NextGENe™
by SoftGenetics
March 2009

# De novo Assembly of Next Generation Sequencing Reads with NextGENe Software using a de Bruijn graph method

Megan Manion, Kevin LeVan, Ni Shouyong and CS Jonathan Liu

## Introduction

Utilization of 2nd generation sequencing systems are lowering sequencing costs, while dramatically increasing the speed and amount of genetic sequencing 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 and the Applied Biosystems SOLiD™ System using sequencing by ligation and di-color tagging give reliable sequence read-outs of 25-75 bps at about 5-200 million reads per sequencing 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 100 kbps. These sequencing 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. In order to produce accurate assemblies, software must be able to correct low frequency errors while maintaining true variations.

NextGENe software includes an assembly method based on the de Bruijn graph technique (2) that is capable of reducing error and resolving repeats. Ideal for the short reads of Illumina and SOLiD System platforms, this method is capable of utilizing paired reads information to assist with proper assembly of large contigs, but can also be used without paired end data.

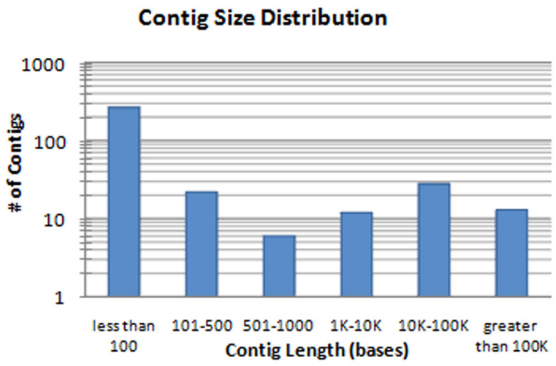


**Figure 1**

**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.

NextGENe also provides a unique, patent-pending Condensation Tool™ that is designed to statistically polish and lengthening short reads by utilizing depth of coverage. Because Next Generation sequencing systems produce data with varying characteristics that are used for numerous applications, NextGENe's Condensation Tool includes multiple methods for reducing systematic errors. The Consolidation and Elongation methods both correct low frequency instrument errors and elongate reads. Elongation is able to maintain original read counts while Consolidation reduces read number by merging identical reads. The Elongation method is recommended for de Bruijn Assembly as well as when paired reads are used and for studies where accurate count numbers are essential such as expression studies.

NextGENe's Condensation Tool also includes an Error Correction method designed to deal with low frequency homopolymer errors for the pyrosequencing reads from the Roche/454 platform.

## Methodology

This assembly method involves using short words (17-31 bps), not entire reads, as indexes to develop the graph which reduces redundancy. These short words are used to generate a hash table. For each short word the software scans the reads for its first occurrence and records its location within the read. Once this is done for all the short words found in the reads, each read can be represented by the short words it contains and its overlaps with other reads. Using this information, reads are mapped as a path along the graph with nodes representing overlaps and arcs between nodes representing links.

# Procedure

1. Open NextGENe's Run Wizard by clicking on the  icon on the main toolbar.
2. Select Instrument Type.
3. Select "de novo assembly" under Application Type.
4. Sequence Condensation and Sequence Assembly are automatically selected under Steps.
    a. Condensation can be deselected to assemble raw reads.
    b. Alternatively, the Elongation method of Condensation can be selected to correct instrument errors and lengthen reads prior to the assembly.
5. Click Next to open the Load Data step.
6. Browse to upload sample file(s).
    a. If not in fasta format, or csfasta format for the SOLiD System, use the Format conversion tool to convert file.
7. Specify output location and folder name.
8. Click Next to open settings.
    a. Select de Bruijn under Assembly Method.
    b. Select appropriate de Bruijn Assembly options for the project.
9. Click Finish and then Run NextGENe to begin processing project.

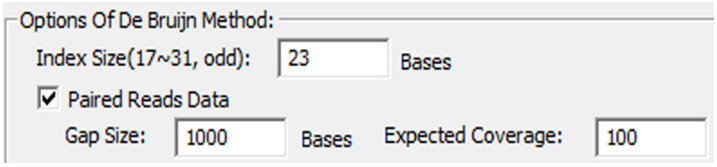


**Figure 2:** De Bruijn Graph Assembly Method settings. The index size indicates the length in nucleotides of the word used in making the graph. This should be set lower for shorter reads. For example, for 36 bp reads an index of 19-25 is recommended. However, lower index sizes will require more memory. Additional options are available to utilize paired read information. Gap size indicates the length of the insert, or fragment library.

# Results

NextGENe Software's de Bruijn graph assembly method is able to accurately assembly short sequence reads from the Illumina Genome Analyzer and the SOLiD System into large contigs up to several hundred thousand bases in length.

```
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 3:** 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

NextGENe provides an easy-to-use software application for de novo assembly of short sequence reads from next generation sequencing systems. The software applies a de Bruijn graph method to assemble short reads into large contigs that can exceed 100 kbps in length (data specific).

NextGENe also includes software modules for SNP and Indel detection (with or without use of paired reads), ChIP-Seq, Transcriptome, small RNA discovery and quantification and Digital Gene Expression studies like SAGE.

# References

1. J Butler et al. 2008. De novo assembly of whole-genome shotgun microreads. Genome Research. 18:810-820.
2. D R Zerbino, E Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research. 18: 821-829.



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