Finding indels with short sequencing reads and de novo assembly with SOLiD reads

Zheng Zhang, Joel Malek, Heather Peckham, and Jon Sorenson, Applied Biosystems, 850 Lincoln Centre Drive, Foster City, California 94404, USA

ABSTRACT

This poster presents algorithm and analysis that demonstrate that we can use short resequencing reads to find indels when we have large coverage and pairing information.

INTRODUCTION

In next generation shot gun sequencing many short reads are produced. One challenge is to find indels (insertion/deletion) in reference genome. If one maps those short reads to a large genome allowing indels, the false positive rate will be high. AB SOLiD sequencing technology includes the ability to generate paired reads whose approximate distances can be obtained (see Figure 1). For large size indel, the deviation from the average distance can be used to derive the existence and size of an indel (Figure 2). Here we focus on finding small to medium size indels using paired reads and alignment directly.

ALGORITHM AND ANALYSIS

It is a 2-step algorithm: (1) Map all reads to the reference genome allowing only small number of mismatches (0-2); (2) In paired reads A and B, if A can be uniquely mapped, but not B, we align B to the small region that is the correct distance from A, allowing one indel in the middle of the reads and 0 or 1mismatch. See Figure 3 for detail.

A detailed statistical analysis based on assumption of random sequence and Baysian theory (whose detail will be presented elsewhere) shows that, with high sequencing coverage rate, and reasonable sequencing error rate, the above algorithm can find almost all medium size (up to 100 bps) deletion and small size insertion (up to 5-8bps) with very low false positive rate.

To analyze FP in more realistic setting, we did some simulations. The result is quite close to the statistical analysis. For insertion, the FP is about 0.01%, and for deletion size up to 100, the FP is 0.007%. For FN, our simulation shows that at 5% raw sequencing error rate, FN rate is 0.0043% at 50X coverage. In summary the result is close to the theoretical result we have found.

For large insertion, we also design a mini-assembly program (see Figure 4). It is a simple assembly tool which use perfect match of at least N base as the criterion for overlap of reads. Then we find the shortest path as our assembly.

Our simulation show that with random reads with uniform distribution on the reference genome, at high coverage (100X), we can find 97% of insertion of up to 1000bp long. At 50X coverage, we can still find 85% of insertion of 500 bp long.

Fragment library with longer read length

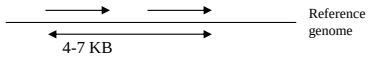
If people have only unpaired reads, one way to find indels is to use coverage map to find locations of possible indels, and then use our indel alignment tool on those regions. When one has longer reads with length up to 50-60 bases, it is possible to do indel finding in the whole genome very fast. The idea is to first search the genome using the first 15 bases without indel (may allow 1 mismatch), then perform indel alignment on the regions anchored by the 15mer hits (Figure 5).

De novo assembly using SOLiD reads

Several groups are designing de novo assembly tools for short sequence reads. The two base coding of SOLiD reads makes it harder to use these tools directly to assemble SOLiD reads. We developed an adaptor for these tools, so people can use generic assembly tools (designed for assembling base sequences) to assemble SOLiD reads and get high quality sequences. Formally, the procedure includes the following steps (Figure 6):

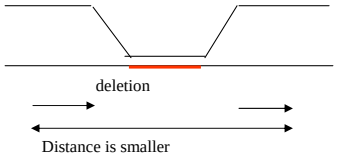
•Get a generic assembler to assemble color reads (treat color code as bases);
•Take the resulting assembly along with its alignments of reads, and use our adaptor to translate it into a nucleotide sequence;
•(Option) Map other reads to the assembly use SOLiD mapping tool, and repeat the previous step.

Figure 1. Paired reads



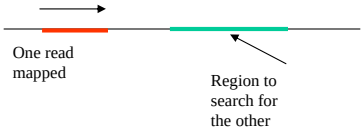
Paired reads are certain distance away from each other. The range of the distance is usually a few Kb. The numbers listed for illustration purpose only.

Figure 2. Large indel size can be derived



A large deletion (red line) affect the sequencing coverage, and paired reads shall consistently shorter distance.

Figure 3. Paired reads help find to find indels

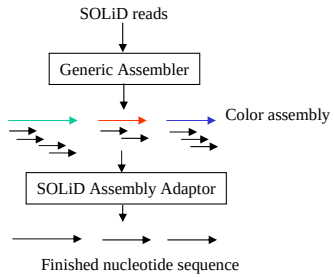


If one read of a pair can be mapped to a place on the reference, one can narrow down the search of the other to a much smaller region.

Figure 5. Indel finding with unpaired reads

The blue lines are the 15mer hits to a read. The indel alignment program will be performed on the red regions.

Figure 6. De novo assembly using SOLiD reads: Generic assembler with adaptor



We built a SOLiD assebly adaptor so any generic assembler can be used to assemble SOLiD reads.

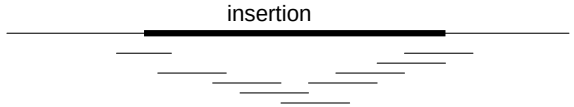
Table 1. Yoruban data set.

#of alignment	Deletion (% match to dbSNP)	Insertions (% to dbSNP)	Total (% to dbSNP)
1+	21966(60%)	32496(45%)	54462(51%)
2+	5821(65%)	9166(51%)	14987(57%)
3+	1730(64%)	3109(53%)	4839(57%)
4+	564(61%)	1107(53%)	1671(56%)

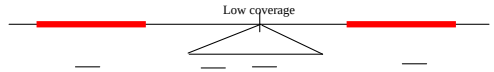
In this table, we list the number of insertions and deletions we predicted in Yoruban data set. The first column is the number of the minimal number of alignments required to predict an indel. The numbers in () list the % of the indels in each group that match one in the dbSNP. The data set has a 4X coverage of the whole human genome.

©Copyright 2008. Applied Biosystems. All rights reserved.

Figure 4. Mini-assembly to find large insertion.



We use a standard assembly routine that requires perfect match of at least N bases for overlap, and then look for a shortest path in the overlap group of reads.



Coverage map and paired reads info can be used to obtain the set of reads that may overlap an insertion.

EXPERIMENT RESULTS

Experiment 1, Real reads and artificial random long insertions.

This experiment is to test how effective can mini-assembly find larger insertions. We sequenced S. Suis genome for 100X. Then we mutated the reference by delete long stretch of DNA, and use the mutated sequence as reference, try to recover the deleted sequence (as insertions) using the set of reads. This estimate the FN for mini-assembly in real setting.

This set show that we can find around 50% of the insertions of up to 200 bp long.

Experiment 2, Real reads to find short indels.

We sequenced one 500KB of ENCODE region. There are two pools of reads. Below we list the number of indels found by our program and the number among them that are matched to some indel in dbSNP.

	#indels	#match dbSNP
Pool 1	**55**	**13**
Pool 2	**138**	**29**

Since we look for high quality alignments and make prediction only if there are many supporting alignments, many of the indel not in dbSNP may well be real (esp. since dbSNP under-represents indels).

Experiment 3, Yoruban data set

We sequenced whole human genome to 4X coverage. The goal is to test out small indel finding program to estimate a FP rate for the procedure. We pulled out a comprehensive set of human indels reported in the dbSNP. In summary, if we require 2 alignments matching to the same place to confirm an indel, our program predicted about 15000 indels (deletions up to 10 bases, and insertions up to 3), and 57% of them matched to some reported ones in dbSNP. If we require only one alignment to make the prediction, we reported more than 54000 indels, and 51% of them matching some one in dbSNP. More details in Table 1. Given that dbSNP is far from listing all indels, our FP rate shall be well lower. Also note that there are little difference in the % matching to dbSNP among indels with different numbers of alignments. This seems to suggest that the FP is really low.

CONCLUSION

In conclusion, using only alignments, we can find medium deletions and small insertions in large genome. Experiments we did show it worked very well. And for large insertions, mini-assembly program we implemented works very well in simulated data, and reasonably well for real data.

ACKNOWLEDGEMENTS

We thank Eugene Spier, Pius Brzoska, Heinz Breu for helpful discussion; Stephen McLaughlin for pulling out relevant dbSNP indels for experiment; AB Beverly group for running the experiments.