Genome Completion by SOLiD™ System Next-Generation Sequencing

Dumitru Brinza, Jorge Duitama, Fiona C Hyland, Eugene G Spier, Asim Siddiqui, Francisco De la Vega, Ellen Beasley, Life Technologies, 850 Lincoln Center, Foster City, CA, 94404

ABSTRACT

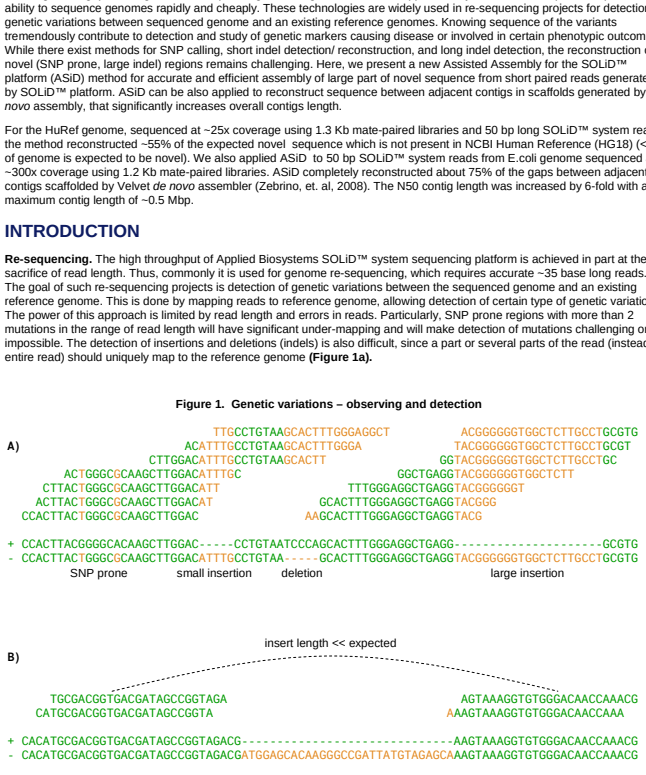
Ultra-high throughput next-generation sequencing technologies, such as the Applied Biosystems SOLiD™ platform, provide the ability to sequence genomes rapidly and cheaply. These technologies are widely used in re-sequencing projects for detection of genetic variations between sequenced genome and an existing reference genomes. Knowing sequence of the variants tremendously contribute to detection and study of genetic markers causing disease or involved in certain phenotypic outcomes. While there exist methods for SNP calling, short indel detection/ reconstruction, and long indel detection, the reconstruction of novel (SNP prone, large indel) regions remains challenging. Here, we present a new Assisted Assembly for the SOLiD™ platform (ASiD) method for accurate and efficient assembly of large part of novel sequence from short paired reads generated by SOLiD™ platform. ASiD can be also applied to reconstruct sequence between adjacent contigs in scaffolds generated by *de novo* assembly, that significantly increases overall contigs length.

For the HuRef genome, sequenced at ~25x coverage using 1.3 Kb mate-paired libraries and 50 bp long SOLiD™ system reads, the method reconstructed ~55% of the expected novel sequence which is not present in NCBI Human Reference (HG18) (<2% of genome is expected to be novel). We also applied ASiD to 50 bp SOLiD™ system reads from E.coli genome sequenced at ~300x coverage using 1.2 Kb mate-paired libraries. ASiD completely reconstructed about 75% of the gaps between adjacent contigs scaffolded by Velvet *de novo* assembler (Zebrino, et. al, 2008). The N50 contig length was increased by 6-fold with a maximum contig length of ~0.5 Mbp.

INTRODUCTION

Re-sequencing. The high throughput of Applied Biosystems SOLiD™ system sequencing platform is achieved in part at the sacrifice of read length. Thus, commonly it is used for genome re-sequencing, which requires accurate ~35 base long reads. The goal of such re-sequencing projects is detection of genetic variations between the sequenced genome and an existing reference genome. This is done by mapping reads to reference genome, allowing detection of certain type of genetic variations. The power of this approach is limited by read length and errors in reads. Particularly, SNP prone regions with more than 2 mutations in the range of read length will have significant under-mapping and will make detection of mutations challenging or impossible. The detection of insertions and deletions (indels) is also difficult, since a part or several parts of the read (instead of entire read) should uniquely map to the reference genome **(Figure 1a).**

Figure 1. Genetic variations – observing and detection



+ Reference sequence | - Sequenced sequence | green –match to reference | orange – mismatch to reference

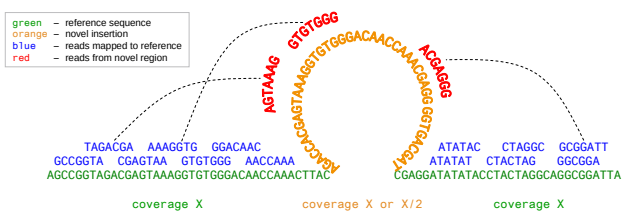
Detection of the above genomic events is simplified by availability of mate-paired reads, since both the fragments in the read and distance between them can be used to uniquely locate the read. In the the presence of mate-paired reads the genetic variations are detected by unique mapping of one of the fragments and ambiguous mapping, no mapping, or mate-paired distance violating mapping of another fragment **(Figure 1b).** The sequences of detected variations are novel comparing to reference, and cannot be reconstructed by mapping the fragments. The challenge is to de novo assemble the novel sequences (orange).

De novo assembly. Recently, it was demonstrated that de novo assembly of small genomes (, e.g., bacterial, fungus, microbial) from short reads generated by next-generation sequencing platforms is possible. The availability of mate-paired library reads allows us to order assembled contigs into scaffolds. Usually, scaffolds are spanning very large regions of the genomes providing sufficient information for large varieties of genomic studies. However, in certain studies, e.g., gene annotation, availability of continuous sequence (contig) is crucial. The challenge is to reconstruct sequence between adjacent contigs and merge them into one large contig.

MATERIALS AND METHODS

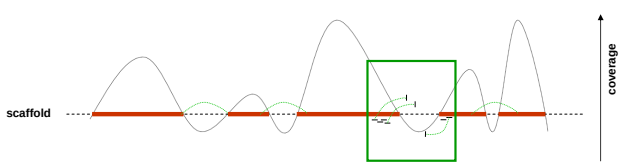
Recently, it was demonstrated that *de novo* reconstruction of low complexity small to medium size genomes (e.g., bacterial genomes) from the short reads generated by next-generation platforms is possible [1]. *De novo* assembly of repetitive or large size genomes remains challenging. To address this, we reduced *de novo* assembly to multiple local assemblies of short novel regions located in the vicinity of well mapped regions (in re-sequencing projects) or between two adjacent contigs (in de novo assembly projects). In re-sequencing projects assembly is built from the fragments of mate-paired reads for which one fragment belongs to well mapped region while another is expected (according to mate-pair information) to lie in the novel region. Prior to assembly, the set of fragments is de novo error corrected using the SOLiD™ Accuracy Enhancement Tool (SAET). SAET reduces the color calling error rate by a factor of 3-5, making it below 1% for SOLiD™ system reads. Correction results in more accurate de novo assembly and increases the length of assembled novel sequences by a factor of 2-3. Corrected fragments are assembled into long contigs using Velvet assembler [2].

Figure 2. ASiD for Re-Sequencing: Fishing reads for *de novo* assembly of novel insert



- Map reads to reference
- Collect hanging mates
- Error correct and de novo assemble reads from novel region
  - Robust error correction and assembly (due to low complexity of the region)

Figure 3. ASiD for De novo assembly: Fishing reads for *de novo* assembly of gapped regions



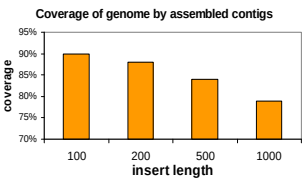
- Assemble scaffolds of entire genome using Velvet assembler
- Map all reads to scaffolds
- For each gap between two contigs collect hanging mates
- Error correct and assemble each subset of reads using low coverage cut-off
- Replace gap with uniquely assembled sequence

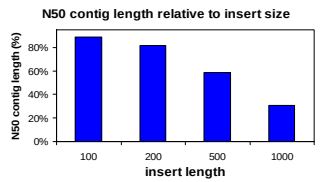
REFERENCES

[1] Chaisson MJ, Brinza D, Pevzner PA. 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Res. 19:336-46.
[2] Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821-9.

RESULTS

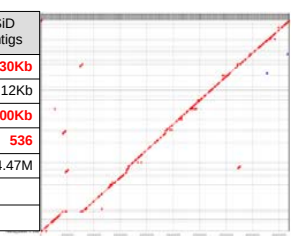
- **Re-sequencing: Reconstruction of novel sequence in human genome**
  - HuRef genome
  - 2x50 insert ~1.3 Kb, coverage (after mapping) 15x
  - hg18 is used as reference
  - Randomly select insertions of length ~100, 200, 500, 1000 (500 cases of each) in hg18 and recreate them using assisted assembly tool – test accuracy by comparing with known HuRef sequence
- **On average 85% of novel sequence is recovered**





- **De novo assembly: Reconstructing gaps between contigs**
  - E.coli length 4.6 Mbp
  - 2x50 insert ~1.3 Kb coverage 300x
- **75% of gaps are filled, 6 fold increase in N50 contig size**

E.Coli (4.6Mb) 2x50 300x coverage	Velvet Contigs	Velvet Scaffolds	ASiD Contigs
N50	**5.2Kb**	200Kb	**30Kb**
Mean contig length	3Kb	15Kb	12Kb
Max contig length	**23Kb**	520Kb	**500Kb**
Number contigs > 100 nt	**1522**	307	**536**
Sum contig length	4.45M	4.47M	4.47M
% of genome covered	97.24%		
Average identity	99.8%		



CONCLUSIONS

- Assisted de novo assembly can identify a large part of the novel content in large genomes sequenced with SOLiD™ system
- We have demonstrated that many reads that did not map to the reference genome could be rescued, error corrected, and assembled into a novel sequence
- Method recovers a large part of previously lost genetic variations; resulting in increased in TP SNP calls
- ASiD can be used for any genome resequencing (bacteria, eukaryotes) to identify insertions not present in the reference
- This method improves bacterial de novo by increasing N50 contig size by factor of 6, and making the quality of de novo assembly similar to the quality of the finished genome

TRADEMARKS/LICENSING