LARGE STRUCTURAL VARIATION DISCOVERY USING LONG READS WITH SEVERAL DEGREES OF ERROR

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By
Ezgi Ebren
December 2020

LARGE STRUCTURAL VARIATION DISCOVERY USING LONG READS WITH SEVERAL DEGREES OF ERROR

By Ezgi Ebren

December 2020

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

Can Alkan(Advisor)

A. Erciiment Çiçek

Tunca Doğan

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

LARGE STRUCTURAL VARIATION DISCOVERY USING LONG READS WITH SEVERAL DEGREES OF ERROR

Ezgi Ebren
M.S. in Computer Engineering
Advisor: Can Alkan
December 2020

Genomic structural variations (SVs) are briefly defined as large-scale alterations of DNA content, copy, and organization. Although significant progress has been made since the introduction of high throughput sequencing (HTS) in characterizing SVs, accurate detection of complex SVs and balanced rearrangements still remains elusive due to the sequence complexity at the breakpoints. Until very recently, the difficulty of read mapping in such regions when the reads were short and the high error rates of long read platforms kept the problem challenging. However, with the introduction of the Pacific Biosciences' High Fidelity (HiFi) sequencing methodology, powerful SV detection and breakpoint resolution became possible as a result of its capability to produce highly accurate ($> 99\%$) long reads ($10 - 20$ kbps).

Here, we introduce DALEK, a novel algorithm that aims to use long-read technologies to discover large structural variations with high break-point resolution. DALEK uses split read and read depth signatures from long read data to discover large ($\geq$ 10 kbps) deletions, inversions and segmental duplications. We also develop methods to detect large SVs in existing high-error Oxford Nanopore Technologies data.

Keywords: Structural Variation, Deletion, Inversion, Segmental Duplication, HiFi Reads, Long Reads.

ÖZET

FARKLI HATA ORANLARINA SAHİP UZUN OKUMALAR İLE BÜYÜK YAPISAL VARYASYON TESPİTİ

Ezgi Ebren
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Can Alkan
Aralık 2020

Genetik yapısal varyasyonlar (YV) kısaca DNA'nın içerik, kopya ve düzenindeki büyük çaplı değişikliklerdir. Her ne kadar yüksek çıktılı dizileme (YÇD) kullanılmaya başlandıktan sonra ciddi oranda aşama kaydedildiyse de, kırılma noktalarındaki dizi karmaşıklığı kompleks YV ve dengeli yeniden-düzenlenmelerin doğru tespiti muamması güçlüğünü sürdürmektedir. Yakın zamana kadar, okumalar kısa olduğunda bahsi geçen bölgelerdeki okuma hizalamasının zorluğu ve uzun okuma platformlarının yüksek hata oranları YV keşfi sürecindeki problemlerin temelini oluşturmaktaydı. Ancak, Pacific Biosciences şirketinin, $> 99\%$ doğruluk payı ve $10-20$ kbps uzunlukta okuma yapabilen High Fidelity (Yüksek Doğrulukta, *HiFi*) dizileme metodunun ortaya çıkmasıyla, etkili YV keşfi ve kırılma noktası çözünürlüğünün iyileştirilmesi mümkün olmuştur.

Biz bu çalışmayla uzun okuma teknolojileri kullanarak yüksek kırılma noktası çözünürlüğüne sahip büyük yapısal varyasyon keşfi yapan özgün bir algoritma olan DALEK'i sunuyoruz. DALEK, uzun okumalardaki ayrık dizi ve dizi derinliği sinyalleri kullanarak büyük ($\geq$ 10 kbps) silinme, inversiyon ve kesitsel duplikasyonları keşfetmektedir. Ayrıca algoritmanın parametrelerine göre halihazırda yüksek hatalı olan Oxford Nanopore Technologies uzun okumalarından da YV tespiti yapabilmektedir.

Anahtar sözcükler: Yapısal Varyasyon, Silinme, Delesyon, İnversiyon, Kesitsel Duplikasyon, HiFi Okuma, Uzun Okuma.

Acknowledgement

First and foremost, I must give the greatest of thanks to my supervisor, Asst. Prof. Dr. Can Alkan. He has been deeply involved and incredibly helpful at all times, not once losing faith in me or the work I do. He supported me in every decision and welcomed every mistake so that they never felt like failure. I will always be in his debt for being an extraordinary guide for these three years.

I wish to acknowledge EMBO grant (IG 2521) and TÜBİTAK 215E172 grant for the financial support of this project.

I would also like to express my gratitude to my thesis committee, Asst. Prof. Dr. A. Ercüment Çiçek and Asst. Prof. Dr. Tunca Doğan, for their consideration and feedback.

In addition, I wish to thank all the past and present members of Alkanlab, their work and experience guided me through this time in one way or the other. Fatma Kahveci and Can Fırtına welcomed me to the lab and gave me their experienced help and support without dismissing any of my elementary inquiries. Fatih Karaoğlanoğlu was not only sitting next to me for many years, but we also shared similar research interests and assisted each other during critical times, which allowed me to improve notably. Halil İbrahim Özercan, Balanur İçen and Zülal Bingöl were the sincerest of colleagues and friends, they supported me both professionally and mentally.

To the people that are my pillars, I owe them my deepest gratitude. To my mother Zerrin, who allowed me to focus on my work while she was there for everything else. My Father Erdinç showed me his support from miles away everyday, helping me have more faith in myself. My lifelong friend, my brother Gökhan made for a great conversation partner during the times I needed a break. And my aunt Zeynep, who has been like another mother to me, always succeeded in coming up with ways to cheer me up and keep me motivated. This work would not have been possible without their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cells that make up organisms are the smallest units of life. The organism carries out its vital functions through transcription, translation and replication of the genetic material found in every living cell, deoxyribonucleic acid (DNA). For centuries, humans strove to find the processes that construct the universe and everything in it, including themselves. The journey of humankind to find their genetic make-up started in the 1870s with the concept of macromolecules. Then, came theories about nucleic acids and their link to proteins and genes [6], until Watson & Crick correctly elucidated the double-helix structure of DNA in 1953 [7]. The building blocks that hold this structure together are the nucleotides, which are composed of 3 parts: A 5-Carbon sugar, a phosphate group, and a nitrogenous base (nucleobase). While the sugar and phosphate group are identical to their respective components in each nucleotide in a DNA molecule, the nucleobases can be either of the 4 different types in a nucleotide: Adenine, Thymine, Guanine or Cytosine. Permutations of the nucleotides made up of these 4 types of bases within the DNA molecule of each organism determine their unique characteristics. The materialization of those characteristics occur through the transcription, translation and replication processes of DNA that utilize RNA and proteins within cells [8].

After establishing that the order of nucleobases in genetic material dictates

the genetic and biochemical characteristics of organisms, the significance of determining this sequence and understanding its implications became clear [9].

1.1 Background information

1.1.1 Sequencing

Genetic Sequencing has been introduced as a means to determine the order of nucleobases in a chain of nucleic acid (DNA or RNA). The initial attempts of First-generation DNA sequencing occurred during the 1970's with parallel efforts of various researchers and laboratories. In 1977, Frederick Sanger developed Sanger sequencing, which, then became the most commonly used technology for years to come [9].

Second-generation (i.e. Next-generation) sequencing differed from the previous technologies in the method used to identify nucleobases which allowed real-time analysis. The biotechnology company 454 Life Sciences produced the first major successful commercial next-generation sequencing machine with significantly improved throughput. However, the Illumina sequencing platform became the most successful among the next-generation sequencing technologies, as well as reducing the initial significantly large cost of sequencing [9].

The progression from second to third-generation sequencing was considered to be the capability to sequence single molecules and eliminate the need for DNA amplification [9]. The SMRT platform by Pacific Biosciences and nanopore sequencers offered by Oxford Nanopore Technologies (ONT) became the two leaders of third-generation sequencing.

1.1.1.1 Second-generation Sequencing Technologies

Sequencing by Synthesis - Illumina: The Illumina sequencing workflow includes 3 steps: Library preparation, sequencing, and data analysis. The library preparation stage ensures the compatibility of the genetic material sample with the sequencing machine. It typically involves fragmenting the sample (DNA/RNA) and adding adapters to the fragment ends. For the next stage of the Illumina sequencing workflow, each fragment is isothermally multiplied and purified.

Illumina sequencing by synthesis is a massively parallel process that works by adding fluorescently tagged nucleotides to the sample DNA/RNA strand and identifying the sequence of bases (reads) depending on the fluorescent signal produced at each addition. After recording read data for the forward strand, the same process repeats for the reverse strand. This method produces paired-end reads that are ~ 150 base pairs (bps) in length.

```
@ERR194147.2493233 HSQ1004:134:C0D8DACXX:2:1108:9660:77816/1
TGATGTCACACTCCTGGGCTTAAGCAATCCTCTTGTCTCCACCCCCCACAGTGCTAGGAGTGAGCCACCATGCCTGGCCCAGTTTTTAATGCTTGGACATC
+
CCCFFFFFHHGFHIIJJIJJIJJIIJIGDHIJIEFGGGIJJGGGGIJIJIIHGGGHE;?C=;=@C@>BACCEEDDCDDDDDD@@@>BBCCCACCCDAC@@:
```

Figure 1.1: A single read from the FASTQ file of the public genome NA12878. Line 1 denotes the sequence identifier. Line 2 shows the sequence in terms of nucleobase symbols. Line 3 is a separator, the plus (+) sign. Line 4 is allocated for the base call quality scores.

Illumina (and many other) FASTQ files that consist of hundreds to millions of such lines, are most commonly utilized in two ways: 1) Performing genome assembly, a computational process of reconstructing the genome from the reads, or 2) By read alignment to a known reference.

1.1.1.2 Third-generation Sequencing Technologies

The two leading third-generation sequencing technologies, SMRT Sequencing by Pacific Biosciences and Nanopore Sequencing by Oxford Nanopore Technologies, have been rapidly evolving since their first introduction.

The latest innovation from Pacific Biosciences, HiFi sequencing, allowed the successful production of long-reads with high ($> 99\%$) accuracy. Oxford Nanopore Technologies also very recently improved their nanopore sequencing method from having error rates of $\sim 20 - 25\%$ to as low as 4%.

Therefore, both technologies have successfully produced the long-awaited highly accurate long-read data whose details are explained in the sections below.

SMRT Sequencing by Pacific Biosciences :

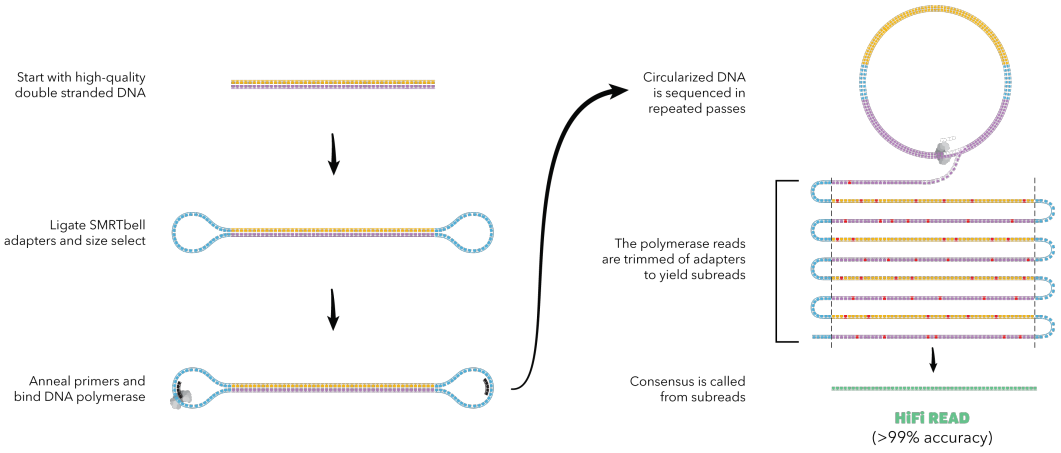


Figure 1.2: Pacific Biosciences High Fidelity read generation flow. Adapted from [1].

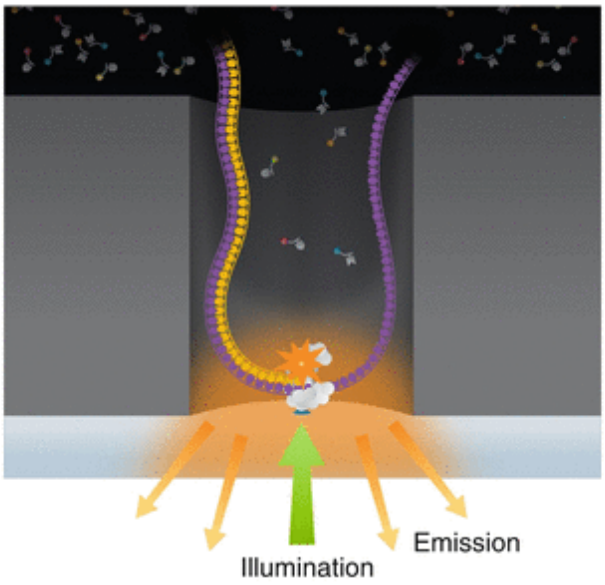


Figure 1.3: *Zero-mode waveguides (ZMWs)* SMRT sequencing works using ZMWs (Zero mode waveguides), which resemble tiny wells. Single DNA molecules are immobilized in the ZMWs before DNA polymerase begins DNA synthesis. Then, it incorporates labeled nucleotides to the growing chain which causes light to be emitted. SMRT sequencing has two modes: Continuous long-read (CLR) mode or Circular Consensus Sequencing (CCS) mode [2]. Adapted from [3]

Nanopore Sequencing by Oxford Nanopore Technologies :

In nanopore sequencing, a protein nanopore is set in an electrically resistant polymer membrane. An ionic current is passed through the nanopore where the passage of the DNA strand creates a disruption in current. Measurement of that current makes it possible to identify individual bases. Nanopore sequencing also requires no amplification and can be parallelized [4].

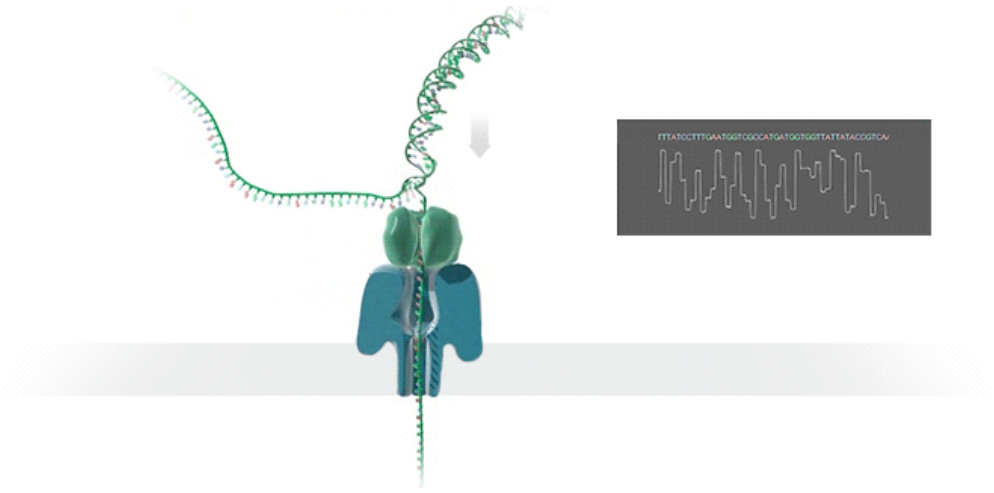


Figure 1.4: Oxford Nanopore Technologies nanopore sequencing overview. Adapted from [4].

1.1.1.3 Sequencing drawbacks

Second-generation sequencing devices primarily generate highly accurate ($\sim$ 99.9%) short reads ($\sim$ 150 bps) with relatively low cost. While high accuracy is desired for most areas of use of sequencing, read length also plays a crucial role for several applications, such as detecting variations in a sequence of nucleobases w.r.t another known sequence.

1.1.2 Structural variations

Briefly defined as genomic alteration larger than 50 base pairs, structural variations (SVs) occur in the form of deletions, insertions, inversions, duplications, transpositions, and translocations. SVs are associated with gene expression variation [10], female fertility [11], susceptibility to HIV infection [12], systemic autoimmunity [13], and genomic disorders like Williams-Beuren syndrome and velocardiofacial syndrome [14, 15]. Therefore it is of utmost importance to characterize the full extent of structural variation to understand phenotypic variation and genetic causes of disease in humans [16]. For accurate SV detection, long and highly accurate reads are needed.

Deletions

Deletions are a type of copy number variation (CNV) where a number of base pairs are missing from the individual's genome with respect to the reference genome of the species.

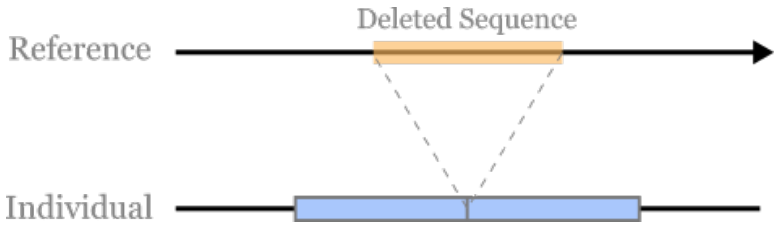


Figure 1.5: Deletion w.r.t the reference genome. Adapted from [5].

Inversions

Inversions are portions of the individual's genome that have inverted in-place with respect to the reference genome of the species.

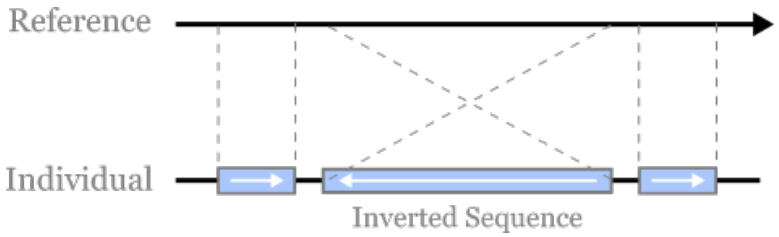


Figure 1.6: Inversion w.r.t the reference genome. Adapted from [5].

Although the effects of inversions on human disease are still not clearly known, they have been recognized to partake in some types of cancer [17] [18], and other diseases. However, due to the ineptitude of previous sequencing technologies, the actual functional effects of inversions on human disease is still not clear. We may now be able to characterize inversions contributing to disease more successfully with whole genome sequencing since genes truncated by a break point will be detectable as opposed to other methods such as exome sequencing [19]. This suggests that structural variation identification is critical for human health among many other areas.

Duplications

Duplications describe the phenomenon that some part of the genome is repeated and pasted in between nucleotides of either the same or a different chromosome. The original copy is in alignment with the reference genome while the new copies are out of place.

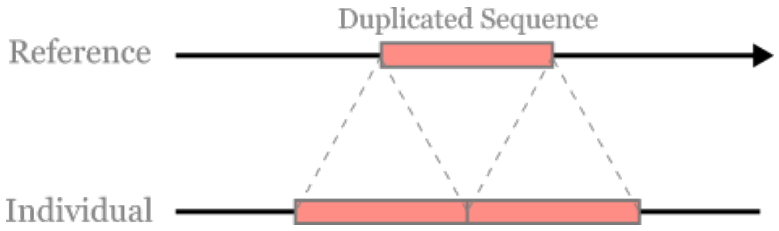


Figure 1.7: Tandem duplication w.r.t the reference genome. Adapted from [5].

There are two classes of duplications. Tandem duplications are CNVs whose copied portions are pasted immediately after the original sequence, and therefore, all copies are in succession. Interspersed duplications comprise all cases outside of tandem duplications, where the copies are located in different regions from the original sequence and each other.

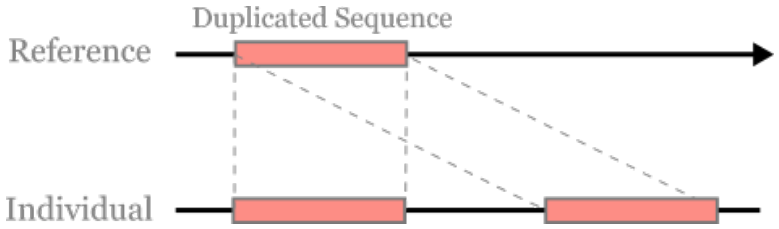


Figure 1.8: Interspersed duplication w.r.t the reference genome. Adapted from [5].

1.2 Existing algorithms for the detection of structural variations

Previously it was possible to characterize only large copy number variations (i.e., deletions and duplications - CNVs) using array comparative genomic hybridization (array CGH) with bacterial artificial chromosome arrays [20], and

then oligonucleotide microarrays [21, 22, 23]. Additionally SNP genotyping arrays were also used to identify CNVs by analyzing B-allele frequencies [24, 25]. Detecting inversions and insertions became possible using fosmid end sequencing [26, 27, 28], and smaller insertions and deletions (indels) were detected by mining alignments of whole genome shotgun (WGS) sequences [29] generated using Sanger technology. However, it remained unfeasible to apply these methods to characterize many genomes due to the high costs.

The cost of WGS data generation decreased substantially after the introduction of high throughput sequencing (HTS) technologies [30]. This led to the development of a plethora of novel algorithms to discover various forms of SVs, although most focused on CNVs. Several tools were extensively used to generate the 1000 Genomes Project SV call set [31, 32, 33]. Among many such algorithms, mrCaNaVaR is able to predict large segmental duplications, deletions and absolute copy numbers of genomic intervals by utilizing read depth information [34]. VariationHunter presents combinatorial algorithms to discover SVs using Next Generation paired-end reads [35], Pindel proposes a pattern growth approach to detect SV breakpoints at single-based resolution from paired-end short reads [36], DELLY uses several signals in paired-end short read data such as short insert paired-ends, long range mate-pairs and split-read alignments [37]. Further algorithmic development made it possible to discover mobile element insertions [38, 39, 40], novel sequence insertions [41], and satellite expansions and contractions [42]. More modern algorithms such as LUMPY [43] and TARDIS [44] aim to characterize several types of SVs simultaneously. Although PBSV, the tool for structural variation discovery developed by Pacific Biosciences, is compatible with HiFi reads, according to the developer, it is most effective for deletions of size up to 100kb, and insertions, inversions and duplications of size up to 10 kb. No additional information on the performance of the tool for SVs with larger sizes could be obtained.

HTS technologies made SV detection easier and more comprehensive than before, yet in addition to the complex nature of genomes, there were still many challenges due to the read lengths and errors associated with these technologies until very recently. For accurate SV detection, long reads with very low error rate

are desired, however, there is a trade-off between read lengths and error rates in most HTS technologies. While the reads generated with Oxford Nanopore sequencers [45] and Pacific Biosciences devices prior to CCS [46] were sufficiently long to discover many types of SVs, the high error rates (12-20%) introduced further challenges in read mapping. This, in turn, made it extremely difficult to find the exact SV breakpoint. On the other hand, Illumina reads have been very accurate with low (<0.1%) error rates, however, they are too short to be confidently mapped to repeat and duplication rich areas of the genomes that are known to harbor most of the SVs [27].

With the emergence of Single Molecule High Fidelity (HiFi) Sequencing by Pacific Biosciences, error rates during read generation have been significantly lowered without compromising read length (10-20 kb). With the potential to provide adequate information for highly accurate and precise SV detection, HiFi sequencing data has been becoming more and more widespread.

1.3 DALEK

The contributions of this thesis are two algorithms for structural variation discovery, DALEK Version I and DALEK Version II.

DALEK Version I is a novel algorithm that aims to leverage complementary strengths of CLR and short read technologies to help correct the different biases inherent in these platforms. DALEK I integrates split read and read depth signatures observed in PacBio continuous long reads that signal an SV event with paired-end signature seen in short read (i.e. Illumina) data to discover large deletions (> 10 Kbp) and inversions (> 50 Kbp) of size up to 10 Mbp. We show that DALEK I can re-identify very large (up to 6 Mbps) inversions in the genome of NA12878 even with low (5X) PacBio depth of coverage.

DALEK Version II uses long reads with several degrees of error to discover

large SVs with high break point resolution. DALEK II examines split read signatures from PacBio CCS or ONT data and filters out potential false discoveries using read depth information. DALEK II discovers both tandem and interspersed segmental duplications (> 50 Kbp) in addition to the large SVs DALEK I is able to detect.

The next chapter of this thesis contains a detailed description of both algorithms, while their individual performance and the quality of the output using real and simulated data sets are assessed in the results chapters. The next chapter includes the discussion of the aforementioned results and the final one a mention of future work.

Chapter 2

Methods

Both DALEK I and II use a combination of split read and read depth signatures with DALEK I additionaly utilizing read pair information [5, 47] to identify SVs. DALEK I relies on the availability of PacBio CLR and Illumina alignments while DALEK II requires BAM files for any one of the PacBio CCS or Nanopore data sets. Additionally, they can use the within-read alignment gaps reported in the CIGAR field [48] for each alignment for the PacBio (or Nanopore for DALEK II) data to find smaller SVs.

Although DALEK I and II are agnostic to the read aligners, in this work we use BLASR [49] for PacBio CLR mapping, Minimap 2 [50] for PacBio CCS and Nanopore mappings, and BWA-MEM [51] for Illumina alignments. We show an overview of the DALEK I and II methodologies in Figure 2.1 and Figure 2.2 respectively.

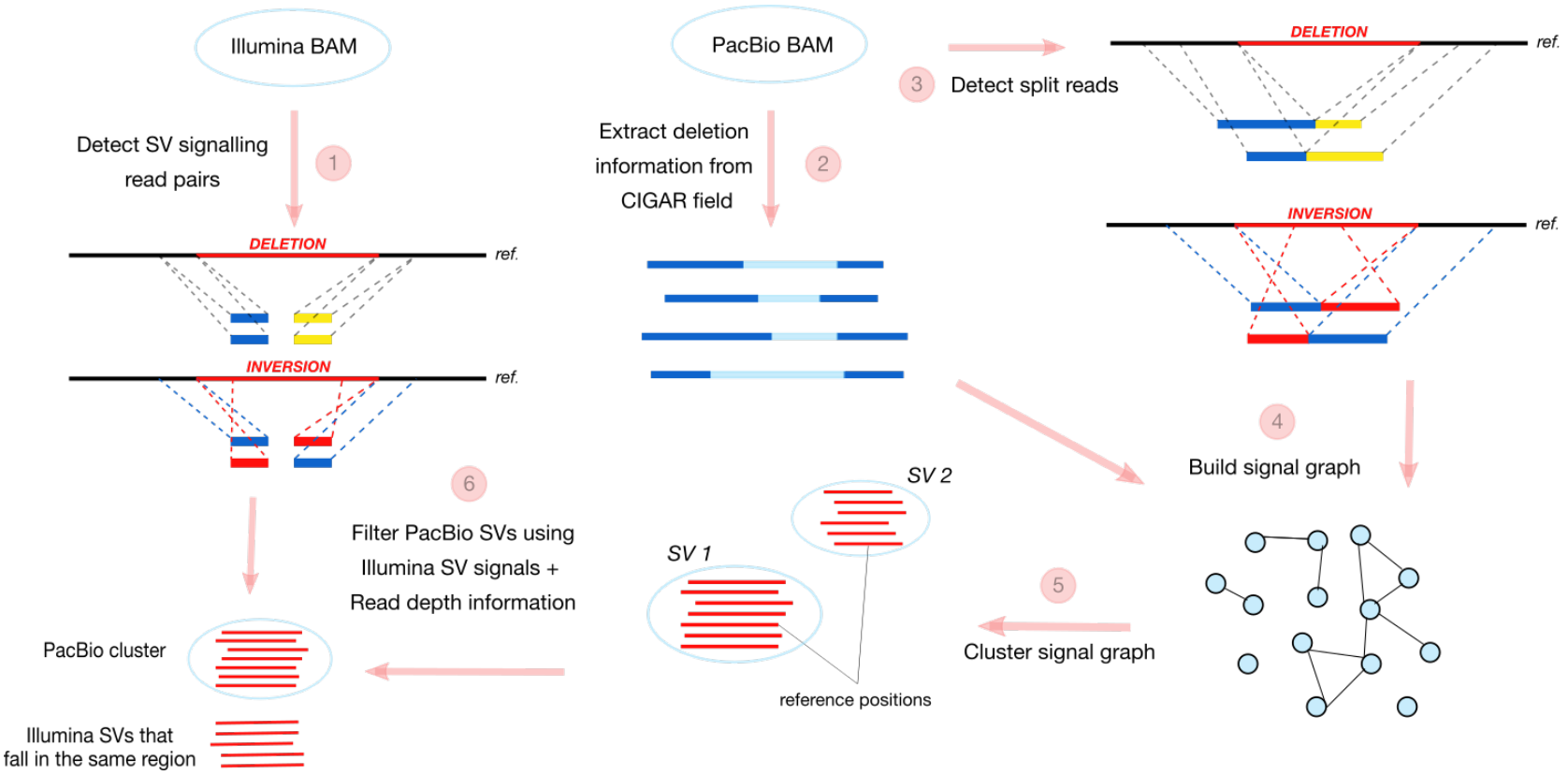


Figure 2.1: Overview of the DALEK I algorithm. 1) Two Illumina read-pairs signaling the same deletion (top) and two Illumina read-pairs signaling a single inversion (bottom). Discordant read pairs are selected as SV signals. 2) PacBio reads with deletions reported in their alignments. Dark blue shows the matching segments and light blue shows the deletion as reported in CIGAR. 3) Two PacBio reads signaling the same deletion (top) and two PacBio reads signaling a single inversion(bottom). 4) Nodes represent SV intervals and an edge is created between two nodes if they have 50% reciprocal overlap. 5) Circles denote different clusters. Each red line is a continuous interval in the reference genome. SV intervals (nodes) are clustered together if they satisfy the (λ, γ)-quasi-clique conditions. 6) PacBio SVs are filtered using Illumina SV signals and read depth information. DALEK I checks whether it can add an edge between an Illumina SV and a PacBio SV cluster centroid. If there is an edge between a PacBio SV cluster centeroid and a number of (depending on the SV type) Illumina SV intervals, and the median copy number of that interval is $\leq$1.5, that cluster is considered a valid SV. Breakpoints are determined using the average of the PacBio map locations.

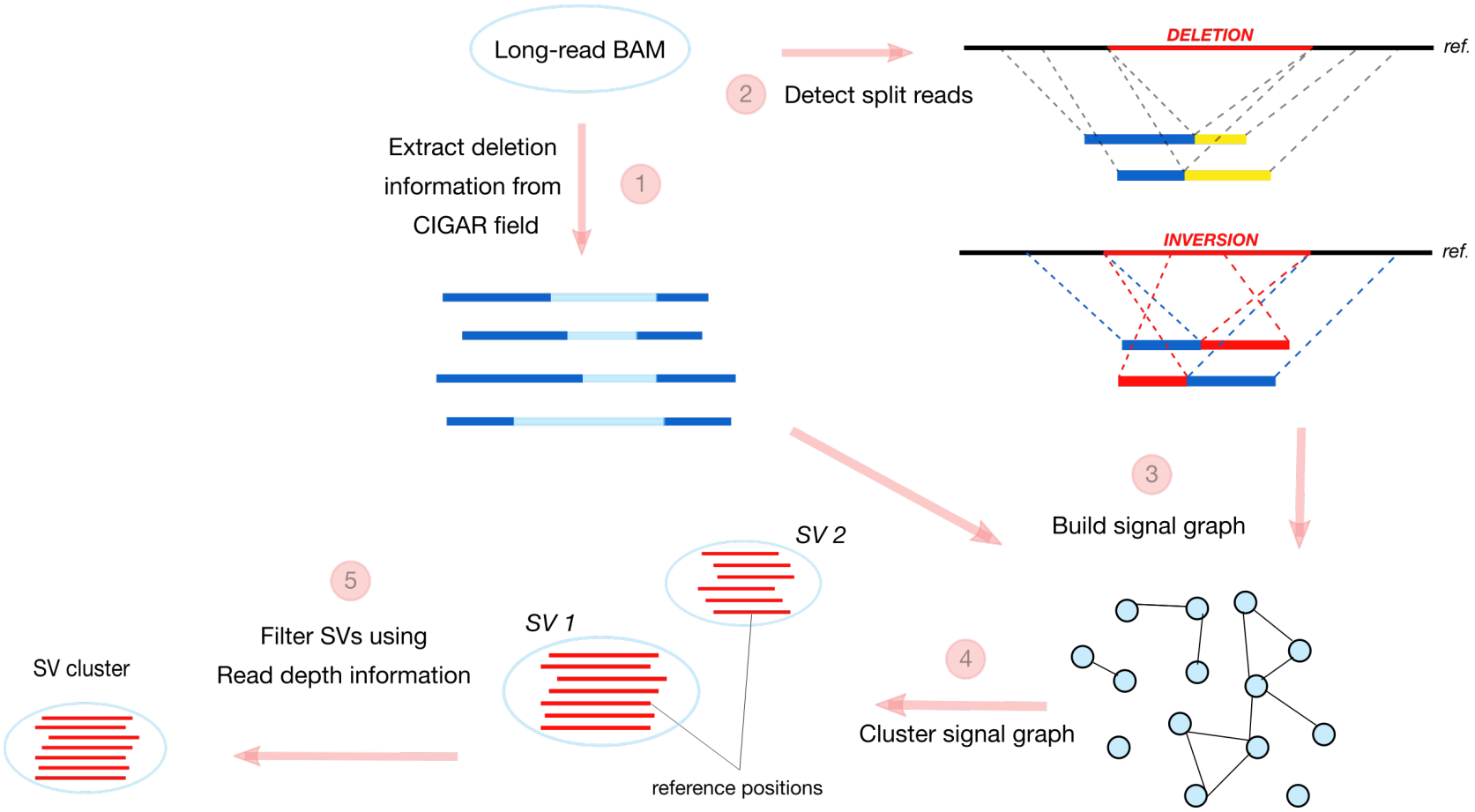


Figure 2.2: Overview of the DALEK II algorithm. 1) PacBio/Nanopore reads with deletions reported in their alignments. Dark blue shows the matching segments and light blue shows the deletion as reported in CIGAR. 2) Two long reads signaling the same deletion (top), and two long reads signaling a single inversion (bottom). 4) SV signal graph. Nodes represent SV intervals and an edge is created between two nodes if they have 50% reciprocal overlap. 5) Circles denote different clusters. Each red line is a continuous interval in the reference genome that was previously selected as an SV signal. SV intervals (nodes) are clustered together if they satisfy the (λ, γ)-quasi-clique conditions.

2.1 SV signal identification and clustering in PacBio alignments

Both DALEK I and II first analyze PacBio alignments, and 1) extract the deletion information from CIGAR field (i.e, operation 'D'), and 2) identify those reads with multiple alternative alignments with soft or hard clips (i.e., operations 'S' and 'H') that are in agreement with each other as *split* (Figure 2.3, Figure 2.5, Figure 2.6, Figure 2.7). Here we only consider the deletion, inversion and segmental duplication signaling split reads. To reduce misalignment artifacts, we do not consider reads that map to satellite regions or segmental duplications in the reference genome. We set default minimum length for deletions as 10 Kbp and inversions and duplications as 50 Kbp.

2.1.1 Signals for different types of SVs

Deletion signals

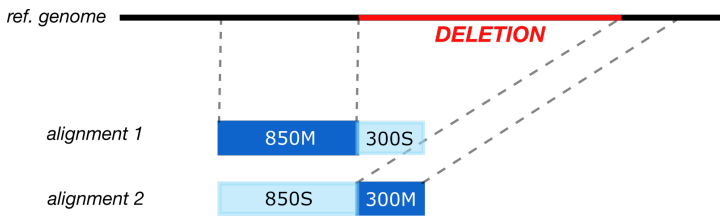


Figure 2.3: Split read signatures used in DALEK I and II. Larger deletions inferred from a split read. The 850 bp prefix of the read is mapped to the proximal break point of the deletion, and the 300 bp suffix of the read is mapped to the distal break point. DALEK I and II track this information using the soft clip annotation in the CIGAR fields of both alignments.

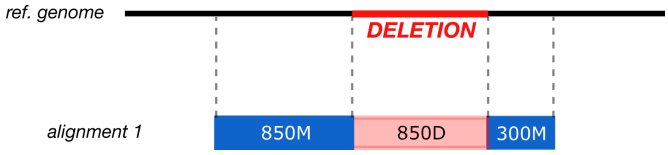


Figure 2.4: Split read signatures used in DALEK I and II. Smaller deletion inferred from CIGAR. A 850 bp deletion reported by the aligner in CIGAR field.

Inversion signals

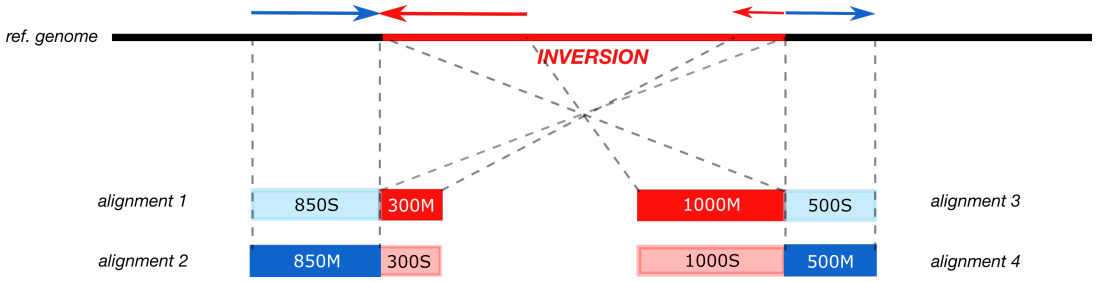


Figure 2.5: Split read signatures used in DALEK I and II. A large inversion marked with two split reads. The 850 bp prefix of the leftmost read is mapped to the proximal break point of the inversion, and the 300 bp suffix of the read is mapped to the distal break point after being inverted. The 500 bp suffix of the rightmost read is mapped to the proximal break point of the inversion, and the 1000 bp prefix of the read is mapped to the distal break point after being inverted.

Segmental duplication signals

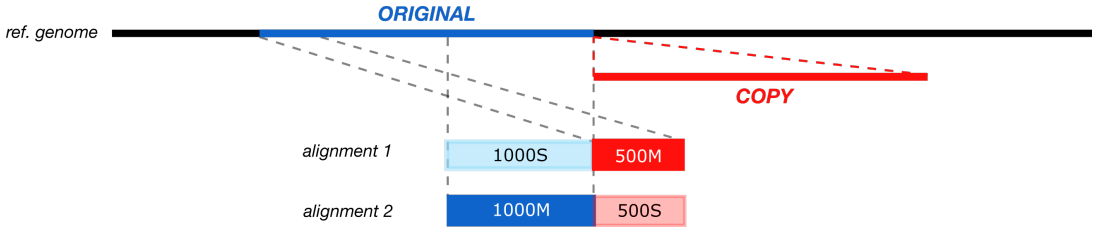


Figure 2.6: Split read signatures used in DALEK II. A large tandem duplication marked with one split read. The 1000 bp prefix of the read is mapped to the proximal break point of the copy, while the 500 bp suffix of the read is mapped to the distal edge of the original sequence.

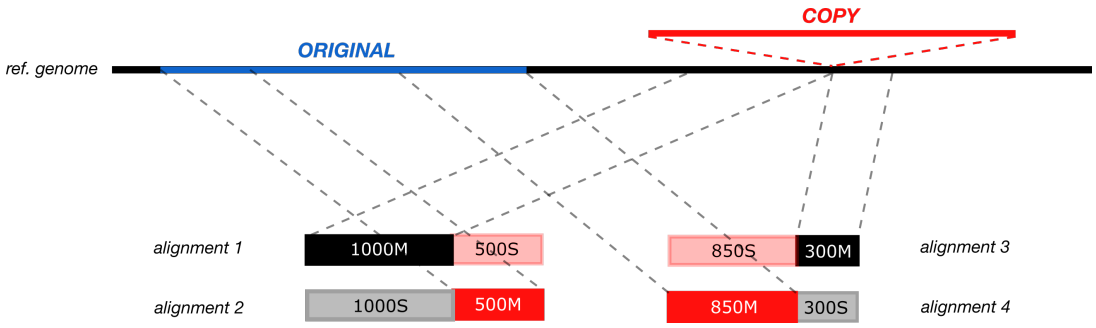


Figure 2.7: Split read signatures used in DALEK II. A large interspersed duplication marked with two split reads. The 1000 bp prefix of the leftmost read and the 300 bp suffix of the rightmost read are mapped to the break point of the copy in order. The 500 bp suffix of the leftmost read is mapped to the beginning while the 850 bp prefix of the rightmost read is mapped to the end of the original sequence.

2.1.2 Clustering

After the initial extraction of CIGAR and split-read driven candidate SV intervals, we build a *signal graph* as follows. We denote each interval as a node, and we create an edge between two nodes if they are likely to represent the same SV event. Formally, we define an SV interval as a 3-tuple $SV_k = (s_k, e_k, t_k)$, where s_k and e_k are the start and end locations, and t_k denotes the SV type (i.e. deletion or inversion). We consider two SV intervals SV_i and SV_j represent the same event if $t_i = t_j$ and there is >50% reciprocal overlap between intervals.

We then use an approximation algorithm to the maximum clique problem, which instead tries to find (λ, γ)-quasi-cliques [52] to identify clusters of SV intervals that show a putative structural variation. Note that quasi-clique is a relaxation on the maximum clique problem, which is known to be NP-Complete [53], and the algorithm by Brunato et al. (2008) extends upon stochastic local search algorithms originally developed for the maximal clique problem [54, 55].

Given an undirected graph $G = (V, E)$, a subgraph $G' = (V', E')$ is a (λ, γ)-quasi-clique if:

$$\forall v \in V', degree_{V'}(v) \geq \lambda \cdot (|V'| - 1) \tag{2.1}$$

$$|E'| \geq \gamma \cdot \binom{|V'|}{2} \tag{2.2}$$

We set $\lambda = 0.5, \gamma = 0.6$ for deletions, and $\lambda = 0.2, \gamma = 0.3$ for inversions in DALEK I by default, which we set after a parameter sweeping experiment. We also calculate the average of each breakpoint of the SVs that make up a cluster. A hypothetical SV with said breakpoints serves as a representative for that cluster, which is used as a means of comparison in the next stages of the algorithm.

2.2 Read pair support with Illumina alignments

Read pair support is only used in the DALEK I algorithm as it utilizes hybrid data while DALEK II only uses long-read data. Hence, as per VariationHunter [35], we first calculate the average size (μ) and standard deviation (σ) of fragment lengths of the Illumina paired-end data. We then discard concordant read pairs where the span (ℓ) is within three standard deviations of the average (i.e. $\mu - 3 \cdot \sigma \leq \ell \leq \mu + 3 \cdot \sigma$). DALEK I then considers the remaining read pairs as discordant and flags them with the type of structural variation, following the standard formulations [35, 47]. There is, however, one exception to the detection of the final Illumina inversions, where we consider an inversion a single read pair (instead of two) that satisfies the above conditions. This approach is favored instead of matching two types of inversion signals (forward pairs and reverse pairs) with each other since it reduces time complexity significantly. In order to adapt the algorithm to this approach, we adjusted the filtering parameters as explained in the next paragraph. Similar to the SV interval definition above, we define an Illumina read pair that signals an SV event as $P_k = (s_k, e_k, t_k)$, where s_k, e_k are start and end locations, and t_k is the SV type. Next, DALEK I hierarchically adds an Illumina read pair P_i to an SV cluster SV_j (calculated in the previous step) if it does not violate the same (λ, γ)-quasi-clique properties.

Finally, DALEK I retains those deletion clusters with at least 20 supporting Illumina read pairs and inversion clusters with at least 70 supporting Illumina read pairs as putative SVs. We set a higher requirement for inversion read pair support because of the higher repeat and duplication content at the breakpoints of balanced rearrangements [27], and because we consider each inversion twice as a forward and a split pair. Note that, if the PacBio reads and their respective alignments were perfect, the split breakpoints would be identical to the inferred SV. However, since the alignments may be inaccurate because of both the high (12-15%) error rate and the repeats around the breakpoints, DALEK I uses the average of the PacBio map locations as the estimated breakpoints.

2.3 Read depth filtering for deletions and SDs

To improve accuracy in the detevtion of copy number variations (CNVs), both DALEK I and II apply a simple read depth based strategy to filter out false positives using PacBio alignments. DALEK I uses the read-depth information for deletions, while DALEK II applies it to both deletion and duplication calls. Briefly, DALEK first calculates read depth *per base* across the genome. Since PacBio Single Molecule Real Time sequencing does not show GC% bias [56], we do not apply any read depth correction.

Let $\mathcal{M}$ denote the average of overall genome read depth and d_i be the depth of the window with genome start index i. As per [34], we calculate the *diploid copy number* of the window starting at position i as:

$$CN_i = 2 \cdot \frac{d_i}{\mathcal{M}}$$

Finally, for any deletion or duplication interval identified in the previous step, we calculate the median copy number of the windows contained within. We discard such SV predictions if the median copy number is ≥ 2 (in practice, > 2.5 to avoid excessive filtering due to errors in PacBio reads for DALEK I).

Chapter 3

DALEK I Results

We tested DALEK I on both simulated data sets and real PacBio and Illumina sequencing data from the genome of NA12878. As the PacBio data set, we used the FASTA files released by the Genome in a Bottle Project [57] at 5X coverage and realigned the reads to human reference genome GRCh38 using the BLASR [49] aligner. We downloaded the Illumina BAM files (50X depth of coverage) from the Platinum Genomes Project [58], aligned to the same genome reference using BWA-MEM [51]. We compared the prediction accuracy of DALEK I with Sniffles [59], and other tools such as DELLY [37] and LUMPY [43] that predict SVs from only the short read WGS data to demonstrate the additional power gained by long read sequencing.

We used the 1000 Genomes Phase 3 release [32] for structural variation as the gold standard for NA12878 deletions to calculate sensitivity and false discovery rate (FDR). For large inversions, we used the InvFEST database [60], and previous studies with experimental validation [61, 62, 63] as the truth set.

3.1 Simulation experiments

We used VarSim [64] to insert 1,755 deletions, 2,245 insertions, 459 inversions, 584 tandem and 260 interspersed duplications (size range 50 bp to 6 Mbp) to human reference genome GRCh37. 110/260 of the interspersed duplications were also inverted. We also included >2.8 million SNPs and ∼194,000 small indels in the simulation. We then generated Illumina reads at 40X coverage using ART [65], and PacBio reads at 10X coverage using PBSIM [66]. We aligned the PacBio simulation using both BLASR and NGM-LR [59], and Illumina simulation using BWA-MEM to GRCh37.

We used DALEK I to detect deletions (>10 Kb) and inversions (>50 Kb) using both Illumina and PacBio data sets simultaneously. To compare DALEK I's performance with other state-of-the-art tools, we ran DELLY and LUMPY using the Illumina data set, and Sniffles on the PacBio data set.

DALEK I detected 373 deletions ranging from 11 Kbp to 450 Kbp, and 394 inversions ranging from 50 Kbp to 9 Mbp (Table 3.1). We found that DALEK I achieved 91% sensitivity and 55% FDR in deletion predictions. Although the sensitivity of DALEK I remained high for inversions (34%), the FDR increased to 88% due to the low coverage simulation of PacBio reads. DALEK I runs very efficiently, and it completed the analysis in only 1 hour using a single CPU and required 3 GB of memory.

3.2 Real Data

3.2.1 NA12878 genome

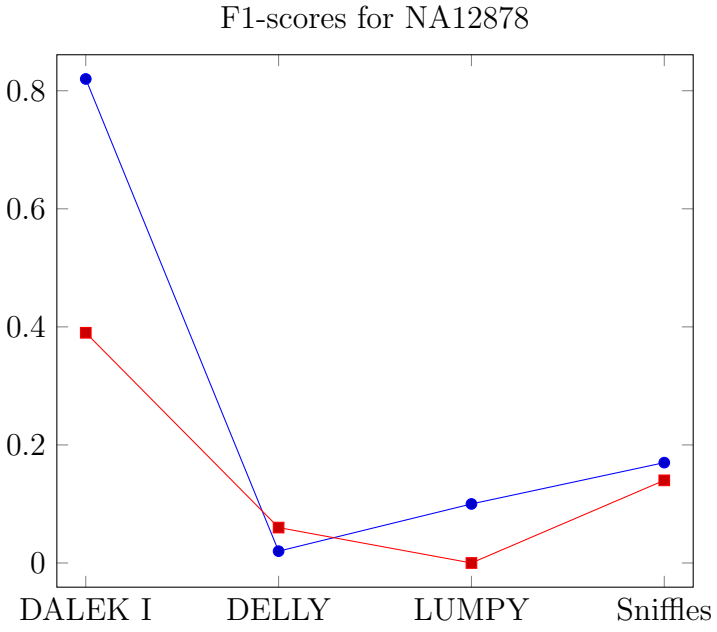
Next we tested DALEK I's performance using real data sets generated from the genome of an individual of Northern European ancestry (i.e. HapMap CEPH), NA12878. We aligned the real PacBio data using BLASR, and the real Illumina

data using BWA-MEM to GRCh38. We found all four previously validated large inversions [61, 67, 63] using DALEK I.

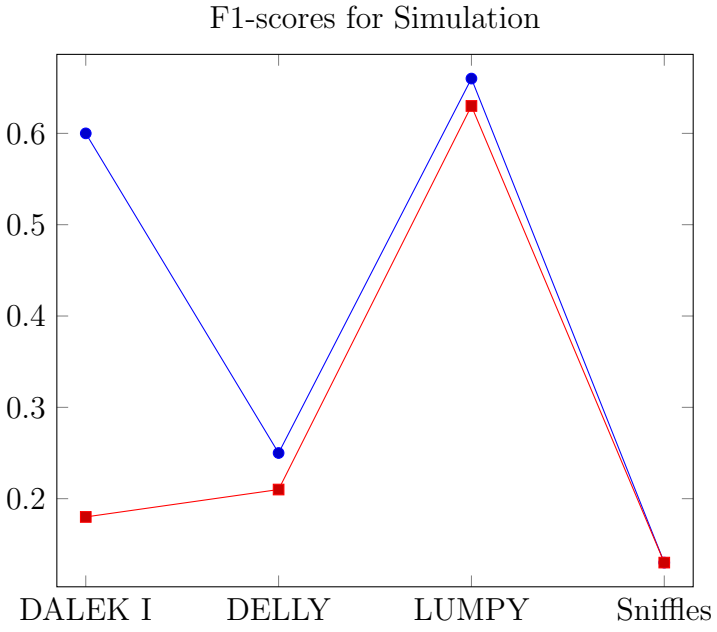
Table 3.1: Summary of prediction results using real (NA12878) and simulated human genomes.

	NA12878			Simulation		
	Predicted	**Sensitivity**	**FDR**	**Predicted**	**Sensitivity**	**FDR**
Large Deletions						
DALEK I	29	78%	13%	373	91%	55%
DELLY	4,513	51%	99%	761	44%	83%
LUMPY	522	44%	94%	480	86%	46%
Sniffles	477	60%	90%	7,331	95%	93%
Large Inversions						
DALEK I	49	40%	61%	394	34%	88%
DELLY	2,480	15%	96%	1,033	84%	88%
LUMPY	16	0%	100%	63	46%	0.05%
Sniffles	510	24%	90%	12,462	97%	83%

We present the Sensitivity and FDR estimates of DALEK I, DELLY, LUMPY, and Sniffles on both NA12878 genome and simulated data sets. Note that DALEK I uses hybrid data, where DELLY and LUMPY use only short reads, and Sniffles uses only long reads. We also note that DELLY and LUMPY do not focus on large genomic variation, therefore DALEK I provides a complementary approach. Sensitivity($\frac{TP}{TP+FN}$), FDR: false discovery rate ($\frac{FP}{TP+FP}$).



The above plot visualizes F1-scores of the SV detection tools DALEK I, DELLY, LUMPY, and Sniffles for the NA12878 genome. Blue indicates deletions and red indicates inversions.



The above plot visualizes F1-scores of the SV detection tools DALEK I, DELLY, LUMPY, and Sniffles for the simulated genome. Blue indicates deletions and red indicates inversions.

We predicted deletions (>10 Kbp) in the genome of NA12878 using DALEK I. We considered the deletions reported by the 1000 Genomes Project [32] to be the gold standard when calculating Sensitivity and false discovery rate (FDR). For large (>50 Kbp) inversions, we used the InvFEST database for this purpose. We compared our results for both the real and simulated datasets to the predictions of DELLY, LUMPY and Sniffles in Table3.1.

In summary, DALEK I detected 29 deletions (>10 Kb) and 49 inversions (>50 Kb) for the NA12878 data set. The 1000 Genomes Project release included 37 deletions of the same size range, where DALEK I predicted 25/29 correctly, achieving 78% sensitivity and 13% false discovery rate for deletions.

On the simulated genome, DALEK I predicted 373 deletions (>10 Kb) and

Table 3.2: Experimentally validated large inversions detected by DALEK I.

	Validated		**DALEK I prediction**	
Chrom	**start**	**end**	**start**	**end**
8	8,239,446	11,922,365	9,059,658	10,581,083
15	30,618,102	32,153,207	30,469,697	32,468,604
16	16,210,619	18,592,220	15,543,271	18,541,799
17	36,446,544	37,890,227	36,156,449	38,314,581

We require >50% reciprocal overlap for a prediction to be called as true. DALEK I is able to accurately predict **all** large (>1.5 Mbp) inversions that were previously experimentally validated [61, 67, 63]

394 inversions (>50 Kb). DALEK I achieved a higher F1-score for both real and simulated deletions. LUMPY performed the best for all SVs in the simulated genome. Although DALEK I outperformed all other tools for the discovery of inversions within the real genome, it failed to do so for the simulated data. This is most likely due to the default graph assignment parameters. As DALEK I becomes less strict in considering cluster sizes while building the signal graph, Sensitivity for inversions is expected to improve accordingly. However, DALEK I should ideally be able to perform consistently with similar parameters across different genomes. In order to test whether this problem is caused by the specific real genome used or not, we plan on conducting tests with other human genomes in the future as well.

DALEK I also makes far less predictions compared to the other tools for any SV in any genome. This may be due to the fact that Illumina data used for DELLY and LUMPY is significantly higher coverage (30X) than the PacBio CLR data (5X) used for primary SV detection for DALEK I. Also, we ran sniffles with default parameters as advised by the authors and that most likely affected the number of predictions it makes. There were no specifications for adjusting parameters based on coverage, therefore we used the default values for this evaluation.

It is difficult to assess the inversion false discovery rate in this genome since we did not perform experimental validation and no gold standard exists for NA12878 inversions. However, confirmed and unconfirmed inversions from the InvFEST

database were used as a means of assessment of DALEK I and comparison with the other tools. As these results show, DALEK I outperforms all other tools in real data. DALEK I also correctly re-identified 4 out of 4 previously validated large inversions [61, 62, 63].

Table 3.3: Run times of the tools we tested on the simulated genome predictions.

Tool	DALEK I	DELLY	LUMPY	Sniffles
Run time (s)	3677	11400	12060	3060

We used UNIX *time command* to calculate run times of each SV tool on the simulated genome. DALEK I and Sniffles are the fastest tools and their run times are comparable.

Chapter 4

DALEK II Results

DALEK II was tested on both simulated data sets and real PacBio High Fidelity sequencing data from the genome of NA19238. As the NA19238 data set, we used the FASTA files released by the 1000Genomes Consortium at 67X coverage and realigned the reads to human reference genome GRCh38 using the Minimap2 [50] aligner. We compared the prediction accuracy of DALEK II with Sniffles [59], PBSV [68], and other tools such as DELLY [37] and LUMPY [43] that predict SVs from only the short read WGS data to demonstrate the additional power gained by long read sequencing again.

Deletions and duplications were validated using dbVar non-redundant callset. We also used gnomAD v2.1.1 truth set for inversions.

4.1 Simulation experiments

For the evaluation of DALEK II, we used the same simulation as DALEK I. We inserted 1,755 deletions, 2,245 insertions, 459 inversions, 584 tandem and 260 interspersed duplications (size range 50 bp to 6 Mbp) to human reference genome GRCh37 using VarSim [64]. 110/260 of the interspersed duplications were inverted and we included >2.8 million SNPs and ∼194,000 small indels in

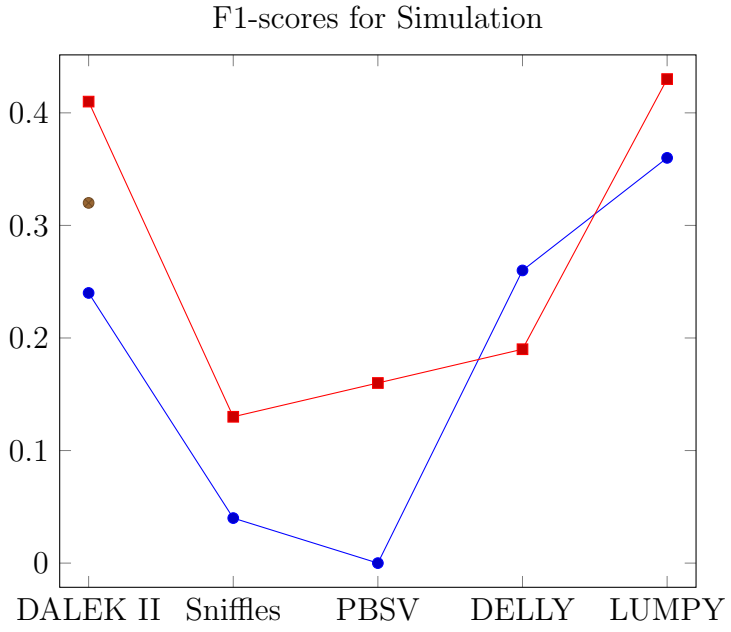
the simulation. We then generated PacBio CCS reads at 38X coverage using PBSIM [66]. We aligned the PacBio simulation using Minimap2 [69] to GRCh37.

We used DALEK II to detect deletions (>100 Kb), inversions (>80 Kb) and duplications (>50 Kb) using HiFi data. To compare DALEK II's performance with other state-of-the-art tools, we ran Sniffles, and PBSV on the PacBio data set and used the previous DELLY and LUMPY results.

Table 4.1: Summary of prediction results using a simulated genome.

	Simulation		
	Predicted	**Sensitivity**	**FDR**
Large Deletions			
DALEK II	23	28%	13%
Sniffles	2	2%	0%
PBSV	0	0%	0%
DELLY	496	85%	85%
LUMPY	292	78%	77%
Large Inversions			
DALEK II	30	26%	7%
Sniffles	9	7%	11%
PBSV	13	9%	23%
DELLY	358	43%	88%
LUMPY	35	30%	21%
Large Duplications			
DALEK II	38	20%	16%

We present the Sensitivity and FDR estimates of DALEK II, Sniffles, PBSV, DELLY, and LUMPY on simulated data sets. Note that DALEK II uses PacBio HiFi data, where DELLY and LUMPY use only short reads. We also note that DELLY and LUMPY do not focus on large genomic variation, therefore DALEK II provides a complementary approach. Sniffles, PBSV, and DeepVariant can perform using HiFi as well. Among all tools, only DALEK II is able to call interspersed duplications. Sensitivity($\frac{TP}{TP+FN}$), FDR: false discovery rate ($\frac{FP}{TP+FP}$).



The above plot visualizes F1-scores of the SV detection tools DALEK II, Sniffles, PBSV, DELLY, and LUMPY for the Simulated genome. Blue indicates deletions, red indicates inversions and brown indicates duplications (It is a single point in this graph).

On the simulated genome, DALEK II predicted 23 deletions (>100 Kb), 30 inversions (>80 Kb) and 38 duplications (>50 Kb).

4.2 Real Data

We tested DALEK II's performance using real data set generated from the genome of the mother of the Yoruba Trio in Ibadan, Nigeria (International Hapmap Project, NA19238). We aligned the real PacBio data using Minimap2 [50] to GRCh38.

We predicted deletions (>100 Kbp), inversions (>80 Kbp) and duplications (>50 Kbp) in the genome of NA19238 using DALEK II. We compared our results for the real data sets to the predictions of PBSV, DELLY, and LUMPY in

Table4.2.

Table 4.2: Summary of prediction results using real (NA19238) human data.

	NA19238		
	Predicted	**TP**	**Precision**
Large Deletions			
DALEK II	11	10	91%
PBSV	0	0	0%
DELLY	192	127	66%
LUMPY	81	49	60%
Large Inversions			
DALEK II	2	1	50%
PBSV	3	1	33%
DELLY	407	37	09%
LUMPY	3	0	0%
Large Duplications			
DALEK II	9	7	78%

We present the true positive (TP) and Precision estimates of DALEK II, PBSV, DELLY and LUMPY on the NA19238 genome. Note that DALEK II uses PacBio HiFi data, where DELLY and LUMPY use only short reads. We also note that DELLY and LUMPY do not focus on large genomic variation, therefore DALEK II provides a complementary approach. PBSV can perform using HiFi as well. Deletions are $>$ 100 kb and inversions are $>$ 80 kb. Deletions and duplications are validated using dbVar non-redundant callset (https://ftp.ncbi.nlm.nih.gov/pub/dbVar/sandbox/sv_datasets/nonredundant/). We also used gnomAD v2.1.1 truth set (https://storage.googleapis.com/gnomad-public/papers/2019sv/gnomad_v2.1_sv.sites.vcf.gz) for inversions. Precision($\frac{TP}{TP+FP}$).

To summarize, DALEK II detected 11 deletions (>100 Kb), 2 inversions (>80 Kb) and 9 duplications (>50 Kb) for the NA19238 genome.

Table 4.3: Run times of the tools we tested on the simulated genome predictions.

Tool	DALEK II	Sniffles	PBSV	DELLY	LUMPY
Run time (s)	6320	7833	4593	11400	12060

We used UNIX *time command* to calculate run times of each SV tool on the simulated genome. DALEK II is the fastest tool after PBSV and their run times are comparable.

Chapter 5

Discussion

Despite the advances made in the last decade, characterization of most forms of structural variants, especially balanced rearrangements, remains an unsolved problem. The lack of detection of all genomic variation in return limits our ability to understand the etiology of genetic diseases, as well as the evolution of species.

Difficulty of structural variation discovery stems from both genome complexity and sequence data inaccuracy, which causes ambiguities in read mapping especially when the reads are short. Long reads promise to help reduce the problems in mapping ambiguity, however some long read technologies still suffer from high error rates. Therefore using a combination of short and long reads simultaneously may increase SV characterization accuracy for specific cases.

In this thesis we presented two algorithms: DALEK I that aims to forge strengths of different sequencing technologies to improve SV discovery, and DALEK II which uses long reads with several degrees of error to discover large SVs with high break point resolution. We show that DALEK I achieves high sensitivity and low false discovery rate by using PacBio and Illumina reads generated from the same genomes even when the PacBio coverage is very low (5X). DALEK II performs on average compared to existing tools, however, it can differenciate between tandem and interspersed duplication unlike the other algorithms.

We also show that both DALEK I and II perform better in terms of sensitivity and require less computational resources when compared to several other tools for SV discovery. Furthermore, contrary to most other similar tools, they work directly on the off-the-shelf alignment (i.e. BAM) files and do not perform any realignments.

In its current form DALEK I can characterize only deletions and inversions, however we plan to extend the algorithm to discover all forms on structural variation. DALEK II will also include insertions, inverted segmental duplications and translocations in the future. Note that, both DALEK I and DALEK II can detect very large inversions up to 10 Mbp. To our knowledge, no other WGS-based SV detection tool can find such large inversions, and until now they were only accessible to very large pooled clone sequencing data or Linked-Reads [63].

Chapter 6

Future Work

6.1 Detection of other types of structural variations

Both DALEK I and II can detect deletions and inversions. DALEK II also detects tandem and interspersed duplications as it is possible to detect duplications effectively using accurate long read data with high coverage. However, there are other large structural variations that are significant in areas such as medicine and health, which include translocations, insertions and inverted duplications. The detection of these variations is more difficult due to the nature of break points in areas of the genome where translocations or inverted duplications are prominent, and also the time requirements to detect insertions. In the future, we will be extending our algorithm to be able to detect all significant large structural variations with high throughput and accuracy.

6.2 Including ONT data in our primary workflow

DALEK II performs best on Pacific Biosciences High Fidelity (HiFi) sequencing data even though it is theoretically able to detect variations using any long read data. We have tested our algorithm on HiFi and Nanopore whole genomes up to this point, however, to be able to make the same claims for Nanopore data as HiFi data, we need to conduct more tests and analyze the problem in more diverse environments. Therefore, one of our more imminent plans of action is to make Nanopore data a standard input of our algorithm as well.

6.3 Break point resolution

While DALEK I performs on next-generation short read and erroneous long read data, DALEK II relies on long read technologies with various degrees of error to detect large variations. Due to this difference in input, their mechanisms of detection also vary slightly, especially in terms of break point resolution. For DALEK I, this stage is handled with a trivial approach through the use of tolerance intervals and simple estimations. We plan on improving on this via methods such as multiple sequence alignment using the short read data. When accurate PacBio HiFi data is given as input to the DALEK II algorithm, break point resolution as high as 99% of the SV length is achieved. However, when erroneous Nanopore data is processed for variation detection, additional information is required for high quality break point resolution due to the imperfections in the original data. Although multiple sequence alignment of highly accurate short read Illumina data of the same genome would be the ideal method to determine exact break points, requiring an additional sequencing data is an additional burden on the user. Therefore, we aim to explore new methods for break point resolution while working with Nanopore data.

6.4 Experiments for parameter restriction

For DALEK I, we have had several issues with inversion discovery on the simulated genome in comparison to the real genome. In order to test whether this problem is caused by the specific real genome used or not, we plan on conducting tests with other human genomes in the future.

As for DALEK II, we aim to test it using a much larger genome pool and this and additional testing will allow us to determine the best performing parameter values across all genomes. We will be submitting a scientific paper of our work on DALEK II on March 2021 and plan on conducting all said tests by that date.

Bibliography

[1] "Hifi reads - highly accurate long-read sequencing," 2020.

[2] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, "Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, pp. 133–138, Jan 2009.

[3] J. Korlach and S. W. Turner, *Zero-Mode Waveguides*, pp. 2793–2795. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[4] "How it works," Jun 2020.

[5] 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.

[6] R. C. Olby, *The Path to the Double Helix The Discovery of DNA*. Dover Publications, 1994.

[7] L. A. Pray.

[8] "Dna is a structure that encodes biological information."

[9] J. M. Heather and B. Chain, "The sequence of sequencers: The history of sequencing dna," *Genomics*, vol. 107, no. 1, p. 1–8, 2016.

[10] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. De Grassi, C. Lee, *et al.*, "Relative impact of nucleotide and copy number variation on gene expression phenotypes," *Science*, vol. 315, no. 5813, pp. 848–853, 2007.

[11] 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.

[12] E. Gonzalez, H. Kulkarni, H. Bolivar, A. Mangano, R. Sanchez, G. Catano, R. J. Nibbs, B. I. Freedman, M. P. Quinones, M. J. Bamshad, K. K. Murthy, B. H. Rovin, W. Bradley, R. A. Clark, S. A. Anderson, R. J. O'connell, B. K. Agan, S. S. Ahuja, R. Bologna, L. Sen, M. J. Dolan, and S. K. Ahuja, "The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility," *Science*, vol. 307, pp. 1434–1440, Mar 2005.

[13] M. Fanciulli, P. J. Norsworthy, E. Petretto, R. Dong, L. Harper, L. Kamesh, J. M. Heward, S. C. Gough, A. De Smith, A. I. Blakemore, *et al.*, "FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity," *Nature genetics*, vol. 39, no. 6, p. 721, 2007.

[14] J. R. Lupski and P. Stankiewicz, "Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes.," *PLoS genetics*, vol. 1, p. e49, Dec. 2005.

[15] J. L. Freeman, G. H. Perry, L. Feuk, R. Redon, S. A. McCarroll, D. M. Altshuler, H. Aburatani, K. W. Jones, C. Tyler-Smith, M. E. Hurles, N. P. Carter, S. W. Scherer, and C. Lee, "Copy number variation: new insights in genome diversity," *Genome Res*, vol. 16, pp. 949–961, Aug 2006.

[16] 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.

[17] P. J. Campbell, S. Yachida, L. J. Mudie, P. J. Stephens, E. D. Pleasance, L. A. Stebbings, L. A. Morsberger, C. Latimer, S. Mclaren, M.-L. Lin, and et al., "The patterns and dynamics of genomic instability in metastatic pancreatic cancer," *Nature*, vol. 467, no. 7319, p. 1109–1113, 2010.

[18] C. M. Croce, "Oncogenes and cancer," *New England Journal of Medicine*, vol. 358, no. 5, pp. 502–511, 2008. PMID: 18234754.

[19] M. Puig, S. Casillas, S. Villatoro, and M. Cáceres, "Human inversions and their functional consequences," *Briefings in Functional Genomics*, vol. 14, pp. 369–379, 05 2015.

[20] 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.

[21] 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.

[22] A. J. Sharp, A. Itsara, Z. Cheng, C. Alkan, S. Schwartz, and E. E. Eichler, "Optimal design of oligonucleotide microarrays for measurement of DNA copy-number," *Hum Mol Genet*, vol. 16, pp. 2770–2779, Nov 2007.

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

[24] S. A. McCarroll, T. N. Hadnott, G. H. Perry, P. C. Sabeti, M. C. Zody, J. C. Barrett, S. Dallaire, S. B. Gabriel, C. Lee, M. J. Daly, D. M. Altshuler, and I. H. C. , "Common deletion polymorphisms in the human genome," *Nat Genet*, vol. 38, pp. 86–92, Jan 2006.

[25] G. M. Cooper, T. Zerr, J. M. Kidd, E. E. Eichler, and D. A. Nickerson, "Systematic assessment of copy number variant detection via genome-wide SNP genotyping," *Nat Genet*, vol. 40, pp. 1199–1203, Oct 2008.

[26] 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.

[27] 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.

[28] J. M. Kidd, N. Sampas, F. Antonacci, T. Graves, R. Fulton, H. S. Hayden, C. Alkan, M. Malig, M. Ventura, G. Giannuzzi, J. Kallicki, P. Anderson, A. Tsalenko, N. A. Yamada, P. Tsang, R. Kaul, R. K. Wilson, L. Bruhn,

and E. E. Eichler, "Characterization of missing human genome sequences and copy-number polymorphic insertions," *Nat Methods*, vol. 7, pp. 365–371, May 2010.

[29] 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.

[30] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends Genet*, vol. 24, pp. 133–141, Mar 2008.

[31] 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. StÃ$\frac{1}{4}$tz, 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] F. Hormozdiari, F. Hach, S. C. Sahinalp, E. E. Eichler, and C. Alkan, "Sensitive and fast mapping of di-base encoded reads," *Bioinformatics*, vol. 27, pp. 1915–1921, Jul 2011.

[39] C. Stewart, D. Kural, M. P. StrÃ¶mberg, J. A. Walker, M. K. Konkel, A. M. StÃ¼tz, A. E. Urban, F. Grubert, H. Y. K. Lam, W.-P. Lee, M. Busby, A. R. Indap, E. Garrison, C. Huff, J. Xing, M. P. Snyder, L. B. Jorde, M. A. Batzer, J. O. Korbel, G. T. Marth, and . G. Project, "A comprehensive

map of mobile element insertion polymorphisms in humans.," *PLoS genetics*, vol. 7, p. e1002236, Aug. 2011.

[40] J. Wu, W.-P. Lee, A. Ward, J. A. Walker, M. K. Konkel, M. A. Batzer, and G. T. Marth, "Tangram: a comprehensive toolbox for mobile element insertion detection.," *BMC genomics*, vol. 15, p. 795, Sept. 2014.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O'Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, and M. Loose, "Nanopore sequencing and assembly of a human genome with ultra-long reads.," *Nature biotechnology*, vol. 36, pp. 338–345, Apr. 2018.

[46] J. Huddleston, M. J. Chaisson, K. Meltz Steinberg, W. Warren, K. Hoekzema, D. S. Gordon, T. A. Graves-Lindsay, K. M. Munson, Z. N. Kronenberg, L. Vives, P. Peluso, M. Boitano, C.-S. Chin, J. Korlach, R. K. Wilson, and E. E. Eichler, "Discovery and genotyping of structural variation from long-read haploid genome sequence data," *Genome research*, Nov. 2016.

[47] 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.

[48] 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.

[49] 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.

[50] H. Li, "Minimap2: pairwise alignment for nucleotide sequences.," *Bioinformatics (Oxford, England)*, vol. 34, pp. 3094–3100, Sept. 2018.

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

[52] M. Brunato, H. H. Hoos, and R. Battiti, *On Effectively Finding Maximal Quasi-cliques in Graphs*, pp. 41–55. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[53] R. M. Karp, "Reducibility among combinatorial problems," in *Complexity of computer computations*, pp. 85–103, Springer, 1972.

[54] R. Battiti and M. Protasi, "Reactive local search for the maximum clique problem 1," *Algorithmica*, vol. 29, pp. 610–637, Apr 2001.

[55] W. Pullan and H. H. Hoos, "Dynamic local search for the maximum clique problem," *J. Artif. Int. Res.*, vol. 25, pp. 159–185, Feb. 2006.

[56] M. Ferrarini, M. Moretto, J. A. Ward, N. Surbanovski, V. Stevanovic, L. Giongo, R. Viola, D. Cavalieri, R. Velasco, A. Cestaro, and D. J. Sargent, "An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome.," *BMC genomics*, vol. 14, p. 670, Oct. 2013.

[57] J. M. Zook, D. Catoe, J. McDaniel, L. Vang, N. Spies, A. Sidow, Z. Weng, Y. Liu, C. E. Mason, N. Alexander, E. Henaff, A. B. R. McIntyre, D. Chandramohan, F. Chen, E. Jaeger, A. Moshrefi, K. Pham, W. Stedman, T. Liang, M. Saghbini, Z. Dzakula, A. Hastie, H. Cao, G. Deikus, E. Schadt, R. Sebra, A. Bashir, R. M. Truty, C. C. Chang, N. Gulbahce, K. Zhao, S. Ghosh, F. Hyland, Y. Fu, M. Chaisson, C. Xiao, J. Trow, S. T. Sherry, A. W. Zaranek, M. Ball, J. Bobe, P. Estep, G. M. Church, P. Marks, S. Kyriazopoulou-Panagiotopoulou, G. X. Y. Zheng, M. Schnall-Levin, H. S. Ordonez, P. A. Mudivarti, K. Giorda, Y. Sheng, K. B. Rypdal, and M. Salit, "Extensive sequencing of seven human genomes to characterize benchmark reference materials," *Scientific data*, vol. 3, p. 160025, June 2016.

[58] M. A. Eberle, E. Fritzilas, P. Krusche, M. Kallberg, B. L. Moore, M. A. Bekritsky, Z. Iqbal, H.-Y. Chuang, S. J. Humphray, A. L. Halpern, S. Kruglyak, E. H. Margulies, G. McVean, and D. R. Bentley, "A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree.," *Genome research*, vol. 27, pp. 157–164, Jan. 2017.

[59] 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, pp. 461–468, June 2018.

[60] 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.

[61] F. Antonacci, J. M. Kidd, T. Marques-Bonet, M. Ventura, P. Siswara, Z. Jiang, and E. E. Eichler, "Characterization of six human disease-associated inversion polymorphisms," *Hum Mol Genet*, vol. 18, pp. 2555–2566, Jul 2009.

[62] F. Antonacci, J. M. Kidd, T. Marques-Bonet, B. Teague, M. Ventura, S. Girirajan, C. Alkan, C. D. Campbell, L. Vives, M. Malig, J. A. Rosenfeld, B. C.

Ballif, L. G. Shaffer, T. A. Graves, R. K. Wilson, D. C. Schwartz, and E. E. Eichler, "A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk," *Nat Genet*, vol. 42, pp. 745–750, Sep 2010.

[63] 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.

[64] 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.

[65] 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.

[66] Y. Ono, K. Asai, and M. Hamada, "PBSIM: PacBio reads simulator—toward accurate genome assembly," *Bioinformatics*, vol. 29, no. 1, pp. 119–121, 2013.

[67] F. Antonacci, M. Y. Dennis, J. Huddleston, P. H. Sudmant, K. M. Steinberg, J. A. Rosenfeld, M. Miroballo, T. A. Graves, L. Vives, M. Malig, L. Denman, A. Raja, A. Stuart, J. Tang, B. Munson, L. G. Shaffer, C. T. Amemiya, R. K. Wilson, and E. E. Eichler, "Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability," *Nat Genet*, vol. 46, pp. 1293–1302, Dec 2014.

[68] PacificBiosciences, "Pacificbiosciences/pbsv," 2020.

[69] 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.

Appendix A

Glossary

DNA: Deoxyribonucleic acid
RNA: Ribonucleic acid
SV: Structural variation
HTS: High Throughput Sequencing
FASTQ: A file format for storing reads with quality information
BAM: Compressed Sequence Alignment Mapping format
CNV: Copy Number Variation
CLR: Continuous long-read
CCS: Circular Consensus Sequencing
HiFi: High Fidelity

Appendix B

Code

Implementation is available at https://github.com/BilkentCompGen/