NextGENe™
by SoftGenetics
September 2008

De novo Assembly of SOLiD™ Sequence Reads with NextGENe™

Kevin LeVan, Jacie Wu, Haitao Ren, 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. These instruments generate billions of bases in a day for just a few thousand dollars. The Applied Biosystems SOLiD System generates more than 9 gigabases per run with reading lengths between 25-35 bps.

De novo sequence assembly with the short reads from genome analyzers producing these short reads presents many challenges (1). With many of the current techniques, it is difficult to assemble the short reads into a contig larger 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 SOLiD system produces reads in color space rather than base space, an additional layer of complexity for *de novo* assembly.

NextGENe, which is a Windows® based program, was developed to assist with these complex issues. NextGENe contains tools to remove low quality reads and trim low quality bases from reads in order to improve assembly accuracy. The software then utilizes a Condensation Tool to statistically polish the short reads, correcting additional errors and simultaneously lengthening the reads. Through this step, NextGENe can more reliably assemble the short reads into contigs of 500 bps to upwards of 50 kbps. An additional benefit is that the original short reads used to generate each assembled contig are recorded to show the copy number and Indel positions. NextGENe is capable of detecting Indels of 1-30 bps.

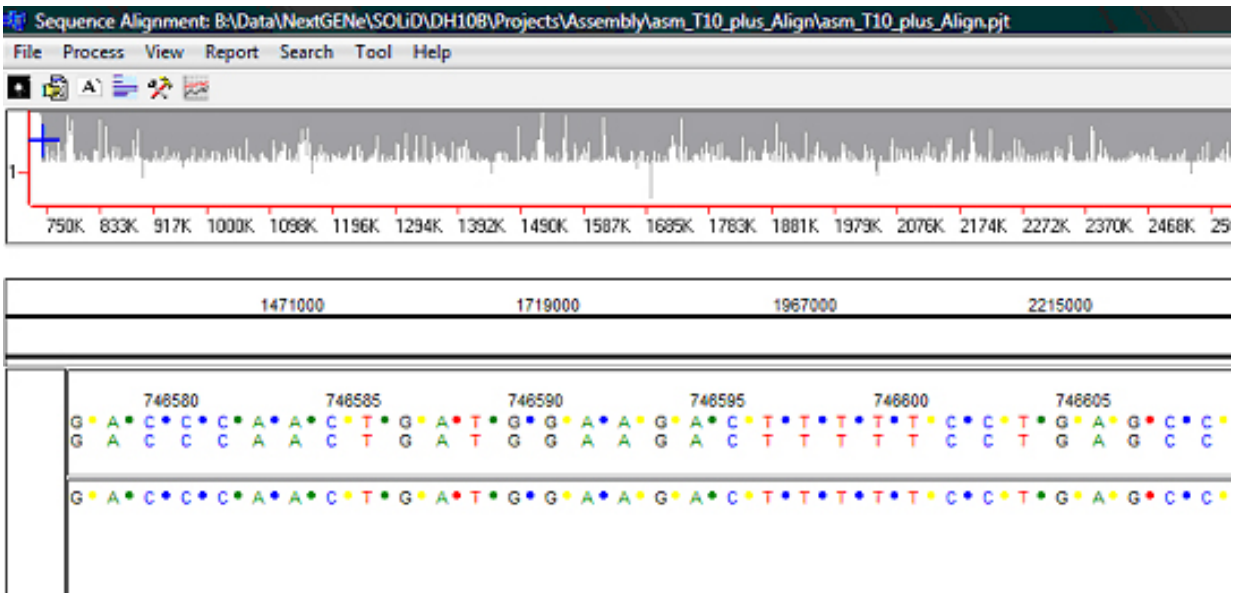


Figure 1: Assembled contigs aligned to the DH10B reference genome to validate the assembly results.

De novo Assembly Methodology

NextGENe statistically polishes datasets of adequate coverage to remove random sequencing errors and increase read lengths with the Condensation Tool. Repeating the Condensation removes systematic errors and further lengthens the sequence reads with each additional cycle. The polished and elongated reads can then be assembled into large contigs while removing redundant reads. Once the dataset has been cleaned to remove low quality reads and ends, the remainder of the process is fully automated through the use of a Run Wizard that guides you through the project configuration.

The first step is utilizing the Condensation Tool to generate the first assembly. All reads with the same 12 bp anchor sequence are collected into a cluster. The two shoulder sequences on either side of the anchor sequence are used to sort the short reads into multiple groups. The consensus sequence in each group is obtained from the short reads. By using the consensus, random sequencing errors are corrected. The ending bases are removed from the consensus when the base is covered by only one sequence read or there is an inconsistency between multiple reads. The 5' sequence has higher weight than that of 3' end because of its higher quality. With 50x coverage within one group, confidence of the condensed sequence is about 99.8%.

The second step is to further assemble the condensed results using a similar process. The software assembles the consensus sequences while tolerating a 1 bp error rate. This second step is conducted by completing multiple cycles of Condensation, each cycle increasing the average contig read length. After the first cycle, short reads of 36 bps assembled to 60 bps. After a second cycled, the 60 bps consensus sequences can be elongated to 100 bps. A third and forth cycle can generate sequences of 180 and 350 bases respectively.

The last step is to reduce the number of redundant sequences and to assemble the highly similar sequences into larger contigs while minimizing assembly error.

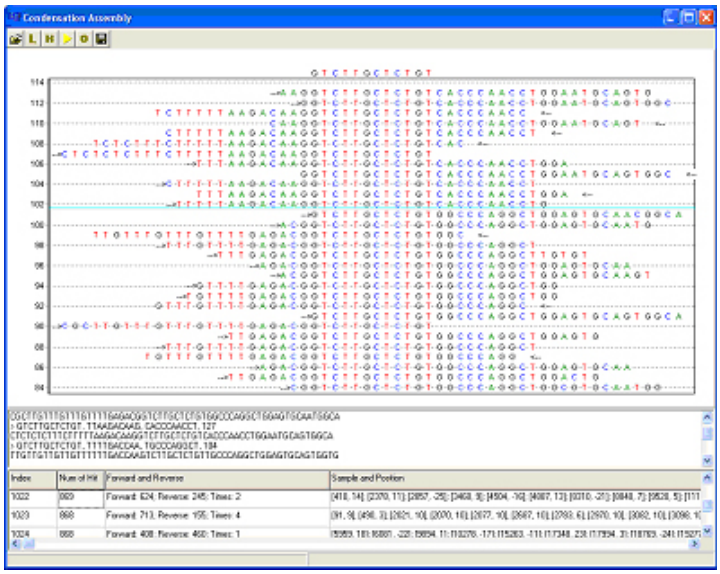


Figure 2: Condensation Assembly Tool elongated the 35 bp reads to approximately 60 bp while removing many of the random errors produced by the instrument.

Procedure

Sample File Preparation

NextGENe removes low quality reads and trims low quality ends. The first color call represents the color change between the last base of the primer and the first base of the sequence. This position doesn't represent the base call in the genome and is used only for translational purposes between color space and base space; so this position is not used.

1. Choose Format Conversion from NextGENe's Tools menu to remove low quality calls.
    a. Choose the SOLiD CSFASTA (Color) File Format Type.
    b. Add the CSFASTA and QUAL files generated for one SOLiD analysis to the Input field.
    c. Browse to an Output Path location to save resultant CSFASTA file. Filename is automatically appended with "_converted".
    d. Add checkmark to desired settings for removing low quality calls. A suggested start would be to remove reads with a Median Score below 13 and to trim when three consecutive bases are below 10.
    e. Click OK. The Format Conversion window will close when resultant file is created.

2. Choose Sequence Operation from NextGENe's Tools menu to remove first base from each read.
    a. Choose the Sequence Trim Operation Type.
    b. Add the CSFASTA file produced by the Format Conversion Tool to the Input field. If no removal of low quality reads is necessary, load the original CSFASTA file.
    c. Browse to an Output Path location to save resultant CSFASTA file. Filename is automatically appended with "_trimmed".
    d. Add checkmark to the Remove Setting and type 1 in First Bases and type 0 in Last Bases.
    e. Click OK. The Sequence Operation window will close when resultant file is created.

NOTE: The Sequence Trim operation removes bases from the CSFASTA file while leaving the QUAL file unmodified. Therefore, the positional information between the two files is no longer available.

Project Configuration

3. Open NextGENe's Run Wizard through the Process menu.
4. From the Application window, select SOLiD for Instrument Type and *de novo* Assembly for Application Type. This enables the Condensation and Assembly steps of NextGENe. Click Next.
5. From the Load Data Window, click Load button next to Sample Files field to add the sample file that has low quality bases removed and first base trimmed from each read.

NOTE: Sample size is limited to 3 million reads or 200 megabytes with a 32-bit Windows® system. Input size increases to 10 million reads with a 64-bit Windows system containing a quad processor and 8GB RAM.

6. Set an Output Path location to save the assembled project files and click Next.
7. Configure Condensation cycles. Set the number of cycles to 4 and click Set. Default settings are ideal for 100X coverage.

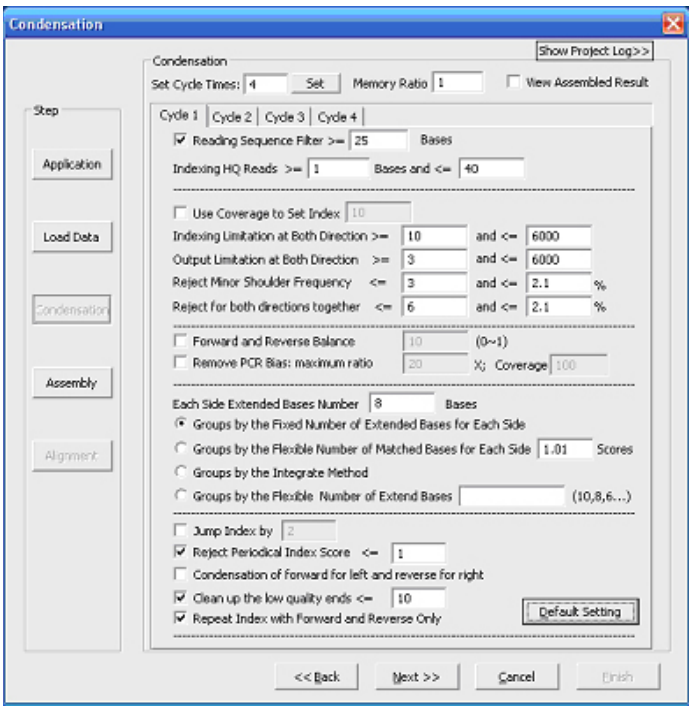


Figure 3: Options for first cycle of Condensation Assembly. The three additional cycles are very similar to one another, and often the default settings are optimal.

NOTE: For a more detailed understanding of each setting, refer to the NextGENe manual.

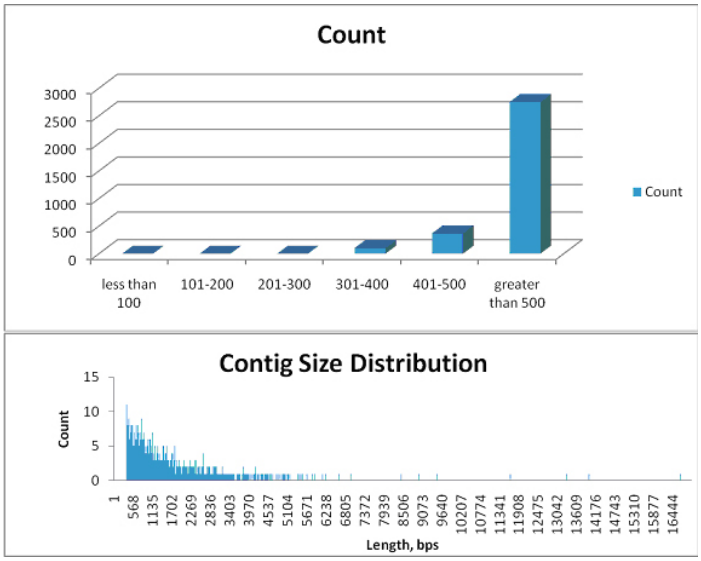
8. Choose Assembly steps. Default settings are ideal for 100X coverage. Click finish and choose to Run NextGENe.
9. The Running Log shows when the project has completed. The resultant project contains several files, including a statistics file and an Assembled Reads file.

Results

As an example, a sample of the K-12 DH10B strain of *E. coli* that was sequenced with the SOLiD System was assembled using NextGENe. This dataset has high quality reads suggestive of about 30X coverage. These single reads had low quality reads removed, low quality ends trimmed, and the first color-call removed from reads through NextGENe tools. The remainder of reads, those of high quality and reliability, were processed through NextGENe's Condensation and Assembly modules to generate larger color-space contigs. Since the DH10B genome is available, the assembled results were then aligned to this genome to validate the assembly results.

Four cycles of Condensation were performed followed by assembly of the condensed results to generate 3220 contigs. Over 85% of these contigs were larger than 500 bps, and the largest is over 16 Kbps, as shown in Figure 4.

Figure 4: Four cycles of condensation produced 3220 contigs ranging in size from 304 to 16,471. Over 85% of the contigs are over 500 bps.



The assembled contigs were aligned to the DH10B reference genome to validate the assembly results (figure 1). Only three of the assembled reads did not match to the DH10B genome, indicating that the contigs produced by NextGENe are very reliable. Figure 4 shows 4,024,290 bases of the 4,686,137 base genome has coverage, approximately 86%. When excluding the duplicated region of this reference genome that is larger than 100 kbps, coverage is 88%.

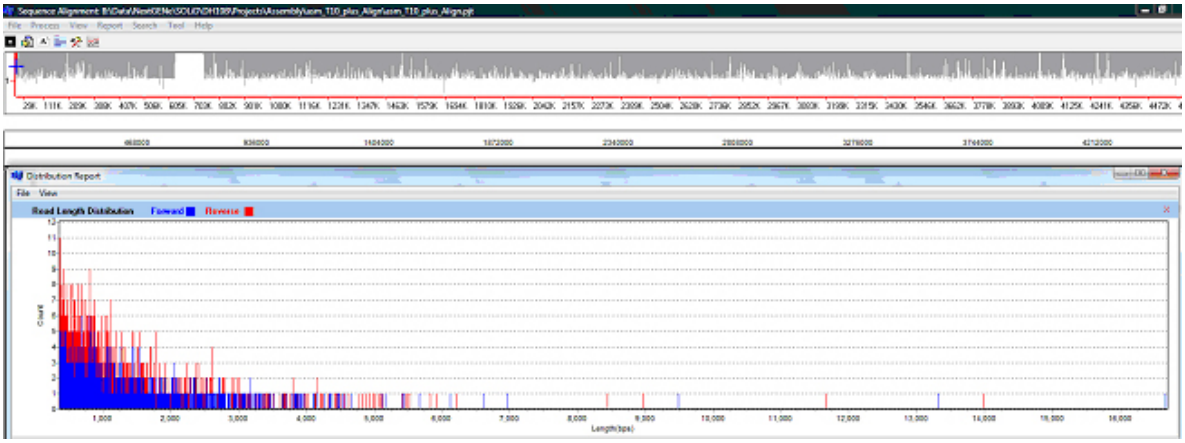


Figure 5: Distribution Report of Forward and Reverse fragments as they align to the DH10B genome. The top graph shows genome coverage – the large region without coverage is a duplicated section of the reference genome. The lower chart shows the size distribution and orientation of reads that are aligned to the DH10B genome.

	Contigs >500	N50 Stats of contigs>1589	All Contig Stats
Number of Contigs	2749	783	3220
Number of Bases	3,867,807	2,035,715	4,072,839
Average Contig Size	1407	2600	1265
Median Contig Size	1092	2270	951
Largest contig	16681		

Discussion

The use of genome analyzers in over 100 articles in less than two years indicates that the development of software to analyze these mega data sets is increasing in demand. Assembly using NextGENe provides an innovative and accurate tool to solve challenges associated with the short reads generated by the SOLiD system.

NextGENe improves contig generation with tools to remove and clean reads. Its unique Condensation feature allows for the simultaneous error correction and elongation of reads prior to assembly, for improved reliability.

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 (Nforward/Nreverse). It is possible that this imbalance is the result of amplification biases or primer/sample impurities. 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.

Depth of coverage is critical for creating reliable contigs of high length. Many biases affect the uniformity of coverage distribution, as shown in Figure 5. Low quality reads will not be used during assembly and these reads may encompass as many as 50% of the reads. Fragmentation of genome may not be uniform, possibly due to nucleotide content, leading to localized regions of lower coverage. All of these factors suggest that higher coverage may lead to better, more reliable assembly results.

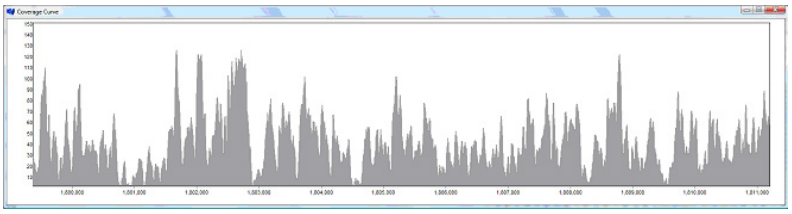


Figure 6: Actual coverage of a sample with an average coverage of 30X often shows many regions with very low coverage, inhibiting assembly in these regions.

In addition to *de novo* assembly, aligning of short sequence reads to assembled contigs is useful for evaluating transcript expression levels. NextGENe comes equipped with several reports including the Expression and Comparison reports. When the Assembled contigs are used as a reference file for the short reads, The Sequence Alignment tool aligns the short reads to each contig and calculates statistics such as coverage. The Expression Report shows several statistics for each contig, including the depth of coverage and the total number of reads. The Comparison Report allows for multiple projects to be loaded simultaneously in order to create a table comparing the expression levels of each project relative to one another or a reference project.

NextGENe is a powerful tool for the assembly of SOLiD's color-space reads in addition to applications including SNP discovery and expression studies. The software is fast, operates on a Windows operating system, and generates reliable results.

Acknowledgements

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

SoftGenetics LLC 200 Innovation Blvd. Suite 235 State College, PA 16803 USA
Phone: 814/237/9340 Fax 814/237/9343
www.softgenetics.com email: info@softgenetics.com