NextGENe®

January 2009

by SoftGenetics

Paired Read Assembly of AB SOLiD™ System and Illumina® Genome Analyzer data with NextGENe® Software

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

Introduction

NextGENe uses a de Bruijn graph method for assembly of paired read (mate pair) data from Next Generation Sequencers such as the Applied Biosystems SOLiD System and the Illumina Genome Analyzer (Solexa). This method involves using short words, not entire reads, as indexes to develop the graph which reduces redundancy. Reads are mapped as a path along the graph with nodes representing overlaps and arcs between nodes representing links. This assembly technique for paired reads (mate pairs) is able to accurately produce large contigs greater than 100 kbps from short next generation sequencing reads (1).

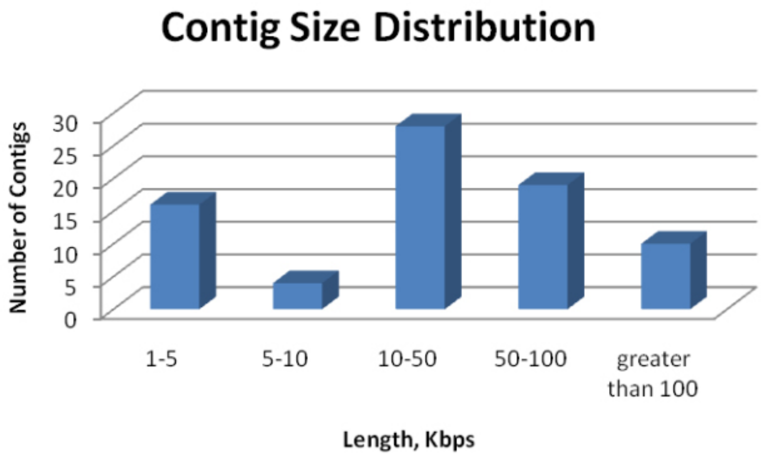


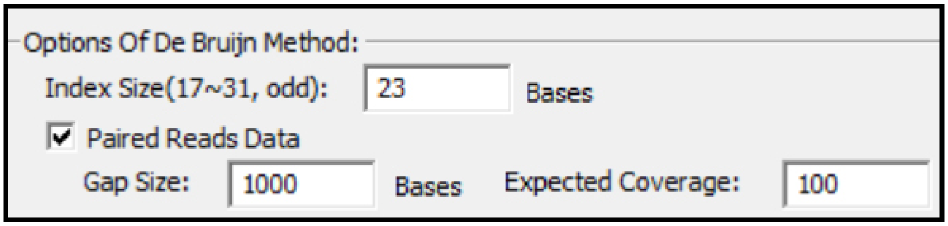
Figure 1: NextGENe is able to produce large contigs, many between 1 kbps and 100 kbps, from paired read data. Additionally, several assembled contigs are generated that exceed 100 kbps. Results shown were obtained using genomic data from E. coli, which has a total genome size of roughly 4.6 Mbp.

The use of paired-end or mate-pair sequence reads is a valuable tool for constructing de novo assemblies from short sequence reads. Next Generation Sequencing platforms have allowed for sequencing paired reads in a shorter time span for lower cost. However, the volume of data produced in the form of short reads with high error rates presents a challenge for data analysis.

Paired read analysis involves the use of DNA fragments containing two regions of sequenced DNA separated by an unsequenced insert of known length. Paired reads enhance assembly of short reads by improving the specificity of the reads since single short (25-36bp) sequencing reads, as produced by next generation technologies, are not significantly unique in the genome for accurate assembly.

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" for Application Type. No condensation step is needed.
4. Click Next to open the Load Data step.
5. Browse to upload both sample files generated for a lane.
    1. If not in fasta format, or csfasta format for SOLiD, use the Format conversion tool to convert file.
6. Specify output location and folder name.
7. Click Next to open settings.
    1. Select De Bruijn for Assembly Method.
    2. Select Paired Reads Data.
    3. Set gap size equal to the expected insert size.
    4. Set coverage equal to the expected depth of coverage.
8. Click Finish and then Run NextGENe to begin processing project.



Options Of De Bruijn Method:

Index Size(17~31, odd): 23 Bases

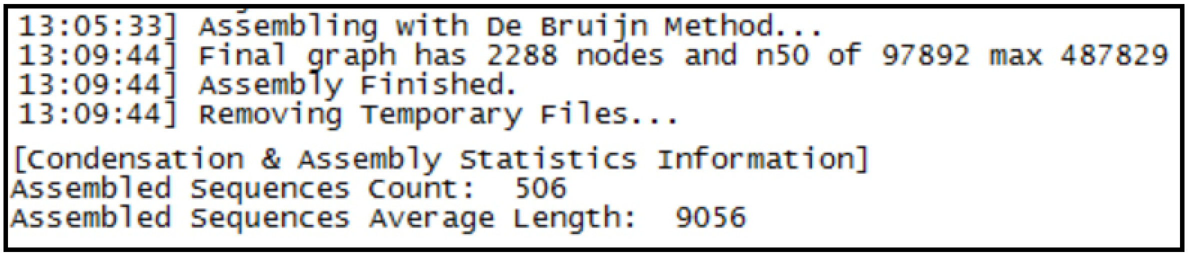
☑ Paired Reads Data

Gap Size: 1000 Bases Expected Coverage: 100

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. The gap size should be set equal to the insert size and expected coverage should be set to the expected depth of coverage.

Results

NextGENe's Sequence Assembly using paired reads is able to accurately assemble short sequence reads into contigs greater than 100 Kbps (data specific).



```
 13:05:33] Assembling with De Bruijn Method...
 13:09:44] Final graph has 2288 nodes and n50 of 97892 max 487829
 13:09:44] Assembly Finished.
 13:09:44] Removing Temporary Files...
[Condensation & Assembly Statistics Information]
Assembled Sequences Count:  506
Assembled Sequences Average Length:  9056
```

Figure 3: Upon completion of an assembly project NextGENe produces a file containing statistical information on the results. For the project shown above the maximum contig length is 487829.

Discussion

NextGENe's de Bruijn assembly method using paired end (mate pair) reads offers an easy-to-use, biologist friendly Windows-based application for producing large contigs greater than 100 Kbps from short sequence reads produced by second generation sequencing technologies. This method allows for the resolution of small repeats and reduction of error to produce accurate assemblies.

NextGENe also includes software application 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. NextGENe has been designed to operate on low cost hardware configurations.

References

1 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.