Understanding The Complexity of Human Structural Genomic Variation Through Multiple Whole Genome Sequencing Platforms

by
Xuefang Zhao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2017

Doctoral Committee:
Assistant Professor Ryan E. Mills, Chair
Professor Tomas Glover
Assistant Professor Jeffrey Kidd
Associate Professor Jun Li
Professor Kerby Shedden

Xuefang Zhao

xuefzhao@umich.edu

ORCID iD: 0000-0003-4036-9577

This dissertation is dedicated to my family,
whose unconditional love, understanding and respect has supported me through.

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. Ryan E Mills for his guidance during the five years of my graduate life. He’s a great mentor, who not only introduced me to the wonder and frustrations of scientific study, but also helped me build confidence in pursuing an independent academic career. He has granted me the freedom to establish scientific questions, explore the underlying truth, and provided timely guidance when I got lost. He has also introduced me to several large collaborative consortiums, where I got to learn the art of achieving goals collaboratively and representing my own work efficiently. In one word, Dr. Mills is the best Ph.D. advisor I could ever think of. I am also grateful to my highly supportive committee members, whose feedback on my research and presenting skills are of great value. In particular, I would like to thank Dr. Jeffery Kidd and Dr. Tomas Glover for the insightful discussions of complex duplications in human genomes, Dr. Jun Li and Dr. Kerby Shedden for their brilliant suggestion on mathematical and statistical model building.

I would also like to thank all my collaborators, for helping me complete my research stories. I want to thank Dr. Jeffery Kidd and Sarah Emery for their contribution to my structural variants detection project, where they conducted the PCR validation of the complex structural variants my algorithm predicted. I also thank the Human Genome Structural Variation Consortium (HGSVC) and the Schizophrenia Genetics and Brain

Somatic Mosaicism Network (BSMN) for providing me the opportunity to contribute to their studies and access to the highly valuable datasets they have created.

I would also like to thank all my friends here at University of Michigan. I am lucky to have Shweta Ramdas, Sai Chen, Shiya Song and Raymond Cavalcante as my classmates, together with whom I took classes, worked for group projects and also killed most of my spare time. They are always there for me, shared the happiness of my life as well as helped me out of the hard time.

Last but not the least, I’d like to thank my husband and my parents for supporting me through all these years. Their love and respect has brought me where I am now, made me fearless for the future, whatever challenges are there waiting for me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Genomic structural variants (SVs) are major sources of genome diversity and closely related to human health, as indicated by numerous studies. In spite of the recent advances in sequencing technology and discovery methodology, there are still considerable amounts of variants in the genome that are partially or completely misinterpreted. This thesis has mainly focused on comprehensively interpreting the structural variants in human genomes by accurately defining the locations and formats of variants with the application of different sequencing platforms. To accomplish this goal, I developed a randomized iterative approach to define all types of SVs, which has shown superior performance in accurately defining complex variants. Next, I built a recurrence based validation pipeline to systematically validate SVs with long read sequences. I conclude with a systematic integration of SVs in multiple individuals discovered by various short read based detecting algorithms, with supportive evidence from orthogonal technologies, which presents to date the most comprehensive SV map in the human genome and the best current technologies allow us to do.

CHAPTER I

Introduction and Background

1.1 INTRODUCTION

No two humans are identical; neither are their genomes. The differences between individual genomes are called genomic variants, which is not only the major drive of evolution but also main reason for various severe human diseases. Genomic variants are summarized by size into three major categorizes: the single nucleotide polymorphisms (SNPs) (Sachidanandam et al. 2001; International HapMap Consortium 2003), small insertions and deletions, i.e. indels, ranging from 1 - 50 bp (Weber et al. 2002; Bhangale et al. 2005; Mills et al. 2006; Mullaney et al. 2010), and large genomic structural variants (SVs) (Iafrate et al. 2004; Tuzun et al. 2005). An individual genome is estimated to carry ~3 million SNPs, 500,000 indels and tens of thousands SVs (Shen et al. 2013). Though there are relatively fewer SVs in the human genome, compared against the smaller indels and SNPs, genomes vary more as a consequence of large SVs because of the large genomic regions involved in such events (Iafrate et al. 2004; Alkan et al. 2011; Kidd et al. 2008; Conrad et al. 2010). At the same time, it has also been pointed out by numerous studies that SVs are closely related with cellular viability as SVs could alter the gene expression by truncating the DNA coding or regulatory regions, or cause fusion of different genes by transporting DNA material (Mertens et al. 2015). Moreover, SVs have been found to play important roles in various diseases raging from neurological and developmental disorders (Pinto et al. 2010; Sebat et al. 2007; Stefansson et al. 2008; McCarthy et al. 2009; Wellcome Trust Case Control Consortium et al. 2010) to large spectrum of cancers (Liu et al. 2015; Quinlan & Hall 2012; Weischenfeldt et al. 2013).

SVs are defined as the removal or rearrangement of genomic regions over 50bp in length, which are canonically categorized into four types: deletion, duplication, inversion and insertion. However, recent studies have revealed the existence of SVs in more complex formats, which have three or more breakpoints involved yet cannot be summarized by any of these canonical forms (Quinlan & Hall 2012), and the most extreme representative of complex SV that involved massive chromosomal shattering and rearrangements, termed 'chromothripsis', has been first described in cancer genomes (Stephens et al. 2011) and later characterized in germline genome (Chiang et al. 2012).

Advances in high throughput next generation sequencing (NGS) technology has made it possible for investigators to quickly sequence individual genomes at high depth, and the emergence of various variant detecting algorithms has further accelerated the discovery and analysis of genomic variants. Most current detection algorithms use pair-end short sequences and detect SVs by interpreting aberrant alignment signals such as abnormally long / short insert sizes against the overall library, aberrant read pair orientations, or split reads that are partially aligned or read depth that deviate from the overall distribution (Zhang et al. 2011). Despite the superior performance of current algorithms in detecting canonical SVs, limitations remain for:

1. SVs in complex formats which produce ambiguous alignment patterns that are beyond the scope of which current methods can detect;

2. SVs in highly repetitive genomic regions, as short reads from such regions usually have multiple alignment spots across the genome so that alignment bias arise.

Long read sequencing technology delivers DNA sequences that are several kilo bases (eg. Single Molecule Real Time sequencing from Pacific Biosciences) or even at the length of whole chromosome (eg.Oxford Nanopore sequencing) (Roberts et al. 2013; Loman et al. 2015). These reads are usually long enough to fully transverse the repetitive regions, thus suffering from little alignment bias and allow for direct comparison against the reference. In theory, all different forms of large genomic SVs could be defined by comparing the long sequences against the reference genome, in spite of the repetitiveness of local reference, as long as the long reads or their assembled contigs are of a low enough error rates. Several studies have been conducted to define variants genome wide with long reads (Pendleton et al. 2015; Chaisson et al. 2014; Shi et al. 2016), and showed significant superiority. However, as a relatively new technology, long read sequencing technology is held back from being widely adopted mainly due to the high cost of current platforms, and the limited number of methods available for the application at low computing cost.

1.2 THESIS OUTLINE

This thesis mainly focuses on understanding the complexity of genomic structural variants by interpreting sequencing data from various platforms including paired end short reads and long read sequences. To accomplish this goal, a short read sequence based SV detecting algorithm was developed to systematically discover SVs in all formats in human genome, and then a long read based SV validation algorithm was implemented to add assessment of the SVs through orthogonal technology. The scope of this thesis was later extended to include an integration of SVs discovered in multiple individuals by 15 different algorithms to reach an inclusive non-redundant SV set in human genome, and a systematic comparison between SVs discovered by different technologies to estimate their relative strength and weakness, which provides guidance for future sequencing projects.

In chapter II, we developed an algorithm that’s capable of comprehensively describing genomic structural variants in both simple and complex formats (Zhao et al. 2016). Instead of the traditional strategies that search for aberrant alignment signals to infer structural changes, this approach works by virtually rearranging segments of the genomes in a randomized fashion and attempting to minimize such aberrations relative to the observed characteristics of the sequence data. In this manner, the rearrangements detected by this approach are expanded to include all the complex types such as multi-deletion and duplication-inversion-deletion events, instead of only focusing on the canonical forms of which rich experience has been accumulated. Moreover, the homologous loci in diploid genome are assessed independently in this approach thus allowing for accurately description of overlap SVs on both alleles.

In chapter III, a recurrence based validation approach was developed to assess the quality of SVs through long read sequences, which directly compares long reads against the reference genome and its alterations guided by the predicted SV through recurrence matrix (preprint at : http://biorxiv.org/content/biorxiv/early/2017/02/24/105817.full.pdf, under review). This method differs from other long read based SV algorithms in the fact that assembly is not required, thus achieving high computing efficiency as well as avoiding the requirement of high sequence depth which is directly correlated with the sequencing cost.

In chapter IV, an integrated set of SVs was produced by combining genomic variants detected in multiple human samples by different sequencing platforms including paired end short read sequencing, long read sequencing and single strand sequencing technologies, the set of which represents the limit of current technology and methodology in defining SVs, and provides insight into the relative power of different technologies. With this work, a systematic integration pipeline was proposed to systematically integrate SVs detected from paired end short libraries by different algorithms, the output of which showed significant performance increase.

1.3 BACKGROUND

1.3.1 Sequencing technology revolution in recent decade

It has been four decades since the 'chain-termination' DNA sequencing technology was first developed by Frederick Sanger and his colleges, where radiolabelled dideoxynucleotides (ddNTPs) are mixed with deoxyribonucleotides (dNTPs) at certain fraction into a DNA replication reaction to cause randomized stops of DNA extension thus produce DNA strands of different length, which can be easily differentiated on a polyacrylamide gel. In this way, DNA sequence can be inferred by running four experiments with ddATP, ddTTP, ddCTP and ddGTP respectively in parallel (Heather & Chain 2016) (Figure 1.1) and inferring the bases by observing their relative position on the gel. Despite the improvements that have been made to this technology, Sanger sequencing remained the most widely used sequencing technology for almost four decades until the emergence of next generation sequencing (NGS).

Like Sanger sequencing, NGS is also a ‘sequence-by-synthesis’ (SBS) technology, while differentiating itself by the high throughput feature that’s accomplished in varied ways according to different platforms. Taking Illumina sequencing, the most commonly adopted platform, for example, DNA is randomly fragmented and ligated with adaptors on both ends and is then loaded onto a lawn of surface-bounded oligomers complementary to the adaptors to be amplified into a clonal cluster. dNTPs, labeled with unique fluorescent for each type, are added to be incorporated into the growing DNA chain through the replication process, which also serve as reversible terminators that prevent new dNTPs from being added until the fluorescents are imaged and then endemically cleaved for further DNA extension (Buermans & den Dunnen 2014). Compared to Sanger sequencing, NGS shows significantly higher efficiency and robustness, with an over 50,000-fold decrease in cost (Goodwin et al. 2016) (Figure 1.2), while keeping the error rate within 1% (Nakamura et al. 2011; Manley et al. 2016) and making it affordable to deep sequence individual genomes for detailed genomic variants discovery or to sequence at population level to explore the evolutionary principles.

Despite these advantages of NGS, it exhibits limitations in extremely complex genomic regions, mainly due to the relatively short sequence length which can easily arise ambiguity when aligned against the reference genome. Third generation sequencing, or long read sequencing, delivers reads in several kilo bases or even longer, that can fully cover the genomic regions where NGS usually show weakness at. Two main systems of long read sequencing are currently adopted: single molecule real-time (SMRT) from Pacific Biosciences (PacBio) and linked read sequencing (eg. 10X Linked Read Sequencing). During SMRT sequencing, the DNA sequences are recorded while dNTPs are incorporated into the templates stabilized in a picoliter well, named zero-mode waveguides (ZMW) (Levene et al. 2003). The current platform produces DNA sequences as long as ~100Kb, long enough to cover most of the complex genomic regions. Though the raw sequences have a high error rate around 15% (Rhoads & Au 2015), the introduction of circular consensus sequencing (Travers et al. 2010), whereby a SMRTbell template consisting of a double-stranded region flanked on both end by single-stranded loops was constructed to allow the DNA polymerase to read through the sequences for multiple times circularly (Figure 1.5). In this way, multiple subreads of the same genomic region could be considered together for consensus with significantly decreased error rates. The other long read technology, i.e. 10X, first clonally amplify and barcode long DNA molecules (~10Kb) and then sequence them through the short read sequencing platform. This way, the sequencing error rate and cost are at the same level to NGS, while the potential bias introduced by GC content and or tandem repeats are inevitable.

The breakthroughs in sequencing technologies have greatly accelerated the development of human genome studies, allowing the variances in human genome to be defined at much finer scales compared to previous array based technologies (Anon n.d.). With these, it has been revealed that human genomes harbor large number of variants in various forms, with size ranging from single base to couple mega-bases. The landscape of genomic variants largely exceeded people’s expectation, with millions of single nucleotide variants (SNVs) and small variants (indels, 2-50 bp) , as well as tens of thousands large structural variants discovered in human population (The 1000 Genomes Project Consortium 2010)

1.3.2 Genomic structural variation in simple and complex forms

Genomic structural variants (SVs) are rearrangement of large genomic regions (<50bp), which are commonly observed in the forms of deletion, duplication, insertion, inversion and translocation (Figure1.3). Deletion is defined as depletion of a genomic region in subject genome compared to reference genome, and insertion is the existence of extra copy of DNA material. Duplication is duplicated copy of genomic region gets inserted either right next to the original copy (tandem) or other regions in the genome(dispersed). Inversion and translocation refer to the change of orientation and physical position of DNA pieces respectively.

Aside from the canonical forms, SVs in more complex patterns are also observed in considerable frequencies with varied degrees of complexity. Complex genomic structural variant (CSVs) are events that consist of three or more breakpoints, and cannot be explained by a single end-joining or DNA exchange event (Quinlan & Hall 2012). For example, a piece of DNA could be duplicated and inserted at a different locus, in either the original or the opposite orientation, with the insertion point harboring micro-deletion/ insertion, or sometimes large deletions. This represents a complex SV that's coupled by duplications and inversion, sometimes with deletions involved as well, which is a relatively straight forward type of CSV. The complexity of CSVs could be surprising, represented by the event termed 'chromothripsis', where a large genomic region could be sheared into tens to hundreds of small genomic pieces followed by massive translocations of those pieces. This phenomenon was first described in cancer genomes (Stephens et al. 2011) and then also characterized in the germline (Chiang et al. 2012).

Except for serving as an important source of genomic variety, SVs can affect the activity of individual cell by altering gene expression at different level or cause fusion between different genes (Sjödin & Jakobsson 2012; Tang & Amon 2013), further causing undesired biological or physiological conditions such as cancers (Campbell et al. 2008), autism-related disorders (Henrichsen et al. 2009; Zhang et al. 2013; Teshima & Innan 2012; Zhang et al. 2009; Brand et al. 2015; Hedges et al. 2012; Kusenda & Sebat 2008; Marshall et al. 2008; Sebat et al.

2007; Henrichsen et al. 2009; Zhang et al. 2013; Teshima & Innan 2012; Zhang et al. 2009) or psychiatric disorders (eg. Schizophrenia (Sekar et al. 2016; Shi et al. 2008; Sebat et al. 2009).

1.3.3 Detection of SVs and CSVs with paired end short library sequences

A good diversity of algorithms have been developed to detect SVs from short read paired end sequences, the underlying approaches can be summarized into four categories: read-pair (RP), read-depth (RD) split-read (SR) and sequence assembly(AS). Some of the currently available SV detecting tools adopt one of these approaches, while most others take multiple of them in proper combination to achieve their specific expectations. Each of the four approaches is briefly described here:

RP approaches compare the insert size and orientation of read pairs in targeted genomic regions against their expected overall distribution estimated from all read pairs aligned to the reference genome. The sizes of DNA fragments produced by paired-end sequencing are usually approximated by Gaussian or bimodal distribution, so that deletions and insertions can be detected with read pairs of aberrantly large or small insert sizes (Pang et al. 2010; Tuzun et al. 2005; Kidd et al. 2008) against the null distribution. Orientation of a pair of reads is expected to be forward-reverse in absence of inversions, so that clusters of forward-forward or reverse-reverse pairs are indicative of existence of inverted structures. Sensitivity of RP methods in detecting SVs is largely decided by the distribution of fragment sizes, and higher sensitivity is achieved with tighter distributions. Most RP based SV detecting algorithms only detect deletions and insertions over 500bp, owing to the difficulty in separating small perturbations in read-pair distance from the normal background variability (Medvedev et al. 2009). Furthermore, RP methods are not applicable for detection of CNVs in low-complexity regions with segmental duplication (Zhao et al. 2013).

RD methods detect copy number variants (CNVs) by searching for changes in the read depth, under the assumption that read depth of a genomic region is positively correlated to the copy number of the region. GC content is usually considered as a major confounding factor that introduces bias to read depth assessment (Benjamini & Speed 2012), so that read depth are usually corrected for GC content before introduced to the assessment. Most RD algorithms would set an appropriate size for a sliding window according to the mean number of reads in each window, and implement a negative-binomial distribution to approximate an over dispersed Poisson distribution of the data. Size of the sliding window is negatively correlated with the breakpoint resolution and the computing cost. With larger window, it is faster for RD approaches to make CNV calls at the sacrifice of breakpoint resolution (Yoon et al. 2009). One of the recent approaches, named *Genome STRiP* (Handsaker et al. 2011) , achieves high accuracy in estimating the copy number of a genomic region by fitting mixed Gaussian models at population level.

Split Read(SP) methods analyze reads that are only partially aligned to the genome (Zhang et al. 2011). The end of the aligned portion could serve as potential breakpoint candidates, thus providing single base resolution for SV detection. Split-read based methods, such as Pindel (Ye et al. 2009), Gustaf (Trappe et al. 2014), SVseq2 (Zhang et al. 2012), and Prism (Jiang et al. 2012), have the advantage of identifying breakpoints at high resolution, but are usually limited to find relatively small SVs.

The sequence assembly(SA) methods first assemble the original sequences and then infer deletions or insertions by directly comparing assemblies against the reference (Simpson et al.

2009; Hajirasouliha, Hormozdiari & Alkan 2010; Li, Fan & Tian 2010 (Nijkamp et al., 2012; Teo et al., 2012). SA method, in theory, is capable of detecting all types of genomic variants at high resolution. However, various limitations such as the alignment bias of short reads and the overwhelming demand on computing sources severely prevent this approach from being commonly adopted.

Each of the approaches described above has its own strengths and weaknesses. Though RP and SP methods can define breakpoints at high resolution, their performance is highly influenced by the alignment quality so that they show severely decreased performance at repetitive genomic regions (Medvedev et al. 2009). RD method is well suited for accurately detecting deletions and tandem duplications, the breakpoint resolution is much lower compared to others. AS-based tools take advantage of not requiring a reference genome, but they usually require extremely large memories and long computing time. In this situation, methods that properly combine multiple approaches are expected to take advantage of the unique features of each approach, while avoid the limitations by complementing each method with another. Most of the current algorithms, such as Delly (Rausch et al. 2012), Lumpy (Layer et al. 2014), Wham (Kronenberg et al. 2015), SVelter (Zhao et al. 2016) and novoBreak (Chong et al. 2016), analyze different types of aberrant alignment signals and combine them in their unique framework to achieve superior performance.

1.3.4 Limitations of current short read based SV detecting algorithms

The SV detecting methods described above has been widely adopted in various studies and proven to work well on simple SVs. However, CSVs are usually misinterpreted as of the ambiguous alignment signatures they produce. For example, inverted duplications are usually misinterpreted as inversion, because of the aberrant oriented read clusters is in similar pattern with the signature of simple inversions (Figure 1.4A). Similarly, when a deletion locates adjacent to an inversion, ambiguous inversion is usually predicted with the deletion being omitted (Figure 1.4B).

CSVs, differentiated from simple SVs mainly by the fact that they have more than two breakpoints involved, could only be fully resolved with the precondition that all breakpoints are properly defined and clustered. The typical methods described above usually show poor performance on CSVs, mainly due to the fact that:

1. These methods only try to define a pair of breakpoint each time and then define SVs accordingly, in which situation only simple event can be reported.

2. The underlying SVs are inferred by examining the pre-defined discordant alignment patterns of read pairs, while the alignment patterns of CNVs are usually unpredictable, thus impossible to be comprehensively resolved with the traditional methods.

To overcome the limitations of current SV discovery approaches, the first focus of this thesis was set to develop a method that could comprehensively detect and resolve genomic SVs in both simple and complex format, which should address the limitations discussed above. This method is unique in the fact that, 1. Groups of two or more breakpoints, instead of breakpoint pairs, were defined simultaneous based on the alignment signatures; 2. Instead of predicting SVs by fitting pre-designed models on the alignment abnormity, a Markov chain Monte Carlo process was implemented instead to allow for data-driven exploration of the optimized underlying structure; 3. The two alleles at the same locus were considered jointly for the assessment, allowing overlap SVs to be predicted together. More details of this method are described in Chapter 2.

1.3.5 Application of long read sequencing technology

Though NGS technology has various advantages including the high throughput, low cost and low error rate, limitations have been shown in accurately characterizing simple repeats and segmental duplications in human genome (Alkan et al. 2010), due to the relatively short sequences delivered. Moreover, short reads are also significantly biased by the GC content in the sequence, which might cause misinterpretation of genomic variants. The long read sequencing technologies, such as the single-molecule real-time (SMRT) (Eid et al. 2009) from Pacific Biosciences (PacBio), or the linked read sequencing from 10X genomics, address these issues by delivering reads that are of several kilo bases to hundreds of kilo bases, which could fully transverse the complex genomic regions so that the biases caused by GC content or the alignment are minimized. Instead of predicting SVs based on the aberrant alignment signature, long reads allow for direct comparison against reference sequences thus providing the possibility to fully resolve all the genomic variants at once, including those in the extreme complex formats such as chromothripsis. At the same time, single base resolution could be achieved. Currently, SMRT sequencing is the mostly commonly adopted long read technology which have been applied on various studies (Pendleton et al. 2015; Chaisson et al. 2014; Shi et al. 2016) and have shown significant advantages in calling genomic variants compared to NGS.

The current application of long read sequences in deciphering genomic structural variants mainly focus on individual or small amount of genomes, represented by studies by Chaisson et al. 2014, Pendleton et al. 2015 and Shi et al. 2016, while the discovery methods usually require long read assembly across the whole genome. Though long reads assembly is of high quality compared to short reads (Carvalho et al. 2016), as they are free of the alignment ambiguousness in complex or repetitive genomic regions, the high computing cost of genome wide assembly still make it affordable to only limited number of researchers.

To explore the potential applications of long read sequences in an efficient way, I first attempted to apply them as an orthogonal validation approach for genomic SVs predicted by other technologies, built an autonomous pipeline that systematically while efficiently assess the quality of SVs, and wrapped it in an light weighted user-friendly tool named VaPoR (http://biorxiv.org/content/early/2017/02/24/105817). Rather than global or local assembly, this method summarizes statistical characteristics of the recurrence matrix produced by direct comparison of individual long sequences versus reference, thus achieving the high efficiency as well as leaving the possibility to be developed to a long read based SV detector. Details of this method will be discussed in Chapter 3.

1.3.6 Integration and comparison of SVs detected by different platforms

The 1000 Genomes Project (1KGP) is an international research consortium founded in 2008, aiming at systematically sequencing thousands of individuals from various populations across the world to build a resource to help to understand the genetic contribution to phenotypes (1000 Genomes Project Consortium et al. 2012; The 1000 Genomes Project Consortium 2010). Different versions of genomic variation maps have been published by the consortium since its foundation (Mills et al. 2011a; 1000 Genomes Project Consortium et al. 2015; Sudmant et al. 2015a), becoming more and more comprehensive in their detailed description of all forms of variants in human genome that were discovered from a collection of samples from different populations. These maps have provided valuable resources to the community and served as an important reference for annotation of disease causal variants, as well as encouraged the development of various genomic variants detecting, genotyping and haplotyping algorithms.

Most of the previous publications from the 1KGP describe SVs discovered by short read sequencing technologies, while more recently the consortium have included other technologies such as SMRT (Rhoads and Au 2015), 10X (Mostovoy et al. 2016), BioNano, Strand-Seq (Falconer & Lansdorp 2013; Falconer et al. 2012) and Hi-C (Belton et al. 2012). These technologies, together with the paired-end NGS were applied to three father-mother-child trios from Han Chinese, Puerto Rican and Yoruba respectively with high depth. Similar with the other studies conducted by 1KGP, an integrated SV map is expected from these sequences, while this analysis is unique in the fact that it also explores the optimized combination of different platforms and the represent the extent to which people can do best in defining genomic variants. Details of these analyses are listed in chapter IV.

1.4 FIGURES

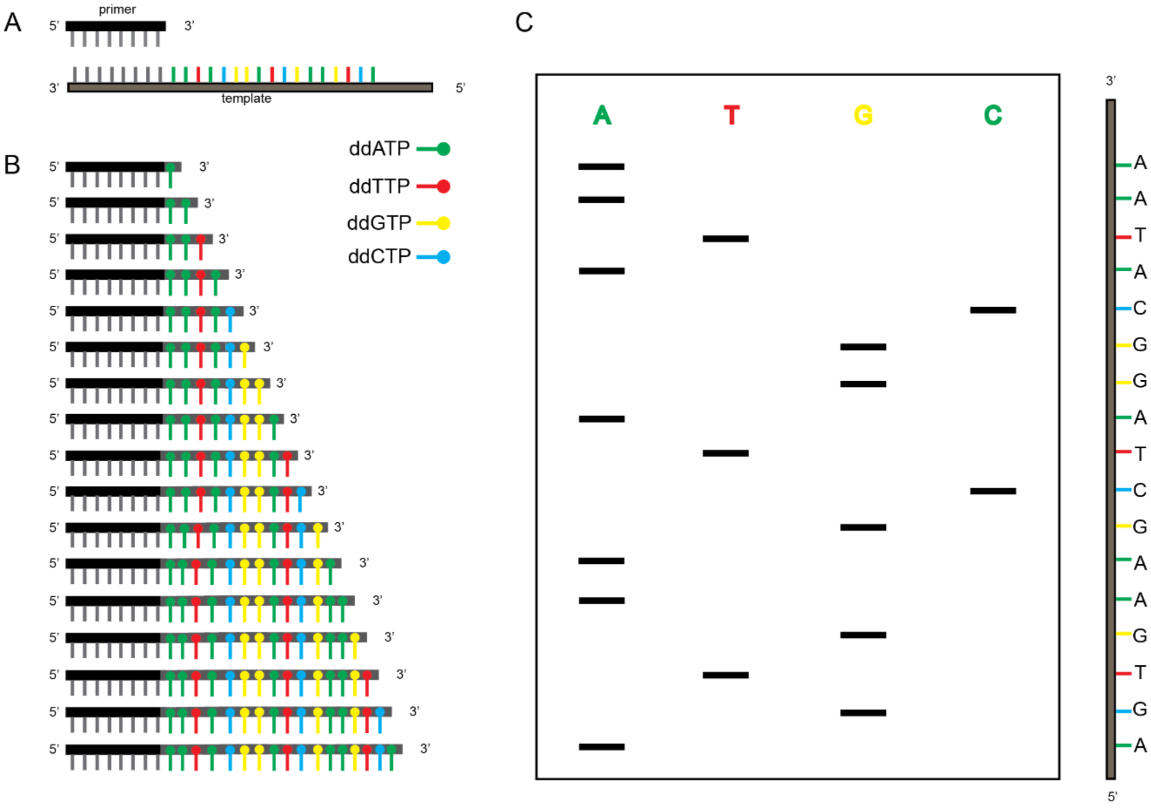
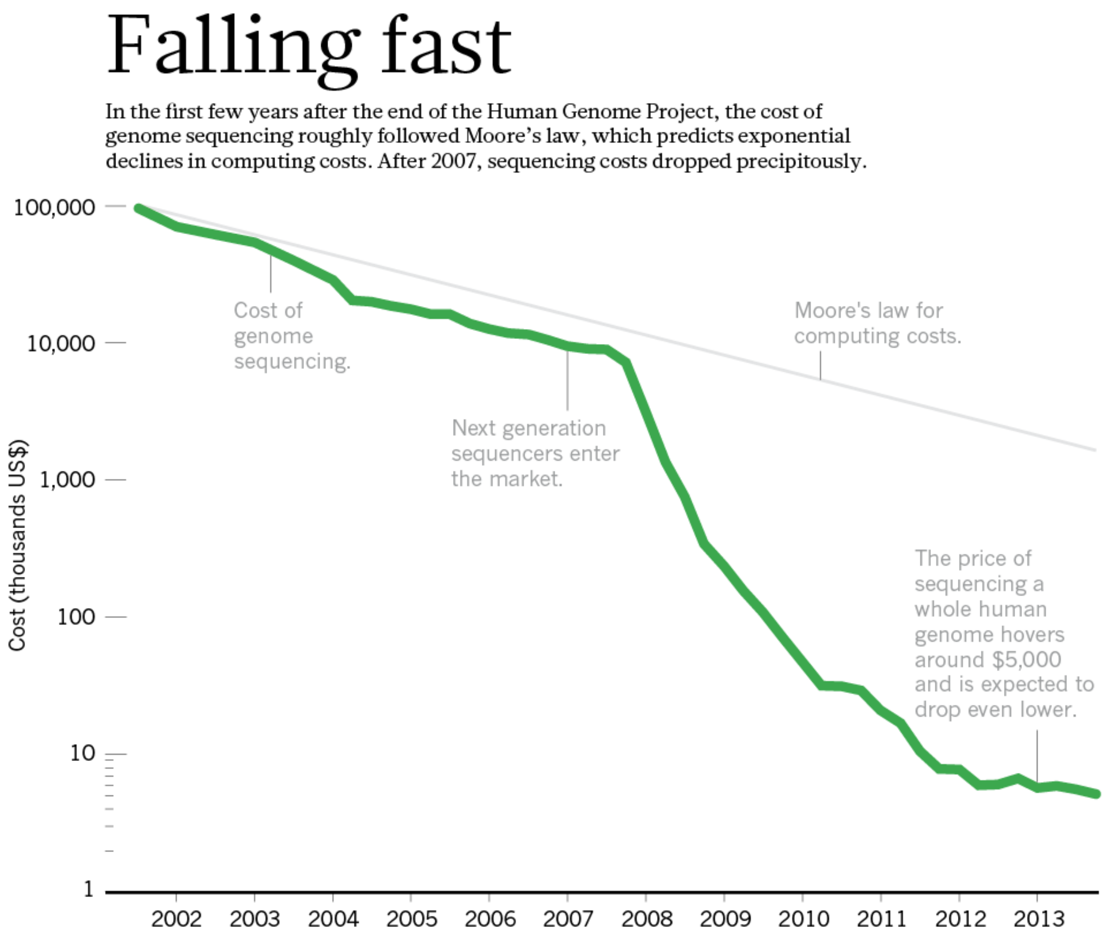


Figure 1.1 First-generation DNA sequencing technologies.

Example DNA to be sequenced (A) is illustrated undergoing Sanger sequencing. (B) Sanger's ‘chain-termination’ sequencing. Radio- or fluorescently-labelled ddNTP nucleotides of a given type, which once incorporated, prevent further extension, are included in DNA polymerization reactions at low concentrations (primed off a 5′ sequence, not shown). Therefore, in each of the four reactions, sequence fragments are generated with 3′ truncations as a ddNTP is randomly incorporated at a particular instance of that base (underlined 3′ terminal characters). (C) Fragments generated from either methodology can then be visualized via electrophoresis on a high-resolution polyacrylamide gel: sequences are then inferred by reading ‘up’ the gel, as the shorter DNA fragments migrate fastest. In Sanger sequencing (left) the sequence is inferred by finding the lane in which the band is present for a given site, as the 3′ terminating labelled ddNTP corresponds to the base at that position. Maxam–Gilbert sequencing requires a small additional logical step: Ts and As can be directly inferred from a band in the pyrimidine or purine lanes respectively, while G and C are indicated by the presence of dual bands in the G and A + G lanes, or C and C + T lanes respectively.



Hayden, Erika Check. "The $1,000 genome." Nature 507.7492 (2014): 294.

Figure 1.2 Change of sequencing cost and throughput during the recent decades.

Since the introduction of next generation sequencing technology, the sequencing cost has decreased by over 50 times, which makes it feasible for deep sequence multiple genomes for a detailed maps of human genomic variants.

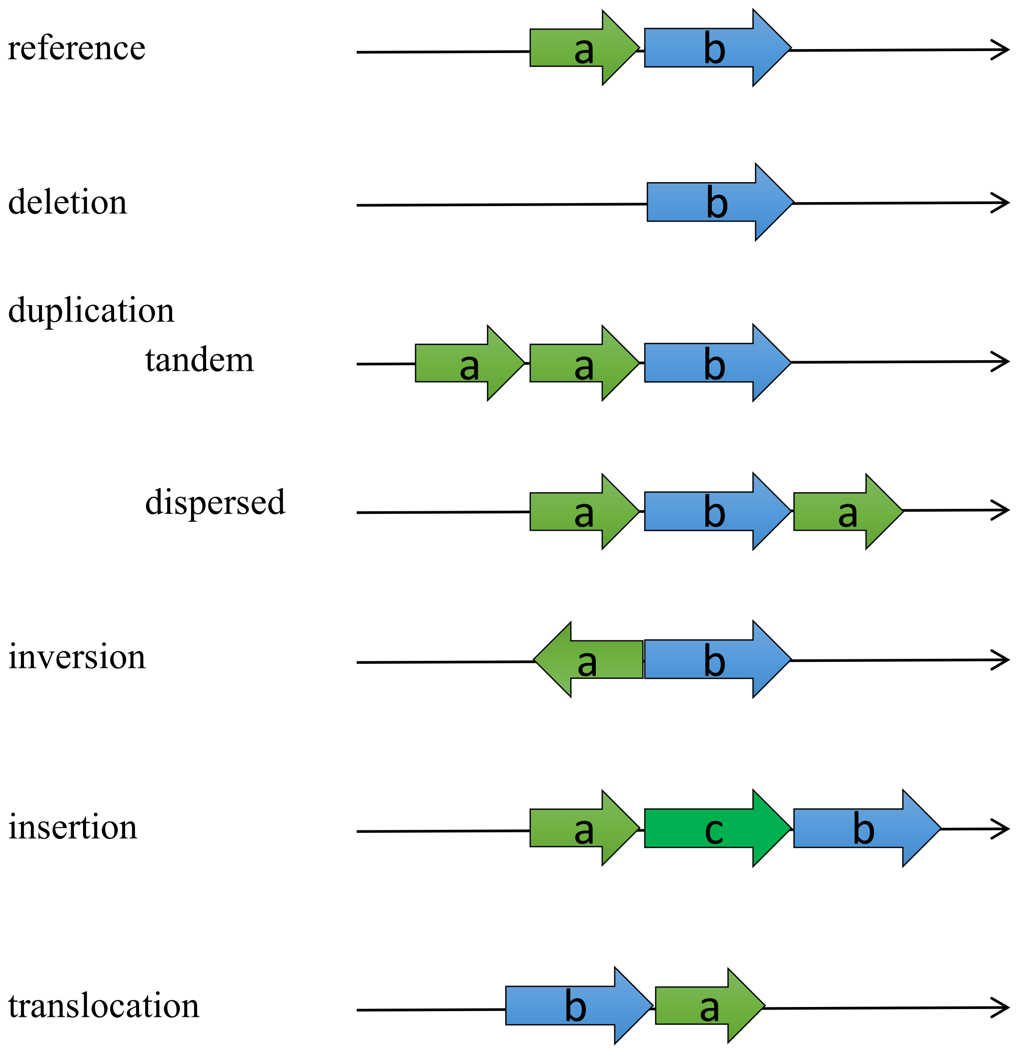


Figure 1.3 Canonical formats of structural variants.

The canonical defined formats of genomic structural variants include deletion, duplication, inversion, insertion and translocation. Deletion is the removal of a piece of DNA, while duplication is the existence of an extra copy of genomic materials. The second copy could either present adjacent to or far from the original copy.

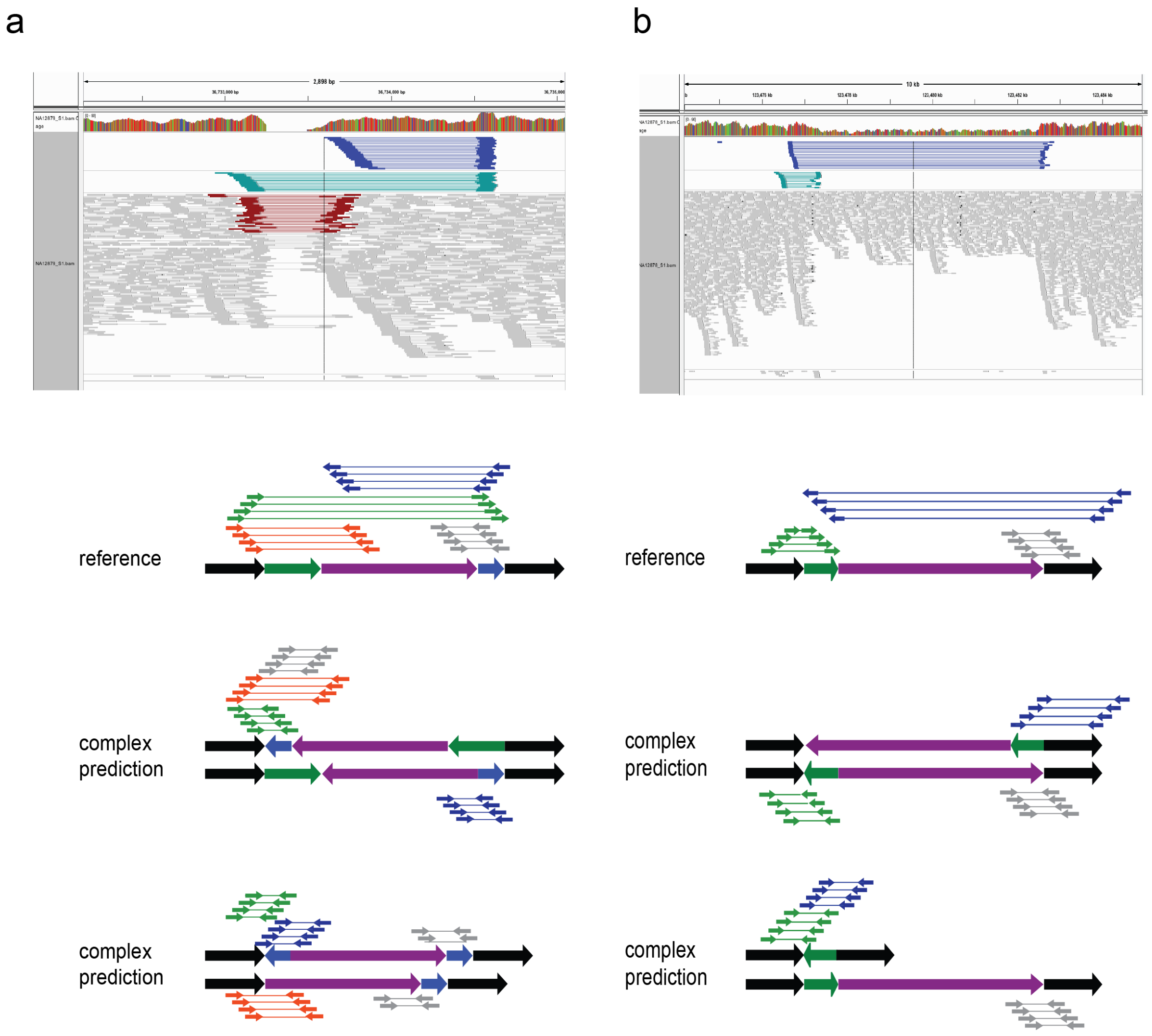
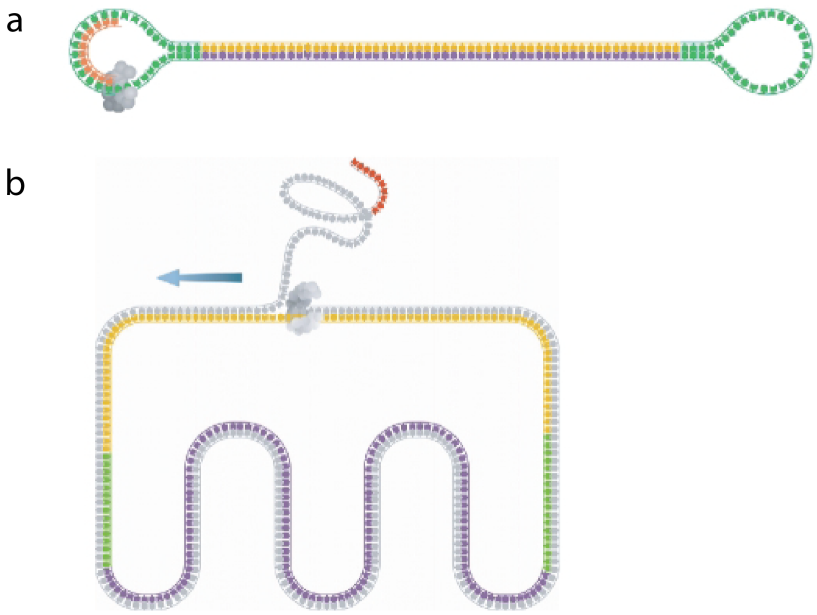


Figure 1.4 Examples of complex structural variants.

(a) shows a complex structural variation where a piece of DNA was duplicated and inserted 1kb upstream in the inverted orientation, while the ~300bp around the insertion point also harbors a homozygous deletion. (b) shows a deletion followed by an inversion.



Travers, Kevin J., et al. "A flexible and efficient template format for circular consensus sequencing and SNP detection." Nucleic acids research (2010): gkq543.

Figure 1.5 Schematic of a SMRTbell™ template.

(a) A SMRTbell template consists of a double-stranded region (the insert) flanked by two hairpin loops. The hairpin loops present a single-stranded region to which a sequencing primer can bind (orange). (b) As a strand-displacing polymerase (gray) extends a primer from one of the hairpin loops, it uses one strand as the template strand and displaces the other. When the polymerase returns to the 5′-end of the primer, it begins strand displacement of the primer and continues to synthesize DNA (moving in the direction of the blue arrow). Therefore, the length of sequence obtained from these templates is not limited by the insert length. Furthermore, the resulting sequence is derived from both sense- and anti-sense strands.

CHAPTER II

Detect and Resolve Complex Genomic Structural Variants

(Zhao, Xuefang, et al. *Genome biology* , 2016.)

2.1 ABSTRACT

Complex chromosomal rearrangements consist of structural genomic alterations involving multiple instances of deletions, duplications, inversions, or translocations that co-occur either on the same chromosome or represent different overlapping events on homologous chromosomes. We present SVelter, an algorithm that first identifies regions of the genome suspected to harbor a complex event and then iteratively rearranges the local genome structure, in a randomized fashion, with each structure scored against characteristics of the observed sequencing data. We show that SVelter is able to accurately reconstruct complex chromosomal rearrangements when compared to well-characterized genomes that have been deep sequenced with both short and long read technologies

. 2.2 KEYWORDS

structural variation

complex structural rearrangements

sequence analysis

human

SV

CNV

2.3 BACKGROUND

Structural variation (SV), defined as chromosomal rearrangements resulting from the removal, insertion, or rearrangement of genomic regions, are natural sources of genetic variation (Zarrei et al. 2015; Mills et al. 2011b; 1000 Genomes Project Consortium et al. 2012) that have also been implicated in numerous human diseases (Brand et al. 2014; Chiang et al. 2012; Stankiewicz & Lupski 2010). There have been extensive studies to discover these genomic aberrations from the whole genomes of humans and other species and numerous algorithms have been developed to accurately identify their prevalence (Chen et al. 2009; Ye et al. 2009; Layer et al. 2014; Handsaker et al. 2011; Rausch et al. 2012; Zhu et al. 2012)). These approaches have primarily focused on simple copy number variants (CNVs; deletions, duplications) or copy neutral (inversions) rearrangements defined by at most two chromosomal breakpoints (BPs) and work by identifying and clustering various signals of discordant alignments from paired-end next generation sequencing data (Alkan et al. 2011). Recent algorithms have begun to integrate signals across multiple features to increase sensitivity (Rausch et al. 2012; Layer et al. 2014; Sindi et al. 2012) and these have been successful in precisely identifying various types of SVs. Knowledge of the underlying structure of the rearrangement is still required, however, in order to properly model these aberrant signals to the correct type of structural variant. For example, clusters of read pairs (RPs) with insert sizes (ISs) larger than expected are typically representative of deleted sequence since this observation is consistent with how the reads would map in the presence of such an event.

While these simple rearrangements are common in the genome, there are additional rearrangements that, while rarer, are far more convoluted. These complex SVs (CSVs) are typically represented by three or more BPs and can arise from a variety of mechanisms including fork stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication (MMBIR) (Fig. 1, reviewed in (Usher & McCarroll 2015). Although fairly common in cancers, their prevalence in germline genomes is gradually becoming more apparent as is their potential role in the pathogenesis of other disease (Handsaker et al. 2015; Brand et al. 2014; Chiang et al. 2012). The complex nature of these events have made them challenging to accurately detect and catalog and many CSVs have been either neglected or misinterpreted by current techniques due to the complexity of the signals shown by the sequencing data. This is primarily due to the limitations of presupposing the types of SVs being considered, as oftentimes the signals from one event are clustered independently from those of another and can lead to contradictory or sometimes even opposing predictions to what is actually present. Under such circumstances, traditional prediction models lose their ability to discriminate between signals and therefore new computational strategies are required to overcome these challenges. Previous endeavors have been made to reconstruct somatic variants in cancer genomes both spatially (Steinberg et al. 2014; Chaisson et al. 2014) and temporally (Pendleton et al. 2015), but require an unaltered "matched" germline genome as an anchor for comparison. Studies into CSVs in the germline itself to date have thus been more limited, though there has been some early work that has profiled the existence of some of the more common types of CSVs including inverted-duplications and deletion-inversions (Parikh et al. 2016).

Here, we present a novel approach, SVelter, to accurately resolve complex structural genomic rearrangements in whole genomes. Unlike previous "bottom up" strategies that search for deviant signals to infer structural changes, our "top down" approach works by virtually rearranging segments of the genomes in a randomized fashion and attempting to minimize such aberrations relative to the observed characteristics of the sequence data. In this manner, SVelter is able to interrogate many different types of rearrangements, including multi-deletion and duplication-inversion-deletion events as well as distinct overlapping variants on homologous chromosomes. The framework is provided as a publicly available software package that is available online (https://github.com/millslab/svelter).

2.4 METHODS

2.4.1 SVelter Algorithm

SVelter takes aligned Illumina paired-end sequence data in sorted BAM format as input as well as the reference genome against which the sequences were aligned and reports all predicted SVs in both a custom format as well as VCFv4.1. Default parameters are chosen to best balance sensitivity and efficiency, though are adjustable for users to best fit their own data. The SVelter framework consists of three major modules: null model determination, breakpoint detection, random iterative rearrangement, and structure scoring. (Figure 2.1)

Null Model Determination

SVelter first filters the reference genome to exclude regions of low mappability from downstream analysis to increase efficiency by avoiding regions where alignments are unreliable. Such regions include gaps and unknown regions in the reference genome (Ns) and these are integrated with previously reported genomic regions identified by ENCODE (ENCODE Project Consortium 2012) (wgEncodeDacMapabilityConsensusExcludable and DukeMapabilityRegionsExcludable obtained from UCSC Genome Browser) that are of low mappability to form a final version of excluded regions. Next, the IS distribution (f_{IS}) is determined by calculating the mean (μ_{IS}) and standard deviation (${\sigma_{IS}}^2$) of all RPs aligned to genomic regions that are either randomly sampled or collected from a set of copy neutral (CN2) genomic regions defined as places in the genome where no polymorphic CNVs, segmental duplications, or repetitive elements have been annotated and thus providing a good estimate of the baseline alignment characteristics (Handsaker et al. 2015). Normal distribution is constructed ($f_{IS} \sim N(\mu_{IS}, {\sigma_{IS}}^2)$). A normal distribution of RD ($f_{RD} \sim N(\mu_{RD}, {\sigma_{RD}}^2)$)) and physical coverage ($f_{PC} \sim N(\mu_{PC}, {\sigma_{PC}}^2)$). are characterized by sliding a fixed-size window (default: 100 bp) across the same genomic region and constructing the sample mean and standard deviation. Alternatively, in situations where the RD is not high enough be approximated as normal (empirically, <10X), SVelter provides options for more complex but less efficient models, i.e. bimodal (fitted by *mixtools*) for IS,

$$f_{IS} \sim p \times N(\mu_{IS1}, {\sigma_{IS1}}^2) + (1-p) \times N(\mu_{IS2}, {\sigma_{IS2}}^2)$$

and negative binomial for read depth and physical coverage:

$$f_{RD} \sim NB(r_{RD}\,, P_{RD}), \qquad where\ r_{RD} = \frac{{\mu_{RD}}^2}{{\sigma_{RD}}^2 - \mu_{RD}}\,, \;\; P_{RD} = 1 - \frac{\mu_{RD}}{{\sigma_{RD}}^2}$$

$$f_{PC} \sim NB(r_{PC}\ , P_{PC}), \qquad where\ r_{PC} = \frac{{\mu_{PC}}^2}{{\sigma_{PC}}^2 - \mu_{PC}}\ ,\ \ P_{PC} = 1 - \frac{\mu_{PC}}{{\sigma_{PC}}^2}$$

Detection and Clustering of Putative Breakpoints

SVelter next scans the input alignment file to define putative breakpoints (BPs) where the sample genome differs from the reference. These are defined through the identification of aberrant read alignments. Clusters of read pairs (RP) showing abnormal insert length or aberrant mapping orientation may indicate breakpoints nearby, while reads with truncated (clipped) split read (SR) alignments are indicative of precise breakpoint positions. SVelter specifically defines aberrant reads as follows:

1. RPs outside expected IS ($\mu_{IS}\ \pm s \times \sigma_{IS}$, where s is the number of standard deviation from the mean, default as 3);

2. RPs that do not have forward reverse pair orientation;

3. SRs with high average base quality (i.e. 20) clipped portion with minimum size fraction of overall read length (i.e. 10 %).

It should be noted that the specific parameters listed were set as default empirically and can be adjusted by the user. Discordant RPs of the within a window of $mean\ IS + 2 \times std$ distance and of the same orientation are clustered together. Next, split reads within this window and downstream of the read direction are collated and the clipped position is considered as a putative breakpoint. If no such reads exist, the rightmost site of forward read clusters or leftmost site of reverse read clusters is assigned instead. For each cluster of aberrant RPs, a BP is assigned if the total number of split reads exceeds 20% of the read depth or the total number of all aberrant reads exceeds 30%. For samples of poorer quality, higher cutoffs might be preferred. Each putative BP will be paired with other BPs that's defined by mates of its supporting reads. BP pairs that intersect or are physically close (<1kb) to each other will be further grouped and reported as a BP cluster for the next step.

Random Iterative Rearrangement

For each BP cluster containing n putative BPs, a randomized iterative procedure is then applied on the n-1 genomic blocks between adjacent BPs. SVelter has three different modules implemented for this step: diploid module (default) that detects structural variants on both alleles simultaneous, heterozygous module that only report high quality heterozygous SVs and homozygous module for high quality homozygous SVs only. For the diploid module, a simple rearrangement (deletion, inversion or insertion) is randomly proposed and applied to each block on one allele while the other allele is kept unchanged and the newly formed structure is scored against the null models of expectation for each feature through the scoring scheme described below. A new structure is then selected probabilistically from the distribution of scores such that higher scores are more likely but not assured. The same approach is then applied to the other allelic structure representing a single iteration overall. For heterozygous and homozygous modules, only one allele is iteratively rearranged while the other allele remains either unchanged or is mirrored, respectively.

The iterative process will terminate and report a final rearranged structure if one of the following situations is met:

1. No changes to a structure after 100 continuous iterations

2. The maximum number of iterations is reached (100,000 as default)

After the initial termination, the structure is reset and the process is repeated for another 100 iterations while avoiding the fixed structure, at which point the highest scoring structure overall is chosen.

Structure Scoring

Assume S_j as the score of rearranged structure j. To estimate S_j, four different characteristics of RP $_i$: IS (IS_{ij}), Pair Orientation (PO_{ij}), RD (RD_{ij}), and Physical Coverage Through a BP (PC_{ij}) would be calculated and integrated. As the distribution of IS, RD, and Physical Coverage has been defined, the density function would be calculated and transformed to log scale:

$$Score\_IS_{ij} = log\ (f_{IS}\ \ (IS_{ij}))$$

$$Score\_RD_{ij} = log\big(f_{RD}\ \ (IS_{RD})\big)$$

$$Score\_PC_{ij} = log\ (f_{PC}\ \ (IS_{PC}))$$

Score of Pair Orientation is specified by the indicator function:

$$Score\_PB_{ij} = \begin{cases} 1, & if\ PO = Forward - Reverse \\ 0, & if\ other\ wise \end{cases}$$

Assuming total number of n pairs of reads are aligned in the targeted genomic region, for each structure j, individual scores of each RP would be integrated to form the structure score:

$$S_i = \sum_{i=1}^{n} Score\_IS_{ij} \times \left(1 + \frac{\sum_{i=1}^{n} Score\_PO_{ij}}{n}\right) + \tau \sum_{i=1}^{n} Score\_RD_{ij} \times \left(1 - \sum_{i=1}^{n} Score\_PC_{ij}\right)$$

where $\tau = \frac{log(f_{IS}\,(\mu_{IS}))}{log(f_{RD}\,(\mu_{RD}))}$ serves as the factor to regulate two parts into same scale.

2.4.2 Performance Assessment

Both simulated and real data were used to evaluate performance of SVelter. To produce simulation datasets, we altered the human GRCh37 reference genome to include both homozygous and heterozygous simple SVs and complex SVs independently while adding micro-insertions and short tandem repeats around the junctions in frequencies consistent with previously reported breakpoint characteristics (Kidd et al. 2010). Details about specific types of SVs simulated are summarized in Table 2.1 - 2.2. Paired-end reads of 101bp with an insert size of 500bp mean and 50bp standard deviation were simulated using wgsim (https://github.com/lh3/wgsim**)** across different read depths (10X, 20X, 30X, 40X, 50X).

For comparisons using real sequence data, we adopted two previously published samples: a haploid hydatidiform mole (CHM1) (Chaisson et al. 2014; Steinberg et al. 2014) and a well-characterized HapMap/1000 Genomes Project sample (NA12878) (Pendleton et al. 2015; Parikh et al. 2016). CHM1 has been deep sequenced by Illumina whole-genome sequence to 40X and by Single Molecule, Real-Time (SMRT) sequencing to 54X, and SVs of the sample have been detected and published by the same group as well (http://eichlerlab.gs.washington.edu/publications/chm1-structural-variation/). NA12878, together with the other 16 members from CEPH pedigree 1463, has been deep sequenced to 50X by Illumina HiSeq2000 system (http://www.illumina.com/platinumgenomes/). Additionally, the

Genome in a Bottle (GIAB) Consortium has published the PacBio sequencing data (20X) of NA12878 and also provided a set of high-confident SV calls (Chaisson et al. 2014; Parikh et al. 2016).

We assessed SVelter against four other algorithms with diverse approaches: Pindel, Delly, Lumpy, and ERDs. We applied these algorithms to both simulated and real data with default settings, except that SVelter's homozygous module was used for CHM1. All algorithms were compared using the same set of excludable regions and were run on the same computing cluster.

Table 2.1 Brief description of simulated simple SVs

Ref Structure	Alt Structure	Description	Number
a/a	/	Homo.DEL	3134
ab/ab	aba/aba	Homo.DUP	2962
a/a	aaaaaaa/aaaaaaa	Homo.DUP.Tandem	2970
a/a	a^/a^	Homo.INV	2936
ab/ab	ba/ba	Homo.TRA	2735
a/a	a/	Het.DEL	3128
ab/ab	ab/aba	Het.DUP	2918
a/a	aaaaaaa/a	Het.DUP.Tandem	2979
a/a	a/a^	Het.INV	2927
ab/ab	ba/ab	Het.TRA	2773

Table 2.2 Brief description of simulated complex SVs

Het				Homo			
Ref Structure,	Alt Structure	Description	Num	Ref Structure	Alt Structure	Description	Num
ab/ab	aab/ab^ab	FROM.NA12879	265	ab/ab	aab/ab^ab	FROM.NA12879	202
ab/ab	ab/aba	INS+DUP	283	ab/ab	aba/aba	INS+DUP	386
ab/ab	ab/aba^	INS+DUP	275	ab/ab	aba^/aba^	INS+DUP	390
ab/ab	ab/b^a^b	INS+DUP	304	ab/ab	b^a^b/b^a^b	INS+DUP	187
ab/ab	ab/b^ab	INS+DUP	294	ab/ab	b^ab/b^ab	INS+DUP	194
abc/abc	aa^b^c^c/abc	dup+INV+DUP	14	abc/abc	a^c/a^c	dup+INV+DUP	243
abc/abc	abc/a^c	INV+DEL	260	abc/abc	a^c^/a^c^	INV+DEL	236
abc/abc	abc/a^c^	INV+DEL	308	abc/abc	aa^b^c^c/aa^b^c^c	INV+DEL	12
abc/abc	abc/aba	INS+DUP+DEL	260	abc/abc	aba/aba	INS+DUP+DEL	479
abc/abc	abc/aba^	INS+DUP+DEL	309	abc/abc	aba^/aba^	INS+DUP+DEL	454
abc/abc	abc/ac^	INV+DEL	255	abc/abc	ac^/ac^	INV+DEL	228
abc/abc	abc/aca^	FROM.NA12881	271	abc/abc	ac^b^a^c/ac^b^a^c	FROM.NA12881	15
abc/abc	abc/b	Multi_DEL+INV	319	abc/abc	aca^/aca^	Multi_DEL+INV	218
abc/abc	abc/b^	Multi_DEL+INV	246	abc/abc	b/b	Multi_DEL+INV	238

abc/abc	abc/ba^c	FROM.NA12882	285	abc/abc	b^/b^	FROM.NA12882	259
abc/abc	abc/c^a^	INV+DEL	294	abc/abc	ba^c/ba^c	INV+DEL	245
abc/abc	abc/c^bc	INS+DUP+DEL	263	abc/abc	c^a^/c^a^	INS+DUP+DEL	229
abc/abc	abc/cbc	INS+DUP+DEL	297	abc/abc	c^bc/c^bc	INS+DUP+DEL	228
abc/abc	ac^b^a^c/abc	dup+INV+DUP	12	abc/abc	cbc/cbc	dup+INV+DUP	243
abcd/abcd	abc/ab^a^d	FROM.NA12880	290	abcd/abcd	abc/ab^a^d	FROM.NA12880	263
abcd/abcd	ad/b^a^d	FROM.NA12883	257	abcd/abcd	ad/b^a^d	FROM.NA12883	277
abcde/abcde	a/abd^e	FROM.NA12878	236	abcde/abcde	a/abd^e	FROM.NA12878	315
abcde/abcde	abcde/aba^e^de	INS+DEL+MultiDUP	257	abcde/abcde	aba^e^de/aba^e^de	INS+DEL+MultiDUP	255
abcde/abcde	abcde/aba^ede	INS+DEL+MultiDUP	267	abcde/abcde	aba^ede/aba^ede	INS+DEL+MultiDUP	273
abcde/abcde	abcde/abae^de	INS+DEL+MultiDUP	266	abcde/abcde	abae^de/abae^de	INS+DEL+MultiDUP	283
abcde/abcde	abcde/abaede	INS+DEL+MultiDUP	261	abcde/abcde	abaede/abaede	INS+DEL+MultiDUP	260
abcde/abcde	abcde/abe^a^de	INS+DEL+MultiDUP	270	abcde/abcde	abe^a^de/abe^a^de	INS+DEL+MultiDUP	282
abcde/abcde	abcde/abe^ade	INS+DEL+MultiDUP	268	abcde/abcde	abe^ade/abe^ade	INS+DEL+MultiDUP	295
abcde/abcde	abcde/abea^de	INS+DEL+MultiDUP	269	abcde/abcde	abea^de/abea^de	INS+DEL+MultiDUP	291
abcde/abcde	abcde/abeade	INS+DEL+MultiDUP	265	abcde/abcde	abeade/abeade	INS+DEL+MultiDUP	241

Assessment of Simulated Simple SVs

For simulated datasets, we compared the performance of each algorithm by calculating their sensitivity and positive predictive values (PPV) on each type of simple SV (deletion, disperse duplication, tandem duplication, inversion). As Lumpy reports breakpoints in terms of range, we calculated the median coordinate of each reported interval and consider it as the breakpoint for downstream comparison. A reported SV would be considered as a true positive (TP) if the genomic region it spanned overlapped with a simulated SV of the same type by over 50% reciprocally. As Delly and Lumpy didn't differentiate tandem and dispersed duplication in their SV report, we compare their reported duplications to both simulated tandem and dispersed duplications independently to calculate sensitivity, but use the entire set of simulated duplications together for the calculation of specificity. In this manner, the result will be biased towards higher TP and TN rates for these approaches. Dispersed duplications reported by Pindel were very rare and as such were processed in the same way as Delly and Lumpy.

Assessment of Real SVs

We initially made use of reported simple and complex SVs in CHM1 and NA12878 as gold standard sets, however the FP rate of each algorithm were high compared to previously published performance. To augment this set, we therefore have developed our own approach to validate simple and complex SVs using PacBio long reads. For each reported SV, we collect all PacBio reads that go through the targeted region and hard clip each read prior to the start of the region. We then compare each read to the local reference and an altered reference reflecting the structure of the reported SV by sliding a 10bp window through the PacBio read and aligning it against the reference sequence. Coordinates of each window are plotted against its aligned position in the form of a dotplot. Theoretically speaking, if a read was sampled from the reference genome, a diagonal line should be observed. However, if a read was sampled from an altered genomic region, a continuous diagonal line would only show when plotted against a correctly resolved sequence. In this manner, shorter SVs (<5kb) can be validated by accessing the deviation of all dots from diagonal. For each PacBio Read, the score:

$$Ratio\ of\ Dis = \frac{\sum_{j=1,2,\ldots,n_0} D_{ij} |\ (i = 0, PacBio\ read\ vs.original\ structure)}{\sum_{j=1,2,\ldots,n_0} D_{ij} |\ (i = 1, PacBio\ read\ vs.altered\ structure)} - 1$$

is assigned, so that a positive Ratio of Dis indicates the priority of altered genome over reference genome, and vise versa. The validation score of an SV is integrated from all PacBio reads spanning through it using an indicator function:

$$Structure\ Score = \frac{\sum I(Ratio\ of\ Dis > 0)}{total\ number\ of\ Pacbio\ Reads}$$

SVs with validation score >0.5 for haploid genome, or >0.3 for diploid genome would be considered validated.

For longer (>5kb) SVs, PacBio reads spanning through the whole targeted region are rarely observed. In this situation, we scored each breakpoint by adding 500bp flanks and assessing each individually. The final validation score is then determined through the collation of matches from all breakpoints.

We reassessed our initial true positive (TP) and false positive (FP) simple calls from each algorithm by combining our PacBio validated SVs from each algorithm together with the reported calls. For simple SVs, we utilized a 50% reciprocal overlap criterion. However, for CSVs we utilized a more complex comparison strategy to take into account that some algorithms will often detect individual parts of a complex rearrangement as distinct events. With each CSV predicted by SVelter, we extracted SVs with over 50% reciprocal overlap from other algorithms and calculated the validation score for each of them using our PacBio validation approach described above. When multiple SVs were extracted from an algorithm, averaged scores were adopted. Validation scores of a CSV from all algorithms were ranked and normalized from 0 to 1 for comparison.

2.5 RESULTS

2.5.1 Overview of SVelter

Our approach predicts the underlying structure of a genomic region through a two-step process. SVelter first identifies and clusters breakpoints (BP) defined by aberrant groups of reads that are linked across potentially related structural events. It then searches through candidate rearrangements using a randomized iterative process by which rearranged structures are randomly proposed and scored by statistical models of expected sequence characteristics (Figure 2.2; Materials and Methods).

SVelter begins by fitting statistical models for insert size (IS) and read depth (RD) based on paired-end sequences sampled from copy neutral genomic regions (Handsaker et al. 2015). Both are modeled as normal distributions for efficiency purposes which is recommended for relatively clean data sequenced at higher depth; however, more accurate but slower models (i.e. binomial for IL and negative binomial for RD) are also available as options for data of lower quality. SVelter then searches for and integrates potential SV signals from read pairs with aberrant insert size, orientation, and/or alignment (soft-clipping). Pairs of BPs are assigned simultaneously, and BP pairs that intersect with each other are further connected to form BP clusters. For each cluster containing n BPs, the n-1 genomic segments defined by adjacent BPs are then rearranged in a randomized iterative process whereby a simple SV (deletion, insertion, inversion) is randomly proposed and applied to all possible segments to assess the viability of each putative change. The initial structure and each subsequent rearranged structure are then scored by examining the impact of each change on various features of the sequence reads in the region, including insert size distribution, sequence coverage, physical coverage, and the relative orientation of the reads. A new structure is then chosen for the next iteration using a probability distribution generated from the structure scores. This continues until the algorithm converges on a final, stable set of rearrangements or a maximum number of iterations is reached.

An important feature of SVelter is that it simultaneously constructs and iterates over two structures, consistent with the zygosity of the human genome. This allows for the proper linking of breakpoint segments on the correct haplotypes, which is crucial for the proper resolution of overlapping structural changes that can often confuse or mislead other approaches. Individual breaks in the genome can then be properly linked and segregated, aiding in downstream genotyping across multiple individual sequences.

The randomized aspect of this approach leads to additional computation cost relative to other SV detection algorithms. We have addressed this by implementing a number of optimizations to increase the overall efficiency of SVelter. First, we limit the number of clustered BPs during the initial breakpoint-linking step in order to manage the number of random combinations at the next step. For regions with potentially higher numbers of linked breakpoints, we form subgroups based on physical distance between adjacent BPs that are later combined. Second, we set an upper and lower bound on the potential copy number (CN) of each segment between BPs using local read depth information and only allow structures containing CN-1 to CN+1 blocks for downstream analysis. Lastly, we have restricted the total number of iterations such that after converging on a stable rearrangement for 100 continuous iterations, we set this structure aside and restart the random iterations for another 100 iterations, at which point the highest scoring structure overall is chosen. This results in a total processing time for SVelter on a re-sequenced human genome with 50X coverage of under 20 hours when run in parallel on a high-performance computing cluster.

2.5.2 Performance Evaluation

We compared SVelter to three popular SV detection algorithms: Delly (Rausch et al. 2012), Lumpy (Layer et al. 2014), Pindel (Ye et al. 2009) and ERDS (Zhu et al. 2012). Both Delly and Lumpy have integrated insert size and split read information into their SV detection strategy, while Pindel implements a pattern grown approach to utilize split read alignments. While there are numerous other algorithms that have been developed for detecting SVs, we focused on these as they have previous published comparisons that can be transitively applied to our results.

Multiple experiments were conducted in order to evaluate our approach. We first simulated genomes of various sequence coverage containing both simple and complex SVs as homozygous and heterozygous events. We next applied these algorithms to the genome of a haploid hydatidiform mole (CHM1) (Chaisson et al. 2014; Steinberg et al. 2014) and also a well-characterized diploid genome (NA12878) (Pendleton et al. 2015; Parikh et al. 2016), both of which had reported high-confident calls as well as long-read Pacific Biosciences (PacBio) sequences available for orthogonal assessment. All algorithms were run either with the recommended settings as provided by the authors (where available) or default settings. Detailed commands for running each algorithm can be found in supplemental materials.

Simulated data

We simulated heterozygous and homozygous non-overlapping simple SVs (deletions, inversions, tandem duplications, dispersed duplications and translocations) of varied sizes into synthetic genomes sequenced at different depths of coverage (10-50X). We then calculated the sensitivity and positive predictive value (PPV) of each algorithm (Figure 2.3A,B, Figure 2.4, Figure 2.5). Overall, SVelter achieves a higher sensitivity and PPV for simple deletions compared to all other algorithms. Comparisons with duplications were more difficult; while all compared approaches can report tandem duplications, for dispersed duplications only SVelter reports both the duplicated sequence and its distal insertion point. We therefore took a conservative approach such that for calculating sensitivity we compared the full set of duplications predicted from each approach to the simulated set of tandem and dispersed events, but limited the false positive analysis to only tandem duplications for the other algorithms. It should be noted that this method of comparison would bias against SVelter to some extent, however under these circumstances SVelter still showed very good sensitivity and positive predictive value in calling dispersed duplications, with slightly worse performance for tandem duplications. For inversions, SVelter showed a comparable accuracy to other the algorithms.

We also simulated specific types of complex rearrangements based on structures recently reported (Sudmant et al. 2015b) as well as our own observations (Table 2.2). Performance comparisons with complex structures are less straightforward than with simple SVs as most algorithms are only designed to identify simple events, and therefore may predict portions of CSVs as independent events. We address this issue by considering the identification and predicted copy number of individual junctions as reported in the entire prediction set of each algorithm (deletions, duplications, inversions) and compared against each simulated complex event collectively, treating predicted non-simulated junctions in the region as false positives (Methods and Materials). SVelter performs consistently better in terms of sensitivity and PPV across almost all types of complex events, including inverted duplications and inversion deletion events (Figure 2.3 C,D).

To evaluate the performance of SVelter on somatic variation in cancer genomes, we made use of both previously generated as well as locally simulated germline and somatic variants from a recent study that describes an approach for detecting complex somatic variants by directly comparing tumor and matched normal sequence reads(Moncunill et al. 2014). We included the four SV detection algorithms described above as well as SMuFin for this comparison and focused on variants located on chr21 and chr22. These predefined sets only contained a small number of large SVs >100bp (18 germline and 14 somatic), and as such the sensitivity and PPV of each algorithm exhibited a loss of granularity. Nevertheless, for germline events SVelter achieved comparable sensitivity with consistently higher PPV when compared against Delly, Lumpy and Pindel, in agreement with the results above; SMuFin does not report germline calls and as such was not included in this comparison. For somatic events, sensitivity of SVelter, Delly and Lumpy are similar and consistently higher than SMuFin, which showed the highest PPV along with SVelter (Figure 2.6).

Real data

To estimate how SVelter performs on real data, we have applied each algorithm to two publicly available datasets: a haploid hydatidiform mole (CHM1) (Chaisson et al. 2014) and a well-characterized diploid genome analyzed by the Genome in a Bottle Consortium (NA12878)(Pendleton et al. 2015; Parikh et al. 2016). Both have been deep sequenced by Illumina short-insert and PacBio long-read sequencing, and provide an excellent foundation for comparing baseline accuracies among approaches. We initially compared deletion calls of each algorithm to the reported set of variants to determine their relative accuracy, however the false discovery rate of each algorithm was abnormally high with respect to previously reported values (Table 2.3), suggesting that the reported deletion set may be overly conservative. We therefore examined the PacBio data directly for each predicted variant using a custom validation approach that utilizes a recurrence strategy to compare each read to both the reference allele as well as a rearranged reference consistent with the predicted structure (Figure 2.7A,B, Methods and Materials). We evaluated this approach using sets of reported deletions in these samples as well as matched random sets located within copy neutral regions and found it to have very high true positive and true negative rates (Figure 2.7C). We also conducted PCR experiments on the predicted breakpoints of three predicted complex rearrangements that were validated with this approach to show convincing evidence for two, with inconclusive results for the third due to high degrees of repetitiveness in the region (Supplemental Figures 2.8-2.11). We then reassessed the earlier deletion predictions made by each algorithm in CHM1 and NA12878 by combining the previously reported deletions in each sample with those having PacBio validation support from our analysis. As expected, we observed a marked increase in accuracy for each algorithm (Figure 2.7D).

We next compared the performance of each algorithm on identifying and resolving CSVs. Given that there are very few reference sets available of known complex rearrangements, we first created a set of non-overlapping candidate CSVs as identified by SVelter in CHM1 and NA12878. We then collected all predictions from each algorithm that overlap that region and scored them using the validation approach above. As many complex rearrangements may be described as a combination of simple SVs, we utilized a ranking approach to compare the individual predictions by assigning 0 to the lowest scores and 0.75 to the highest scores (see Methods and Materials). We observed a significant enrichment of SVelter predictions with high validation scores, indicative of its efficacy in correctly resolving CSVs (Figure 2.12A). An example is shown in Figure 2.12B, which depicts a summary of sequence read alignments for a region on chromosome 1 in CHM1 containing multiple deletions as well as a local translocation. Using standard read clustering algorithms, the signals present might suggest the presence of a tandem duplication overlapping with large deletions. However, this is not consistent with the haploid nature of CHM1, and comparisons with long PacBio sequence reads that overlap the region show the true structure (Figure 2.12C), which when aligned to a rearranged reference using SVelter predictions shows a full length alignment (Figure 2.12D). A comparison with other algorithms indicates that their predictions are indeed consistent with analyzing each aberrant read cluster independently of each other and result in a combination of tandem duplications, deletions, and inversions (Figure 2.12E).

Table 2.3 True positive and false positive deletion calls made by each algorithm on NA12878 and CHM1, based on previous reported calls as well as our custom PacBio validation approach.

		Method	SV	TP	Ref. Calls	Total Calls	FP	TPR	PPV
NA12878	Simple Deletions called by each algorithm	SVelter	DEL	1753	2316	2988	1235	0.7569	0.5867
		Delly	DEL	1029	2316	1499	470	0.4443	0.6865
		Lumpy	DEL	1755	2316	2740	985	0.7578	0.6405
		Pindel	DEL	1833	2316	2349	516	0.7915	0.7803
	Simple Deletions compared to NIST set + Pacbio Validation Set	SVelter	DEL	2625	3970	2988	363	0.6612	0.8785
		Delly	DEL	1400	3970	1499	99	0.3526	0.9340
		Lumpy	DEL	2370	3970	2740	370	0.5970	0.8650
		Pindel	DEL	2236	3970	2349	113	0.5632	0.9519
CHM1	Simple Deletions compared to Chaisson et.al Set	SVelter	DEL	1149	3588	1890	741	0.3202	0.6079
		Delly	DEL	224	3588	773	549	0.0624	0.2898
		Lumpy	DEL	845	3588	1670	825	0.2355	0.5060
		Pindel	DEL	1131	3588	1672	541	0.3152	0.6764
	Simple Deletions compared to Chaisson et.al + Pacbio Validation Set	SVelter	DEL	1636	4425	1890	254	0.3697	0.8656
		Delly	DEL	609	4425	773	164	0.1376	0.7878
		Lumpy	DEL	1354	4425	1670	316	0.3060	0.8108
		Pindel	DEL	1521	4425	1672	151	0.3437	0.9097

Computational Runtime

We compared the overall executable runtime of the different software packages using a single chromosome from NA12878. For each algorithm, we initialized the analysis using a previously aligned sequence in BAM format and used the respective procedures necessary for each approach to result in a variant call file (see Methods and Materials). Delly was observed to complete the fastest, followed by Lumpy. Pindel and SVelter were both considerably slower and were comparable in their runtime (Table 2.4). It should be noted that some algorithms (e.g. Lumpy) can perform faster with optimized alignment strategies (Chiang et al. 2014), however this was not included in our assessment.

Table 2.4 Run time of different algorithms on isolated post-processed alignment file of chromosome 21 from NA12878

	Algorithm	mem /core(GB)	time / core (seconds)
chr21 of NA12878	SVelter	20	8647.94
	Delly	20	810.86
	Lumpy	20	1105.10
	Pindel	20	7220.31

2.5.3 Examination of Identified SVs in CHM1 and NA12878

We examined the full set of identified simple and complex SVs in both CHM1 and NA12878. As expected, we rediscovered many previously reported deletions, duplications and inversions (Table 2.5). In some cases, we were also able to identify dispersed duplications that were incorrectly identified as overlapping tandem duplication and deletion events in prior reports (Figure 2.13a, Supplemental Figure 2.14). Furthermore, we found a recurrence of particular types of CSVs, including inverted-duplication and deletion-inversion events (Figure 2.13b,c,d, Supplemental Figures 2.15-2.17) suggesting that they are likely more common than previously thought. However, there were numerous other CSVs that could not be coalesced into a single classification and may provide future insight into new mechanisms for SV formation.

Table 2.5 Predicted SV Types in CHM1 and NA12878 by SVelter.

SV Type	CHM1	NA12878
Simple DEL	1890 (0.86)	2988 (0.84)
Simple DUP	1872 (0.28)	1232 (0.41)
Tandem	1725(0.27)	1184 (0.41)
Dispersed	147 (0.37)	48 (0.40)
Simple INV	14 (0.50)	106 (0.76)
Simple TRA	6 (0.67)	3 (0.67)
DEL+DUP	6 (0.50)	39 (0.49)
DEL+INV	5 (0.40)	16 (0.44)
DEL+TRA	3 (0.67)	3 (0.67)
DUP+INV	188 (0.64)	141 (0.44)
DEL+DUP+INV	8 (0.50)	34 (0.32)
Other	27 (0.37)	369 (0.70)

Numbers in parenthesis indicate percentage of calls with PacBio validation support. The remaining calls either were not able to be assayed with our approach or were invalidated.

2.6 DISCUSSION

We have described an integrative approach, SVelter, that can identify both simple and complex structural variants through an iterative randomization process. We show that it has an improved or comparable accuracy to other algorithms when detecting deletions, duplications and inversions but has the additional capability of correctly interpreting and resolving more complex genomic rearrangements with three or more breakpoints. Furthermore, SVelter can resolve structural changes on parental haplotypes individually, allowing for the correct stratification of multiple overlapping SVs. SVelter achieves this by forgoing the assumption of specific patterns of read alignment aberrations as associated with individual rearrangements and instead allowing the underlying sequence itself to dictate the most probable structure.

The ability to accurately identify CSVs in whole genome sequence data is a significant advancement, as currently many such regions are either missed or identified as individual errant events. For example, in our investigation of NA12878 we identified many disperse duplications that were previously reported as overlapping deletion and tandem duplication events as well as other simple deletions and inversions that were in fact part of a larger complex rearrangement (Figure 2.12). Such regions could be, in part, responsible for the observed discrepancies when comparing different SV algorithms with each other as well as other platforms such as array-CGH (Pinto et al. 2011). Our observations are also consistent with recent findings by the 1000 Genomes Project (Sudmant et al. 2015b), however their analysis required the use of multiple long-read sequencing technologies including PacBio and Moleculo to interpret the regions while SVelter is able to correctly resolve the regions from short-insert Illumina sequences alone. Although long-read technologies are very well suited for such an application, their use is currently limited to small-scale projects and there have been estimates that over 300,000 genomes will be sequenced using Illumina short-insert reads in 2015 alone. Thus, approaches like SVelter that perform accurately on such data sets are likely to have a larger impact on correctly reporting complex structural genomic aberrations.

One limitation of SVelter is that even with our efficiency enhancements it still exhibits a longer processing time with respect to the other SV algorithms compared here. This is in part due to the randomization strategy but is also owing to the inclusion of a read coverage component, which is not modeled in the other approaches we compared against but contributes to the overall increased accuracy of SVelter. Recent advances have made it possible to analyze a high coverage human genome from sequence to variant calling and annotation in half a day (Chiang et al. 2014), and such applications are very useful for diagnostic applications where speed is a critical component. Nevertheless, the enhanced ability of SVelter to correctly resolve overlapping and complex rearrangements relative to other approaches makes it very useful for projects where the accurate detection of such regions is important. Another limitation of SVelter is that in its current form it has a reduced ability to delineate heterogeneous data, such as commonly found when sequencing cancer genomes. This is due to its expectation of a specific ploidy when iterating between multiple haplotypes. Future work in this area will focus on creating a dynamic structure that can allow different levels of heterogeneity or mosaicism.

2.7 CONCLUSIONS

We have developed and applied a new approach to accurately detect and correctly interpret both simple and complex structural genomic rearrangements. Our comparisons to existing algorithms and data sets show that SVelter is very well suited to identifying all forms of balanced and unbalanced structural variation in whole genome sequencing data sets.

2.8 SOFTWARE AND DATA AVAILABILTY

The software package SVelter is available for download at https://github.com/mills-lab/svelter, and additional documentation regarding specific software usage and parameters, supporting files, algorithm comparisons and simulated data sets are provided at this site.

Sequence data used in this analysis were obtained from the following resources:

CHM1 – Resolving the complexity of the human genome using single-molecule sequencing (http://eichlerlab.gs.washington.edu/publications/chm1-structural-variation/) (Chaisson et al. 2014)

NA12878 – Genome in a Bottle Consortium (https://sites.stanford.edu/abms/giab) (Pendleton et al. 2015), Illumina Platinum Genomes (http://www.illumina.com/platinumgenomes/)

2.9 ACKNOWLEDGMENTS

This project was supported in part through funds from the University of Michigan, the NIH/NHGRI (1R01-HG007068-01A1), and NIH/Common Fund (DP5OD009154).

2.10 FIGURES

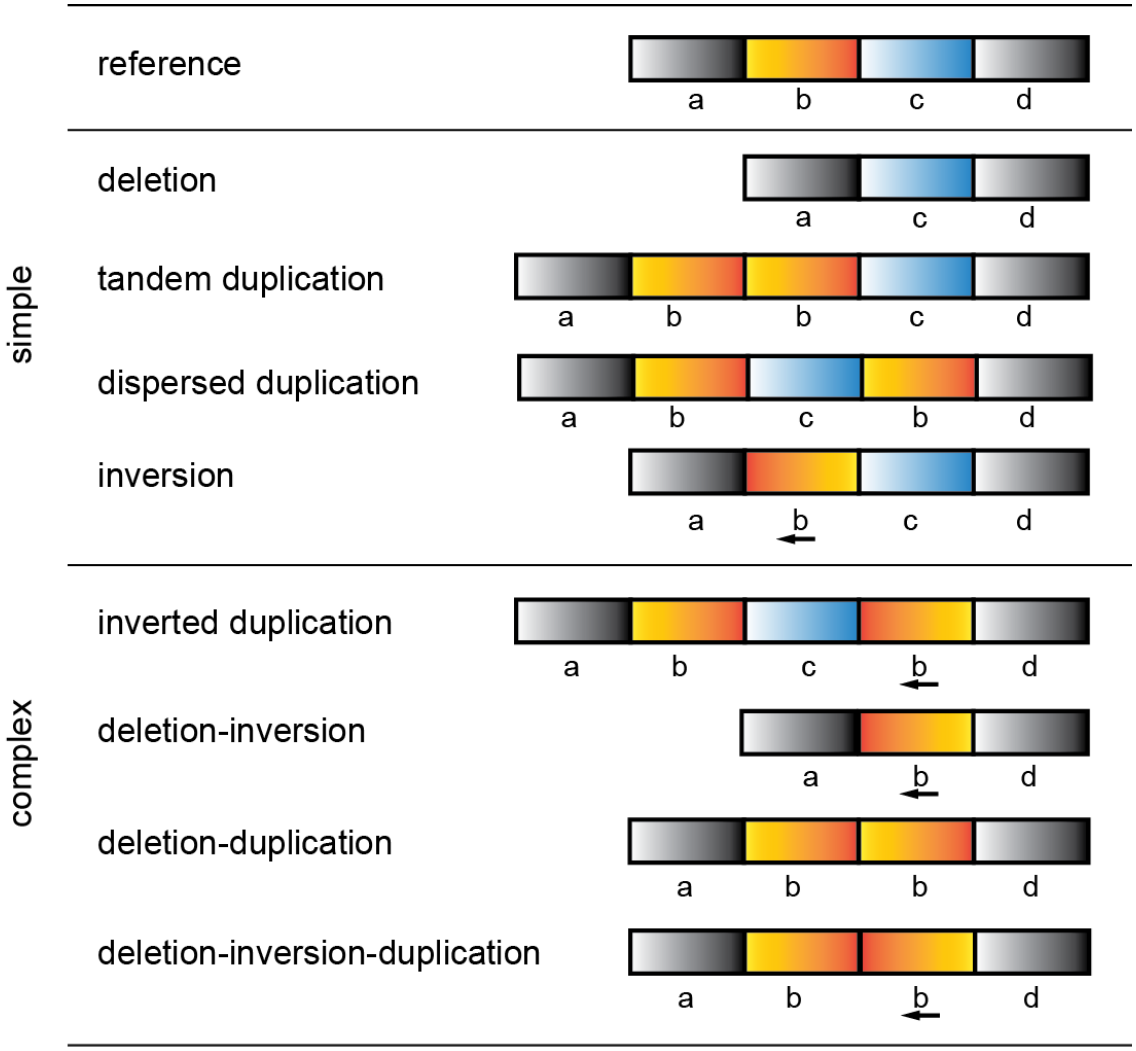


Figure 2.1 Illustration of simple and complex rearrangements, as compared to an unaltered reference genome.

Simple rearrangements are typically defined by two breakpoints, although dispersed duplications include an additional breakpoint at the insertion site. Examples of commonly observed complex structural variants with three or more breakpoints are provided but are not inclusive.

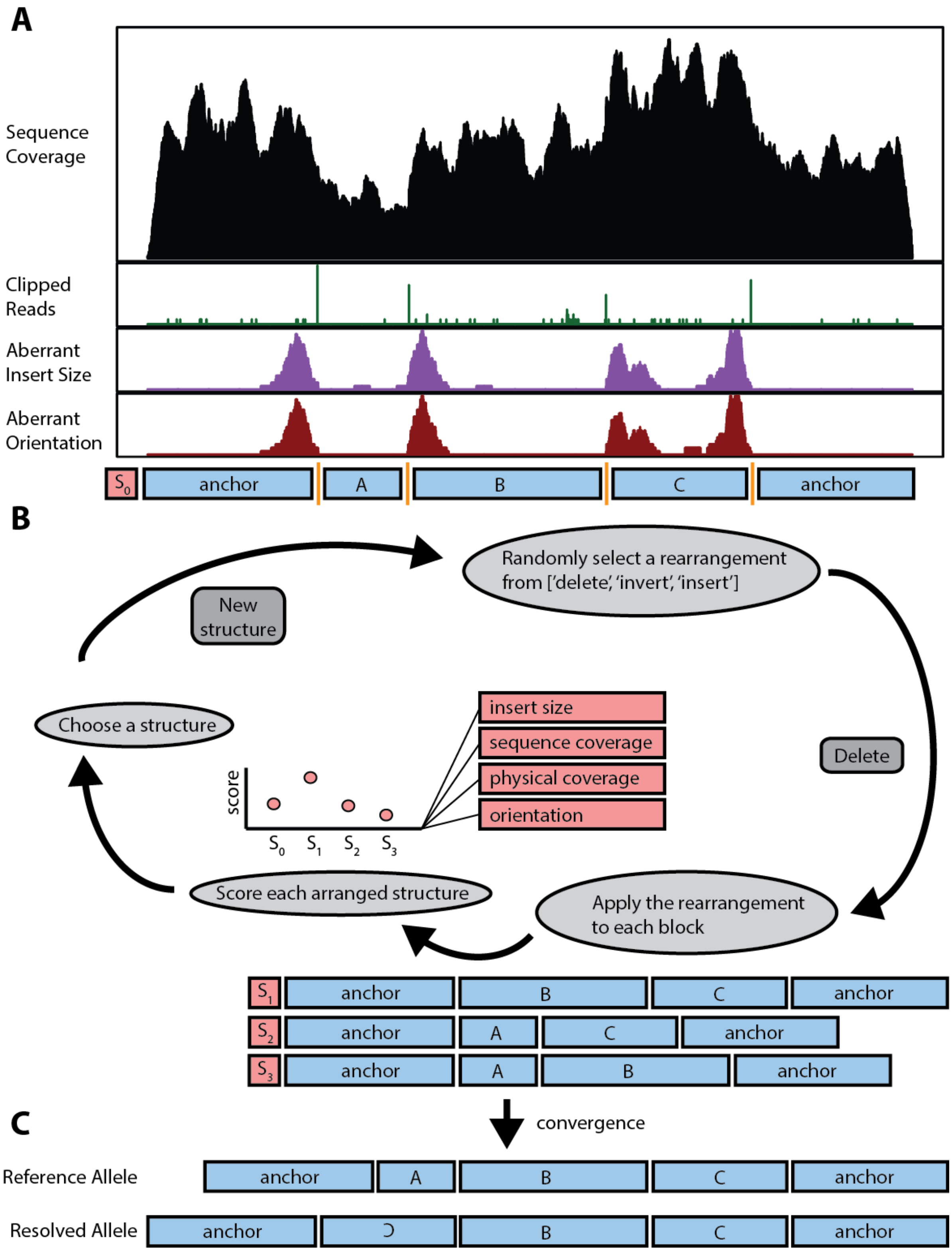

A
Sequence Coverage
Clipped Reads
Aberrant Insert Size
Aberrant Orientation
S_0
anchor
A
B
C
anchor
B
Randomly select a rearrangement from ['delete', 'invert', 'insert']
New structure
Choose a structure
insert size
sequence coverage
physical coverage
orientation
Delete
score
S_0
S_1
S_2
S_3
Score each arranged structure
Apply the rearrangement to each block
S_1
anchor
B
C
anchor
S_2
anchor
A
C
anchor
S_3
anchor
A
B
anchor
convergence
C
Reference Allele
anchor
A
B
C
anchor
Resolved Allele
anchor
Ɔ
B
C
anchor

Figure 2.2 Overview of computational strategy for identifying structural variation in whole genome sequences.

(A) SVelter first scans the genome and identifies clusters of aberrant read characteristics. These are used to create a putative set of breakpoint positions. (B) The segments between breakpoints are then iteratively rearranged and scored against null models of sequence expectations. (C) T he final converged structure is reported as the predicted structural rearrangement for the region.

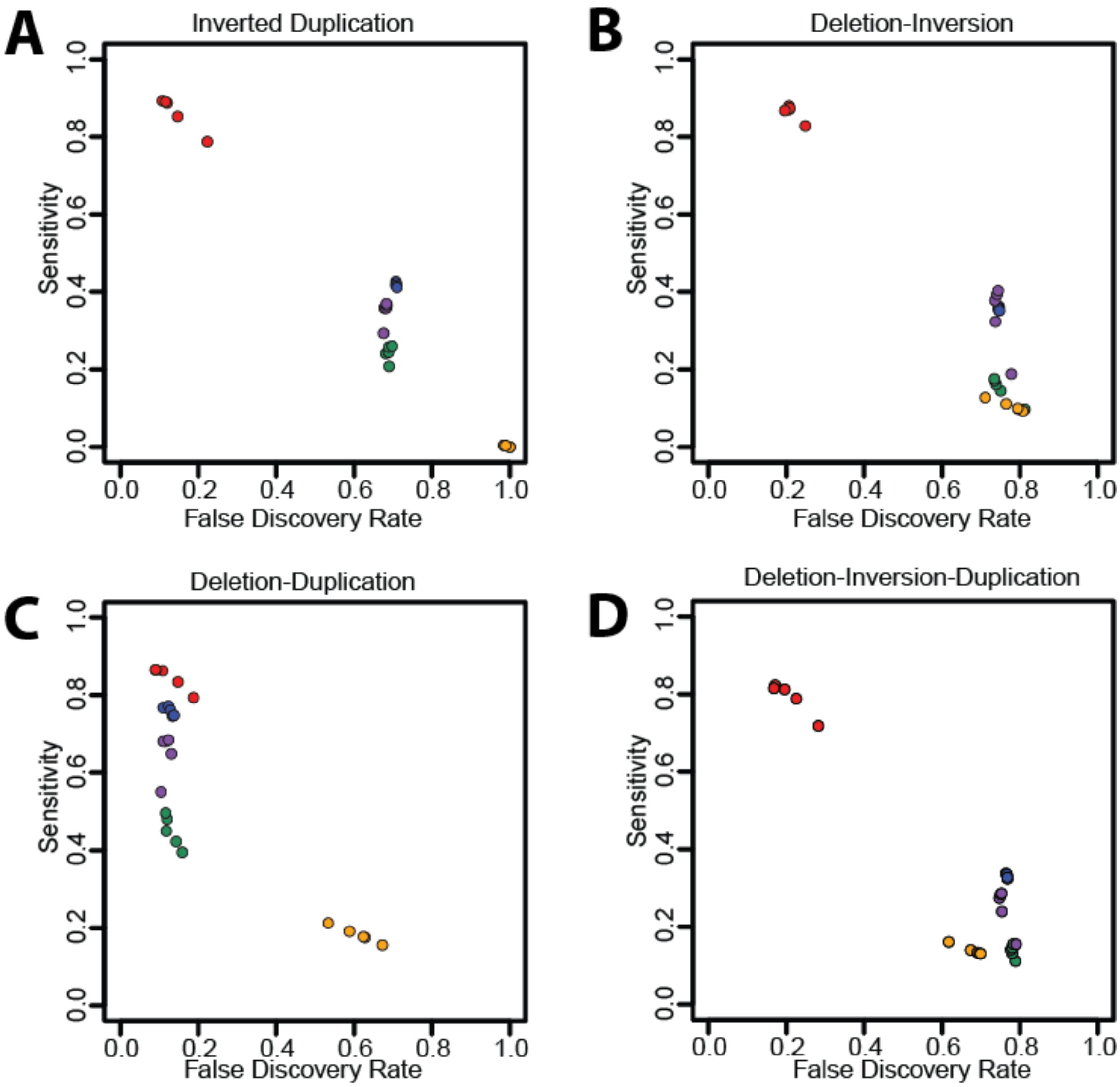


Figure 2.3 Assessment of complex structural variation accuracy using simulated data sets.

Sensitivity and false discovery rates for SVelter (red), Delly (blue), Lumpy (Green), Pindel (purple), and ERDS (yellow) on simulated (A) inverted duplications, (B) deletion inversions, (C) deletion duplications and (D) deletion-inversion-duplication events.

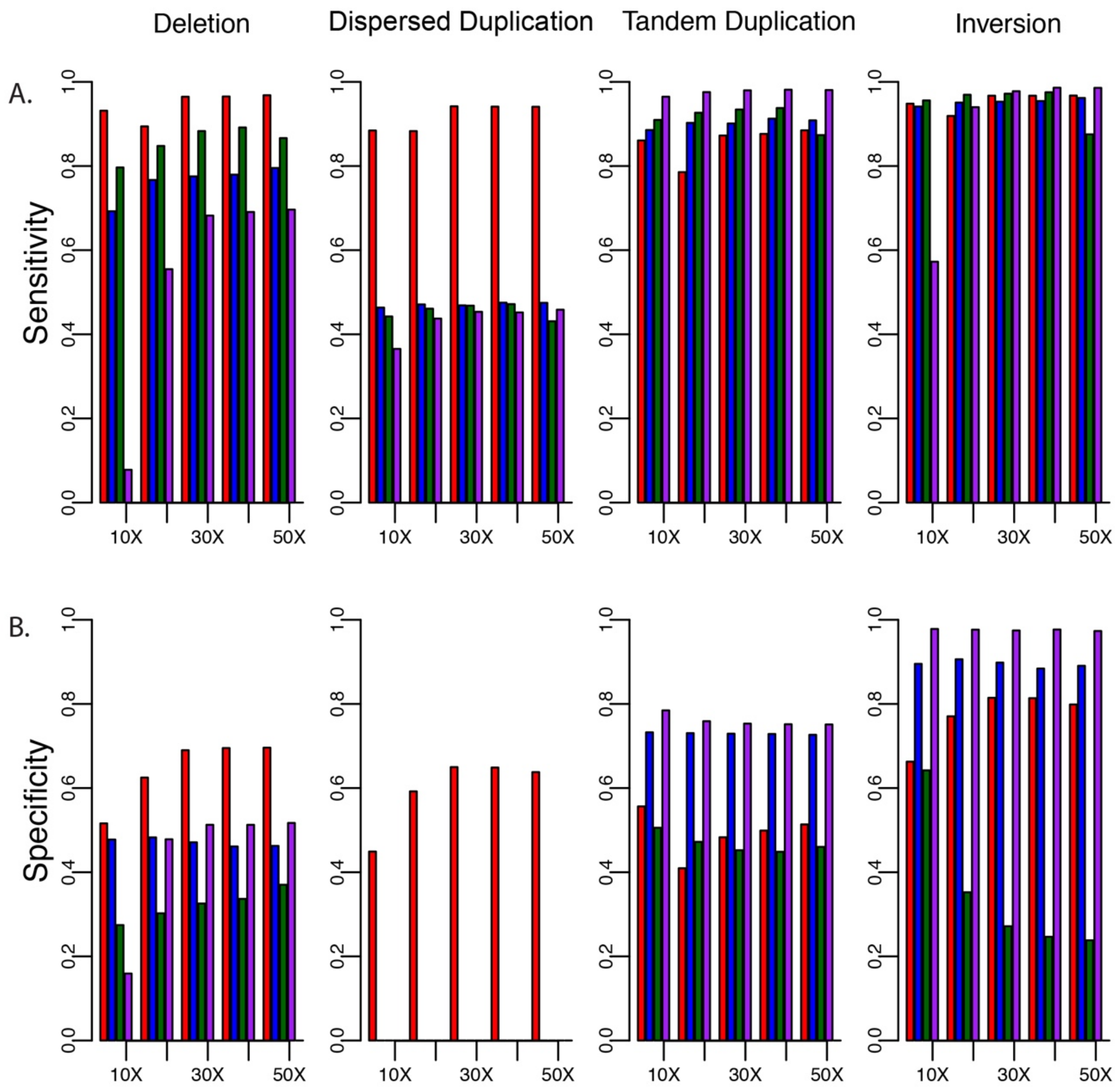


Figure 2.4 Assessment of accuracy on simulated homozygous events.

(A) Sensitivity and (B) positive predictive values for SVelter (red), Delly (blue), Lumpy (Green), and Pindel (purple) across different simple SV types and sequence coverage on combined simulated homozygous and heterozygous events. For dispersed duplications, only SVelter was considered for positive predictive values and all predictions by other algorithms that did not overlap simulated results were considered only for the tandem duplication category.

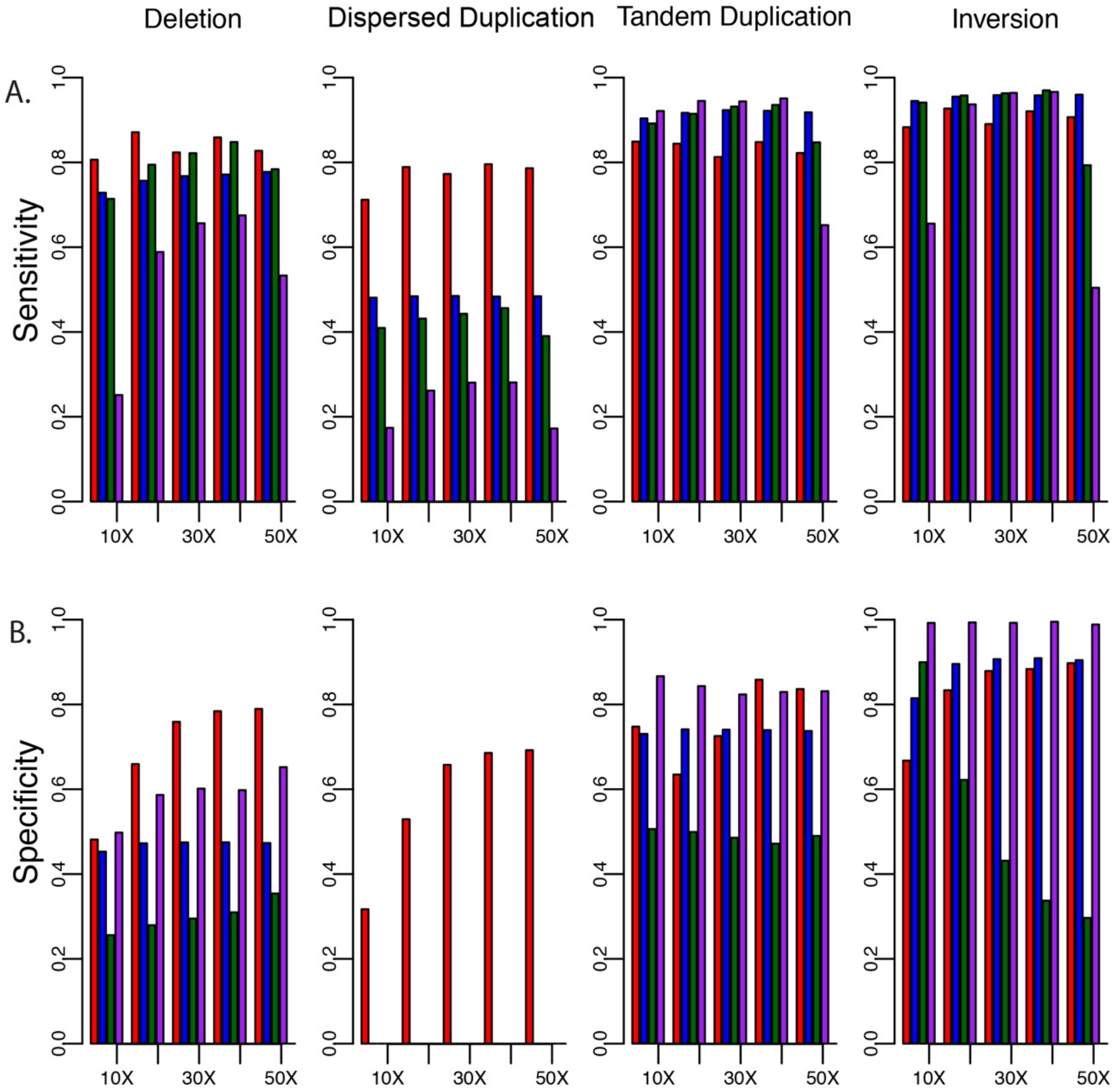


Figure 2.5 Assessment of accuracy on simulated heterozygous events.

(A) Sensitivity and (B) positive predictive values for SVelter (red), Delly (blue), Lumpy (green), and Pindel (purple) across different simple SV types and sequence coverage on combined simulated homozygous and heterozygous events. For dispersed duplications, only SVelter was considered for positive predictive values and all predictions by other algorithms that did not overlap simulated results were considered only for the tandem duplication category.

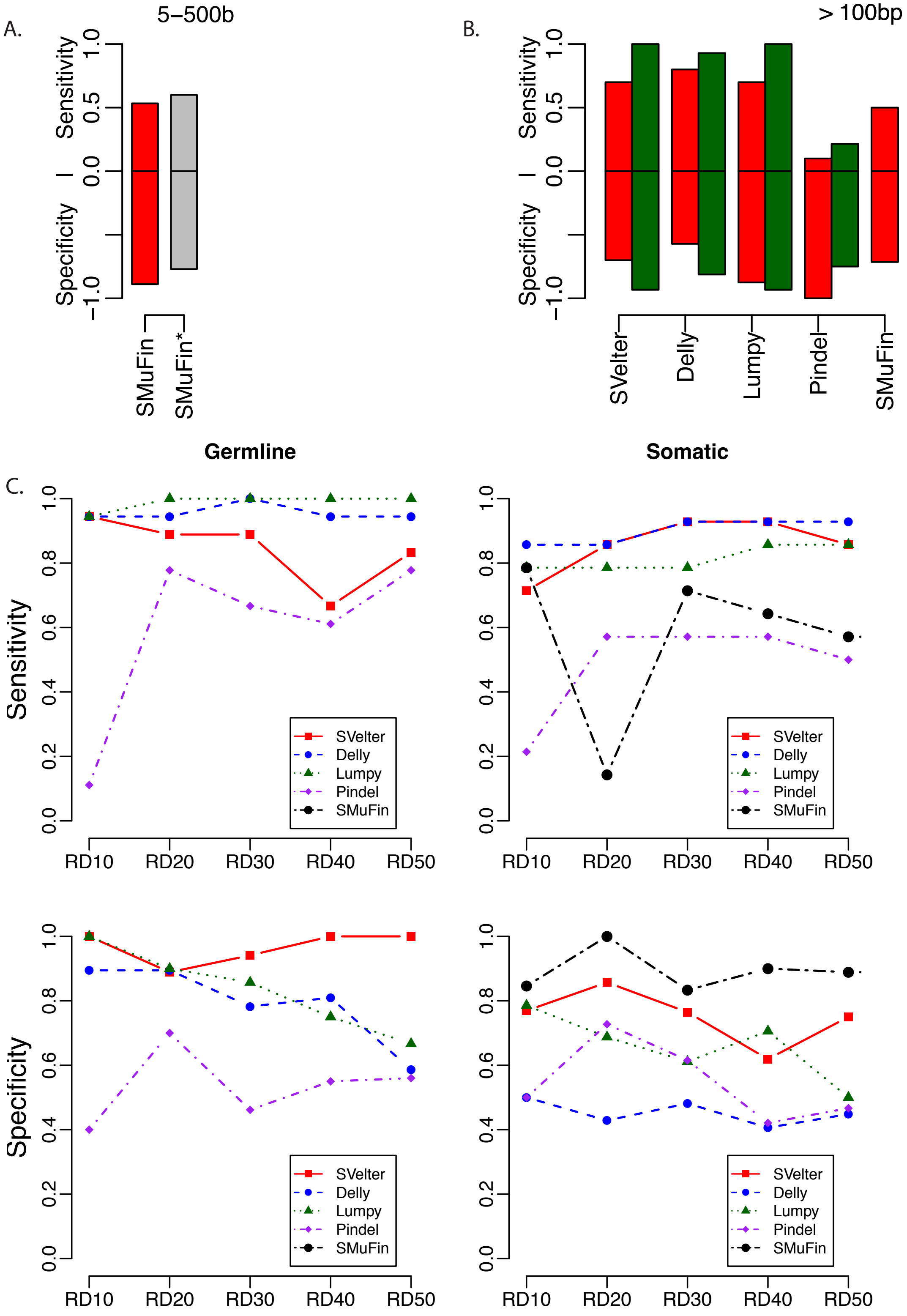

A.
5–500b
Sensitivity
Specificity
1.0
0.5
0.0
−1.0
SMuFin
SMuFin*
B.
> 100bp
Sensitivity
Specificity
1.0
0.5
0.0
−1.0
SVelter
Delly
Lumpy
Pindel
SMuFin
C.
Germline
Somatic
Sensitivity
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
RD10
RD20
RD30
RD40
RD50
SVelter
Delly
Lumpy
Pindel
SMuFin

Figure 2.6 Assessment of accuracy on simulated tumor and matched normal genomes.

(A) Sensitivity and specificity of SMuFin as applied locally on chr22 (left panel) or estimated from the authors original publication from the entire genome (Supplemental Figure 2 in Moncunill et al. 2014) for comparison of predicted SVs ranging from 5 to 500bp at RD30 (right panel, *). This shows consistency with both our application of this algorithm and between single chromosome and whole genome results. (B) Sensitivity and specificity of multiple algorithms on chr22 at RD30 for SVs over 100bp. (C). Sensitivity and specificity of multiple algorithms on chr21 and chr22 at different coverage using simulated matched germline and somatic data generated locally using the Moncunill et al set of variant calls and simulation strategy. SMuFin is absent in germline data as it only reports somatic events.

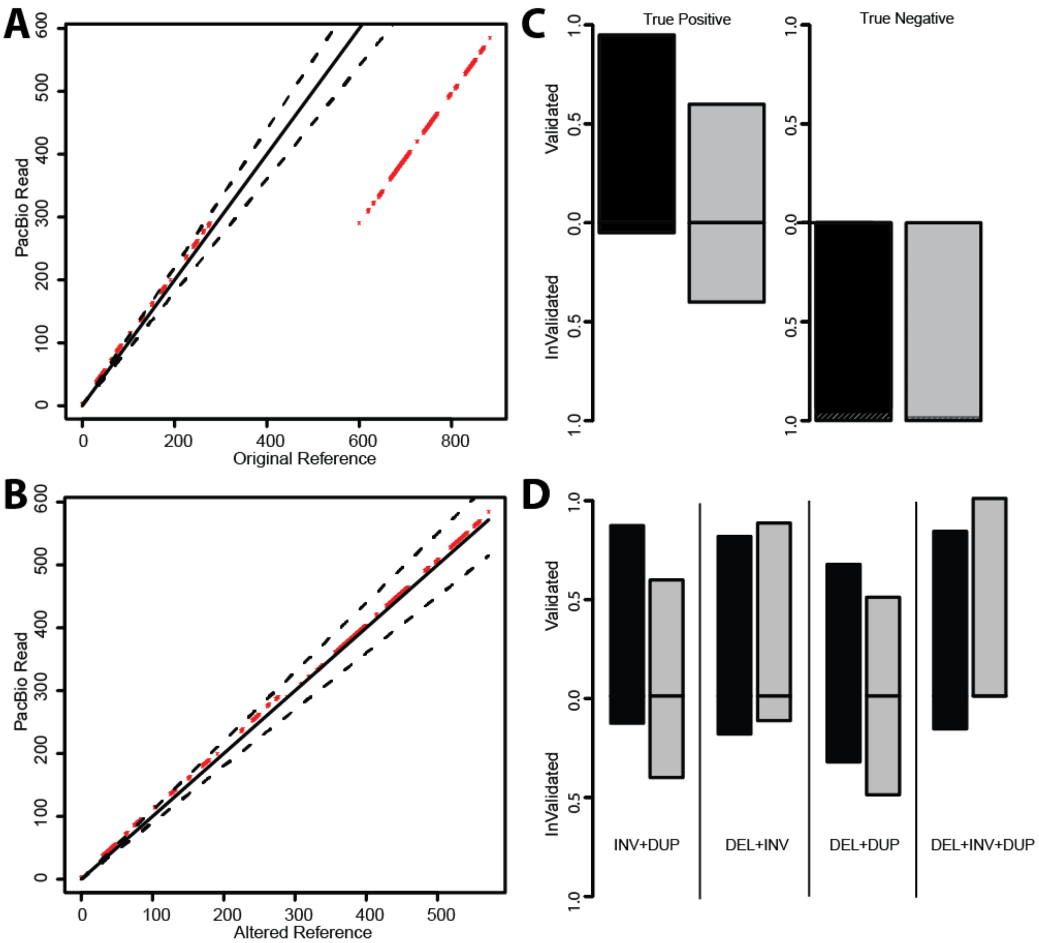


Figure 2.7 Overview and application of PacBio validation approach to human data.

(A) D ot plot of example region containing a simple deletion u sing a single PacBio read against the reference genome. Red dots indicate matches between sequences and dashed black lines delineate 10% deviance from the diagonal. (B) Dot plot of same region using an altered reference incorporating the deletion event. (C) Fraction of true positive calls using validation approach on published deletions in NA12878 (black) and CHM1 (grey) and CN2 regions as negative controls. Dashed black lines indicate regions that could not be assessed due to lack of PacBio reads to interrogate. (D) Assessment of specific predicted complex structures by SVelter using PacBio reads in NA12878 (black) and CHM1 (grey).

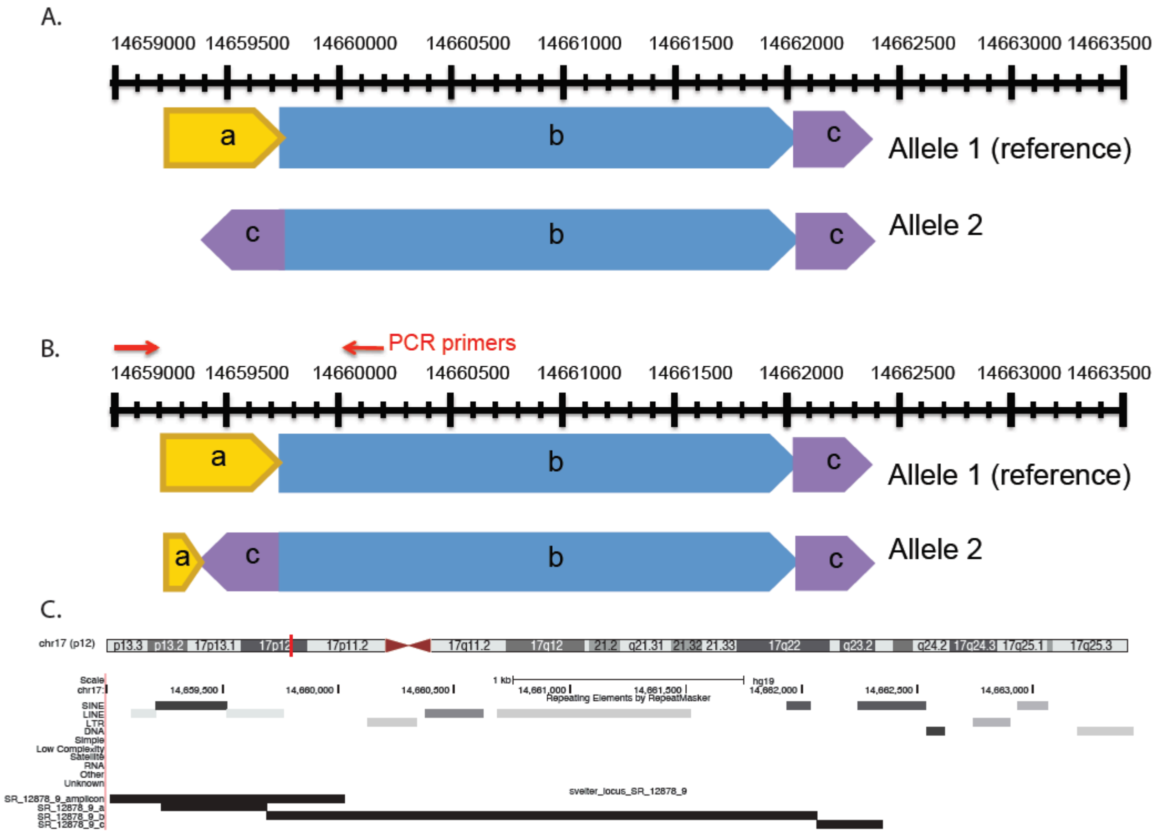


Figure 2.8 Validation of inverted duplication (chr17_14659237_14662064_14662349_14662516_C) on at locus chr17:14659000-14663500.

(A) Predicted structure of CSV. (B) PCR primer strategy and resulting observed structured. Primer sequences S91: GTGCACAGGATTGCTTCTGA, S92: TGTGTGGCTTTGACCACAAT. (C) Graphical representation of predicted structure and observed PCR product.

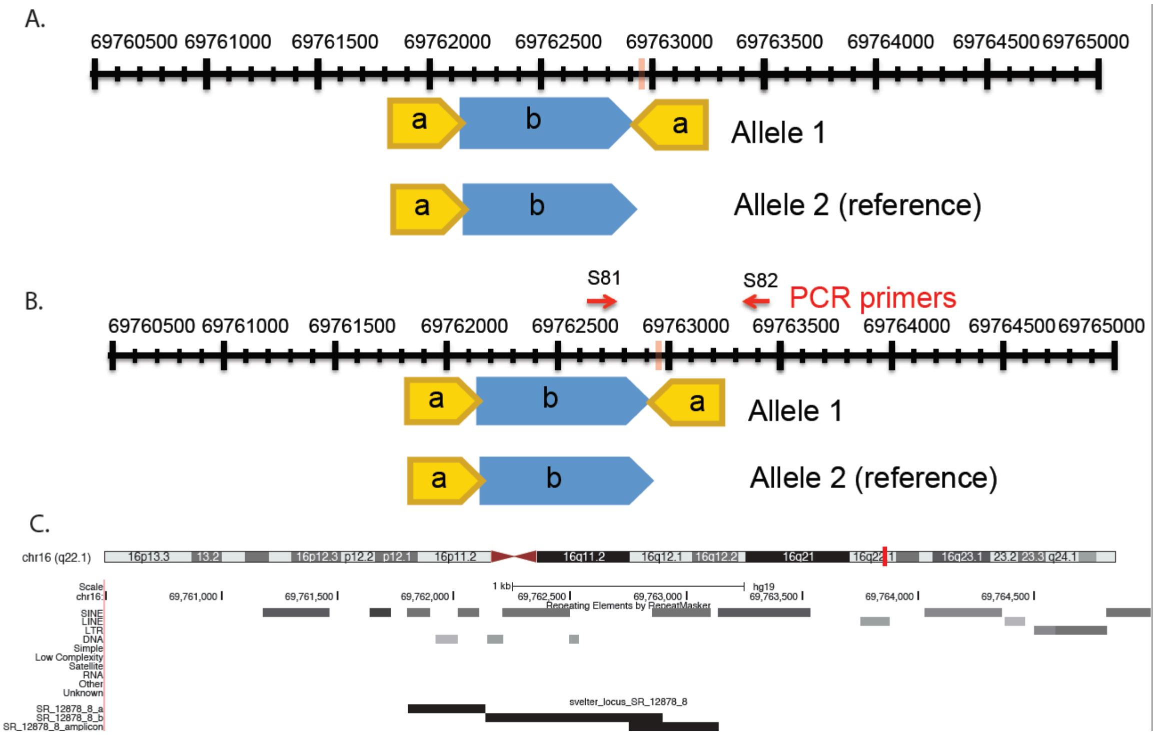


Figure 2.9 Validation of inverted duplication (chr16_69761804_69762136_69762896_69766900_C) on at locus chr16:69760500-69765500.

(A) Predicted structure of CSV. (B) PCR primer strategy and resulting observed structured. Primer sequences S81: CCCATCCCAAGTCATCTCAT, S82: AAATGTCTGTCTTTACCACTGTGTAG. (C) Graphical representation of predicted structure and observed PCR product.

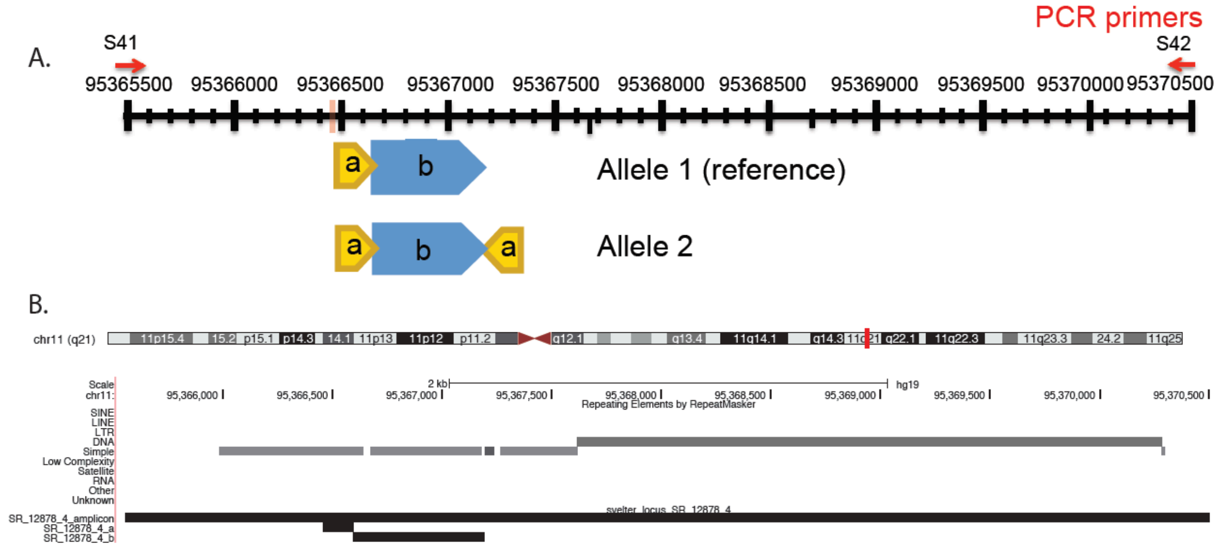
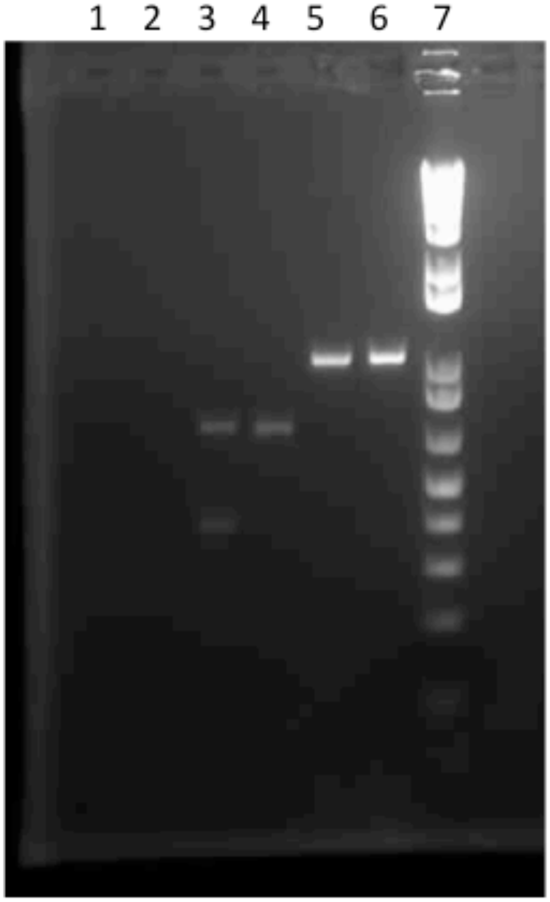


Figure 2.10 Attempted validation of inverted duplication (chr11_95366462_95366593_95367193_C) at locus chr11:95365500-95375000

This region contained numerous repetitive elements making direct PCR validation difficult such that no distinct bands were produced resulting in an inconclusive validation status. (A) PCR primer strategy and predicted structured. Primer sequences S41: GCCAGGCAGTCAGAATTAGC, S42: TCCCTGAGGACAGGAACAAC. (B) Graphical representation of predicted structures.



1. NA12878 with primers S41-S42
2. NA19240 with primers S41-S42
3. NA12878 with primers S91-S92
4. NA19240 with primers S91-S92
5. NA12878 with primers S81-S82
6. NA19240 with primers S81-S82
7. Ladder

Figure2.11 Electrophoresis GEL image of PCR products for each primer pair in two different samples (NA12878, NA19240), as outlined in Supplemental Figures 1-3.

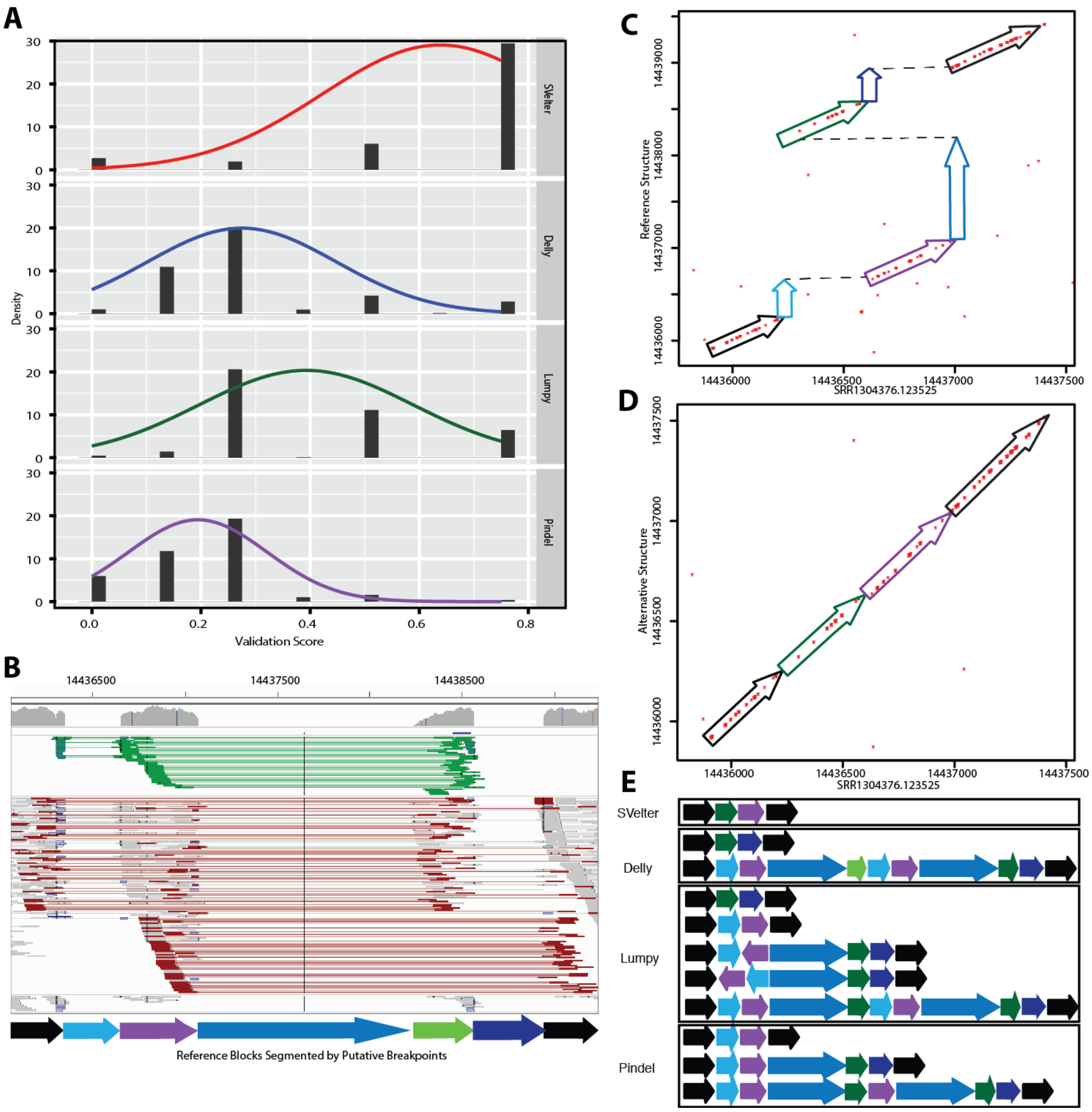


Figure 2.12 Evaluation of complex structural variation predictions.

(A) Validation scores of complex structural variation predicted in NA12878 from all algorithms ranked and normalized from 0 to 1 for comparison. For approaches with multiple predicted SVs in a region, average scores from each prediction were averaged. (B) I GV screenshot of example complex region in CHM1 (chr1:14435000-1444000) containing multiple deletions (blue shaded arrows) and a translocated region (green arrow), with surrounding anchor regions in black. Light green lines in IGV indicate read pairs with reverse-forward orientation, while red lines indicate read pairs with aberrant insert size length. (C) D ot plot of region between an individual PacBio read (SRR1304376.123525) against the reference sequence. Colored arrows correspond to segments indicated in (B). (D) D ot plot of altered reference sequence implementing predicted rearrangements by SVelter. (E) S chematic of predictions by each SV algorithm with respect to segments indicated in (B). For approaches with multiple predictions overlapping the region, each predicted SV is show independently.

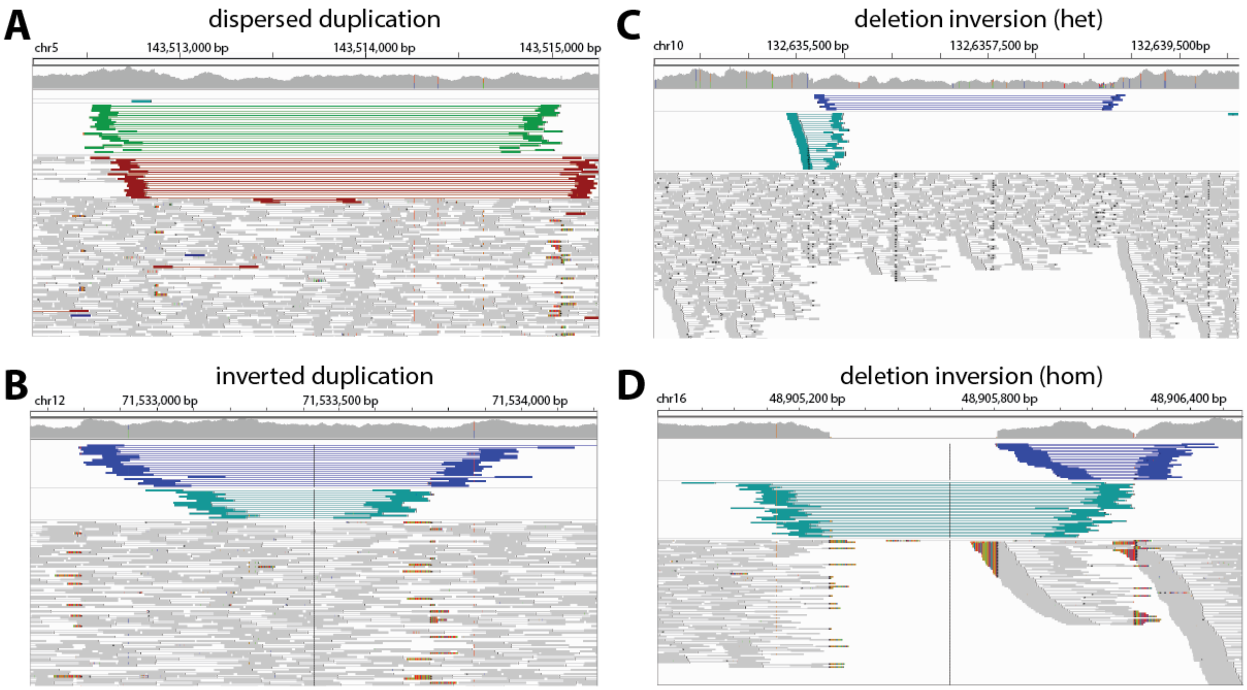


Figure 2.13 Examples of various types of complex structural variation in NA12878 identified by SVelter.

(A) I GV screenshot of disperse duplication event predicted by SVelter. Line colors as described in Figure 4. Such regions are typically identified as an overlapping tandem duplication and deletion. (B) Example of inverted duplication event. Blue lines in IGV indicated reverse-reverse read pair orientation while dark green lines indicate forward-forward read pair orientation. (C) R egion with heterozygous inversion and deletion rearrangement. (D) Region with homozygous inversion and deletion rearrangement. All regions shown had PacBio sequences consistent with predicted SVelter structures and were misclassified by other approaches (Supplemental Figures 8-11)

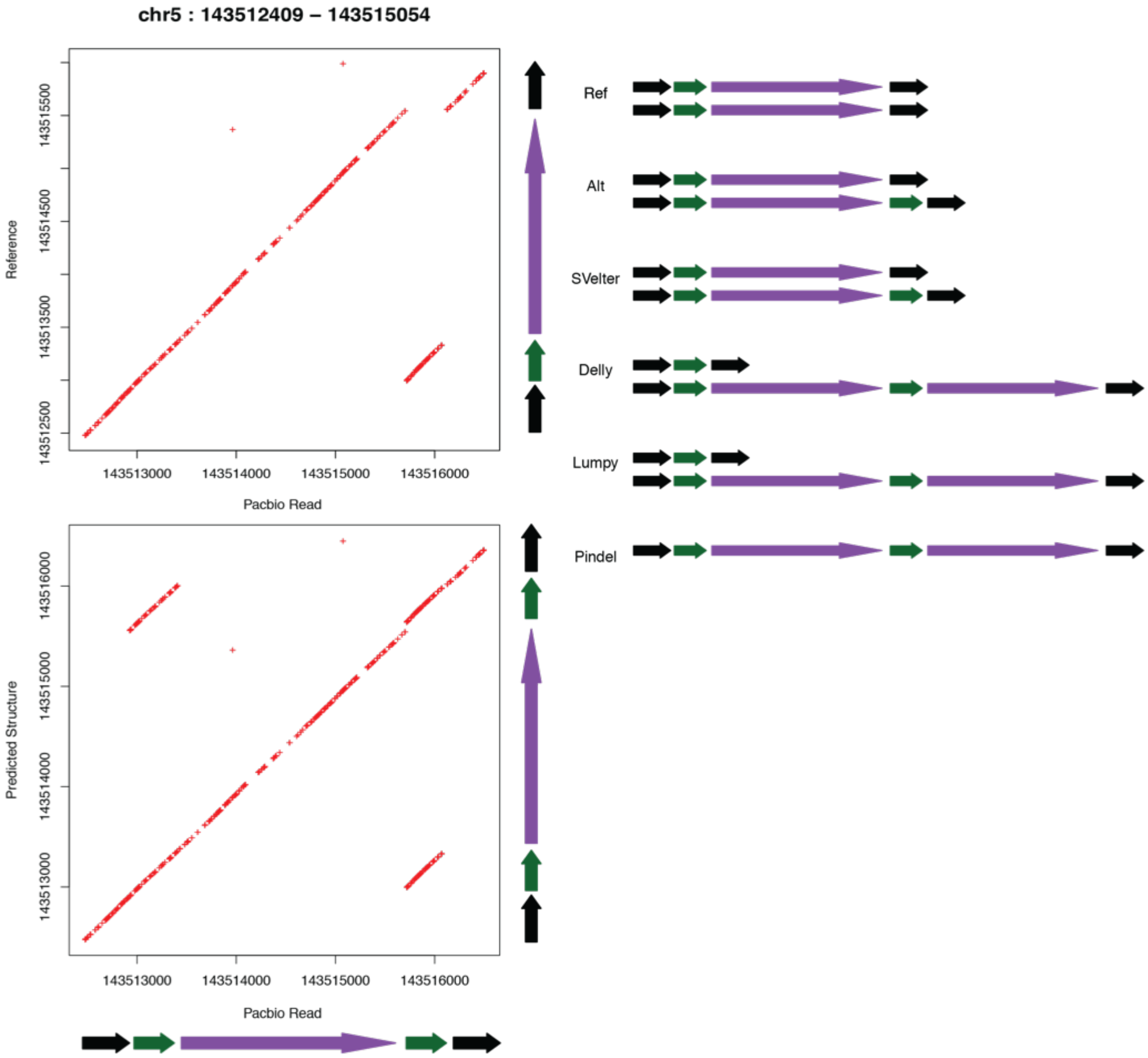


Figure 2.14 Dot plot of a PacBio read (fa716c69_55756_0) from NA12878 against both unaltered reference sequence (chr5:143512409-143515054) and modified reference sequence containing the predicted rearrangement.

Colored arrows on the right side indicate reference and alternative structures as diploid arrangements, as well as predictions from each individual algorithm. Ploidy for individual approaches is based on reported genotypes where available.

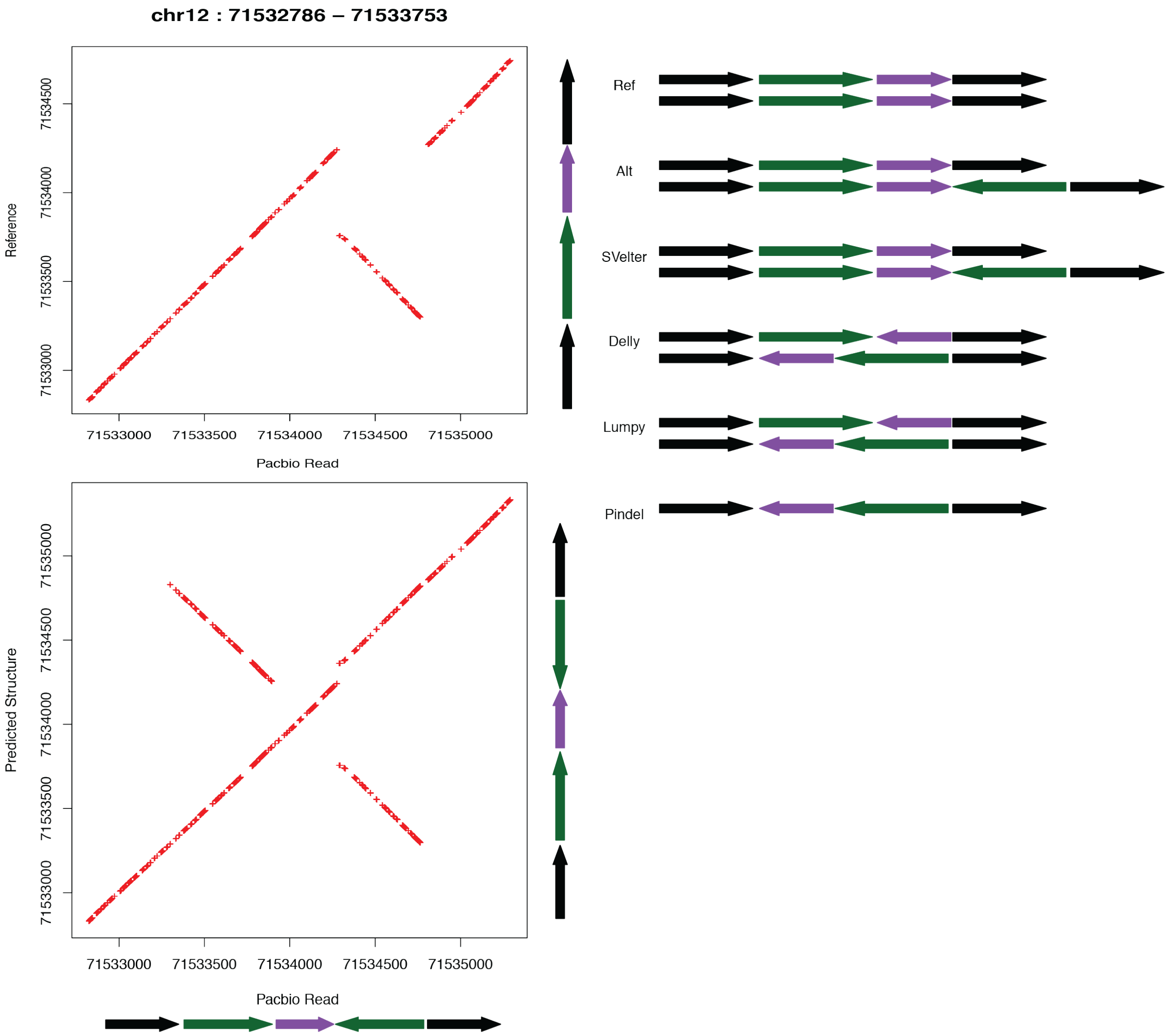


Figure 2.15 Dot plot of a PacBio read (5308fbec_46356_10341) from NA12878 against both unaltered reference sequence (chr12:71532786-71533753) and modified reference sequence containing the predicted rearrangement.

Colored arrows on the right side indicate reference and alternative structures as diploid arrangements, as well as predictions from each individual algorithm. Ploidy for individual approaches is based on reported genotypes where available.

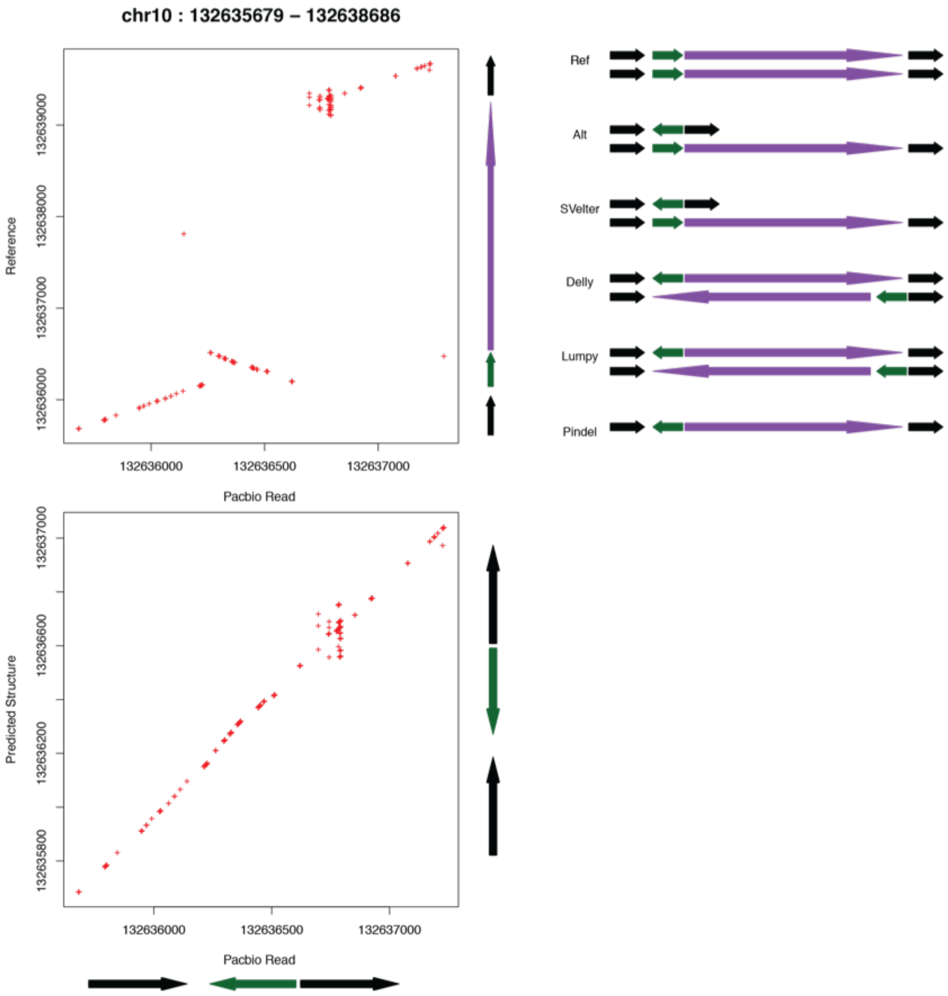


Figure 2.16 Dot plot of a PacBio read (325320e3_146839_203) from NA12878 against both unaltered reference sequence (chr10:132635679-132638686) and modified reference sequence containing the predicted rearrangement.

Colored arrows on the right side indicate reference and alternative structures as diploid arrangements, as well as predictions from each individual algorithm. Ploidy for individual approaches is based on reported genotypes where available.

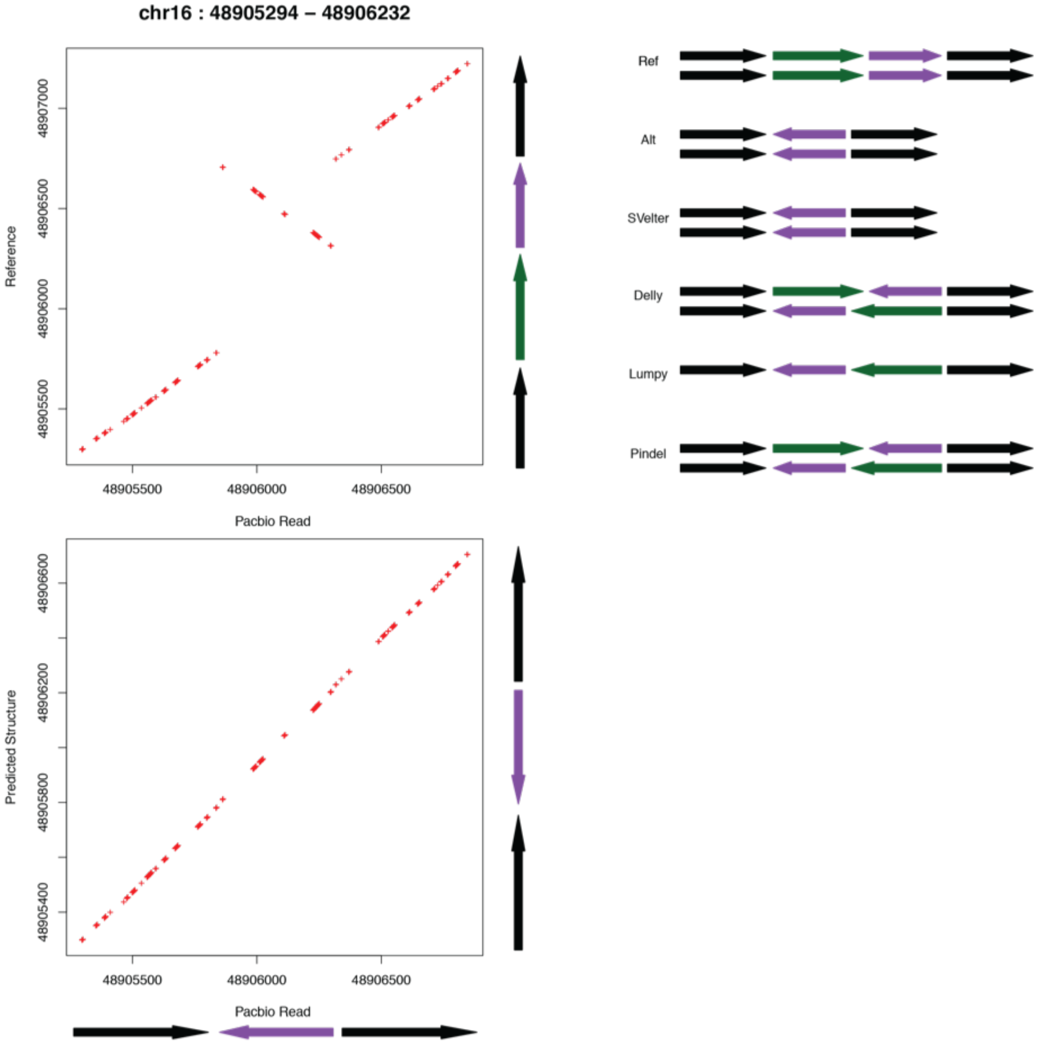


Figure 2.17 Dot plot of a PacBio read (fffb5d0d_36049_18160) from NA12878 against both unaltered reference sequence (chr16:48905294-48906232) and modified reference sequence containing the predicted rearrangement.

Colored arrows on the right side indicate reference and alternative structures as diploid arrangements, as well as predictions from each individual algorithm. Ploidy for individual approaches is based on reported genotypes where available.

CHAPTER III

Validate Structural Variants Through Long Read Sequencing Technology

(Zhao, Xuefang et al. Under review at GigaScience)

3.1 ABSTRACT

3.1.1 Background

Although numerous algorithms have been developed to identify structural variation (SVs) in genomic sequences, there is a dearth of approaches that can be used to evaluate their results. This is significant, as the accurate identification of structural variation is still an outstanding yet unsolved problem in genomics. The emergence of new sequencing technologies that generate longer sequence reads can, in theory, provide direct evidence for all types of SVs regardless of the length of region through which it spans. However, current efforts to use these data in this manner require the use of large computational resources to assemble these sequences as well as visual inspection of each region.

3.1.2 Results

Here we present VaPoR, a highly efficient algorithm that autonomously validates large SV sets using long read sequencing data. We assessed the performance of VaPoR on SVs in both

simulated and real genomes and reported a high-fidelity rate for overall accuracy across different levels of sequence depths. We show that VaPoR can interrogate a much larger range of SVs while still matching existing methods in terms of false positive validations and providing additional features considering breakpoint precision and predicted genotype. We further show that VaPoR can run quickly and efficiency without requiring a large processing or assembly pipeline.

3.1.3 Conclusions

VaPoR serves as a high efficient long read based validation approach for genomic SVs that requires relatively low read depth and computing resources and thus will provide utility with targeted or low-pass sequencing coverage for accurate SV assessment.

3.2 KEYWORDS

structural variation

copy number variation

sequence analysis

3.3 INTRODUCTION

Structural variants (SVs) are one of the major forms of genetic variation in humans and have been revealed to play important roles in numerous diseases including cancers and neurological disorders (Brand et al. 2014; Chiang et al. 2012). Various approaches have been developed and applied to paired-end sequencing to detect SVs in whole genomes (Rausch et al. 2012; Layer et al. 2014; Zhao et al. 2016; Chong et al. 2016) , however individual algorithms often exhibit complementary strengths that sometimes lead to disagreements as to the precise structure of the underlying variant. The emergence of long read sequencing technology, such as Single Molecule Real-Time (SMRT) sequencing from Pacific Biosciences (PacBio) (Rhoads and Au 2015; Travers et al. 2010), can deliver reads ranging from several to hundreds of kilobases and provide direct evidence for the presence of an SV. Current strategies make use of de novo assembly to create long contigs with minimized error rate (Chaisson et al. 2014; Pendleton et al. 2015; Shi et al. 2016), and then predict SVs, usually with single base resolution, through direct comparison of the assembly against the reference. Though such approaches are powerful, they require both a very high sequencing depth and significant computing power and are currently impracticable for many ongoing research studies.

The additional information obtained from using long reads can still be leveraged to improve variant calling, however. Indeed, such approaches have already been implemented to combine high depth Illumina sequencing with lower depth PacBio reads to improve error correction and variant calling in the context of de novo genome assembly (Koren et al. 2012). With structural variation, the current state of the art is to use long reads to manually assess potential SVs using subsequent recurrence (dot) plots (Huddleston et al. 2016), where the sequences are compared against the reference through a fixed size sliding window (k-mer) and the matches are plotted for visual inspection. The k-mer method is of higher robustness compared against the direct sequences comparison (Carvalho et al. 2016), which is why these types of dot plots have been used for decades to examine the specific features of sequence alignments (Gibbs and McIntyre 1970).

However, they require manual curation and, coupled with the computational costs of sequence assembly, are time-consuming and inefficient at scale for the high throughput validation of large sets of SVs.

Here, we present a high-speed long read based assessment tool, VaPoR, that scores each SV prediction by autonomously analyzing the recurrence of windows within a local read against the reference genome in both their original and rearranged format per the prediction. A positive score of each read on the altered reference, normalized against the score of the read on the original reference, supports the predicted structure. A baseline model is constructed as well by interrogating the reference sequence against itself at the query location. We show that our approach can quickly and accurately distinguish true from false positive predictions of both simple and complex SVs as well as their underlying genotypes and is also able to assess the breakpoint accuracy of individual algorithms.

3.4 DATA DESCRIPTION

3.4.1 Simulated Data:

Non-overlapping simple deletions, inversions, insertions and duplications as well as complex structural variants as previously categorized (Zhao et al. 2016) were independently incorporated into GRCh38 in both heterozygous and homozygous states, excluding regions of known difficulties of the genome as described from the ENCODE project (ENCODE Project Consortium 2012). Detailed descriptions of each simulated SV types simulated are summarized in Tables 3.1- 3.2. We applied PBSIM (Ono et al. 2013) to simulate the modified reference sequences to different read depth ranging from 2X to 70X with a parameters difference-ratio of 5:75:20, length-mean 12000, accuracy-mean 0.85 and model_qc model_qc_clr. Simulated data can be obtained from https://umich.box.com/v/vapor.

3.4.2 Real Data

We applied VaPoR to a set of diverse samples (HG00513 from CHS, HG00731 and HG00732 from PUR, NA19238 and NA19239 from YRI) that were initially sequenced by the 1000 Genomes Project and for which a high-quality set of SVs were reported in the final phase of the project (Sudmant et al. 2015). These samples were recently re-sequenced using PacBio and therefore provides a platform for assessing VaPoR on known data. The 1000 Genomes Project (1KGP) Phase 3 data were obtained from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/integrated_sv_map/ and lifted over to GRCh38. PacBio sequence data were obtained from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/.

We have also compared VaPoR against the long read validation approach developed by Layer et al. (Layer et al. 2014), which requires both PacBio and Moleculo long sequences for full evaluation of SVs. These comparisons made use of NA12878, one of few samples that have been sequenced with various technologies including Illumina NGS, PacBio and Moleculo with a truth SV set included in the 1KGP Phase 3 report. The PacBio and the Moleculo sequences of this individual were obtained from : http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20131209_na12878_pacbio/si/ and http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/supporting/NA12878/moleculo/alignment/ respectively.

Table 3.1 Number of homozygous SVs simulated for each type on different chromosomes

sv_type	del	ins	inv	tan_dup	dis_dup	del_dup	del_inv	dup_inv_ins	del_dup_inv
chr1	352	29	83	132	17	156	25	152	169
chr2	345	44	72	143	13	144	24	197	181
chr3	253	34	61	110	17	152	19	147	129
chr4	267	16	54	95	9	132	13	150	153
chr5	273	21	74	84	13	154	19	142	119
chr6	252	28	57	87	20	123	23	128	128
chr7	211	31	60	76	8	101	20	119	114
chr8	165	23	50	79	12	97	14	91	101
chr9	175	24	47	75	7	76	17	95	101
chr10	196	15	46	57	12	94	20	79	87
chr11	192	26	54	64	12	105	10	98	83
chr12	159	30	49	48	7	98	6	95	87
chr13	145	18	34	43	9	87	11	77	78
chr14	167	13	34	52	7	76	10	83	67
chr15	121	8	37	44	6	69	15	72	71
chr16	133	13	31	46	7	55	8	69	58
chr17	108	11	20	28	6	60	9	66	56
chr18	102	11	20	42	7	73	11	60	58
chr19	82	13	22	21	2	46	5	46	32
chr20	85	10	22	28	2	44	9	42	36
chr21	89	3	16	15	0	32	6	31	29
chr22	63	4	17	26	3	37	7	30	36
chrX	225	21	55	80	9	109	16	109	96
chrY	88	5	22	37	5	42	5	34	44

Table 3.2 Number of heterozygous SVs simulated for each type on different chromosomes

sv_type	del	ins	inv	tan_dup	dis_dup	del_dup	del_inv	dup_inv _ins	del_dup _inv
chr1	348	45	91	103	20	166	27	184	149
chr2	306	27	77	105	22	184	25	158	185
chr3	274	27	59	90	12	133	17	127	124
chr5	265	24	69	81	14	148	18	132	116
chr6	231	21	59	76	12	126	24	135	136
chr7	230	19	54	76	7	126	29	110	100
chr8	171	17	44	77	12	104	14	99	90
chr9	174	26	54	62	4	103	13	100	94
chr10	170	17	47	57	13	75	13	105	78
chr11	180	23	44	63	13	94	21	100	100
chr12	201	23	39	60	12	83	12	101	98
chr13	147	24	38	63	10	80	14	77	68
chr14	127	20	31	52	11	78	9	73	77
chr15	156	17	30	41	7	80	12	65	83
chr16	123	8	30	43	7	61	7	52	63
chr17	108	9	26	39	2	61	7	54	48
chr18	109	10	24	44	4	54	11	58	54
chr19	81	10	22	21	5	43	6	49	38
chr20	96	11	25	31	4	52	6	44	50
chr21	43	5	15	26	6	29	3	21	31
chr22	77	7	17	22	8	42	2	29	28
chrX	227	18	49	60	16	116	19	123	112

3.5 RESULTS

We assessed the performance of VaPoR on both simulated sequences and real genomes from the 1000 Genomes Project to assess the following characteristics: sensitivity and false discovery rate on validating structural variants in simple and complex structures; sensitivity of VaPoR on validating different levels of predicted breakpoint efficacy; stratification of VaPoR scores by genotype; and time and computational cost of VaPoR.

3.5.1 VaPoR on Simulated Data

We applied VaPoR to simulated simple deletions, inversions, insertions and duplications as well as complex structural variants and first assessed the proportion of SVs that VaPoR is capable of interrogating (i.e. passed VaPoR QC). We found that VaPoR can successfully evaluate >80% of insertions, >85% deletion-duplications and >90% SVs in all other categories when the read depth is 10X or higher. We then assessed the sensitivity and false discovery rate (FDR) at different VaPoR score cutoffs and found that when considering different types of SVs across various read depths from 2X to 70X, most of the SV types can achieve a sensitivity >90% with false discovery rate <10% at a VaPoR score cutoff of 0.15 (Figures 3.3-3.4). We further observed that there were no significant changes of sensitivity or false discovery rate once the read depth was at or above 20X and is consistent across different SV types (Figure 3.2, Figures 3.5-3.6, Table 3.3).

Table 3.3 Sensitivity and false discovery rate of VaPoR on simulated SVs.

Sensitivity of heterozygous simulations

	DEL	INS	INV	TANDUP	DISDUP	DEL_DUP	DEL_INV	DUP_INV	DEL_DUP_INV
RD_2	0.45(0.70)	0.55(0.59)	0.56(0.74)	0.36(0.78)	0.47(0.75)	0.11(0.41)	0.71(0.70)	0.62(0.62)	0.32(0.69)
RD_5	0.62(0.90)	0.79(0.77)	0.74(0.92)	0.59(0.93)	0.62(0.91)	0.26(0.68)	0.88(0.88)	0.74(0.84)	0.52(0.89)
RD_10	0.78(0.94)	0.91(0.79)	0.91(0.94)	0.76(0.94)	0.90(0.94)	0.57(0.86)	0.97(0.90)	0.87(0.91)	0.75(0.93)
RD_20	0.86(0.94)	0.93(0.82)	0.97(0.94)	0.90(0.93)	0.96(0.95)	0.80(0.94)	1.00(0.90)	0.93(0.93)	0.92(0.93)
RD_30	0.89(0.94)	0.89(0.89)	0.98(0.94)	0.96(0.93)	0.97(0.96)	0.82(0.94)	0.98(0.90)	0.95(0.94)	0.92(0.93)
RD_50	0.91(0.95)	0.83(0.95)	0.97(0.94)	0.98(0.93)	0.98(0.96)	0.83(0.94)	1.00(0.90)	0.96(0.94)	0.93(0.94)
RD_70	0.90(0.95)	0.83(0.96)	0.99(0.94)	0.99(0.93)	0.98(0.96)	0.82(0.94)	1.00(0.90)	0.96(0.94)	0.94(0.94)

False discovery rate(FDR) of heterozygous simulations

	DEL	INS	INV	TANDUP	DISDUP	DEL_DUP	DEL_INV	DUP_INV	DEL_DUP_INV
RD_2	0(0.72)	0.01(0.57)	0.01(0.73)	0.01(0.75)	0.05(0.76)	0.03(0.44)	0.34(0.69)	0.07(0.57)	0.24(0.71)
RD_5	0(0.93)	0.02(0.78)	0.02(0.93)	0.02(0.93)	0.06(0.94)	0.09(0.73)	0.36(0.88)	0.1(0.85)	0.41(0.92)
RD_10	0(0.94)	0.02(0.81)	0.02(0.94)	0.02(0.94)	0.07(0.95)	0.2(0.89)	0.37(0.91)	0.11(0.91)	0.41(0.94)
RD_20	0(0.94)	0.02(0.85)	0.01(0.94)	0.03(0.93)	0.08(0.95)	0.23(0.94)	0.38(0.91)	0.11(0.93)	0.29(0.94)
RD_30	0(0.94)	0.03(0.92)	0.01(0.94)	0.04(0.93)	0.06(0.95)	0.24(0.94)	0.37(0.91)	0.1(0.94)	0.28(0.94)
RD_50	0(0.94)	0.02(0.96)	0.01(0.94)	0.06(0.93)	0.07(0.95)	0.24(0.94)	0.38(0.91)	0.09(0.94)	0.27(0.94)
RD_70	0(0.94)	0.03(0.96)	0.01(0.94)	0.07(0.92)	0.09(0.95)	0.23(0.95)	0.39(0.91)	0.09(0.94)	0.29(0.94)

Sensitivity of homozygous simulations

	DEL	INS	INV	TANDUP	DISDUP	DEL_DUP	DEL_INV	DUP_INV	DEL_DUP_INV
RD_2	0.76(0.62)	0.92(0.55)	0.85(0.72)	0.62(0.77)	0.71(0.75)	0.11(0.41)	0.96(0.69)	0.91(0.60)	0.36(0.60)
RD_5	0.86(0.86)	0.96(0.74)	0.94(0.92)	0.83(0.92)	0.91(0.92)	0.25(0.68)	0.99(0.85)	0.95(0.82)	0.54(0.81)
RD_10	0.92(0.93)	0.98(0.76)	0.97(0.94)	0.94(0.94)	0.95(0.93)	0.61(0.87)	0.99(0.89)	0.97(0.89)	0.79(0.91)
RD_20	0.95(0.94)	0.98(0.79)	0.99(0.94)	0.99(0.93)	0.97(0.93)	0.84(0.92)	1.00(0.90)	0.97(0.92)	0.94(0.93)
RD_30	0.95(0.94)	0.97(0.80)	0.99(0.94)	0.99(0.93)	0.96(0.95)	0.86(0.93)	1.00(0.90)	0.98(0.92)	0.95(0.93)
RD_50	0.96(0.94)	0.96(0.81)	0.99(0.94)	1.00(0.93)	0.96(0.95)	0.86(0.93)	1.00(0.90)	0.98(0.93)	0.96(0.93)
RD_70	0.95(0.94)	0.95(0.83)	0.99(0.94)	0.99(0.93)	0.97(0.95)	0.86(0.93)	1.00(0.90)	0.97(0.93)	0.96(0.93)

False discovery rate(FDR) of homozygous simulations

	DEL	INS	INV	TANDUP	DISDUP	DEL_DUP	DEL_INV	DUP_INV	DEL_DUP_INV
RD_2	0(0.64)	0(0.55)	0.01(0.67)	0.01(0.66)	0.04(0.75)	0.02(0.42)	0.32(0.64)	0.08(0.54)	0.27(0.66)
RD_5	0(0.91)	0.02(0.75)	0.01(0.92)	0.01(0.92)	0.08(0.94)	0.08(0.7)	0.35(0.89)	0.09(0.83)	0.42(0.92)
RD_10	0(0.93)	0.01(0.79)	0.01(0.94)	0.02(0.93)	0.12(0.92)	0.19(0.88)	0.38(0.9)	0.12(0.9)	0.37(0.93)
RD_20	0(0.93)	0.03(0.85)	0.01(0.94)	0.04(0.93)	0.08(0.94)	0.25(0.93)	0.38(0.9)	0.1(0.92)	0.29(0.93)
RD_30	0(0.93)	0.02(0.9)	0.01(0.94)	0.06(0.92)	0.09(0.95)	0.24(0.93)	0.37(0.9)	0.09(0.92)	0.26(0.93)
RD_50	0(0.93)	0.03(0.95)	0.01(0.94)	0.07(0.93)	0.09(0.95)	0.23(0.93)	0.38(0.9)	0.09(0.93)	0.25(0.93)
RD_70	0(0.93)	0.02(0.95)	0.01(0.94)	0.07(0.92)	0.09(0.96)	0.24(0.93)	0.38(0.9)	0.09(0.93)	0.26(0.93)

3.5.2 VaPoR on 1000 Genomes Project Samples

We next examined SVs reported on chr1 of 5 individuals from the 1000 Genomes Project (1000 Genomes Project Consortium et al...) to assess the sensitivity of VaPoR on real genomes (Table 3.4). We first observed that >95% of deletions and insertions could be successfully evaluated by VaPoR. For inversions, there were a limited number of events reported but at maximum only 1 event failed the VaPoR quality control per individual. A sensitivity of >90% was achieved for deletions (Figure 3.7a) and >80% for insertions (Figure 3.7b) at the VaPoR score cutoff of 0.15. To examine the false validation rate of VaPoR, we modified reported events on chr2 to appear at the same coordinates on chr1 and assessed them as though they were real events using the same sequence data set. VaPoR validated very few deletions or inversion and <10% of insertions. We further assessed the performance of VaPoR to validate SVs with varying degrees of breakpoint accuracy. Real coordinates were artificially shifted each direction by -1000 to 1000 base pairs and re-assessed with VaPoR for both simulated and real samples. In both cases, VaPoR exhibited a robust validation score up to approximately 200bp overall, with some slight differences observed between different SV types (Figure 3.7c,d, Figures 3.8-3.9).

We also compared VaPoR against a long-read validation approach developed in conjunction with Lumpy (Layer et al. 2014) using SVs on chr1 of NA12878 reported by the 1000 Genomes Project Phase 3. VaPoR achieved a sensitivity of 72% for deletions and 86% for insertions, while the Lumpy-associated approach was only able to assess 11% and 0% respectively. Both approaches exhibited a very low false validation rate when synthetically assigning the variants to chr2, with 0 for all SV types by the Layer et al approach and varying between 0 and 2.5% for VaPoR (Table 3.5).

Table 3.4 Sensitivity and false discovery rate of different SV types

	deletion	insertion	inversion
Sample	Sens/FDR	Sens/FDR	Sens/FDR
HG00513	0.96/0.00 (0.94[1])	0.80/0.05 (0.93)	0.50/0.00 (0.71)
HG00731	0.94/0.00 (0.96)	0.85/0.07 (0.97)	0.60/0.00 (1.00)
HG00732	0.92/0.00 (0.98)	0.92/0.08 (0.96)	0.33/0.00 (0.86)
NA19238	0.90/0.00 (0.93)	0.88/0.10 (0.96)	1.00/0.00 (1.00)
NA19239	0.87/0.02 (0.95)	0.73/0.09 (0.96)	0.33/0.00 (1.00)

[1] Proportion of SVs passed VaPoR QC, as listed in brackets, are counted for events on chr1 and chr2 together.

Table 3.5 Sensitivity and false discovery rate of SVs on chr1 in NA12878, compared against two validation approaches: VaPoR and the long read validation approach by Layer et al.

	Sens		FDR	
NA12878	Layer et al.	VaPoR	Layer et al.	VaPoR
deletion	10.66%	71.90%	0.00%	1.46%
inversion	66.67%	50.00%	0.00%	0.00%
insertion	0.00%	86.25%	0.00%	2.50%

3.5.3 Discrimination of SV types and genotypes

We identified a small number of SVs in the high quality 1000 Genomes set that did not validate with VaPoR. Previous studies have shown that complex rearrangements are often misclassified as simple structural changes (Rausch et al. 2012, Huddleston et al. 2016), and indeed upon manual inspection these appeared to consist of multiple connected rearrangements. For example, we observed a reported inversion in HG00513 and NA19239 on chromosome 1 (chr1:239952707-239953529) that was invalidated by VaPoR; an investigation into the long-reads aligned in the region showed the signature of an inverted duplication (Figure 3.10a) which, when incorporated into a modified reference that location, matched almost exactly with the read sequence (Figure 3.10b).

We further explored the distribution of VaPoR scores for this region and others across the sample set and observed clear delineations between allelic copy number when fit with a Gaussian mixture model allowing for the generation of genotype likelihoods for each site (Figure 3.10c). These tracked with our expected genotypes for the inverted duplication on chr1 across the 5 individuals queried while showing no support for the originally predicted inversion (Figure 3.10d). This shows that VaPoR is not only able to accurately genotype variants but can also distinguish between similar but distinct SV predictions in the same region.

3.5.4 Runtime and efficiency

The computation runtime of VaPoR was assessed using 2 Intel Xeon Intel Xeon E7-4860 processors with 4GB RAM each on both simulated and real genomes. The runtime of simulated event was observed to increase linearly with read depth (Table 3.6). For events sequenced up to 20X, VaPoR takes ~3 seconds to assess a simple SV and ~5s for a complex event. The assessment of real samples sequenced at 20X required ~1.4 seconds to assess a simple deletion or insertion and ~6 seconds for an inversion (Figure 3.11).

Table 3.6 Averaged computing time required for VaPoR to validate an SV estimated in seconds.

Read Depth	Average (s)
RD_2	1.73
RD_5	1.93
RD_10	2.55
RD_20	3.67
RD_30	4.73
RD_50	5.48
RD_70	6.16

3.6 DISCUSSION

Here we present an automated assessment approach, named VaPoR, for exploring various features of predicted genomic structural variants using long read sequencing data. VaPoR directly compares the input reads with the reference sequences with relatively straightforward computational metrics, thus achieving high efficiency in both run time and computing cost. VaPoR exhibits high sensitivity and specificity in both simulated and real genomes, with the capability of discriminating partially resolved SVs either consisting of similar but incorrect SV types at the same location or correct SVs with offset breakpoints. Furthermore, we show that VaPoR performs well at low read depths (5-10X), thus providing the option of systematically assessing large-scale SVs with a lower sequencing cost.

3.7 METHODS

3.7.1 VaPoR Workflow

VaPoR takes in aligned sequence reads in BAM format and predicted SVs (>50bp) in various formats including VCF and BED. Evaluation of an SV is performed by comparing long reads that go through the event against reference sequences in two formats: (a) the original human reference to which the sample is aligned and (b) a modified reference sequence altered to match the predicted structural rearrangement. A recurrence matrix is then derived by sliding a fixed-size window with 1bp step through each read to mark positions where the read sequence and reference are identical. The matching patterns are then assessed as to the validity of the SV as described below and a validation score is reported. Given the large variance of SVs lengths, each SV is stratified into one of two groups: smaller SVs that can be completely encompassed within multiple (>10 by default) long sequences and larger events that are rarely covered by individual long reads, with different statistical model applied. The VaPoR workflow is briefly summarized in Figure 3.1.

Small Variants Assessment:

For an SV *k* in sample *s* that is covered by *n* reads, the recurrence matrix between each read and the reference sequences in original (R_o) and altered (R_a) format is calculated. The vertical distance between each record ($x_{i,k,s,Rx}$, $y_{i,k,s,Rx}$) in matrix *x* and the diagonal ($x_{i,k,s,Rx}$, $x_{i,k,s,Rx}$) line is calculated as $d_{i,k,s,Rx}$ = abs($x_{i,k,s,Rx}$ - $y_{i,k,s,Rx}$), and the average distance of all records would be exported as the score of each matrix:

$$Score_{k,s,Rx} = \sum_{i=1}^{m} d_{i,k,s,Rx} \;/\; m,$$

where m is the total number of records in the matrix. Sequences that share higher identity with the read shall have a lower $\mathrm{Score}_{k,s,Rx}$, such that the score of each read is normalized as:

$$Score_{k,s,R} = Score_{k,s,R_o} \;/\; Score_{k,s,R_a} - 1,$$

where a positive $\mathrm{Score}_{k,s,R}$ represents the superiority of the predicted structure versus the original and vise versa for negative $\mathrm{Score}_{k,s,R}$, with one exceptional case where there exists a duplicated structure in the predicted SV such that the predicted structure would show higher $\mathrm{Score}_{k,s,R}$ due to the multi-alignment of duplicated segments. To correct for duplications, VaPoR adopts the directed distance $d_{i,k,s,Rx} = x_{i,k,s,Rx} - y_{i,k,s,Rx}$ instead such that the distance contributed by centrosymmetric duplicated segments would offset each other.

Large Variants Assessment:

For larger SVs where there are few, if any, long reads that can transverse the predicted SV, VaPoR assesses the quality of each predicted junction instead using:

$$Score_{k,s,Rx} = \frac{\sum_{i=1}^{m} I = \begin{cases} 1, if\ abs(x_{i,k,s,Rx} - y_{i,k,s,Rx}) < 0.15 * x_{i,k,s,Rx} \\ 0, otherwise \end{cases}}{m},$$

where a larger $\mathrm{Score}_{\mathrm{k,s,Rx}}$ represents higher similarity between the read and the reference sequence. The normalized scores of each read is then defined as:

$$Score_{k,s,R} = Score_{k,s,R_a} \,/\, Score_{k,s,R_o} - 1,$$

VaPoR Score Calculation:

With a score assigned to each read spanning through the predicted structural variants, the VaPoR score is summarized as:

$$\mathrm{Score}_{\mathrm{k,s}} = \frac{\sum_{\mathrm{R}=1}^{\mathrm{n}} \mathrm{I} = \begin{cases} 1, \mathrm{if}\ \ \mathrm{Score}_{\mathrm{k,s,R}} > 0 \\ 0, \mathrm{otherwise} \end{cases}}{\mathrm{n}}$$

to represent the proportion of long reads supporting predicted structure.

The highest supportive score ($\max(\mathrm{Score}_{\mathrm{k,s,R}})$) is also reported as a reference for users to meet the specific requirement of their study design, for which we recommend 0.1 as the cutoff.

Flexible window size:

By default, VaPoR uses a window size of 10bp and requires an exact match between sequences, though these can be changed to user-defined parameters. However, many regions of the genome contain repetitive sequences resulting in an abundance of spurious matches in the recurrence matrix, thus introducing bias to the assessment. To address this, VaPoR adopts a quality control step by iteratively assessing the reference sequence against itself and tabulating the proportion of matches along the diagonal. The window size initially starts at 10bp and iteratively increases by 10bp until either (a) the proportion of matches on the diagonal exceeds 40% and the current window size is kept or (b) the window size exceeds 40bp whereby the event will be labeled as 'non-assessable and excluded from the evaluation.

3.8 AVAILABILITY AND REQUIREMENTS

Project name: VaPoR

Project home page: https://github.com/millslab/vapor

Operating systems: Linux, OS X

Programming languages: Python, R

Other requirements: Python v2.7.8+, rpy2, HTSeq, samtools v0.19+, pyfasta v0.5.2+, and pysam 0.9.1.4+.

3.9 ACKNOWLEDGEMENTS

We thank the Human Genome Structural Variation Consortium (HGSVC) for generating and providing the deep PacBio sequencing. We also thank Yuanfang Guan and Kerby Shedden for discussions over specific statistical considerations.

3.10 FIGURES

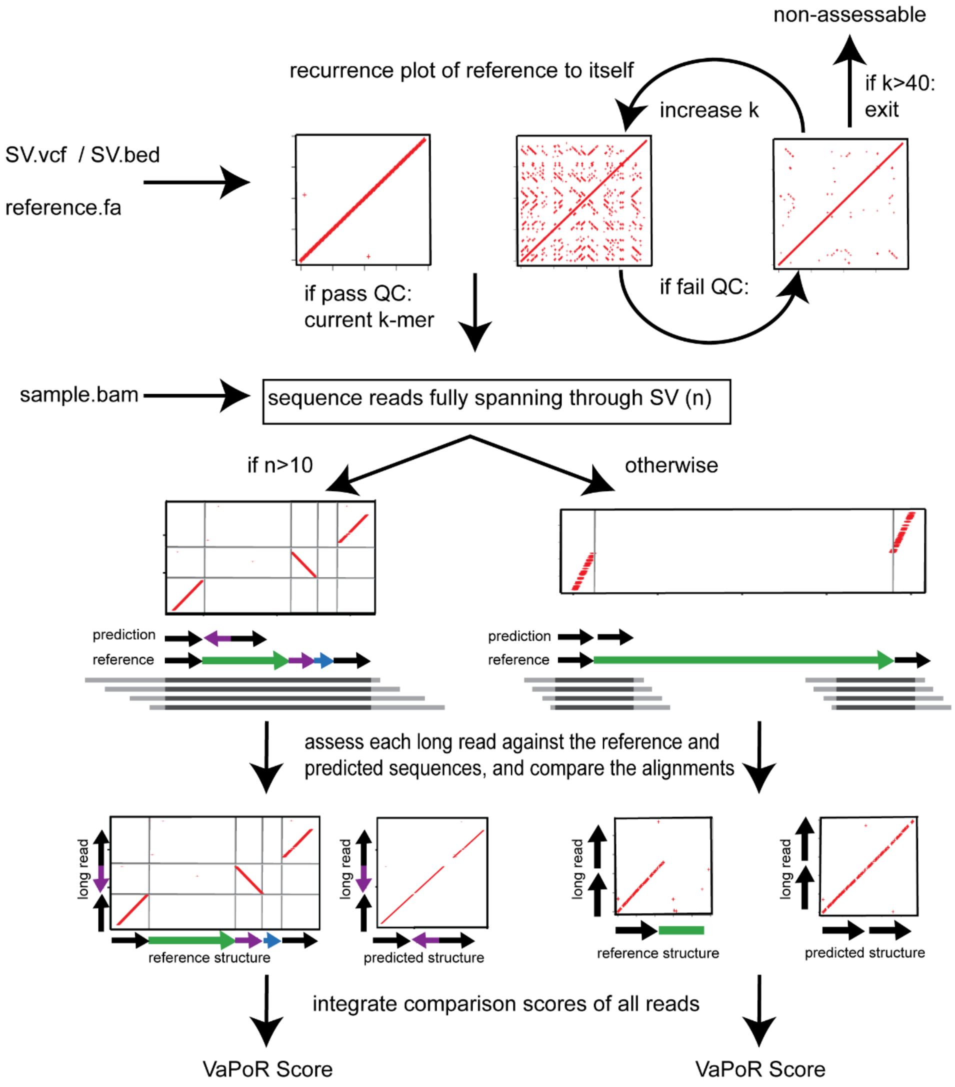


Figure 3.1 Flowchart describing the VaPoR algorithm.

As input, the algorithm requires a set of structural variants in either VCF or BED format, a series of long reads and/or sequence contigs in BAM format, and the corresponding reference sequence. VaPoR then interrogates each variant individually at its corresponding reference location, assesses the quality of the region and assigns a score.

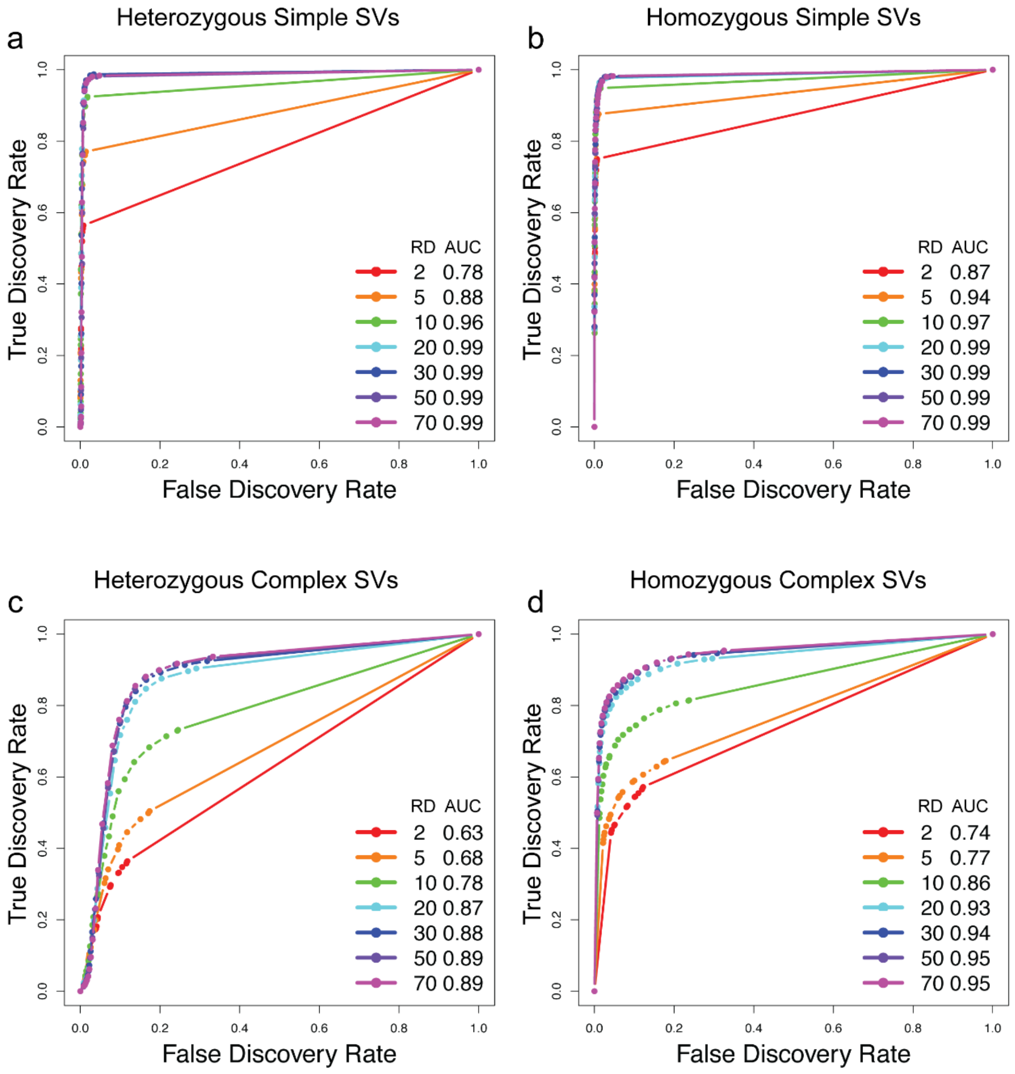


Figure 3.2 Accuracy of VaPoR on simulated heterozygous and homozygous SVs at varying degrees of sequence coverage and VaPoR score cut-offs.

Receiver operator curves (ROC) are shown for simple deletions, duplications and inversions (a,b) as well as complex rearrangements including inverted duplications and deletion-inversion rearrangements (c,d).

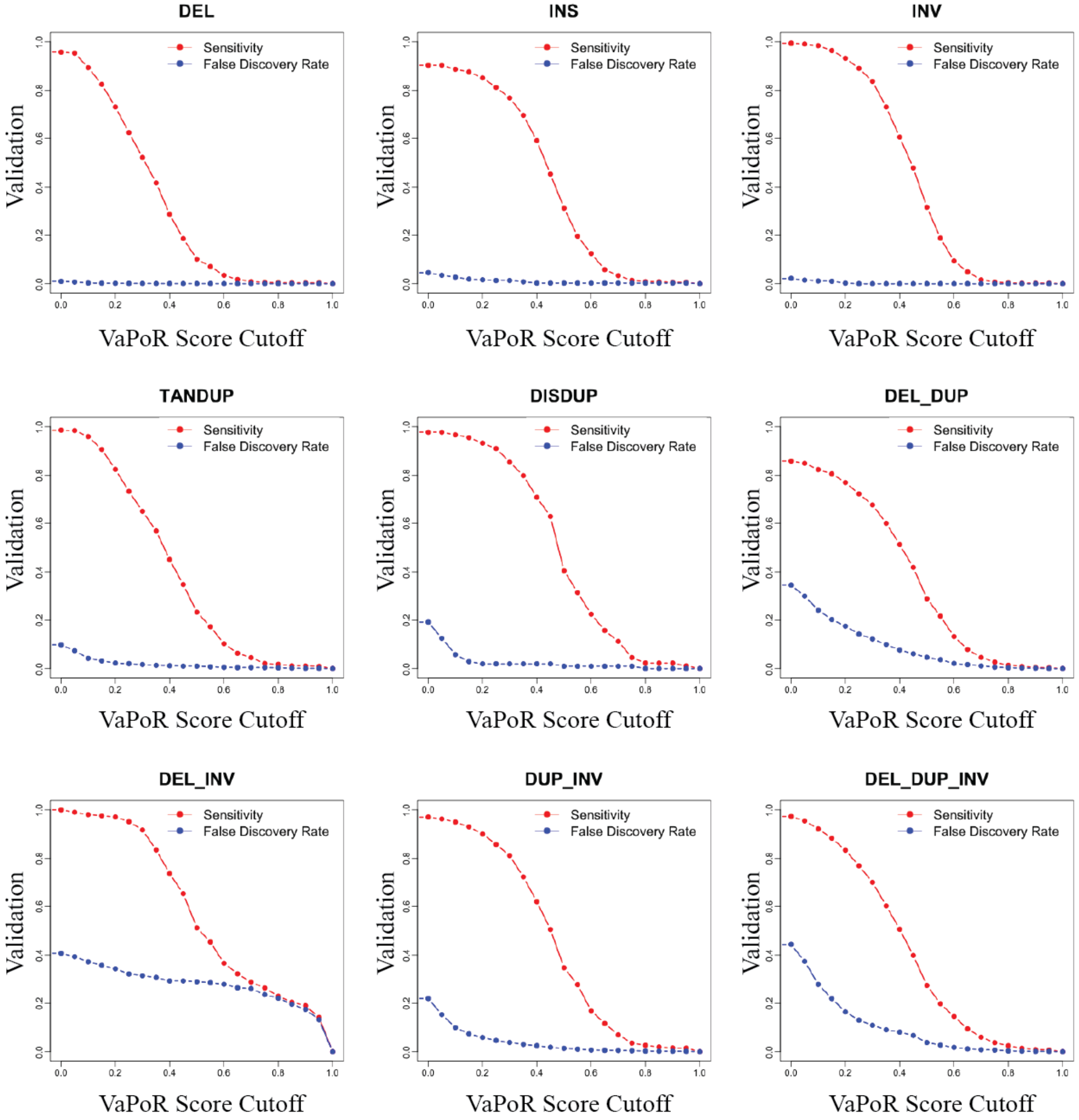


Figure 3.3 Sensitivity and false discovery rate (FDR) of validating heterozygously simulated structural variants calculated at different VaPoR score cutoff.

Sensitivity and FDR both decreases with the cutoff increasing, with >90% sensitivity and <10% FDR achieved at cutoff=0.1

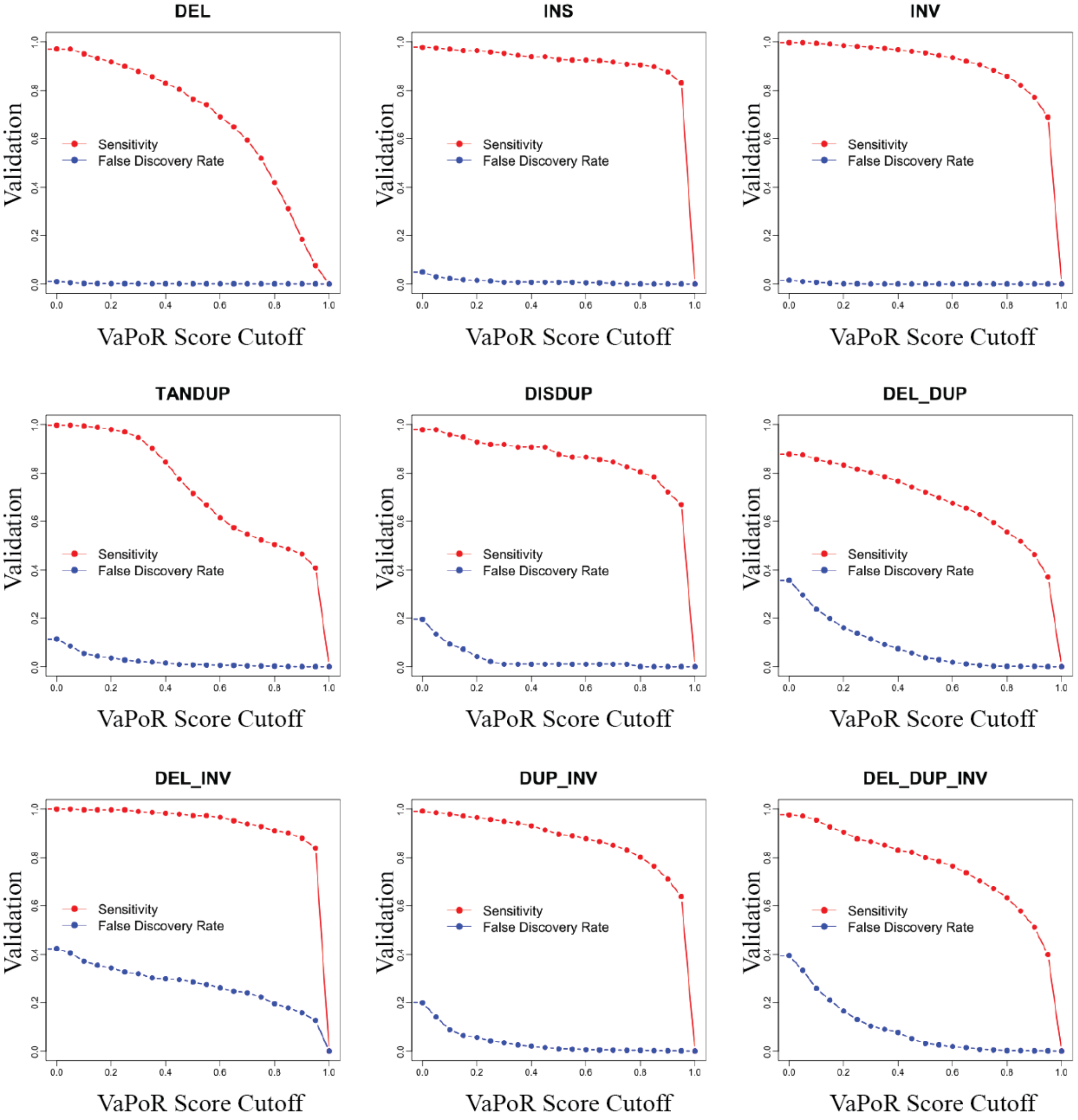


Figure 3.4 Sensitivity and false discovery rate (FDR) of validating homozygously simulated structural variants calculated at different VaPoR score cutoff.

Sensitivity and FDR both decreases with the cutoff increasing, with >90% sensitivity and <10% FDR achieved at cutoff=0.1 – 0.25

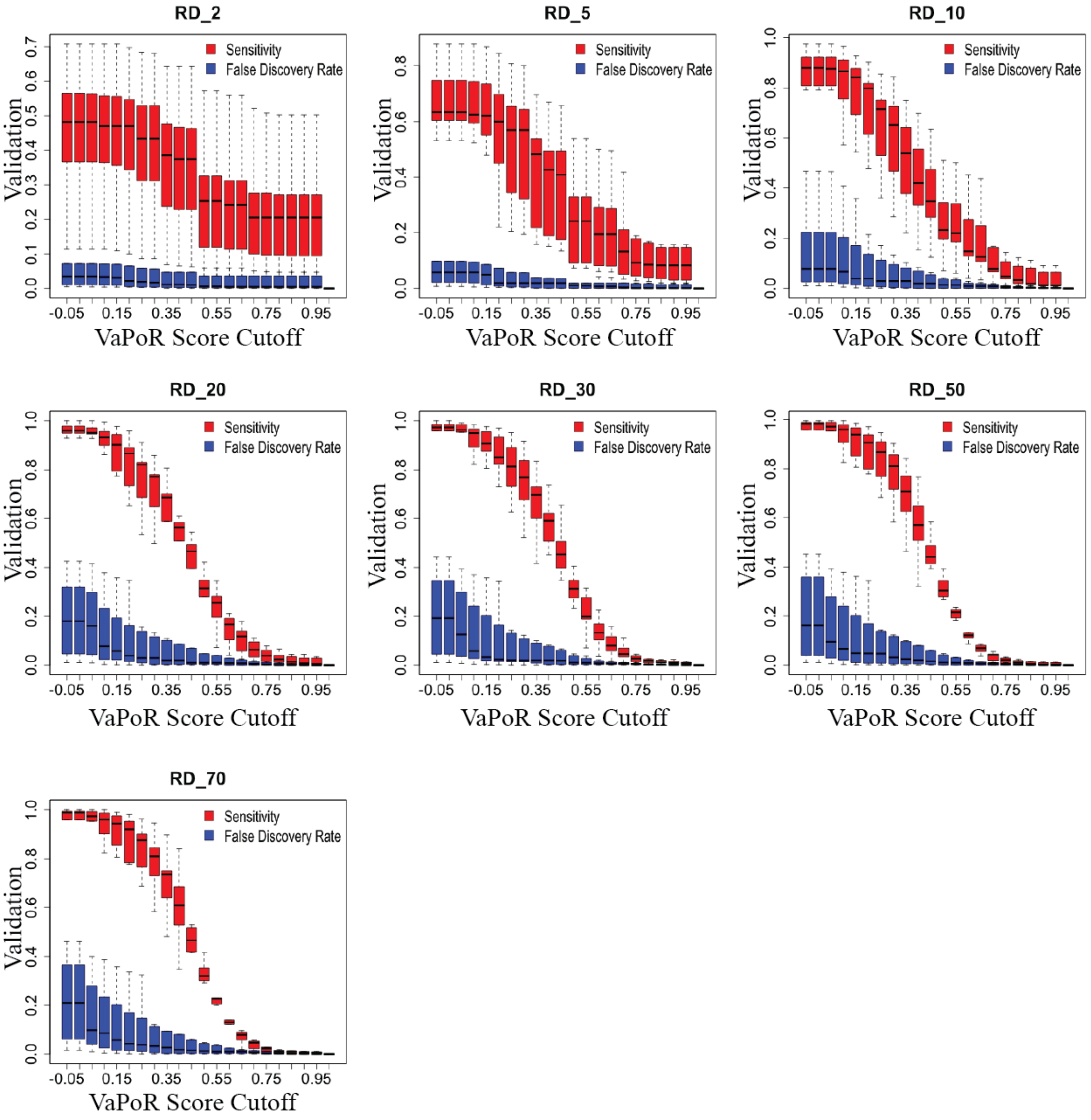


Figure 3.5 Sensitivity and false discovery rate (FDR) of VaPoR on validating heterozygous structural variants plotted across different read depth (RD).

Similar pattern was observed from RD = 20 to 70.

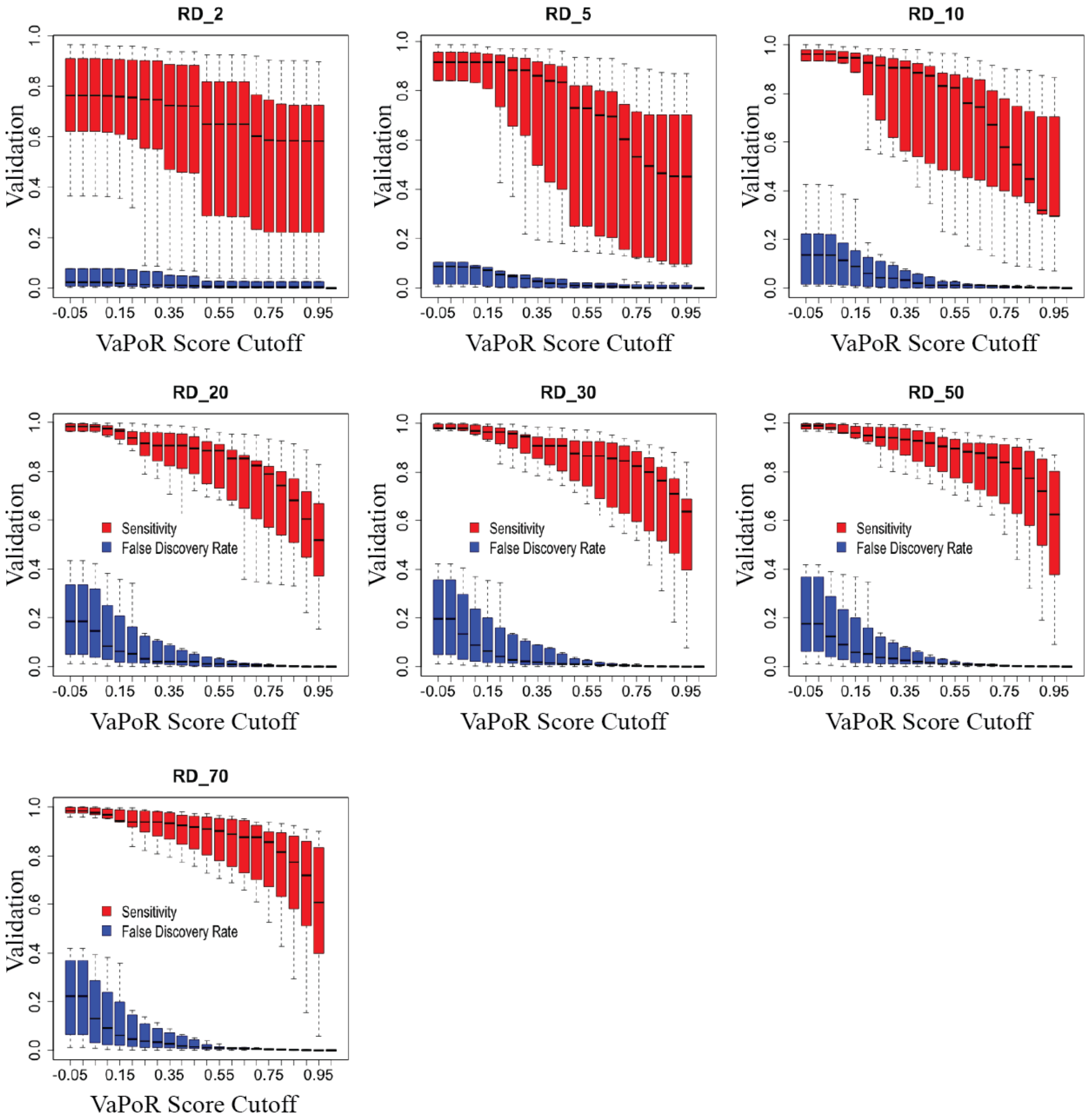


Figure 3.6 Sensitivity and false discovery rate (FDR) of VaPoR on validating homozygous structural variants plotted across different read depth (RD).Similar pattern were observed from RD = 20 to 70.

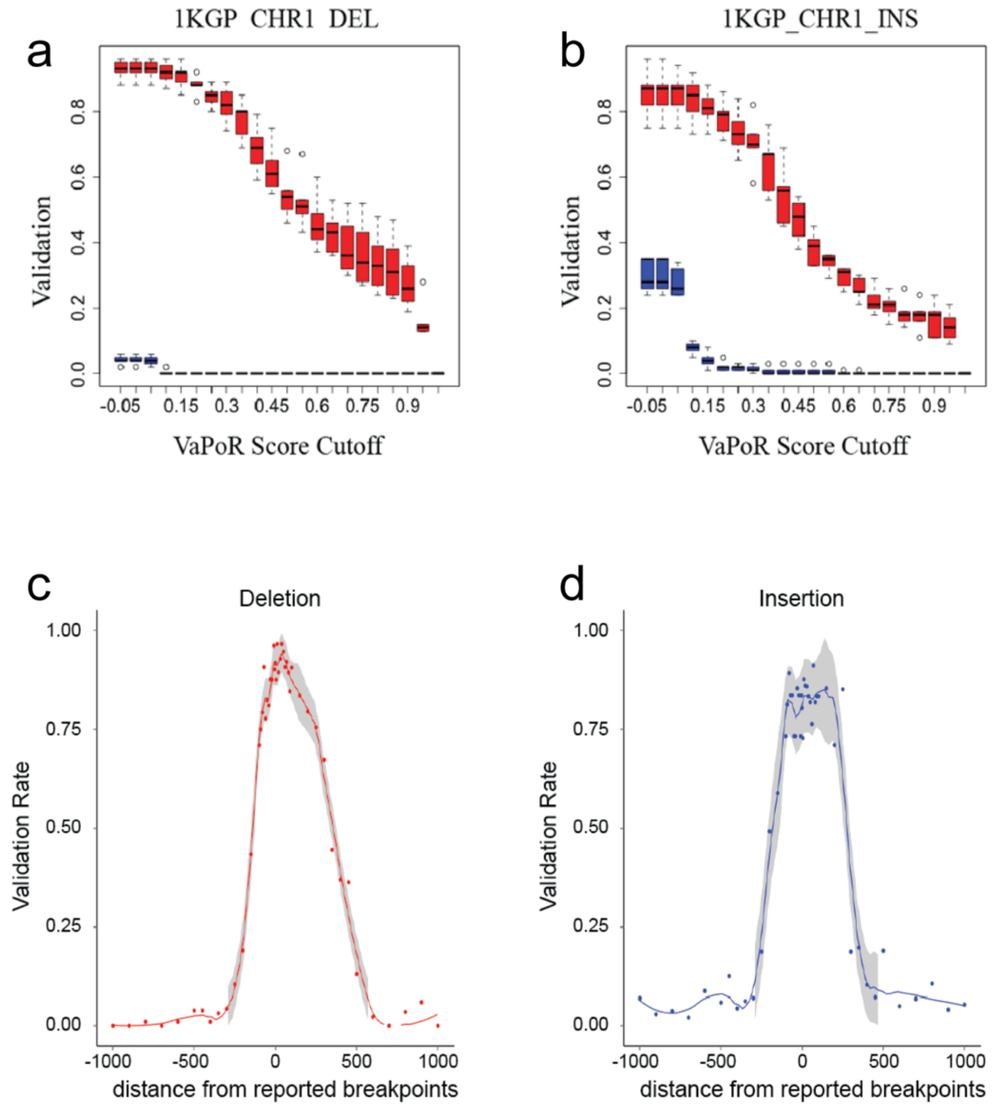


Figure 3.7 Validation rate and breakpoint accuracy of VaPoR on the 1000 Genomes Projects phase 3 calls.

VaPoR was applied on 5 individuals with reported SVs as a truth set: HG00513, HG00731, HG00732, NA19238, NA19239. The validation rate of deletions (a) and insertions (b) are shown here across different cutoff scores for VaPoR. Robustness to breakpoint accuracy was assessed using fake breakpoints deviated from the real ones by different base pair distances for deletions (c) and insertions (d).

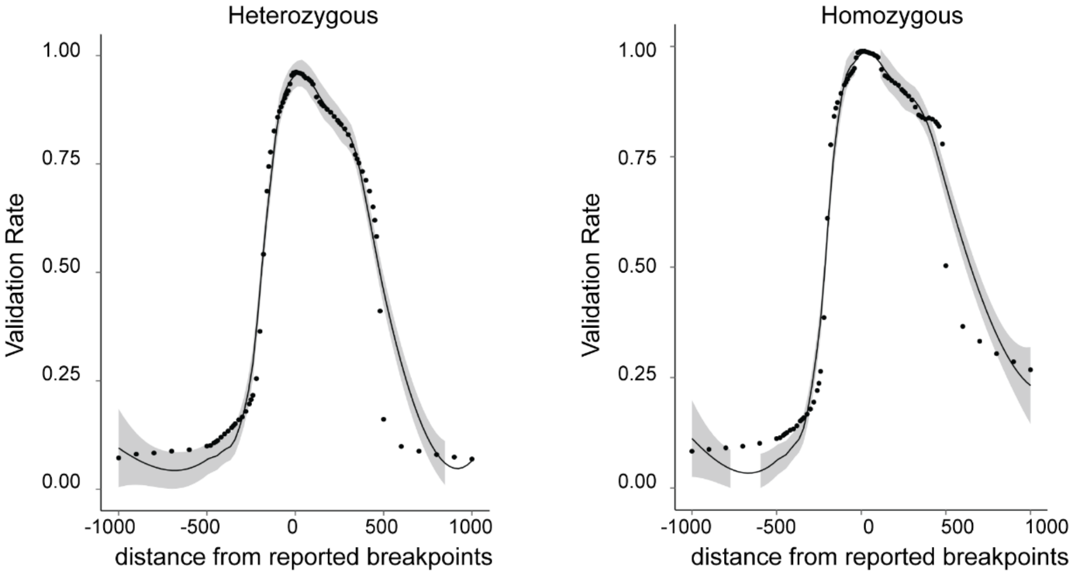
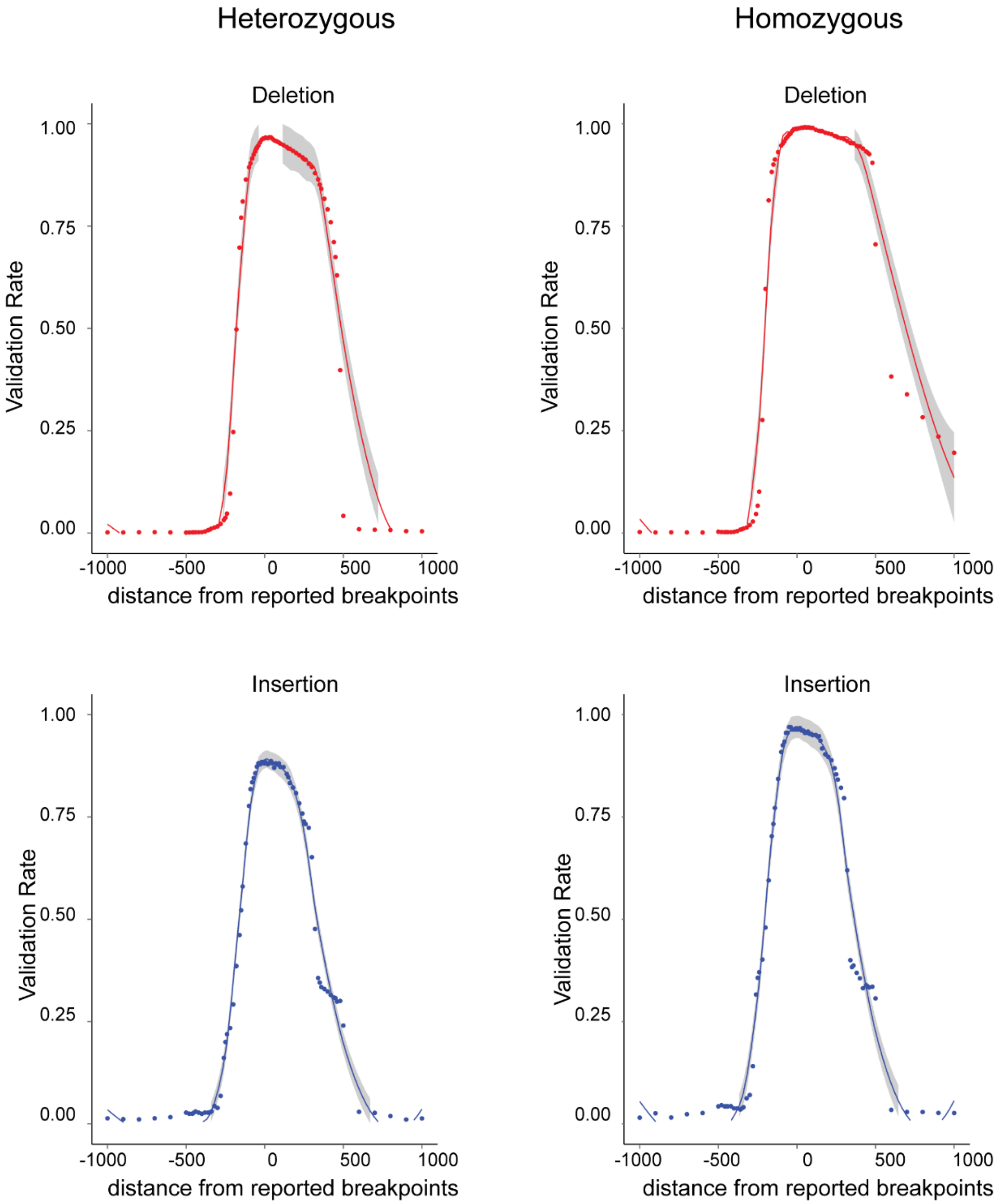


Figure 3.8. Figure 0.8 Plot of validation rate when validating the simulated SVs with fake breakpoints deviated from the real ones by different bases.

Validation rates are averaged from simulated deletion, insertion, inversion and tandem duplication at 30X coverage.


Heterozygous
Homozygous
Deletion
Deletion
Insertion
Insertion
Validation Rate
1.00
0.75
0.50
0.25
0.00
-1000
-500
0
500
1000
distance from reported breakpoints

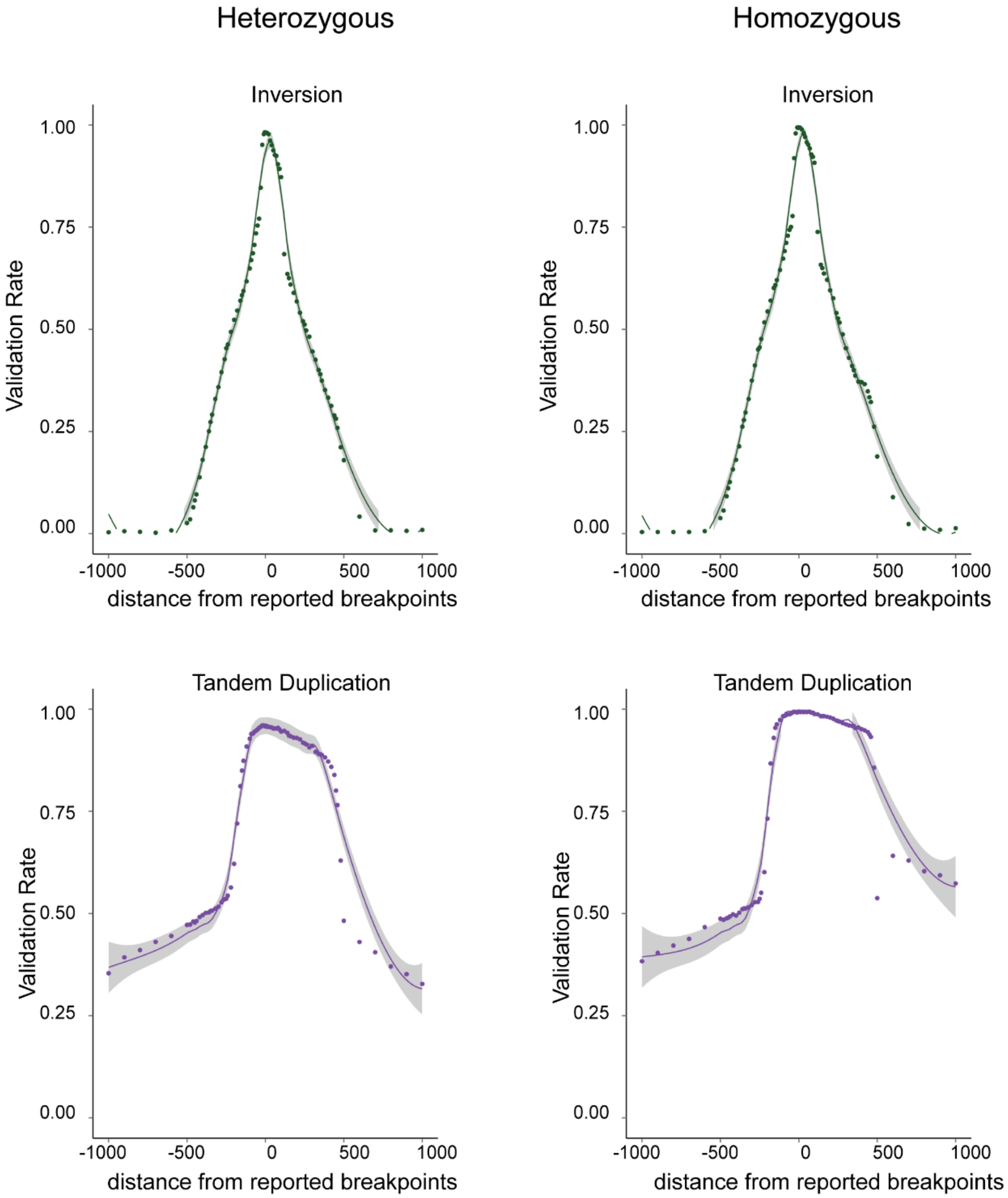


Figure 3.9 Plot of validation rate when validating the simulated SVs with fake breakpoints deviated from the real ones by different bases.

Validation rates are shown for simulated deletion, insertion, inversion and tandem duplication at 30X coverage.

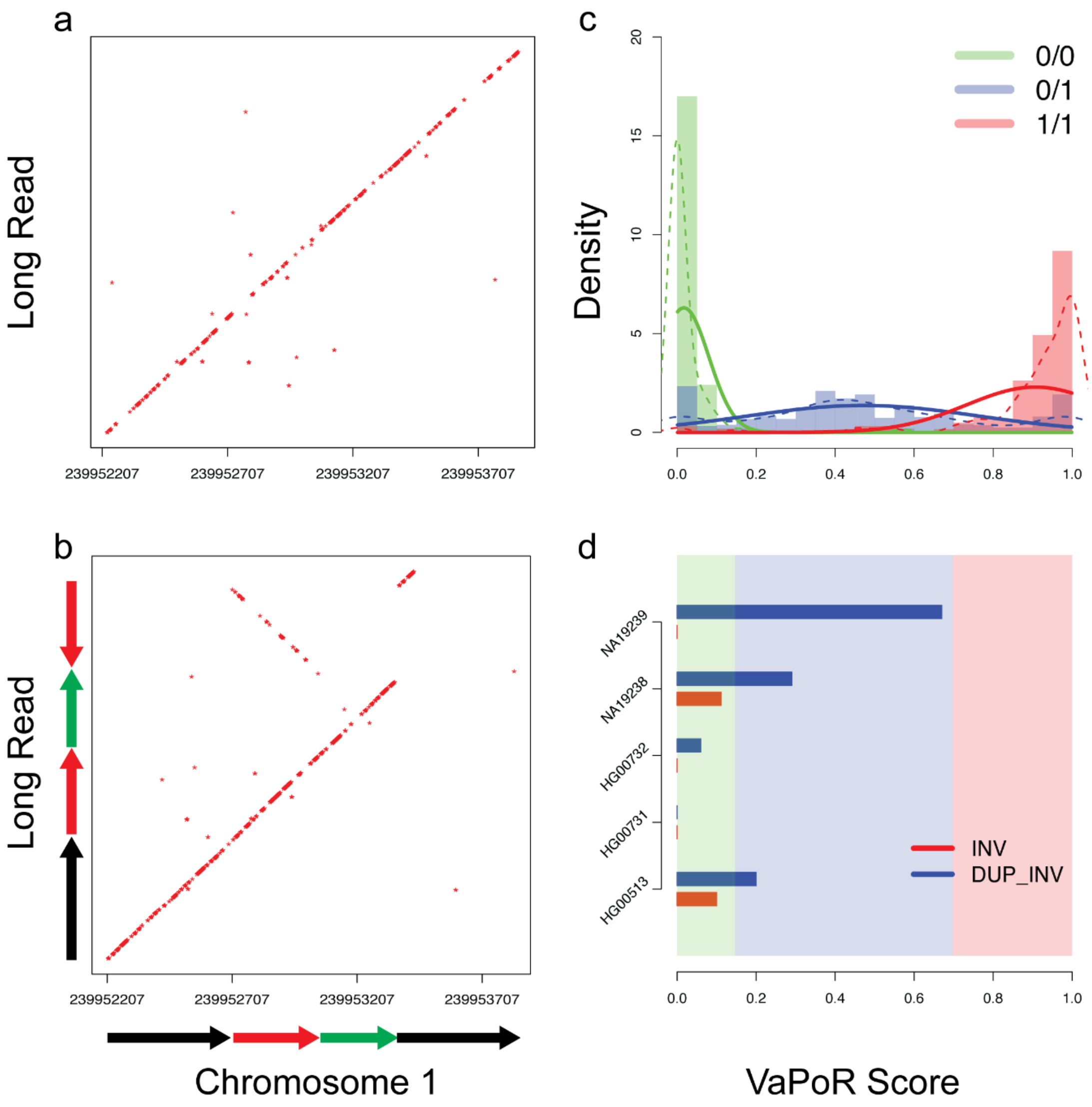

a
Long Read
239952207
239952707
239953207
239953707
b
Long Read
239952207
239952707
239953207
239953707
Chromosome 1
c
Density
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
0/0
0/1
1/1
d
NA19239
NA19238
HG00732
HG00731
HG00513
INV
DUP_INV
0.0
0.2
0.4
0.6
0.8
1.0
VaPoR Score

Figure 3.10 Validation and genotyping of assessed regions using VaPoR.

(a) Recurrence plot of reference genome (GRCh38) to an aligned long read in NA19239 (m150208_160301_42225_c100732022550000001823141405141504_s1_p0/3831/0_12148) for a reported inversion at position chr1:239952707-239953529. The signature is consistent with an inverted duplication structure. (b) Recurrence plot of a different read (m150216_212941_42225_c100729442550000001823151505141565_s1_p0/106403/0_13205) against the same location, consistent with a non-variant (reference) structure. (c) Distribution of VaPoR scores on all reported SVs on chr1 in samples HG00513, HG00731, HG00732, NA19238, NA19239, stratified by color (solid) and modeled with a Gaussian mixture model (dashed). (d) VaPoR scores of SV above now stratified by color as indicated in (c) for both reported inversion (red) and predicted inverted duplication (blue).

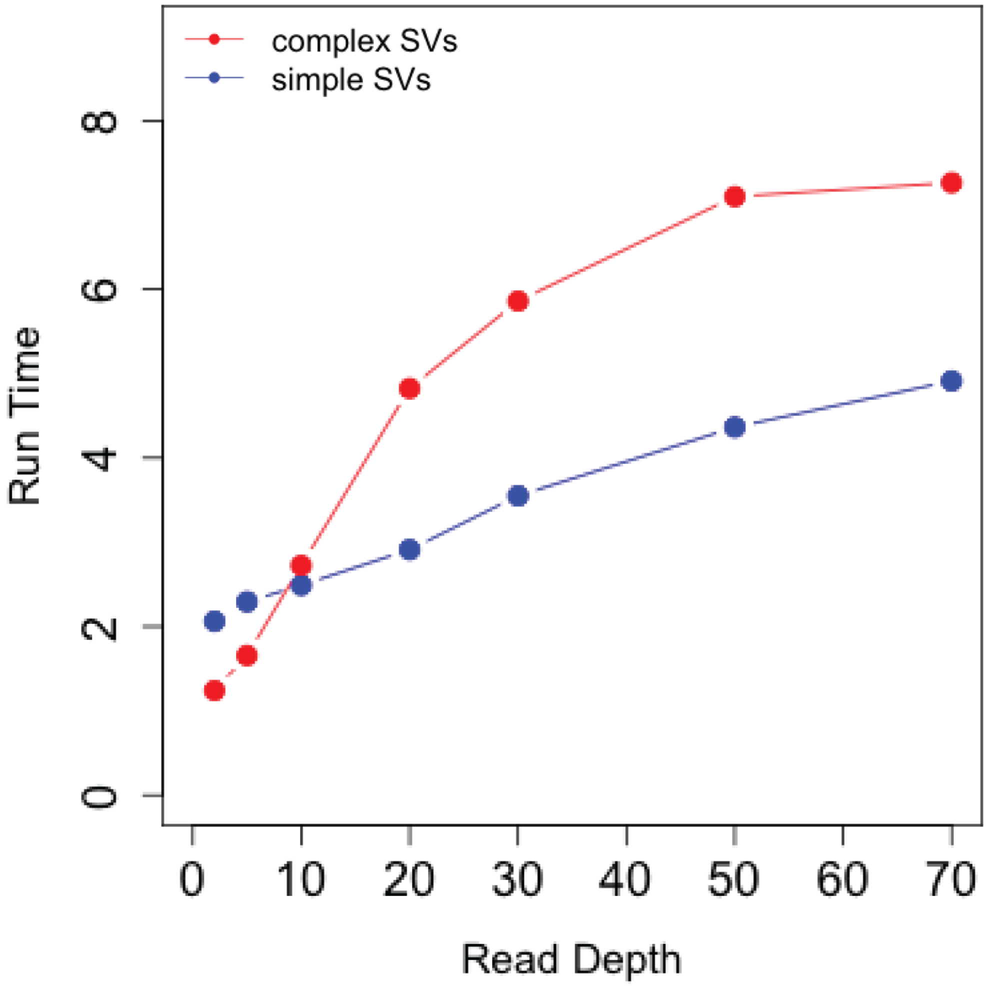


Figure 3.11 Run time summarization of VaPoR.

Averaged run time (seconds) of each simulated SV summarized and plotted at different read depth. Simple and complex SVs are estimated separately, shown in red and blue lines respectively.

CHAPTER IV

Integration of SVs Predicted by Different Platforms

4.1 ABSTRACT

In addition to serving as an important portion of genome diversity, genomic structural variants are also indicated to closely relate to large spectrum of human diseases ranging from developmental disorders to various cancers. In spite of the advances in sequencing technology and discovering methodology in recent decade, limitations remain for accurately yet comprehensively discovering SVs at whole genome scale through next generation sequencing data (NGS). This is an overall result of the relatively short reads from NGS, the complexity of human genome and the inevitable false discovery rate of individual discovering method. However, the discovery rate can be significantly improved by either properly integrating SVs called by multiple algorithms, or adding additional information from orthogonal sequencing technologies such as long read sequences.

Here, we presented an integration of genomic variants summarized from multiple genomic variants discovering algorithms, which showed significant improvement in true discovery rate while still keeping the low false discovery rate. Moreover, increased accuracy was achieved in defining breakpoints with confident intervals assessed and reported. We have also

conducted a systematic comparison between genomic variants discovered from multiple sequencing platforms including short read and long read sequencing, where we showed that long reads are with outstanding superiority in discovering small variants while paired-end short sequences are more capable of describing complex forms of variants that have long distance translocations of DNA materials involved.

4.2 KEYWORDS

structural variants integration

short read sequencing

long read sequencing

breakpoint accuracy

complex genomic structural variants

4.3 INTRODUCTION

The appearance of high throughput next generation sequencing (NGS) has dramatically brought down the cost of human genome sequencing (Goodwin, McPherson, and Richard McCombie 2016), thus greatly accelerating the overall study of human genomic variants. Short read pair end sequencing technology, represented by Illumina HiSeq, are widely used to detect genomic structural variants (SVs), with a large amount of data produced and various methods developed. However, it’s not yet feasible to accurately describe all SVs in any individual genome according to alignment biases resulting from the relatively short reads produced, as well as the inevitable false positive / negative rate of existing algorithms introduced by the randomness while selecting sequences from the pool of DNA segments.

Long read sequencing, eg. the single cell real time sequencing (SMRT) from the Pacific Biosciences (PacBio), that deliver reads that are several kilobases or longer, could fully cover the ambiguous genomic regions such as segmental duplications or long simple repeats, thus providing direct insight into those regions, with multiple published work demonstrating the superiority of long reads in detecting SVs compared to NGS (Pendleton et al. 2015; Chaisson, Wilson, and Eichler 2015; Shi et al. 2016). Except for the being able to accurately define SVs in genomic regions that are of high complexity or repetitiveness, directly comparing long sequences against the reference genome, in theory, allows for SVs in all forms to be defined with breakpoints at single base resolution. However, the current cost of long read sequencing, as well as the limited options of detecting algorithms, prevents it to be widely applied onto most of the research projects. Instead, NGS still remains the most commonly adopted platform with much larger amount of data already produced and served as references for future studies.

The 1000 Genomes Project have been dedicated in developing comprehensive SV sets across multiple populations, with large amount of investment in deep sequencing human samples with diverse genomic background. Recently, three father-mother-child trios from Han Chinese (CHS), Puerto Ricans (PUR) and Yoruba (YRI) were deep sequenced with various sequencing technologies such as Illumina pair end short libraries, PacBio long reads (Rhoads and Au 2015), 10X linked read sequencing (“[PDF]One System, One Workflow, Powerful New Sequencing

Applications," n.d.), BioNano (Lu, Giordano, and Ning 2016) and Strand-Seq (Falconer and Lansdorp 2013) . The availability of data sequenced with multiple platforms makes it possible to systematically assess the performance of current algorithms, with support from orthogonal technique as the external adjudication.

In this work, we first developed an integration pipeline that summarized SVs detected by numerous short read based SV detecting methods, the resulting set of which showed significantly increased discovering power with reduced false discovery rate, as well as breakpoints defined at higher resolution. At the same time, we have also integrated SV set discovered from long sequences, i.e. PacBio sequences. We conducted a systematic comparison between the sets proposed by different platforms, and showed that:

1. long sequences have more power in defining small deletions and insertions, which were usually reported at high / single base breakpoint resolution.
2. short sequences, as with more mature methodology, showed higher capability in defining complex SVs that have multi-step rearrangements involved, and tracing the origin of long-distance translocations.

The integrated SVs proposed in this work represents the best set that current technology and methodology can achieve, which not only refresh out vision of the frequencies and formats of SVs in individual human genome, but also sheds lights on the strength and weakness of current available technologies, which serve as valuable instructions for the community while deciding on sequencing platforms to adopt.

In the following text, the algorithms that detect SVs from Illumina paired end short library sequences are referred to as 'Illumina caller' for short and the corresponding SV sets are named 'Illumina SVs' or 'Illumina set', while the methods and sets related to PacBio long seqeunces are named as 'PacBio caller' and 'PacBio SVs' in abbreviation.

4.4 RESULTS

4.4.1 Data overview

9 individuals from three father-mother-child trios were deep sequenced to ~170X by Illumina paired end short insert sequencing. Children (HG00514, HG00733, NA19240) were sequenced up to ~40X by long read PacBio sequencing, while the parents' genome were sequenced to ~20X. Details of the sequencing platforms were summarized in table 4.1.

15 different algorithms (Rausch et al. 2012; Layer et al. 2014; Kronenberg et al. 2015; Zhao et al. 2016; Chen et al. 2016; Michaelson and Sebat 2012; Chong et al. 2016; Handsaker et al. 2015; Ye et al. 2009; Hormozdiari et al. 2010; Collins et al. 2017) were applied to predict SVs with the Illumina paired end sequences, including two read depth based CNV callers (digital CGH, GenomeStrip) and two mobile element insertion(MEI) detecting algorithms (MELT, Tardis). Most algorithms predict SVs over 100 bp, with the exception that Pindel and Manta also report small indels through their split read module. Number of calls in each type predicted by different algorithms were integrated in Table 4.2

4.4.2 Individual caller assessment

We first assessed the performance of each short read SV detecting algorithms against the PacBio integration set through 50% reciprocal overlap, where we found that at maximum 17% PacBio proposed variants were successfully recapitulated by individual algorithm (Table 4.3), while the relative specificity ranges from 25-90%. We have also compared the breakpoints called through different platforms, by calculating the distance between breakpoints of Illumina SVs and the PacBio SV that shared >50% reciprocal overlap with it (Figure 4.2). Most of the algorithms predict SVs with breakpoints within 20bp from a PacBio SV, with few exceptions like dCGH and HOLMES that only focuses on large events or make predictions from long insert libraries.

Table 4.1 Summarization of sequencing technology applied to the three trios

	Avg. cov,	Avg. frag. len.(bp)	Avg. cov.	Avg. frag. len.(bp)	Avg. cov.	Avg. frag. len.(bp)
Han Chinese	HG00512		HG00513		HG00514	
PacBio	22.04	9,060	17.39	9,855	41.8	10,120
Illumina short insert	167	679	170	680	178	720
Illumina liWGS	154.26	3,417	162.53	3,329	138.11	3,339
10X Chromium	40	88 K	41	73 K	71	72 K
BioNanoGenomics	88	281 K	125	268 K	147	304 K
Tru-Seq	5.22	4,195	3.31	5,198	1.44	4,825
Puerto Rican	HG00731		HG00732		HG00733	
PacBio	23.005	9,545	23.08	9,410	39.43	10,119
Illumina short insert	177	702	162	673	169	703
Illumina liWGS	156.21	3,452	142.68	3,537	188.59	3,751
10X Chromium	39	108 K	44	82 K	79	88 K
BioNanoGenomics	82	260 K	112	258 K	142	285 K
Tru-Seq	4.81	4,137	2.97	5,002	2.6	5,077
Yoruban	NA19238		NA19239		NA19240	
PacBio	18.21	5,702	16.5	5,420	37.67	5,619
Illumina short insert	174	712	174	707	165	668
Illumina liWGS	154.28	3,506	153.73	3,433	178.81	3,509
10X Chromium	43	100 K	43	91 K	85	108 K
BioNanoGenomics	90	286 K	151	300 K	113	285 K
Tru-Seq	4.02	5,116	2.69	5,418	4.17	5,129

Table 4.2 Summarization of SVs detected by each algorithm, categorized by individual SV types.

Algorithm	Technology	SV_Type	Total SVs	DEL	DUP	INS	INV	Other
dCGH	Illumina	DEL,DUP, CPX	1820	1094	726	0	0	0
GenomeStrip	Illumina	DEL, DUP	2846	2379	467	0	0	0
Delly	Illumina	DEL,DUP, INV, CPX	5307	4873	258	159	17	0
NovoBreak	Illumina	DEL,DUP,INV	6375	5974	89	0	312	0
Pindel	Illumina	DEL,DUP,INV	14348	11740	0	2608	0	0
retroCNV	Illumina	DUP	19	0	0	19	0	0
SVelter	Illumina	DEL,DUP,INV,CPX	23145	12375	7868	0	344	2558
VH	Illumina	DEL	4787	4787	0	0	0	0
Wham	Illumina	DEL,DUP,INV,INS	6117	5018	731	0	368	0
Lumpy	Illumina	DEL,DUP,INV	12067	8760	3006	0	301	0
ForestSV	Illumina	DEL,DUP	1117	1103	14	0	0	0
Manta	Illumina	DEL,DUP,INV,INS	21000	14711	1739	3294	1256	0
MELT	Illumina	MEI, DEL	5867	1770	0	4097	0	0
Tardis_MEI	Illumina	MEI	4125	0	0	4125	0	0
HOLMES	Illumina, jumping lib	DEL,DUP,INV,INS,CPX	1046	784	151	0	111	0

Table 4.3 Pseudo sensitivity and specificity of individual short read SV detecting algorithms, compared against PacBio SVs.

	Discovery rate			Extra calls		
	HG00514	HG00733	NA19240	HG00514	HG00733	NA19240
Manta	0.17	0.16	0.17	0.65	0.68	0.65
Pindel	0.15	0.14	0.15	0.54	0.56	0.57
lumpy	0.1	0.09	0.11	0.56	0.56	0.61
SVelter	0.08	0.07	0.08	0.39	0.37	0.39
wham	0.08	0.07	0.08	0.83	0.84	0.82
MELT	0.07	0.07	0.08	0.89	0.92	0.84
VH	0.07	0.06	0.07	0.86	0.88	0.85
novoBreak	0.05	0.05	0.06	0.75	0.74	0.74
Delly	0.05	0.04	0.05	0.87	0.88	0.85
GenomeStrip	0.02	0.02	0.02	0.54	0.52	0.54
liWGS	0.01	0.01	0.01	0.69	0.72	0.63
ForestSV	0.01	0.01	0.01	0.28	0.3	0.28
dCGH	0.01	0.01	0.01	0.24	0.24	0.26
retroCNV	0	0	0	0	0	0
Tardis	0	0	0	0	0	0

4.4.3 Overview of Illumina integrated SV set

The integration of Illumina SVs consists of 44505 unique SVs that are 50bp or larger, defined across all 9 samples (Table 4.4), with each individual genome carrying ~20,000 SVs. Most of the merged SVs were assigned with unified SV types such as simple deletions, duplications, inversions and insertions. However, there are ~5% events that have ambiguous SV types where different algorithms predict different types of variants, which indicate the possibility of complex SVs. A detailed tabulation of SVs for each individual is summarized in table 4.4.

4.4.4 Primary quality controls on the integrated set

The quality of the integrated Illumina SVs was assessed in terms of their locations and sizes, as well as the redundancies where the same SV were represented by multiple records in the set. In brief, there are 421 SVs in the integrated set that fell within either telomere or centromere regions, and 594 over 1Mb in size, both of which were labeled as LowQual to avoid the potential confusion they would introduce for downstream analysis.

The redundancy rate was defined as the percentage of SVs that overlap with another in the set, which was estimated as 32% (n=14373. However, the vast majority of the redundancies were singletons, those contributed by a single algorithm (n=13845, 96%), which could be considered as either false discoveries or outliers that failed to be merged because of the offset breakpoints. On this other side, a singleton could be well represented by a cluster that's merged from SVs called by multiple callers, if they share overlaps. With these, singletons that overlap with a cluster were also labeled as LowQual.

After the primary quality controls, 70% deletions(n=18678), 46% duplications(n=2347) and 34% inversions(n=387) were kept as ‘PASS’, while almost all insertions, and ~70% or higher portions of the ambiguous types were also kept (Table 4.5). In summary, we have 32266 SVs, spanning 29896 non-overlapping genomic regions, kept after the primary quality control step, while still have a redundancy rate of 7% which requires more careful inspection with the aid of external validation approaches.

The frequencies and formats of SVs discovered by short sequences in each individual genome were listed in table 4.6, where we showed that ~8,000 deletions (including ~1,000 ALU and ~70 LINE1 deletions) and ~3,500 insertions (including ~1,200 ALU, ~180 LINE1, ~100 SVA and 0-3 HERVK) that are 50bp or larger were defined per genome. At the same time, the short sequences also discovered 300-400 duplications, ~1,200 multi-allelic copy number variants, ~130 inversions and 80-100 complex SVs.

Table 4.4 Summary of SV set integrated from short read based SV detecting algorithms

	DEL	DUP	INV	INS	DEL+ DUP	DEL +INV	DUP +INV	DEL+ DUP+ INV	Total
HG00512	10796	2048	422	3902	891	84	35	40	18218
HG00513	10704	2055	416	3961	908	93	38	46	18221
HG00514	10892	2021	436	3956	875	91	40	44	18355
HG00731	10936	2041	392	3939	908	86	43	46	18391
HG00732	10689	1976	456	3923	904	94	38	46	18126
HG00733	10698	2056	411	3974	871	91	40	45	18186
NA19238	12153	2137	452	4609	981	102	37	41	20512
NA19239	12093	2117	473	4552	928	93	36	45	20337
NA19240	12218	2154	468	4660	946	99	36	46	20627
all	26849	5099	1146	9309	1794	178	66	64	44505

Table 4.5 Number of SVs in each type that passed the primary quality control.

	DEL	DUP	INV	INS	DEL;DUP	DEL;INV	DUP;INV	DEL;DUP;INV
ALL	26848	5099	1146	9309	1795	178	66	64
PASS	18678	2347	387	9293	1325	136	46	54
PASS/ALL	69.57%	46.03%	33.77%	99.83%	73.82%	76.40%	69.70%	84.38%

Table 4.6 Number of SVs in each type in each individual geome

	DEL	DEL	DEL	INS	INS	INS	INS	INS	DUP	CNV	INV	CPX
SAMPLE		ALU	LINE1		ALU	LINE1	SVA	HERVK				
HG00512	7,410	988	61	3,198	1,115	173	93	2	309	1,265	128	80
HG00513	7,471	993	63	3,314	1,141	184	97	1	330	1,265	124	88
HG00514	7,546	990	64	3,321	1,134	179	96	2	310	1,265	122	90
HG00731	7,573	993	68	3,234	1,132	182	91	0	303	1,265	124	86
HG00732	7,450	979	62	3,270	1,162	180	92	2	292	1,263	145	80
HG00733	7,475	1,003	68	3,284	1,180	169	91	1	316	1,264	134	92
NA19238	8,567	1,148	75	3,825	1,472	192	109	3	352	1,265	131	95
NA19239	8,539	1,177	76	3,765	1,448	195	101	2	347	1,264	145	89
NA19240	8,629	1,179	78	3,983	1,536	188	113	2	373	1,264	134	102

4.4.5 Systematic comparison of SVs discovered by different platforms

The deletions and insertions detected by the Illumina sequences in each individual were first compared against those predicted by PacBio long sequences, where deletions that share over 50% reciprocal overlap and insertions that have an insert point within 20bp and predicted insert length deviates within 20% were considered as the same event. The results of this comparison are displayed in Figure 4.3, where we found ~4000 deletions and ~2000 insertions shared by both platforms. There are 4000 – 5000 deletions and 15000 - 16000 insertions uniquely discovered by PacBio technology. We further compared the distributions of SV length between both platforms. As shown in Figure 4.4, the two platforms discover about the same number of deletions that are larger than 300bp, while PacBio sequences have significantly more small deletions (<300bp), and insertions of the full length spectrum than Illumina.

We also checked the overall distribution of overlap portions between Illumina and PacBio SVs, where an Illumina SV was paired up with the PacBio SV that shared the largest reciprocal overlap, if present. The distribution of number of SVs versus the reciprocal overlap range are shown in Figure4.5A, where we observe that 39% PacBio SVs were covered by SVs in the integration set by >50 RO, which comprises 52% of the Illumina integration set. Comparing to results from individual Illumina callers, we observed a 20% increase in sensitivity. The singletons and clusters in the integration set were differentiated in this comparison, where we observed that proportion of clusters increases with the reciprocal overlap, indicating a relatively higher quality of the clusters.

We next examined the genotype concordance between Illumina and PacBio SVs as an external assessment matrix. For each pair of SVs that share over certain reciprocal overlap, the agreement in number of alternative alleles is counted as their genotype concordance. Thus, 2 is assigned if the pair of SVs have the same genotype, 1 if homozygous and heterozygous SV predictions were made respectively, and 0 if homozygous reference is predicted in one SV while homozygous alternative is in the other. As shown in Figure 4.3B, genotype concordance increases with reciprocal overlap, indicating a high fidelity of the genotypes of integrated SVs.

To evaluate the PacBio SV set, we have included external algorithms for the assessment, which include SV calls from two hybrid algorithms, i.e. HySA(Fan et al. 2017) and cloudSV, that combine both short and long sequences for the discovery of genomic variants, as well as two SV validation approaches i.e. VaPoR and Graphite, that assess the quality of SVs by seeking for evidence from long reads and short reads respectively. Each of the four algorithms were treated equally for the assessment, and we showed that the external support also increases with the reciprocal overlap between the Illumina and PacBio SVs.

As deletions and insertions are the major components of SVs in both the Illumina and the PacBio set, these two types were compared between the platforms for an overall estimation of their similarity and difference. Deletions sharing over 50% reciprocal overlap and insertions with insert point within 20bp and insert length differ less than 20% are considered as overlaps. As shown in Fugure 4.3D, that there are 7,000 to 8,000 deletions per individual that are uniquely discovered by PacBio, while 3,000 to 3,500 unique to Illumina. These two platforms share 4,000-5,000 deletions in their discovery set. For insertions, there are ~ 16,000 PacBio unique events while only ~2,200 Illumina unique ones, indicating the possible limitation of Illumina sequencing technology of SV detecting algorithms in deciding the insertions. However, it should be noted that insertions are sometimes labeled as duplications instead by Illumina algorithms, which might explain the significantly smaller number of insertions in the Illumina set.

4.5 METHOD

With the high depth Illumina sequences, a total of 15 differen Illumina algorithms (Handsaker et al. 2011; Kronenberg et al. 2015; Rausch et al. 2012; Layer et al. 2014; Zhao et al. 2016; Chong et al. 2016; Michaelson and Sebat 2012; Chen et al. 2016; Ye et al. 2009) have been applied with the parameters either specified by the algorithm developer or default. SVs callers were applied in parallel on 9 individuals with results integrated in vcf formats. At the same time, , Dr. Chaisson and Dr. Bashir in the consortium have also produced an integrated set of SVs detected by PacBio sequences. These are the two main sets that were compared and then combined to reach a final discovery set, with additional validations provided by two hybrid (short read and long read) SV discovery algorithms, i.e. HySA(Fan et al. 2017) and cloudSV (not yet published) and two SV validation approaches (Graphite, and VaPoR [http://biorxiv.org/content/early/2017/02/24/105817]).

4.5.1 Integration of SVs detected by different short library based algorithms

To achieve an integrated Illumina SV set with optimized breakpoint resolution, a two-layer breakpoint focused integration pipeline was developed (Figure4.1), where the breakpoint precision was first assessed against SVs predicted by long PacBio reads (PacBio SVs) and then clustered based on the estimated varying ranges.

To assess the breakpoint precision of a Illumina caller, the SV predictions contributed by this algorithm were first compared against the integrated PacBio SVs by 50% reciprocal overlap, then the distance between the matched Illumina and PacBio SVs were collected to approximate the distribution of the breakpoint precision, with 10% and 90% quantile assigned as the confident interval (CI).

The next step was to cluster breakpoints by overlapping their breakpoint variation, where breakpoints from all Illumina callers, as long as their CIs overlap, were clustered to from initial breakpoint groups. For each group, the minimized common region shared by all CIs, is there's any, is assigned as the consensus CIs for the merged breakpoint, with the most frequently proposed breakpoint assigned as the consensus. In the situation where not a common consensus CI can be derived, a pseudo-kernel density model was adopted instead to assign the consensus breakpoints (figure2). Where the number of intervals that span through each breakpoint was counted as its 'density', and the consensus breakpoints are assigned at the peaks. Intervals in the group will then be assigned to their closest consensus breakpoint.

With consensus breakpoints decided, the next step is to pair them up for complete description of SVs such as deletions, duplications and inversions where two breakpoints are required to characterize an event. The pair up were conducted by linking breakpoints where their supportive breakpoints come from the same event.

Insertions are different from deletions, duplications and inversions in the aspect that only one insertion point is required for complete characterization, with the insertion length left as optional, depending on the characteristics of each algorithm. In this situation, the confident interval of insertion point accuracy cannot be defined by 50% RO comparison. A different approach is adopted to estimate the breakpoint accuracy, where each short-read predicted insertion point will be compared to its closest PacBio predicted breakpoint, with the distances (if < 1KB) collected and confident interval assigned accordingly. Insertion points were clustered in the same way described above and no BP pairing step is necessary here.

4.5.2 Integration of SVs with and without single base resolution

Unlike other SV detecting algorithms that mainly depend on the read pair and read depth approach, GATK(McKenna et al. 2010) has a special design implemented for the discovery of smaller event that are under 100 bases, i.e. indels, while with the capability to decide accurate breakpoints. Similarly, Pindel and Delly have also included split read modules to accurately define indels with single base resolution reported. A quick integration was conducted to directly merge indels predicted by these three algorithms by combining the events at exactly the same locus and assigning the most frequently proposed genotypes as the consensus for each individual, the set of which were later added to the Illumina integration set described above based on these rules:

1. Any small indels with <50bp DNA bases deleted or inserted were included

2. Any indels >50bp were first compared against the SV integration set, with both SV position and length considered together. Deletions in both sets, if overlap by >80% reciprocal overlap and breakpoints within 20bp, were merged together. For insertions, differences in insertion sites and insert length were both reqiured to be within 10bp to be merged.

3. Any event in the indels and SVs that do not overlap, are kept together for downstream quality controls.

4.5.3 Integrate SVs detected by different sequencing platforms

Comparison between the Illumina and PacBio SV sets were conducted by checking: the reciprocal overlap of SVs in both sets, genotype concordance, and number of external supports. See results for more details.

To integrate the two main sets, support vector machine(SVM) models were iteratively trained on each of them, which was started by manually picking up the 'best' and 'worst' group of each set. For PacBio SVs, 'best' sets are defined as: SVs that share >90% reciprocal overlap with a Illumina merged non-singleton SV and have external support from at least two of the validation approaches, while 'worse' set were those that do not overlap with any Illumina merged SVs neither were supported by any external tools. Similarly for Illumina SVs, the subgroups with extreme qualities have been selected based on the reciprocal overlap with PacBio SVs as well as number of external supportive.

The initial SVM model was trained with the extreme subgroups with the 'best' set labeled as 1 and the 'worst' as -1, with these features kept as variables: number of supportive ILL callers for each ILL integrated SV, reciprocal overlap of an SV with the closest HySA prediction and cloudSV prediction respectively, VaPoR and Graphite validation scores respectively. The model was then applied to all SVs. In the following interations, SVM model were trained on subsets of SVs randomly selected from those with positive prediction scores (labeled 1) and negative scores (labeled -1), and applied to the whole SV set. The SVMs were conducted by the algorithm implemented in R package e1071 (Dimitriadou et al. 2004), and the difference of models between adjacent iterations were estimated by calculating the Pearson correlation coefficient(Benesty et al. 2009) of their weight vectors, used to examine whether the model have reached the convergence.

With SVM model trained on both set independently, SVs with positive prediction values were kept and merged as the first tier integration.

4.6 DISCUSSION

In this work, we first proposed an efficient pipeline to integrate SVs discovered by different methods, with the capability to estimate the accuracy of the final consensus breakpoints. The integrated set of SVs, compared to those proposed by single algorithms, showed significant increased predicting power as well as breakpoint precision. With this method, we provided a comprehensive set of SVs (over 50 bp) that were integrated from 15 different short read based SV detecting algorithms, which includes 18,630 deletions and 8,630 insertions, 2,069 duplications, 348 inversions as well as 1,503 complex SVs that have mutli-step accumulative rearrangement involved.

Moreover, we have also systematically compared the SVs discovered from different sequencing platforms, i.e. Illumina short read paired end sequencing and PacBio long sequences. As illustrated in Figure 4.3, nearly half of the Illumina deletions and insertions were shared by the PacBio discoveries, which consist ~30% of the overall PacBio deletions and only ~10% of the PacBio insertions. With a closer insight into the length distribution as shown in Figure 4.4, we showed that PacBio long sequences show significantly higher power in discovering small deletions (<300bp) and insertions at full length spectrum. However, it should be noticed that most of the short read based algorithms report duplications instead of insertions, which partially explain the fact that number of the insertions discovered by short sequencers are depleted except for the ALUs

(~300bp peak) and LINE1s(~6Kb peak). Rather than considered as reduced discovering power of short sequences, the strength lies in the fact that the duplications clearly stated the origin of the inserted sequences, which could greatly aid the downstream analysis such as functional annotation of duplicated genomic factors.

In spite of the large number of SVs that were uniquely proposed by PacBio sequences, we have also found ~100 SVs per individual that were, instead, missed by them. One of the possible explanation here is the relatively low read depth of long sequences because of their current cost, so that the assembly might fail in certain regions due to the lack of enough reads. The other aspect to consider, is the fact that long sequences is relatively new compared to short sequences, for which the methodology is less mature so that the complex SVs that have multi-step rearrangements (especially long-distance translocations) involved are usually partially resolved as simple deletions or insertions without the origins of insertion sequences clearly stated.

In conclusion, we showed in this study that the long read sequences have significant superiority, compared to the short reads, in defining small deletions and insertion in terms of the number of events as well as the accuracy of defining breakpoints. However, as with the limitation of current long read based methodology, the short reads show stronger capability in defining complex genomic structural variants that have multiple breakpoints and rearrangements involved. At the same time, the fact that most of the currently available long read based SV discovery pipelines require global or local assembly presents as a barrier for most researchers to independently conduct SV discoveries from long sequences.

4.7 FIGURES

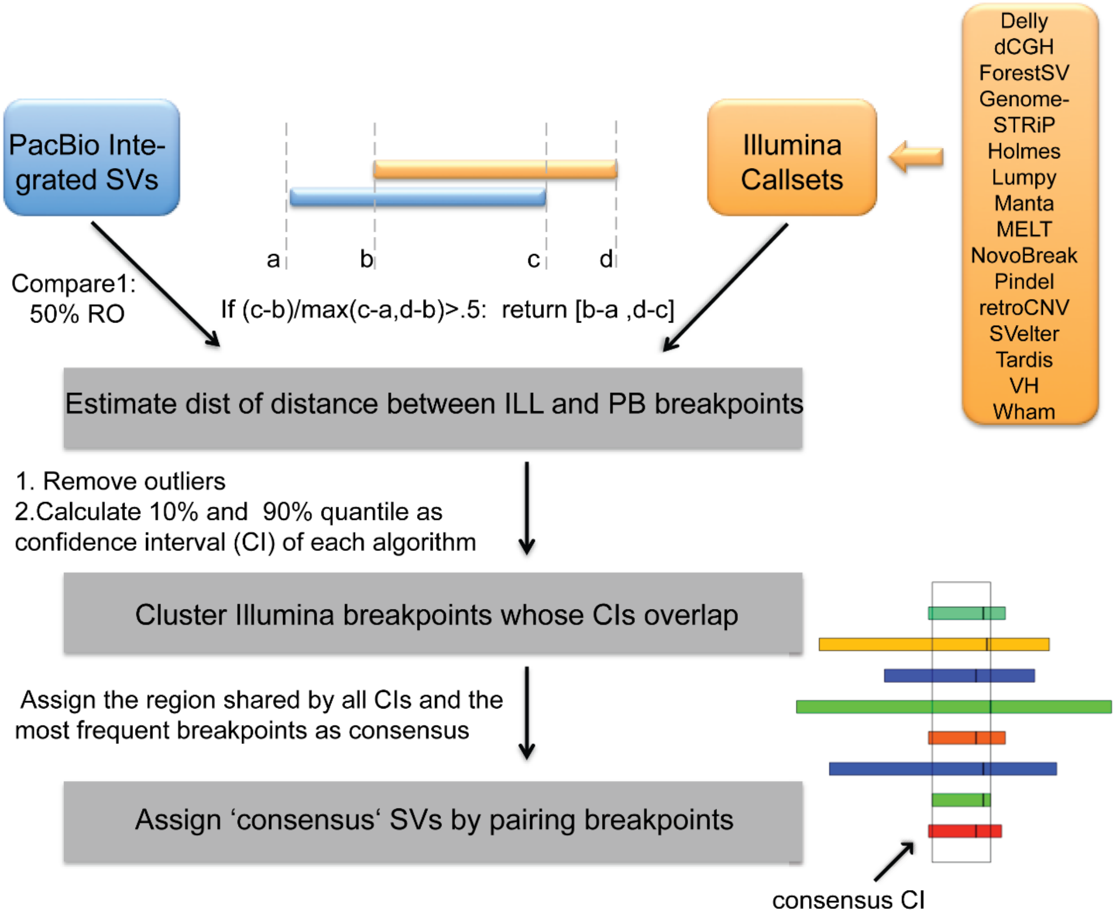

PacBio Integrated SVs
a
b
c
d
Illumina Callsets
Delly
dCGH
ForestSV
Genome-STRiP
Holmes
Lumpy
Manta
MELT
NovoBreak
Pindel
retroCNV
SVelter
Tardis
VH
Wham
Compare1: 50% RO
If (c-b)/max(c-a,d-b)>.5: return [b-a ,d-c]
Estimate dist of distance between ILL and PB breakpoints
1. Remove outliers
2.Calculate 10% and 90% quantile as confidence interval (CI) of each algorithm
Cluster Illumina breakpoints whose CIs overlap
Assign the region shared by all CIs and the most frequent breakpoints as consensus
Assign 'consensus' SVs by pairing breakpoints
consensus CI

Figure 4.1 Integration pipeline of combining SVs predicted by multiple algorithms.

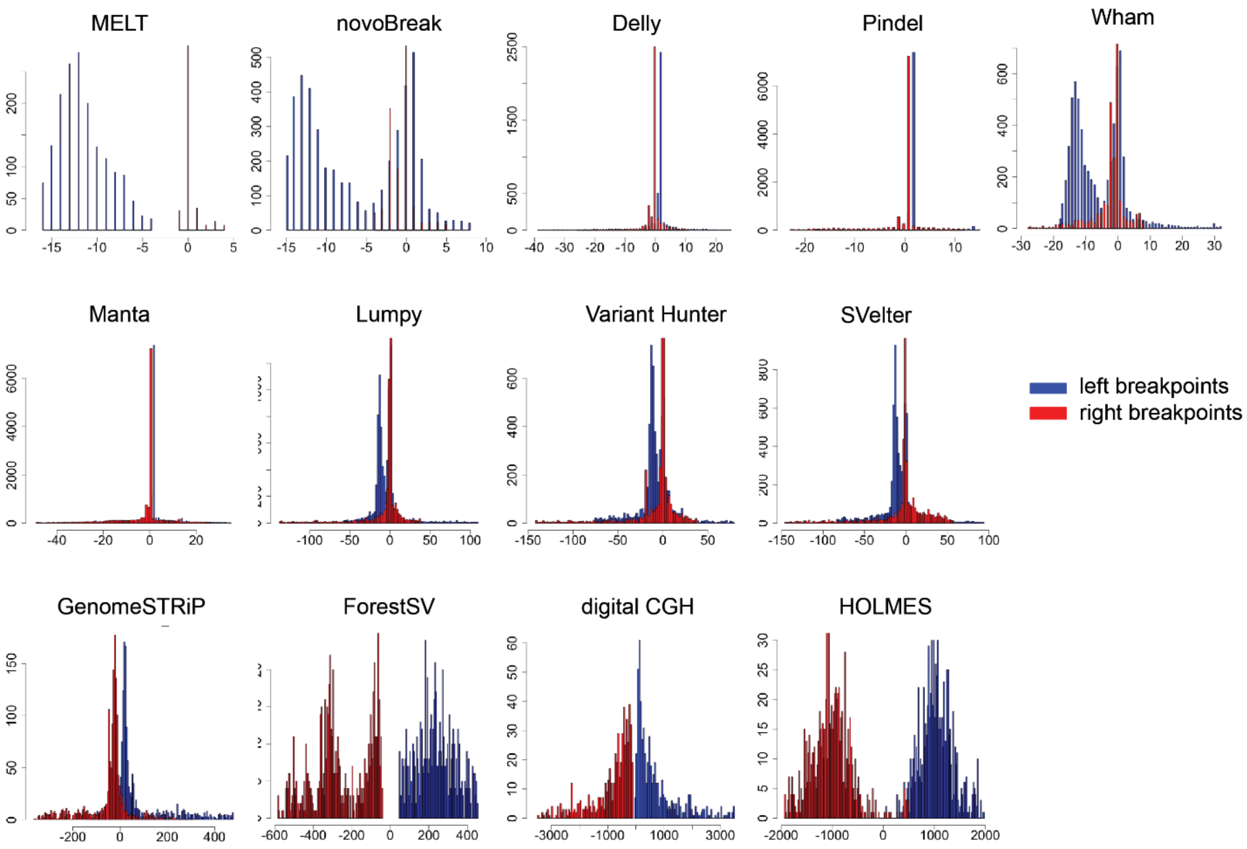


Figure 4.2 Breakpoint precision of different Illumina SV discovery algorithms.

The distributions describe the distance between Illumina and PacBio breakpoints, with left (blue) and right(red) breakpoints calculated separately. The algorithms were ranked by their breakpoint precision from top left to bottom right, with the confident interval ranging from less than 10 bp to over 1Kb.

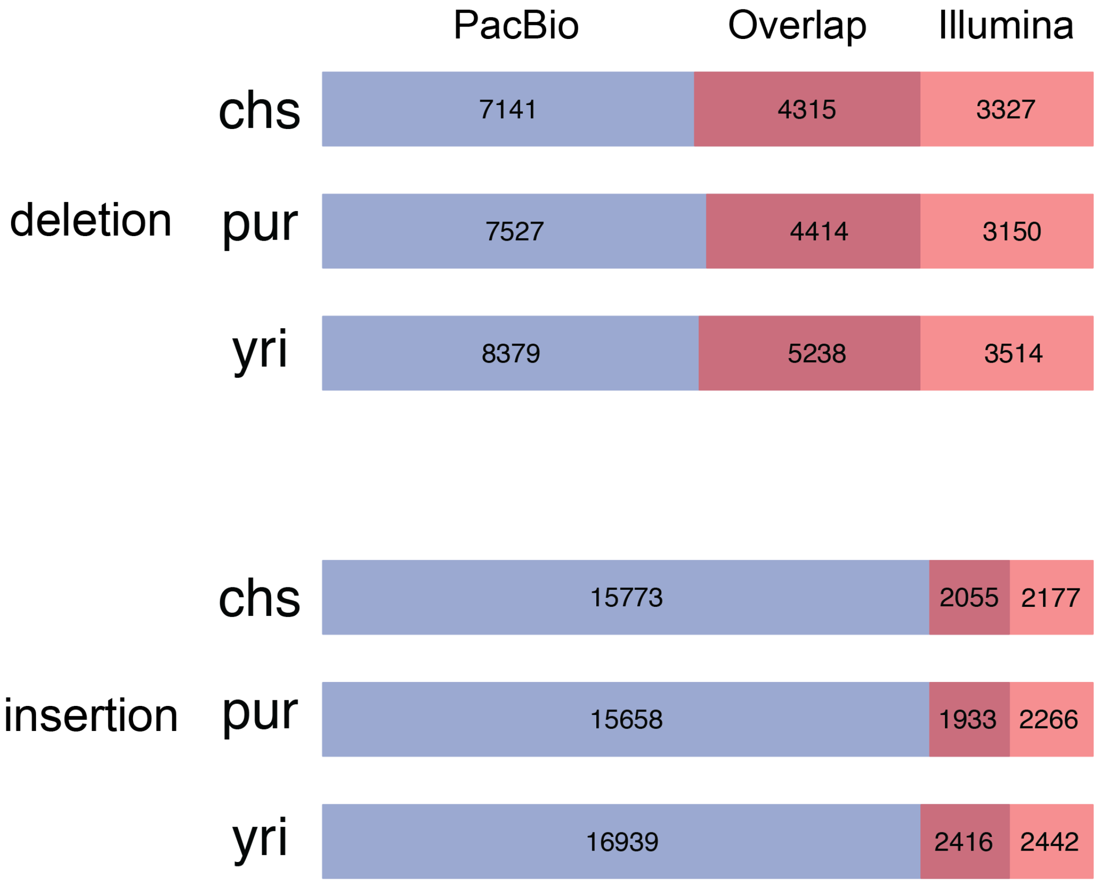


Figure 4.3 Number of deletions and insertions that were uniquely discovered by PacBio and Illumina technologies respectively, and those shared by both technologies.

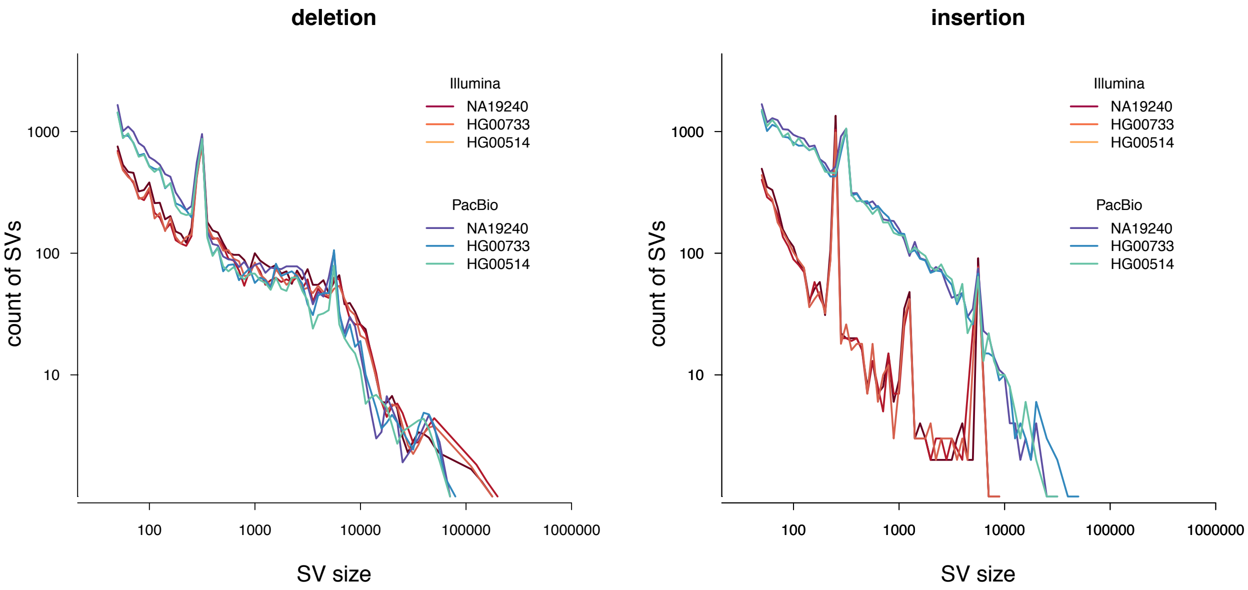
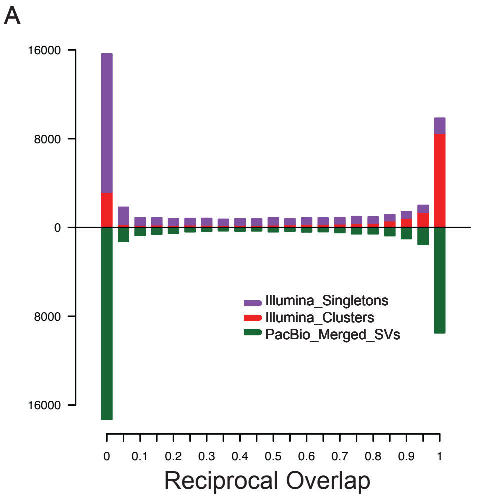


Figure 4.4 Distributions of lengths of deletions and insertions discovered in each child from the three trios, by Illumina and PacBio technology.



RO	PB	Perc	ILL_C	ILL_S	Perc
0	17280	45.61%	3178	6776	33.08%
0-0.1	712	50.80%	207	233	37.18%
0.1-0.2	529	53.78%	223	163	39.85%
0.2-0.3	339	55.70%	231	199	42.74%
0.3-0.4	326	57.35%	225	185	45.38%
0.4-0.5	387	59.17%	245	200	48.17%
0.5-0.6	400	61.09%	314	172	51.22%
0.6-0.7	459	63.30%	331	168	54.45%
0.7-0.8	568	66.31%	408	185	58.34%
0.8-0.9	1005	70.91%	861	211	64.49%
0.9-1	9494	100.00%	8491	652	100.00%

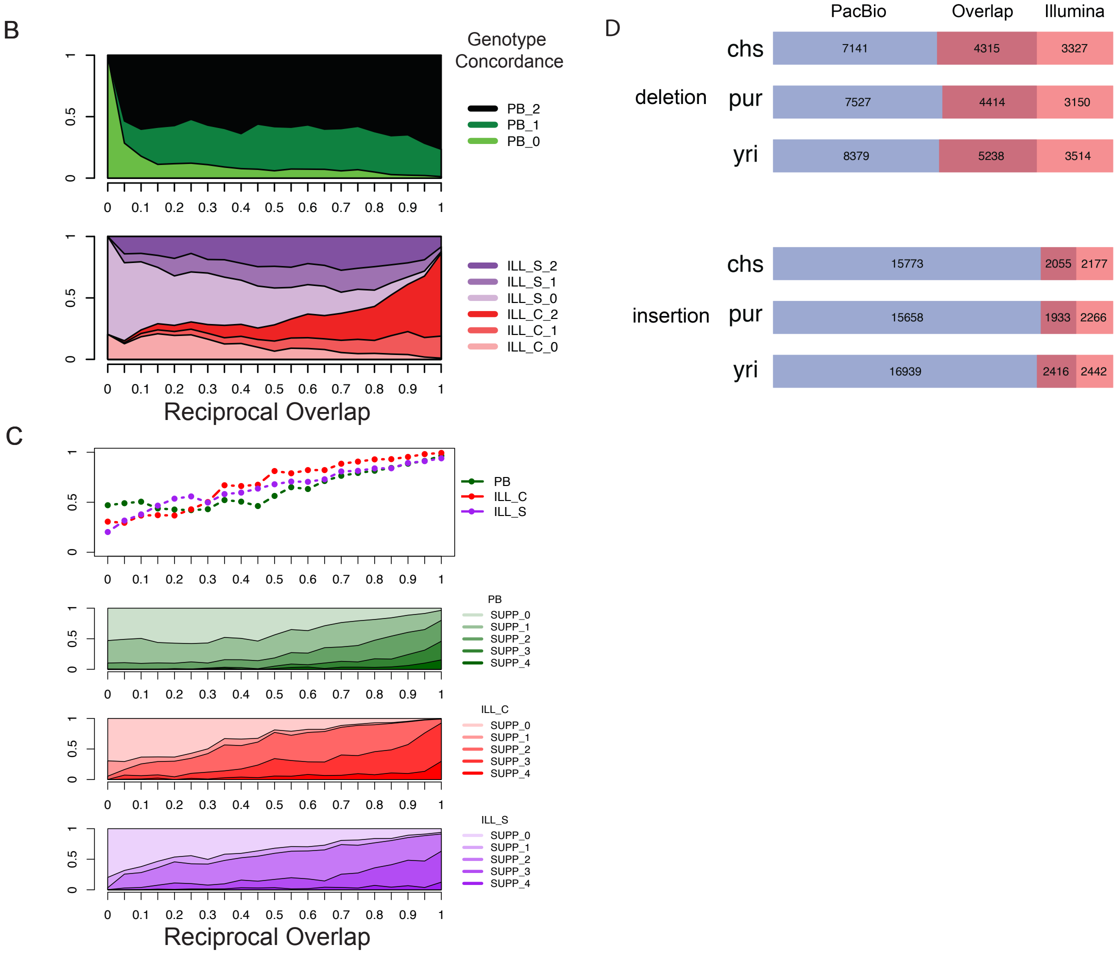


Figure 4.5 Comparison between Illumina and PacBio SVs.

(A) Reciprocal overlap distribution of SVs from both platforms. Most of the SVs from both sets either overlap with >90% reciprocal, or do not interference at all (reciprocal overlap=0). 39% of the PacBio SVs, and 52% of the Illumina SVs share over 50% reciprocal overlap SV from the other set. (B) Genotype concordance comparison. Top panel shows the relative proportion of PacBio SVs that have 2 (PB_2, black), 1(PB_1, dark green) and 0 (PB_0, green) alleles in concordance with Illumina SVs, segmented by the range of reciprocal overlap between PacBio and Illumina SVs. Bottom panel represent the same feather in Illumina SVs, with singletons and clusters described separately. (C) Proportion of SVs supported by at least 1 external validating approach is plotted in the top panel, with the relative proportion of PacBio and Illumina SVs that are supported by 0-4 external validators described in the following three panels. (D) Unique and shared deletions and insertions detected from Illumina and PacBio respectively. There are constantly more SVs discovered by PacBio across the three trios compared to Illumina set.

CHAPTER V

Conclusion and Future Direction

5.1 CONCLUSION

I have shown in this dissertation that the successful discovery and interpretation of genomic structural variants (SVs) is an important yet not fully established area, though SVs consist as a significant component of genomic diversity, and contribute as important causal factors for numerous human diseases. Owing to the relatively low error rate and cost of next generation sequencing, it remains the most popularly adopted platform with large cohorts of datasets produced or being processed, as well as various algorithms developed to fit the genomic variant discovering purposes. However, the following challenges remain for comprehensively discovering SVs across the whole genome with pair-end short insert libraries, which this thesis has focused on developing methodologies to address.

1. Accurately describe SVs in complex formats where multiple rearrangements happen simultaneously or accumulatively at the same genomic locus.

2. Define SVs at the diploid level where the rearrangements on each allele could be accurately described.

3. Detect variants in complex genomic regions that are of high repetitiveness, where the short reads are usually aligned with decreased quality due to the decreased sequencing quality and the increased ambiguity arising from such regions, although still harboring a significant number of variants.

To provide proper solutions to these challenges, this thesis has dived deep into the development of new methods to describe SVs in all formats across whole genome, while also exploring the proper independent and combined application of different sequencing platforms.

In Chapter II, an integrative randomization approach was developed to accurately describe the genomic structural variants in both simple and complex formats, with comparable performance achieved in detecting canonical simple SVs against published format while significant superiority in describing CSVs. Instead of predicting SVs by statically recognizing aberrant alignment patterns, this method searches for the optimized structure through an iterative stochastic process where in each iteration a randomly proposed rearrangement is applied to the current structure and the alignment patterns are compared to decide the superior one. In each iteration, homologous alleles were considered independently to allow read pairs be realigned to the optimized location, so that the overlap events where different variants happen simultaneously on the same genomic location can be accurately detected. With this approach, we found that a large amount of complex SVs were misinterpreted from previous study. For the future direction, application of this method to pathological genomes allows systematically examination of the SV complexity in diseased versus healthy genomes and could potentially reveal new disease causal mechanism.

In Chapter III, a long read based SV validation algorithm was developed and implemented in a user-friendly software named VaPoR. This tool evaluates the quality of predicted SV by assessing the recurrence matrix of long read against reference genome, and is capable of discriminating the partially resolved SVs such as variants with inaccurate breakpoints or wrong formats from the fully resolved high quality predictions. This method avoids assembly of long reads, which is a popularly adopted approach for SV discovery using long reads, to achieve high efficiency while also kept the accuracy by implementing carefully designed statistical models. For the future extension, this method could potentially evolve to a stand along SV caller, with delicate matrix transforming algorithms included.

In Chapter IV, an integration pipeline was established to combine SVs predicted by multiple algorithms with quality control steps carefully designed. We presented the set of SVs integrated from 15 different short read SV callers, which has shown significantly increased sensitivity with decreased false discovery rate. In total, we discovered seven to eight thousand deletions and three to four thousand insertions per individual genome from short sequences. At the same time, ~1,350 SVs in complex / ambiguous formats were discovered per individual, consisting ~10% of all the SVs, implying that CSVs consist as an important portion of the genomic variants.

Moreover, we have also conducted systematic comparison between SVs proposed by different sequencing platforms, i.e. short sequences represented by Illumina paired end sequencing and long sequences established by PacBio, where we verified the idea that long sequences have significantly increased power in describing small SVs with higher, most times single base resolutions, as was stated in previous publications (Pendleton et al. 2015; Chaisson et al. 2014; Shi et al. 2016). However, we have also shown that SVs in complex formats are usually mis- or partially interpreted by long sequences due to the pre-mature methodology nowadays.

With each sequencing platform exhibiting relative advantages and disadvantages, the proper combination of different platforms would potentially present SVs both comprehensively and accurately.

5.2 FUTURE DIRECTION

Most of the effort that has been put onto this thesis work was towards the goal of accurately describing structural variation in human genomes. Instead of pursuing a minimized false discovery rate, as is the focus of most other research projects, the objective of this thesis has put emphases on the sensitivity side and trying to build the most comprehensive set and testing the limit to which extent the most SVs can be correctly described.

As was shown in Chapter IV of this thesis, most of the current genomic structural variants discovery studies focus on understanding SVs in canonical forms, i.e. deletions, insertions and inversions, among which approximately 10% are actually complex events that have multiple genomic pieces involved. Comprehensively defining and interpreting complex SVs helps us piece together different genomic factors as well as get us a finer view of the potential impact of SVs on gene function, thus being especially meaningful for locating disease causal variants in pathogenic genomes. However, more carefully designed methods and analysis pipelines are required to accurately describe such events, where multiple factors outside of the genomic sequences such as evolutionary pressure and the penetration rate should also be comprehensively considered. I have been fortunate to obtain vast experience in sequence analysis and method development, but have been exposed relatively less to these analysis at the population scale, where I would want to gain more experience during my future training.

The other major challenge of locating pathogenic variants remains with cancer genomes, which are usually of high heterogeneity level so that SV discovery models build on the germline diploid genomes are of impacted power in such cases. For these questions, ambitious yet realistic models are required, where the heterogeneity should be well address while the issue of over fitting issues should also be properly avoided. Though challenging, the first chapter of this thesis has provided a foundation, based on how an unbalanced model with weighted frequencies of each allele could be modified from the original.

With vast experiences accumulated in genomic variants discovery, I view the main focus of my future research shifting to interpreting genomic variants in pathological genomes. As has been discussed in the beginning of this thesis, SVs have been revealed by numerous studies to be closely related to neurological disorders such as autism and schizophrenia, as well as large spectrum of cancers. Systematically discovery of the SVs that are unique to, or significantly enriched in diseased genomes, compared against healthy controls, helps locating the potential causal variants thus aiding revealing the pathogenic mechanisms. Future advances in both sequencing technology and computational innovations will pave the way for new understanding in these areas.

BIBLIOGRAPHY

1000 Genomes Project Consortium et al., 2015. A global reference for human genetic variation. Nature, 526(7571), pp.68–74.

1000 Genomes Project Consortium et al., 2012. An integrated map of genetic variation from 1,092 human genomes. Nature, 491(7422), pp.56–65.

Alkan, C., Coe, B.P. & Eichler, E.E., 2011. Genome structural variation discovery and genotyping. Nature reviews. Genetics, 12(5), pp.363–376.

Alkan, C., Sajjadian, S. & Eichler, E.E., 2010. Limitations of next-generation genome sequence assembly. Nature methods, 8(1), pp.61–65.

Anon, Microarray-based Comparative Genomic Hybridization (aCGH) | Learn Science at Scitable. Available at: http://www.nature.com/scitable/topicpage/microarray-based-comparative-genomic-hybridization-acgh-45432 [Accessed March 20, 2017].

Benjamini, Y. & Speed, T.P., 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic acids research, 40(10), pp.e72–e72.

Bhangale, T.R. et al., 2005. Comprehensive identification and characterization of diallelic insertion--deletion polymorphisms in 330 human candidate genes. Human molecular genetics, 14(1), pp.59–69.

Brand, H. et al., 2014. Cryptic and complex chromosomal aberrations in early-onset neuropsychiatric disorders. American journal of human genetics, 95(4), pp.454–461.

Brand, H. et al., 2015. Paired-Duplication Signatures Mark Cryptic Inversions and Other Complex Structural Variation. American journal of human genetics, 97(1), pp.170–176.

Buermans, H.P.J. & den Dunnen, J.T., 2014. Next generation sequencing technology: Advances and applications. Biochimica et biophysica acta, 1842(10), pp.1932–1941.

Campbell, P.J. et al., 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nature genetics, 40(6), pp.722–729.

Carvalho, A.B., Dupim, E.G. & Goldstein, G., 2016. Improved assembly of noisy long reads by k-mer validation. Genome research, 26(12), pp.1710–1720.

Chaisson, M.J.P. et al., 2014. Resolving the complexity of the human genome using single-molecule sequencing. Nature, 517(7536), pp.608–611.

Chen, K. et al., 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature methods, 6(9), pp.677–681.

Chiang, C. et al., 2012. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. Nature genetics, 44(4), pp.390–7, S1.

Chiang, C. et al., 2014. SpeedSeq: Ultra-fast personal genome analysis and interpretation. bioRxiv.

Collins, Ryan L., Harrison Brand, Claire E. Redin, Carrie Hanscom, Caroline Antolik, Matthew R. Stone, Joseph T. Glessner, et al. 2017. “Defining the Diverse Spectrum of Inversions, Complex Structural Variation, and Chromothripsis in the Morbid Human Genome.” Genome Biology 18 (1): 36.

Conrad, D.F. et al., 2010. Origins and functional impact of copy number variation in the human genome. Nature, 464(7289), pp.704–712.

Eid, J. et al., 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. Science, 323(5910), pp.133–138.

Falconer, E. et al., 2012. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nature methods, 9(11), pp.1107–1112.

Falconer, E. & Lansdorp, P.M., 2013. Strand-seq: a unifying tool for studies of chromosome segregation. Seminars in cell & developmental biology, 24(8-9), pp.643–652.

Goodwin, S., McPherson, J.D. & Richard McCombie, W., 2016. Coming of age: ten years of next-generation sequencing technologies. Nature reviews. Genetics, 17(6), pp.333–351.

Handsaker, R.E. et al., 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nature genetics, 43(3), pp.269–276.

Handsaker, R.E. et al., 2015. Large multiallelic copy number variations in humans. Nature genetics, 47(3), pp.296–303.

Heather, J.M. & Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), pp.1–8.

Hedges, D.J. et al., 2012. Evidence of novel fine-scale structural variation at autism spectrum disorder candidate loci. Molecular autism, 3(1), p.2.

Henrichsen, C.N., Chaignat, E. & Reymond, A., 2009. Copy number variants, diseases and gene expression. Human molecular genetics, 18(R1), pp.R1–8.

Iafrate, A.J. et al., 2004. Detection of large-scale variation in the human genome. Nature genetics, 36(9), pp.949–951.

International HapMap Consortium, 2003. The International HapMap Project. Nature, 426(6968), pp.789–796.

Jiang, Y., Wang, Y. & Brudno, M., 2012. PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. Bioinformatics , 28(20), pp.2576–2583.

Kidd, J.M. et al., 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell, 143(5), pp.837–847.

Kidd, J.M. et al., 2008. Mapping and sequencing of structural variation from eight human genomes. Nature, 453(7191), pp.56–64.

Kusenda, M. & Sebat, J., 2008. The role of rare structural variants in the genetics of autism spectrum disorders. Cytogenetic and genome research, 123(1-4), pp.36–43.

Layer, R.M. et al., 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome biology, 15(6), p.R84.

Levene, M.J. et al., 2003. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. Science, 299(5607), pp.682–686.

Liu, B. et al., 2015. Structural variation discovery in the cancer genome using next generation sequencing: computational solutions and perspectives. Oncotarget, 6(8), pp.5477–5489.

Loman, N.J., Quick, J. & Simpson, J.T., 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nature methods, 12(8), pp.733–735.

Manley, L.J., Ma, D. & Levine, S.S., 2016. Monitoring Error Rates In Illumina Sequencing. Journal of biomolecular techniques: JBT, 27(4), pp.125–128.

Marshall, C.R. et al., 2008. Structural variation of chromosomes in autism spectrum disorder. American journal of human genetics, 82(2), pp.477–488.

McCarthy, S.E. et al., 2009. Microduplications of 16p11.2 are associated with schizophrenia. Nature genetics, 41(11), pp.1223–1227.

Medvedev, P., Stanciu, M. & Brudno, M., 2009. Computational methods for discovering structural variation with next-generation sequencing. Nature methods, 6(11 Suppl), pp.S13–20.

Mertens, F. et al., 2015. The emerging complexity of gene fusions in cancer. Nature reviews. Cancer, 15(6), pp.371–381.

Mills, R.E. et al., 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome research, 16(9), pp.1182–1190.

Mills, R.E. et al., 2011a. Mapping copy number variation by population-scale genome sequencing. Nature, 470(7332), pp.59–65.

Mills, R.E. et al., 2011b. Mapping copy number variation by population-scale genome sequencing. Nature, 470(7332), pp.59–65.

Moncunill, V. et al., 2014. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. Nature biotechnology, 32(11), pp.1106–1112.

Mullaney, J.M. et al., 2010. Small insertions and deletions (INDELs) in human genomes. Human molecular genetics, 19(R2), pp.R131–6.

Nakamura, K. et al., 2011. Sequence-specific error profile of Illumina sequencers. Nucleic acids research, 39(13), p.e90.

Pang, A.W. et al., 2010. Towards a comprehensive structural variation map of an individual human genome. Genome biology, 11(5), p.R52.

Parikh, H. et al., 2016. svclassify: a method to establish benchmark structural variant calls. BMC genomics, 17, p.64.

Pendleton, M. et al., 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nature methods, 12(8), pp.780–786.

Pinto, D. et al., 2011. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nature biotechnology, 29(6), pp.512–520.

Pinto, D. et al., 2010. Functional impact of global rare copy number variation in autism spectrum disorders. Nature, 466(7304), pp.368–372.

Quinlan, A.R. & Hall, I.M., 2012. Characterizing complex structural variation in germline and somatic genomes. Trends in genetics: TIG, 28(1), pp.43–53.

Rausch, T. et al., 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics , 28(18), pp.i333–i339.

Rhoads, A. & Au, K.F., 2015. PacBio Sequencing and Its Applications. Genomics, proteomics & bioinformatics, 13(5), pp.278–289.

Roberts, R.J., Carneiro, M.O. & Schatz, M.C., 2013. The advantages of SMRT sequencing. Genome biology, 14(7), p.405.

Sachidanandam, R. et al., 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature, 409(6822), pp.928–933.

Sebat, J. et al., 2007. Strong association of de novo copy number mutations with autism. Science, 316(5823), pp.445–449.

Sebat, J., Levy, D.L. & McCarthy, S.E., 2009. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. Trends in genetics: TIG, 25(12), pp.528–535.

Sekar, A. et al., 2016. Schizophrenia risk from complex variation of complement component 4. Nature, 530(7589), pp.177–183.

Shen, H. et al., 2013. Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. PloS one, 8(4), p.e59494.

Shi, L. et al., 2016. Long-read sequencing and de novo assembly of a Chinese genome. Nature communications, 7, p.12065.

Shi, Y.Y. et al., 2008. A study of rare structural variants in schizophrenia patients and normal controls from Chinese Han population. Molecular psychiatry, 13(10), pp.911–913.

Sindi, S.S. et al., 2012. An integrative probabilistic model for identification of structural variation in sequencing data. Genome biology, 13(3), p.R22.

Sjödin, P. & Jakobsson, M., 2012. Population genetic nature of copy number variation. Methods in molecular biology , 838, pp.209–223.

Stankiewicz, P. & Lupski, J.R., 2010. Structural variation in the human genome and its role in disease. Annual review of medicine, 61, pp.437–455.

Stefansson, H. et al., 2008. Large recurrent microdeletions associated with schizophrenia. Nature, 455(7210), pp.232–236.

Steinberg, K.M. et al., 2014. Single haplotype assembly of the human genome from a hydatidiform mole. Genome research, 24(12), pp.2066–2076.

Stephens, P.J. et al., 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell, 144(1), pp.27–40.

Sudmant, P.H. et al., 2015a. An integrated map of structural variation in 2,504 human genomes. Nature, 526(7571), pp.75–81.

Sudmant, P.H. et al., 2015b. An integrated map of structural variation in 2,504 human genomes. Nature, 526(7571), pp.75–81.

Tang, Y.-C. & Amon, A., 2013. Gene copy-number alterations: a cost-benefit analysis. Cell, 152(3), pp.394–405.

Teshima, K.M. & Innan, H., 2012. The Coalescent with Selection on Copy Number Variants. Genetics, 190(3), pp.1077–1086.

The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. Nature, 467(7319), pp.1061–1073.

Trappe, K. et al., 2014. Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. Bioinformatics , 30(24), pp.3484–3490.

Travers, K.J. et al., 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic acids research, 38(15), p.e159.

Tuzun, E. et al., 2005. Fine-scale structural variation of the human genome. Nature genetics, 37(7), pp.727–732.

Usher, C.L. & McCarroll, S.A., 2015. Complex and multi-allelic copy number variation in human disease. Briefings in functional genomics, 14(5), pp.329–338.

Weber, J.L. et al., 2002. Human diallelic insertion/deletion polymorphisms. American journal of human genetics, 71(4), pp.854–862.

Weischenfeldt, J. et al., 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. Nature reviews. Genetics, 14(2), pp.125–138.

Wellcome Trust Case Control Consortium et al., 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature, 464(7289), pp.713–720.

Ye, K. et al., 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics , 25(21), pp.2865–2871.

Yoon, S. et al., 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome research, 19(9), pp.1586–1592.

Zarrei, M. et al., 2015. A copy number variation map of the human genome. Nature reviews. Genetics, 16(3), pp.172–183.

Zhang, F. et al., 2009. Copy number variation in human health, disease, and evolution. Annual review of genomics and human genetics, 10, pp.451–481.

Zhang, H. et al., 2013. Gene copy-number variation in haploid and diploid strains of the yeast Saccharomyces cerevisiae. Genetics, 193(3), pp.785–801.

Zhang, J., Wang, J. & Wu, Y., 2012. An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. BMC bioinformatics, 13 Suppl 6, p.S6.

Zhang, Z.D. et al., 2011. Identification of genomic indels and structural variations using split reads. BMC genomics, 12, p.375.

Zhao, M. et al., 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC bioinformatics, 14 Suppl 11, p.S1.

Zhao, X. et al., 2016. Resolving complex structural genomic rearrangements using a randomized approach. Genome biology, 17(1), p.126.

Zhu, M. et al., 2012. Using ERDS to infer copy-number variants in high-coverage genomes. American journal of human genetics, 91(3), pp.408–421.