

de novo Assembly of Short Sequence Reads with NextGENe™ Software

Jacie Wu, Kevin LeVan, Teresa Snyder-Leiby, Haigou He, Haitao Ren, Igor Wojciechowski, Shouyong Ni, and ChangSheng Jonathan Liu

Introduction

Next generation DNA sequencing strategies have lowered the costs of sequencing, increasing the speed and amount of information gathered. It is common for an instrument to generate 3 billion bases in a couple of days for just a few thousand dollars. The Illumina® Genome Analyzer utilizing the Solexa sequencing technology, Applied Biosystem SOLiD™ System and the Helicos™ Genetic Analysis System from Helicos Biosciences Corporation give reliable sequence read-outs of 25-35 bps and about 5-100 million reads.

De novo sequence assembly with the short reads from the genome analyzers presents many challenges (1). With many of the current techniques, it is difficult to assemble the short reads into a large contig of 1 to 5 kb. These techniques often create many false alignments due to two major issues; short reads with high base calling errors and ambiguity within the genome. The short reads with SNPs and Indels are often discarded, which is problematic in the determination of copy number variations in applications such as chromatin immunoprecipitation (ChIP), gene expression and transcriptome studies.

NextGENe sequence assembler was developed to solve the current problems. The software is able to assemble the short reads into contigs of 0.5 kb to 5 kb, where contigs end with repeat sequences. It uniquely aligns these contigs to a reference genome. The short reads used in the assembly of a contig are recorded to show the copy number and Indel positions. NextGENe is capable of detecting Indels of 1-30 bps.

de novo Assembly Methodology

NextGENe statistically polishes high coverage (20-100x) datasets to remove random sequencing errors and roughly double the read lengths with the use of the Condensation Assembly Tool (Patent Pending). Repeating the Condensation removes systematic errors and further lengthens the sequence reads. The polished and elongated reads can then be assembled into large contigs while removing redundant reads.

The first step is utilizing the Condensation Assembly Tool to generate the first assembly. All of the reads with the same anchor sequence of 12 bps are collected into a cluster. The two shoulder sequences of 10 bps are used to sort the short reads into multiple groups. The consensus sequence in each group is obtained from the short reads. The ending bases are ignored from the consensus when the base has covered only one sequence read or inconsistency between multiple reads. The 5' sequence has higher weight than that of 3' end because of quality. With 50x coverage, confidence of the condensed sequence is about 99.8%. Then all of the possible anchor sequences with 16.7 million possibilities are calculated.

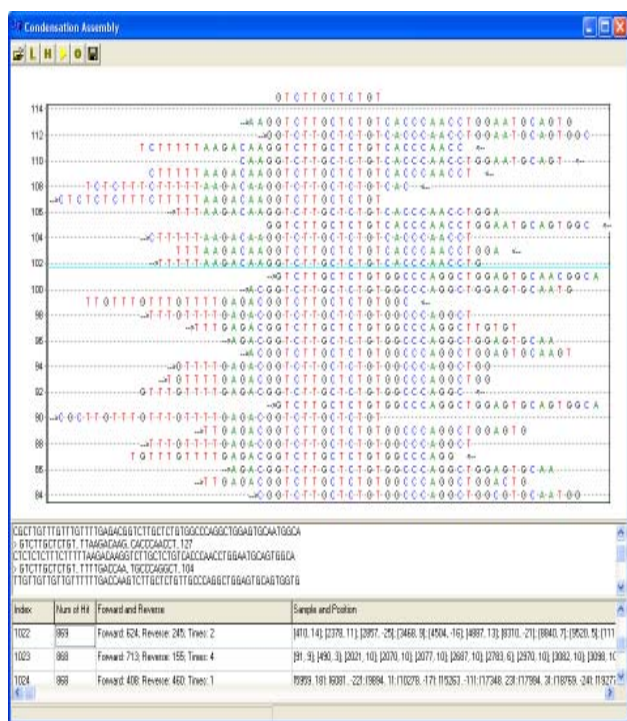


Figure 1: Condensation Assembly Tool elongated the 35 bp reads to approximately 60 bp while removing many of the random errors produced by the instrument.

Procedure

First Condensation

1. Open the Condensation Assembly Tool and click on the Open Folder button.
2. Click the Add button and choose the sample file.
3. Set the Load Sample Section value to the number of reads analyzed simultaneously. **NOTE:** Data input is limited to 3 million reads or 200 megabytes with a 32-bit Windows® system. Input size increases to 10 million short reads with a 64-bit Windows system with 8GB RAM.

4. Click the Options button and set options according to Figure 2.
5. Click the Save button.
6. When condensation is complete, a message will appear showing the start and end time of analysis. Click OK to view results.

NOTE: The Condensation Assembly Tool generates a condensed file that contains the elongated reads and an uncondensed file that contains all reads that were not used for elongation (often the reads containing errors and repeat sequences). Other files may also be created. Depending on the number of reads simultaneously analyzed, the condensed reads may be parsed into multiple files.

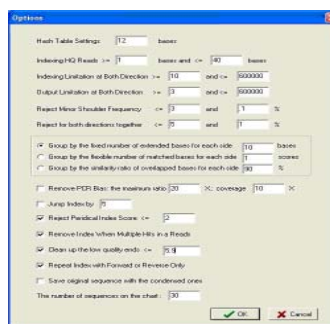


Figure 2: Options for first cycle of Condensation Assembly.



Figure 3: Options for remaining cycles of Condensation.

Additional Condensation Cycles

7. In the Condensation Assembly Tool, click the Add button and choose the Condensed sample file produced by the previous cycle.
8. Click the Options button and set options according to Figure 3.
9. Click the Save button and name the file in a manner recognizable as cycle number.
10. Repeat the Additional Condensation Cycle steps for a total of four additional cycles.

Generate Large Contigs

11. Assemble the highly similar sequences into larger contigs and reduce the number of redundant sequences with the proprietary algorithm of NextGENe software.

Results

We analyzed data supplied by Professor Jim Knowles, University of South California, of a patient who has a defective x chromosome. Within a 250 kb region, an additional few kilobases are found when using slab gel. Using the Condensation Assembly to make contigs, 5 cycles extended the short reads to 0.5 – 1 kb. The final contig distribution is shown in Figure 4.

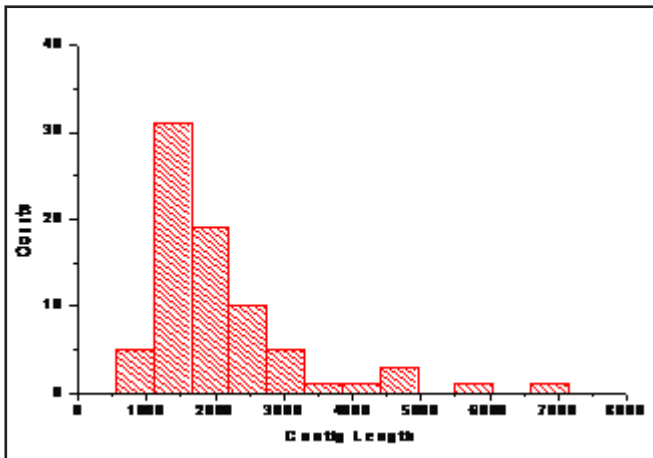


Figure 4: Contig distribution after sequence assembly. The short sequence reads of 36 bps were condensed and elongated using 5 cycles of the Condensation Assembly Tool. The median contig length is about 2.5 kb.

Discussion

When using template DNA that is double-stranded, regions of the genome often show a directional bias. It is thought that the number of forward and reverse short sequence reads should be balanced. This imbalance has been shown to reach a 100-5000 ratio ($N_{\text{forward}}/N_{\text{reverse}}$). It is possible that this imbalance is the result of cluster amplification biases from the adaptors in the Solexa technique, primer or sample impurities, or PCR bias. In cases where these biases exist, a large portion of the imbalanced reads can be rejected, improving the analysis efficiency and accuracy.

Repeat sequences are problematic in assembly of short reads. NextGENe analyzes the collection of reads containing the same repeat and aligns them at the 5' end of the repeat sequence. Reverse sequences are treated in the same manner, helping to accurately identify both sides of the repeat.

The use of next-generation sequencing in over 100 articles in less than two years indicates that the development of software to analyze these mega data sets will be increasing in demand. *De novo* synthesis using NextGENe provides an innovative, accurate tool to solve problems associated with short reads from these genome analyzers. The software is compatible with data files from Illumina Genome Analyzer, the Applied Biosystem SOLiD™ System and the Genome Sequencer FLX System from Roche Applied Science.

Acknowledgements

We would like to thank Victor Velculescu, Kimmel Cancer Center, Johns Hopkins University, Zemin Deng and James Knowles, Keck School of Medicine, University of Southern California, for their collaboration in the development of this software.

References

1. Jonathan Butler et al. *de novo* assembly of whole-genome shotgun microreads. Genome Research. March 2008.

Additional Application Notes for the Analysis of “Next Generation Data” with NextGENe Software:

- **Analysis of Digital Gene Expression Using Short Sequence Reads with NextGENe Software.**
- **SNP and Micro Indel Detection with NextGENe Software.**
- *de novo* **Assembly of Short Sequence Reads with NextGENe Software.**
- **Transcriptome Analysis Using NextGENe Software.**
- **NextGENe Software Tools for Analysis of Protein-DNA Interactions by ChIPSeq.**

Request a copy today: info@biogene.com