Bioinformatics Advance Access published April 12, 2010

# Detection and characterization of novel sequence insertions using paired-end next-generation sequencing

Iman Hajirasouliha[1,*], Fereydoun Hormozdiari[1,*], Can Alkan[2,3,*], Jeffrey M. Kidd[5] , Inanc Birol[1,4] , Evan E. Eichler[2,3,†] and S. Cenk Sahinalp[1,†]

[1]The Lab for Computational Biology, Simon Fraser University, Burnaby, BC, Canada
[2]Department of Genome Sciences, University of Washington, Seattle, WA, USA
[3]Howard Hughes Medical Institute, Seattle, WA, USA
[4] BC Cancer Agency, Genome Science Center, Vancouver, BC, Canada
[5] Stanford University School of Medicine, Stanford, CA, USA

Associate Editor: Dr. Alex Bateman

**ABSTRACT**

**Motivation:** In the past few years, human genome structural variation discovery has enjoyed increased attention from the genomics research community. Many studies were published to characterize short insertions, deletions, duplications, and inversions, and associate copy number variants (CNVs) with disease. Detection of new sequence insertions requires sequence data, however, the "detectable" sequence length with read-pair analysis is limited by the insert size. Thus longer sequence insertions that contribute to our genetic makeup are not extensively researched.

**Results:** We present NovelSeq: a computational framework to discover the content and location of long novel sequence insertions using paired-end sequencing data generated by the next-generation sequencing platforms. Our framework can be built as part of a general sequence analysis pipeline to discover multiple types of genetic variation (SNPs, structural variation, etc.), thus it requires significantly less computational resources than *de novo* sequence assembly. We apply our methods to detect novel sequence insertions in the genome of an anonymous donor and validate our results by comparing with the insertions discovered in the same genome using various sources of sequence data.

**Availability:** The implementation of the NovelSeq pipeline is available at `http://compbio.cs.sfu.ca/strvar.htm`.

**Contact:** eee@gs.washington.edu; cenk@cs.sfu.ca

## 1 INTRODUCTION

It is estimated that 19-40 Mb of human genomic sequence is missing from the human genome reference assembly (Li *et al.*, 2009). Although the Human Genome Project (HGP) revolutionized the field of genomics, human sequences that are not represented in the reference genome leads to incomplete genome analyses. The missing sequences can even harbor undiscovered exons or other types of sequences of functional importance. There is a need to discover the loci and content of so-called "novel sequence insertions" to build a more comprehensive human reference genome to better analyze the genomes of individuals from many different populations.

To date, one of the more promising methods to characterize longer DNA segments that are not represented in the human reference genome has been building sequence assemblies from unmapped fosmid clone ends sequenced with the traditional Sanger-based capillary sequencing (Kidd *et al.*, 2008) and, then, sequencing the entire fosmid clone (Kidd *et al.*, 2010). However, the higher cost of the capillary sequencing is prohibitive to characterize genomes of thousands of individuals. Next-generation sequencing technologies make sequencing of thousands of genomes possible, and for the first time, they give us the opportunity to discover novel sequences across many human populations in order to build better genome assemblies (or "pan genomes" (Li *et al.*, 2009)). Various computational methods were developed in the recent years to characterize structural variation, including deletions, insertions, inversions, and duplications, among human individuals using next-generation sequencing (NGS) platforms (Medvedev *et al.*, 2009). Characterization of longer novel sequences remained elusive due to the shorter insert size and sequence length associated with the NGS methods. For example, applying the end-sequence profiling approach (Volik *et al.*, 2003; Tuzun *et al.*, 2005; Kidd *et al.*, 2008) one cannot discover insertions $>$100 bp when 200 bp insert size is used with the Illumina platform (Bentley *et al.*, 2008; Hormozdiari *et al.*, 2009; Chen *et al.*, 2009). Currently, the only method applicable for the discovery of long novel insertions using NGS technologies is *de novo* sequence assembly (Simpson *et al.*, 2009; Chaisson and Pevzner, 2008; Li *et al.*, 2010). However, this approach requires large computational resources and requires further processing to anchor the sequences to the reference genome.

Here we present a computational framework to discover the content of novel sequence insertions using the NGS platforms.

*These authors contributed equally

†Corresponding authors.

We test our methods with the high-coverage (42X) short-insert sequence library generated from the genome of a Yoruba African individual (NA18507) sequenced using the Illumina platform (Bentley *et al.*, 2008). We validate the content of the predicted novel sequence insertions by comparing with sequences generated from fosmid end-sequence assembly (Kidd *et al.*, 2008), full fosmid sequencing (Kidd *et al.*, 2010), and *de novo* sequence assembly of the same Illumina WGS library (Li *et al.*, 2009). We show that our methods are reliable, and together with the cost optimizations introduced by the NGS platforms, they can be efficiently used to characterize the DNA sequences missing from the reference assembly to obtain a more complete picture of human genome diversity.

A "novel sequence insertion" refers to an insertion of a sequence into the donor genome where no subsequence with high similarity to the inserted sequence exists in the reference genome. We aim to identify novel sequence insertions in a high-coverage sequenced donor genome through our computational pipeline NovelSeq.

Note that the insertions of repeat sequences such as SINEs and LINEs, and segmental duplications do not constitute as novel sequence insertions since paralogs of the same repeat sequence exists elsewhere in the reference genome assembly. Therefore, the algorithms presented here will not be able to predict such repeat sequence insertions unless the inserted sequence is highly divergent from other existing copies. For algorithms specifically designed for repeat sequence (or more formally, transposon) insertion detection, see the recent paper by Hormozdiari *et al.* (2010).

In Section 2, we will present the general approach of the NovelSeq pipeline divided into five different phases. In Section 3, we will give the details of our algorithms, and finally in Section 4, we will discuss the results of the NovelSeq pipeline.

## 2 APPROACH

*(a) Paired-end read mapping:* The computational pipeline begins by mapping the WGS paired-end reads onto the reference genome using mrFAST (Alkan *et al.*, 2009) and identifying *orphan* reads and *one end anchored (OEA)* reads. The paired-end reads where neither end-read [1] sequences can be mapped (with more than $95\%$ sequence identity) to the reference genome are classified as *orphan* reads. Following the nomenclature previously described (Kidd *et al.*, 2008, 2010), if only one end-read is mapped onto the reference genome, such paired-end reads are classified as OEA.

A hypothesis that can explain the existence of these orphan and OEA paired-end reads in a sequenced donor genome is as follows. The unmapped reads of the OEA pairs and the orphan paired-end sequences both belong to novel sequence insertions (Figure 1(a)).

*(2) Orphan assembly and contamination removal:* Using available *de novo* assembly algorithms such as EULER-SR (Chaisson and Pevzner, 2008) and ABySS (Simpson *et al.*, 2009), we assemble all orphan reads into longer contigs. These contigs may later be identified as novel insertion sequences in the donor genome. In addition, we perform an initial screening of the contigs using BLAST (Altschul *et al.*, 1990) and remove any contig that contains sequences from known contaminants (e.g. Epstein-Barr virus, E. coli, cloning vectors, etc.). As a second test to remove the mapping artifacts, we remove the contigs that can be aligned to the reference genome with a sequence identity of more than $99\%$. One reason that such contigs were generated from the reads classified as orphans due to low-quality sequence at the tails of the reads, and thus remained unmapped. However, those reads can still be assembled into reliable contigs since both ABySS and EULER build de Bruijn graphs from 25 bp subsequences of the reads, effectively discarding the sequence tails causing the mapping artifacts.

*(c) OEA read clustering:* We use a novel *clustering algorithm* mrCAR (micro-read Cluster Anchored Reads) to cluster the OEA reads based on their mapping orientations and locations in the reference genome such that those OEA reads that support the same insertion in the donor genome are grouped together. Note that for each potential novel sequence insertion prediction, there exists a group of OEA read alignments with $'+'$ orientation (denoted as $OEA+$, the single end-read that has an alignment on the reference genome is aligned to the forward strand), and a second group of OEA read alignments with $'-'$ orientation (denoted as $OEA-$, the single end-read is aligned to the reverse strand). In the remainder of this paper, we use the term *OEA cluster* to describe the two groups of OEA reads that are both mapped to different strands yet support the same novel sequence insertion. Also note that for all pairs of $OEA+$ and $OEA-$ clusters that support the same insertion, the $OEA+$ cluster should be mapped to the proximal location, and the corresponding $OEA-$ cluster should be mapped to the distal location.

The objective of mrCAR is to identify the OEA clusters efficiently such that, with a minimum number of novel sequence insertion prediction, all OEA paired-end reads are "explained" (i.e. for every OEA paired-end read $oea_i$, there exists an insertion prediction that is supported by $oea_i$).

*(d) The local assembly of the OEA clusters:* We assemble all unmapped end-reads in the OEA clusters that were created in the previous step into two OEA contigs using a local assembly routine, mrSAAB (micro-read Strand-Aware Assembly Builder). For each OEA cluster, the goal is to assemble the unmapped reads in each $OEA+$ cluster into a single contig (i.e. $OEA+$ *contig*) and the unmapped reads in each $OEA-$ into another single contig (i.e. $OEA$ *contig*).

*(e) Anchoring orphan contigs using the OEA contigs:* In the final stage of the NovelSeq pipeline, we aim to merge the OEA contigs (from both $+$ and $-$ strands) with the orphan contigs. Through this merging step, we both provide more read support for the orphan contigs and obtain the approximate anchoring position of the novel sequence insertion to the reference genome.

Our merging algorithm mrBIG (micro-read Big Insertion Gluer) aims to report the maximum number of orphan contigs

[1] Each end sequence of a paired-end read is referred to as *end-read*.

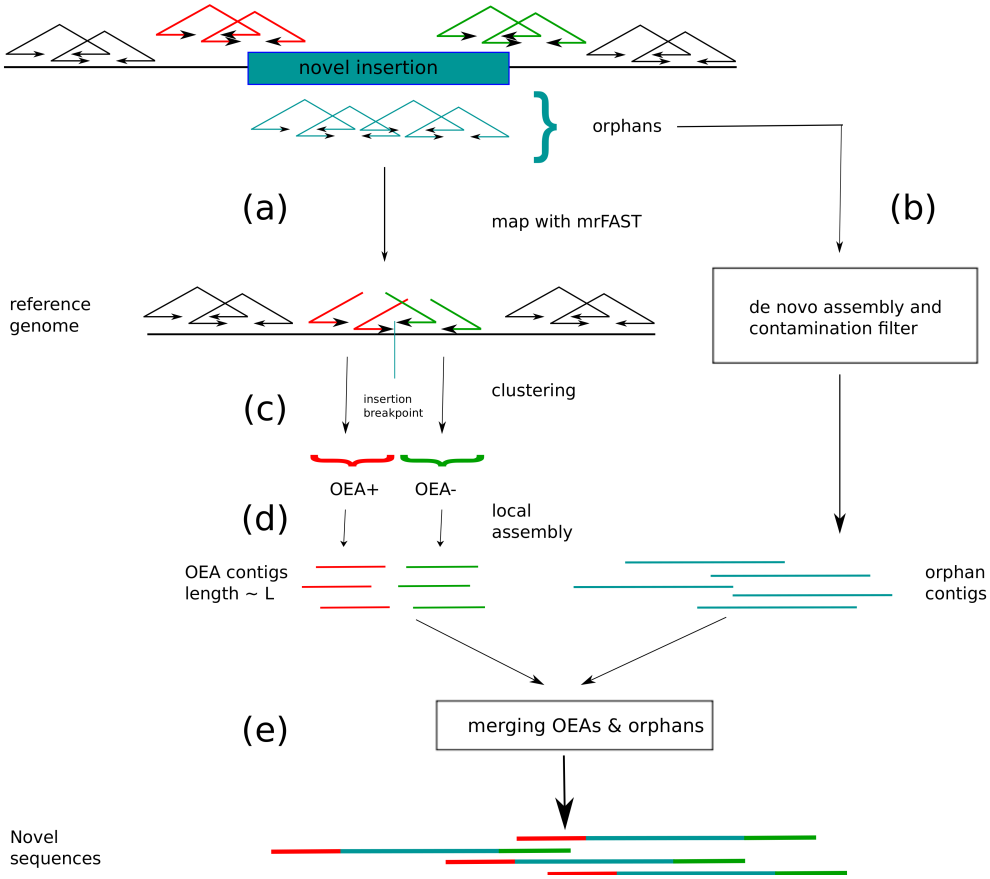



**Fig. 1.** The overall approach of the NovelSeq pipeline. (a) We start by mapping the paired-end reads to the reference genome and then classify the paired-end reads as OEA and orphan reads. (b) We then assemble the orphan paired-end reads using available *de novo* assembly algorithms and screen the contigs for possible contamination. (c) We cluster the OEA reads into groups and find the insertion locus supported by each OEA cluster. (d) We assemble the unmapped end-read in each OEA cluster (the OEA reads with different orientation of mapping should be assembled independently). (e) Finally, we merge the orphan contigs and OEA contigs to anchor the orphan contigs to the reference genome.

that can be merged with OEA contigs with *high support* (defined as the length and sequence identity of the overlapping basepairs, see Section 3.4). mrBIG is based on an algorithm for maximum weighted matching in bipartite graphs (West, 2001).

## 3 METHOD

### 3.1 Notations and definitions

Here we present the notations and definitions that we use in the rest of this paper. We define the set of paired-end reads of a sequenced donor genome as $R = \{pe_1, pe_2, \cdots, pe_n\}$. Each paired-end read $pe_k$ can be mapped to multiple locations on the reference genome. The set of all alignments of $pe_k$ is defined as $Align(pe_k) = \{a_1 pe_k, a_2 pe_k, \cdots, a_j pe_k\}$. Structural variation discovery algorithms using read-pair analysis start by calculating the observed distance between the two end-reads of a paired-end read. This distance is referred to as the *insert size* (denoted by $InsSize$). The $InsSize$ is assumed to be in a range of $[\Delta_{\min}, \Delta_{\max}]$ and can be calculated as previously described (Tuzun *et al.*, 2005).

An alignment of a paired-end read to the reference genome is *concordant*, if the distance between the aligned end-reads is within the expected range of $[\Delta_{\min}, \Delta_{\max}]$, and the paired-end alignment orientation is $+-$ (i.e. the end read which was aligned on the proximal location is on the $+$ strand, and its matepair is mapped to a distal location on the $-$ strand).

The set of one end anchored reads is represented as $OEA$ and the set of orphan reads is represented as $Orph$. Note that $Orph, OEA \subset R$. The end-reads in $OEA$ can also be mapped to multiple locations on the reference genome. For all $pe \in OEA$, alignment of $pe$ is defined as $ape = (loc(ape), or(ape))$, where $loc(ape)$ is the map location and $or(ape) \in \{+, -\}$ is the alignment orientation of the mapped end-read.

### 3.2 Clustering the OEA reads

In this section we formally describe a greedy algorithm, named mrCAR, to identify the OEA clusters. We first mathematically formulate the conditions required by a group of OEA reads that support the same novel insertion. Next, similar to the approach introduced in (Hormozdiari *et al.*, 2009) to cluster the discordant paired-end reads, we present an efficient greedy algorithm to find the minimum number of OEA clusters such that all OEA reads would "support" at least one insertion (i.e. a maximum parsimonious explanation of all OEA reads (Hormozdiari *et al.*, 2009)). We remind the reader that although

the map location of an OEA read serves as a guide to detect the insertion breakpoint of the novel sequence, the possibility of multiple map locations for an OEA read makes detecting the correct position a challenging task.

*Clustering rules:* A set of OEA reads $clu \subset OEA$ supports the same insertion if the following conditions hold:

- For every pair of OEA read alignments $\rho_F \in clu$ and $\rho_R \in clu$ (without loss of generality we assume that $\rho_F$ aligns to the forward and $\rho_R$ aligns to the reverse strand), the map location of $\rho_F$ is proximal to the map location of $\rho_R$.
- The *maximum* pairwise distance between the map locations of the OEA reads in $clu$ with the same mapping orientation must be less that the maximum $InsSize$, $\Delta_{\max}$.
- The distance between the map locations of two OEA reads with different mapping orientations should not exceed twice the maximum $InsSize$, $2 \cdot \Delta_{\max}$.

Note that an OEA cluster $c$ is called a "maximal valid cluster" if no more OEA read alignment can be added to $c$ that all the conditions noted above remain valid. Through an iterative method, we find all such maximal valid clusters in polynomial time. We first order all OEA read alignments based on their $loc$ value, and then traverse the genome from left to right. For each genome position $k$, we consider a window of size $2\Delta_{max}+1$ centered at k. Every OEA alignment inside the first half of the window with a $+$ orientation, and every OEA alignment on the second half of the window with a $-$ orientation is considered as one potential maximal valid cluster. Finally, a pairwise comparison is performed between all overlapping clusters detected in the previous step and only the maximal clusters are reported.

*Selecting the minimum number of clusters:* We define the Maximum Parsimonious Insertion Detection (MPID) problem as follows. Given a set of OEA clusters where each cluster potentially indicates a novel insertion, our goal is to select the minimum number of clusters (i.e. to minimize the total number of insertions) such that all OEA reads are aligned to the reference genome. We model this problem as a set cover problem and provide an $O(\log n)$ approximation solution. Note that the set of all OEA reads is the *universe* of elements, and the clusters created in the previous step are the sets that are selected to cover this universe. MPID is a necessary step since an OEA read can be present in multiple clusters.

### 3.3 Local assembly of the OEA clusters

The next step is to assemble the unmapped reads of OEA clusters that were created by the clustering algorithm and selected by the set cover approach. In each cluster, the OEA reads with mates that map to the $+$ strand and the reads with mates that map to the $-$ strand should be assembled into $OEA+$ and $OEA-$ contigs independently. However, the available *de novo* assemblers including EULER and ABySS do not provide the option of assembling the reads of only a single strand [2]. Using single end-reads, both ABySS and EULER consider the reverse complements of the read sequences as well. We therefore develop a local assembly routine that makes use of the fact that all unmapped reads from a single OEA cluster originate from the single strand reciprocal to the mapping orientation of the anchored reads from the same cluster. During the traversal of the assembly graph, we do not allow two consecutive OEA reads such that the mapping locations of their mates (from the corresponding paired-end reads) are *too far* from each other. The map location order of the anchored reads dictates the approximate positions of the unmapped reads in the local OEA assembly. The confidence interval for this position information depends on the $InsSize$ distribution.

Our local assembly routine is based on the standard overlap-layout-consensus graph approach. Note that this routine can also be implemented with an Eulerian path approach using a de Bruijn graph (e.g. through a modification to ABySS or EULER). Next, we briefly present this routine.

*Traversal of the overlay graph:* We first construct the overlay graph for all unmapped reads in an OEA cluster whose mates are anchored to the same strand.

Note that there will be two disjoint assembly graphs representing two different strands for each OEA cluster. Given a pair of nodes $u, v$ in the overlay graph (representing two OEA reads), we add a weighted directed edge connecting $u$ with $v$ if there exists an overlap between the suffix of $u$ and the prefix of $v$. The assigned weight of the noted edge will be a function of the suffix-prefix overlap between them. We implemented a greedy heuristic to find an *assembly* of the reads using both the edge weights and the extra information of the mapping locations of the other mates.

### 3.4 Merging the OEA and orphan contigs

Given the set of OEA and orphan contigs, we aim to find the maximum number of orphan contigs that can be merged with OEA contigs. We do not allow an orphan contig to merge with a pair of OEA contigs ($oea_+$ and $oea_-$) if the score of the prefix/suffix match between the two ends of the orphan contig and $oea_+$ and $oea_-$ is less than a user-defined threshold.

We mathematically model this problem as a maximum-weight bipartite matching problem, and give an exact solution based on the Hungarian method West (2001).

Let $Orph_{co} = \{or_1, or_2, \cdots, or_k\}$ be a set of orphan contigs and $OEA_{co} = \{oea_1, oea_2, \cdots, oea_v\}$ be a set of OEA contigs where $oea_j$ is a pair of two OEA contigs from the local assembly of the OEA cluster with id $j$. (i.e., $oea_i = (oea_{i_+}, oea_{i_-})$). We aim to assign each element in $Orph_{co}$ (e.g. $or_i \in Orph_{co}$) to an element in $OEA_{co}$ (e.g. $oea_j \in OEA_{co}$) such that the summation of (i) the alignment score between the prefix of $or_i$ and the suffix of $oea_{j_+}$ and (ii) the alignment score between the suffix of $or_i$ and the prefix of $oea_{j_-}$ is maximized.

We reduce this problem to the maximum-weight matching problem in a bipartite graph $G(U, V, E)$ where $G$ is defined as follows (Figure 2:

- $\forall or_i \in Orphco : \exists u_i \in U$
- $\forall oea_j \in OEA_{co} : \exists v_j \in V$
- The weight of edge $(u_i, v_j)$ is a function of the overlap between the the first $\Delta_{\max}$ basepairs of $or_i$ and $oea_{j_+}$ and the overlap between the last $\Delta_{\max}$ basepairs of $or_i$ with $oea_{j_-}$.

## 4 EXPERIMENTAL RESULTS

We tested our framework using the whole-genome shotgun (WGS) sequence library generated from the genome of an anonymous Yoruba African donor (NA18507) generated with the Illumina Genome Analyzer platform (Bentley *et al.*, 2008). The genome of NA18507 has been previously studied by many groups (Hormozdiari *et al.*, 2009; Alkan *et al.*, 2009; Lee *et al.*, 2009; Chen *et al.*, 2009) to discover structural variation and copy number polymorphism. This dataset contains approximately 3.5 billion sequence reads ($\sim$

[2] Personal communication with the developers of these tools

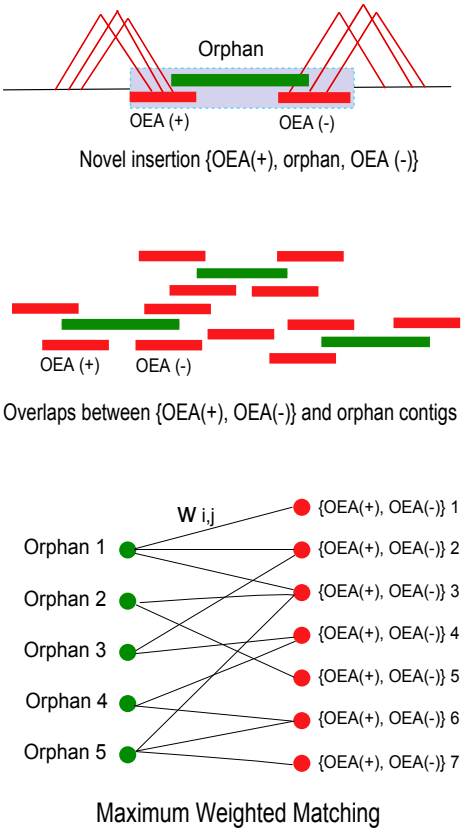



**Fig. 2.** Merging the orphan contigs with OEA clusters. Note that each OEA cluster is in fact composed of two contigs with different orientations that together represent an insertion. Each orphan contig is shown as a green node and each OEA cluster (as a 2-tuple) is represented with a red node. The edge weights are assigned as the total alignment score of suffix/prefix matches between the OEA clusters and the orphan contigs.

1.7 billion pairs) of length $36 - 41$bp with an $InsSize$ of $\sim$ 209bp (Bentley *et al.*, 2008; Hormozdiari *et al.*, 2009). The $InsSize$ distribution of this dataset was previously presented in (Hormozdiari *et al.*, 2009).

### 4.1 Novel sequence insertion map

*Preprocessing.* Similar to the pre-screeing methodology used in (Hormozdiari *et al.*, 2009), we removed any paired-end reads from consideration if either (or both) end sequence has an average *phred* (Ewing *et al.*, 1998) quality value less than 20, or if either (or both) sequence contains more than 2 unknown (i.e. $N$) nucleotides.

*Mapping to the reference genome.* After the preprocessing step, we mapped all the remaining $\sim 2.2$ billion end sequences to the human genome reference assembly (UCSC build 36) using *mrFAST* (Alkan *et al.*, 2009), allowing for edit distance $\leq 2$. Note that mrFAST returns *all* possible map locations of read sequences, thus an OEA read can be aligned to multiple locations in the reference genome. In total, $15,173,562$ pairs of reads ($30,347,124$ end-sequences) were identified as *orphans*, while $83,662,790$ reads were identified as *OEA*s.

*Orphan assembly.* Using ABySS (Simpson *et al.*, 2009), we assembled the orphan paired-end reads into $4,154$ contigs of size $\geq$200 bp (N50 = 995). In the rest of this paper we call these contigs as *ABySS contigs*. As an independent assessment, we also generated the sequence assembly of the orphans using the EULER (Chaisson and Pevzner, 2008) algorithm, which we call *EULER contigs*. EULER returned $4,564$ contigs of size $\geq$200 bp (N50 = 730).

*Contamination removal.* Next, we screened the orphan contigs to test for contamination. Using BLAST (Altschul *et al.*, 1990), we compared the orphan contigs with the `nt` database [3], and removed the contigs that align to consensus sequences of known contaminants (E. coli, bacteriophage, herpesvirus, plasmid, Epstein-Barr, bacteria, etc.) from further consideration. In total, 39 contigs were removed from the ABySS contig set as contamination, where the majority were due to Epstein-Barr, a virus commonly used for cell immortalization. Figure 3 shows the length distribution of the ABySS contigs of length $\geq$200 bp after the contamination removal. Note that out of 4,115 ABySS contaminant-free contigs ($\geq$ 200bp), 1,984 are $\geq$500 bp and 778 are $\geq$1 Kbp in size. Among the EULER contaminant-free contigs, $1,690$ are $\geq$ 500 bp and 582 are $\geq$1 Kbp.

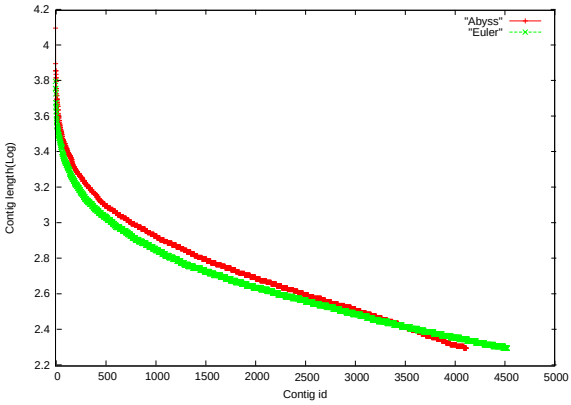



**Fig. 3.** Length distribution (log scale) of the ABySS (red) and EULER (green) contigs ($\geq$200 bp).

We then mapped the orphan contigs to the human genome reference assembly (both build35 and build36) using BLAST in order to remove the orphan contigs with high sequence identity with the reference genome. 493 of ABySS contigs of length $\geq$200 bp could be mapped onto either build35 or build36 with more than 99% sequence identity (548 of EULER contigs). We removed such contigs from consideration in the remainder of the NovelSeq pipeline. See Step 2 in Section 2 for the explanation of this filtering. The remaining ABySS contigs

[3] `http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastdb.html`

(n=3,622, Figure 4(a)) had a total length of 2.66Mbp while the remaining EULER contigs (n=3,977, Figure 4(a)) had the a total length of 2.37Mbp of the sequence.

*OEA clustering and orphan anchoring.* We used our clustering algorithm followed by the set cover approach to cluster the OEA reads, and obtained $10,560$ sets of OEA clusters with a *high support* [4] on each side (i.e. both $+$ and $-$ strands). Each side (or strand) of the detected OEA clusters were independently assembled using our local assembly routine, mrSAAB. Resulting OEA contigs were then processed together with the orphan contigs in the last phase of the NovelSeq pipeline, mrBIG. In summary, we anchored $130$ EULER contigs and $113$ ABySS contigs independently to the reference genome using the NovelSeq pipeline. In the merging phase of the orphan and OEA contigs (mrBIG), NovelSeq requires the alignment score between the orphan contig and the OEA contig to be $\geq$50. The alignment score is calculated as the score of the local alignment under affine gap model, where the *match* score is $+1$, *mismatch* penalty is $-1$, and *gap* penalties are $-16$ and $-4$ for gap opening and gap extension, respectively. The minimum requirement for the alignment score is a user defined parameter in the NovelSeq pipeline. Clearly, the lower alignment score one chooses at the merging phase, the more orphan contigs can be anchored to the reference assembly.

Recently, Kidd et al. end sequenced all fosmid clones ($\sim$40 Kbp each) generated from the genome of the same individual (NA18507) using the traditional Sanger method and built a map of novel insertions with high quality sequence information (Kidd *et al.*, 2008). We used this dataset to test the accuracy of the NovelSeq pipeline. As shown in Table 1, we anchored $>$70% of the orphan contigs (with high sequence identity to a novel sequence insertion detected by fosmids) to locations concordant with the fosmid-based predictions. Our concordance rate increases to 78% for ABySS contigs of length $\geq$500bp. Note that some of the fosmid sequences were not anchored to the human genome reference assembly, thus we were not able to test the accuracy of the loci we predicted for the contigs that are highly identical to such fosmid sequences.

We need to re-emphasize that anchoring a novel insertion is not an easy task if there are repeat sequences (that also are not represented in the reference genome) at the flanks of the inserted sequence. Note that the dataset used here is generated by the Illumina platform and the insert size is very small (average size 209 bp, standard deviation 8.25 bp). Any anchoring strategy that utilizes the OEA concept would fail to do so in such cases, since the OEA read pair will be too short to span over the flanking repeat if the repeat length is larger than the insert size (for example an Alu element is typically 300bp). For a more reliable OEA/orphan anchoring step, longer insert sizes are required.

[4] We considered the OEA clusters supported by $\geq$10 OEA reads in both strands, where $\geq$20 OEA reads were required to support the cluster in at least one strand.

| **NA18507** | # merged | | same locus | | different locus | |
|---|---|---|---|---|---|---|
| minimum length | $500bp$ | $200bp$ | $500bp$ | $200bp$ | $500bp$ | $200bp$ |
| ABySS | 78 | 113 | 37 | 50 | 10 | 21 |
| EULER | 85 | 130 | 35 | 51 | 14 | 23 |

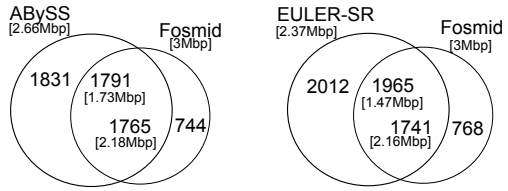
**Table 1.** This table shows two different result sets depending on the minimum length of the orphan contigs considered for the merging phase. For both ABySS and EULER contigs, we show the number of orphan contigs that are merged with an OEA contig (and hence anchored) with an alignment score $\geq$50. *Same locus* (table header) indicates the number of orphan contigs with high sequence identity to a novel insertion sequence detected by fosmids and loci in concordance with the fosmid-based predictions. *Different locus* (table header) indicates the number of orphan contigs with high sequence identity to a novel insertion sequence detected by fosmids but with loci not in concordance with the fosmid-based predictions.

## 4.2 Comparison of the orphan contigs with the NA18507 fosmid shotgun sequence library

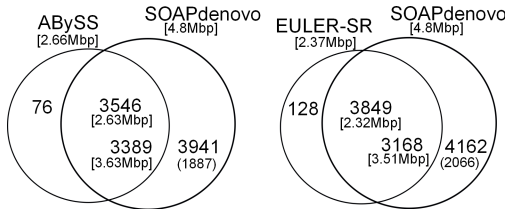
We compare the sequence content of both ABySS and EULER contigs with a set of $2,509$ sequence contigs assembled from one end anchored fosmid end sequences as previously described by Kidd et al. (Kidd *et al.*, 2008). This fosmid resource was end-sequenced using capillary technology, and in the remainder of this paper, we denote the sequence assembly generated from this dataset as *fosmid contigs*. Using cross_match (Green, 2010) with default parameters, we observed that $1,789$ ($\sim$71%) fosmid contigs overlap with the ABySS contigs, and $1,754$ ($\sim$70%) fosmid contigs overlap with the EULER contigs. Figure 4(a) shows the comparison between ABySS and EULER contigs against the fosmid contigs. Next, we compared both ABySS and EULER orphan contigs with a total of 4.8 Mbp of novel sequence in NA18507 sequences found by a recent study by Li et al. (Li *et al.*, 2009) (n=7,330; Figure 4(b)) based on whole-genome *de novo* sequence assembly using SOAPdenovo (Li *et al.*, 2010). The reader can easily verify that *de novo* sequence assembly using the entire next-generation shotgun sequence read library requires extensive computational resources that are not needed by our method. The high amount of overlap between ABySS and EULER contigs with the contigs found by Li et al. (Li *et al.*, 2009) also validates the sequence content of ABySS and EULER contigs. Figure 4(c) depicts the comparison between ABySS and EULER contigs and the SOAPdenovo(Li *et al.*, 2009) and fosmid contigs.

Note that a close inspection of the sequences detected by SOAPdenovo and missed by ABySS and EULER revealed that 2,054 contigs missed by ABySS and 2,096 contigs missed by EULER are $<$200bp, that we removed from consideration in our analysis. We further analyzed the contigs found by SOAPdenovo and missed by ABySS, and using BLAST, we found that 119 contigs can be aligned to sequences from known contaminants (the majority to Epstein-Barr) with $>$90% sequence identity, requiring at least 80 bp alignment length (total of 136 Kbp). 97/119 contigs are $>$200 bp, the longest contig is 6,765 bp. Note that when we used blast, with paremeters identical to the ones used for the analyes of ABySS

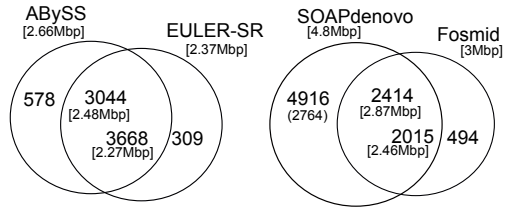
and EULER contigs, only 92 of SOAPdenovo contigs were aligned to either build35 or build36.




(a) The comparison of ABySS and fosmid contigs (left), and the comparison of EULER and fosmid contigs (Kidd *et al.*, 2008) (right)




(b) The comparison of ABySS and SOAPdenovo contigs (Li *et al.*, 2009) (left), and the comparison of EULER and SOAPdenovo contigs (right)




(c) The comparison of ABySS and EULER contigs (left), and the comparison of SOAPdenovo (Li *et al.*, 2009) and fosmid contigs (Kidd *et al.*, 2008) (right)

**Fig. 4.** Venn diagrams depicting pairwise comparisons of novel sequence assemblies generated by ABySS, EULER, SOAPdenovo (Li *et al.*, 2009, 2010), and fosmid end-sequences using *phrap*. Note that we provide two numbers at the intersections, corresponding to the numbers of contigs in each set that are almost identical to the contigs in the reciprocal set. We also provide the total length of those contigs in brackets. The numbers given in parenthesis, next to SOAPdenovo, correspond to the number of contigs with at least 200bp.

### 4.3 Comparison with WGS libraries and the Venter genome

Finally, we used BLAST to compare the contaminant-free orphan contigs generated by ABySS (n=4,115) and EULER (n=4,525) with the WGS library generated from the genome of the same individual (NA18507) using Sanger sequencing, WGS library generated from the genome of Craig Venter (Levy *et al.*, 2007), as well as the sequence assembly of the Venter genome (HuRef (Levy *et al.*, 2007). In Table 2 we also provide comparisons against human genome reference assembly (both build35 and build36). We consider two category of 99% and 95% sequence identity to call a hit in the database search. In addition, we provide the comparison statistics separated by the minimum contig length (i.e. $\geq$200bp and $\geq$500bp). We observe that the novel sequences detected in NA18507 genome are also found in the Venter genome, suggesting that these sequences correspond to rare deletions in the reference genome assembly.

| **NA18507** | | $\geq 200bp$ | | $\geq 500bp$ | |
|---|---|---|---|---|---|
| | **database** | 95% | 99% | 95% | 99% |
| ABySS | build35 | 616 | 481 | 236 | 174 |
| ABySS | build36 | 611 | 475 | 222 | 159 |
| ABySS | NA18507 (fosmid end-seq.) | 2305 | 1944 | 1253 | 1076 |
| ABySS | Venter WGS | 3028 | 2938 | 1811 | 1798 |
| ABySS | HuRef | 3815 | 3763 | 1512 | 1488 |
| EULER | build35 | 670 | 530 | 123 | 100 |
| EULER | build36 | 660 | 522 | 114 | 92 |
| EULER | NA18507 (fosmid end-seq.) | 2530 | 2169 | 1055 | 933 |
| EULER | Venter WGS | 4193 | 4131 | 1542 | 1536 |
| EULER | HuRef | 3272 | 3127 | 1329 | 1309 |

**Table 2.** The comparison of NA18507 orphan contigs with the WGS libraries and the Venter genome. For different cases, the number of orphan contigs with a high similarity to each library is given. Contigs that were aligned to build35 or build36 were also included.

## 5 DISCUSSION AND CONCLUSION

The completion of the Human Genome Project in 2003 was a major step towards understanding our genetic makeup. Although it is invaluable for genome research, the reference genome assembly is not a global representative of all haplotypes and a host of human genome sequences remain missing. Due to the cost of traditional sequencing technologies, the characterization of such sequences, commonly referred to as "novel insertion sequences" (or alternatively deletion alleles in the reference genome) remained elusive. However, with the introduction and continuous evolution of next-generation sequencing technologies, it is now possible to detect and characterize these sequences in the hopes of building a human "pan-genome" (Li *et al.*, 2009). *De novo* sequence assembly methods (Simpson *et al.*, 2009; Chaisson and Pevzner, 2008; Li *et al.*, 2010) are developed to address the computational challanges of this issue, however, one needs to invest significantly in computational resources due to the excessive memory and CPU requirements of such methods. We designed our pipeline, NovelSeq, to efficiently assemble the novel sequence insertions and build maps of insertion by anchoring the sequences back into the reference genome assembly. An important aspect of our framework is that it can be applied as a post-processing step after the completion of read mapping to analyze other types of genetic variation such as SNP and structural variation discovery. We validated our predictions

by comparing the sequence content and the anchor position independently assessed by other groups using (i) fosmid end sequence analysis, (ii) full fosmid sequencing, and (iii) *de novo* sequence assembly using data generated from the genome of the same individual. In addition, we compared the sequence content of our novel sequence predictions with the WGS dataset and the assembly of the Venter genome. The significant overlap between the sequences detected in two different genomes suggest rare deletions in the reference genome.

To better understand the human genome variation and evolution, as well as genotype-phenotype associations, we need to build comprehensive genome assemblies. The next-generation sequencing platforms now give us the opportunity to target genomes from many populations, as exemplified by the 1000 Genomes Project (`http://www.1000genomes.org`). The next challenge will be the full characterization of these "novel insertions" to discover new promoters, exons, and other functional elements.

## ACKNOWLEDGMENTS

We would like to thank S. Jackman for helpful discussions with respect to ABySS, T. Brown, L. Brunner, and D. Yorukoglu for their help in manuscript preparation, and the anonymous reviewers for their comments. We would greatly acknowledge T. Marques for the help on wet-lab validation. This work was supported, in part, by the Natural Sciences and Engineering Research Council of Canada (NSERC), Bioinformatics for Combating Infectious Dieseases (BCID), Genome BC, and Michael Smith Foundation for Health Research grants to S.C.S, and U.S. National Institutes of Health grant HG004120 to E.E.E. E.E.E. is an investigator of the Howard Hughes Medical Institute.

## REFERENCES

Alkan, C., Kidd, J.M., Marques-Bonet, T., Aksay, G., Antonacci, F., et al. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics* **41:** 1061–1067.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J. Molec. Biol.* **215:** 403–410.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59.

Chaisson, M.J. and Pevzner, P.A. (2008) Short read fragment assembly of bacterial genomes. *Genome Res.* **18:** 324–330.

Chen, K., Wallis, J., McLellan, M., Larson, D., Kalicki, J., et al. (2009) Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6:** 677 – 681.

Ewing, B., Green, P. (1998) Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome Res* **8:** 186–94.

Green, P. (2010) cross-match, at http://www.phrap.org.

Hormozdiari, F., Alkan, C., Eichler, E.E., Sahinalp, S.C. (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* **19:** 1270–1278.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., et al. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453:** 56–64.

Kidd, J.M., Sampas, N., Antonacci, F., Graves, T., Fulton, R., et al. (2010) Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature Methods*, **to appear**.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420–426.

Lee, S., Hormozdiari, F., Alkan, C., Brudno, M. (2009) Modil: detecting small indels from clone-end sequencing with mixtures of distributions. *Nature Methods* **6:** 473 – 474.

Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H. *et al.* (2009) Building the sequence map of the human pan-genome. *Nature Biotech.* **28:** 57–63.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X. *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 265–272.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., et al. (2007) The diploid genome sequence of an individual human.*PLoS Biol* **5:** e254.

Medvedev, P., Stanciu, M., Brudno, M. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* **6:** 13–20.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J. E., Jones, S.J., M., and Birol, I. (2009) ABySS: A parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123.

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

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., et al. (2005) Fine-scale structural variation of the human genome. *Nat Genet* **37:** 727–32.

Volik, S., Zhao, S., Chin, K., Brebner, J.H., Herndon, D.R., et al. (2003) End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc Natl Acad Sci U S A* **100:** 7696–7701.

West, D.B. (2001) Introduction to Graph Theory (2nd Edition). Prentice Hall.