Identifying structural variants using linked-read sequencing data

Rebecca Elyanow[2], Hsin-Ta Wu[2], and Benjamin J. Raphael[1,*]

[1]Department of Computer Science, Princeton University, Princeton, NJ 08540.

[2]Center for Computational Molecular Biology, Brown University, Providence, RI 02912.

[*]Correspondence: braphael@princeton.edu

Abstract

Structural variation, including large deletions, duplications, inversions, translocations, and other rearrangements, is common in human and cancer genomes. A number of methods have been developed to identify structural variants from Illumina short-read sequencing data. However, reliable identification of structural variants remains challenging because many variants have breakpoints in repetitive regions of the genome and thus are difficult to identify with short reads. The recently developed linked-read sequencing technology from 10X Genomics combines a novel barcoding strategy with Illumina sequencing. This technology labels all reads that originate from a small number (˜5-10) DNA molecules ˜50Kbp in length with the same molecular barcode. These barcoded reads contain long-range sequence information that is advantageous for identification of structural variants. We present Novel Adjacency Identification with Barcoded Reads (NAIBR), an algorithm to identify structural variants in linked-read sequencing data. NAIBR predicts novel adjacencies in a individual genome resulting from structural variants using a probabilistic model that combines multiple signals in barcoded reads. We show that NAIBR outperforms several existing methods for structural variant identification – including two recent methods that also analyze linked-reads – on simulated sequencing data and 10X whole-genome sequencing data from the NA12878 human genome and the HCC1954 breast cancer cell line. Several of the novel somatic structural variants identified in HCC1954 overlap known cancer genes.

1 Introduction

Recent whole genome sequencing (WGS) analysis of human genomes has shown that *structural variation*, including insertions, deletions, duplications, and rearrangements of genomic segments greater than 50bp, are a key component of human variation (Sudmant *et al.*, 2015). Collectively, structural variants affect a larger portion of the human genome than single nucleotide variants (Pang *et al.*, 2010). Inherited germline structural variants have been implicated in several diseases including Crohn's disease, rheumatoid arthritis, Type I diabetes, and autism (Mccarroll *et al.*, 2009; Weisenfeld *et al.*, 2016; Sebat *et al.*, 2007). In addition, somatic structural variants are common in cancer genomes (Chittenden *et al.*, 2008; Rausch *et al.*, 2013). These include deletions of tumor suppressor genes and amplifications of oncogenes which can promote aggressive cell growth and drive the development of cancer. Cancer genomes can also

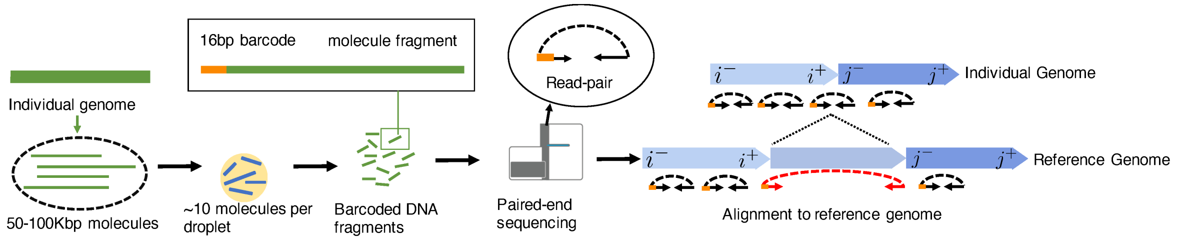

Figure 1: (Left) Linked-read sequencing with the 10X Genomics Chromium platform begins by fragmenting the individual genome into large DNA molecules, which are isolated into individuals beads that contain several large molecules and sequencing reagents. Within the bead, molecules are sheared into smaller fragments (500 bp) and labeled with a 16bp barcode indicating its bead of origin. Illumina paired-read sequencing of each fragment results in barcoded paired-end reads. (Right) Alignment of read-pairs to a reference genome results in concordant reads (black) and discordant reads (red). Discordant reads indicate candidate novel adjacencies that are a result of structural variants that distinguish individual genomes from the reference genome.

undergo dramatic rearrangement events such as chromothripsis, the shattering and random repair of chromosomes in a single catastrophic event (Stephens *et al.*, 2011), or chromoplexy (Baca *et al.*, 2013), both of which result in a large number of complex structural variants in a cancer genome.

The identification of structural variants from high-throughput DNA sequencing data is generally more challenging than the identification of single nucleotide variants. This difficulty is primarily a result of the the fact that many structural variants are significantly longer than the DNA sequence reads produced by current (second generation) DNA sequencing technologies, whose read lengths are ˜300-500 nucleotides. In addition, such reads are too short for *de novo* genome assembly. Thus structural variants are inferred from atypical, or aberrant, alignments of reads to a reference genome.

Numerous methods have been developed over the past several years to identify different types of structural variants from read alignments. Each of these methods use some combination of three signals that can be extracted from read alignments: *discordant read-pairs*, *split-reads*, and *read depth* (Figure S12a). A discordant read-pair is a pair of reads from the same fragment/insert whose alignments to the reference genome have distance and/or orientation that differ from expected if the entire fragment was contiguous on the reference genome. A split read is a read with no contiguous alignment to the reference genome, but rather with at least two partial alignments to the reference. (In practice, only a single partial, or clipped, alignment may be reported by the read alignment software.) Discordant read pairs and split reads are signatures of a *novel adjacency* in the individual genome; that is, two intervals that are non-adjacent in the reference genome are adjacent in an individual genome. Methods that rely on discordant paired reads and/or split reads include BreakDancer (Chen *et al.*, 2009), GASV (Sindi *et al.*, 2009a), VariationHunter (Hormozdiari *et al.*, 2010), Pindel (Ye *et al.*, 2009), DELLY (Rausch *et al.*, 2012), and LUMPY (Layer *et al.*, 2014); many others are reviewed in Tattini *et al.* (2015). Read depth is the (normalized) number of reads that map to a particular region of the genome. Read depth can be used to identify copy-number aberrations, such as deletions and duplications, by identifying regions of unexpectedly low or high coverage in the genome. Examples of read depth methods include BIC-Seq (Xi *et al.*, 2010), and CNVnator (Abyzov *et al.*, 2011). In addition, methods such as GASVPro (Sindi *et al.*, 2012) and SV-Bay (Iakovishina *et al.*, 2016) combine signals from discordant read-pairs and read depth signals to identify structural variants. Local assembly approaches such as SvABA (Wala *et al.*, 2017), and novoBreak (Chong *et al.*, 2017) achieve nucleotide level resolution of novel adjacencies. However, local assembly approaches require the identification of candidate regions for assembly using the signals described above; thus, local assembly generally increases specificity more than sensitivity. In addition, assembly-based approaches requires high coverage and is typically more time consuming than read-pair or read depth based methods (Alkan *et al.*, 2011).

The fundamental limitations in structural variant detection and whole-genome assembly are the fact that the human genome is diploid and highly repetitive (Mak *et al.*, 2016). Short reads have low signal to assign variants to haplotypes and to identify structural variants whose breakpoints lie in repetitive sequences. While algorithms can help extract information from this low signal, longer reads provide stronger signal. New *3rd-generation* sequencing technologies developed by Pacific Biosciences and Oxford Nanopore produce much longer reads (exceeding 10Kbp). However, these technologies are practically limited by their high per-base error rate and high cost compared to Illumina short-read sequencing. Additionally, structural variants that are larger than the average read size or that fall in repetitive regions still remain difficult to identify using these technologies (Weischenfeldt *et al.*, 2013). An alternative sequencing technology called *linked-read sequencing* was recently developed by 10X Genomics, and commercialized in their Chromium platform. In this technology, long DNA molecules, $50 - 100$Kbp in size are partitioned into one of several million droplets using microfluidics. Each droplet contains a small number (˜10) of molecules (Weisenfeld *et al.*, 2016). The molecules in each droplet are sheared into smaller fragments and labeled with a 16bp molecular barcode that is unique to each droplet. The fragments are then amplified and sequenced using Illumina paired-end sequencing protocol (Figure 1). 10X Genomics' linked-read technology thus provides both the low error rate and low cost of Illumina sequencing as well as long-range sequencing information provided by 3rd-generation sequencing technologies (Figure S12b). The technology is similar in some respects to the strobe sequencing technology that was prototyped by Pacific BioSciences but never commercially released (Ritz *et al.*, 2010, 2014).

Here, we introduce Novel Adjacency Identification with Barcoded Reads (NAIBR, pronounced ”neighbor”), a method that identifies novel adjacencies resulting from structural variants in an individual genome from linked-read sequencing data. NAIBR combines a novel split-read type signal from linked-reads with traditional signals of structural variants in the underlying paired-reads in the data. We demonstrate that NAIBR outperforms existing structural variant detection algorithms – both paired-read methods and two recently developed methods (Zheng *et al.*, 2016; Spies *et al.*, 2016) for linked-reads – using simulated and real linked-read sequencing data. NAIBR also leverages haplotype phasing information from linked-reads, improving the detection of heterozygous structural variants.

2 Methods

Consider two genomes, a *reference genome* and an *individual genome*, each represented by an interval, $G = [1, n]$ and $G' = [1, n']$ respectively. We let $[i^-, i^+]$ denote an interval in the genome, where i^- and i^+ indicate the start and end of the interval respectively. We consider a structural variant to be any difference between an individual genome and a reference genome due to DNA breakage and repair, that results in the joining of two non-adjacent intervals $[i^+, i^-]$ and $[j^+, j^-]$ in the reference genome. The ends of these intervals may be joined in one of four orientations, and we indicate the four possible *novel adjacencies* in the individual genome by the pairs of joined ends: (i^+, j^-),(i^+, j^+),(i^-, j^-), or (i^-, j^+). Note that novel adjacencies are formed by most of the usual structural variants (segmental deletions/insertions, inversions, and translocations), with the notable exception of the deletion of a chromosome to the telomere.

NAIBR aims to identify such novel adjacencies using linked-read sequencing data. Our algorithm differs from previously published methods in that it incorporates signals from both paired-end reads and linked-reads into a unified model. Before defining the model, we will describe the signals we observe in paired-end and linked-read data and how these signals are combined to identify novel adjacencies arising from structural variants in an individual genome.

2.1 Paired-end sequencing data

In Illumina paired-end sequencing, chromosomes are sheared into small fragments and size selected such that most fragment lengths are within the interval $[l_{\min}, l_{\max}]$. Each fragment is sequenced from both ends from opposite DNA strands; thus one read will originate from the forward (+) strand and one from the reverse (-) strand. Paired reads are aligned to the reference genome. Each aligned read, x, is represented by a tuple $x = (l_x, r_x, o_x, q_x)$, where l_x is the leftmost position of x in the reference genome, r_x is the rightmost position in the reference genome, $o_x \in$ {+,-} is the orientation of x, and q_x is the mapping probability of x. We define a read-pair to be the ordered pair $\langle x, y\rangle$, where read x has the smaller starting coordinate. A read-pair $\langle x, y\rangle$ is *concordant* provided the distance between aligned reads $f = r_y - l_x$ isbetween $l_{\min}$ and $l_{\max}$ and the orientations are $o_x = +, o_y = -$. Concordant reads are consistent with the fragment aligning contiguously to the reference genome with no rearrangement. A read pair $\langle x, y\rangle$ not satisfying this condition is *discordant*. Discordant read-pairs arise from either (1) errors in sequencing and/or alignment or (2) novel adjacencies in an individual genome with respect to the reference.

2.2 Linked-read sequencing data

The linked-read sequencing technology developed 10X Genomics adds a layer of structure to Illumina paired-end sequencing by tagging each DNA fragment with a barcode prior to sequencing (Figure 1). Barcoded read-pairs are aligned to the reference using a linked-read aware aligner called Lariat (Bishara *et al.*, 2015). Lariat processes all read-pairs from a single barcode simultaneously, using the knowledge that reads originate from a small number of long molecules. Using this prior knowledge, it finds more unique mappings than other tools and can map to highly repetitive regions.

Read-pairs $p = \langle x, y\rangle$ originating from the same long molecule will each be tagged with the same barcode $b_p \in \mathbb{N}$. Because molecules are partitioned into droplets uniformly at random, the likelihood of assigning the same barcode to two molecules from nearby locations on the reference genome is low. Thus, we assume that read-pairs with the same barcode that map near each other on an individual genome are likely to have originated from the same long molecule. We partition such read-pairs into sets called *linked-reads*.

A linked-read is a set of concordant read-pairs that share the same barcode and have a maximum distance of δ from another read-pair in the linked-read. Each linked-read is assumed to have originated from a contiguous strand of DNA in the reference genome. The set of all concordant read-pairs in the genome is partitioned into linked-reads such that any two read-pairs $p = \langle x, y\rangle$ and $p' = \langle x', y'\rangle$, where $l_x < l'_x$, are both partitioned into the same linked-read if $b_p = b_{p'}$, and $l_{x'} - r_y \leq \delta$ (see Figure 2a).

Any pair of linked-reads sharing a barcode may have originated from a molecule that is split with respect to the reference genome due to the presence of a novel adjacency in an individual genome. We define a *candidate split*

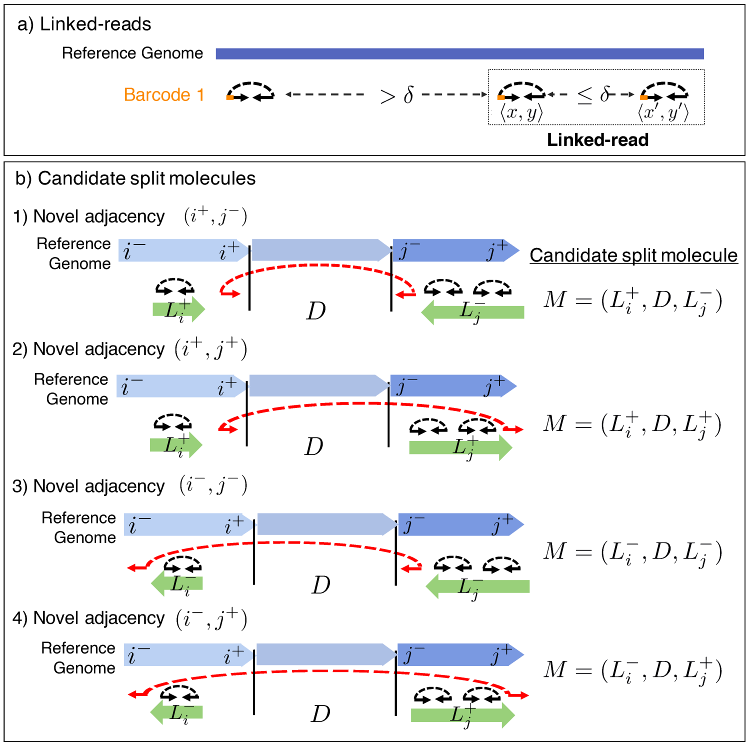

Figure 2: (a) A linked-read is defined by read-pairs separated by a distance $\leq \delta$ on the reference genome. (b) Linked-reads L_i and L_j may have originated from one of 4 candidate split molecules, each supporting a novel adjacency with a different orientation. $M = (L_i^+, D, L_j^-)$ supports a novel adjacency (i^+, j^-) and indicates that the end of linked-read L_i is adjacent to the start of linked-read L_j (the arrows points to the location of the novel adjacency).

molecule as a tuple $M = (L_1, D, L_2)$, where L_1 and L_2 are linked-reads and D is a set of discordant read pairs, satisfying the following conditions: (1) All read pairs in L_1, L_2, and D share the same barcode. (2) All discordant read pairs in D have the same orientations, and the distance between any two discordant read-pairs in D is at most $l_{\max}$. Formally, for read pairs $p = \langle x, y \rangle, p' = \langle x', y' \rangle \in D$, $|x - x'| < l_{\max}$ and $|y - y'| < l_{\max}$. (3) The linked reads are located within δ of the discordant read-pairs in D, in the direction consistent with the orientation of D. To indicate the last condition, we assign an orientation to each linked read in M. For example, $M = (L_1^+, D, L_2^-)$ is a candidate split molecule provided: the *rightmost* position, $r_{L_1} = \max\{r_y \mid \langle x, y \rangle \in L_1^+\}$, of L_1 is within δ of discordant read-pairs in D and the *leftmost* position, $l_{L_2} = \min\{l_x \mid \langle x, y \rangle \in L_2^-\}$, of L_2 is within δ of discordant read-pairs in D. See Figure 2). Note that we also define a candidate split molecule in the case where D is the empty set. In this case, a candidate split molecule can be formed for any of the four possible orientations of linked-reads L_1 and L_2: $(L_1^+, \emptyset, L_2^-), (L_1^+, \emptyset, L_2^+), (L_1^-, \emptyset, L_2^-), (L_1^-, \emptyset, L_2^+)$.

We say that a candidate split molecule M *supports* a novel adjacency provided that the distances and orientations of the linked and discordant reads in M are consistent with the novel adjacency. For example, candidate split molecule $M = (L_1^+, D, L_2^-)$ supports the novel adjacency (i^+, j^-) provided: (1) The orientation of the candidate split molecule matches the orientation of the novel adjacency. (2) Each read-pair $p = \langle x, y \rangle \in D$, is at most $l_{\max}$ from breakends i and j: $i - l_{\max} \leq r_x \leq i$ and $j \leq l_y \leq j + l_{\max}$. (3) Oriented linked-reads L_i^+ and L_j^- are at most a distance δ from positions i and j respectively: $i - \delta \leq r_{L_1} \leq i$ and $j \leq l_{L_2} \leq j + \delta$.

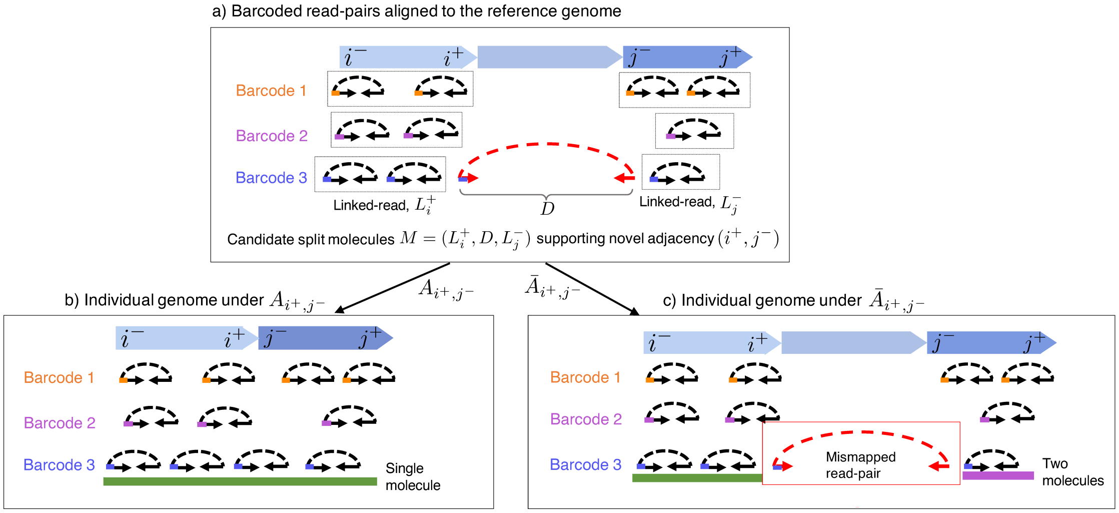
a) Barcoded read-pairs aligned to the reference genome
i^-
i^+
j^-
j^+
Barcode 1
Barcode 2
Barcode 3
Linked-read, L_i^+
D
Linked-read, L_j^-
Candidate split molecules $M = (L_i^+, D, L_j^-)$ supporting novel adjacency (i^+, j^-)
A_{i^+,j^-}
$\bar{A}_{i^+,j^-}$
b) Individual genome under A_{i^+,j^-}
c) Individual genome under $\bar{A}_{i^+,j^-}$
Single molecule
Mismapped read-pair
Two molecules

Figure 3: (a) A candidate split molecule for novel adjacency (i^+, j^-) consists of a linked-read L_i^+ a linked-read L_j^-, within a distance δ of position j and a set of discordant reads D. Barcoded reads aligned to the reference genome may originate from an individual genome that either contains a novel adjacency (A_{i^+,j^-}) or does not contain a novel adjacency ($\bar{A}_{i^+,j^-}$). (b) Under the alternative hypothesis A_{i^+,j^-} that i^+ and j^- are adjacent in an individual genome, reads in barcode 3 are close and are likely to have originated from a single molecule. (c) Under the null hypothesis $\bar{A}_{i^+,j^-}$ that i and j are non-adjacent in an individual genome, reads in barcode 3 are separated by a large distance and are likely to have originated from two molecules. Barcode 2 contains a read that is discordant under $\bar{A}_{i^+,j^-}$ and therefore assumed to be mismapped.

According to the definitions above, for a given barcode and novel adjacency (i^+, j^-), there is at most one candidate split molecule that supports this novel adjacency. We define the set $\mathscr{M}$ to be the set of all candidate split molecules supporting a novel adjacency (i^+, j^-). For ease of exposition, we will describe the model below for a novel adjacency of the form (i^+, j^-), but the model may be applied to novel adjacencies with any of the four orientations.

2.3 Likelihood ratio score

We use a likelihood ratio score to evaluate the evidence support a novel adjacency. Given a potential novel adjacency (i^+, j^-), let A_{i^+,j^-} be the event of a novel adjacency (i^+, j^-) in an individual genome and let event $\bar{A}_{i^+,j^-}$ to be the absence of this novel adjacency. Let $\mathscr{M}$ be the set of all candidate split molecules supporting novel adjacency (i^+, j^-). We compare the likelihood $P(\mathscr{M} \mid A_{i^+,j^-})$ of A_{i^+,j^-} and the likelihood $P(\mathscr{M} \mid \bar{A}_{i^+,j^-}$ of $\bar{A}_{i^+,j^-}$ given the set of observed candidate split molecules (Figure 3) using the log-likelihood ratio

$$\Lambda_{i^+,j^-} = \log \frac{P(\mathscr{M} \mid A_{i^+,j^-})}{P(\mathscr{M} \mid \bar{A}_{i^+,j^-})}. \tag{1}$$

We report novel adjacencies with log-likelihood ratio, $\Lambda_{i^+,j^-} > c$ (selection of c is described in the Supplement) as the set of predicted novel adjacencies in the individual genome.

We now describe how we compute each of the terms in the log-likelihood ratio. First, we assume that the locations of molecules with different barcodes are independent. Combining this assumption with the requirement that candidate split molecules with different barcodes must originate from different molecules, we have that $P(\mathscr{M} \mid A_{i+,j-})$ and $P(\mathscr{M} \mid A_{i+,j-})$ are each the product of the individual probabilities of observing each candidate split molecule $M = (L_i^+, D, L_j^-) \in \mathscr{M}$. Next, a candidate split molecule M may have either originated from a single molecule that spans the interval $[i^+, j^-]$ or from two molecules, each of which does not span the interval $[i^+, j^-]$. Let E_M be the event that candidate split molecule M originates from a single molecule and let $E_{L_i^+}$ and $E_{L_j^-}$ be the events that linked-reads L_i^+ and L_j^- originate from unique molecules. Thus, the probability of observing candidate split molecule M is the probability of two disjoint events: that M originates from one molecule (E_M), or that L_i^+ and L_j^- originate from two different molecules ($E_{L_i^+} \cap E_{L_j^-}$). Thus, the log-likelihood ratio $\Lambda_{i+,j-}$ is calculated as follows,

$$\Lambda_{i+,j-} = \frac{\prod\limits_{M \in \mathscr{M}} \left[P(E_M \mid A_{i+,j-}) + P(E_{L_i^+} \cap E_{L_j^-} \mid A_{i+,j-})\right]}{\prod\limits_{M \in \mathscr{M}} \left[P(E_M \mid \bar{A}_{i+,j-}) + P(E_{L_i^+} \cap E_{L_j^-} \mid \bar{A}_{i+,j-})\right]}.$$

We now describe how we calculate the probabilities of E_M and $E_{L_i^+} \cap E_{L_j^-}$ given the the events $A_{i+,j-}$ and $\bar{A}_{i+,j-}$. Given a candidate split molecule $M = (L_i^+, D, L_j^-)$ supporting a novel adjacency (i^+, j^-), $P(E_M \mid A_{i+,j-})$ is the probability that M was sequenced from a single molecule when i^+ and j^- are adjacent in the individual genome; correspondingly, $P(E_M \mid \bar{A}_{i+,j-})$ is the probability that M was sequenced from a single molecule when i^+ and j^- are not adjacent in the individual genome. We model the sequencing of a molecule as the generation of mapped reads through three sequential processes: (1) molecule size selection, (2) molecule sequencing, and (3) read mapping. We assume that the fragmentation of chromosomes into long molecules and the sequencing of reads from a molecule are independent processes. This assumption of independence is reasonable because sequencing occurs after molecules are sheared into short fragments.

Thus, we model the probability $P(E_M \mid A_{i+,j-})$ as,

$$\begin{aligned} P(E_M \mid A_{i+,j-}) &= P(S(M) = s) \cdot P(R(M) = \rho) \cdot P(Q_M), \\ \text{where } S(M) &= \text{the size of molecule } M, \\ R(M) &= \text{the sequencing rate of molecule } M, \text{ and} \\ Q_M &= \text{the event that all reads in } M \text{ are correctly mapped,} \end{aligned}$$

and the probability $P(E_M \mid \bar{A}_{i+,j-})$ as,

$$\begin{aligned} P(E_M \mid \bar{A}_{i+,j-}) &= P(S(M) = \bar{s}) \cdot P(R(M) = \bar{\rho}) \cdot P(Q_{M \setminus D}) \\ \text{where } Q_{M \setminus D} &= \text{the event that reads in } M \setminus D \text{ are correctly} \\ &\quad \text{mapped and reads in } D \text{ are mismapped.} \end{aligned}$$

We model the size $S(M)$ of a molecule M by a negative binomial distribution, $P(S(M) = \cdot) = NB(\cdot)$, with parameters estimated from the collection of all aligned linked reads. We assume that the vast majority of linked-reads in the data originate from molecules that are contiguous with respect to the reference genome. Formally, the size of a linked-read L is $r_L - l_L$, where $r_L = \max\{r_y \mid \langle x, y \rangle \in L\}$ and $l_L = \min\{l_x \mid \langle x, y \rangle \in L\}$ are the rightmost and leftmost positions, respectively, of paired reads in L. This approximation tends to slightly underestimate the true length of a molecule due to missing reads at the ends of the molecule. Under $A_{i+,j-}$, the size of candidate split molecule M is approximately $s = (i - l_{L_i}) + (r_{L_j} - j)$, the sum of the portions of the molecule aligning to the left and right of the novel adjacency. Similarly, under $\bar{A}_{i+,j-}$, the size of M is approximated by $\bar{s} = r_{L_j} - l_{L_i}$, the distance between the leftmost position in L_i and the rightmost position in L_j. For the case where i^+ and j^- are on different chromosomes, $P(S(M) = \bar{s}) = 0$.

We model the sequencing rate $R(M)$ of a molecule M by a Gamma distribution, $P(R(M) = \cdot) = \Gamma(\cdot)$. The negative binomial and Gamma distributions provide a good fit to the empirical distributions of molecule size and sequencing rate respectively (Figure S2), but other distributions can be used (Table S1). The sequencing rate of a candidate split molecule M under A_{i^+,j^-} is approximated by $\rho = \frac{|L_i^+ \cup D \cup L_j^-|}{s}$, the number of reads sequenced from M normalized by the size s of M. The sequencing rate of M under $\bar{A}_{i^+,j^-}$ is approximated by $\bar{\rho} = \frac{|L_i^+ \cup L_j^-|}{\bar{s}}$. Under $\bar{A}_{i^+,j^-}$, we exclude D from the set of reads sequenced from M because these reads are assumed to be mismapped.

We now define the probabilities $P(Q_M)$ and $P(Q_{M \setminus D})$ of mapping reads to the reference genome. Under A_{i^+,j^-}, all reads in M are correctly mapped, an event with probability

$$P(Q_M) \approx \prod_{\langle x,y \rangle \in L_i^+ \cup D \cup L_j^-} (\frac{q_x + q_y}{2}),$$

where q_x and q_y are the mapping probabilities of reads x and y obtained from the alignment software. We chose to approximate the mapping probability of the read-pair $p = \langle x, y \rangle$ to be the average of the mapping probabilities of each read, which we found to perform well on data. Under $\bar{A}_{i^+,j^-}$, concordant reads in L_i^+ and L_j^- are correctly mapped and discordant reads in D are mismapped, an event with probability,

$$P(Q_{M \setminus D}) \approx \prod_{\langle x,y \rangle \in D} (1 - \frac{q_x + q_y}{2}) \cdot \prod_{\langle x,y \rangle \in L_i^+ \cup D \cup L_j^-} (\frac{q_x + q_y}{2}).$$

We calculate the probability $P(E_{L_i^+} \cap E_{L_j^-}) \mid A_{i^+,j^-})$ that L_i^+ and L_j^- were sequenced from two different molecules as the product of the probabilities that L_i^+ and L_j^- were sequenced from independent molecules with corresponding sizes s_i, s_j and sequencing rates p_i, p_j, and that the reads from the two molecules were then properly mapped to the reference genome, the latter event denoted by event Q_M. Formally,

$$\begin{aligned} P(E_{L_i^+} \cap E_{L_j^-}) \, |A_{i^+,j^-}) = & \\ & P(S(L_i) = s_i) \cdot P(R(L_i) = \rho_i) \cdot \\ & P(S(L_j) = s_j) \cdot P(R(L_j) = \rho_j) \cdot P(Q_M), \end{aligned}$$

where $s_k = r_{L_k} - l_{L_k}$ and $\rho_k = \frac{|L_k|}{s_k}$.

Under $\bar{A}_{i^+,j^-}$, the probabilities of molecule size selection and sequencing remain the same. However any discordant reads in D will be mismapped under $\bar{A}_{i^+,j^-}$, resulting in the last term being the probability $P(Q_{M \setminus D})$.

2.4 Incorporating haplotype phase

10X Genomics provides phasing as part of its alignment pipeline, using linked-reads to phase SNPs into large phase blocks. Each position i in the reference genome is assigned a phase block m_i and any SNP within phase block m_i will be assigned to either haplotype 1 or haplotype 2. Zheng *et al.* (2016) report that the current software can phase $> 95\%$ of SNPs to a phase blocks of size > 0.5Mb, with an average error rate of 0.03%. Let $m_x \in \mathbb{N}$ denote the phase block of a read x, aligned to the reference genome, and let $h_x \in \{1, 2\}$ denote the haplotype of read x. We define $\mathscr{M}^{\alpha,\beta} \subseteq \mathscr{M}$ be the subset of candidate split molecules $M = (L_i^+, D, L_j^-)$ such that the haplotype h_{L_i} of linked-read L_i is α, the haplotype h_{L_j} of linked-read L_j is β, and for all read pairs $\langle x, y \rangle$ in D, $h_x = \alpha$ and $h_y = \beta$.

On the linked-read datasets described in the Results below, we found that $> 99\%$ of linked-reads ($\delta = 10\text{Kbp}$) contain read-pairs that are assigned to a unique phase block and haplotype. We omit the small number of linked-reads containing read-pairs from multiple haplotypes or phase blocks from further analysis.

We define a haplotype-specific log-likelihood ratio,

$$\Lambda^{\alpha,\beta}_{i+,j-} = \log \frac{P(\mathscr{M}^{\alpha,\beta} \mid A_{i+,j-})}{P(\mathscr{M}^{\alpha,\beta} \mid \bar{A}_{i+,j-})}. \tag{2}$$

Each novel adjacency in a diploid genome is the result a heterozygous structural variant on haplotype 1, a heterozygous structural variant on haplotype 2, or a homozygous structural variant affecting both haplotypes. If positions i and j are on the same phase block $m_i = m_j$, then the haplotypes of i and j must match. Thus, we calculate the log-likelihood ratio $\Lambda^h_{i+,j-}$ to be the maximum of $\Lambda^{1,1}_{i+,j-}$, $\Lambda^{2,2}_{i+,j-}$, and $\Lambda_{i+,j-}$, corresponding to a novel adjacency resulting from a heterozygous structural variant on haplotype 1, a heterozygous structural variant on haplotype 2, or a homozygous structural variant,

$$\Lambda^h_{i+,j-} = \max(\Lambda^{1,1}_{i+,j-}, \Lambda^{2,2}_{i+,j-}, \Lambda_{i+,j-}). \tag{3}$$

However, if positions i and j are not on the same phase block, $m_i \neq m_j$, then the haplotypes of i and j may not match. For example, SNPs assigned to haplotype 1 of phase block m_i may originate from the same chromosome as SNPs assigned haplotype 2 of phase block m_j. In this case, we calculate the log-likelihood ratio $\Lambda^h_{i+,j-}$ as,

$$\Lambda^h_{i+,j-} = \max(\Lambda^{1,1}_{i+,j-}, \Lambda^{2,2}_{i+,j-}, \Lambda^{1,2}_{i+,j-}, \Lambda^{2,1}_{i+,j-} \Lambda_{i+,j-}), \tag{4}$$

which accounts for the ambiguity in haplotype assignment. NAIBR reports both the phased log-likelihood ratio $\Lambda^h_{i+,j-}$ as well as the inferred haplotype for each novel adjacency.

3 Results

We assess NAIBR's ability to detect novel adjacencies in simulated and real 10X long-read sequencing data and benchmark against 5 other methods: Long Ranger (Zheng *et al.*, 2016), GROC-SVs (Spies *et al.*, 2016), GASV (Sindi *et al.*, 2009b), GASVPro (Sindi *et al.*, 2012), and LUMPY (Layer *et al.*, 2014). We chose these methods for comparison because they utilize different combinations of signals to identify and rank novel adjacencies. Long Ranger is 10X Genomics' structural variant detection program. Long Ranger identifies novel adjacencies by computing overlapping pairs of linked-reads and computes a likelihood score based on the number of overlaps observed in the data. GROC-SVs is also designed for linked-read sequencing data. GROC-SVs identifies structural variants by performing local assembly on barcoded reads. GROC-SVs operates in two steps. First it assigns p-value to novel adjacencies, and then it performs local assembly, labelling novel adjacencies as assembled or unassembled. In some cases, unassembled variants have smaller p-values than assembled variants. GASV, GASVPro, and LUMPY analyze paired-end sequencing data. When running these algorithms on 10X Genomics data we ignore barcodes and treat the data as Illumina paired-end sequencing data. GASV uses only discordant read-pairs and ranks novel adjacencies by the number of supporting discordant read-pairs. GASVPro uses a combination of discordant read-pairs and breakend read depth to assign a log-likelihood score to each predicted novel adjacency. LUMPY uses discordant and split reads to call novel adjacencies and reports a p-value for each adjacency.

We benchmark NAIBR against each method on both simulated and real data. Reported novel adjacencies are ranked from highest to lowest confidence according to the metrics used by each method. We run each method using its default parameters. More specifics can be found in the Supplement.

We simulate several types of structural variants – including duplications, deletions, translocations, and inversions – on chromosomes 17 and 18 of the human reference genome hg19. To assess NAIBR's ability to detect novel adjacencies that occur on a single haplotype, we simulate two test genomes, one that contains 400 homozygous structural variants and one that contains 400 different structural variants on each haplotype. Translocations and inversions create more than 1 novel adjacency in the simulated genome, resulting in 508 homozygous novel adjacencies in the first simulated genome and 1027 heterozygous novel adjacencies in the second simulated genome. We simulate linked-read sequencing to 30X coverage. Details on simulation can be found in the Supplement. Figure 4a shows the precision-recall curve for all 5 methods run on the 30X test dataset containing 508 homozygous novel adjacencies. NAIBR has the highest recall of all methods, correctly identifying $479/508$ homozygous novel adjacencies. GASV correctly identified $309/508$ variants, however GASV reported several thousand variants, resulting in very low precision at high values of recall. LUMPY performed similarly to GASV, correctly identifying $308/508$ true variants, however the algorithm only reports variants with high probability scores according to their scoring metric, resulting in lower recall. GASVPro identified as many true variants as NAIBR at 50% recall, but only reported $289/508$ true variants in total, compared to 479 reported by NAIBR.

Long Ranger and GROC-SVs are each designed to utilize linked-reads, however both methods are limited to the identification of certain types of variants. Long Ranger reports variants larger than 30Kbp and GROC-SVs only reports variants larger than 10Kbp. The simulated dataset contains 369 novel adjacencies larger than 10Kbp and 269 structural variants larger than 30Kbp. Both Long Ranger and GROC-SVs perform with lower precision than NAIBR. For GROC-SVs, $38/39$ of assembled novel adjacencies were present in the truth set, showing that the local assembly approach has high precision. However an additional 59 true novel adjacencies were predicted by GROC-SVs but failed to assemble, indicating that local assembly removes many true positives. We perform the same comparison on a simulated dataset containing 800 heterozygous novel adjacencies with similar results. We also compare the runtime and memory usage of NAIBR to other methods and find that NAIBR outperforms other linked-read methods (see Supplement). These results demonstrate that NAIBR's incorporation of both linked-read and paired-end read data improves performance over other methods without significant additional time or memory requirements.

3.1 Benchmarking on NA12878

To assess NAIBR's ability to detect known variants from a real dataset, we obtained whole-genome sequencing data of individual NA12878 of the 1000 Genomes Project from 10X genomics (https://support.10xgenomics.com/genome-exome/datasets/
NA12878_WGS_210). The data was sequenced on the Chromium platform to 35X sequencing coverage with the Illumina Hiseq2500 to produce 845 million 98bp reads with mean insert size of 340bp and a mean molecule size of 68Kbp. We used the 2,950 validated novel adjacencies in NA12878 reported in Layer *et al.* (2014) as the truth set. Novel adjacencies in this dataset were validated by split-read mapping analysis of independent long-read sequencing data from PacBio or Illumina Moleculo platforms.

Figure 4b shows the precision-recall curves each method against 121 validated structural variants larger than

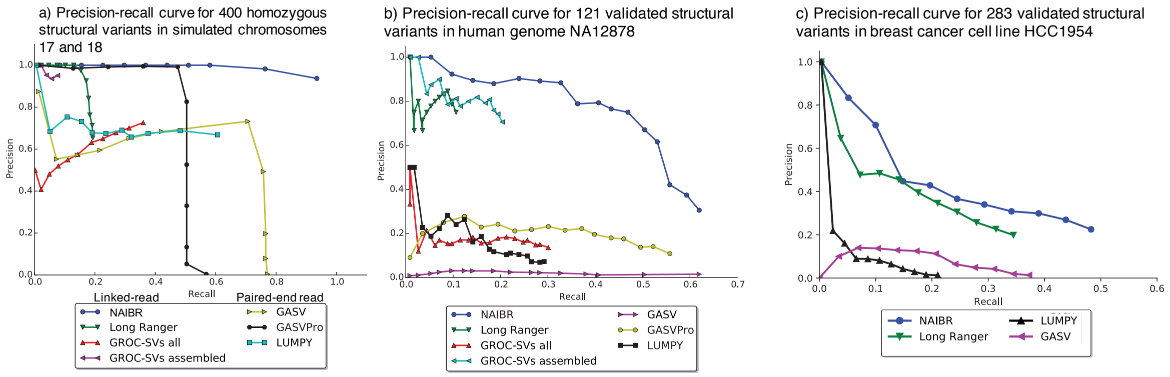

Figure 4: a) Precision-recall curve for NAIBR, Long Ranger, GROC-SVs, GASV, GASVPro, and LUMPY on 30X simulated data from chromosomes 17 and 18, containing 400 homozygous structural variants. b) Precision-recall curve for NAIBR, GASV, GASVPro, LUMPY, GROC-SVs, and Long Ranger evaluated against the set of 123 validated structural variants $>$ 7Kbp from NA12878 (Layer *et al.*, 2014). c) Precision-recall curve for NAIBR, GASV, GASVPro, LUMPY, GROC-SVs, and Long Ranger evaluated against the set of validated structural variants $\geq$ 30Kbp from breast cancer cell line HCC1954 (Bignell *et al.*, 2007; Stephens *et al.*, 2009; Galante *et al.*, 2011).

7Kbp. NAIBR correctly predicted $73/121$ novel adjacencies $>$ 7Kbp. GASV also correctly predicted 73 novel adjacencies but displayed poor precision. GROC-SVs reported variants that it was able to assemble using its local assembly pipeline as well as variants that were not assembled. The local assembly step of GROC-SVs drastically reduces the number of reported false positives but also fails to assemble several predictions matching the truth set. Long Ranger reported 17 novel adjacencies, correctly identifying $12/32$ variants larger than 30Kbp from the truth set, compared to $17/32$ variants larger than 30Kbp detected by NAIBR. Figure S9 shows the precision-recall curves for each of the 5 methods on different structural variant sizes, ranging from 50bp-30Kbp. NAIBR outperforms other linked-read methods at detecting variants between 50bp and 10Kbp and outperforms paired-end methods at detecting variants larger than 5Kbp. NAIBR outperforms both linked-read and paired-end methods at detecting variants between 1Kbp and 10Kbp. In summary, NAIBR detects large structural variants ($>$ 1Kbp) with better precision than paired-end callers GASV, GASVPro, and LUMPY and has higher recall and precision for variants smaller than 10Kbp compared to linked-read callers Long Ranger and GROC-SVs.

3.2 Tumor cell line HCC1954

We test NAIBR's ability to detect somatic structural variants in tumor cell line HCC1954T. The cell line was derived from a grade 3 invasive ductal carcinoma and sequenced by 10X Genomics to $35X$ coverage with a mean molecule size of 85Kbp. The matched normal HCC1954N was sequenced by 10X Genomics to $35X$ coverage with a mean molecule size of 88Kbp. We identify novel adjacencies in both HCC1954T and HCC1954N using 4 different methods: NAIBR, Long Ranger, GASV, and LUMPY. We formed a set of 369 true novel adjacencies by combining PCR-validated novel adjacencies from three previous studies: Bignell *et al.* (2007), Stephens *et al.* (2009), and Galante *et al.* (2011).

NAIBR identifies 142 PCR-validated novel adjacencies, significantly more than Long Ranger (100), GASV (117), and LUMPY (55) (Figure S6). NAIBR also demonstrates better precision at all levels of recall than other methods (Figure 4c). Notably, GASV has significantly lower precision than Long Ranger and NAIBR, predicting over ten times as many novel adjacencies with lower recall than the other methods (Figure S6). NAIBR significantly outperforms Long Ranger at identifying duplications and interchromosomal events (Figure S5), identifying 72 duplications compared to 50 identified by Long Ranger and 53 interchromosomal events compared to 37 predicted by Long Ranger. Over 30% of the breakends predicted by NAIBR lie within 50bp of the breakends of the PCR-validated novel adjacencies, compared to approximately 10% of breakends predicted by Long Ranger and GASV (Figure S7).

Several of the novel adjacencies in HCC1954T that were not identified by Long Ranger and GASV affect known oncogenes and tumor suppressors. For example, a novel adjacency between Chr11:93153935 and Chr11:93160223 occurs within the gene CCDC67, potentially leading to loss of function of the gene. CCDC67 has recently been identified as a tumor suppressor gene (Yin *et al.*, 2016; Zhu *et al.*, 2014). A novel adjacency between Chr14:89829140 and Chr7:155683934 affects the forkhead transcription factor checkpoint suppressor, CHES1. CHES1 expression has been shown to be reduced across many cancer types (Huot *et al.*, 2014). A novel adjacency between Chr11:69059340 and Chr11:69089741, surrounding the MYEOV gene (Chr:69061613-69064754), is potentially a result of a duplication of MYEOV. MYEOV has been shown to be amplified in breast cancer patients and has been identified as a candidate oncogene (Janssen *et al.*, 2002).

4 Discussion

We present NAIBR, a probabilistic algorithm for the identification of novel adjacencies using linked-read sequencing data. Linked-read sequencing combines low per-base error rate of short-read sequencing technologies with long-range linking information of long-read technologies. Linked-read sequencing offers drastically improved mapping and phasing results compared to paired-end sequencing (Bishara *et al.*, 2015) with similar cost, making it an attractive option for researchers. NAIBR is one of the first algorithms that identifies structural variation by using signals unique to linked-read sequencing data. NAIBR uses discordant read-pairs obtained from paired-end reads combined with candidate split molecule obtained from linked-reads to identify and rank novel adjacencies. NAIBR detects novel adjacencies with higher accuracy and precision than existing methods on both simulated and real linked-read sequencing data.

We also demonstrate NAIBR's ability to predict somatic novel adjacencies from cancer data by applying it to cell line HCC1954. Several novel adjacencies detected by NAIBR were not identified by other methods, including novel adjacencies affecting tumor suppressor genes CCDC67 and CHES1 and candidate oncogene MYEOV. While some of the novel adjacencies predicted by NAIBR might be verified PCR, it is possible that long-read sequencing data (e.g.from PacBio or Oxford Nanopore) would be required, since these adjacencies were not readily apparent from short-read sequencing data.

As future work, we plan to incorporat.epse additional signals, such as read depth and split-reads, into our probabilistic model for identifying novel adjacencies. Our algorithm can also be extended by performing local assembly on linked-reads supporting novel adjacencies to reconstruct structural variants, as is done by GROC-SVs (Spies *et al.*, 2016).

The utility of linked-reads in identifying novel adjacencies between nearby positions on the reference genome (such as a small deletion) is limited by the fact that each molecule is sequenced to low coverage, introducing large gaps between read-pairs sequenced from the same molecule. Thus, with the current Chromium technology from 10X Genomics, linked-reads provide substantial additional power to detect large structural variants but provide little additional power above that of paired-end sequencing in the detection of small structural variants. However, linked-read sequencing is a very new technology and is likely to improve in the coming years. As molecules are sequenced to higher coverage, NAIBR's ability to detect novel adjacencies using candidate split molecules will improve significantly, enabling the identification of novel adjacencies arising from smaller structural variants.

Funding

This work is supported by a US National Science Foundation (NSF) CAREER Award (CCF-1053753) and US National Institutes of Health (NIH) grants R01HG005690, R01HG007069, and R01CA180776 to BJR. BJR is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, an Alfred P. Sloan Research Fellowship.

NAIBR Software

Software is available at compbio.cs.brown.edu/software

5 Bibliography

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). Cnvnator: an approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. *Genome research*, **21**(6), 974–984.

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

Baca, S. C., Prandi, D., Lawrence, M. S., Mosquera, J. M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T. Y., Ghandi, M., *et al.* (2013). Punctuated evolution of prostate cancer genomes. *Cell*, **153**(3), 666–677.

Bignell, G. R., Santarius, T., Pole, J. C., Butler, A. P., Perry, J., Pleasance, E., Greenman, C., Menzies, A., Taylor, S., Edkins, S., *et al.* (2007). Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome research*, **17**(9), 1296–1303.

Bishara, A., Liu, Y., Weng, Z., Kashef-Haghighi, D., Newburger, D. E., West, R., Sidow, A., and Batzoglou, S. (2015). Read clouds uncover variation in complex regions of the human genome. *Genome research*, **25**(10), 1570–1580.

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L., and Mardis, E. R. (2009).

BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, **6**(9), 677–81.

Chittenden, T., Chittenden, T., Howe, E., Howe, E., a.C. Culhane, a.C. Culhane, Sultana, R., Sultana, R., Taylor, J., Taylor, J., Holmes, C., Holmes, C., Quackenbush, J., and Quackenbush, J. (2008). Functional classification analysis of somatically mutated genes in human breast and colorectal cancers. *Genomics*, **455**(7216), 1061–1068.

Chong, Z., Ruan, J., Gao, M., Zhou, W., Chen, T., Fan, X., Ding, L., Lee, A. Y., Boutros, P., Chen, J., *et al.* (2017). novobreak: local assembly for breakpoint detection in cancer genomes. *Nature methods*, **14**(1), 65–67.

Galante, P. A., Parmigiani, R. B., Zhao, Q., Caballero, O. L., De Souza, J. E., Navarro, F. C., Gerber, A. L., Nicolás, M. F., Salim, A. C. M., Silva, A. P. M., *et al.* (2011). Distinct patterns of somatic alterations in a lymphoblastoid and a tumor genome derived from the same individual. *Nucleic acids research*, **39**(14), 6056–6068.

Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2010). Next-generation variationhunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, **26**(12), i350–i357.

Huot, G., Vernier, M., Bourdeau, V., Doucet, L., Saint-Germain, E., Gaumont-Leclerc, M.-F., Moro, A., and Ferbeyre, G. (2014). Ches1/foxn3 regulates cell proliferation by repressing pim2 and protein biosynthesis. *Molecular biology of the cell*, **25**(5), 554–565.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature genetics*, **36**(9), 949–951.

Iakovishina, D., Janoueix-Lerosey, I., Barillot, E., Regnier, M., and Boeva, V. (2016). Sv-bay: structural variant detection in cancer genomes using a bayesian approach with correction for gc-content and read mappability. *Bioinformatics*, **32**(7), 984–992.

Janssen, J. W., Imoto, I., Inoue, J., Shimada, Y., Ueda, M., Imamura, M., Bartram, C. R., and Inazawa, J. (2002). Myeov, a gene at 11q13, is coamplified with ccnd1, but epigenetically inactivated in a subset of esophageal squamous cell carcinomas. *Journal of human genetics*, **47**(9), 460–464.

Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome biology*, **15**(6), R84.

Luo, R. (2017). https://github.com/aquaskyline/lrsim.

Mak, A. C., Lai, Y. Y., Lam, E. T., Kwok, T.-P., Leung, A. K., Poon, A., Mostovoy, Y., Hastie, A. R., Stedman, W., Anantharaman, T., *et al.* (2016). Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics*, **202**(1), 351–362.

Mccarroll, S. a., Huett, A., Kuballa, P., Chilewski, S. D., Goyette, P., Zody, M. C., Hall, J. L., Brant, S. R., Judy, H., Duerr, R. H., Silverberg, M. S., Taylor, K. D., Rioux, J. D., Daly, M. J., and Xavier, R. J. (2009). NIH Public Access. **40**(9), 1107–1112.

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. a., Conrad, D. F., Park, H., Hurles, M. E., Lee, C., Venter, J. C., Kirkness, E. F., Levy, S., Feuk, L., and Scherer, S. W. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome biology*, **11**.

Rausch, T., Zichner, T., Schlattl, A., St??tz, A. M., Benes, V., and Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**(18), 333–339.

Rausch, T., Jones, D. T. W., Zapatka, M., Stütz, A. M., Weischenfeldt, J., Jäger, N., Remke, M., Shih, D., Paul, A., Pfaff, E., Tica, J., Wang, Q., Massimi, L., Witt, H., Pleier, S., Cin, H., Hawkins, C., Beck, C., Von, A., Hans, V., Brors, B., Eils, R., Scheurlen, W., Benes, V., Kulozik, A. E., Witt, O., and Martin, D. (2013). NIH Public Access. *NIH Public Access*, **148**, 59–71.

Ritz, A., Bashir, A., and Raphael, B. J. (2010). Structural variation analysis with strobe reads. *Bioinformatics*, **26**(10), 1291–1298.

Ritz, A., Bashir, A., Sindi, S., Hsu, D., Hajirasouliha, I., and Raphael, B. J. (2014). Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics*, **30**(24), 3458–3466.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., *et al.* (2007). Strong association of de novo copy number mutations with autism. *Science*, **316**(5823), 445–449.

Sindi, S., Helman, E., Bashir, A., and Raphael, B. J. (2009a). A geometric approach for classification and comparison of structural variants. *Bioinformatics*, **25**(12), 222–230.

Sindi, S., Helman, E., Bashir, A., and Raphael, B. J. (2009b). A geometric approach for classification and comparison of structural variants. *Bioinformatics*, **25**(12), i222–i230.

Sindi, S. S., Onal, S., Peng, L. C., Wu, H.-T., and Raphael, B. J. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol*, **13**(3), R22.

Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., Zook, J. M., Salit, M., West, R. B., Batzoglou, S., and Sidow, A. (2016). Genome-wide reconstruction of complex structural variants using read clouds. *bioRxiv*, page 074518.

Stephens, P. J., McBride, D. J., Lin, M.-L., Varela, I., Pleasance, E. D., Simpson, J. T., Stebbings, L. A., Leroy, C., Edkins, S., Mudie, L. J., *et al.* (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**(7276), 1005.

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., Pleasance, E. D., Lau, K. W., Beare, D., Stebbings, L. A., *et al.* (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *cell*, **144**(1), 27–40.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., *et al.* (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**(7571), 75–81.

Tattini, L., D?Aurizio, R., and Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in bioengineering and biotechnology*, **3**.

Wala, J., Bandopadhayay, P., Greenwald, N., O'Rourke, R., Sharpe, T., Stewart, C., Schumacher, S. E., Li, Y., Weischenfeldt, J., Yao, X., *et al.* (2017). Genome-wide detection of structural variants and indels by local assembly. *bioRxiv*, page 105080.

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, **14**(2), 125–138.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D., and Jaffe, D. B. (2016). Direct determination of diploid genome sequences. *bioRxiv*, page 070425.

Xi, R., Luquette, J., Hadjipanayis, A., Kim, T.-M., and Park, P. J. (2010). Bic-seq: a fast algorithm for detection of copy number alterations based on high-throughput sequencing data. *Genome biology*, **11**(S1), O10.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (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), 2865–2871.

Yin, D.-T., Xu, J., Lei, M., Li, H., Wang, Y., Liu, Z., Zhou, Y., and Xing, M. (2016). Characterization of the novel tumor-suppressor gene ccdc67 in papillary thyroid carcinoma. *Oncotarget*, **7**(5), 5830.

Zheng, G. X., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., *et al.* (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology*.

Zheng, G. X., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., *et al.* (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology*.

Zhu, X.-J., Liu, J., Xu, X.-Y., Zhang, C.-D., and Dai, D.-Q. (2014). Novel tumor-suppressor gene epidermal growth factor-containing fibulin-like extracellular matrix protein 1 is epigenetically silenced and associated with invasion and metastasis in human gastric cancer. *Molecular medicine reports*, **9**(6), 2283–2292.