BIOINFORMATICS HITSEQ PAPER

Vol. 26 no. 10 2010, pages 1291–1298
doi:10.1093/bioinformatics/btq153

Sequence analysis

Advance Access publication April 8, 2010

Structural variation analysis with strobe reads

Anna Ritz[1,†], Ali Bashir[2,†] and Benjamin J. Raphael[1,3,*]

[1]Department of Computer Science, Brown University, Providence, RI 02912, [2]Pacific Biosciences, 1505 Adams Drive, Menlo Park, CA 94025 and [3]Center for Computational Molecular Biology, Brown University, Providence, RI 02912, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Structural variation including deletions, duplications and rearrangements of DNA sequence are an important contributor to genome variation in many organisms. In human, many structural variants are found in complex and highly repetitive regions of the genome making their identification difficult. A new sequencing technology called *strobe sequencing* generates *strobe reads* containing multiple subreads from a single contiguous fragment of DNA. Strobe reads thus generalize the concept of paired reads, or mate pairs, that have been routinely used for structural variant detection. Strobe sequencing holds promise for unraveling complex variants that have been difficult to characterize with current sequencing technologies.

Results: We introduce an algorithm for identification of structural variants using strobe sequencing data. We consider strobe reads from a test genome that have multiple possible alignments to a reference genome due to sequencing errors and/or repetitive sequences in the reference. We formulate the combinatorial optimization problem of finding the minimum number of structural variants in the test genome that are consistent with these alignments. We solve this problem using an integer linear program. Using simulated strobe sequencing data, we show that our algorithm has better sensitivity and specificity than paired read approaches for structural variation identification.

Contact: braphael@brown.edu

Received on March 4, 2010; revised on April 2, 2010; accepted on April 5, 2010

1 INTRODUCTION

Identifying the DNA sequence differences that distinguish individuals is a major challenge in genetics. Recent whole-genome sequencing and microarray measurements have shown that copy number variants (insertions, duplications and deletions) and balanced rearrangements, such as inversions and translocations, are common in most organisms including human (Sharp *et al.*, 2006), mouse (Egan *et al.*, 2007), fly (Dopman and Hartl, 2007) and yeast (Faddah *et al.*, 2009). These larger differences in DNA sequences are commonly referred to as *structural variants*. The Database of Genomic Variants (Iafrate *et al.*, 2004) currently (winter 2010) lists nearly 30 000 copy number variants and nearly 900 inversion variants in the human genome. Although some of these variants are redundant and/or erroneous, it is clear that structural variation is an important component of human genome variation. In fact, there are more total base pairs in human genome affected by structural variants than single nucleotide polymorphisms (SNP; Redon *et al.*, 2006). Both common and inherited structural variants and *de novo* structural variants have recently been linked to a number of human diseases (Girirajan *et al.*, 2010; Greenway *et al.*, 2009; Marshall *et al.*, 2008). Moreover, somatic structural variants are common in cancer genomes and lead to altered regulation of oncogenes and tumor suppressor genes (Albertson *et al.*, 2003) and the creation of novel fusion genes (Mitelman *et al.*, 2004).

Much of the recent excitement surrounding structural variation stems from better measurement technologies. In particular, End Sequence profiling (Raphael *et al.*, 2003; Volik *et al.*, 2003), also known as paired read mapping (Korbel *et al.*, 2007; Tuzun *et al.*, 2005), has been used to identify structural variants in both normal and cancer genomes. In paired read mapping, DNA fragments from a test genome are sequenced from both ends, and these sequences (reads) are mapped to a reference genome. Paired reads, or mate pairs, with discordant alignments identify inversions, translocations, transpositions, insertions, deletions and other rearrangements that distinguish the test genome from the reference genome. A number of methods have been introduced to identify structural variants from paired read sequencing data (Bashir *et al.*, 2008; Chen *et al.*, 2009; Hormozdiari *et al.*, 2009; Korbel *et al.*, 2009; Lee *et al.*, 2008; Quinlan *et al.*, 2010; Raphael *et al.*, 2003).

Structural variants vary widely in size and complexity, and are more difficult to characterize than SNPs. Many are associated with repeated sequences in the genome (Korbel *et al.*, 2007), complicating their detection and characterization. In extreme cases, the variants themselves have highly repetitive or complex organization relative to the reference genome. For example, different lists of variants have been identified in the same individual using older clone-based sequencing (Kidd *et al.*, 2008) and various next-generation sequencing platforms (Bentley *et al.*, 2008; Hormozdiari *et al.*, 2009; Korbel *et al.*, 2007). Characterizing these complex variants requires longer reads, longer fragments, or both.

Pacific Biosciences recently demonstrated *strobe sequencing* technology (Turner, 2009). A *strobe read*, or *strobe*, consists of multiple *subreads* from a single contiguous fragment of DNA. These subreads are separated by a number of 'dark' nucleotides (called *advances*), whose identity is unknown (Fig. 1). A strobe with two subreads is analogous to a paired read, while strobes with more than two subreads provide additional information for structural variant detection. Thus far, Pacific Biosciences has demonstrated strobe reads with lengths up to 10 kb with 2–4 subreads each of 50–200 bp. Additional improvements are expected as technology matures.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

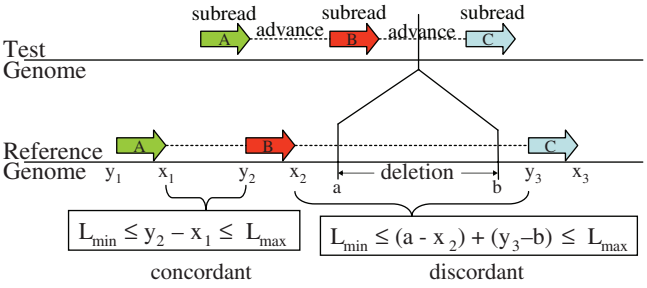


Fig. 1. A strobe read is a sequence of subreads and advances. When aligned to a reference genome, pairs of adjacent subreads can either be labeled as concordant or discordant, according to a derived minimum and maximum allowed length of an advance.

Strobe sequencing provides reads from long, contiguous fragments of DNA with low input DNA requirements, a feature missing from current short insert sequencing technologies. However, a consideration with strobe reads are the per-base error rates. While the capabilities of the Pacific Biosciences commercial machine (expected later in 2010) are not yet known, we conservatively assume that this error rate will be higher than existing next-generation sequencing technologies. Thus, realizing the advantages of strobe reads for structural variation detection demands new algorithms that utilize information from multiple subreads while allowing higher single-base error rates.

We introduce an algorithm to identify and characterize complex structural variants with strobe reads by considering multiple possible alignments for each subread. We formulate the combinatorial optimization problem of selecting an alignment for each subread of every strobe read so that the total number of structural variants in the test genome is minimized. This generalizes a formulation that has proved successful for paired read analysis (Hormozdiari *et al.*, 2009). We show how to reduce the problem to an optimization problem on directed graphs, and derive an integer linear program (ILP) for the problem. We apply our method to simulated strobe read sequencing data. We find that strobe reads outperform paired reads for structural variation detection. In particular, at a fixed sensitivity level strobe reads have nearly double the specificity of paired reads.

2 APPROACH

A *strobe read*, or *n-strobe*, $S=(R_1,A_1,R_2,A_2,\ldots,A_{n-1},R_n)$ is an alternating sequence of *subreads* R_i, whose sequence is known, and *advances* A_j, unknown sequences of 'dark' bases. The sequencing technology does not identify precisely the length, or number of dark bases, in each advance. However, the length of an advance is related to the time that the DNA polymerase incorporates dark bases. Thus, a distribution $f(a)$ for the length is obtained from the data. From this distribution, we derive a minimum and maximum allowed length of an advance which we denote by $L_{\min}$ and $L_{\max}$, respectively.

Suppose one has strobe reads from a test genome and wants to identify structural variants that distinguish this genome from a reference genome. The first step in structural variation identification is to align the strobe reads to the reference genome. Each subread R_i is aligned independently and corresponds to an interval $[y_i,x_i]$ on the reference genome. Suppose that $([y_1,x_1],\ldots,[y_n,x_n])$ is a sequence of alignments for the subreads of an n-strobe S. We say that an adjacent pair of subread alignments $([y_i,x_i],[y_{i+1},x_{i+1}])$ is *concordant* provided alignments are on the *same* strand and in the same order on the reference[1] and $L_{\min}\le|y_{i+1}-x_i|\le L_{\max}$. Otherwise, we say that the pair of subread alignments is *discordant*. Similarly, we say that a strobe read alignment is concordant if all pairs of adjacent subread alignments are concordant; otherwise, the strobe read is discordant.

A discordant pair $([y_i,x_i],[y_{i+1},x_{i+1}])$ suggests a structural variant in the test genome. For example, the breakpoints of a deletion from a pair of alignments mapped to the positive strand of the reference genome satisfy the equation

$$L_{\min}\le(a-x_i)+(y_{i+1}-b)\le L_{\max} \quad (1)$$

See Figure 1. Similar equations hold for inversions, translocations and other variants (Bashir *et al.*, 2008).

A collection of discordant strobe reads indicate the *same* structural variant, if they have discordant pairs that simultaneously satisfy the equation above for a particular choice of breakpoints a and b.

Now consider the case when subreads of a strobe read do not have unique alignments to the reference genome because of repetitive sequences in the reference genome and/or sequencing errors. In this case, the selection of an alignment for a subread determines a candidate structural variant. Our intuition is that breakpoints of true structural variants will be contained in many strobe reads. Thus, we aim to choose an alignment for each subread of every strobe read so that the resulting set of structural variants is *optimal* according to some objective function. This problem has been considered in the paired read case by Lee *et al.* (2008) and Hormozdiari *et al.* (2009). In particular, Hormozdiari *et al.* (2009) define the maximum parsimony objective function of choosing alignments to *minimize* the number of predicted structural variants.

Below we consider the equivalent problem for strobe reads, show a reduction to an optimization problem on directed graphs and derive an ILP for the problem.

3 METHODS

3.1 Problem formulation

Consider an n-strobe $S=(R_1,A_1,R_2,A_2,\ldots,A_{n-1},R_n)$ consisting of subreads R_i and advances A_j. For a subread R_i, let $M(R_i)$ denote the set of locations in the reference genome where R_i aligns, with each element of $M(R_i)$ being an interval $[y,x]$ on the reference genome. An alignment for S is obtained by selecting an alignment $m_i\in M(R_i)$ for each subread R_i. Thus, let $M(S)=M(R_1)\times M(R_2)\times\cdots\times M(R_n)$ be the set of alignments for S.

For $m\in M(S)$, let $B(m)$ be the set of genomic breakpoints indicated by m. If m is concordant, then $B(m)=\emptyset$. Note that these breakpoints are only approximately defined according to the uncertainty in the advance lengths [e.g. according to Equation (1)]. We define the following problem.

n-Strobe Minimum Breakpoints Problem. *Given alignment sets $M(S_1),\ldots,M(S_K)$ for n-strobes $S_1,\ldots,S_K$, find a set of breakpoints $\mathcal{B}$ of minimum cardinality such that for all $k=1,\ldots,K$ there is an $m_k\in M(S_k)$ with $B(m_k)\subseteq\mathcal{B}$.*

This problem is NP-hard, as was shown for the paired read case ($n=2$) by a reduction to the Set Cover problem (Hormozdiari *et al.*, 2009). Below, we reformulate this problem for general n as an ILP.

[1]The definition of concordant is different for paired read sequencing technologies (e.g. Illumina) that produce reads on opposite DNA strands.

3.2 Graph construction

We derive our ILP from a directed graph $G=(V,E)$ that represents breakpoints shared by multiple strobe reads. We begin with a graph whose edges are alignments of a subread and vertices are advances between subreads. Formally, consider an individual strobe read S. We represent the set of all possible alignments for S with a directed graph $G_S=(V_S,E_S)$. The vertex set $V_S=\cup_{i=1}^{n-1}\big(M(R_i)\times M(R_{i+1})\big)\cup\{\alpha_S,\beta_S\}$ is the set of all possible pairs of alignments of adjacent subreads, with an additional source vertex α_S and sink vertex β_S corresponding to the start of the first subread and the end of the last subread, respectively. We refer to vertices that are not sources or sinks as *internal* vertices. The edge set $E_S=\cup_i M(R_i)$, where each $m=[y,x]\in M(R_i)$ corresponds to a directed edge (y,x). The alignments for strobe read S are exactly the set of paths in G_S from α_S to β_S.

In the graph construction, we use concordant pairs to reduce the number of alignments that we consider for subreads. In particular, if there exists a concordant pair for subreads R_i and R_{i+1}, then we ignore all discordant pairs for these subreads R_i with lower alignment score. For every concordant pair of alignments $m_i\in M(R_i)$ and $m_{i+1}\in M(R_{i+1})$, we consider the vertex $v=(m_i,m_{i+1})$, add edges from all incoming neighbors to v to all outward neighbors from v and finally remove v from the graph.

Now we form a graph $G=(V,E)$ by merging vertices in the graphs G_S whose alignments are consistent with a single breakpoint (a,b) according to (1). We compute the vertices to merge using GASV (Sindi *et al.*, 2009), a program that efficiently computes whether paired reads indicate the same structural variant using a computational geometry algorithm. In each merged vertex, we store the identities of strobe and subread alignments from the original vertices.

The n-Strobe Minimum Breakpoints Problem reduces to finding a subgraph H of minimum cardinality (fewest number of vertices) such that H contains a path from source α_k to sink β_k for all $k=1,\ldots,K$. Note that H will always contain the source and sink vertices.

3.3 ILP formulation

The graph formulation suggests that finding the subgraph H might be solved as a *fixed charge* network flow problem. In particular, the graph formulation is suggestive of a fixed charge multi-commodity network flow problem (Crainic *et al.*, 2001), with each strobe representing a distinct commodity. However, our problem differs from this and related problems in that we need to maintain separate accounting of each strobe entering and exiting a vertex rather than merely accounting for the net flow as in a multi-commodity flow problem.

Nevertheless, we are able to formulate our problem as an ILP, motivated by an ILP for the fixed charge flow problem (Hochbaum and Segev, 1989). For each vertex $v_i\in G$, we define binary indicator variables p_i such that $p_i=1$ if and only if v_i is in the optimal solution. Similarly, we introduce variables q_{ij}^k, which represent the flow across edge (v_i,v_j) for strobe k. Lastly, we define N_{i+}^k and N_{i-}^k as the outward and inward neighbors, respectively, of vertex v_i and strobe k.

Our ILP for minimizing structural variants is then given by,

$$\min \sum_i p_i \tag{2}$$

$s.t.$

$$\begin{aligned} &p_i\in\{0,1\} \quad \forall i \\ &0\le q_{ij}^k\le p_i \quad \forall i,j,k \\ &0\le q_{ij}^k\le p_j \quad \forall i,j,k \end{aligned} \tag{3}$$

$$\sum_{j\in N_{i+}^k} q_{ij}^k - \sum_{j\in N_{i-}^k} q_{ji}^k = \begin{cases} 1 & \text{if } i=\alpha_{S_k} \\ -1 & \text{if } i=\beta_{S_k} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Recall that each vertex v_i corresponds to a breakpoint. perhaps common to multiple strobes. The objective (2), minimizes the number of vertices, thus minimizing the number of breakpoints. In the optimal solution, each edge q_{ij}^k must have weight 0 or 1—it is either used once or not used at all per strobe. However, this does not need to be enforced as an explicit binary integer constraint. Instead, constraint (3) bounds q_{ij}^k between 0 and 1. Note that the flow for any edge q_{ij} can only be non-zero if the v_i and v_j are in the optimal solution. Constraint (4) ensures that each strobe has a valid sequence of subread alignments.

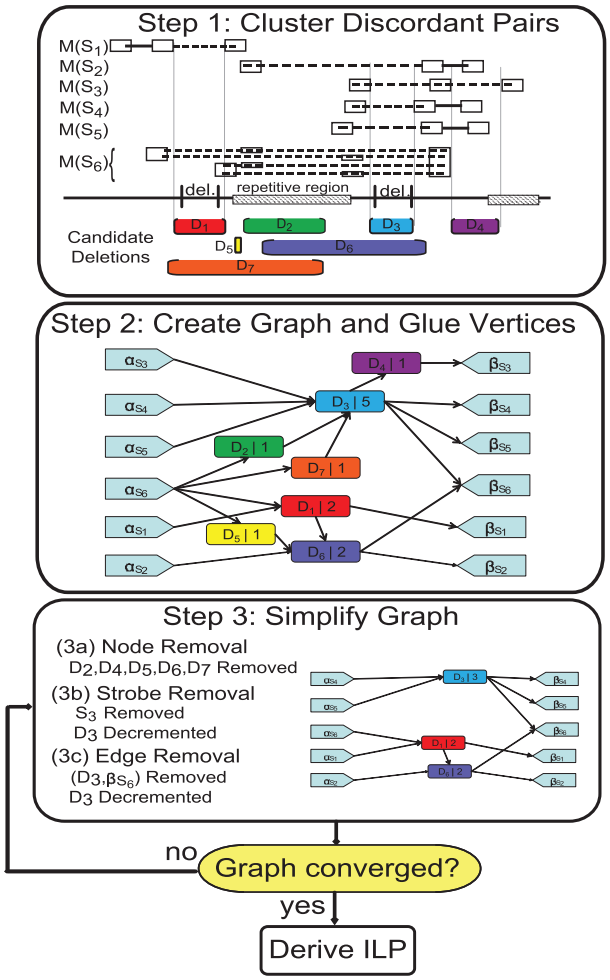


Fig. 2. Construction of graph for 3-strobes $S_1,\ldots,S_6$. In Step 1, the discordant pairs (dotted lines) are clustered, producing seven candidate deletions. In Step 2, a graph is created with a source vertex α and a sink vertex β for each strobe and one vertex for each cluster (candidate deletion). The number on each vertex indicates the number of strobes with a discordant pair consistent with the deletion. In Step 3, the graph is simplified using heuristics described in Section 3.4. Step 3 is repeated until the graph has converged, upon which it is used as input to the ILP.

3.4 Graph simplification

In order to improve performance, we developed several heuristics to simplify the graph G before solving the ILP (Fig. 2).

(1) *Vertex removal: remove vertices in G that are supported by fewer than Δ strobe reads.* We define the *support* of a vertex as the number of strobe reads with paths through it. We require that a breakpoint be supported by at least Δ strobe reads. Thus, we remove internal vertices with support less than Δ. Note that we must count the number of strobe reads that travel through the vertex, rather than the number

of discordant pairs that cluster together, because the same strobe read might have multiple discordant pairs that support the same breakpoint. We also remove all incoming and outgoing edges of a removed vertex.

(2) *Strobe removal: remove a strobe read S in G with no path from source α_S to sink β_S.* Following vertex removal, some strobe reads will no longer have a path from source to sink. We remove each of these strobe reads from the graph. We use a dynamic program to efficiently check the path constraint for each strobe.

(3) *Edge removal: remove edges corresponding to subread alignments for strobe read S that are on a path from source α_S to sink β_S.* Following vertex removal, each strobe S will have at least one path from source to sink, but may have extraneous alignments that do not lie on any path from source to sink. We remove such edges from the graph, since they cannot appear in the ILP solution due to the flow constraint. The dynamic program for strobe removal is also used for this step.

We iteratively perform these three operations on G until no more vertices, strobe reads or edges are removed.

3.5 Benchmarking

Here, we present the computations used to assess sensitivity and specificity in detecting a set of known deletions. Consider a set of deletions defined by intervals $P=\{[a_1,b_1],\ldots,[a_{|P|},b_{|P|}]\}$ known to be present in the test genome. Given a set $B=\{[c_1,d_1],\ldots,[c_{|B|},d_{|B|}]\}$ of intervals corresponding to a prediction returned by the ILP, we compute two different Receiver Operating Characteristic (ROC)-like plots.

(1) *Variant-based 'ROC'.* We count the number of deletions $[a_i,b_i]\in P$ that have a non-empty intersection with some interval $[c_j,d_j]\in B$ and

$$(d_j-c_j)\leq(b_i-a_i)+L_{\max},$$

i.e. the prediction length must be smaller than the deletion length plus the largest allowed length of an advance. Let the number of such variants be V. The true positive rate is $\frac{V}{|P|}$. and the number of false positives $|B|-V$. Note that this is not a true ROC curve because we do not compute a false positive rate, but rather report the number of false positives (hence the quotes).

(2) *Pair-based ROC.* We count the number of discordant pairs that support variants. Let $D(P)$ be the set of discordant pairs that support variants in P, and let $\overline{D(P)}$ be the remaining discordant pairs. Similarly, let $D(B)$ be the discordant pairs that support variants in B. The true positive rate is then $\frac{|D(B)\cap D(P)|}{|D(P)|}$ and the false positive rate is $\frac{|D(B)\cap\overline{D(P)}|}{|\overline{D(P)}|}$.

4 RESULTS

We applied our algorithm to simulated strobe sequencing data from Pacific Biosciences. The Pacific Biosciences simulator models the errors in their single-molecule sequencing technology. Specifically, the simulator models the higher rate of insertions and deletions (using a roughly equivalent ratio of each) relative to miscall errors in subreads that is typically seen in their data (Eid *et al.*, 2009). We analyze two datasets: structural variants identified in the Venter whole-genome assembly (Levy *et al.*, 2007) and a synthetic complex rearrangement in highly repetitive regions. For each dataset, we generated 3-strobe reads of length 3 kb, consisting of three subreads of 200 bp separated by 1200 bp advances. We then introduced error into the subreads in each strobe using Pacific Biosciences's error simulator. Since the capabilities of the Pacific Biosciences commercial machine (expected later in 2010) are not yet known, we conservatively assume a sequencing error rate of 5%.

4.1 Comparison to paired reads

For each dataset, we compare the results with strobe reads to those obtained via paired reads. To remove differences due to read alignment, we construct paired reads using subsets of the subreads in each strobe. We consider two sets of paired reads:

(1) *Paired Read Library (1.6 kb fragment length)*: for each n-strobe $S=(R_1,A_1,R_2,A_2,\ldots,R_{n-1},A_{n-1},R_n)$, we define the set of pairs $\mathcal{L}(S)=\{(R_i,R_{i+1})\}$.

(2) *Mixed Paired Read Library (1.6 and 3 kb fragment lengths)*: we define the set of pairs $\mathcal{A}(S)=\{(R_i,R_j):j>i\}$.

The Paired Read Library dataset corresponds to a paired read dataset generated by a single size selection, while the Mixed Paired Read Library dataset corresponds to multiple fragment sizes. Note that for Illumina and ABI SOLiD machines, the latter requires preparation of multiple sequencing libraries.

Given a set of strobe reads with physical coverage c, the Paired Read Library will have approximately the same physical coverage as strobe reads with twice as many reads. The Mixed Paired Read Library will have physical coverage $2c$ with three times as many reads. We subsample the reads in the Mixed Paired Read Library to achieve physical coverage c.

4.2 Variant detection on Venter Chromosome 17

We simulated a test chromosome based on known rearrangements from Chromosome 17 of the Venter genome (Levy *et al.*, 2007), following the procedure presented in Chen *et al.* (2009). Given a list of 17 376 insertions, deletions and inversions on Chromosome 17, we concatenated intervals from the hg18 human reference corresponding to the rearrangements. We then simulated 3 kb strobe reads at 10×, 20× and 30× coverage, producing 262 355, 524 709 and 787 064 strobe reads, respectively. After aligning the subreads to Chromosome 17 of hg18 using Pacific Biosciences's in-house aligner, BLASR, we generated a Paired Read library and a Mixed Paired Read library as described in Section4.1. BLASR is designed to quickly align large reads and is tolerant to a wide range of sequencing errors (M.Chaisson, personal communication).

4.2.1 Deletions We considered the 124 deletions greater than 120 bp in Venter and computed the variant-based ROC curve by varying Δ, the minimum support of a vertex (Fig. 3). In terms of 'Area under the ROC Curve' (AROC) values, strobe reads outperform paired reads for all three coverages. Moreover, strobe reads outperform Mixed Paired Reads for 10× and 30× coverages, where the advantage in AROC for the Mixed Paired Read library is a result of slightly better specificity at extremely low (<20%) values of sensitivity. On average, at fixed values of sensitivity, strobe reads make 57.18% ± 4.282 fewer false positive predictions than paired reads. At 20× coverage and a maximum sensitivity of 87.10% for strobe reads, 90.32% for the Paired Read library and 92.74% for the Mixed Paired Read library, strobe reads make 45.13% fewer false positive predictions than the Paired Read library, and 61.53% fewer false positive predictions than the Mixed Paired Read library.

It is possible that some of the predictions in the ILP solutions are accurate, but we fail to identify them as a true positives because the breakpoint coordinates might be shifted due to nearby structural variants. Of the considered deletions, 39 have another event (at least 50 bp in size) within the corresponding advance.

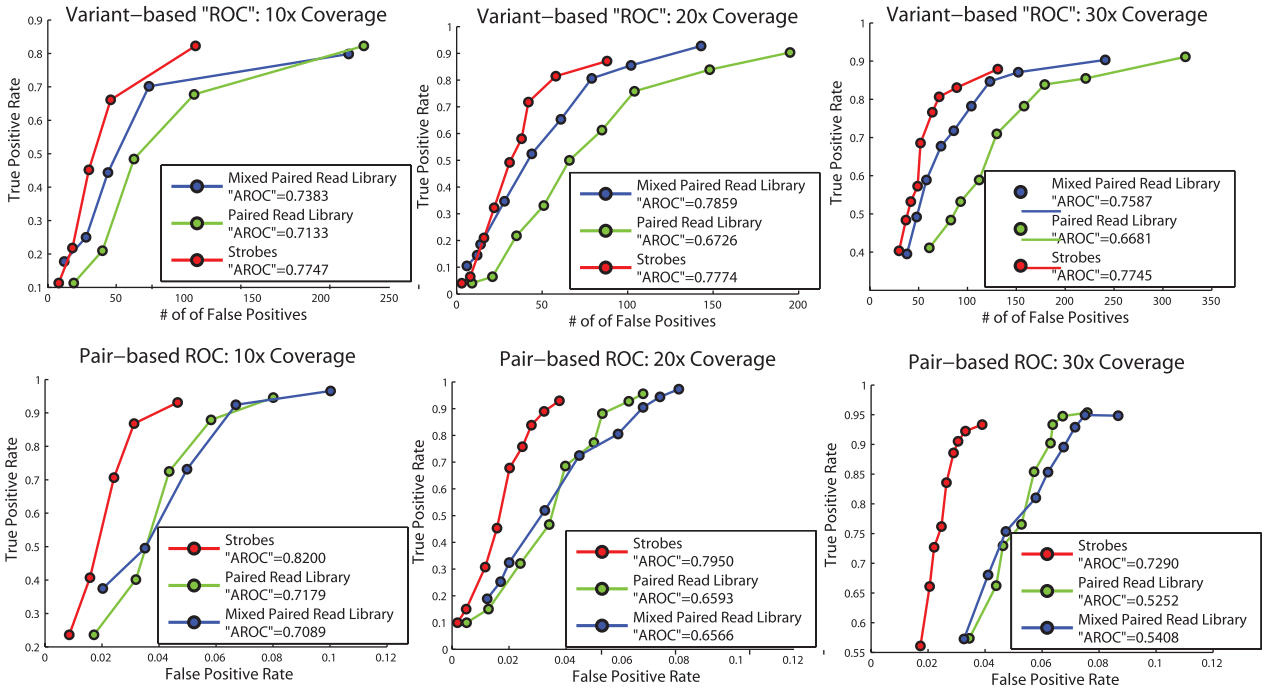


Fig. 3. ROC curves for the Venter simulation. The top row is a variant-based 'ROC' curve for 124 deletions $\geq$120 bp, which computes the number of real deletions out of 124 that appear in the solution. The bottom row is a pair-based ROC curve for the same 124 deletions, which counts the number of discordant pairs that support the real deletions (see Section 3.5). The reported 'AROC' is the area under the curves normalized by the maximum x-value in each plot. For 10× coverage, Δ ranges from 2 to 10 in steps of 2. For 20× and 30× coverage Δ ranges from 4 to 20 in steps of 2.

Thus, we considered the pairs that span the 124 deletions in each dataset and computed the 'pair-based' ROC curve (Fig. 3). Strobe reads outperform paired reads and mixed paired reads for all three coverages (AROC). Strobe reads decrease the false positive rate by an average of 50.83% ± 4.83 compared with the Paired Read datasets and 56.07% ± 8.11 compared with the Mixed Paired Read libraries at fixed sensitivity.

The size and topology of the graph G used as input to the ILP varies between the Strobe, Paired Read and Mixed Paired Read libraries, affecting the runtime of the ILP. In particular, the Strobe datasets construct graphs with fewer edges and vertices than the Paired Read and Mixed Paired Read libraries. For example, with support $\Delta = 8$ there are ∼65% fewer edges and internal vertices for 10× coverage; ∼80% fewer edges and internal vertices for 20× coverage; and ∼83% fewer edges and internal vertices for 30× coverage (Table 1). Note that this does not necessarily indicate that the Strobe dataset generates fewer clusters that are used to construct the graph; Strobe and Paired Read datasets predict the same number of clusters at each coverage (which is expected due to the Paired Read library construction in Section 4.1); and the Mixed Paired Read library uses a smaller set of clusters to construct the graph G. In addition to the size of the input graph G, the graph topology is different between the Strobe datasets and the Paired Read libraries. In general, graphs with many connected components are easily parallelizable, as each connected component could be run independently of the others. While all graphs contain many connected components, the graphs of Paired Read and Mixed Paired Read libraries contain connected components with many more internal vertices than graphs with Strobe data. For example, at 10× coverage with $\Delta = 4$, the largest connected component for the Paired Read library contained 350 internal vertices while the largest connected component for the Strobe dataset contained only six internal vertices. One such connected component from the Strobe dataset with 10× coverage is shown in Figure 4.

4.2.2 Inversions Only four inversions appear on Venter Chromosome 17 with lengths detectable by the simulated sequencing data. Our method detects the two longest inversions of these four for the Strobe and Paired Read libraries, while the Strobe dataset predicts 22% fewer false positives (Table 2). Here, the true positive rate computation described in Section 3.5 does not apply; instead we counted an inversion if the mutual intersection between the inversion interval and the prediction interval was >50%, following Chen *et al.* (2009).

4.2.3 Comparison with short read sequencing To illustrate the advantages of the longer fragments and subreads of strobe sequencing, we compared strobe reads to simulated paired read data that approximates the fragments sizes and read lengths that are routinely obtained with short read, short insert sequencing technologies (Fig. 5). Using the same Venter Chromosome 17, we simulated 200 bp fragments with 50 bp reads using the wgsim program from SAMtools (Li *et al.*, 2009) at 20× coverage. The simulated fragments have a base error rate of 0.02, a mutation rate of 0.001, 10% indels and 30% probability that an indel is extended. We aligned the reads using Burrows-Wheeleer Alignment (BWA)

Table 1. Clustering, graph construction, and ILP solution statistics for the Venter Chromosome 17 with $\Delta = 8$

Pairs	Clustering statistics			Graph statistics				ILP sol. statistics	
	No. of Discordant	No. of Deletions	No. of clusters ≥8	No. of strobe reads or pairs		No. of internal vertices	No. of edges	No. of vertices in solution	No. of correct
				Removed	Retained				
10× Strobe	50 554	12 762	145	362	410	50	978	45	27
10× Paired Read	50 543	12 762	145	321	518	143	2638	66	26
10× Mixed Paired Read	35 881	5984	83	330	542	83	1650	59	31
20× Strobe	101 367	25 701	807	389	1585	159	4182	131	89
20× Paired Read	101 352	25 698	807	265	1974	785	19 270	198	94
20× Mixed Paired Read	70 022	11 896	372	288	1976	365	8658	179	100
30× Strobe	151 053	38 711	1492	545	2540	229	7784	171	100
30× Paired Read	151 034	38 705	1492	345	3240	1465	42 322	283	104
30× Mixed Paired Read	107 577	19 451	810	430	3028	802	21 650	228	105

The clustering statistics include the number of discordant pairs after removing concordant alignments, the number of these discordant pairs that are deletions (their lengths are greater than $L_{\max}$ and they have proper orientation) and the number of clusters. The graph statistics include the number of strobe reads or paired ends that are removed from the graph, the number that are retained, the number of internal vertices and the number of edges in the graph. The solution statistics report the number of internal vertices in the final ILP solution and the number of these vertices that are in the list of 124 deletions ≥120 bp.

(Li and Durbin, 2009). We then ran GASV (Sindi *et al.*, 2009) on the discordant pairs that map uniquely to the reference.

Since VariationHunter (Hormozdiari *et al.*, 2009) utilizes reads with non-unique alignments, we considered discordant pairs that have multiple alignments to the reference. We considered reads with low mapping quality (≤10) for BWA and aligned them with Novoalign, an aligner that has higher sensitivity than BWA at the cost of a longer running time (Krawitz *et al.*, 2010). We considered up to 100 alignments for reads aligned with Novoalign. At maximum specificity for strobe reads, where the true positive rate is 0.87, VariationHunter achieves a true positive rate of 0.35 and GASV achieves a true positive rate of 0.34 at approximately the same sensitivity (between 85 and 95 false positives).

We emphasize that this comparison is limited by its use of simulated data, and by our use of VariationHunter and GASV without further post-processing. For example, the VariationHunter publication (Hormozdiari *et al.*, 2009) describes several additional steps used to achieve better performance. Additionally, uncontrolled simulation parameters such as the fragment length and the subread length affect the comparison, and explicitly comparing the performance of different types of sequencing platforms is beyond the scope of this article.

4.3 Variant detection in repetitive regions

Repetitive regions in the genome are notoriously difficult for structural variant detection. To test the ability of strobe reads to capture breakpoints near repetitive regions, we constructed a 11.6 kb sequence with two translocations by concatenating three different transposons from hg18: a 6 kb L1-family LINE (chr2:181406133-181413161), a 503 bp Alu (chr7:66854543-66856104) and a 3 kb L2-family LINE (chr15:87930634-87933678), each flanked by 500 bp. From this sequence, we generated 10 simulations with 10× coverage using the same sized 3 kb strobe reads as above and introduced 5% error. After aligning the subreads to hg18, we generated the paired read datasets as above and ran our method.

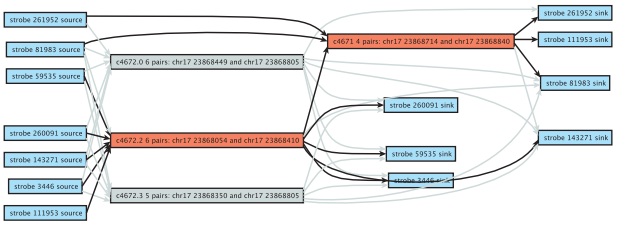


Fig. 4. A connected component from the solution of the 10× coverage strobe dataset with $\Delta = 4$. Edges in the final solution are in black. Six strobe reads support the first variant in the solution, four strobe reads support the second variant in the solution and two strobe reads span both variants.

Table 2. Predicted Inversions for Venter Chromosome 17

Left Breakpoint Coordinate	Length (bp)	Strobe	Paired Read	Mixed Paired Read
5 826 739	552			
40 566 233	1151	✓	✓	
55 552 838	3557	✓	✓	✓
57 999 778	472			
Total no. of predicted inversions		23	96	52

Inversions that appear in the solution for the Strobe, Paired Read, and Mixed paired read libraries with 20× coverage and $\Delta = 10$. The Strobe and Paired Read libraries detect two of the four inversions, while the Mixed Paired Read library detects only the longest inversion.

The Strobe datasets and Paired Read libraries report similar true positive rates on average, while the Strobe dataset reports fewer false positives (Fig. 6). Many subreads in the simulations align to

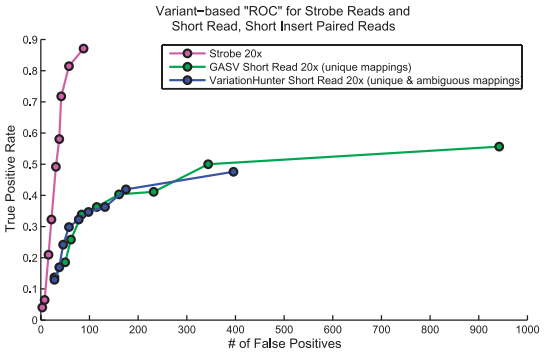
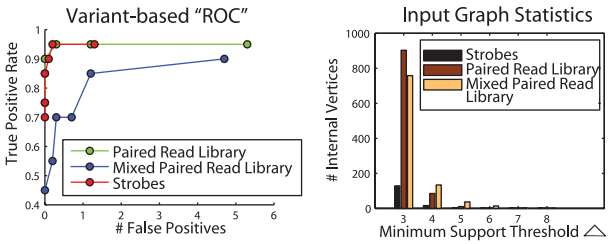


Fig. 5. Comparison of strobe reads to simulated short read, short insert paired reads for 124 deletions ≥120 basepairs. Δ ranges from 4 to 20 in steps of 2 for all curves.



Dataset	Avg # Input Vertices*	Avg # Final Predictions	Avg # Correct Predictions
Strobe	16 ± 25.88	2.1 ± 0.58	1.9 ± 0.32
Paired Read	85 ± 106.12	3.1 ± 1.37	1.9 ± 0.32
Mixed P.R	133 ± 183.20	2.9 ± 1.60	1.7 ± 0.48

* # of internal vertices in the input to the ILP.

Fig. 6. Statistics for 10 10× coverage simulations of a highly repetitive region. (Top left) ROC curve of average number of false positives and average true positive rate, varying Δ from 3 to 8. (Top Right) Distribution of the number of internal vertices in the generated graph.

hundreds of different regions on the genome, causing the Paired Read libraries to contain many internal vertices in the input graph for the ILP. However, the number of vertices in the graph is greatly reduced using strobe reads, indicating that many pairs in these repetitive regions are eliminated by incorporating concordant information (Fig. 6).

5 DISCUSSION

Structural variants vary widely in size and complexity, and thus are generally more difficult to characterize than SNPs. Sensitive and specific detection of structural variation in human genomes from next-generation sequencing data remains a challenge. This is due to both technological limitations (in read length, error rates and insert sizes) and biological factors. Structural variants in human are: (i) enriched for repetitive sequences near their breakpoints (Kidd *et al.*, 2008); (ii) may overlap or have complex architectures; and (iii) recurrent (but not identical) variants may exist at the same locus (Perry *et al.*, 2008; Scherer *et al.*, 2007). We have shown that strobe reads have advantages in sensitivity and specificity over paired reads for structural variant detection. Since a single strobe can resolve multiple breakpoints, inference of duplications and rearrangements become more direct. The method handles both intra-chromosomal (Section 4.2) and inter-chromosomal (Section 4.3) events, making it suitable for genome-wide analyses. While we have examined deletions, inversions and translocations in simulated strobe read data, explicitly testing the detection power of insertions using strobe reads remains future work. An interesting question is whether strobe reads will detect more complex variants that are beyond the abilities of current sequencing technologies.

We showed a preliminary comparison of simulated strobe sequencing versus simulated short read, short insert sequencing. Although this comparison demonstrates an advantage for strobes, a more rigorous assessment with real sequencing data is required to obtain a definite comparison of these technologies.

An inherent limitation in any approach is the accuracy of the underlying alignments. Most second generation sequencing technologies produce short reads with increasing error rates near the end of the reads and few insertions. The Pacific Biosciences technology differs in that it has predominantly long reads, uniform error over the entire read and an insertion/deletion heavy error model (Eid *et al.*, 2009). For these reasons, different alignment approaches have been suggested that use k-mer seeding followed by alignment extension (Li and Durbin, 2010). In our study, mappings were ranked using a naive Smith–Waterman score, assuming uniform quality of all sequenced bases. As mapping and scoring for single-molecule sequencing technologies mature, the proposed ILP can be changed to use more sophisticated scoring functions, increasing sensitivity (by ensuring true alignments are included) and specificity (by reducing spurious alignments).

Current and new technologies are continually increasing the read length (or subread length in the case of strobe reads). Longer subreads imply a higher frequency of *split* reads (subreads which span a breakpoint). Such events can be readily incorporated into the ILP by converting subreads with non-overlapping mappings (relative to the subread) into *psuedo*-strobe reads, leading to paths of variable length from source to sink for a given strobe read. The advance in this case becomes the distance between the end of the first alignment and the start of second alignment on the subread. This allows the method to be used as a generic framework for evaluating genome-wide structural variation given any form of sequencing data.

The suggested model is designed to handle germ-line mutations. In scenarios with heterogeneous data or when somatic mutations are abundant, such as in cancer, the assumptions in our model could be limiting. For example, the notion of minimizing breakpoints and clustering breakpoint junctions may be incorrect. In areas of elevated chromosome instability, each individual cancer cell may undergo different mutations leading to slightly (or greatly) different breakpoint boundaries. Additionally, breakpoint clustering is significantly more challenging as some genomic aberrations may only be present in a small fraction of cells. Integrating copy number information on a per breakpoint basis could provide added power in detecting these lower frequency events.

6 CONCLUSIONS

We introduce a combinatorial algorithm for structural variant identification from strobe read sequencing data. We show that strobe

reads outperform paired reads on simulated sequencing data. In particular, strobe reads have nearly doubled the specificity at fixed sensitivity for structural variation prediction. In the near future, we will test our algorithm on real sequencing data.

ACKNOWLEDGEMENTS

We thank Mark Chaisson for providing the alignment software for strobe reads. We thank Luke Peng for his assistance in generating simulated data.

Funding: National Science Foundation Graduate Research Fellowship (to A.R.); Career Award at the Scientific Interface from the Burroughs Wellcome Fund (to B.J.R.).

Conflict of Interest: Ali Bashir is an employee at Pacific Biosciences.

REFERENCES

Albertson,D. *et al.* (2003) Chromosome aberrations in solid tumors. *Nat. Genet.*, **34**, 369–376.

Bashir,A. *et al.* (2008) Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput. Biol.*, **4**, e1000051.

Bentley,D. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Chen,K. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.

Crainic,T. *et al.* (2001) Bundle-based relaxation methods for multicommodity capacitated fixed charge network design. *Discrete Appl. Math.*, **112**, 73–99.

Dopman,E. and Hartl,D. (2007) A portrait of copy-number polymorphism in Drosophila melanogaster. *Proc. Natl Acad. Sci. USA*, **104**, 19920–19925.

Egan,C. *et al.* (2007) Recurrent DNA copy number variation in the laboratory mouse. *Nat. Genet.*, **39**, 1384–1389.

Eid,J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Faddah,D. *et al.* (2009) Systematic identification of balanced transposition polymorphisms in Saccharomyces cerevisiae. *PLoS Genet.*, **5**, e1000502.

Girirajan,S. *et al.* (2010) A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat. Genet.*, **42**, 203–209.

Greenway,S. *et al.* (2009) De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nat. Genet.*, **41**, 931–935.

Hochbaum,D. and Segev,A. (1989) Analysis of a flow problem with fixed charges. *Networks*, **19**, 291–312.

Hormozdiari,F. *et al.* (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.*, **19**, 1270–1278.

Iafrate,A. *et al.* (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.

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

Korbel,J. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.

Korbel,J. *et al.* (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.*, **10**, R23.

Krawitz,P. *et al.* (2010) Microindel detection in short-read sequence data. *Bioinformatics*, **26**, 722–729.

Lee,S. *et al.* (2008) A robust framework for detecting structural variations in a genome. *Bioinformatics*, **24**, 59–67.

Levy,S. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Marshall,C. *et al.* (2008) Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.*, **82**, 477–488.

Mitelman,F. *et al.* (2004) Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat. Genet.*, **36**, 331–334.

Perry,G. *et al.* (2008) The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.*, **82**, 685–695.

Quinlan,A.R. *et al.* (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.* [Epub ahead of print].

Raphael,B. *et al.* (2003) Reconstructing tumor genome architectures. *Bioinformatics*, **19** (Suppl. 2), i162–i171.

Redon,R. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Scherer,S. *et al.* (2007) Challenges and standards in integrating surveys of structural variation. *Nat. Genet.*, **39**, 7–15.

Sharp,A.J. *et al.* (2006) Structural variation of the human genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 407–442.

Sindi,S. *et al.* (2009) A geometric approach for classification and comparison of structural variants. *Bioinformatics*, **25**, i222–i230.

Turner,S. (2009) *Personal Genomes* (conference talk). Cold Spring Harbor, NY.

Tuzun,E. *et al.* (2005) Fine-scale structural variation of the human genome. *Nat. Genet.*, **37**, 727–732.

Volik,S. *et al.* (2003) End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc. Natl Acad. Sci. USA*, **100**, 7696–7701.