing errors (gaps, etc) or uneven coverage can make the graph non-Eulerian. Even if not, genomic repeats produce many possible walks (i.e., fragmented assemblies) [144].

All these assemblers create thousands to millions of contigs depending on the data. To help improve assembly contiguity, scaffolding algorithms are used that is the process of ordering and orienting these contigs with respect to each other using various data such as paired-ends. Usually assembler have scaffolding feature but there are also standalone scaffolding algorithms such as SSPACE [145], Opera [146], SCARPA [147], BESST [148] and LINKS [149].

1.6 Structural Variation Discovery Signatures

To detect structural variations, ideally two assembled genomes are needed; a genome that we seek to detect SVs (donor) and a second genome with no variation (the reference). This way, a direct comparison of these two genomes will reveal the genetic variations trivially. However, because of the limitations of the current technology, we only have the reference genome correctly assembled (not 100% though) and chunks of the donor genome aligned to it (billions of reads). Therefore, we need to rely on signatures to detect structural variations. As Figure 1.4 shows, sequencing machines generate two reads from both ends (start and end) of a fragment and these reads are called mate-pairs or paired-end reads. The distance between these two mates, called "insert size", is the major data we have in order to utilize the sequence signatures in general.

There are four main signatures used to find SVs; (1) read-pair, (2) read-depth, (3) split read, (4) assembly; and all of these methods are based on the principle of aligning billions of reads to a reference genome and identifying potential SV events [150, 151]. However, there are complications with respect to specificity and sensitivity. The main problem lies in the repeated nature of the human genome. It comprises long segmental duplications and repeats where most of the potential SVs intervene with them. When using unique mapping of the reads, sensitivity decreases, whereas ambiguous mappings increases the false positive count. Additionally, incompleteness of the reference genome causes more problems in accurate detection since these missing portions are mostly duplications. Second, many SVs are complex with many rearrangements at the same site. In

addition, tight breakpoint resolution is often difficult to achieve with specific sequence signatures, besides, most of the smaller SVs (50 bp to 1 Kbp in length) cannot be resolved with short read sequencing [9]. Finally, short read sequencing approaches suffer from the GC bias (regions with elevated G plus C bases have higher read depth [152]).

First attempts to detect structural variations relied on similar approaches used in capillary sequencing [2, 153] and were able to detect only some basic types of SVs such as insertions, deletions, inversions and tandem duplications by using one of the sequence signatures. The following algorithms are some of the earlier approaches; (1) Read-pair signature based; VariationHunter [154], PEMer [155], BreakDancer [156], MoDIL [157], SVDetect [158], GASV [159] (2) Read depth signature based; CNVnator [160], EWT [161], mrCaNaVaR [98] (3) Split read signature based; Pindel [162], AGE [163] (4) Assembly based; NovelSeq [164], Cortex [139], SOAPdenovo [135].

On the other hand, newer approaches mostly utilize more than one sequence signatures to detect SV events and some are also able to characterize more complex variations. Table 1.1 shows state-of-the-art short-read sequencing algorithms that use Illumina platform to detect SVs by tracking multiple SV discovery signatures.

1.6.1 Read-pair

Read-pair (RP) method is the most widely employed approach and was first introduced in [153] for bacterial artificial chromosome end sequences generated from the breast cancer cell line MCF-7. Later, it was applied to discover germline genetic variation using fosmid paired-end sequencing [2]. Then, with the introduction of NGS, it was applied to short-read sequencing with 454 FLX platform [176] and then to Illumina.

The general strategy is based on aligning the paired-end reads to the reference genome and observing the distance, called "insert size", between the pairs as

Table 1.1: SV discovery algorithms that use short reads.

	Signatures	SV Types								
		DEL	INS	INV	MEI	NUMT	TRA	Duplication		
								TDUP	ISP	ISP-INV
CNVer(2010) [165]	RP, RD	✓	X	X	X	X	X	✓	X	X
inGAP-sv(2011) [166]	RP, RD	✓	✓	✓	X	X	✓	✓	X	X
DELLY(2012) [167]	RP, SR	✓	X	✓	X	X	✓	✓	X	X
GASVPro(2012) [168]	RP, RD	✓	X	✓	X	X	✓	X	X	X
PeSV-Fisher(2013) [169]	RP, RD	✓	✓	✓	X	X	✓	✓	X	X
LUMPY(2014) [170]	RP, RD, SR	✓	X	✓	X	X	✓	✓	X	X
Manta(2015) [171]	RP, SR	✓	✓	✓	X	X	✓	✓cannot distinguish		
SV-BAY(2016) [172]	RP, RD	✓	✓	✓	X	X	✓	✓	X	X
Pamir(2017) [173]	RP, SR, AS	X	✓	X	X	X	X	X	X	X
TARDIS(2017) [174]	RP, RD, SR	✓	✓	✓	✓	✓	X	✓	✓	✓
SvABA(2018) [175]	RP, SR, AS	✓	✓	✓	X	X	X	✓cannot distinguish		

State-of-the art short-read sequencing algorithms that use Illumina platform are summarized in the table. We limit the algorithms to the ones tracking more than one sequence signature as none of the algorithms that use a single signature is comprehensive and newer approaches employ multiple signatures to increase sensitivity and specificity. Thus, for each algorithm, we give the sequence signatures it utilizes (RP, RD, SP, AS denote read-pair, read-depth, split read and assembly, respectively) and whether it is able to discover SV types of deletion (DEL), insertion (INS), inversion (INV), mobile element insertion (MEI), nuclear mitochondria insertion (NUMT), translocation (TRA), tandem duplication (TDUP), interspersed segmental duplication in forward (ISP) and reverse (ISP-INV) directions.

shown in Figure 1.5. Figure 1.6 shows the distribution of insert sizes (or span sizes) for NA11930 genome. If the insert size of a paired-end read is below the expected distance ($\sim$ 170 bp in this example), there is a high possibility of an insertion event and larger insert size relative to the expected threshold ($\sim$ 500 bp) is an indication of a deletion event relative to the reference genome.

In broad terms, algorithms that use read-pair sequencing use the term "discordant" for read-pairs whose insert size is smaller/larger than the expected interval when mapped to the reference genome. Also, "concordant" read-pairs are those having expected insert size, i.e., when the distance between the aligned reads is within the expected range. Therefore the algorithms that make use of read-pair method mostly deal with the discordant read-pairs as they indicate possible SV loci.

In addition to observing the insert size, if one of the pairs has an unexpected orientation, it's likely the result of an inversion [2] and these read-pairs are also

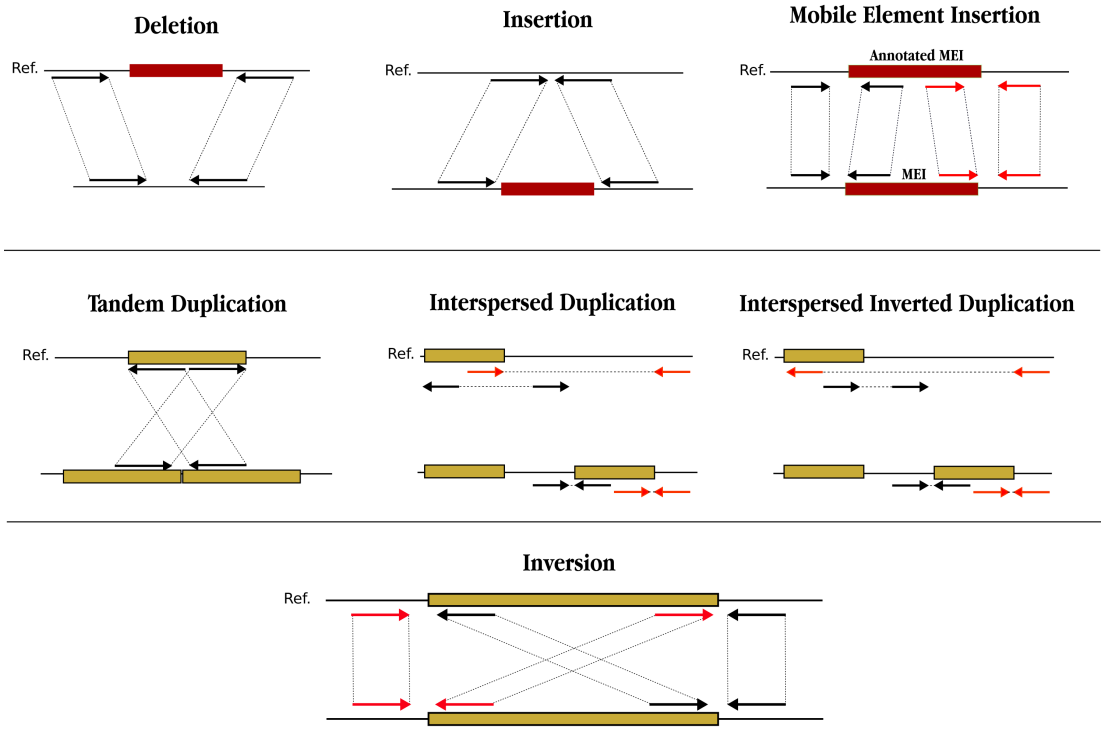


Figure 1.5: Read-pair signatures of each SV event is shown. Reads are mapped to the reference genome and the distance between the mate-pairs or the orientation of the mapped reads determine whether a potential SV is implied or not.

classified as discordant. Note that reads should typically map in forward-reverse $(+/-)$ orientation, which is considered correct mapping, i.e., the left mate is mapped to the "+" strand, and the right mate is mapped to the "−" strand. However, for inversions, one of the ends has flipped orientation, so $+/+$ or $-/-$ mappings are tracked.

In case of tandem duplications, both ends map in everted $(-/+)$ orientation. For mobile element insertions, the approach is different; if one of the mate pairs fall inside the annotated transposons, we see a potential mobile element insertion event. Finally, for interspersed segmental duplications, two types of mappings should be detected; $+/-$ and $-/+$ for direct and $+/+$, $-/-$ for inverted segmental duplications. However, these signals might not be enough for detecting interspersed segmental duplications as they are very similar to other types of SV events, so different types of signals or post-processing based on a likelihood model

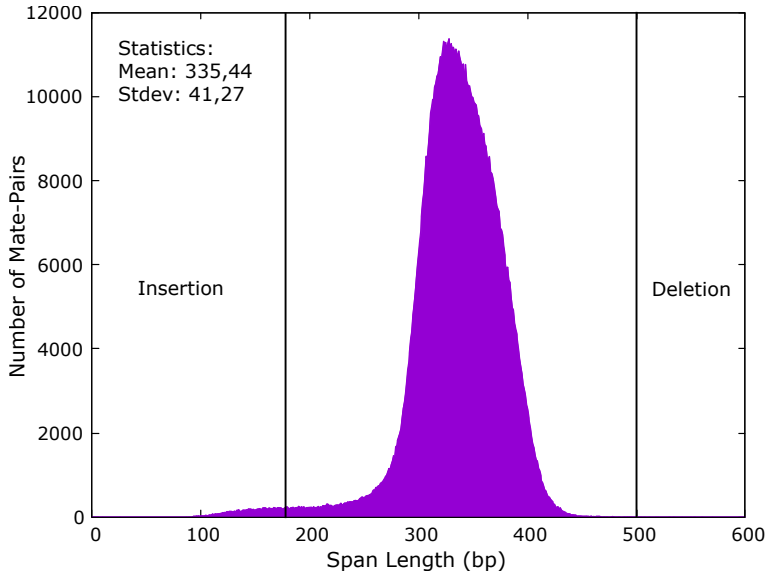


Figure 1.6: Span size distribution of read-pairs sampled from NA11930 genome. When the read-pairs are mapped to the reference genome, distance between them is a possible indication of a genomic variation. If the distance is below or above the expected cut-off values; $min = mean - (4 \times stdev) = 170$, $max = mean + (4 \times stdev) = 500$ in this case; then these read-pairs are called "discordant" and they are the ones considered as potential SVs [2].

might be needed.

1.6.2 Read-depth

Read-depth (RD) signature was first applied to old capillary sequencing data in order to identify duplications in human genome by [67]. It was later applied to HTS data to define rearrangements in cancer [177, 178] and then segmental duplication and absolute copy-number maps in human genomes [98, 179]. The general idea is simple and depends on the assumption that number of reads mapping to any region follows a Poisson distribution. Thus, by analyzing the divergence from this distribution reveals deletions or duplications in the sample. As displayed in Figure 1.7, deleted regions will have lower read-depth, whereas duplicated regions will have higher read-depth compared to regions of diploid nature.

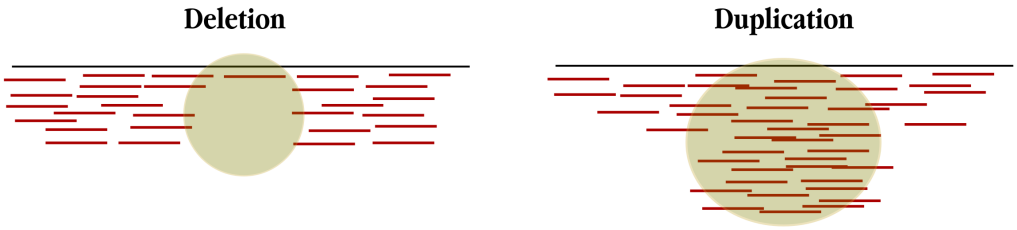


Figure 1.7: Read-depth signatures of SV events are shown. When reads are mapped a region, divergence from the distribution implies a deletion (decrease in read-depth) or a duplication (increase in read-depth) event.

The accuracy of this method is highly correlated to the coverage of the dataset; with lower coverages, accurate SV detection will not be possible. Compared to read-pair approach, detection of larger SVs are possible with this approach, although smaller events that RP is able to detect with lower coverages might not be detected by RD signals [150]. As expected, breakpoint resolution of the method is poor. Finally, RD approach is highly affected by the sequencing bias of HTS technology, as some regions are over or under sampled mostly due to the GC bias.

1.6.3 Split read

The first application of split read (SR) signature was based on Sanger sequencing [37], and Pindel [162] was the first to apply it to HTS data. According to this method; when a read maps to a reference genome, there are some chunks of the read, mostly at the beginning or at the end, called soft clips, that cannot be mapped correctly. Split read approach involves remapping these chunks into the reference with gaps indicating a possible SV event (Figure 1.8).

This approach depends on read-length and is more successful when the reads are longer because they will be more likely to span SV breakpoints. Additionally, remapping of each split will be more accurate when read length increases as the difficulty of the alignment decreases.

1.6.4 Assembly

As described in Section 1.5, by assembling reads into DNA fragments (contigs), one can detect any type of genomic variations by comparing them directly to the reference genome. However, assembly (AS) is a hard problem such that the human genome is highly repetitive and consists of too many duplications [180, 181], which results in low-quality contigs. Thus, these drawbacks are prohibitive for using assembly to detect structural variations, although there are some assemblers that characterize SVs using *de novo* assembly, local or reference guided.

1.7 Contributions

In this dissertation, we present TARDIS (Toolkit for automated and rapid discovery of structural variants) that uses high-throughput sequencing technology to detect various types of genomic variations. TARDIS harbors novel algorithms to accurately characterize both simple and complex SV types including deletions, novel sequence insertions, inversions, transposon insertions, nuclear mitochondria insertions, tandem duplications and interspersed segmental duplications in direct or inverted orientations using short read whole genome sequencing datasets. Our algorithms make use of multiple sequence signatures including read-pair, read-depth and split read to find near-exact loci of each variation while resolving ambiguities among various putative SVs: 1) at the same locations signaled by different sequence signatures, and 2) in different locations signaled by the same mapping information. Additionally, TARDIS is able to analyze more than one possible mapping location of each read to overcome the problems associated with repeated nature of genomes. Finally, we extended TARDIS to be able to utilize Linked-Read information of 10x Genomics to overcome some of the constraints of short-read sequencing technology. TARDIS is the first method to predict insertion locations of complex SV events including tandem, direct and inverted interspersed segmental duplications. Using simulated and real data sets, we showed that it outperforms state-of-the-art methods in terms of specificity and demonstrates

comparable sensitivity for all types of SVs, and achieves considerably high true discovery rate for segmental duplications.

Before giving the details of our approach in the following chapters, we briefly describe two other state-of-the-art SV discovery tools, namely DELLY [167] and LUMPY [170] that we used to compare our results with. DELLY uses read-pair and split read signatures to characterize deletions, inversions, tandem duplications and translocations. It utilizes an undirected graph based paired-end clustering, where each node in the graph denotes a paired-end read. The edges between the nodes indicate that both paired-end reads support the same SV and the edge weights denote the difference between the predicted SV sizes of the mapping locations. It assumes that the graph contains one fully connected component for each SV and it could thus be identified by computing the connected components of the graph. However, this is an ideal condition, which is not possible in most cases, thus they identify maximum cliques heuristically in the components. In order to fine map the genomic rearrangements at single-nucleotide resolution, they utilize this information as input to their split read analysis.

LUMPY, on the other hand, has a module based probabilistic framework, which is able to discover deletions, inversion, tandem duplications and translocations. It harbors read-pair, split read and a generic module to convert SV signals from each alignment to probability distributions assuming that they reflect the uncertainty in the reference genome that can be a potential end of a breakpoint, i.e., the split read module maps the output of a split read sequence alignment algorithm. The algorithm simply maps the evidences from the alignment signals to breakpoint intervals and the overlapping ones are put into the same cluster by integrating their probabilities. Finally, breakpoint regions with sufficient evidence is returned as SV predictions (If the alignments are not given as input, they utilize SAMBLASTER [182] to detect discordant reads and split reads using the input BAM file). Sites of known variants can also be provided to LUMPY as prior knowledge in order to improve sensitivity.

In this chapter, we briefly surveyed the field of computational genomics by

describing genomic variations, technologies of the past and the present, and approaches used to detect genomic variations. Finally, we outlined our contributions shortly. In Chapter 2, we give an overview of our approach to the problem of structural variation discovery with TARDIS. In Chapter 3, details of structural variations and our approach to characterize them are given in detail. We present our results in Chapter 4 and finally conclude the dissertation by exploring potential future research directions throughout Chapter 5.

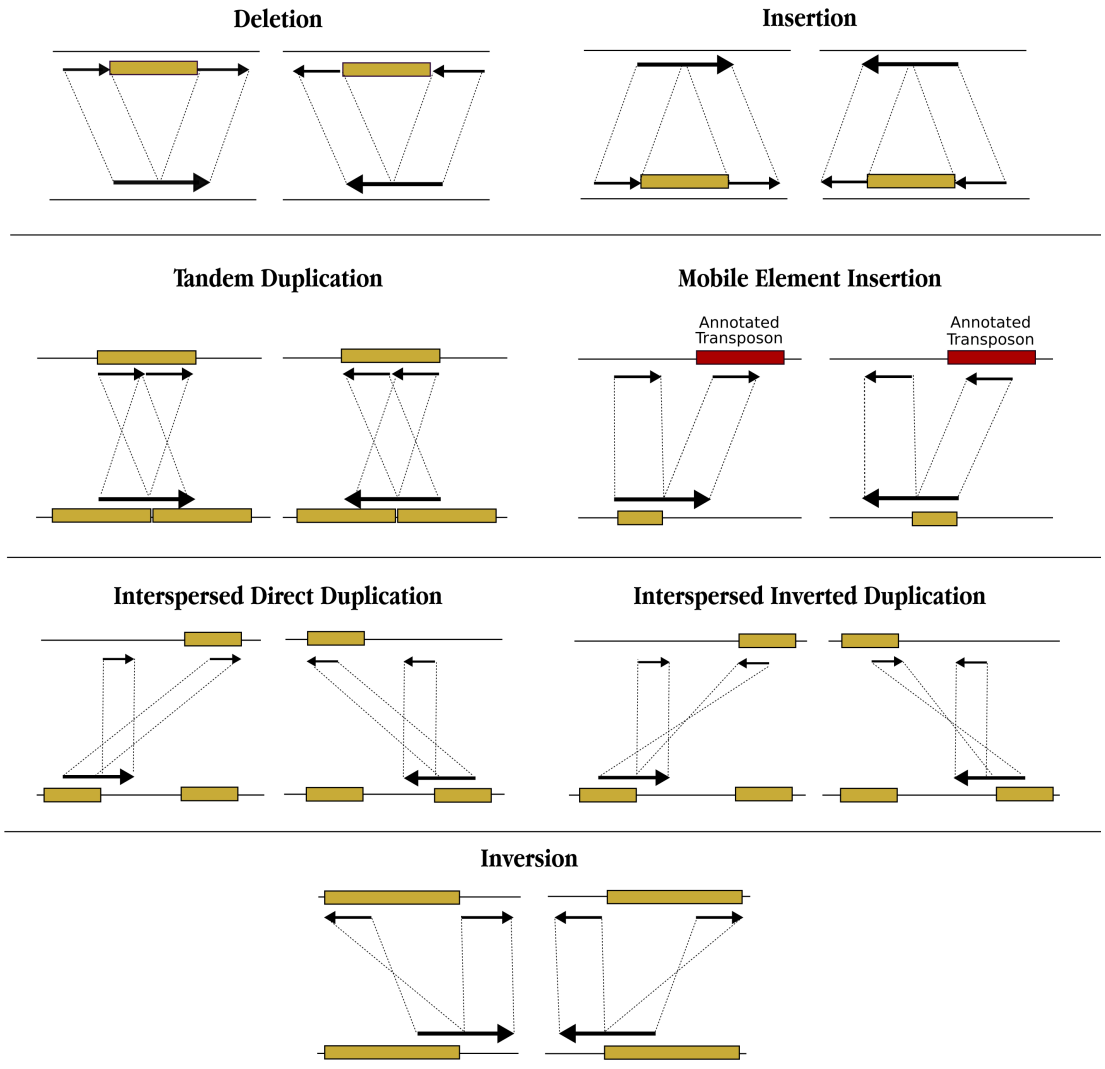


Figure 1.8: Split read signatures of SV events are displayed. Here, unmapped reads are split and each fragment is remapped to the reference genome to observe potential SVs.

Chapter 2

Overview of TARDIS: Toolkit for Automated and Rapid Discovery of Structural Variants

2.1 Introduction

2.1.1 Motivation

Genomic variations are known to be the prominent source of genetic diseases. These variations range from single nucleotide polymorphisms (SNPs) that affect a single nucleotide as substitution, to small insertions/deletions (INDELs) up to 50 bp, structural variations that affect more than 50 bp and larger chromosomal alterations that alter the whole chromosome.

Beginning with the introduction of high throughput sequencing platforms, and later 1000 Human Genome Project (1000GP), several researchers focused on characterizing SVs in human genome. Thus a plenty of algorithms have been developed. Compared to the previous approaches to detect SVs that involved using

BAC arrays, oligonucleotide array comparative genomic hybridization, SNP microarrays and finally Sanger sequencing, speed and cost of HTS platforms are unimaginably low and they are able to detect plenty of SVs. As a matter of fact, with the completion of 1000GP, over $65,000$ SVs [20, 21] were reported, which was made possible with the algorithms that utilize HTS platforms.

However, these algorithms have considerable drawbacks as well; first, although they possess acceptable sensitivity, they have many false predictions. Second, they are able to discover only simple types of SVs such as insertions, deletions and short inversions as they cannot accurately characterize complex SV events. Indeed, this is one of the reasons for lower specificity they possess.

Earlier algorithms rely on using only one of the possible sequence signatures (read pair, read depth, split read or assembly). This results in fewer detectable SV types since each signature is capable of detecting only some specific SV classes. Additionally, using a single sequence signature reduces the precision of the algorithms as they have no way of verifying detected SVs by taking advantage of multiple evidence. On the other hand, newer approaches combine more than one sequence signature, however the second signature is mostly used as a post-processing approach for verification. Thus, since they aim to characterize only a few types of SVs, they do not try to resolve conflicting SVs within the same loci.

Another weakness of these algorithms is that most of them consider only high confidence alignments and dismiss reads with lower quality alignments or reads with multiple possible mapping locations. Considering the repetitive nature of human genome, this naturally decreases the sensitivity of the algorithms within the repeated segments.

2.1.2 Our approach

Here we introduce TARDIS, a toolkit for automated and rapid discovery of SVs using both whole genome sequencing data (WGS) generated by Illumina platform and Linked-Read sequencing of 10x Genomics (10xG) [87]. However it can easily

be extended to support long read sequencing such as PacBio or Oxford Nanopore.

The general framework to using paired-end reads to detect SVs between a reference and a donor genome was first formulated by [153] and [2]. It was based on a simple idea with two steps; (1) Mapping the paired-end reads to the reference genome; (2) Observing the orientations and distance of the discordant mappings.

When the read pairs are mapped to the reference genome, there are two signs that we track; span size and orientation of the pairs. First, the distance between the pairs, called insert size or span size, is expected to be in some range $[\delta_{min}, \delta_{max}]$. Read pairs within this range are categorized as concordant and the ones that are below or above are called discordant. Indeed, discordant read-pairs suggest potential SV regions where larger or smaller insert sizes indicate deletions or insertions respectively. Second, orientation of the read-pairs are expected to be correct, i.e., $+/-$ is the correct orientation for Illumina platform. Accordingly, the read-pairs with unexpected orientation are also designated as potential SVs such that $-/-$ or $+/+$ are inversions and $-/+$ are tandem duplications.

Similar to most modern SV callers, TARDIS integrates multiple sequence signatures including read-pair, read-depth and split-read to accurately characterize both simple variants such as novel insertions, deletions, inversions, mobile element insertions, nuclear mitochondria insertions; and complex variants such as tandem duplications, direct and inverted interspersed segmental duplications (SDs) accurately. Besides, it resolves ambiguities such as; (1) Two or more SVs reported to be at the same location but signaled by different signatures; (2) Two or more SVs in different locations signaled by the same mapping information. However, current SV callers are incapable of characterizing several forms of complex SVs such as tandem or interspersed segmental duplications (SDs) [183, 184] with the exception of read-depth based methods that can only identify the existence of SDs [98, 179], but cannot detect breakpoints. TARDIS, on the other hand, is the first method to characterize insertion locations of segmental duplications using HTS data.

In this chapter we describe our approach to the problem of structural variation discovery using HTS and leave the details of SV discovery to Chapter 4. The next section elucidates how we handle the mapping information (all possible mappings of the reads to the reference vs. unique mappings) given as input to TARDIS. Then the details of the Maximum Parsimony Structural Variation approach (MPSV), adapted from [154, 185, 186] are given. Finally, we describe our approach using Linked-Read data.

Before proceeding, we give a brief description of the MPSV approach; It is a two-step process, where the first involves creating clusters with paired-end reads that suggest the same potential SV. Then in the second step, final alignments of each read-pair are determined by eliminating read-pairs from the clusters step-by-step. This is similar to the classical Set-Cover problem that is known to be NP-Complete, however [154] provides a greedy algorithm with an approximation factor of $O(logn)$ using only the read-pair signature. TARDIS builds upon this approach, using the same objective function, but it also includes methodological and heuristic novelties; (1) incorporates split-read signature and adds novel paired-end reads to the clusters; (2) uses a probabilistic model that makes use of read-depth signature to assign a likelihood score to each potential SV.

We should also note that TARDIS is implemented in C using HTSlib and it is suitable for cloud use. Source code is also freely available at `https://github.com/BilkentCompGen/tardis`.

2.2 Read Mapping

In order to discover genomic alternations in a genome, the initial step constitutes mapping the reads belonging to a donor genome to the reference. Therefore, one can observe the variations from the reference genome. This preliminary step is achieved by using a read mapping algorithm. There are currently a vast amount of read mappers, which can be categorized based on various aspect such as the data structures they utilize, their output types, etc. Indeed, for SV discovery

algorithms, there is an important distinction between these read mappers. Some of them report only the best or an arbitrary mapping of the reads in case of a tied map locations such as MAQ [96], BWA [116], Bowtie [95] and some report all possible map locations such as mrFAST [98] and mrsFAST [100].

As we already know that most SVs lay within the repeated regions, considering all possible mapping locations of the reads is of crucial importance. There are currently some algorithms that utilize this information [187, 154, 188, 168]. Therefore the number of hits they have increase, although their specificity decreases proportionally as some of these map locations might be incorrect. In addition to this, running time of these type of algorithms inherently increase, as searching for multiple map locations is a costly operation in terms of memory consumption.

In TARDIS, we possess two distinct modes of operation where the "Quick Mode" works only with the best mappings of each read-pair and "Sensitive Mode" deals with all the potential mappings.

2.2.1 Quick Mode

In order to interpret the mappings produced by a read mapper, TARDIS requires the mapper to utilize the well-known and most-widely used Sequence Alignment/MAP (SAM) format and it needs the compressed and indexed BAM version [189] as input. The BAM file is composed of an optional header, that contains information about the reference that the individual has been mapped to, and an alignment section consisting of the actual alignments. Quick Mode of TARDIS utilizes the mapping information produced by BWA [116] or a similar aligner as input. It works by reading each mapping sequentially and deciding whether the mapping is a potential SV or not. Additionally read count and read-depth information is retained for read-depth analysis and yet, soft-clipped reads are manipulated for split reads.

We should note that BWA also reports a few alternative mapping locations

of some of the reads in XA tag of BAM file. In TARDIS, we have an option to analyze these locations. However, this mode naturally increases the false positive count in parallel to true positives.

2.2.2 Sensitive Mode

TARDIS is able to resolve all potential mappings of the dataset, so the second mode employed by TARDIS is called Sensitive Mode. This mode makes use of mrFAST [98] (or mrsFAST [100], which supports multi-threaded run) for read mapping purpose. As TARDIS runs in Quick Mode by default, invoking "–sensitive" parameter is enough to enable the Sensitive Mode. We have built in functions to run mrFAST within TARDIS, or it can be run seperetaly. mr/mrsFAST typically reports all the potential mapping locations of discordant paired-end reads in a seperate file. Table 2.1 describes all the field available in the DIVET file produced by mr/mrsFAST and how TARDIS handles them.

Table 2.1: Mandatory fields of mrFAST output and how TARDIS handles them.

Field	**Description**
readN	Distinct name of the read
chroN	Chromosome name where the read is mapped
$ML_{1,L}$	Left-side mapping location of the first pair
$ML_{1,R}$	Right-side mapping location of the first pair
OR_1	Orientation of the first pair
chroN2	Chromosome name where the read's pair is mapped
$ML_{2,L}$	Left-side mapping location of the second pair
$ML_{2,R}$	Right-side mapping location of the second pair
OR_2	Orientation of the second pair
SVtype	Type of variation the mapping of paired-end suggests INS, DEL, INV
ED	Total edit distance of this paired-end (left-end plus right-end)
AvgPhred	Average Phred Score of this mapping used to prune some mappings
ProbBasedPhred	Prob. of the mapping solely based on Phred Score and edit distance

Although reporting all possible mapping locations of paired-end reads is time-consuming [98], it is an offline process that needs to be done once per BAM file. It should be noted that in order to resolve the read depth and read count of the dataset, Sensitive Mode still requires the BAM file along with the DIVET.

2.3 SV Discovery via Maximum Parsimony

Briefly, TARDIS aims to minimize the total number of SVs gathered from all the discordant read-pairs and split-reads. To formulate the problem formally, let the discordant read-pairs/split reads be represented as;

$$D = \{rp_1, rp_2, \ldots, rp_n\}, \tag{2.1}$$

where rp_i corresponds to a discordant read-pair or split-read. Each of these rp_i have more than one mappings in the reference genome when multiple mapping is enabled (i.e., Sensitive Mode) and these are represented with

$$rp_i = \{a_1 rp_i, a_2 rp_i, \ldots, a_n rp_i\} \tag{2.2}$$

Each of these mappings $a_n rp_i$ is a 5-tuple that keeps map locations and orientation;

$$a_n rp_i = (rp_i, (L_l(a_n rp_i), L_r(a_n rp_i)), (R_l(a_n rp_i), R_r(a_n rp_i)), or(a_n rp_i)), \tag{2.3}$$

where $(L_l(a_n rp_i), L_r(a_n rp_i))$ and $(R_l(a_n rp_i), R_r(a_n rp_i))$ are the start and end locations of the left, right pairs respectively and $or(a_n rp_i) \in \{+/-, +/+, -/-, -/+\}$ corresponds to the orientation of the alignment as formulated in [154]. As noted before, $+/-$ mappings designate no inversion, $+/+$, has right pair inverted, $-/-$ has left pair inverted and $-/+$ has both pairs inverted.

2.3.1 Building clusters

We cluster the alignments, $a_n rp_i$, that support the same particular potential SV in terms of loci and type as:

$$VCLUS_i = \{a_{i'_1} rp_{i_1}, a_{i'_2} rp_{i_2}, \ldots, a_{i'_l} rp_{i_l}\} \tag{2.4}$$

These are also called "valid clusters", where all the mappings in $VCLUS_i$ support the same particular SV and each discordant mapping satisfies a set of rules based on the type of SV it supports such as insertion, deletion, inversion, MEI, NUMT, tandem duplication, interspersed segmental duplication, inverted interspersed segmental duplication. The details of these rules are given in Chapter 3.

Similar to the valid cluster formulation, a "maximal valid cluster", is a valid cluster where no valid superset exists [190, 154, 185, 164, 186]. This can be computed in polynomial time as formulated initially by [154]:

1. Let $MPOS = \{MPOS_1, MPOS_2, \ldots, MPOS_n\}$, a collection of maximal intersecting intervals where the interval $(I_{i,j} = [l(a_n rp_i), r(a_n rp_i)])$ corresponds to each read-pair/split-read alignment $a_n rp_i$. Then the maximal intersecting interval is computed as follows; after sorting the intervals, scan them from left to right, adding each interval, that intersects with all the previously added intervals, to $MPOS_1$. Proceeding with $MPos_2$, include the members of $MPos_1$, except the one that has the leftmost right end and iterate. At each step i, eliminate $MPos_i$ if it ends up to be a proper subset of $MPos_{i-1}$.
2. For interspersed duplications, we employ a further step to join mappings in both $+/+$ and $-/-$ ($+/-$ and $-/+$ for inverted duplications) orientations inside the same sets.
3. For each maximal overlapping set $MPOS_i$ found in Step 1, create all the overlapping maximal subsets $MPOS_{i_1}, MPOS_{i_2}$, etc. that the insertion

size they suggest is overlapping (Necessary for detecting inversions and interspersed duplications only).

4. Among all the discovered sets, remove any set that is a proper subset of another chosen set.

This way, final clusters have been formed, which are the main data source for the set-cover phase.

2.3.2 Set-Cover approximation to find putative SVs

The second phase of TARDIS utilizes set-cover approximation to find putative SVs. It uses a similar objective function with the set-cover combinatorial problem that is known to be NP-Complete. Given a universe $U = \{u_1, u_2, \ldots, u_n\}$ and a collection of subsets of U as $S = \{s_1, s_2, \ldots, s_n\}$, the set-cover problem asks to find the smallest subset of S whose union covers all the elements ($u_i \in U$) of U. Similarly in TARDIS, we have clusters given as input that are made up of read-pairs that signal the same particular SV and our aim is to cover all the reads in each cluster. More formally, we want to compute a unique mapping for each discordant read-pairs such that the total number of SVs is minimized, or similarly the average number of read-pairs that support an SV is maximized, such that each paired-end read maps to a location in at most one selected valid cluster [154].

Since the set-cover problem is NP-Complete, an $O(logn)$ approximation to the optimal solution is suggested such that in each iteration, a cluster consisting of maximum number of uncovered read-pairs is selected until all the read-pairs are covered [154, 191].

However, occasionally a read-pair may signal two different SVs, i.e., same read-pair in different clusters, because read-pair signatures of some SV events are exactly the same as depicted in Figure 2.1. On the other hand, deletions and duplications should be verified using a read depth signature as for deletions, a

decrease in read depth and for duplications, an increase in read depth similar to Figure 1.7 should be observed.

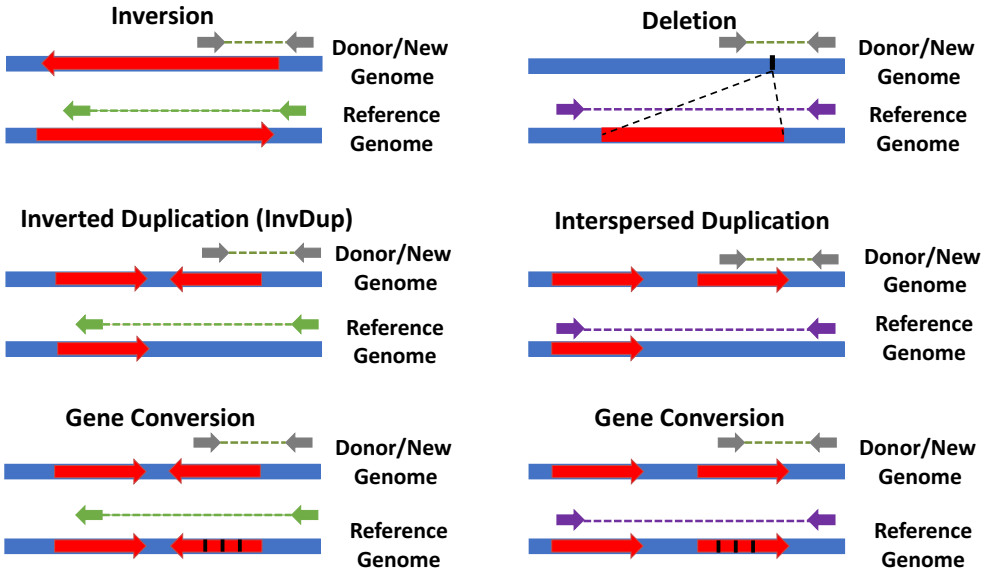


Figure 2.1: Read-pair sequence signatures of some SV events are the same such as inversions, interspersed inverted duplications and gene conversions. Similarly deletions, interspersed direct duplications and gene conversions also show the same signature.

To overcome these issues, we developed a probabilistic model that makes use of read-depth signature to assign a likelihood score to each putative SV.

2.3.2.1 Read Depth based likelihood model to improve TARDIS calls

To score each potential SV, we use a probabilistic model based on read depth signature. These scores, called CNV scores, are utilized in cluster selection. So, in the set-cover phase of TARDIS, we utilize an iterative approach; (1) at each step, we select the cluster with the best CNV score since their likelihood of being a true SV event is higher, (2) we assign the relative discordant paired-end reads to the selected SV and remove them from all the other clusters. This iterates until all the paired-end reads are covered.

Before proceeding to our probabilistic model, we should define the information that TARDIS keeps track of for each maximal valid cluster S_i for $1 \leq i \leq n$ in the set $SV = \{S_1, S_2, \ldots, S_n\}$:

- observed read depth and read pair (d_i, p_i), where d_i is the total observed read depth, and p_i is the number of discordantly mapped read pairs.
- potential duplicated, deleted or inverted regions (α_i, β_i).
- potential breakpoints γ_i.
- potential SV type.

Assuming observed read depth and number of discordant read pairs follow a Poisson distribution [192], $\lambda > 0$,

$$\text{Poisson}(\lambda, x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{2.5}$$

where, λ is the expected read depth or the number of read pairs, and x is the observed read depth or the number of read pairs respectively. However, we assume that the expected read depth or read pairs for some SV events might be zero, so we approximate this probability by,

$$\text{Poisson}(0, x) = \varepsilon^x \tag{2.6}$$

for a small $\varepsilon > 0$ (e.g. $\varepsilon = 0.01$ for read pairs and $\varepsilon = 0.001$ for read depth).

For each cluster S_i, we define a random variable $state_i \in \{0, 1, 2\}$ where the state of S_i is *homozygous* if $state_i = 2$, *heterozygous* if $state_i = 1$, and *no event* if $state_i = 0$. We also define a random variable $type_i$, which represents the SV type for S_i.

Given $state_i = k$ and $type_i = \delta$, the likelihood of S_i can be calculated as:

$$\begin{aligned} L_i(\delta, k) &= P(S_i \mid \delta, k) \\ &= P(\textit{ read depth of } S_i \mid \delta, k) \cdot P(\textit{ read pairs of } S_i \mid \delta, k) \\ &= \text{Poisson}(d_i, \lambda_d) \cdot \text{Poisson}(p_i, \lambda_p) \\ &= \frac{\lambda_d^{d_i} e^{-\lambda_d}}{d_i!} \cdot \frac{\lambda_p^{p_i} e^{-\lambda_p}}{p_i!} \end{aligned} \tag{2.7}$$

where λ_d is the expected read depth of S_i given $type_i = \delta, state_i = k$ and λ_p is the expected read pairs of S_i given $type_i = \delta, state_i = k$.

λ_d is calculated based on $(type_i, state_i)$ and the expected read depth within the region (α_i, β_i) normalized with respect to its GC% content using a sliding window of size 100 bp, denoted by $E_d[(\alpha_i, \beta_i)]$. Note that read depth normalization is a necessity for short read sequencers like Illumina, since they have bias within the regions with elevated GC content (proportion of bases G + C) as these regions have higher read depth [152]. We calculate λ_p based on the $(type_i, state_i)$ and the expected number of discordantly mapped read pairs around the potential breakpoint γ_i, denoted by $E_p[\gamma_i]$.

For instance, if an SV event is categorized as homozygous deletion, read depth inside the potential deleted region (α_i, β_i) is expected to be ~ 0 and the expected number of discordantly mapped read-pairs is approximately close to the expected number of reads containing the potential breakpoint, i.e $E_p[\gamma_j]$. On the other hand, for heterozygous deletions, read depth and the number of discordantly mapped read pairs is assumed to be half the expected value. For the values when

there is no event at the potential SV region, read-depth is supposed to be equal to expected and number of discordantly mapped read-pairs close to zero.

Similar approach is applied for the values λ_d, λ_p of inversions, insertions, MEIs and duplications. Table 2.2 displays the values of λ_d, λ_p for each $(type_i, state_i)$ using $E_d[(\alpha_i, \beta_i)]$ and $E_p[\gamma_i]$. Note that even though the formulations for λ_d, λ_p look the same for all types of duplications, the likelihood score will be different since the potential regions (α_i, β_i) are different based on the categorized type of the event being considered. Furthermore, the read-pair support and signature will be distinct for each type of duplications, which is the key in resolving the particular duplication type.

Table 2.2: Formulation for λ_d and λ_p for maximum valid cluster S_i

SV Type	State	λ_d	λ_p
Deletion	*homozygous*	0.0	$E_p[\gamma_i]$
	heterozygous	$0.5 \cdot E_d[(\alpha_i, \beta_i)]$	$0.5 \cdot E_p[\gamma_i]$
	no event	$E_d[(\alpha_i, \beta_i)]$	0.0
Inversion	*homozygous*	$E_d[(\alpha_i, \beta_i)]$	$E_p[\gamma_i]$
	heterozygous	$E_d[(\alpha_i, \beta_i)]$	$0.5 \cdot E_p[\gamma_i]$
	no event	$E_d[(\alpha_i, \beta_i)]$	0.0
Insertion	*homozygous*	$E_d[(\alpha_i, \beta_i)]$	$E_p[\gamma_i]$
	heterozygous	$E_d[(\alpha_i, \beta_i)]$	$0.5 \cdot E_p[\gamma_i]$
	no event	$E_d[(\alpha_i, \beta_i)]$	0.0
Transposon Insertion (MEI)	*homozygous*	$E_d[(\alpha_i, \beta_i)]$	$E_p[\gamma_i]$
	heterozygous	$E_d[(\alpha_i, \beta_i)]$	$0.5 \cdot E_p[\gamma_i]$
	no event	$E_d[(\alpha_i, \beta_i)]$	0.0
Nuclear Mitochondria Insertion (NUMT)	*homozygous*	$E_d[(\alpha_i, \beta_i)]$	$E_p[\gamma_i]$
	heterozygous	$E_d[(\alpha_i, \beta_i)]$	$0.5 \cdot E_p[\gamma_i]$
	no event	$E_d[(\alpha_i, \beta_i)]$	0.0
Inverted Duplication	*homozygous*	$2 \cdot E_d[(\alpha_i, \beta_i)]$	$E_p[\gamma_i]$
	heterozygous	$1.5 \cdot E_d[(\alpha_i, \beta_i)]$	$0.5 \cdot E_p[\gamma_i]$
	no event	$E_d[(\alpha_i, \beta_i)]$	0.0
Direct Duplication	*homozygous*	$2 \cdot E_d[(\alpha_i, \beta_i)]$	$E_p[\gamma_i]$
	heterozygous	$1.5 \cdot E_d[(\alpha_i, \beta_i)]$	$0.5 \cdot E_p[\gamma_i]$
	no event	$E_d[(\alpha_i, \beta_i)]$	0.0
Tandem Duplication	*homozygous*	$2 \cdot E_d[(\alpha_i, \beta_i)]$	$E_p[\gamma_i]$
	heterozygous	$1.5 \cdot E_d[(\alpha_i, \beta_i)]$	$0.5 \cdot E_p[\gamma_i]$
	no event	$E_d[(\alpha_i, \beta_i)]$	0.0

2.3.2.2 SV weights

For each potential SV, we calculate a score to represent how likely the SV prediction is correct given the observed signature considering the homozygous state and heterozygous state (i.e., 1/1 or 0/1 respectively) separately. Then we select the larger value to approximate the likelihood of that prediction being correct.

We define the score as log likelihood ratio of the putative SV being true given the observed data over it being false. Note that we use log function to avoid numerical errors. The score of potential SV S_i is defined as follows:

$$score(S_i) = \frac{\max\left(\log L_i(\delta_i, k=1), \log L_i(\delta_i, k=2)\right)}{\log L_i(\delta_i, k=0)} \tag{2.8}$$

where δ_i is the potential SV type of S_i. Again, $k = 0, 1, 2$ implies that the state of S_i is *no event*, *heterozygous*, *homozygous* respectively.

Finally, the normalized weight of each cluster can be calculated as:

$$weight(S_i) = \frac{score(S_i)}{E_p[\gamma_i]} \tag{2.9}$$

2.4 Using Linked-Read Information

Currently, the most widespread approach to overcome the limitations of short-read sequencing, with a cost-effective alternative to single molecule long-read sequencing, is called Linked-Read sequencing or 10xG. This approach involves microfluid partitioning of large DNA molecules (typically 10-100 Kbp) into roughly 10^6 pools (or GEMs) that contain fragments of the large DNA molecules (each GEM will receive 1/6000 of the genome, which is ∼500 Kb), where the pools have unique barcode information and the fragments inside the pools have the same barcodes. These reads are then sequenced using Illumina platform, retaining long

range information with short-reads as illustrated in Figure 2.2.

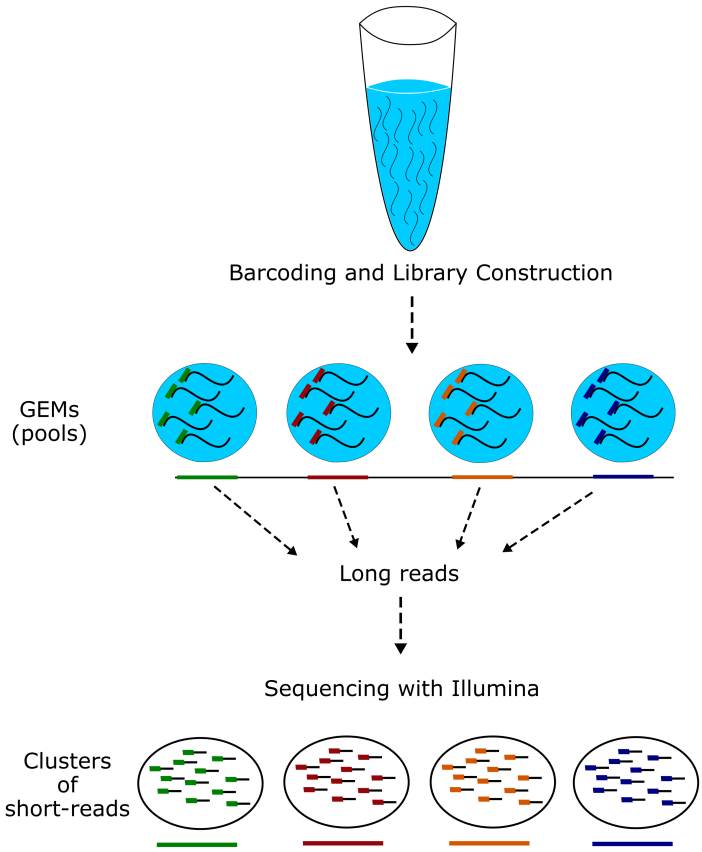


Figure 2.2: Details of 10x technology. Each step of 10x pipeline is briefly depicted. Firstly, 1 ng of high molecular weight DNA is extracted from the sample and distributed to approximately 10^6 pools, where they are barcoded and subjected to priming and polymerase amplification. After the library preparation process, they undergo Illumina sequencing process.

TARDIS has an option to use linked-reads for SV detection triggered with "–10x" option.

The approach we use in 10x mode of TARDIS is similar to that of Common-LAW's (Common Loci structural Alteration Widgets) [186], comparing multiple individuals against the reference genome simultaneously to increase the accuracy of SV discovery. Traditionally, to perform comparative studies involving multiple individuals, a two step process is utilized; each individual is compared with the

reference genome and the list of SVs are compared against each other. CommonLaw, on the other hand, simultaneously compares each genome against the reference to increase the accuracy. The method involves generalizing the MPSV problem to multiple genomes; creating a set of maximal SV clusters based on maximum parsimony. For each SV, if the support coming from multiple donor genomes is higher, then the SV is more likely to be correct. This approach is shown to increase the SV discovery especially among related genomes such as family trios or ethic groups.

In order to formulate our approach to the problem; given the discordant paired-end reads in Equation 2.1, each rp_i is assigned a barcode b_i. We align the reads using mrFAST to allow multiple mappings for each rp_i as given in Equation 2.2.

Our aim is to find the most parsimonious assignment of barcoded discordant reads to the SV clusters. Note that, clusters of discordant paired-end reads that support the same SV breakpoints may have reads with different barcodes.

We previously defined the notion of SV weight utilized in TARDIS (Chapter 2.2); We give a SV score to each maximal cluster using a probabilistic model based on read depth signature and the type of SV. On the contrary, when using Linked-Reads of 10xG, the score of the clusters are not only dependent on our likelihood model but also on what we call the homogeneousness score of a cluster.

Similar to the chi-squared test for identifying and measuring heterogeneity, we define a homogeneousness score for a given SV cluster s. Let s contain n reads with a total of m barcodes $\{b_1, b_2 \ldots, b_m\}$ where $m \leq n$ and let B_i be the subset of reads in s all with the barcode b_i.

We define $h(s)$, the homogeneousness score of s as follows:

$$h(s) = \frac{1}{n^2} \sum i|B_i|, \tag{2.10}$$

where $|B_i|$ is the size of the set B_i.

Thus, modified SV weight becomes

$$weight(s) = w(s) \cdot h(s), \tag{2.11}$$

where w(s) is the previously calculated weight of cluster s that ignores the barcode information and $h(s)$ is the homogeneousness score of cluster s.

As the barcoded reads are given as input to TARDIS in BAM format, the mappings suggested are the best (unique) mappings produced by the read mapper of 10x Genomics, Long Ranger [89]. LongRanger outputs a BAM file, containing position-sorted, aligned reads and attaching the barcode information to each read with a tag field of BX. This tag is composed of nucleotides with exactly 16 digits long, i.e., AGAATGGTCTGCATCG.

Our Linked-Read approach also works in Sensitive Mode that utilizes multiple mapping locations of the reads in order to increase the accuracy. Thus, we use mrFAST or mrsFAST to remap the linked-reads to the reference genome in order to obtain all possible mapping locations of the reads. Since mr/mrsFAST does not have the capability to input barcode information, we hide the barcode information for each read by appending it to the read name as follows; As the barcode is exactly 16 nucleotides long, they fit in the first 4 bytes of unsigned long, so the unsigned long value is appended to the read name with width of 20 digits, zero padded, i.e., $< read - name >< zero - padding >< encoded - barcode >$.

Once all FASTQ files are created and sorted, TARDIS runs mr/mrsFAST for each pair of FASTQ files and creates a DIVET file in the end that contains the barcode information appended to each read. By using a reverse-engineering, we decode the barcodes of each read and utilize them in set-cover phase of TARDIS.

Chapter 3

Structural Variation Discovery with TARDIS

3.1 Introduction

Before the advent of sequencing technologies, the methods to detect genetic variations involved microscopic observations that are ∼3 Mb or more in size, which included aneuploidies, rearrangements, heteromorphisms and fragile sites [193]. Since then, with the advances in experimental molecular biology techniques and sequencing-based approaches, the spectrum of genetic variations have been broadened. These variations involved single nucleotide changes (SNPs), small insertions/deletions (INDELs), repetitive elements that include mini and macro satellites (STRs) and larger variations such as cytogenetic alterations. However, intermediate level, so called submicroscopic variations that range from 50 bp up to 5 Mb in size have been underestimated until the development of BAC arrays such as arrayCGH [194] and ROME [195]. Furthermore, with the help of NGS techniques, the amount of discovered SVs have been increased tremendously and breakpoint resolutions have been improved.

Currently there are several resources that provide a catalog of SVs in human

genome such as 1000 Genomes Project (1000GP) [20, 21], Database of Genomic Variants (DGV) [196], Human Polymorphic Inversion Database (InvFEST) [197], DECIPHER [198]. These databases mostly report all the calls that are detected by some algorithms. However, some recent project such Genome in a Bottle (GIAB) [199, 200] also report high confidence call sets, produced based on multiple platforms such as short-read sequencing, long-read sequencing, etc.

In this chapter, for each distinct SV type that TARDIS can characterize, we give a brief overview and then describe the approach we utilize in order to detect the variation in the clustering phase of TARDIS.

3.2 Characterizing Various Types of SV

3.2.1 Discovering deletions and insertions

It is known that most of the structural variations in human genome involves deletions or insertions, indeed they are currently the best characterized SV types [2, 62]. According to the results of 1000GP, the median number of deletion alleles per human genome is $2,788$, estimated with 88% sensitivity. There are several phenotypic disorders caused by deletions; Williams-Beuren syndrome [201, 202], Smith-Magenis syndrome [203, 204], Carney complex [205], Miller-Dieker lissencephaly syndrome [206], Hereditary neuropathy with pressure palsies [207], cri du chat syndrome [208], Prader-Willi syndrome [209], Crohn's disease [25], developmental delay [210], autism [28, 26, 29, 30] and by insertions; Charcot-Marie-Tooth neuropathy [211], Tay-Sachs disease [212].

In Figure 3.1, we show a sample deletion event discovered by TARDIS. The figure depicts that paired-end reads are mapped to the reference genome where they span the SV event. It is also shown that the read depth is decreased within the breakpoint intervals.

As mentioned before, a valid cluster is a set of alignments of discordant read

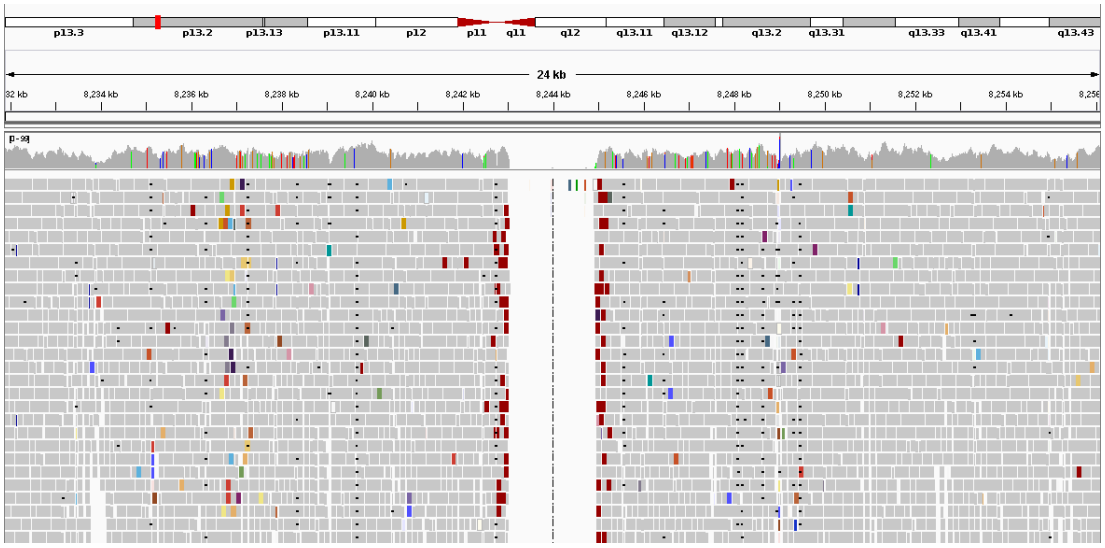


Figure 3.1: Figure shows the IGV [3, 4] visualization of a deletion event predicted by TARDIS within 19:8,231,867-8,256,118 for CHM1 genome [5, 6]. Absence or decrease of read-depth within the breakpoint is an indication of a deletion.

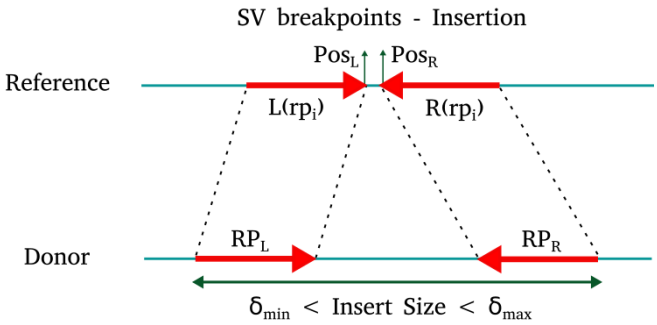
pairs and/or split reads that signal the same particular SV event denoted by;

$$VClus_i = \{a_k rp_{i_1}, a_k rp_{i_2}, ..., a_k rp_{i_n}\} \tag{3.1}$$

where the read pair $a_k rp_i$ is aligned to the reference sequence and clustered in ith cluster (k denotes that rp_i has k different alignments for the cases where multiple alignment is used). Note that read pairs are allowed to be included in other clusters also. This is limited by an upper bound parameter where the default value is decided to be 10 (higher values have the risk of increasing memory consumption enormously for some data sets). Figure 3.2 shows the insertion and deletion signatures of read pairs when mapped to the reference genome.

Pos_L and Pos_R shows left and right breakpoints respectively, $L(rp_i)$ and $R(rp_i)$ are the mappings of left (RP_L) and right (RP_R) ends and we assume that the fragment sizes are in the range $[\delta_{min}, \delta_{max}]$. We note that δ_{min} and δ_{max} are calculated using the following formulations:

A



B

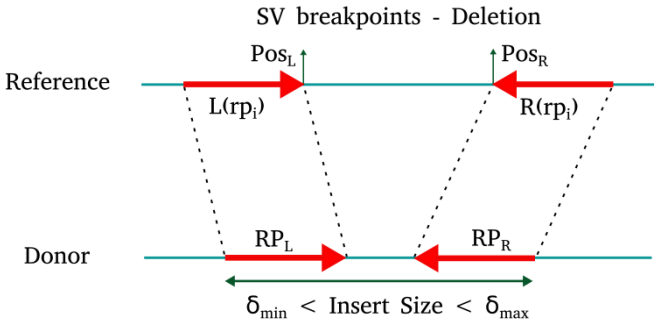


Figure 3.2: Read pairs mapped to the reference genome A) insertion signature, B) deletion signature.

$$
\begin{aligned}
concordant_{min} &= mean - (4 \times stdev) \\
concordant_{max} &= mean + (4 \times stdev) \\
\delta_{min} &= concordant_{min} - (2 * readLength) \\
\delta_{max} &= concordant_{max} - (2 * readLength)
\end{aligned}
\tag{3.2}
$$

where the mean is calculated using the span size distribution of the read-pairs mapped to the reference genome as depicted in Figure 1.6 and read-length is the length of the reads in the BAM file.

We scan the genome from beginning to end and consider each position as a potential breakpoint denoted as P_{Br} to create maximum sets of overlapping intervals. We evaluate all sets of read-pairs where the mates map in $+/-$ stands both

for deletions and insertions. For deletions, we cluster the read-pairs within the interval $[P_{Br} - \delta_{max}, P_{Br}]$ and for insertions within $[P_{Br} - (R(rp_i) - L(rp_i)), P_{Br}]$. More formally, we require the following conditions to be satisfied for insertion:

$$\begin{aligned} &L(rp_i) \leq Pos_L \\ &R(rp_i) \geq Pos_R \\ &\delta_{min} - InsLen < R(rp_i) - L(rp_i) < \delta_{max} - InsLen \end{aligned} \tag{3.3}$$

and the following conditions for deletion:

$$\begin{aligned} &L(rp_i) \leq Pos_L \\ &R(rp_i) \geq Pos_R \\ &\delta_{min} + DelLen < R(rp_i) - L(rp_i) < \delta_{max} + DelLen \end{aligned} \tag{3.4}$$

Once we have the set of reads (a new element must be added in order to denote it as a newer cluster), we follow the steps described in Section 2.3.1.

3.2.2 Characterizing inversions

In contrast to copy number variations such as deletions, insertions and duplications, inversions do not cause any gain or loss of genetic material but they change the orientation of genomic segments. Therefore, this type of variations are known as balanced rearrangements. Detection of inversions is made possible with the introduction of paired-end sequencing since the array-based approaches are only able to detect copy number differences. Thus, inversions were poorly studied before 2005 [62]. [2, 213, 214] and [215] are among the first approaches to detect inversions in human genome utilizing the human genome assembly. Today, with the introduction of HTS, the number of detected inversions increased, however, they are not characterized as well as the other SV events. Indeed, Phase 3 of 1000GP [20] suggests 37 inversions per individual with only 32% sensitivity. On the other hand, InvFEST database [197], which harbors inversions reported in

the literature, currently has 86 validated inversions. The reason behind the drawbacks of inversion discovery lies in the fact that breakpoints usually lie within repeated regions and they are not subject to read-depth signature as balanced rearrangements do not alter read-depth [92]. There are many studies revealing the association between inversions and phenotype such as Hemophilia A [216], Hunter syndrome [217] and disruption of the emerin gene in Emery-Dreifuss muscular dystrophy [218]. On the other hand, most of the studies show that inversions have no effect on the individual but increases the risk of diseases related to further rearrangements in the offspring such as microdeletions [219, 193]. Some of these phenotypes that are observed in the offspings of a parent carrying an inversion are Williams-Beuren syndrome [220], developmental delay [221], Sotos Syndrome [222], Angelman Syndrome [193], etc.

Figure 3.3 shows an inversion event detected by TARDIS where the green arrows show $+/+$ mappings and blue arrows show $-/-$ mappings as they are the signatures for an inverted segment within the genome as given in Figure 3.4.

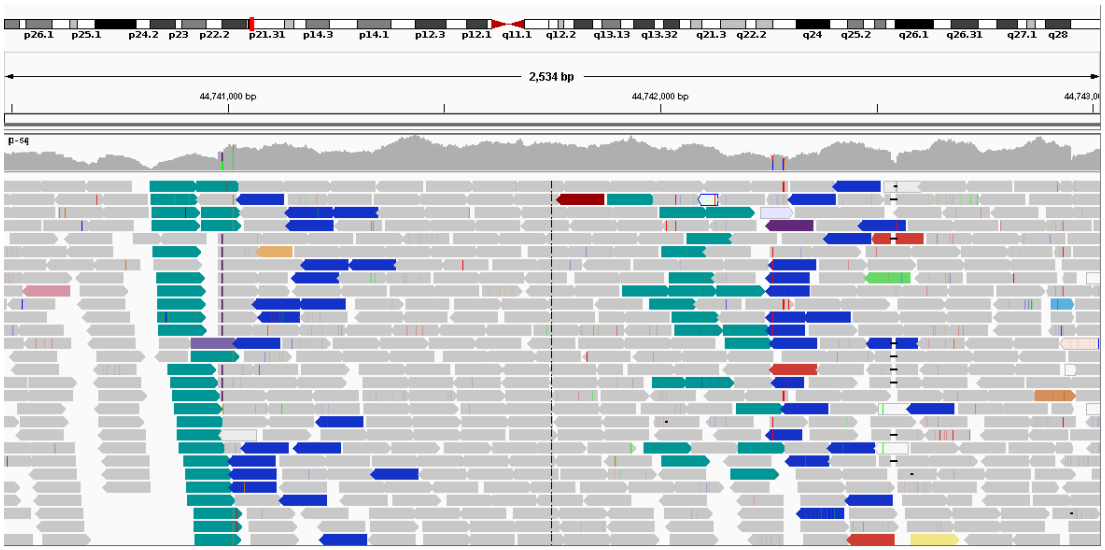


Figure 3.3: Figure shows the IGV output of an inversion event predicted by TARDIS within 3:44,740,482-44,743,019 for CHM1 genome.

Similar to deletions and insertions, we scan the genome from beginning to end and consider each position as a potential breakpoint denoted as P_{Br}. We evaluate all sets of read-pairs where the mates map in $+/+$ or $-/-$ stands and

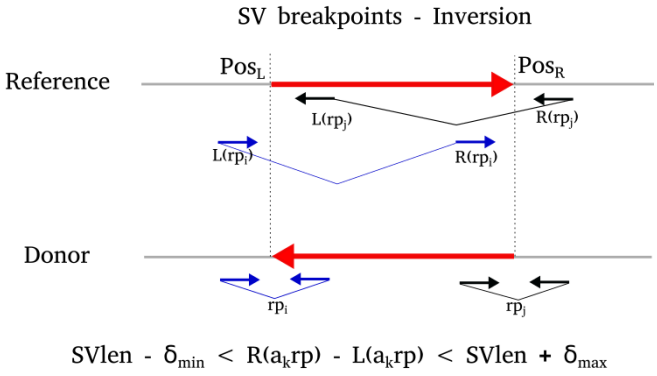


Figure 3.4: Inversion signature of the read pairs mapped to the reference genome.

cluster the read-pairs within the interval $[P_{Br} - \delta_{max}, P_{Br}]$ for $+/+$ mappings and $[P_{Br}, P_{Br} + \delta_{max}]$ for $-/-$ mappings, so we cluster inversions within the range $SVlen - \delta_{max} \leq R(a_k rp_i) - L(a_k rp_i) \leq SVlen + \delta_{max}$.

Finally, there are some conditions to be satisfied for inversions:

$$\begin{aligned} L(rp_i) \leq Pos_L \leq R(rp_i) \quad \text{and} \quad or(rp_i) = +/+ \\ Pos_R \geq R(rp_i) \quad \text{and} \quad or(rp_i) = +/+ \\ L(rp_i) \leq Pos_R \leq R(rp_i) \quad \text{and} \quad or(rp_i) = -/- \\ Pos_L \leq L(rp_i) \quad \text{and} \quad or(rp_i) = -/- \end{aligned} \tag{3.5}$$

3.2.3 Transposon insertions

Transposable genetic elements (TEs), also known as transposons or mobile element insertions (MEIs), are repetitive or movable DNA segments that occupy nearly half ($\sim 44\%$) of the human genome [11, 223]. Most of these elements are currently inactive but the active ones ($< 0.05\%$) including Alu, L1 and SVA families, still contribute to genetic diversity among individuals by generating novel insertions. Studies reveal that mobile element insertions are also related to various diseases such as cancer [224], hemophilia A [225], muscular dystrophy [226]

and are also related to the creation [227] and expansion of interspersed segmental duplications [228]. 1000GP results suggest 1,218 mobile element insertion sites per individual in human genome (mostly Alu insertions) [20]. There is also a database of repetitive elements in eukaryotic genomes called Repbase [229], which contains 46,248 MEI sequences as of the end of 2017.

In this section, we present our approach utilized in TARDIS to discovering mobile element insertions inside the genome. Briefly, we try to characterize MEIs including Alu, L1 and SVA (any type of MEI can be detected by TARDIS given necessary annotation) in the genome where a segment of DNA is copied to another location inter or intra chromosome. With this approach we do not only detect MEIs but also decrease our false negative calls as some of the mobile element insertions might mistakenly be categorized as other forms of SVs. The alignment on the left of Figure 3.5 seems to be a deletion event, however this is shown to be a false prediction when the duplicated region in the donor genome is not considered, thus, this event should be categorized as a transposon insertion.

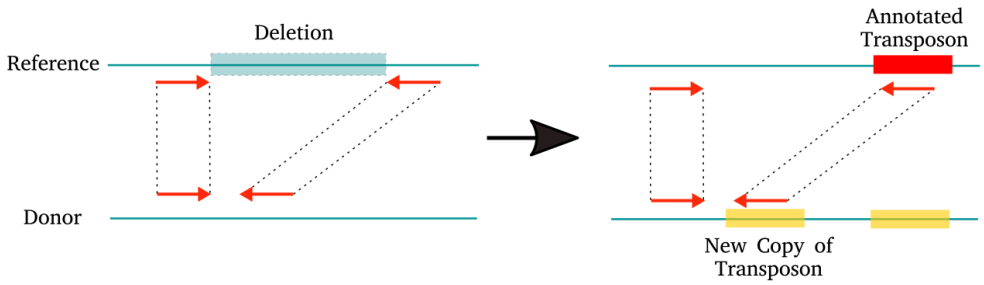


Figure 3.5: An example of a false SV prediction is depicted in the figure. There is a deletion event in the left mapping when the duplication in the genome is not considered. We need to check wheter any of the pairs hit the annotated transposon interval in order to make a correct prediction since the MEI insertions can be underestimated.

MEI clustering is slightly different from the clustering approach we utilize for other types of SVs. Here, we check whether any of the ends for each paired-end read is inside the annotated transposons. We iterate through the genome from start to end considering each P_{Br} as a potential breakpoint and group the reads within the intervals of $[P_{Br} - \delta_{max}, P_{Br} + \delta_{max}]$, where P_{Br} denotes the locus where

the transposon insertion occurs. Figure 3.6 shows a representation of the MEI clustering.

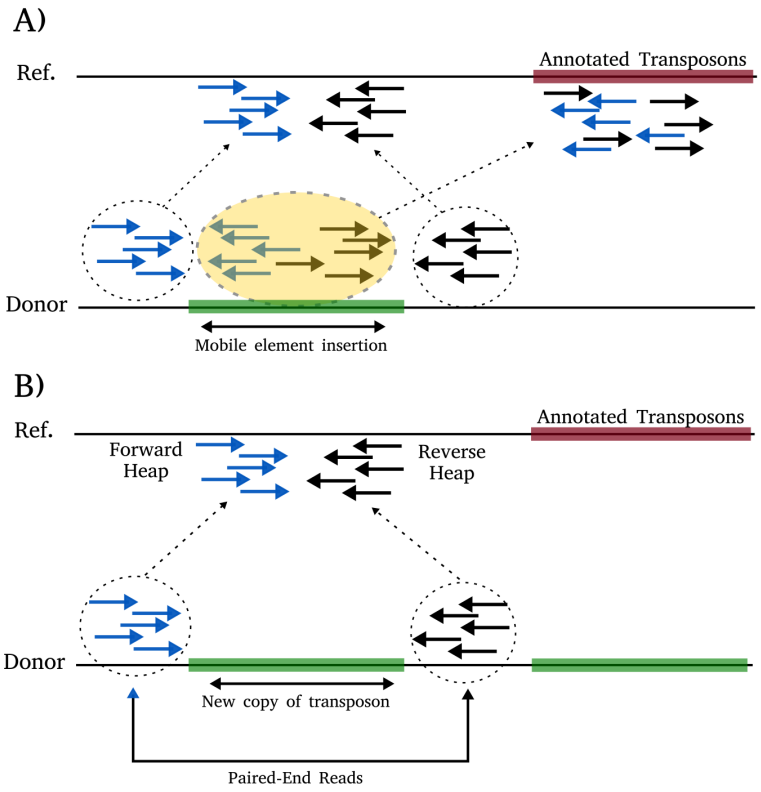


Figure 3.6: The figure depicts the overview of MEI clustering approach we utilize in TARDIS. (A) We first check the paired-end reads where one end maps to an annotated transposon and the other to elsewhere within the genome. (B) For such cases, we cluster the pairs that map to elsewhere in the the genome based on their orientations within an interval. Then we bring forward and reverse pairs together inside the same cluster and treat them as paired-end reads in order to detect the insertion breakpoints.

As the figure shows, we make use of two distinct heaps for reads in forward and reverse orientations. These reads are obtained from read-pairs where one of the ends map to an annotated transposon and the other end maps to any location inside the reference genome. Thus, our aim is to bring the reads in forward and reverse orientations into the same cluster as if they are novel paired-end reads that span the insertion breakpoints.

It should also be noted that [185] is the first to formulate MEIs as depicted in Figure 3.7 for direct and in Figure 3.8 for inverted orientations. These cases are categorized based on the position of P_{Br} and the orientation of the annotated genomic segment. Figures 3.7 and 3.8 show that the genomic segment between Pos_L and Pos_R is inserted into P_{Br} in direct and inverted orientations respectively. It should be noted that any one of these conditions should hold for a transposon insertion event.

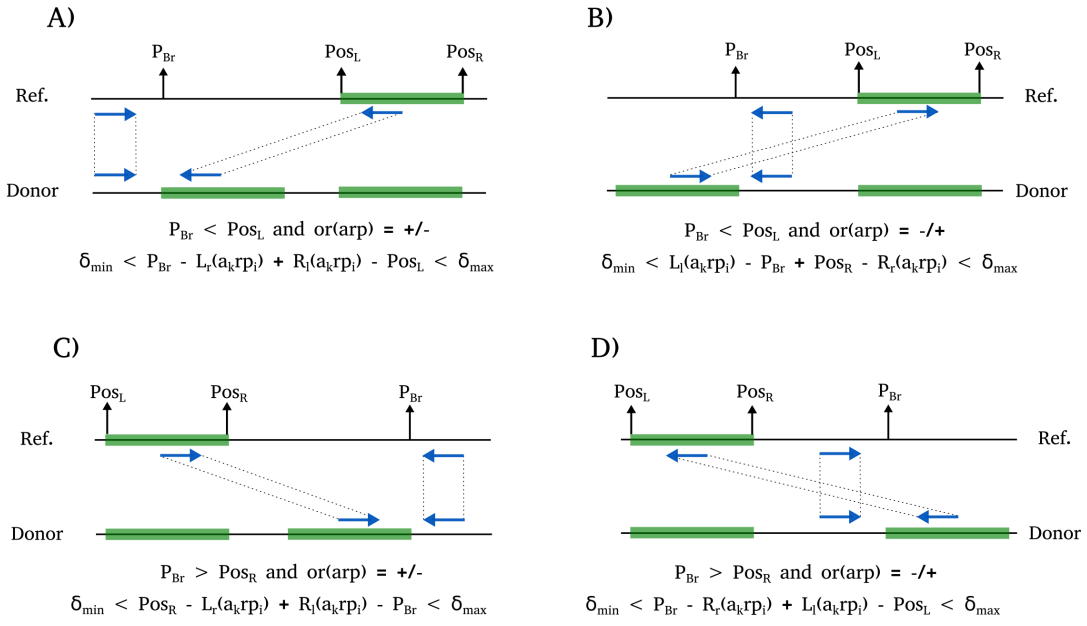


Figure 3.7: There are four different cases for mobile elements (copy events) in direct orientation. The cases are based on the position of P_{Br}, and orientation of the mappings.

3.2.4 Nuclear mitochondria (NUMT) insertions

It is known that there is an ongoing transfer of genetic information from the mitochondrial DNA into the nuclear genome of eukaryotes [230, 231] at a rate of $5.1 - 5.6 \times 10^{-6}$ per germ cell per generation [232]. These insertions are related

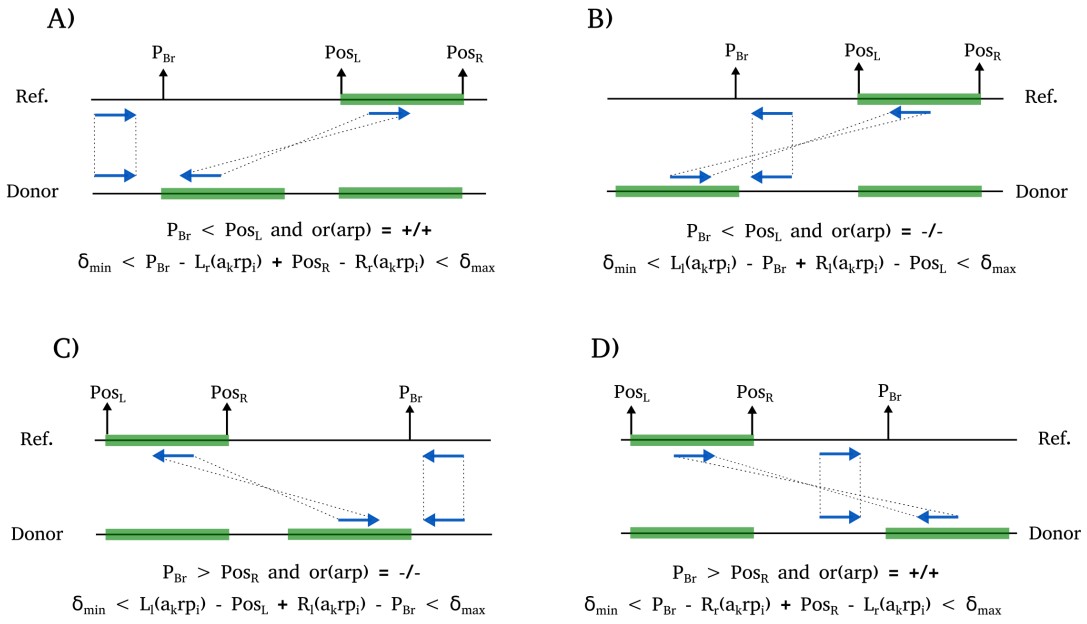


Figure 3.8: There are 4 different cases for mobile elements (copy events) in inverted orientation. The cases are based on the position of P_{Br}, and orientation of the mappings.

to various genetic disorders such as Pallister-Hall syndrome [233], mucolipidosis IV [234], and they are related to mitochondrial diseases [235] and heteroplasmy [236]. Furthermore, they harbor some footprints regarding the genetic history of humans. We also note that Phase 3 of 1000GP reports 5.3 NUMTs per individual [20].

With TARDIS we are able to detect mitochondria insertions in the nuclear genomes with varying sizes and it is the only SV caller with such capability except dinumt [237], which is the specific to NUMT insertion detection using HTS technology. Figure 3.9 briefly depicts our approach for NUMT insertion detection. Similar to MEI discovery, we check whether any of the pairs for each paired-end read map to mitochondria. We cluster end reads that map in forward or reverse strands elsewhere in the genome within the intervals of $[P_{Br} - \delta_{max}, P_{Br} + \delta_{max}]$. Here, P_{Br} denotes the locus where the NUMT insertion occurs. Our aim is to gather forward and reverse mapped reads inside the same cluster and treat them

as novel paired-end reads.

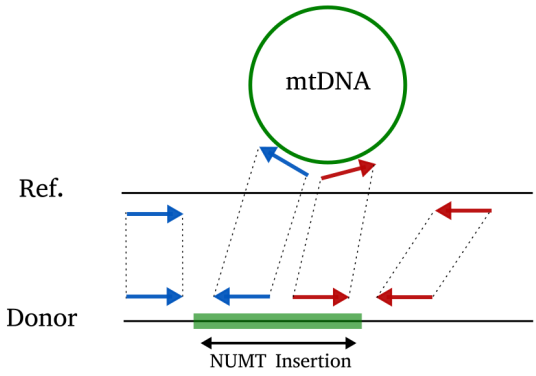


Figure 3.9: For NUMT insertion, we check whether any of the pair maps to mitochondria. Such cases is an indication of NUMT insertions within the genome.

3.2.5 Duplications

Segmental duplications are forms of copy number variations where a segment spanning 1 Kb to 6 Mb of a DNA is duplicated and copied elsewhere in the genome and still retains $> 90\%$ sequence similarity to the original copy. If the segment is placed adjacent to the original copy, it is categorized as a tandem duplication event. If positioned elsewhere, it is called interspersed duplication as shown in Figure 1.1 [67]. $\sim 5\%$ of the human genome is covered with SDs [193]) and that plays an important role in genomic differences such as gene duplications among species that results in phenotypic variations or diseases. Furthermore, duplications are the major source of evolution as duplicated segments create diversity and new genes are formed over time where the short term consequences are the genetic diseases [238], but also adaptive evolution [239]. There are numerous diseases related to duplications such as schizophrenia, epilepsy, intellectual disability, development delays [240].

The approaches to detect duplications are based on experimental methods such as FISH, array-CGH and sequencing based computational methods. However, there is a lack of SV detection tools that distinguish duplications such as tandem

or interspersed. Phase 3 of 1000GP [7] provides a catalog of SVs that primarily focused on characterizing deletions, insertions, and mobile element transpositions, however, it also generated a set of inversion calls. A careful analysis shows that a substantial fraction of the predicted inversions are in fact complex rearrangements that include duplications, inverted duplications and deletions within an inverted segment (Figure 3.10). This is because the read pair signatures that signal such complex SVs are exactly the same with simple SVs as shown in Figure 2.1.

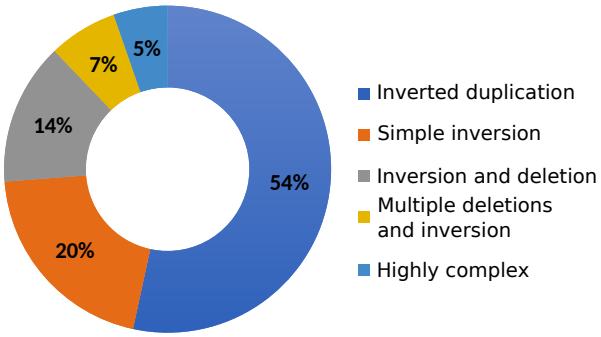


Figure 3.10: Relative abundance of complex SVs among the inversion calls reported in the 1000 Genomes Project [7]. 54% of predicted inversions are in fact inverted duplications and only 20% are correctly predicted as simple inversions.

In this section we describe novel algorithms to accurately characterize complex SVs such as tandem or interspersed segmental duplications. Note that TARDIS is the first method to distinguish duplications as tandem or interspersed in direct or inverted orientations.

3.2.5.1 Tandem Duplications

Tandem duplication is the duplication of a DNA segment that is copied adjacent to the original copy. These SVs are easier to detect than interspersed SDs by short read sequencing approaches as the duplicated and original copies are adjacent to each other, which creates a $-/+$ orientation signature that is distinct among the other types of events. Figure 3.11 depicts the IGV visualization of a tandem duplication discovered by TARDIS. The orientations of the read pairs (shown in

green) clearly demonstrate the tandem duplication signature that we seek. It should be noted that the read depth is decreased within the breakpoint intervals as described in Figure 1.7.

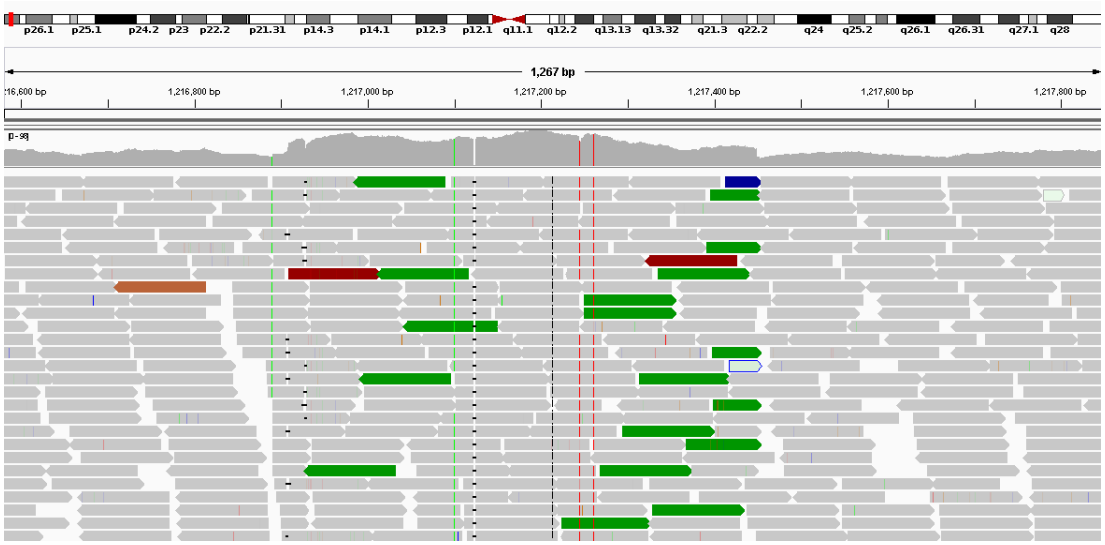


Figure 3.11: Figure shows the IGV output of a tandem duplication event predicted by TARDIS within 3:1,216,580-1,217,848 for CHM1 genome.

Clustering of these read-pairs are similar to deletions. We first fetch each alignment in the BAM file and store the discordant paired-end reads and split-reads that are in $-/+$ orientation. Then we scan the genome from start to end by considering each position as a potential breakpoint denoted as P_{Br} to create maximum sets of overlapping intervals. The clusters we create for tandem duplications are within the interval $[P_{Br} - \delta_{max}, P_{Br}]$. Figure 3.12 shows the detailed description of the sequence signature that we utilize in TARDIS. Note that Pos_L is the left and Pos_R is the right end of the duplicated segment and $L_l(a_k rp_i)$, $L_r(a_k rp_i)$ are the k^{th} mapping of the left and right end of read-pair (rp_i) respectively. Similarly $R_l(a_k rp_i)$ and $R_r(a_k rp_i)$ are the left and right mappings of the right read-pair.

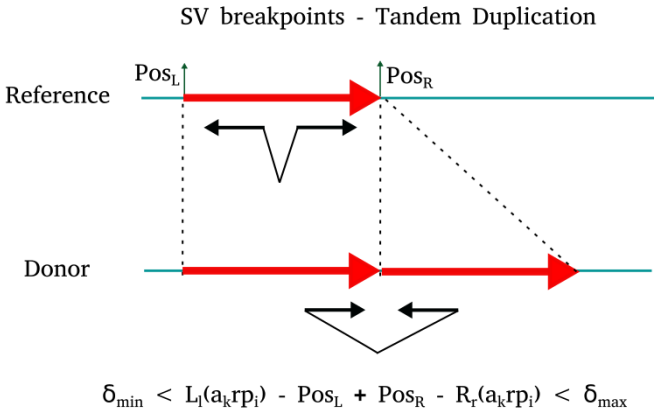


Figure 3.12: Tandem duplication signature of the read pairs mapped to the reference genome.

3.2.5.2 Interspersed Segmental Duplications

Unlike tandem duplications, the duplicated segment for interspersed duplications are placed away from the original copy in direct or inverted orientation (we use inverted duplication and interspersed inverted segmental duplication interchangeably throughout the text).

Figure 3.13 shows the IGV visualization of interspersed segmental duplications in (A) direct and (B) inverted orientations. Due to the fact that the sequence signatures of inversions and inverted duplications are exactly the same, characterization of such events using paired-end read signatures simultaneously is very challenging (Likewise deletion/tandem duplication and direct duplication signatures are also the same). Therefore it is easy to make false predictions. TARDIS is the first approach to characterize these complex SV events.

3.2.5.2.1 Inverted duplications: We assume the fragment sizes for read pairs are in the range $[\delta_{min}, \delta_{max}]$, and we denote the insertion breakpoint of the duplication as P_{Br} and the locus of the duplicated sequence is $[Pos_L, Pos_R]$ (Figure 3.14A). Similar to the approach we utilize for the other SV events, we scan

the genome from beginning to end, and we consider each position as a potential duplication insertion breakpoint P_{Br}.

We consider all sets of read pairs where both mates map in the same strand (i.e., $+/+$ and $-/-$) within interval $[P_{Br} - \delta_{max}, P_{Br}]$ and $[P_{Br}, P_{Br} + \delta_{max}]$ respectively as clusters that potentially signal an inverted duplication.

3.2.5.2.2 Interspersed direct duplications: We create the valid clusters in a way similar to the inverted duplications with the exception of the required read mapping properties. For direct duplications we require each mate of a read pair to map in opposing strands (i.e., $+/-$ and $-/+$).

The clusters for inverted duplication encompasses both $+/+$ and $-/-$ mappings as the signature for this event involves both types of mappings. Similarly, we gather the read-pairs of $+/-$ and $-/+$ inside the same cluster since only one signature is not enough to decide the type of SV event. Separating inversions from inverted duplications or deletions and tandem duplications from direct duplications is done with the probabilistic approach we utilize in the set-cover step described previously.

3.3 Incorporating Split Read Information To Improve SV Calls

Split reads can be defined as the reads that are split into multiple segments and partially mapped to the reference genome. Therefore, it is evident that detection of split-reads begin during the read mapping step where the mapping information is produced.

We formally define the read mapping problem as follows. Given a read R and a reference genome S, our aim is to detect all or the best mappings of R within S with some error threshold. This problem is called "approximate string

matching problem" as most read mappers allow some gaps and mismatches that are caused due to genetic mutations or sequencing errors up to some threshold based on a metric such as Hamming distance. As given in Figure 3.15, some reads are aligned to the reference genome with multiple alignments/mappings allowed. Among these mappings, some of them are exact, where all the nucleotides exactly match, and some are partial, where some nucleotides do not match. The score of exact matches are higher than the partial mappings, so read mappers aim to find these higher scoring alignments.

We utilize CIGAR, one of the mandatory fields of SAM/BAM format, as the main source of data for split read detection. More precisely, when the reads are mapped to the reference genome, due to some INDELs, gaps, etc, mappings are not exact and CIGAR string indicates the details of the mapping (Figure 3.16).

Note that clipped reads involve sequences whose ends are clipped-off as they are not aligned to the reference genome. There are two types of clippings; (1) Soft clipping, where the clipped part is not aligned to the reference but the unaligned sequence is present in the mapping; (2) Hard clipping, where the clipped part is not aligned too however, clipped segment is also not available in the mapping.

In summary, split reads suggest potential SV breakpoints when the unmapped (clipped) segment of the read is remapped to somewhere else and the distance between the main read and the remapped segment spans the putative structural variation event.

3.3.1 Detection and clustering of split reads

As described above, in order to detect structural variations, where the reads span the putative SV breakpoints, checking the split-read signature is currently the most relevant way. Furthermore, the accuracy of the approach increases with larger reads because they will more likely to span SV breakpoints.

As TARDIS works independently for each chromosome, detection of split-reads

for an arbitrary chromosome can be described as follows:

1. **Read the reference sequence:** The first step involves reading the DNA sequence of the chromosome into the memory to create a hash index. The collision-free hash table whose size is 4^k is used to store overlapping k-mers of 10 bps from the reference sequence using a linked list structure. The function uses a simple two-bit encoding scheme.
2. **Store the soft-clips using the mapping information:** As we iterate through each mapping in TARDIS, we check whether the read includes an unmapped segment at the beginning or at the end. This is the soft-clip information that resides within the CIGAR string of the mapping (A match is denoted by M and a soft clip is denoted by S). We keep these mappings in a linked list structure for further processing in the following steps.
3. **Remapping soft-clips to the reference genome:** In this step, we utilize the mappings that harbor soft-clipped segments that are collected in Step 2. The split coming from the soft-clipped chunk of the original read is remapped to the reference genome in order to create the second pair of the mapping. Inherently, the first pair is the original read excluding the soft-clipped chunk. This is shown in Figure 3.17.

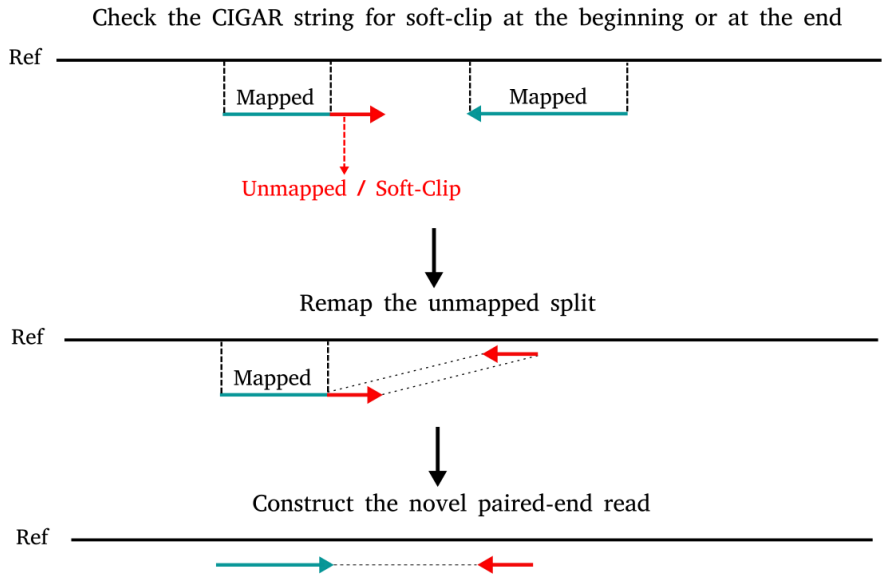


Figure 3.17: Mapping soft-clips to the reference genome.

In order to map the split to the reference genome, we have a requirement such that the distance between the split and the original read is within some predefined range +/− 100,000 bps. We believe that this bound is enough to detect potential SV regions because most events happen within such a range and larger values will likely to increase false positives.

Briefly, we initially hash the first k nucleotides of the split using the same approach that we use to hash the k-mers of the reference genome. Once we locate the linked list of the hash value within the hash table, we compare the split with the reference by using the Hamming distance metric. Here, we don't search for exact matches only, but we also allow for some mismatches. Note that we use the same procedure to search for the reverse mapping of the split, i.e., 3' to 5'. Finally we return up to 11 distinct mapping locations of the split.

4. **SV Type Resolution** Once we have obtained the possible mapping locations of the split, the tricky part is to decide the type of SV event that these novel paired-end reads suggest. To make this decision, we consider a bunch of signatures as illustrated in Figure 3.18. These signatures are dependent

on, (1) The orientation of the original read and split; (2) Mapping location of the split, i.e., before or after the original read, (3) The location of the soft-clip, i.e., at the beginning or at the end of the read.

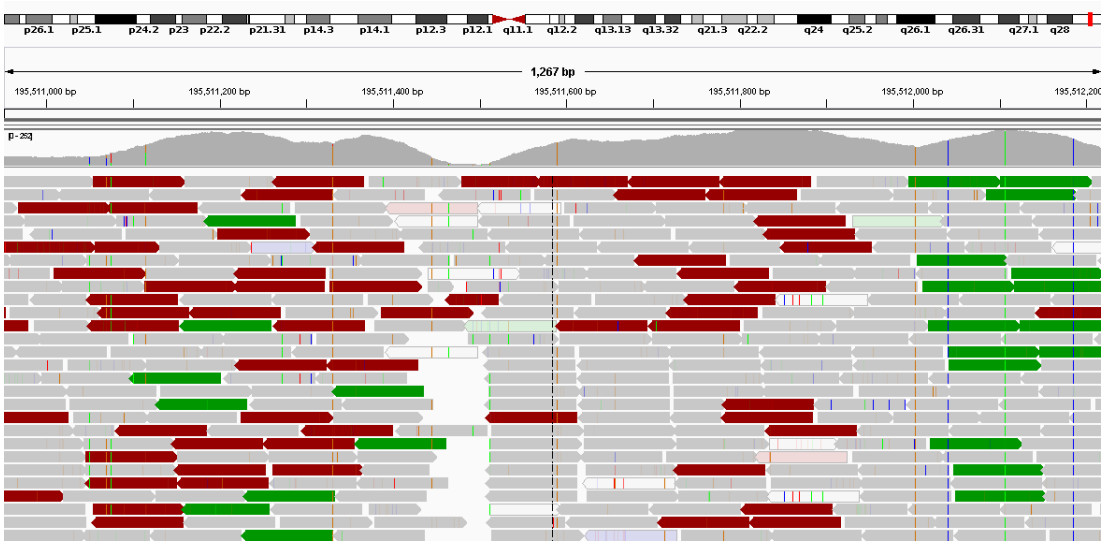
We finally add a wrong-map-window to the start and end locations of the reads and utilize them in the clustering in conjunction with the other paired-end reads.

3.3.2 Runtime and memory usage of split reads

Split read operation is memory intensive because it involves a read mapping step, i.e., multiple mappings. Therefore, TARDIS needs to keep the reference chromosome's DNA sequence and the related hash table in memory. While keeping the reference sequence for chromosome 1 takes around 250 MB of memory, the hash table, which keeps each 10 bp k-mers in a linked list require around 2,5 GB of memory.

The most intensive operation involves comparing the split with the original genome sequence using Hamming distance. The reason for this is that each split's first 10 bps is hashed and the value of the hash is compared with the k-mers of the reference genome, which is stored in a linked list structure. Because of possible collisions, the linked list grows exponentially and comparison operation requires high amounts of computation.

A



B

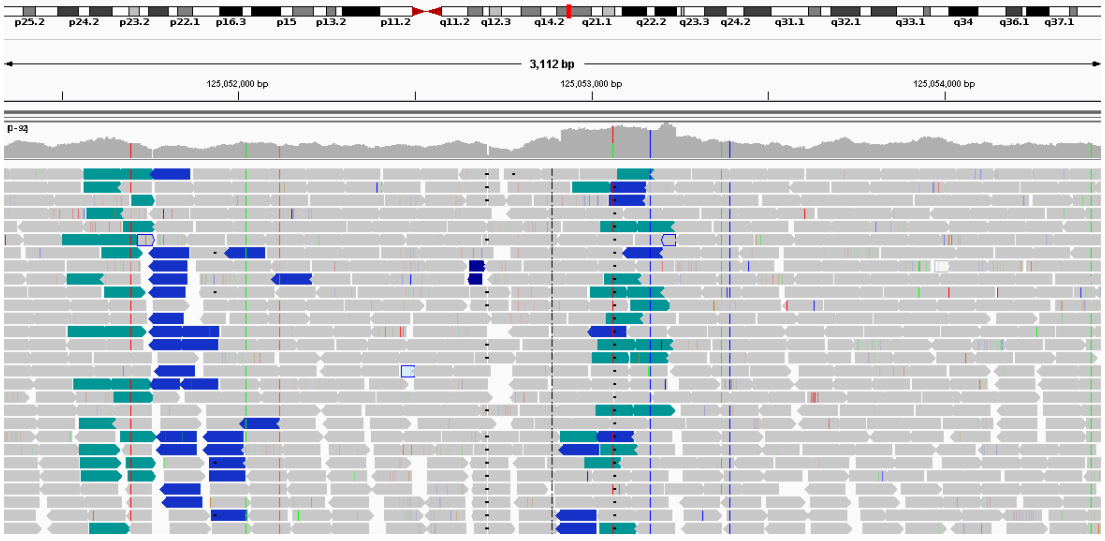
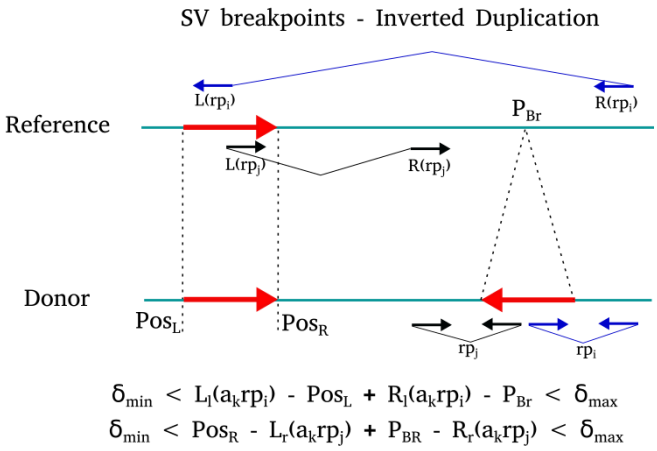


Figure 3.13: IGV visualization of interspersed SD in A) direct orientation and B) inverted orientation. It should be clear that the signature in (A) is +/− and −/+, in (B) −/− and +/+. The first one is exactly the same as the signature of deletion and tandem duplication, the second one as inversion.

A



B

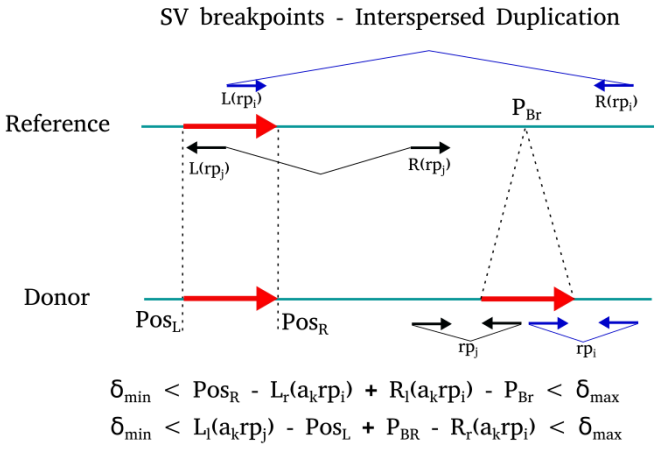


Figure 3.14: The sequence signatures for interspersed SDs in (A) inverted (B) direct orientations.

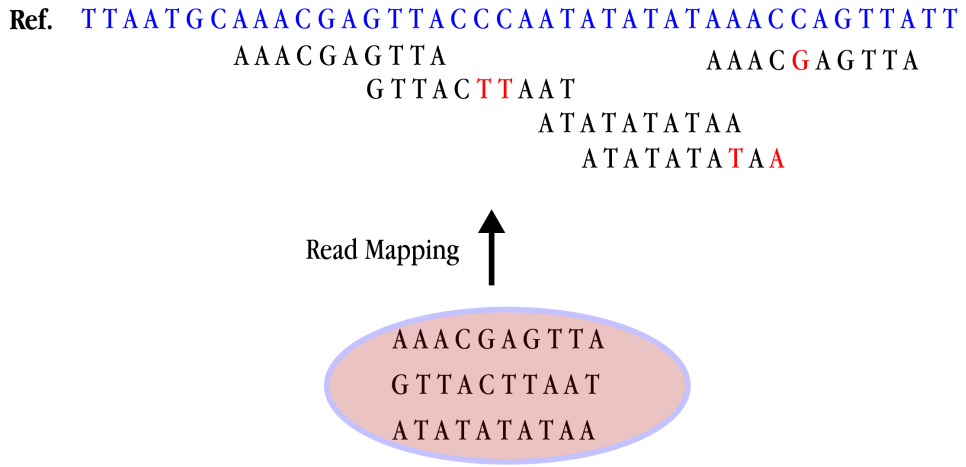


Figure 3.15: Figure shows some reads mapped to the reference genome with multiple mappings allowed. We also show how the reads align with some mismatches allowed. The nucleotides in red color are the mismatches.

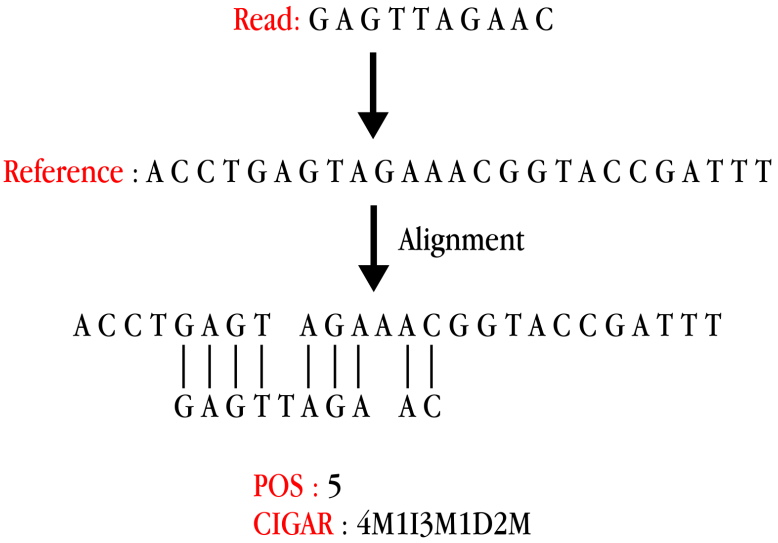


Figure 3.16: When aligning a read to a reference genome, some bases match, some don't. SAM/BAM specification outputs this information in a CIGAR string. The position of the read aligned to the reference is 0-based starting position of the alignment.

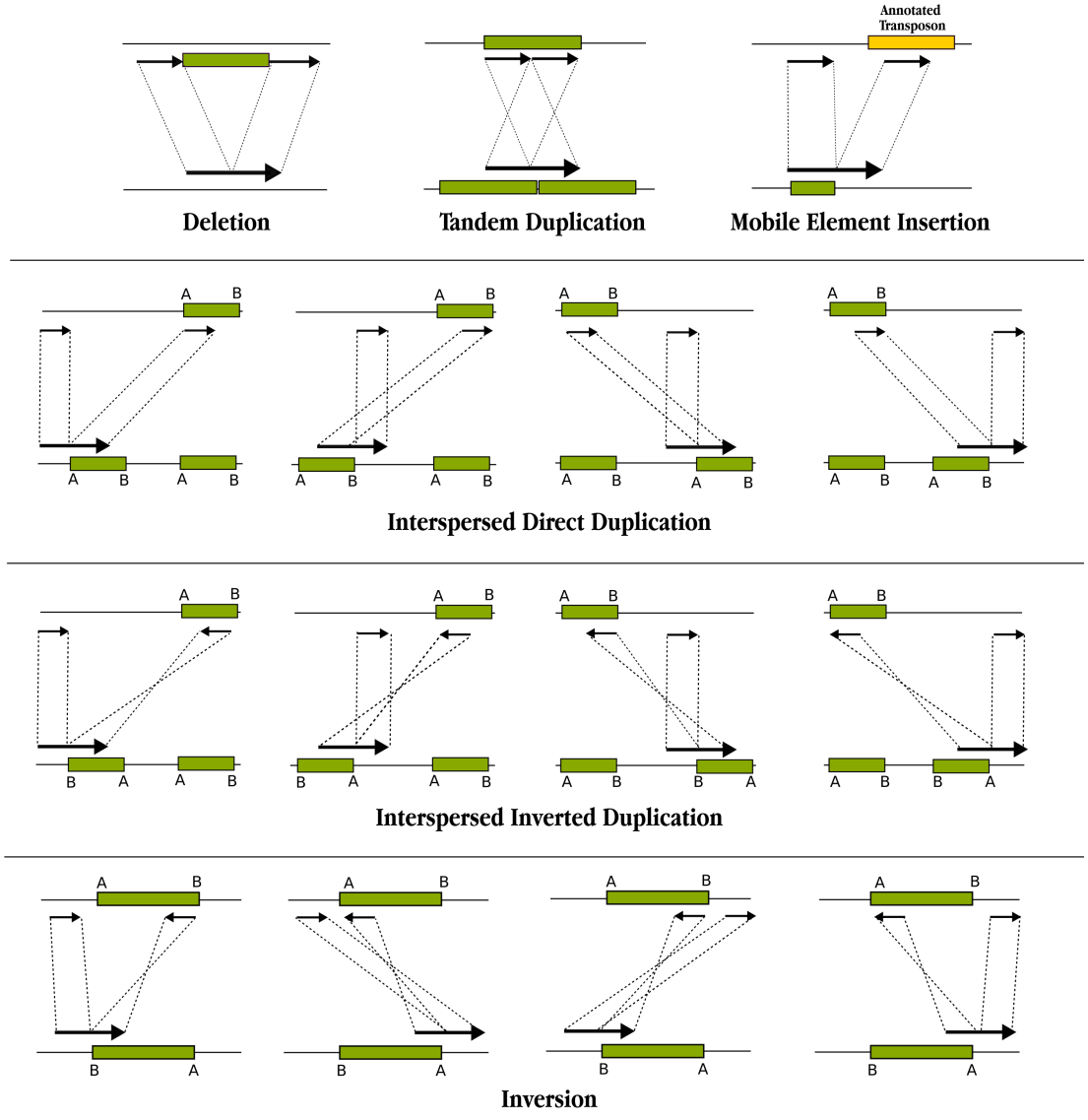


Figure 3.18: Split read signatures used by TARDIS to characterize SV types of deletion, novel sequence insertion, transposon insertion, inversion, tandem duplication and interspersed segmental duplication in direct and inverted orientations. Briefly, when a read is mapped to the reference genome, the SV is hidden inside the read and this is resolved by splitting the read into two segments.

Chapter 4

Results

In this chapter, we analyze the performance of TARDIS by utilizing both simulation and real data experiments in order to benchmark the accuracy for deletion, inversion, mobile element insertion, nuclear mitochondria insertion, tandem duplication and interspersed segmental duplication discovery. These experiments involve comparison against other state-of-the-art SV discovery tools, namely LUMPY [170] and DELLY [167]. To validate our results using real data, we mostly utilized long reads generated with single molecule real time (SMRT) technology (i.e. PacBio) to be able to cross-validate and compare our predictions with an orthogonal technology. These call sets can be regarded as the gold standard for us.

In the following two sections, we present the results for simulation and real data experiments using Quick mode of TARDIS. Then in Section 4.3, we evaluate the performance of Sensitive Mode compared to Quick Mode. Next, the results involving Linked-Read data of 10xG is presented and finally we analyze the time and memory requirements of the each tool in Section 4.5.

4.1 Simulation

In order to evaluate the performance of TARDIS, we developed a new simulator called CNVSim to simulate five types of SVs including deletions, inversions, tandem duplications, inverted duplications and interspersed direct duplications. We simulated a deletion by removing a segment from the reference genome, an inversion by replacing a segment by its reverse complement and a tandem duplication by replacing a segment by two copies of itself. Finally, an interspersed duplication is placed by inserting a copy of itself either to the left or to the right of the original segment (For interspersed inverted duplications, the copy is inserted in reverse direction).

We simulated SVs of random lengths selected uniformly between 500 bps and 10 Kbp. For interspersed duplications, the distance from the new paralog to the original copy is chosen uniformly random between 5,000 bps and 50 Kbp. All segments are sampled randomly from the well-defined (i.e., no assembly gaps) regions in the reference genome, and guaranteed to be non-overlapping. Each simulated SV can be in homozygous or heterozygous state.

Based on the human reference genome (GRCh37), we simulated total of 1,200 SVs including 400 deletions, 200 inversions, 200 tandem duplications, 200 inverted duplications and 200 interspersed direct duplications. We then simulated WGS data at four depth of coverages 10X, 20X, 30X, 60X using wgsim (`https://github.com/lh3/wgsim`). We mapped the reads back to the human reference genome (GRCh37) using BWA-MEM [116]. Finally we obtained structural variation call sets using TARDIS, DELLY [167] and LUMPY [170].

Table 4.1 shows the true positive rate (TPR) and false discovery rate (FDR) of TARDIS compared to DELLY and LUMPY on the simulated data. The sensitivity of TARDIS is comparable to others for deletions and inversions, but TARDIS achieved a substantially higher TDR for duplications as the other tools are not able to characterize interspersed duplications. Additionally, TARDIS suffered very low FDR compared to the other tools we tested.

Table 4.1: Summary of simulation predictions by TARDIS, LUMPY and DELLY.

SV Type	Cov.	TARDIS			DELLY			LUMPY		
		#Predicted	FDR	TPR	#Predicted	FDR	TPR	#Predicted	FDR	TPR
Deletion	10X	373	**0.063**	0.933	383	0.312	**0.958**	316	0.315	0.790
	20X	380	**0.036**	0.950	387	0.329	**0.968**	377	0.327	0.943
	30X	384	**0.047**	0.960	389	0.330	**0.973**	379	0.328	0.948
	60X	386	**0.052**	0.965	391	0.330	**0.978**	383	0.329	0.958
Inversion	10X	194	0.025	0.970	197	0.482	**0.985**	189	**0.000**	0.945
	20X	196	0.011	0.980	197	0.495	**0.985**	193	**0.000**	0.965
	30X	199	0.003	**0.995**	198	0.495	0.990	194	**0.000**	0.970
	60X	199	0.009	**0.995**	198	0.495	0.990	194	**0.000**	0.970
Duplication	10X	560	**0.004**	**0.933**	300	0.204	0.500	245	0.202	0.408
	20X	576	**0.010**	**0.960**	309	0.202	0.515	299	0.205	0.498
	30X	580	**0.004**	**0.967**	309	0.204	0.515	301	0.202	0.502
	60X	582	**0.018**	**0.970**	311	0.205	0.518	301	0.206	0.502

We show the true positive rate/recall and false discovery rates (TPR and FDR) of TARDIS, LUMPY, and DELLY at different depths of coverage from 10X to 60X for Deletions, Inversions, and Segmental Duplications. Total number of SVs are 1,200; 400 deletions, 200 inversions, 600 duplications. Note that LUMPY and DELLY can not predict interspersed segmental duplications, therefore these tools miss such events. TARDIS consistently shows low FDR with comparable sensitivity. In our simulation, the length of each SV is generated uniformly random between 500 bp and 10 Kbp.

Furthermore, TARDIS can classify duplications into tandem, interspersed directed duplication and inverted duplication. However, DELLY and LUMPY are not designed to characterize interspersed segmental duplications, therefore we cannot provide comparisons. Table 4.2 shows the TDR, FDR, and the exact count of the number of True/False predictions for each type of segmental duplication.

4.2 Real Data Experiments

In addition to simulations, we also conducted experiments by utilizing real data and decided on using CHM1, CHM13 [5, 6] for haploid genomes and NA12878 for diploid genome. These datasets were sequenced using both the Illumina platform at higher depths and long reads generated with single molecule real time (SMRT) technology (i.e. PacBio). Our motivation behind this was to be able to cross-validate and compare our predictions with an orthogonal technology.

Table 4.2: Characterization of different types of segmental duplications using TARDIS on simulated data.

Duplication Type	Coverage	Total SVs	#Missed	#True	TPR	#True	FDR
Inverted Interspersed Duplication	10X	200	10	190	0.950	2	0.010
	20X	200	7	193	0.965	4	0.019
	30X	200	7	193	0.965	2	0.009
	60X	200	7	193	0.965	14	0.047
Direct Interspersed Duplication	10X	200	18	182	0.910	1	0.004
	20X	200	8	192	0.960	1	0.003
	30X	200	7	193	0.965	1	0.003
	60X	200	6	194	0.970	2	0.006
Tandem Duplication	10X	200	16	184	0.920	14	0.057
	20X	200	11	189	0.945	15	0.050
	30X	200	8	192	0.960	6	0.017
	60X	200	6	194	0.970	11	0.028

TARDIS can classify duplications into tandem, interspersed directed duplication and inverted duplication. However, DELLY and LUMPY are not designed to characterize these complex SVs. This table shows the true positive rate (recall) and false discovery rate (TPR and FDR respectively) of TARDIS for each type of duplication.

For haploid genome analysis, we downloaded short read HTS data generated from two haploid cell lines, namely CHM1 and CHM13. We mapped the reads to human reference genome (GRCh37) using BWA-MEM [116]. We also obtained call sets generated with PacBio data from the same genomes [183], which we use as the true call set to compare with our predictions. Similar steps were applied to the diploid NA12878 genome.

First, in Table 4.3, we show the properties of these datasets.

4.2.1 Deletions

First, we compared deletion predictions of TARDIS in CHM1, CHM13 and NA12878 genomes against the call sets generated by DELLY [167] and LUMPY [170]. We restricted the SV size to be >100 bp and required > 50% reciprocal overlap for two deletions to be considered the same using BEDTools [241]. We assume that the PacBio data sets [9] are the gold standard for our comparison as shown in Table 4.4. The results suggest that TARDIS employs the lowest

Table 4.3: Properties of the datasets we utilized in our eperiments.

Genome	Mean	Stdev	Read-length	Cov.	#Reads	#Concordant	#Discordant	#Libraries
CHM1	321	108	100 bp	42x	950,737,331	947,692,059	3,045,272	1
CHM13	356	119	150 bp	42x	655,538,698	646,895,477	8,643,221	23
NA12878	319	81	100 bp	54x	1,351,332,053	1,347,464,192	3,867,861	1

Properties of the genomes that we used in our experiments are given. We decide whether a read is concordant or discordant based on mean and standard deviation of the dataset. Each dataset has one or more libraries corresponding to the number of different lanes that the genome is sequenced. Thus each library has specific mean, standard deviation and read-length (They mostly have very close values, thus we give the average values for mean and standard deviation of CHM13 genome, which has 23 libraries). Finally, coverage (Cov.) refers to the average depth of reads. We should note that the accuracy and cost of sequencing increases parallel to the coverage.

FDR among the three tools with comparable sensitivity overall.

Table 4.4: Comparison of deletion accuracy between TARDIS, LUMPY and DELLY using CHM1, CHM13 and NA12878 data sets

	CHM1					CHM13					NA12878				
	#True	#False	Precision	Recall	F-score	#True	#False	Precision	Recall	F-score	#True	#False	Precision	Recall	F-score
TARDIS	1279	304	**0.81**	0.32	0.46	1595	655	0.71	0.40	**0.51**	2218	622	**0.78**	0.60	**0.68**
LUMPY	1518	1063	0.59	0.38	0.46	1401	457	**0.75**	0.35	0.48	2344	1456	0.62	0.63	0.62
DELLY	1651	1488	0.53	**0.41**	0.46	1794	1865	0.49	**0.45**	0.47	2536	2703	0.48	**0.69**	0.57

We compared deletion accuracy (>100 bp) of TARDIS, LUMPY and DELLY using CHM1, CHM13 and NA12878 data sets. Assuming that PacBio calls are the gold standard, we compared the result with [9] for CHM1 and CHM13, and with MT Sinai callset (data available in GIAB repository) for NA12878. The total number of calls for CHM1, CHM13 and NA12878 are 4016, 3946 and 3698 respectively. Precision is calculated as $\frac{TP}{TP+FP}$ and recall as $\frac{TP}{TP+FN}$, where TP = true positive, FP = false positive, FN = false negative. Finally, we calculated F-score with $2 \times \frac{Precision \times Recall}{Precision+Recall}$ formulation.

Additionally, we provide a comparison of TARDIS and LUMPY predictions for the diploid NA12878 genome given in Figure 4.1a and a size distribution histogram given in Figure 4.1b. The peaks at 300 bp and 5,900 bp corresponds to Alu and L1 mobile element deletions respectively.

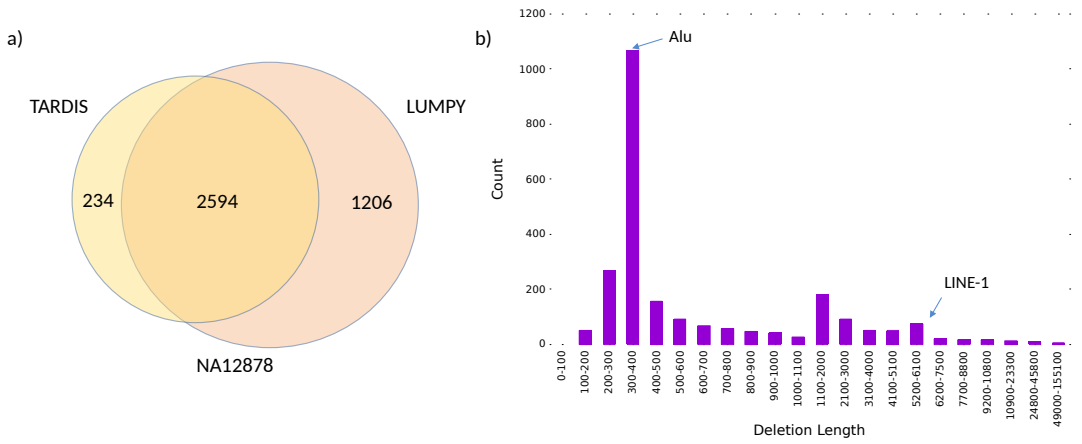


Figure 4.1: Comparison of deletion accuracy (>100 bp) between TARDIS and LUMPY using NA12878 genome (a). We also provide a deletion length histogram (b) exhibiting the expected peaks at 300 bp and 5,900 bp for ALU and L1 deletions

4.2.2 Inversions

For the case of inversions >100 bp, Figure 4.2(a) and (b) shows ROC curves by utilizing the predictions on CHM1 and CHM13 genomes. It is obvious that TARDIS achieves better area under the curve (AUC) statistics. Additionally, we evaluated the 50 highest scoring set of inversion predictions of each tool generated for CHM1 genome. Briefly, we used a reference-guided assembly of PacBio reads generated from the same genome [183] and mapped the contigs to the loci of interest (Figure 4.2) (c)). Compared to LUMPY and DELLY, TARDIS again achieves better AUC here, however, we note that the main reason for DELLY and LUMPY curves being closer to that of TARDIS for low number of false calls is because there were several predictions for which corresponding contigs did not exist in the assembled genome, therefore omitted from this plot.

Similar to CHM1/13 results, for diploid NA12878 genome, TARDIS outperformed state-of-the-art methods for inversion predictions given in Figure 4.3.

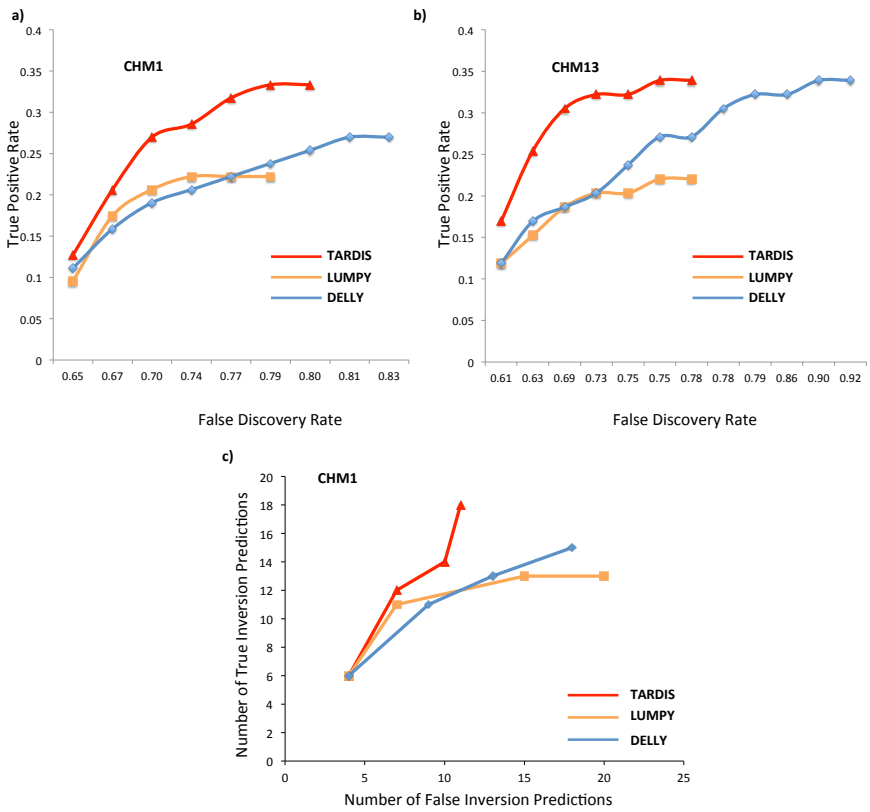


Figure 4.2: Receiver operator characteristic (ROC) curve for the comparison of inversion predictions on CHM1 and CHM13 datasets. Overall, TARDIS achieves better area under the curve (AUC) statistics compared to the other tools. (a), (b) comparison of CHM1 and CHM13 predicted inversions using PacBio reads based on BLASR mappings. (c) validation of top predicted inversion of different tools using local assembly of the PacBio reads of CHM1.

4.2.3 Duplications

In Table 4.5, we provide the full set of the 50 highest scoring segmental duplications that TARDIS predicts in CHM1 genome together with *in silico* validation using the corresponding PacBio-based assembly. Almost all of the predicted duplications, except one, were validated using long reads. Note that in most cases TARDIS assigned the correct subtype of duplications (inverted, direct or tandem duplication) to the prediction. It only gives one false call and three interspersed duplications that are wrongly assigned to tandem duplications. As expected, the highest number of segmental duplications in the top 50 were tandem duplications

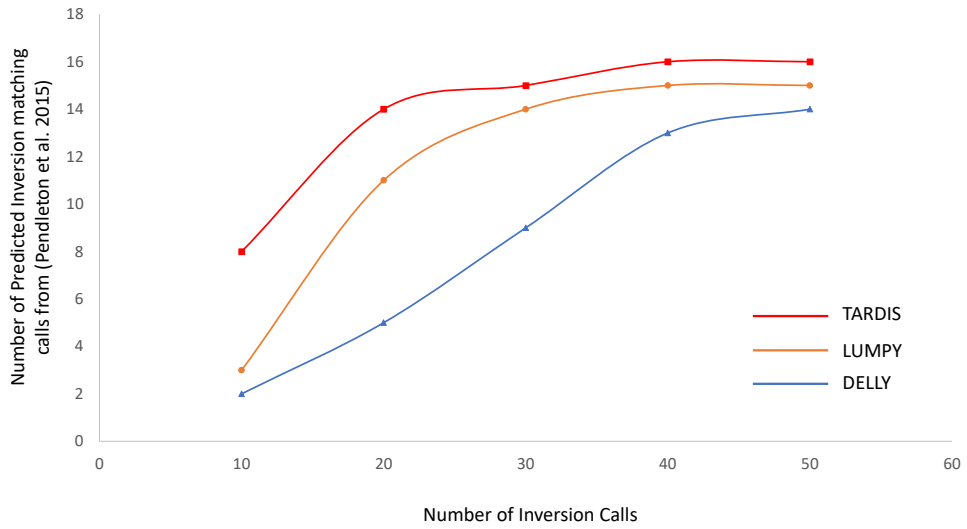


Figure 4.3: Comparison of top inversion prediction on NA12878 sample against predicted and validated set of inversion of the same samples using PacBio data from [8]

($> 50\%$ of all duplications).

We should also state that TARDIS discovered a large inverted duplication in NA12878 genome validated by PacBio data that is generated from the same sample shown in Figure 4.4. The interesting point about this inverted duplication is that it is larger than 10 Kbp and the distance between locus of insertion and the duplicated region is also larger, which shows a potential start of a new segmental duplication.

4.2.4 Insertions

To evaluate the mobile element insertion discovery performance of TARDIS, we utilized CHM1 and CHM13 genomes and compared our results with the orthogonal PacBio predictions given in Figure 4.5. We should emphasize that the additional calls discovered by TARDIS and missing in PacBio data might actually be real and simply false negatives in the PacBio predictions. When we compared

Table 4.5: 50 highest scoring segmental duplications predicted by TARDIS in the CHM1 genome.

Duplication Insertion Locus	TARDIS Dup. Type	TARDIS Score	Validation (PacBio)	Duplication Insertion Locus	TARDIS Dup. Type	TARDIS Score	Validation (PacBio)
chr11 63,698,518 - 63,702,043	Direct	0.000139	True	chr2 37,928,244 - 38,101,822	Tandem	0.000073	N/A
chr3 194,542,832 - 194,546,551	Direct	0.000147	True	chr20 60,032,847 - 60,033,402	Tandem	0.000118	True
chr5 143,512,368 - 143,515,435	Direct	0.000189	True	chr1 207,097,488 - 207,097,792	Tandem	0.000143	True
chr4 190,606,509 - 190,610,728	Direct	0.000356	True (Tandem)	chr5 3,323,854 - 3,324,308	Tandem	0.000150	N/A
chr20 2,359,601 - 2,360,962	Direct	0.000418	True	chr7 2,554,438 - 2,554,794	Tandem	0.000157	True
chr9 112,285,745 - 112,286,960	Direct	0.000422	True	chr12 110,099,331 - 110,099,745	Tandem	0.000164	True
chr19 4,511,103 - 4,511,949	Direct	0.000453	True (Tandem)	chr6 168,052,169 - 168,052,467	Tandem	0.000164	True
chr17 46,615,511 - 46,617,628	Direct	0.000466	True	chr16 86,008,690 - 86,009,146	Tandem	0.000174	True
chr18 69,711,699 - 69,713,216	Direct	0.000469	True	chr10 127,513,387 - 127,513,671	Tandem	0.000181	True
chr6 160,877,581 - 160,956,646	Direct	0.000484	N/A	chr14 106,049,119 - 106,049,358	Tandem	0.000181	True
chr2 10,825,652 - 10,827,218	Inverted	0.000118	True	chr17 80,317,606 - 80,318,018	Tandem	0.000181	N/A
chr3 43,834,994 - 43,836,299	Inverted	0.000123	True	chr20 62,720,019 - 62,720,214	Tandem	0.000181	True
chr2 125,051,481 - 125,053,239	Inverted	0.000127	True	chr9 132,158,786 - 132,159,087	Tandem	0.000181	N/A
chr14 67,169,917 - 67,171,999	Inverted	0.000146	True	chr10 132,974,718 - 132,975,317	Tandem	0.000190	True
chr2 72,440,066 - 72,441,647	Inverted	0.000159	True	chr12 13,164,410 - 13,164,785	Tandem	0.000190	True
chr10 127,190,469 - 127,197,324	Inverted	0.000190	True	chr8 2,215,816 - 2,216,235	Tandem	0.000201	N/A
chr9 107,816,536 - 107,817,623	Inverted	0.000200	True	chr6 44,012,337 - 44,012,939	Tandem	0.000211	True
chr17 36,350,020 - 36,407,396	Inverted	0.000208	False	chr9 34,681,543 - 34,681,898	Tandem	0.000266	True
chr12 71,532,693 - 71,534,000	Inverted	0.000318	True	chr6 35,754,611 - 35,766,730	Tandem	0.000273	True
chr1 114,645,854 - 114,654,623	Inverted	0.000334	True	chr20 59,567,846 - 59,590,250	Tandem	0.000287	True
chr18 11,508,829 - 11,511,479	Inverted	0.000353	True	chr20 62,123,611 - 62,124,191	Tandem	0.000355	True
chr5 115,346,294 - 115,351,084	Inverted	0.000390	True	chr18 77,831,328 - 77,831,783	Tandem	0.000369	N/A
chr7 31,586,823 - 31,590,394	Inverted	0.000437	True	chrX 417,957 - 418,352	Tandem	0.000369	True
chr19 15,785,635 - 15,888,539	Inverted	0.000485	True (Tandem)	chr20 42,325,185 - 42,325,572	Tandem	0.000399	True
				chr10 127,940,156 - 127,940,689	Tandem	0.000452	True
				chr3 197,117,149 - 197,117,806	Tandem	0.000463	N/A

Here we list the insertion locations of the top 50 scoring segmental duplications in CHM1 genome. All predictions are sorted by the SV score (lower is better). If the validation is N/A, that means the incorrect prediction from PacBio data, which will be skipped in the comparison. TARDIS only gives one false call and three interspersed duplications that are wrongly assigned to tandem duplications.

these calls with the polymorphic MEIs in dbRIP, we found out that over 30% of them are indeed correct calls. Moreover, our further analysis revealed that most of the MEI events discovered by PacBio that TARDIS missed were found within other repeats, which makes it very challenging to accurately map short reads.

For NUMT insertions, we assessed the performance of TARDIS using CHM1 and NA12878 genomes. It is known that most of the NUMT insertions occur within similar loci throughout the genomes. Dayama et al. [237] analyzed 999 individuals from 1000 HGP [18] and Human Genome Diversity Project (HGDP) [242] to create a map of these polymorphic insertions. In total, there are 256 NUMT insertion locations for these 999 individuals and TARDIS achieves very high precision; 100% (3 out of 3) of CHM1 and 87.5% (7 out of 8) of NA12878 NUMT insertions predicted by TARDIS hits one of those loci.

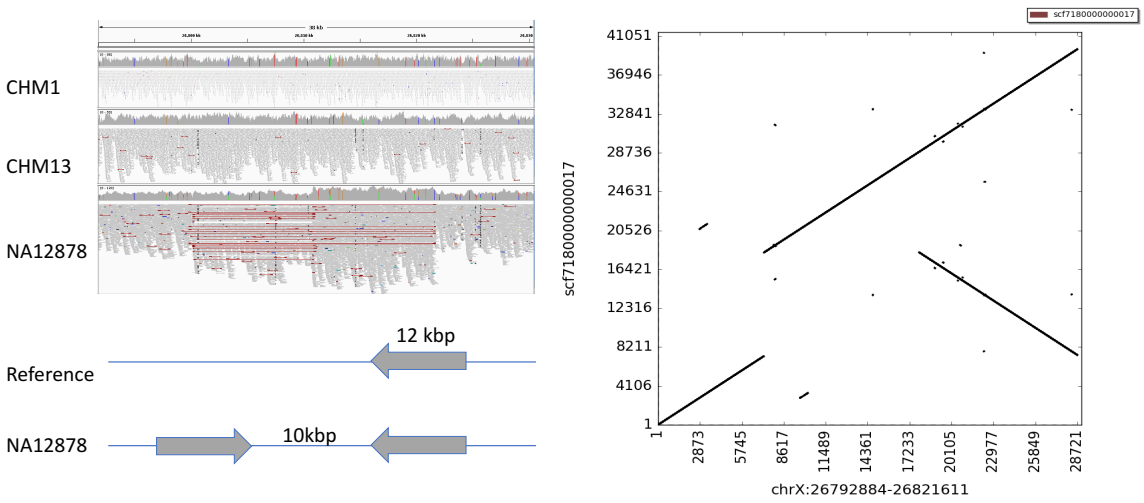


Figure 4.4: a) Illumina signature for an inverted duplication, b) PacBio validation.

We should state that novel sequence insertions are very difficult to detect with short-read sequencing technology. TARDIS is able to characterize these SVs, but because the span-size distribution of read-pairs is not uniform (mostly skewed to the left) and insert-size of read-pairs suggesting an insertion event is small, these read-pairs may also be classified as concordant. Therefore most of the read-pairs suggesting an insertion event are not considered for clustering. Thus, the number of insertion SVs that TARDIS discovers is relatively small. To overcome this limitation, *de novo* assembly such as Pamir [243] will be the genuine solution.

4.3 Sensitive Mode

In order to evaluate the performance of TARDIS in Sensitive Mode, we utilized the simulation dataset that we presented in Section 4.1. We also conducted real data experiments by using the haploid CHM1 genome that we used in Quick Mode experiments.

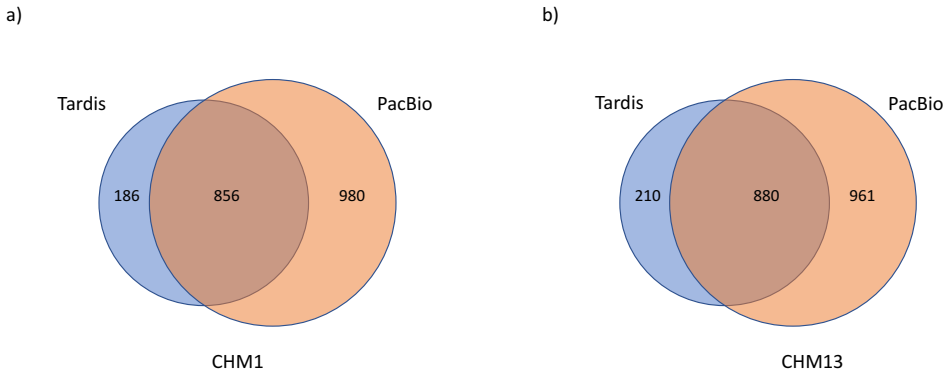


Figure 4.5: Alu insertion predictions in CHM1 and CHM13 datasets, compared against PacBio calls [9].

The importance of Sensitive Mode is that TARDIS searches for almost all potential mapping locations (we allow a read to map up to 500 distinct locations) of each read in the reference genome (Quick Mode utilizes a single mapping location; best or an arbitrary mapping). Since the number of ambiguous mappings naturally increases, we utilize a pruning step based on Phred Quality Score [244] that involves dividing the Phred Score of a read to sum of the Phred Scores of all reads and prune the ones <0.001. We should also note that Sensitive Mode does not have the split read mapping capability, it only uses read pair and read depth signatures. The SV types that this mode discovers are also limited to deletions, novel sequence insertions, inversions, tandem duplications and mobile element insertions.

Table 4.6: Comparison of simulation predictions for Sensitive and Quick Mode of TARDIS.

SV Type	Total SVs	Sensitive Mode			Quick Mode		
		#True	#Miss	#False	#True	#Miss	#False
Deletion	400	388	12	202	384	16	19
Inversion	200	198	2	1	199	1	1
Tandem Duplication	200	194	6	201	192	8	6

We show the number of true, false and missed predictions of TARDIS for deletions, inversions and tandem duplications in Sensitive and Quick Modes.

The results given in Table 4.6 show that both modes yield similar true discovery rates, nevertheless there is only a minor advantage of Sensitive Mode in

deletions and tandem duplications. However, the number of false predictions in deletions and tandem duplications is higher for the Sensitive Mode compared to the Quick Mode. The reason behind this is that the signatures for interspersed duplications, deletions and tandem duplications are the same. Moreover, Sensitive Mode does not utilize interspersed duplication clustering, so TARDIS in Sensitive Mode mispredicts some of the interspersed duplications as deletions and tandem duplications.

Finally, we compared Quick and Sensitive Modes using real data sets of CHM1 for deletion and inversion predictions as given in Table 4.7.

As opposed to the simulations, Quick mode achieves better recall and precision rate for deletion. On the other hand, Sensitive Mode exhibits better precision for inversion predictions.

Table 4.7: Comparison of real data (CHM1 genome) predictions for Sensitive and Quick Modes of TARDIS.

SV Type	Total SVs	Quick Mode		Sensitive Mode	
		Precision	Recall	Precision	Recall
Deletion	4016	**0.81**	**0.32**	0.73	0.24
Inversion	75	0.11	**0.27**	**0.24**	0.13

We present precision and recall rates of TARDIS for deletions and inversions in Sensitive and Quick Modes for CHM1 genome.

4.4 Linked-Reads

TARDIS is also able to use Linked-Read information by utilizing 10x Genomics data. In this section, we evaluate the performance of TARDIS in 10x mode by using simulation and real data sets.

For simulation, we used VarSim [245] to create 2,852,839 SNPs, 194,250 INDELs, 1,755 deletions, 2,245 insertions, 459 inversions, 584 tandem duplications and 150 direct, 110 inverted interspersed segmental duplications. Note that SV lengths are within 50 bp and 6 Mbp. Since VarSim does not generate interspersed duplications, we randomly changed a subset of simulated tandem duplications to

interspersed duplications in the simulated VCF file, a generic format for storing variant data [246], outputted by VarSim. We then generated Illumina WGS reads at 40X depth of coverage with ART [247] and 10xG Linked-Reads at 50X coverage with LRSim [248]. Finally, we mapped the reads to reference genome (GRCh37) using BWA-MEM [116] for WGS and Long Ranger [89] for Linked-Read data. The results for deletions and inversions are given in Table 4.8.

Table 4.8: Evaluation of Linked-Read performance of TARDIS.

	Deletion			Inversion		
	#True	#False	#Miss	#True	#False	#Miss
TARDIS-Sensitive-10x	1539	**369**	236	398	**7**	61
TARDIS-Quick	1589	1724	186	383	88	76
TARDIS-Quick-10x	1590	1802	185	383	80	76
DELLY	**1767**	8204	**8**	**458**	4382	**1**

In order to evaluate deletion and inversion performance of TARDIS with Linked-Reads, we compared TARDIS in Sensitive (TARDIS-Sensitive-10x) and in Quick (TARDIS-Quick-10x) modes by utilizing 10x functionality against DELLY and TARDIS without the 10x functionality (TARDIS-Quick). We omitted the results of TARDIS Sensitive because it yielded almost the same results with the 10x mode.

As shown in the table, we compared the performance of TARDIS using 10x data in both Sensitive and Quick Modes against DELLY (operates only with WGS data) and TARDIS without utilizing 10xG data. We should note that there was no improvement in TARDIS Sensitive with 10xG compared to TARDIS Sensitive without 10xG, therefore we omitted those results from the table.

In general, results suggest that there is only a small improvement with 10x in Quick Mode both for inversions and deletions, yet with increased number of false calls. As the Sensitive mode does not have the capability of characterizing interspersed duplications, we disabled that function for the Quick Mode in our experiment in order to make the comparison equally likely. This is the reason behind the high number of false calls that TARDIS presents in Quick Mode compared to the Sensitive Mode. Those false calls are characterized correctly as interspersed duplication when the interspersed duplication clustering is utilized. Also it is noteworthy that the number of false calls in Sensitive Mode is very low since TARDIS Sensitive Mode was able to classify these calls correctly.

Next, we assessed the performance of TARDIS with real data and used

NA12878 [199], NA24385 Ashkenazi son of 10x Genomics dataset that is available under the GIAB FTP repository and CHM1 genome generated with 10xG Linked-Reads. Additionally, in order to compare our results with true call sets, we utilized high confidence deletion call set generated by GIAB for NA24385, MT Sinai for NA12878 and PacBio callsets for CHM1 genome [183].

The number of true calls increased, by approximately 3%, 2% and 1% for NA24385, NA12878 and CHM1 respectively when 10x mode is utilized. However, our false prediction count also increased by $\sim$ 2% for NA24385 and $\sim$ 10% for NA12878. On the other hand, we observed no increase with CHM1 dataset.

4.5 Time and Memory Consumption

Finally, we compared time and memory consumption of TARDIS, LUMPY and DELLY (Table 4.9). We benchmarked each tool on the same dataset generated from the CHM1 genome, which possesses 42X coverage and mapped to the reference human genome (GRCh37). Results show that TARDIS is substantially faster than the other tools, however it requires much more memory when used with split-read mapping (TARDIS-SC). Further inspection revealed that much of the memory requirement was caused by interspersed duplication clustering.

We also benchmarked Sensitive Mode of TARDIS using 32 threads and observed that it requires much more memory and running time compared to Quick Mode. The reason behind this is the read mapping step performed with mrFAST prior to the SV discovery. SV discovery step only requires 1h 50m requiring 10GB of memory.

Note that the speed and memory requirements were calculated using the same computing server [1].

[1]Intel(R) Xeon(R) CPU E7- 4830 @ 2.13GHz : 4 CPUs * 8 cores each = 32cores total 512 GB RAM

Table 4.9: Performance comparison in terms of time and memory.

	CPU time	Peak memory usage(GB)
TARDIS-noSC	1h 40m	7 GB
TARDIS-SC	2h 28m	16 GB
LUMPY	8h 41m	7 GB
TARDIS-Sensitive	26h	77 GB
DELLY	32h 19m	0.3 GB

Comparison of performance for TARDIS, LUMPY and DELLY for SV discovery in CHM1 genome (42X depth of coverage). TARDIS-SC and TARDIS-noSC denotes TARDIS with split-read mapping enabled (default) and not enabled (–no-soft-clip parameter is invoked) respectively. TARDIS-Sensitive is the Sensitive Mode utilized in TARDIS that harbors read mapping with mrFast.

Chapter 5

Conclusion and Discussion

In this dissertation we introduce novel algorithms to structural variation discovery problem using high-throughput sequencing technology. For this reason, we developed a new tool called TARDIS. The novelty of our approach can be summarized as follows; (1) We integrate multiple sequence signatures including read-pair, read-depth and split-read to identify and cluster potential SV regions for various types of SVs (deletions, novel sequence insertions, inversions, tandem and interspersed segmental duplications, mobile element insertions and nuclear mitochondria insertions) under the assumption of maximum parsimony; (2) TARDIS is the first method to distinguish complex SV events including tandem, direct and inverted interspersed segmental duplications; (3) Using simulated and real data sets, we showed that TARDIS outperforms state-of-the art methods in terms of specificity and demonstrates comparable sensitivity for all types of SVs, and achieves considerably high true discovery rate for segmental duplications. (4) TARDIS is able to utilize Linked-Read data of 10x Genomics to overcome the limitations of short-read sequencing technology.

We compared our experimental results against DELLY and LUMPY algorithms. These approaches do not use read depth signature, thus their performance within deletion regions are lower (LUMPY has the option to utilize read depth information determined by an external tool and given it as an input). Similarly,

duplication performance of these tools are highly affected by the lack of read depth signature analysis and interspersed duplication characterization. Thus, these weaknesses not only increase their FDR for tandem duplications but they are also not able to distinguish duplications into subtypes. Additionally, TARDIS is able utilize transposon annotations to discover MEIs and inhibits false predictions within these regions. Similarly, we use repeat and gap annotations in order not to utilize reads mapping to these regions. Therefore, our FDR is significantly lower compared to these tools as they totally dismiss these information. As a result, although the number of true calls predicted by TARDIS is relatively low compared to DELLY and LUMPY, our FDR is significantly lower than these tools in general and we have the highest F-score in general. We should also note that increasing the number of true calls is always possible within our framework by utilizing multiple mapping information of a read, however, our concerns are based on increased FDR. Thus, TARDIS achieves highest precision overall and comparable recall with these algorithms and that's what we expect to see.

Based on the simulation results, TARDIS achieved lowest FDR overall against LUMPY and DELLY with similar sensitivity for each type of SV. Additionally, TARDIS is the only method to classify duplications as tandem, interspersed direct and interspersed inverted with >90% true discovery rate for various depth of coverages. With real data, TARDIS still possessed lowest FDR among the other tools for deletion predictions in general with the highest F-score. It also achieved highest AUC statistics for inversion in CHM1/13 genomes. For the case of duplications, we compared our predictions against the orthogonal PacBio callsets since no other method is able to distinguish duplications as tandem or interspersed segmental duplication in direct and inverted orientations. We observed that in most cases TARDIS assigned the correct subtype of duplications to the prediction. For mobile element insertions, we were able to verify most of our predictions with PacBio call set and the rest with dbRIP. We also discerned that the MEIs discovered by PacBio and we miss are within other repeats which makes it almost impossible to detect with short-reads. Additionally, we discovered NUMT insertions in CHM1 and NA12878 genomes with high precision. Lastly, for novel sequence insertions, we saw that short-read sequencing seems impractical and *de*

novo assembly is required since most of the read-pairs suggesting an insertion event is classified as concordant based on the distribution of read pairs. However, TARDIS still detects some novel insertions accurately.

We also assessed the performance of Sensitive Mode. The difficulty here is in handling excess amount of mappings where most of them are ambiguous. In spite of this, TARDIS still achieves high precision although our recall is lower compared to Quick Mode for real data experiments. Simulation results are more promising, possessing relatively higher TDR than the Quick mode for deletions and inversions. On the other hand, deletions and inversions suffer from high false predictions, which is caused by the lack of segmental duplication clustering in Sensitive Mode.

We have an option to utilize Linked-Reads in both Sensitive and Quick Modes. For simulations on Linked-Reads, Sensitive Mode showed no improvement whereas Quick Mode has minor advantage. On the other hand, Quick Mode showed better improvement in terms of true discovery rate when we ran it with 10xG for real data experiments. However it still needs enhancements to decrease the false prediction count and increase recall.

Finally, we performed a comparison in terms of run time and memory consumption. TARDIS Quick Mode was substantially faster than the other tools, however it required much more memory. With further analysis, we saw that this consumption is due to split read mapping and interspersed segmental duplication clustering that need further improvements.

5.1 Future Work

We developed TARDIS as a structural variation discovery tool that is able to characterize most of the existing SV types including the complex ones. However there are still potential improvement opportunities.

First, due to the nature of short-read sequencing technology, detection of specific SV types is very challenging including novel sequence insertions and large inversions. Integrating *de novo* assembly to TARDIS will increase our discovery rate for novel sequence insertions. Additionally, local *de novo* assembly signature will help it achieve better accuracy. For the case of inversions, TARDIS achieves better accuracy among the other algorithms however, false prediction rates are still too high compared to other types of SVs. Presence of large inversions inside the genomes is one of the reasons for this. In order to solve this problem, our linked-read approach can be enhanced to characterize such events and this will likely increase our accuracy for inversions.

Second of all, although simulation experiments demonstrated potential efficacy of TARDIS in segmental duplication predictions, those that are generated from real genomes need to be experimentally verified to fully understand the power and shortcomings of our algorithms. We can then apply TARDIS to thousands of genomes that were already sequenced as part of various projects, such as the 1000 Genomes Project to advance our understanding of the SV spectrum in human genomes.

Third, multiple mapping strategies of TARDIS that include Sensitive Mode or by utilization of XA tag in BWA need improvements. A way to solve the ambiguous mappings will boost our recall and precision. In addition to this, inter-chromosome mobile element insertion, nuclear mitochondria insertion and most importantly clustering interspersed segmental duplications need to be integrated to our Sensitive Mode approach.

Fourth, algorithms to detect somatic structural variation discovery can be developed and integrated to TARDIS. Given a structural variation, we want to know if that SV is a somatic variant (i.e., it appears in tumor tissue and not in normal tissue) using variant allele frequency (VAF). We can then apply this to cancer genomes.

Finally, third generation sequencing technologies including PacBio and Oxford Nanopore that have larger read lengths and higher accuracy can to be utilized

in our framework. Algorithms that use a combination of short-read technology that have high coverage and long-reads data can be developed. This type of formulation will allow us to predict broader range of SV types with much higher accuracy.

Bibliography

[1] Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed 06-2018.

[2] E. Tuzun, A. J. Sharp, J. A. Bailey, R. Kaul, V. A. Morrison, L. M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, M. V. Olson, and E. E. Eichler, "Fine-scale structural variation of the human genome," *Nat Genet*, vol. 37, pp. 727–732, Jul 2005.

[3] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov, "Integrative genomics viewer," *Nature biotechnology*, vol. 29, no. 1, p. 24, 2011.

[4] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, "Integrative genomics viewer (igv): high-performance genomics data visualization and exploration," *Briefings in bioinformatics*, vol. 14, no. 2, pp. 178–192, 2013.

[5] K. M. Steinberg, V. A. Schneider, T. A. Graves-Lindsay, R. S. Fulton, R. Agarwala, J. Huddleston, S. A. Shiryev, A. Morgulis, U. Surti, W. C. Warren, D. M. Church, E. E. Eichler, and R. K. Wilson, "Single haplotype assembly of the human genome from a hydatidiform mole," *Genome Res*, vol. 24, pp. 2066–2076, Dec 2014.

[6] J. Huddleston, S. Ranade, M. Malig, F. Antonacci, M. Chaisson, L. Hon, P. H. Sudmant, T. A. Graves, C. Alkan, M. Y. Dennis, R. K. Wilson, S. W. Turner, J. Korlach, and E. E. Eichler, "Reconstructing complex regions

of genomes using long-read sequencing technology," *Genome Res*, vol. 24, pp. 688–696, Apr 2014.

[7] P. H. Sudmant, S. Mallick, B. J. Nelson, F. Hormozdiari, N. Krumm, J. Huddleston, B. P. Coe, C. Baker, S. Nordenfelt, M. Bamshad, L. B. Jorde, O. L. Posukh, H. Sahakyan, W. S. Watkins, L. Yepiskoposyan, M. S. Abdullah, C. M. Bravi, C. Capelli, T. Hervig, J. T. S. Wee, C. Tyler-Smith, G. van Driem, I. G. Romero, A. R. Jha, S. Karachanak-Yankova, D. Toncheva, D. Comas, B. Henn, T. Kivisild, A. Ruiz-Linares, A. Sajantila, E. Metspalu, J. Parik, R. Villems, E. B. Starikovskaya, G. Ayodo, C. M. Beall, A. Di Rienzo, M. F. Hammer, R. Khusainova, E. Khusnutdinova, W. Klitz, C. Winkler, D. Labuda, M. Metspalu, S. A. Tishkoff, S. Dryomov, R. Sukernik, N. Patterson, D. Reich, and E. E. Eichler, "Global diversity, population stratification, and selection of human copy-number variation," *Science*, vol. 349, no. 6253, 2015.

[8] M. Pendleton, R. Sebra, A. W. C. Pang, A. Ummat, O. Franzen, T. Rausch, A. M. Stütz, W. Stedman, T. Anantharaman, A. Hastie, H. Dai, M. H.-Y. Fritz, H. Cao, A. Cohain, G. Deikus, R. E. Durrett, S. C. Blanchard, R. Altman, C.-S. Chin, Y. Guo, E. E. Paxinos, J. O. Korbel, R. B. Darnell, W. R. McCombie, P.-Y. Kwok, C. E. Mason, E. E. Schadt, and A. Bashir, "Assembly and diploid architecture of an individual human genome via single-molecule technologies," *Nature methods*, vol. 12, pp. 780–786, Aug. 2015.

[9] J. Huddleston and E. E. Eichler, "An incomplete understanding of human genetic variation," *Genetics*, vol. 202, pp. 1251–1254, Apr 2016.

[10] J. D. Watson and F. H. Crick, "The structure of dna," in *Cold Spring Harbor symposia on quantitative biology*, vol. 18, pp. 123–131, Cold Spring Harbor Laboratory Press, 1953.

[11] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, Feb 2001.

[12] I. H. G. S. Consortium *et al.*, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, p. 931, 2004.

[13] I. H. Consortium *et al.*, "The international hapmap project," *Nature*, vol. 426, no. 6968, p. 789, 2003.

[14] The 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, Oct 2010.

[15] S. C. Schuster, W. Miller, A. Ratan, L. P. Tomsho, B. Giardine, L. R. Kasson, R. S. Harris, D. C. Petersen, F. Zhao, J. Qi, C. Alkan, J. M. Kidd, Y. Sun, D. I. Drautz, P. Bouffard, D. M. Muzny, J. G. Reid, L. V. Nazareth, Q. Wang, R. Burhans, C. Riemer, N. E. Wittekindt, P. Moorjani, E. A. Tindall, C. G. Danko, W. S. Teo, A. M. Buboltz, Z. Zhang, Q. Ma, A. Oosthuysen, A. W. Steenkamp, H. Oostuisen, P. Venter, J. Gajewski, Y. Zhang, B. F. Pugh, K. D. Makova, A. Nekrutenko, E. R. Mardis, N. Patterson, T. H. Pringle, F. Chiaromonte, J. C. Mullikin, E. E. Eichler, R. C. Hardison, R. A. Gibbs, T. T. Harkins, and V. M. Hayes, "Complete Khoisan and Bantu genomes from southern Africa," *Nature*, vol. 463, pp. 943–947, Feb 2010.

[16] D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. F. Johnson, T. Maricic, J. M. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E. E. Eichler, M. Stoneking, M. Richards, S. Talamo, M. V. Shunkov, A. P. Derevianko, J.-J. Hublin, J. Kelso, M. Slatkin, and S. Pääbo, "Genetic history of an archaic hominin group from Denisova Cave in Siberia," *Nature*, vol. 468, pp. 1053–1060, Dec 2010.

[17] R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna,

P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Pääbo, "A draft sequence of the Neandertal genome," *Science*, vol. 328, pp. 710–722, May 2010.

[18] The 1000 Genomes Project Consortium, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, pp. 56–65, Nov 2012.

[19] C. Alkan, P. Kavak, M. Somel, O. Gokcumen, S. Ugurlu, C. Saygi, E. Dal, K. Bugra, T. Güngör, S. C. Sahinalp, N. Ozören, and C. Bekpen, "Whole genome sequencing of turkish genomes reveals functional private alleles and impact of genetic interactions with Europe, Asia and Africa," *BMC Genomics*, vol. 15, no. 1, p. 963, 2014.

[20] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. Hsi-Yang Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. Paolo Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. Jasmine Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, . G. P. C. , R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, and J. O. Korbel, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, pp. 75–81, Sep 2015.

[21] The 1000 Genomes Project Consortium, "A global reference for human genetic variation," *Nature*, vol. 526, pp. 68–74, Sep 2015.

[22] W.-H. Li and L. A. Sadler, "Low nucleotide diversity in man.," *Genetics*, vol. 129, no. 2, pp. 513–523, 1991.

[23] I. S. M. W. Group *et al.*, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature*, vol. 409, no. 6822, p. 928, 2001.

[24] E. J. Hollox, U. Huffmeier, P. L. Zeeuwen, R. Palla, J. Lascorz, D. Rodijk-Olthuis, P. C. Van De Kerkhof, H. Traupe, G. De Jongh, M. Den Heijer, *et al.*, "Psoriasis is associated with increased β-defensin genomic copy number," *Nature genetics*, vol. 40, no. 1, p. 23, 2008.

[25] K. Fellermann, D. E. Stange, E. Schaeffeler, H. Schmalzl, J. Wehkamp, C. L. Bevins, W. Reinisch, A. Teml, M. Schwab, P. Lichter, *et al.*, "A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to crohn disease of the colon," *The American Journal of Human Genetics*, vol. 79, no. 3, pp. 439–448, 2006.

[26] H. C. Mefford, S. Clauin, A. J. Sharp, R. S. Moller, R. Ullmann, R. Kapur, D. Pinkel, G. M. Cooper, M. Ventura, H. H. Ropers, *et al.*, "Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy," *The American Journal of Human Genetics*, vol. 81, no. 5, pp. 1057–1069, 2007.

[27] E. Gonzalez, H. Kulkarni, H. Bolivar, A. Mangano, R. Sanchez, G. Catano, R. J. Nibbs, B. I. Freedman, M. P. Quinones, M. J. Bamshad, *et al.*, "The influence of ccl3l1 gene-containing segmental duplications on hiv-1/aids susceptibility," *Science*, vol. 307, no. 5714, pp. 1434–1440, 2005.

[28] I. Helbig, H. C. Mefford, A. J. Sharp, M. Guipponi, M. Fichera, A. Franke, H. Muhle, C. de Kovel, C. Baker, S. von Spiczak, K. L. Kron, I. Steinich, A. A. Kleefuss-Lie, C. Leu, V. Gaus, B. Schmitz, K. M. Klein, P. S. Reif, F. Rosenow, Y. Weber, H. Lerche, F. Zimprich, L. Urak, K. Fuchs, M. Feucht, P. Genton, P. Thomas, F. Visscher, G.-J. de Haan, R. S.

Møller, H. Hjalgrim, D. Luciano, M. Wittig, M. Nothnagel, C. E. Elger, P. Nürnberg, C. Romano, A. Malafosse, B. P. C. Koeleman, D. Lindhout, U. Stephani, S. Schreiber, E. E. Eichler, and T. Sander, "15q13.3 microdeletions increase risk of idiopathic generalized epilepsy," *Nat Genet*, vol. 41, pp. 160–162, Feb 2009.

[29] E. E. Eichler and A. W. Zimmerman, "A hot spot of genetic instability in autism," *N Engl J Med*, vol. 358, pp. 737–739, Feb 2008.

[30] A. Hoischen, N. Krumm, and E. E. Eichler, "Prioritization of neurodevelopmental disease genes by discovery of new mutations," *Nature neuroscience*, vol. 17, no. 6, p. 764, 2014.

[31] H. Nagasaki, T. Mochizuki, Y. Kodama, S. Saruhashi, S. Morizaki, H. Sugawara, H. Ohyanagi, N. Kurata, K. Okubo, T. Takagi, *et al.*, "Ddbj read annotation pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data," *DNA research*, vol. 20, no. 4, pp. 383–390, 2013.

[32] L. C. Francioli, A. Menelaou, S. L. Pulit, F. Van Dijk, P. F. Palamara, C. C. Elbers, P. B. Neerincx, K. Ye, V. Guryev, W. P. Kloosterman, *et al.*, "Whole-genome sequence variation, population structure and demographic history of the dutch population," *Nature genetics*, vol. 46, no. 8, p. 818, 2014.

[33] D. F. Gudbjartsson, P. Sulem, H. Helgason, A. Gylfason, S. A. Gudjonsson, F. Zink, A. Oddson, G. Magnusson, B. V. Halldorsson, E. Hjartarson, *et al.*, "Sequence variants from whole genome sequencing a large group of icelanders," *Scientific data*, vol. 2, p. 150011, 2015.

[34] U. Consortium *et al.*, "The uk10k project identifies rare variants in health and disease," *Nature*, vol. 526, no. 7571, p. 82, 2015.

[35] I. H. Consortium *et al.*, "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, p. 1299, 2005.

[36] I. H. Consortium *et al.*, "A second generation human haplotype map of over 3.1 million snps," *Nature*, vol. 449, no. 7164, p. 851, 2007.

[37] R. E. Mills, C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui, W. S. Pittard, and S. E. Devine, "An initial map of insertion and deletion (INDEL) variation in the human genome," *Genome Res*, vol. 16, pp. 1182–1190, Sep 2006.

[38] F. S. Collins, M. L. Drumm, J. L. Cole, W. K. Lockwood, G. V. Woude, and M. C. Iannuzzi, "Construction of a general human chromosome jumping library, with application to cystic fibrosis," *Science*, vol. 235, no. 4792, pp. 1046–1049, 1987.

[39] S. T. Warren, F. Zhang, G. R. Licameli, and J. F. Peters, "The fragile x site in somatic cell hybrids: an approach for molecular cloning of fragile sites," *Science*, vol. 237, no. 4813, pp. 420–423, 1987.

[40] E. Karakoc, C. Alkan, B. J. O'Roak, M. Y. Dennis, L. Vives, K. Mark, M. J. Rieder, D. A. Nickerson, and E. E. Eichler, "Detection of structural variants and indels within exome data," *Nat Methods*, vol. 9, no. 2, pp. 176–178, 2012.

[41] G. Narzisi, J. A. O'rawe, I. Iossifov, H. Fang, Y.-h. Lee, Z. Wang, Y. Wu, G. J. Lyon, M. Wigler, and M. C. Schatz, "Accurate de novo and transmitted indel detection in exome-capture data using microassembly," *Nature methods*, vol. 11, no. 10, p. 1033, 2014.

[42] T. Willems, M. Gymrek, G. Highnam, D. Mittelman, Y. Erlich, . G. P. Consortium, *et al.*, "The landscape of human str variation," *Genome research*, vol. 24, no. 11, pp. 1894–1904, 2014.

[43] C. E. Pearson, K. N. Edamura, and J. D. Cleary, "Repeat instability: mechanisms of dynamic mutations," *Nature Reviews Genetics*, vol. 6, no. 10, p. 729, 2005.

[44] M. Gymrek, D. Golan, S. Rosset, and Y. Erlich, "lobSTR: A short tandem repeat profiler for personal genomes," *Genome research*, vol. 22, pp. 1154–1162, June 2012.

[45] T. Willems, D. Zielinski, J. Yuan, A. Gordon, M. Gymrek, and Y. Erlich, "Genome-wide profiling of heritable and de novo str variations," *Nature methods*, vol. 14, no. 6, p. 590, 2017.

[46] 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," *Nat Genet*, vol. 36, pp. 949–951, Sep 2004.

[47] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Månér, 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*, vol. 305, pp. 525–528, Jul 2004.

[48] A. J. Sharp, D. P. Locke, S. D. McGrath, Z. Cheng, J. A. Bailey, R. U. Vallente, L. M. Pertz, R. A. Clark, S. Schwartz, R. Segraves, V. V. Oseroff, D. G. Albertson, D. Pinkel, and E. E. Eichler, "Segmental duplications and copy-number variation in the human genome," *Am J Hum Genet*, vol. 77, pp. 78–88, Jul 2005.

[49] 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. González, M. Gratacòs, 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 2006.

[50] K. K. Wong, N. S. Dosanjh, L. R. Kimm, Z. Cheng, D. E. Horsman, C. MacAulay, R. T. Ng, C. J. Brown, E. E. Eichler, W. L. Lam, *et al.*, "A comprehensive analysis of common copy-number variations in the human genome," *The American Journal of Human Genetics*, vol. 80, no. 1, pp. 91–104, 2007.

[51] C. Alkan, B. P. Coe, and E. E. Eichler, "Genome structural variation discovery and genotyping," *Nat Rev Genet*, vol. 12, pp. 363–376, May 2011.

[52] P. Stankiewicz and J. R. Lupski, "Genome architecture, rearrangements and genomic disorders," *TRENDS in Genetics*, vol. 18, no. 2, pp. 74–82, 2002.

[53] B. B. de Vries, R. Pfundt, M. Leisink, D. A. Koolen, L. E. Vissers, I. M. Janssen, S. van Reijmersdal, W. M. Nillesen, E. H. Huys, N. de Leeuw, *et al.*, "Diagnostic genome profiling in mental retardation," *The American Journal of Human Genetics*, vol. 77, no. 4, pp. 606–616, 2005.

[54] 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. L. Knight, and E. E. Eichler, "Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome," *Nat Genet*, vol. 38, pp. 1038–1042, Sep 2006.

[55] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, *et al.*, "Strong association of de novo copy number mutations with autism," *Science*, vol. 316, no. 5823, pp. 445–449, 2007.

[56] D. Pinto, A. T. Pagnamenta, L. Klei, R. Anney, D. Merico, R. Regan, J. Conroy, T. R. Magalhaes, C. Correia, B. S. Abrahams, *et al.*, "Functional impact of global rare copy number variation in autism spectrum disorders," *Nature*, vol. 466, no. 7304, p. 368, 2010.

[57] D. Malhotra and J. Sebat, "Cnvs: harbingers of a rare variant revolution in psychiatric genetics," *Cell*, vol. 148, no. 6, pp. 1223–1241, 2012.

[58] C. G. A. R. Network *et al.*, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, p. 1061, 2008.

[59] F. Mitelman, B. Johansson, and F. Mertens, "The impact of translocations and gene fusions on cancer causation," *Nature Reviews Cancer*, vol. 7, no. 4, p. 233, 2007.

[60] D. Locke, R. Segraves, R. Nicholls, S. Schwartz, D. Pinkel, D. Albertson, and E. Eichler, "Bac microarray analysis of 15q11–q13 rearrangements and the impact of segmental duplications," *Journal of medical genetics*, vol. 41, no. 3, pp. 175–182, 2004.

[61] A. Itsara, G. M. Cooper, C. Baker, S. Girirajan, J. Li, D. Absher, R. M. Krauss, R. M. Myers, P. M. Ridker, D. I. Chasman, *et al.*, "Population analysis of large copy number variants and hotspots of human genetic disease," *The American Journal of Human Genetics*, vol. 84, no. 2, pp. 148–161, 2009.

[62] A. J. Sharp, Z. Cheng, and E. E. Eichler, "Structural variation of the human genome," *Annu Rev Genomics Hum Genet*, vol. 7, pp. 407–442, 2006.

[63] C. A. Heid, J. Stevens, K. J. Livak, and P. M. Williams, "Real time quantitative pcr.," *Genome research*, vol. 6, no. 10, pp. 986–994, 1996.

[64] I. Bieche, M. Olivi, M.-H. Champème, D. Vidaud, R. Lidereau, and M. Vidaud, "Novel approach to quantitative polymerase chain reaction using real-time detection: application to the detection of gene amplification in breast cancer," *International Journal of Cancer*, vol. 78, no. 5, pp. 661–666, 1998.

[65] A. M. Maxam and W. Gilbert, "A new method for sequencing dna," *Proc Natl Acad Sci U S A*, vol. 74, pp. 560–564, Feb 1977.

[66] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. H. III, P. M. Slocombe, and M. Smith, "Nucleotide sequence of bacteriophage x174 dna," *Nature*, vol. 265, pp. 687 – 695, February 1977.

[67] J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler, "Recent segmental duplications in the human genome," *Science*, vol. 297, pp. 1003–1007, Aug 2002.

[68] M. L. Metzker, "Sequencing technologies—the next generation," *Nature reviews genetics*, vol. 11, no. 1, p. 31, 2010.

[69] T. D. Harris, P. R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. DiMeo, J. W. Efcavitch, *et al.*, "Single-molecule dna sequencing of a viral genome," *Science*, vol. 320, no. 5872, pp. 106–109, 2008.

[70] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, *et al.*, "Real-time dna sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009.

[71] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, *et al.*, "The potential and challenges of nanopore sequencing," *Nature biotechnology*, vol. 26, no. 10, p. 1146, 2008.

[72] H. Buermans and J. Den Dunnen, "Next generation sequencing technology: advances and applications," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1842, no. 10, pp. 1932–1941, 2014.

[73] C. Bleidorn, "Third generation sequencing: technology and its potential impact on evolutionary biodiversity research," *Systematics and biodiversity*, vol. 14, no. 1, pp. 1–8, 2016.

[74] H. Lee, J. Gurtowski, S. Yoo, M. Nattestad, S. Marcus, S. Goodwin, W. R. McCombie, and M. Schatz, "Third-generation sequencing and the future of genomics," *BioRxiv*, p. 048603, 2016.

[75] D. Senol Cali, J. S. Kim, S. Ghose, C. Alkan, and O. Mutlu, "Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions," *Briefings in bioinformatics*, 2018.

[76] 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, no. 1, p. 341, 2012.

[77] M. Miyamoto, D. Motooka, K. Gotoh, T. Imai, K. Yoshitake, N. Goto, T. Iida, T. Yasunaga, T. Horii, K. Arakawa, *et al.*, "Performance comparison of second-and third-generation sequencers using a bacterial genome with two chromosomes," *BMC genomics*, vol. 15, no. 1, p. 699, 2014.

[78] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, pp. 333–351, May 2016.

[79] C.-S. Chin, P. Peluso, F. J. Sedlazeck, M. Nattestad, G. T. Concepcion, A. Clum, C. Dunn, R. O'Malley, R. Figueroa-Balderas, A. Morales-Cruz, *et al.*, "Phased diploid genome assembly with single-molecule real-time sequencing," *Nature methods*, vol. 13, no. 12, p. 1050, 2016.

[80] C.-S. Chin, D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, *et al.*, "Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data," *Nature methods*, vol. 10, no. 6, p. 563, 2013.

[81] K. Berlin, S. Koren, C.-S. Chin, J. P. Drake, J. M. Landolin, and A. M. Phillippy, "Assembling large genomes with single-molecule sequencing and locality-sensitive hashing," *Nature biotechnology*, vol. 33, no. 6, p. 623, 2015.

[82] A. C. English, W. J. Salerno, and J. G. Reid, "Pbhoney: identifying genomic variants via long-read discordance and interrupted mapping," *BMC bioinformatics*, vol. 15, no. 1, p. 180, 2014.

[83] J. Huddleston, M. J. Chaisson, K. M. Steinberg, W. Warren, K. Hoekzema, D. Gordon, T. A. Graves-Lindsay, K. M. Munson, Z. N. Kronenberg, L. Vives, *et al.*, "Discovery and genotyping of structural variation from long-read haploid genome sequence data," *Genome research*, vol. 27, no. 5, pp. 677–685, 2017.

[84] F. J. Sedlazeck, P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. von Haeseler, and M. C. Schatz, "Accurate detection of complex structural variations using single-molecule sequencing," *Nature Methods*, vol. 15, no. 6, pp. 461–468, 2018.

[85] X. Fan, M. Chaisson, L. Nakhleh, and K. Chen, "Hysa: a hybrid structural variant assembly approach using next-generation and single-molecule sequencing technologies," *Genome research*, vol. 27, no. 5, pp. 793–800, 2017.

[86] G. X. Zheng, B. T. Lau, M. Schnall-Levin, M. Jarosz, J. M. Bell, C. M. Hindson, S. Kyriazopoulou-Panagiotopoulou, D. A. Masquelier, L. Merrill, J. M. Terry, *et al.*, "Haplotyping germline and cancer genomes with high-throughput linked-read sequencing," *Nature biotechnology*, vol. 34, no. 3, p. 303, 2016.

[87] Y. Mostovoy, M. Levy-Sakin, J. Lam, E. T. Lam, A. R. Hastie, P. Marks, J. Lee, C. Chu, C. Lin, Ž. Džakula, H. Cao, S. A. Schlebusch, K. Giorda, M. Schnall-Levin, J. D. Wall, and P.-Y. Kwok, "A hybrid approach for de novo human genome sequence assembly and phasing," *Nat Methods*, May 2016.

[88] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, p. 333, 2016.

[89] P. Marks, S. Garcia, A. M. Barrio, K. Belhocine, J. Bernate, R. Bharadwaj, K. Bjornson, C. Catalanotti, J. Delaney, A. Fehr, *et al.*, "Resolving the full spectrum of human genome variation using linked-reads," *BioRxiv*, p. 230946, 2017.

[90] N. Spies, Z. Weng, A. Bishara, J. McDaniel, D. Catoe, J. M. Zook, M. Salit, R. B. West, S. Batzoglou, and A. Sidow, "Genome-wide reconstruction of complex structural variants using read clouds.," *Nature methods*, vol. 14, pp. 915–920, Sept. 2017.

[91] L. C. Xia, J. M. Bell, C. Wood-Bouwens, J. J. Chen, N. R. Zhang, and H. P. Ji, "Identification of large rearrangements in cancer genomes with barcode linked reads.," *Nucleic acids research*, Nov. 2017.

[92] M. Eslami Rasekh, G. Chiatante, M. Miroballo, J. Tang, M. Ventura, C. T. Amemiya, E. E. Eichler, F. Antonacci, and C. Alkan, "Discovery of large

genomic inversions using long range information," *BMC genomics*, vol. 18, p. 65, Jan. 2017.

[93] T. J. Treangen and S. L. Salzberg, "Repetitive dna and next-generation sequencing: computational challenges and solutions," *Nature Reviews Genetics*, vol. 13, no. 1, p. 36, 2012.

[94] C. Firtina and C. Alkan, "On genomic repeats and reproducibility," *Bioinformatics*, vol. 32, no. 15, pp. 2243–2247, 2016.

[95] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol*, vol. 10, p. R25, Mar 2009.

[96] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome Res*, vol. 18, pp. 1851–1858, Nov 2008.

[97] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J Mol Biol*, vol. 147, pp. 195–197, Mar 1981.

[98] C. Alkan, J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J. O. Kitzman, C. Baker, M. Malig, O. Mutlu, S. C. Sahinalp, R. A. Gibbs, and E. E. Eichler, "Personalized copy number and segmental duplication maps using next-generation sequencing," *Nat Genet*, vol. 41, pp. 1061–1067, Oct 2009.

[99] N. Homer, B. Merriman, and S. F. Nelson, "Bfast: an alignment tool for large scale genome resequencing," *PloS one*, vol. 4, no. 11, p. e7767, 2009.

[100] F. Hach, F. Hormozdiari, C. Alkan, F. Hormozdiari, I. Birol, E. E. Eichler, and S. C. Sahinalp, "mrsfast: a cache-oblivious algorithm for short-read mapping," *Nature methods*, vol. 7, no. 8, p. 576, 2010.

[101] M. David, M. Dzamba, D. Lister, L. Ilie, and M. Brudno, "Shrimp2: sensitive yet practical short read mapping," *Bioinformatics*, vol. 27, no. 7, pp. 1011–1012, 2011.

[102] H. Xin, D. Lee, F. Hormozdiari, S. Yedkar, O. Mutlu, and C. Alkan, "Accelerating read mapping with FastHASH," *BMC Genomics*, vol. 14 Suppl 1, p. S13, 2013.

[103] NovoAlign. Available at: http://www.novocraft.com/. Accessed 06-2018.

[104] M. Li, B. Ma, D. Kisman, and J. Tromp, "Patternhunter ii: Highly sensitive and fast homology search," *Journal of bioinformatics and computational biology*, vol. 2, no. 03, pp. 417–439, 2004.

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

[106] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature methods*, vol. 9, pp. 357–359, Mar. 2012.

[107] H. Cheng, H. Jiang, J. Yang, Y. Xu, and Y. Shang, "Bitmapper: an efficient all-mapper based on bit-vector computing," *BMC bioinformatics*, vol. 16, no. 1, p. 192, 2015.

[108] H. Xin, D. Lee, F. Hormozdiari, S. Yedkar, O. Mutlu, and C. Alkan, "Accelerating read mapping with fasthash," in *BMC genomics*, vol. 14, p. S13, BioMed Central, 2013.

[109] V. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[110] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J Mol Biol*, vol. 48, pp. 443–453, Mar 1970.

[111] M. Alser, H. Hassan, H. Xin, O. Ergin, O. Mutlu, and C. Alkan, "Gatekeeper: a new hardware architecture for accelerating pre-alignment in dna short read mapping," *Bioinformatics*, vol. 33, no. 21, pp. 3355–3363, 2017.

[112] P. Flicek and E. Birney, "Sense from sequence reads: methods for alignment and assembly," *Nature methods*, vol. 6, no. 11s, p. S6, 2009.

[113] C. Trapnell and S. L. Salzberg, "How to map billions of short reads onto genomes," *Nature biotechnology*, vol. 27, no. 5, p. 455, 2009.

[114] S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, and M. C. Frith, "Adaptive seeds tame genomic sequence comparison," *Genome research*, vol. 21, no. 3, pp. 487–493, 2011.

[115] M. J. Chaisson and G. Tesler, "Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory," *BMC Bioinformatics*, vol. 13, p. 238, 2012.

[116] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," *arXiv preprint arXiv:1303.3997*, 2013.

[117] G. Myers, "Efficient local alignment discovery amongst noisy long reads," in *International Workshop on Algorithms in Bioinformatics*, pp. 52–67, Springer, 2014.

[118] I. Sović, M. Šikić, A. Wilm, S. N. Fenlon, S. Chen, and N. Nagarajan, "Fast and sensitive mapping of nanopore sequencing reads with graphmap," *Nature communications*, vol. 7, p. 11307, 2016.

[119] C.-L. Xiao, Y. Chen, S.-Q. Xie, K.-N. Chen, Y. Wang, Y. Han, F. Luo, and Z. Xie, "Mecat: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads," *nature methods*, vol. 14, no. 11, p. 1072, 2017.

[120] B. Liu, Y. Gao, and Y. Wang, "Lamsa: fast split read alignment with long approximate matches," *Bioinformatics*, vol. 33, no. 2, pp. 192–201, 2017.

[121] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 1, p. 7, 2018.

[122] K. M. Steinberg, V. A. Schneider, C. Alkan, M. J. Montague, W. C. Warren, D. M. Church, and R. K. Wilson, "Building and improving reference genome assemblies," *Proceedings of the IEEE*, vol. 105, pp. 422–435, Mar. 2017.

[123] K.-J. Räihä and E. Ukkonen, "The shortest common supersequence problem over binary alphabet is NP-complete," *Theoretical Computer Science*, vol. 16, no. 2, pp. 187–198, 1981.

[124] G. G. Sutton, O. White, M. D. Adams, and A. R. Kerlavage, "TIGR assembler: A new tool for assembling large shotgun sequencing projects," *Genome Science and Technology*, vol. 1, no. 1, pp. 9–19, 1995.

[125] M. de la Bastide and W. R. McCombie, "Assembling genomic dna sequences with phrap," *Current Protocols in Bioinformatics*, pp. 11–4, 2007.

[126] S. Batzoglou, D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. P. Mesirov, and E. S. Lander, "ARACHNE: a whole-genome shotgun assembler," *Genome Res*, vol. 12, pp. 177–189, Jan 2002.

[127] J. C. Mullikin and Z. Ning, "The phusion assembler," *Genome Res*, vol. 13, pp. 81–90, Jan 2003.

[128] X. Huang, J. Wang, S. Aluru, S.-P. Yang, and L. Hillier, "PCAP: a whole-genome assembly program," *Genome Res*, vol. 13, pp. 2164–2170, Sep 2003.

[129] E. W. Myers, "The fragment assembly string graph," *Bioinformatics*, vol. 21 Suppl 2, pp. ii79–ii85, Sep 2005.

[130] M. J. Chaisson and P. A. Pevzner, "Short read fragment assembly of bacterial genomes," *Genome Res*, vol. 18, pp. 324–330, Feb 2008.

[131] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de Bruijn graphs," *Genome Res*, vol. 18, pp. 821–829, May 2008.

[132] M. J. Chaisson, D. Brinza, and P. A. Pevzner, "De novo fragment assembly with short mate-paired reads: Does the read length matter?," *Genome Res*, vol. 19, pp. 336–346, Feb 2009.

[133] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol, "ABySS: a parallel assembler for short read sequence data," *Genome Res*, vol. 19, pp. 1117–1123, Jun 2009.

[134] S. Boisvert, F. Laviolette, and J. Corbeil, "Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies," *J Comput Biol*, vol. 17, pp. 1519–1533, Nov 2010.

[135] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang, "De novo assembly of human genomes with massively parallel short read sequencing," *Genome Res*, vol. 20, pp. 265–272, Feb 2010.

[136] S. Gnerre, I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe, "High-quality draft assemblies of mammalian genomes from massively parallel sequence data," *Proc Natl Acad Sci U S A*, vol. 108, pp. 1513–1518, Jan 2011.

[137] P. Medvedev, S. Pham, M. Chaisson, G. Tesler, and P. Pevzner, "Paired de Bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers," *J Comput Biol*, vol. 18, pp. 1625–1634, Nov 2011.

[138] J. A. Chapman, I. Ho, S. Sunkara, S. Luo, G. P. Schroth, and D. S. Rokhsar, "Meraculous: de novo genome assembly with short paired-end reads," *PloS one*, vol. 6, no. 8, p. e23501, 2011.

[139] Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. McVean, "De novo assembly and genotyping of variants using colored de Bruijn graphs," *Nat Genet*, vol. 44, pp. 226–232, Feb 2012.

[140] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner, "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing," *J Comput Biol*, vol. 19, pp. 455–477, May 2012.

[141] E. Georganas, A. Buluç, J. Chapman, S. Hofmeyr, C. Aluru, R. Egan, L. Oliker, D. Rokhsar, and K. Yelick, "HipMer: an extreme-scale de novo genome assembler," in *Proceedings of the International Conference for High*

Performance Computing, Networking, Storage and Analysis, p. 14, ACM, 2015.

[142] H. Li, "Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences," *Bioinformatics*, vol. 32, no. 14, pp. 2103–2110, 2016.

[143] F. Hormozdiari, *Structural variation discovery: the easy, the hard and the ugly*. PhD thesis, Applied Science: School of Computing Science, 2011.

[144] C. Kingsford, M. C. Schatz, and M. Pop, "Assembly complexity of prokaryotic genomes using short reads," *BMC bioinformatics*, vol. 11, no. 1, p. 21, 2010.

[145] M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano, "Scaffolding pre-assembled contigs using SSPACE," *Bioinformatics*, vol. 27, pp. 578–579, Feb 2011.

[146] S. Gao, W.-K. Sung, and N. Nagarajan, "Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences," *J Comput Biol*, vol. 18, pp. 1681–1691, Nov 2011.

[147] N. Donmez and M. Brudno, "SCARPA: scaffolding reads with practical algorithms," *Bioinformatics*, vol. 29, pp. 428–434, Feb 2013.

[148] K. Sahlin, F. Vezzi, B. Nystedt, J. Lundeberg, and L. Arvestad, "BESST–efficient scaffolding of large fragmented assemblies," *BMC Bioinformatics*, vol. 15, p. 281, 2014.

[149] R. L. Warren, C. Yang, B. P. Vandervalk, B. Behsaz, A. Lagman, S. J. Jones, and I. Birol, "Links: Scalable, alignment-free scaffolding of draft genomes with long reads," *GigaScience*, vol. 4, no. 1, p. 35, 2015.

[150] P. Medvedev, M. Stanciu, and M. Brudno, "Computational methods for discovering structural variation with next-generation sequencing," *Nat Methods*, vol. 6, pp. S13–S20, Nov 2009.

[151] 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. K. 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. Stütz, 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 . G. Project, "Mapping copy number variation by population-scale genome sequencing," *Nature*, vol. 470, pp. 59–65, Feb 2011.

[152] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, "Substantial biases in ultra-short read data sets from high-throughput dna sequencing," *Nucleic acids research*, vol. 36, no. 16, p. e105, 2008.

[153] S. Volik, S. Zhao, K. Chin, J. H. Brebner, D. R. Herndon, Q. Tao, D. Kowbel, G. Huang, A. Lapuk, W.-L. Kuo, *et al.*, "End-sequence profiling: sequence-based analysis of aberrant genomes," *Proceedings of the National Academy of Sciences*, vol. 100, no. 13, pp. 7696–7701, 2003.

[154] F. Hormozdiari, C. Alkan, E. E. Eichler, and S. C. Sahinalp, "Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes," *Genome Res*, vol. 19, pp. 1270–1278, Jul 2009.

[155] J. Korbel, A. Abyzov, X. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder, and M. Gerstein, "PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data," *Genome Biol*, vol. 10, p. R23, Feb 2009.

[156] 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," *Nat Methods*, vol. 6, pp. 677–681, Sep 2009.

[157] S. Lee, F. Hormozdiari, C. Alkan, and M. Brudno, "MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions," *Nat Methods*, vol. 6, pp. 473–474, Jul 2009.

[158] B. Zeitouni, V. Boeva, I. Janoueix-Lerosey, S. Loeillet, P. Legoix-Né, A. Nicolas, O. Delattre, and E. Barillot, "Svdetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data," *Bioinformatics*, vol. 26, no. 15, pp. 1895–1896, 2010.

[159] S. Sindi, E. Helman, A. Bashir, and B. J. Raphael, "A geometric approach for classification and comparison of structural variants," *Bioinformatics*, vol. 25, pp. i222–i230, June 2009.

[160] A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein, "Cnvnator: an approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing," *Genome research*, vol. 21, no. 6, pp. 974–984, 2011.

[161] 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 Res*, vol. 19, pp. 1586–1592, Sep 2009.

[162] 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*, vol. 25, pp. 2865–2871, Nov 2009.

[163] A. Abyzov and M. Gerstein, "Age: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision," *Bioinformatics*, vol. 27, no. 5, pp. 595–603, 2011.

[164] I. Hajirasouliha, F. Hormozdiari, C. Alkan, J. M. Kidd, I. Birol, E. E. Eichler, and S. C. Sahinalp, "Detection and characterization of novel sequence insertions using paired-end next-generation sequencing," *Bioinformatics*, vol. 26, pp. 1277–1283, May 2010.

[165] P. Medvedev, M. Fiume, M. Dzamba, T. Smith, and M. Brudno, "Detecting copy number variation with mated short reads," *Genome Res*, vol. 20, pp. 1613–1622, Nov 2010.

[166] J. Qi and F. Zhao, "ingap-sv: a novel scheme to identify and visualize structural variation from paired end mapping data," *Nucleic acids research*, vol. 39, no. suppl_2, pp. W567–W575, 2011.

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

[168] S. S. Sindi, S. Onal, L. C. Peng, H.-T. Wu, and B. J. Raphael, "An integrative probabilistic model for identification of structural variation in sequencing data," *Genome Biol*, vol. 13, no. 3, p. R22, 2012.

[169] G. Escaramís, C. Tornador, L. Bassaganyas, R. Rabionet, J. M. Tubio, A. Martínez-Fundichely, M. Cáceres, M. Gut, S. Ossowski, and X. Estivill, "Pesv-fisher: identification of somatic and non-somatic structural variants using next generation sequencing data," *PLoS One*, vol. 8, no. 5, p. e63377, 2013.

[170] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall, "LUMPY: a probabilistic framework for structural variant discovery," *Genome Biol*, vol. 15, no. 6, p. R84, 2014.

[171] X. Chen, O. Schulz-Trieglaff, R. Shaw, B. Barnes, F. Schlesinger, M. Källberg, A. J. Cox, S. Kruglyak, and C. T. Saunders, "Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications," *Bioinformatics*, vol. 32, no. 8, pp. 1220–1222, 2015.

[172] D. Iakovishina, I. Janoueix-Lerosey, E. Barillot, M. Regnier, and V. Boeva, "Sv-bay: structural variant detection in cancer genomes using a bayesian approach with correction for gc-content and read mappability," *Bioinformatics*, vol. 32, no. 7, pp. 984–992, 2016.

[173] P. Kavak, Y.-Y. Lin, I. Numanagic, H. Asghari, T. Güngör, C. Alkan, and F. Hach, "Discovery and genotyping of novel sequence insertions in many sequenced individuals," *Bioinformatics (Oxford, England)*, vol. 33, pp. i161–i169, July 2017.

[174] A. Soylev, C. Kockan, F. Hormozdiari, and C. Alkan, "Toolkit for automated and rapid discovery of structural variants," *Methods*, vol. 129, pp. 3–7, 2017.

[175] J. A. Wala, P. Bandopadhayay, N. F. Greenwald, R. O'Rourke, T. Sharpe, C. Stewart, S. Schumacher, Y. Li, J. Weischenfeldt, X. Yao, *et al.*, "Svaba: genome-wide detection of structural variants and indels by local assembly," *Genome research*, vol. 28, no. 4, pp. 581–591, 2018.

[176] J. O. Korbel, A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, B. E. Taillon, Z. Chen, A. Tanzer, A. C. E. Saunders, J. Chi, F. Yang, N. P. Carter, M. E. Hurles, S. M. Weissman, T. T. Harkins, M. B. Gerstein, M. Egholm, and M. Snyder, "Paired-end mapping reveals extensive structural variation in the human genome," *Science*, vol. 318, pp. 420–426, Oct 2007.

[177] 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. W. 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," *Nat Genet*, vol. 40, pp. 722–729, Jun 2008.

[178] D. Y. Chiang, G. Getz, D. B. Jaffe, M. J. T. O'Kelly, X. Zhao, S. L. Carter, C. Russ, C. Nusbaum, M. Meyerson, and E. S. Lander, "High-resolution mapping of copy-number alterations with massively parallel sequencing," *Nat Methods*, vol. 6, pp. 99–103, Jan 2009.

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

[180] X. She, Z. Jiang, R. A. Clark, G. Liu, Z. Cheng, E. Tuzun, D. M. Church, G. Sutton, A. L. Halpern, and E. E. Eichler, "Shotgun sequence assembly and recent segmental duplications within the human genome," *Nature*, vol. 431, no. 7011, p. 927, 2004.

[181] C. Alkan, S. Sajjadian, and E. E. Eichler, "Limitations of next-generation genome sequence assembly," *Nat Methods*, vol. 8, pp. 61–65, Jan 2011.

[182] G. G. Faust and I. M. Hall, "Samblaster: fast duplicate marking and structural variant read extraction," *Bioinformatics*, vol. 30, pp. 2503–2505, Sep 2014.

[183] M. J. P. Chaisson, J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig, F. Hormozdiari, F. Antonacci, U. Surti, R. Sandstrom, M. Boitano, J. M. Landolin, J. A. Stamatoyannopoulos, M. W. Hunkapiller, J. Korlach, and E. E. Eichler, "Resolving the complexity of the human genome using single-molecule sequencing," *Nature*, vol. 517, pp. 608–611, Jan 2015.

[184] M. J. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. Rodriguez, L. Guo, R. L. Collins, X. Fan, J. Wen, R. E. Handsaker, S. Fairley, Z. N. Kronenberg, X. Kong, F. Hormozdiari, D. Lee, A. M. Wenger, A. Hastie, D. Antaki, P. Audano, H. Brand, S. Cantsilieris, H. Cao, E. Cerveira, C. Chen, X. Chen, C.-S. Chin, Z. Chong, N. T. Chuang, D. M. Church, L. Clarke, A. Farrell, J. Flores, T. Galeev, G. David, M. Gujral, V. Guryev, W. Haynes-Heaton, J. Korlach, S. Kumar, J. Y. Kwon, J. E. Lee, J. Lee, W.-P. Lee, S. P. Lee, P. Marks, K. Valud-Martinez, S. Meiers, K. M. Munson, F. Navarro, B. J. Nelson, C. Nodzak, A. Noor, S. Kyriazopoulou-Panagiotopoulou, A. Pang, Y. Qiu, G. Rosanio, M. Ryan, A. Stutz, D. C. Spierings, A. Ward, A. E. Welsch, M. Xiao, W. Xu, C. Zhang, Q. Zhu, X. Zheng-Bradley, G. Jun, L. Ding, C. L. Koh, B. Ren, P. Flicek, K. Chen, M. B. Gerstein, P.-Y. Kwok, P. M. Lansdorp, G. Marth, J. Sebat, X. Shi, A. Bashir, K. Ye, S. E. Devine, M. Talkowski, R. E. Mills, T. Marschall, J. Korbel, E. E. Eichler, and C. Lee, "Multi-platform discovery of haplotype-resolved structural variation in human genomes," *bioRxiv*, 2017.

[185] 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*, vol. 26, pp. i350–i357, Jun 2010.

[186] F. Hormozdiari, I. Hajirasouliha, A. McPherson, E. E. Eichler, and S. C. Sahinalp, "Simultaneous structural variation discovery among multiple paired-end sequenced genomes," *Genome Res*, vol. 21, pp. 2203–2212, Dec 2011.

[187] S. Lee, E. Cheran, and M. Brudno, "A robust framework for detecting structural variations in a genome," *Bioinformatics*, vol. 24, pp. i59–i67, Jul 2008.

[188] A. R. Quinlan, R. A. Clark, S. Sokolova, M. L. Leibowitz, Y. Zhang, M. E. Hurles, J. C. Mell, and I. M. Hall, "Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome," *Genome Res*, vol. 20, pp. 623–635, May 2010.

[189] 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*, vol. 25, pp. 2078–2079, Aug 2009.

[190] A. Bashir, S. Volik, C. Collins, V. Bafna, and B. J. Raphael, "Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer," *PLoS Comput Biol*, vol. 4, p. e1000051, Apr 2008.

[191] V. V. Vazirani, *Approximation algorithms*. Springer Science & Business Media, 2013.

[192] E. S. Lander and M. S. Waterman, "Genomic mapping by fingerprinting random clones: a mathematical analysis," *Genomics*, vol. 2, no. 3, pp. 231–239, 1988.

[193] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics*, vol. 7, no. 2, p. 85, 2006.

[194] L. E. Vissers, B. B. de Vries, K. Osoegawa, I. M. Janssen, T. Feuth, C. O. Choy, H. Straatman, W. van der Vliet, E. H. Huys, A. van Rijk, *et al.*, "Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities," *The American Journal of Human Genetics*, vol. 73, no. 6, pp. 1261–1270, 2003.

[195] R. Lucito, J. Healy, J. Alexander, A. Reiner, D. Esposito, M. Chi, L. Rodgers, A. Brady, J. Sebat, J. Troge, *et al.*, "Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation," *Genome research*, vol. 13, no. 10, pp. 2291–2305, 2003.

[196] J. R. MacDonald, R. Ziman, R. K. Yuen, L. Feuk, and S. W. Scherer, "The database of genomic variants: a curated collection of structural variation in the human genome," *Nucleic acids research*, vol. 42, no. D1, pp. D986–D992, 2013.

[197] A. Martínez-Fundichely, S. Casillas, R. Egea, M. Ràmia, A. Barbadilla, L. Pantano, M. Puig, and M. Cáceres, "InvFEST, a database integrating information of polymorphic inversions in the human genome," *Nucleic Acids Res*, vol. 42, pp. D1027–D1032, Jan 2014.

[198] H. V. Firth, S. M. Richards, A. P. Bevan, S. Clayton, M. Corpas, D. Rajan, S. Van Vooren, Y. Moreau, R. M. Pettett, and N. P. Carter, "Decipher: database of chromosomal imbalance and phenotype in humans using ensembl resources," *The American Journal of Human Genetics*, vol. 84, no. 4, pp. 524–533, 2009.

[199] J. M. Zook, B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide, and M. Salit, "Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls," *Nature biotechnology*, vol. 32, no. 3, p. 246, 2014.

[200] J. Zook, J. McDaniel, H. Parikh, H. Heaton, S. A. Irvine, L. Trigg, R. Truty, C. Y. McLean, F. M. De La Vega, M. Salit, *et al.*, "Reproducible integration

of multiple sequencing datasets to form high-confidence snp, indel, and reference calls for five human genome reference materials," *bioRxiv*, p. 281006, 2018.

[201] A. K. Ewart, C. A. Morris, D. Atkinson, W. Jin, K. Sternes, P. Spallone, A. D. Stock, M. Leppert, and M. T. Keating, "Hemizygosity at the elastin locus in a developmental disorder, williams syndrome," *Nature genetics*, vol. 5, no. 1, p. 11, 1993.

[202] U. Francke, "Williams-beuren syndrome: genes and mechanisms," *Human Molecular Genetics*, vol. 8, no. 10, pp. 1947–1954, 1999.

[203] A. C. Smith, L. McGavran, J. Robinson, G. Waldstein, J. Macfarlane, J. Zonona, J. Reiss, M. Lahr, L. Allen, E. Magenis, *et al.*, "Interstitial deletion of (17)(p11. 2p11. 2) in nine patients," *American journal of medical genetics*, vol. 24, no. 3, pp. 393–414, 1986.

[204] F. Greenberg, V. Guzzetta, R. M. de Oca-Luna, R. E. Magenis, A. Smith, S. F. Richter, I. Kondo, W. B. Dobyns, P. I. Patel, and J. R. Lupski, "Molecular analysis of the smith-magenis syndrome: a possible contiguous-gene syndrome associated with del (17)(p11. 2).," *American journal of human genetics*, vol. 49, no. 6, p. 1207, 1991.

[205] S. A. Boikos and C. A. Stratakis, "Carney complex: pathology and molecular genetics," *Neuroendocrinology*, vol. 83, no. 3-4, pp. 189–199, 2006.

[206] K. Toyo-Oka, A. Shionoya, M. J. Gambello, C. Cardoso, R. Leventer, H. L. Ward, R. Ayala, L.-H. Tsai, W. Dobyns, D. Ledbetter, *et al.*, "14-3-3ε is important for neuronal migration by binding to nudel: a molecular explanation for miller–dieker syndrome," *Nature genetics*, vol. 34, no. 3, p. 274, 2003.

[207] P. F. Chance, M. K. Alderson, K. A. Leppig, M. W. Lensch, N. Matsunami, B. Smith, P. D. Swanson, S. J. Odelberg, C. M. Disteche, and T. D. Bird, "Dna deletion associated with hereditary neuropathy with liability to pressure palsies," *Cell*, vol. 72, no. 1, pp. 143–151, 1993.

[208] J. Overhauser, X. Huang, M. Gersh, W. Wilson, J. McMahon, U. Bengtsson, K. Rojas, M. Meyer, and J. J. Wasmuth, "Molecular and phenotypic mapping of the short arm of chromosome 5: sublocalization of the critical region for the cri-du-chat syndrome," *Human molecular genetics*, vol. 3, no. 2, pp. 247–252, 1994.

[209] R. D. Nicholls, S. Saitoh, and B. Horsthemke, "Imprinting in prader–willi and angelman syndromes," *Trends in Genetics*, vol. 14, no. 5, pp. 194–200, 1998.

[210] D. Wilson, J. Burn, P. Scambler, and J. Goodship, "Digeorge syndrome: part of catch 22.," *Journal of Medical Genetics*, vol. 30, no. 10, pp. 852–856, 1993.

[211] M. H. Brewer, R. Chaudhry, J. Qi, A. Kidambi, A. P. Drew, M. P. Menezes, M. M. Ryan, M. A. Farrar, D. Mowat, G. M. Subramanian, *et al.*, "Whole genome sequencing identifies a 78 kb insertion from chromosome 8 as the cause of charcot-marie-tooth neuropathy cmtx3," *PLoS genetics*, vol. 12, no. 7, p. e1006177, 2016.

[212] E. C. Landels, I. Ellis, A. Fensom, P. Green, and M. Bobrow, "Frequency of the tay-sachs disease splice and insertion mutations in the uk ashkenazi jewish population.," *Journal of medical genetics*, vol. 28, no. 3, pp. 177–180, 1991.

[213] 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, no. 4, p. e56, 2005.

[214] H. Stefansson, A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson, J. Barnard, A. Baker, A. Jonasdottir, A. Ingason, V. G. Gudnadottir, N. Desnica, A. Hicks, A. Gylfason, D. F. Gudbjartsson, G. M. Jonsdottir, J. Sainz, K. Agnarsson, B. Birgisdottir, S. Ghosh, A. Olafsdottir, J.-B. Cazier, K. Kristjansson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, A. Kong, and K. Stefansson, "A common inversion under selection in Europeans," *Nat Genet*, vol. 37, pp. 129–137, Feb 2005.

[215] J. M. Kidd, G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci, E. Haugen, T. Zerr, N. A. Yamada, P. Tsang, T. L. Newman, E. Tüzün, Z. Cheng, H. M. Ebling, N. Tusneem, R. David, W. Gillett, K. A. Phelps, M. Weaver, D. Saranga, A. Brand, W. Tao, E. Gustafson, K. McKernan, L. Chen, M. Malig, J. D. Smith, J. M. Korn, S. A. McCarroll, D. A. Altshuler, D. A. Peiffer, M. Dorschner, J. Stamatoyannopoulos, D. Schwartz, D. A. Nickerson, J. C. Mullikin, R. K. Wilson, L. Bruhn, M. V. Olson, R. Kaul, D. R. Smith, and E. E. Eichler, "Mapping and sequencing of structural variation from eight human genomes," *Nature*, vol. 453, pp. 56–64, May 2008.

[216] D. Lakich, H. H. Kazazian Jr, S. E. Antonarakis, and J. Gitschier, "Inversions disrupting the factor viii gene are a common cause of severe haemophilia a," *Nature genetics*, vol. 5, no. 3, p. 236, 1993.

[217] M.-L. Bondeson, N. Dahl, H. Malmgren, W. J. Kleijer, T. Tönnesen, B.-M. Carlberg, and U. Pettersson, "Inversion of the ids gene resulting from recombination with ids-related sequences in a common cause of the hunter syndrome," *Human molecular genetics*, vol. 4, no. 4, pp. 615–621, 1995.

[218] K. Small, J. Iber, and S. T. Warren, "Emerin deletion reveals a common x-chromosome inversion mediated by inverted repeats," *Nature genetics*, vol. 16, no. 1, p. 96, 1997.

[219] L. Feuk, "Inversion variants in the human genome: role in disease and genome architecture," *Genome medicine*, vol. 2, no. 2, p. 11, 2010.

[220] L. R. Osborne, M. Li, B. Pober, D. Chitayat, J. Bodurtha, A. Mandel, T. Costa, T. Grebe, S. Cox, L.-C. Tsui, *et al.*, "A 1.5 million–base pair inversion polymorphism in families with williams-beuren syndrome," *Nature genetics*, vol. 29, no. 3, p. 321, 2001.

[221] H. Stefansson, A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson, J. Barnard, A. Baker, A. Jonasdottir, A. Ingason, V. G. Gudnadottir, *et al.*, "A common inversion under selection in europeans," *Nature genetics*, vol. 37, no. 2, p. 129, 2005.

[222] N. Kurotaki, N. Harada, O. Shimokawa, N. Miyake, H. Kawame, K. Uetake, Y. Makita, T. Kondoh, T. Ogata, T. Hasegawa, *et al.*, "Fifty microdeletions among 112 cases of sotos syndrome: low copy repeats possibly mediate the common deletion," *Human mutation*, vol. 22, no. 5, pp. 378–387, 2003.

[223] R. E. Mills, E. A. Bennett, R. C. Iskow, and S. E. Devine, "Which transposable elements are active in the human genome?," *Trends Genet*, vol. 23, pp. 183–191, Apr 2007.

[224] Y. Miki, I. Nishisho, A. Horii, Y. Miyoshi, J. Utsunomiya, K. W. Kinzler, B. Vogelstein, and Y. Nakamura, "Disruption of the apc gene by a retrotransposal insertion of l1 sequence in a colon cancer," *Cancer research*, vol. 52, no. 3, pp. 643–645, 1992.

[225] E. Sukarova, A. Dimovski, P. Tchacarova, G. Petkov, and G. Efremov, "An alu insert as the cause of a severe form of hemophilia a," *Acta haematologica*, vol. 106, no. 3, pp. 126–129, 2001.

[226] E. Kondo-Iida, K. Kobayashi, M. Watanabe, J. Sasaki, T. Kumagai, H. Koide, K. Saito, M. Osawa, Y. Nakamura, and T. Toda, "Novel mutations and genotype-phenotype relationships in 107 families with fukuyama-type congenital muscular dystrophy (fcmd)," *Human Molecular Genetics*, vol. 8, no. 12, pp. 2303–2309, 1999.

[227] 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, *et al.*, "Mobile elements create structural variation: analysis of a complete human genome," *Genome research*, pp. gr–091827, 2009.

[228] J. A. Bailey, G. Liu, and E. E. Eichler, "An alu transposition model for the origin and expansion of human segmental duplications," *Am J Hum Genet*, vol. 73, pp. 823–834, Oct 2003.

[229] W. Bao, K. K. Kojima, and O. Kohany, "Repbase update, a database of repetitive elements in eukaryotic genomes," *Mobile DNA*, vol. 6, no. 1, p. 11, 2015.

[230] M. Fukuda, S. Wakasugi, T. Tsuzuki, H. Nomiyama, K. Shimada, and T. Miyata, "Mitochondrial dna-like sequences in the human nuclear genome: characterization and implications in the evolution of mitochondrial dna," *Journal of molecular biology*, vol. 186, no. 2, pp. 257–266, 1985.

[231] T. Mourier, A. J. Hansen, E. Willerslev, and P. Arctander, "The human genome project reveals a continuous transfer of large mitochondrial fragments to the nucleus," *Molecular biology and evolution*, vol. 18, no. 9, pp. 1833–1837, 2001.

[232] D. Leister, "Origin, evolution and genetic effects of nuclear insertions of organelle dna," *TRENDS in Genetics*, vol. 21, no. 12, pp. 655–663, 2005.

[233] C. Turner, C. Killoran, N. S. Thomas, M. Rosenberg, N. A. Chuzhanova, J. Johnston, Y. Kemel, D. N. Cooper, and L. G. Biesecker, "Human genetic disease caused by de novo mitochondrial-nuclear dna transfer," *Human genetics*, vol. 112, no. 3, pp. 303–309, 2003.

[234] E. Goldin, S. Stahl, A. M. Cooney, C. R. Kaneski, S. Gupta, R. O. Brady, J. R. Ellis, and R. Schiffmann, "Transfer of a mitochondrial dna fragment to mcoln1 causes an inherited case of mucolipidosis iv," *Human mutation*, vol. 24, no. 6, pp. 460–465, 2004.

[235] Y.-G. Yao, Q.-P. Kong, A. Salas, and H.-J. Bandelt, "Pseudo-mitochondrial genome haunts disease studies," *Journal of medical genetics*, 2008.

[236] J. B. Stewart and P. F. Chinnery, "The dynamics of mitochondrial dna heteroplasmy: implications for human health and disease," *Nature Reviews Genetics*, vol. 16, no. 9, p. 530, 2015.

[237] G. Dayama, S. B. Emery, J. M. Kidd, and R. E. Mills, "The genomic landscape of polymorphic human nuclear mitochondrial insertions," *Nucleic acids research*, vol. 42, no. 20, pp. 12640–12649, 2014.

[238] Y. Ji, E. E. Eichler, S. Schwartz, and R. D. Nicholls, "Structure of chromosomal duplicons and their role in mediating human genomic disorders," *Genome research*, vol. 10, no. 5, pp. 597–610, 2000.

[239] G. H. Perry, N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, F. A. Villanea, J. L. Mountain, R. Misra, *et al.*, "Diet and the evolution of human amylase gene copy number variation," *Nature genetics*, vol. 39, no. 10, p. 1256, 2007.

[240] T. Marques-Bonet, S. Girirajan, and E. E. Eichler, "The origins and impact of primate segmental duplications," *Trends in Genetics*, vol. 25, no. 10, pp. 443–454, 2009.

[241] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, pp. 841–842, Mar 2010.

[242] A. R. Martin, H. A. Costa, T. Lappalainen, B. M. Henn, J. M. Kidd, M.-C. Yee, F. Grubert, H. M. Cann, M. Snyder, S. B. Montgomery, *et al.*, "Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture," *PLoS genetics*, vol. 10, no. 8, p. e1004549, 2014.

[243] P. Kavak, Y.-Y. Lin, I. Numanagić, H. Asghari, T. Güngör, C. Alkan, and F. Hach, "Discovery and genotyping of novel sequence insertions in many sequenced individuals," *Bioinformatics*, vol. 33, no. 14, pp. i161–i169, 2017.

[244] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, "Base-calling of automated sequencer traces usingphred. i. accuracy assessment," *Genome research*, vol. 8, no. 3, pp. 175–185, 1998.

[245] J. C. Mu, M. Mohiyuddin, J. Li, N. Bani Asadi, M. B. Gerstein, A. Abyzov, W. H. Wong, and H. Y. K. Lam, "VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications," *Bioinformatics*, vol. 31, pp. 1469–1471, May 2015.

[246] 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, *et al.*, "The variant call format and vcftools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011.

[247] W. Huang, L. Li, J. R. Myers, and G. T. Marth, "ART: a next-generation sequencing read simulator," *Bioinformatics*, vol. 28, pp. 593–594, Feb 2012.

[248] R. Luo, F. J. Sedlazeck, C. A. Darby, S. M. Kelly, and M. C. Schatz, "LRSim: a linked-reads simulator generating insights for better genome partitioning.," *Computational and structural biotechnology journal*, vol. 15, pp. 478–484, 2017.

Appendix A

Data and Code Availability

TARDIS is available under BSD 3-clause license at `https://github.com/BilkentCompGen/tardis`, and the CNVSim simulator is available at `https://github.com/LeMinhThong/CNVSim`.

Data sets for short-read sequencing: NA12878 WGS data set can be downloaded from the Illumina Platinum Genomes Project at `https://www.illumina.com/platinumgenomes.html`. SRA IDs for CHM1 and CHM13 are SRP044331 and SRP080317, respectively. GenBank assembly accession numbers for CHM1 and CHM13 assemblies are GCA 000306695.2 and GCA 000983455.2.

Data sets for 10x Genomics Platform is available via the Genome in a Bottle Project FTP site at `ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/10XGenomics/` for Ashkenazim trio son (HG002), at ftp://`ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/10Xgenomics_ChromiumGenome_LongRanger2.1_09302016/NA12878_hg19/` for NA12878 and the CHM1 genome generated with 10xG Linked-Reads is available at `https://support.10xgenomics.com/de-novo-assembly/datasets/2.0.0/chm`