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.

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.

**Results**

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

**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**


Trademarks are property of their respective owners.