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 (NGS) 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 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 (Kidd et al., 2008; Korbel et al., 2007; Tuzun et al., 2005; Volik et al., 2003), one cannot discover insertions >100 bp when 200 bp insert size is used with the Illumina platform (Bentley et al., 2008; Chen et al., 2009; Hormozdiari et al., 2009). Currently, the only method applicable for the discovery of long novel insertions using NGS technologies is de novo sequence assembly (Chaisson and Pevzner, 2008; Li et al., 2010; Simpson et al., 2009). 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. We test our methods with the high-coverage (42×) 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 whole-genome shotgun (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

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 >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\(^1\) sequences can be 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 (Fig. 1a).

Orphan assembly and contamination removal: using available de novo assembly algorithms such as EULER-SR

\(^1\)Each end sequence of a paired-end read is referred to as end-read.
Anchoring orphan contigs using the OEA contigs

The local assembly of the OEA clusters

The OEA cluster, the goal is to assemble the unmapped reads in each mrSAAB (micro-read Strand-Aware Assembly Builder). For each end-reads in the OEA clusters that were created in the previous step, we both provide more read support for the orphan contigs and 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 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 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 article, 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 orientation of the mapped end-read. 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 ⊂ R. The end-reads in OEA can also be mapped to multiple locations on the reference genome. For all pe ∈ OEA, alignment of pe is defined as ape = (lok(ape), or(ape)), where lok(ape) is the map location and or(ape) ∈ {+, −} 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 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 ⊂ OEA supports the same insertion if the following conditions hold:

- For every pair of OEA read alignments pe and pe′ ∈ clu with pe and pe′ to the reverse strand, the map location of pe is proximal to the map location of pe′.
- The maximum pairwise distance between the map locations of the OEA reads in clu with the same mapping orientation must be less than the maximum InsSize, Δmax.
- The distance between the map locations of two OEA reads with different mapping orientations should not exceed twice the maximum InsSize, 2Δmax.

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 2Δmax + 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 contains 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 $+ \text{ strand}$ and the reads with mates that map to the $- \text{ strand}$ 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.\footnote{Personal communication with the developers of these tools.}

Using single end-reads, both ABySS and EULER consider the reverse complements of the read sequences as well. In each cluster, the OEA reads with mates that map to the 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 unmapped reads in the local OEA assembly. The confidence interval for this position information depends on the $\text{InS} \text{ize}$ 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 $u$ and $v$ 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 (or $+ \text{ and } - \text{ contigs}$) if the score of the prefix–suffix match between the two ends of the orphan contig and of $+ \text{ and } - \text{ contigs}$ 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 $\text{Orph}_h = \{o_r, o_r, \ldots, o_r\}$ be a set of orphan contigs and $\text{OEA}_A = \{o_{rA}, o_{rA}, \ldots, o_{rA}\}$ be a set of OEA contigs where $o_{rA}$ is a pair of two OEA contigs from the local assembly of the OEA cluster with $ij$ (i.e. $o_{rA} = (o_{rA}, o_{rA})$). We aim to assign each element in $\text{Orph}_h$ (e.g. $o_r \in \text{Orph}_h$) to an element in $\text{OEA}_A$ (e.g. $o_{rA} \in \text{OEA}_A$) such that the summation of (i) the alignment score between the prefix of $o_r$ and the suffix of $o_{rA}$, and (ii) the alignment score between the suffix of $o_r$ and the prefix of $o_{rA}$ 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 (Fig. 2):

- $\forall u_r \in \text{Orph}_h, \exists v_r \in U$
- $\forall o_{rA} \in \text{OEA}_A, \exists v_{rA} \in V$
- The weight of edge $(u_r, v_{rA})$ is a function of the overlap between the first $\Delta_{\text{first}}$ base pairs of $o_r$ and $o_{rA}$, and the overlap between the last $\Delta_{\text{last}}$ base pairs of $o_r$ with $o_{rA}$.

4 EXPERIMENTAL RESULTS

We tested our framework using the 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 (Alkan et al., 2009; Chen et al., 2009; Hormozdiari et al., 2009; Lee et al., 2009) to discover structural variation and copy number polymorphism. This dataset contains $\sim$3.5 billion sequence reads ($\sim$1.7 billion pairs) of length 36–41 bp with an $\text{InS} \text{ize}$ of $\sim$209 bp (Bentley et al., 2008; Hormozdiari et al., 2009). The $\text{InS} \text{ize}$ distribution of this dataset was previously presented in Hormozdiari et al. (2009).
ABySS contigs of length \( \geq \) size we assembled the orphan paired-end reads into 4154 contigs of
Using ABySS (Simpson et al., 2009), 2009), allowing for edit distance \( \leq 2 \). Note that mrFAST returns all possible map locations of real sequences, thus an OEA read can be aligned to multiple locations in the reference genome. In total, 15173562 pairs of reads (30347124 end-sequences) were identified as orphans, while 83662700 reads were identified as OEA.

4.1 Novel sequence insertion map

4.1.1 Preprocessing Similar to the prescreening methodology used in Hormondian et al. (2009), we removed any paired-end reads from consideration if either (or both) end sequence has an average quality value less than 20, or if either (or both) sequence contains more than 2 unknown (i.e. \( N \)) nucleotides.

4.1.2 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), 2009), allowing for edit distance \( \leq 2 \). Note that mrFAST returns all possible map locations of real sequences, thus an OEA read can be aligned to multiple locations in the reference genome. In total, 15173562 pairs of reads (30347124 end-sequences) were identified as orphans, while 83662700 reads were identified as OEA.

4.1.3 Orphan assembly Using ABySS (Simpson et al., 2009), 2009), we compared the orphan contigs with the nt database (http://www.ncbi.nlm.nih.gov/nt) 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 \( \geq 2000 \) after the contamination removal. Note that out of 4115 ABySS contaminant-free contigs (\( \geq 200 \) bp), 1984 are \( \geq 500 \) bp and 778 are \( \geq 1 \) kb in size. Among the EULER contaminant-free contigs, 1690 are \( \geq 500 \) bp and 582 are \( \geq 1 \) kb.

![Fig. 3. Length distribution (log scale) of the ABySS (red) and EULER (green) contigs (\( \geq 200 \) bp).](image)

4.1.4 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 (http://www.ncbi.nlm.nih.gov/nt) 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 \( \geq 2000 \) after the contamination removal. Note that out of 4115 ABySS contaminant-free contigs (\( \geq 200 \) bp), 1984 are \( \geq 500 \) bp and 778 are \( \geq 1 \) kb in size. Among the EULER contaminant-free contigs, 1690 are \( \geq 500 \) bp and 582 are \( \geq 1 \) kb.

![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 phmap. 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 one strand.](image)

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. Of ABySS contigs, 493 of length \( \geq 200 \) bp could be mapped onto either build35 or build36 with \( \geq 99 \% \) 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 (\( n = 3622 \), Fig. 4a) had a total length of 2.66 Mb, while the remaining EULER contigs (\( n = 3977 \), Fig. 4a) had the a total length of 2.37 Mb of the sequence.

4.1.5 OEA clustering and orphan anchoring We used our clustering algorithm followed by the set cover approach to cluster the OEA reads, and obtained 10560 sets of OEA clusters with a high support\(^1\) 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

\(^{1}\)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.
Table 1. Two different result sets depending on the minimum length of the orphan contigs considered for the merging phase

<table>
<thead>
<tr>
<th>NA18507 # merged Same locus Different locus</th>
<th>ABySS</th>
<th>EULER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum length (bp)</td>
<td>500</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>200</td>
</tr>
</tbody>
</table>

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 ≥ 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.

We compare the sequence content of both ABySS and EULER insertions with high-quality sequence information (Kidd et al. 2009). The reader can easily verify that de novo 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. (2009) also validates the sequence content of ABySS and EULER contigs. Figure 4c 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 2054 contigs missed by ABySS and 2096 contigs missed by EULER are < 200 bp, which 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 kb). Of the 119 contigs, 97 are > 200 bp, the longest contig is 6765 bp. Note that when we used blast, with parameters identical to the ones used for the analyses of ABySS and EULER contigs, only 92 of SOAPdenovo contigs were aligned to either build35 or build36.

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 = 4415) and EULER (n = 4525) 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. ≥ 200 and ≥ 500 bp).

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.

5 DISCUSSION AND CONCLUSION

The completion of the HGP 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 NGS 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 (Chaisson and Pevzner, 2008; Li et al., 2010; Simpson et al., 2009) are developed to address the computational challenges 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.

To better understand the human genome variation and evolution, as well as genotype–phenotype associations, we need to build comprehensive genome assemblies. The NGS 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.

**ACKNOWLEDGEMENTS**

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 also thank T. Marques-Bonet for helpful discussions. E.E.E is an investigator of the Howard Hughes Medical Institute.

**Funding** Natural Sciences and Engineering Research Council of Canada (NSERC to S.C.S. in parts); Bioinformatics for Combating Infectious Diseases (BCID to S.C.S. in parts); Genome BC (to S.C.S. in parts); Michael Smith Foundation for Health Research grants (to S.C.S. in parts); US National Institutes of Health (grant U01HG005209-02 to E.E.E.) (grant HG004120 to E.E.E.).

**Conflict of Interest** EEE is an SAB member of the Pacific Biosciences.

**REFERENCES**


