NextGENe™
by SoftGenetics
February 2009

de novo Assembly of Short Sequence Reads with NextGENe™ Software & Condensation Tool™

Megan Manion, Haitao Ren, Yiqiong Jacie Wu, Kevin LeVan, Shouyong Ni, and Changsheng Jonathan Liu

Introduction

Utilization of 2nd generation sequencing systems are lowering sequencing costs, while dramatically increasing the speed and amount of information gathered. It is common for a single instrument to generate 3 billion bases in a couple of days for around few thousand dollars. The Illumina® Genome Analyzer utilizing the Solexa sequencing by synthesis technology, Applied Biosystems SOLiD™ System using sequencing by ligation and di-color tagging and the Genome Sequencer FLX System from Roche Applied Science (454 Sequencing) using pyrosequencing technology give reliable sequence read-outs of 25-75 bps for Illumina and SOLiD reads and up to 500 bp for Roche/454 reads at about 5-100 million reads per run.

de novo sequence assembly of the short reads from next generation 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 more than 1 kbps. 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 for SNP/Indel detection as well as for the determination of copy number variations in applications such as chromatin immunoprecipitation (ChIP), Digital Gene Expression studies (DGE) and transcriptome analyses.

NextGENe software was developed to assist researchers in resolving these inherent issues in analyzing next generation sequencing data. NextGENe's unique Condensation Tool™ is used to polish and lengthen short sequence reads into fragment sizes that are more unique and accurate. Short reads such as those from the Illumina Genome Analyzer System and the Applied Biosystems' SOLiD instrument are often not unique within the genome being analyzed. By clustering similar reads containing a unique anchor sequence, data of adequate coverage are condensed, the short reads are lengthened and reads containing instrument errors are filtered from the analysis.

Multiple cycles of the Condensation Tool can be used to increase read lengths up to 400 bps, preparing the reads for accurate de novo assembly. The short reads used in the assembly of a contig are recorded so that copy number and Indel position information is not lost and can be utilized for downstream analysis. NextGENe's novel Condensation Tool allows for the detection of Indels of 1-30 bps. (See SNP&INDEL application note)

The Assembly Tool is then used to assemble the short reads into contigs of 0.5 kb to greater than 100kbs.

NextGENe offers multiple methods for de novo assembly in order to provide accurate assemblies of reads from different instruments and in different forms (mate pair reads vs. single reads). The first method uses a de Bruijn graph technique (2). Ideal for short reads of Illumina and SOLiD platforms, this method is capable of utilizing paired read information to assist with proper assembly of large contigs, but can also be used without paired end data. This assembly method is quite memory intensive, and some large datasets require 32GB of RAM to assemble with the de Bruijn method. When choosing to use the de Bruijn method of assembly, the condensation steps are omitted.

NextGENe also offers an alternative method of assembly that is less memory intensive than the de Bruijn method and may be more ideal from some sets of data. When completing multiple cycles of condensation and coverage information is not needed, the Method 2 option can be used for assembly of large contigs.

For the longer reads of Roche/454 datasets, NextGENe offers a 454 Assembler. This assembly method is tailored to accurately assemble 454 reads. The error correction feature from the Condensation step can be used prior to assembly to correct homopolymer errors and improve accuracy and length of assembled contigs.

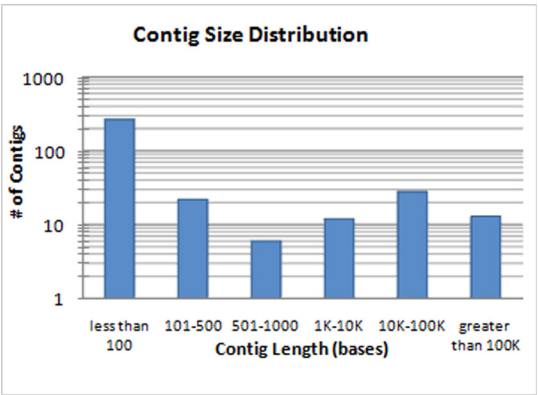

Figure 1

Figure 1: Contig size distribution after de Bruijn assembly with Illumina sequence reads of E. coli. The N50 contig length is 204665 bp with a maximum contig length of 560819 bp. The assembly produced 347 large contigs covering 4.6 Mbp.

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 Tool. Multiple Condensation cycles remove systemic errors and further lengthen the sequence reads. Different methods of condensation are available for varying datasets and applications.

All of the reads with the same anchor sequence of 12 bps are clustered. The two shoulder sequences of 10 bps are used to sort the read clusters into multiple groups. The consensus sequence of each group is obtained from the short reads. Ending bases are ignored from the consensus when the base has covered only one sequence read or inconsistency exists between multiple reads. 5' sequence has a higher weight than that of 3' end because of difference in base call quality. With 50x coverage, confidence of the condensed sequence is about **99.8%**. A short sequence read may be used multiple times in Condensation.

Multiple condensation cycles further lengthen the reads and reduce errors. After the first cycle, short reads are assembled to contigs that are typically double the original read length. After a few more cycles, the condensed reads are extended to about 400 bps.

These short contigs can then be assembled further by NextGENe's Assembly Tool, and, depending on number of repeats and quality of data, contigs of more than 100 Kbps can be generated.

Figure 2: After one cycle, the Condensation Tool elongated the 35 bp reads to approximately 60 bp while removing many of the random errors produced by the instrument.

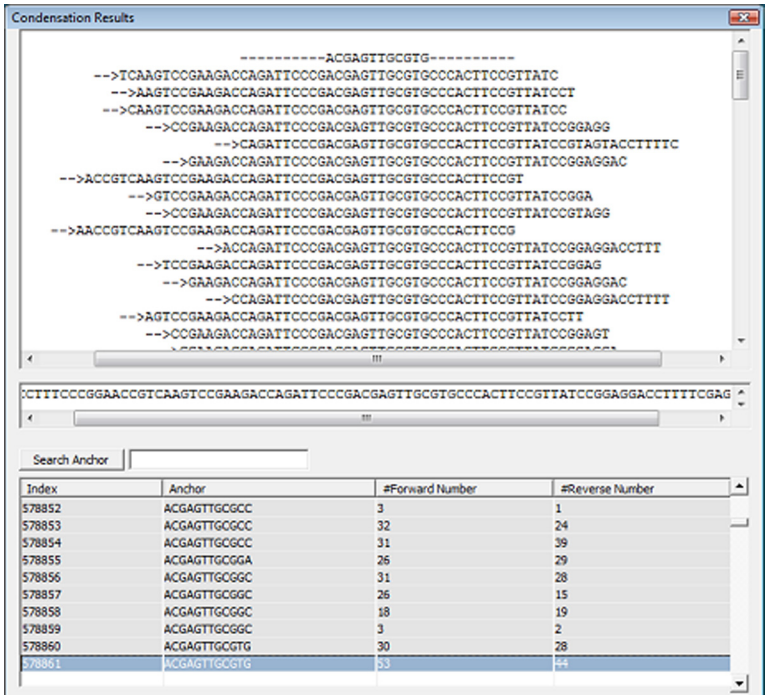

Figure 2

Procedure

1. Open NextGENe's Run Wizard by clicking on the  icon in the main toolbar.
2. Select Instrument Type.
3. Select "de novo Assembly" under Application Type.
   a. Sequence Condensation and Sequence Assembly are automatically selected under Steps.
   b. Note:Sequence Condensation should be deselected when using de Bruijn Assembly method.
4. Click Next to upload Sample files.
5. Browse to select sample file(s).
   a. If sample file is not in fasta format, use the Format Conversion Tool to convert to fasta.
   b. 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.
6. Specify output location and file name.
7. Click Next to continue to Condensation Settings (when applicable). For projects that do not use Condensation clicking Next will open the Sequence Assembly settings.
8. Fill in the appropriate information about the dataset being used in the Condensation Settings.
   a. Choose Condensation Type.
   b. Click Open Advanced Settings to view settings in detail.
9. Click Next to continue to Sequence Assembly settings.
10. Choose appropriate Assembly settings.
11. Click Finish and select Run NextGENe to begin processing project, or Create More Projects to set up more projects or select Exit Wizard to close the Run Wizard.

Additional Condensation Cycles

Increasing the number of condensation cycles will elongate the reads further, and remove more of the systemic instrument errors.

Generate Large Contigs

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

SoftGenetics LLC 100 Oakwood Ave, Suite 350 State College, PA 16803 USA
Phone: 814/237/9340 Fax 814/237/9343
www.softgenetics.com email: info@softgenetics.com

Assembly Settings

The first step in choosing assembly settings is to select which Final Assembler Method to use. This will determine which of the detailed settings are available (not grayed out).

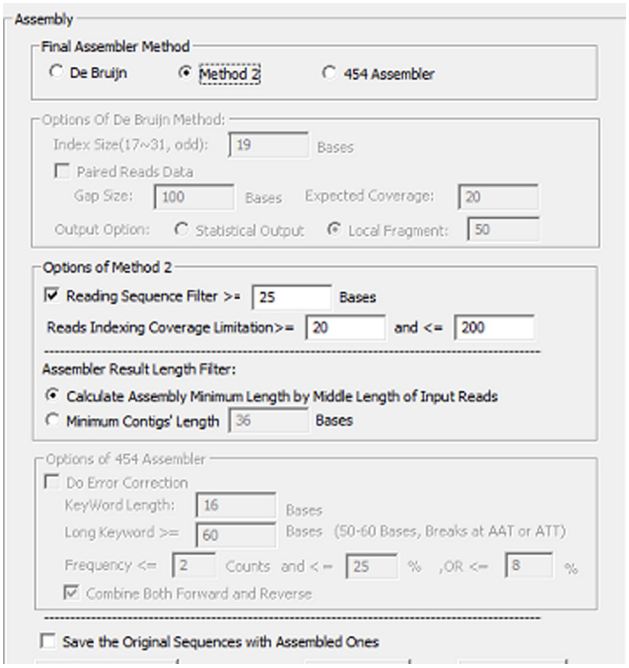

Figure 3

Figure 3: Numerous Assembly options are available to allow researchers to develop ideal settings best suited for their dataset and application.

Results

Conducting de novo assembly with short sequences generates many contigs of varying length. The number and length of these contigs is dependent on many variables including the number of repeat regions and the quality of data. NextGENe first uses the quality scores for each base to remove and trim the low quality reads and bases from the data files. Secondly, the software can perform the condensation of the short reads to cyclically increase the read lengths up to 400 bps. Once the reads have been elongated to this size, NextGENe performs a final assembly which produces the large contigs, some cases in excess of 100 Kbps.

```
Assembling with De Bruijn Method...
Final graph has 1794 nodes and n50 of 204665 max 560819
Assembly Finished.
Removing Temporary Files...

[Condensation & Assembly Statistics Information]
Assembled Sequences Count:  347
Assembled Sequences Average Length:  13263
```

Figure 4

Figure 4: Following completion of an assembly project, NextGENe produces a file that contains statistical information regarding the process. This figure shows information about the de Bruijn assembly of E. coli short read data from the Illumina Genome Analyzer.

Discussion

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 the same manner, helping to accurately identify both sides of the repeat.

de novo assembly 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 Biosystems SOLiD™ System and the Genome Sequencer FLX System from Roche Applied Science.

NextGENe software can be used for analysis of Next-Gen sequencer data for a variety of applications including SNP/Indel detection,
ChIP-Seq, Small RNA detection and quantification and Expression studies such as SAGE (Serial Analysis of Gene Expression) and DGE (Digital Gene Expression).

Acknowledgements

We would like to thank Oleg Evgrafov and James Knowles from the Keck School of Medicine of University of Southern California and Caroline Obert from St. Jude Children's Research Hospital for their collaboration with the development of this software.

References

1. Jonathan Butler et al. De novo assembly of whole-genome shotgun microreads. Genome Research. March 2008.
2. D R Zerbino, E Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genomic Research. 18: 821-829.

Trademarks are property of their respective owners.

SoftGenetics LLC 100 Oakwood Ave, Suite 350 State College, PA 16803 USA
Phone: 814/237/9340 Fax 814/237/9343
www.softgenetics.com email: info@softgenetics.com