NextGENe™
by SoftGenetics
March 2009

Reducing Error in Next Generation Sequencing Data with NextGENe Software's Condensation Tool™

Megan Manion, Kevin LeVan, Shouyong Ni, Yiqiong Jacie Wu and CS Jonathan Liu

Introduction

The development of next generation sequencing technologies such as the Genome Analyzer FLX by Roche Applied Science (454 Sequencing), the SOLiD™ system from Applied Biosystems and the Illumina Genome Analyzer have drastically lowered sequencing costs while increasing the speed and quantity of information gathered. The sheer volume of information generated poses a challenge for data analysis. In addition, these massively parallel sequencing systems produce short reads with high error rates which create further difficulty for de novo assemblies, SNP and Indel detection and many other applications. Fortunately, the large depth of coverage available in second generation sequencing data can be utilized to deal with these problems.

NextGENe provides a unique, patent-pending Condensation Tool that is designed to overcome the challenges of high throughput sequencing technologies by lengthening reads and statistically removing instrument errors by taking advantage of depth of coverage.

Condensation Results

	Before Condensation	After Condensation	Percent Change
Average Read Length	34bp	53bp	+56%
Total Read Count	6,171,828	624,404	-90%
Average Error Rate	2%	<0.05%	--
% of Reads Matched to Reference	33%	93%	+182%

Figure 1: In comparing the alignment of raw reads to a reference sequence with the alignment of condensed reads, read length is increased, total read count is decreased, coverage is increased and errors are removed to allow for much better matching to the reference. The data shown is a dataset of human transcriptome sequence reads from the Illumina Genome Analyzer aligned to the whole human transcriptome reference.

Short reads such as those from the Illumina Genome Analyzer and the Applied Biosystems SOLiD System are often not unique within the genome being analyzed. So, by lengthening reads and filtering errors, the Condensation Tool drastically improves the accuracy of de novo assemblies and increases the length of assembled contigs.

Assembly Results

	Without Condensation	With Condensation	Percent Change
Average Length of Assembled Contigs	34bp	1082bp	+224%
Total Number of Assembled Contigs	14678	4358	-70%
Matching to Reference Genome	76%	92%	+21%

Figure 2: Assembly Results are significantly improved when Condensation is used. Fewer contigs of greater length are generated. Also, assembled contigs are more accurate, as seen by aligning assembly results to the reference genome.

Because Next Generation sequencing systems produce data with varying characteristics that are used for numerous applications, NextGENe's Condensation Tool includes three different methods for reducing systematic errors.

Consolidation and Elongation 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 when paired reads are used and for studies where accurate count numbers are essential such as expression studies.

Error Correction is a Condensation method designed to deal with low frequency errors for longer Roche/454 reads. It is ideal for correcting homopolymer errors and other base calls errors produced by the pyrosequencing technique.

Methodology

Consolidation and Elongation

The Consolidation and Elongation functions of the Condensation Tool are used to polish and lengthen short sequence reads into fragments that are more accurate and manageable. By clustering similar reads containing a unique anchor sequence and two flanking shoulder sequences, data of adequate coverage are condensed, the short reads are lengthened and reads containing instrument errors are corrected. In creating the consensus sequence, the 5' sequence is given a higher weight than that of the 3' end because of the difference in base call quality. This stage helps to prepare data for analysis in applications such as SNP/Indel detection by statistically removing many of the errors while maintaining the allele frequency information.

The unique anchor sequence, or index, is a 12 base fragment that is found in several of the reads. All reads containing this exact sequence are clustered together. Often, many of the reads within a cluster contain homologous shouldering nucleotides both upstream and downstream of the anchor sequence. The cluster of reads can be sorted by these flanking shoulder regions into groups of similar sequence. The consensus of these groups is often close to twice the original length of the reads. The quality of the consensus sequences are significantly improved compared to the original reads when high coverage reads are used with higher statistical weight given to the 5' sequence. Reads containing irregular variations are grouped into a separate file when the variation is found at a high frequency.

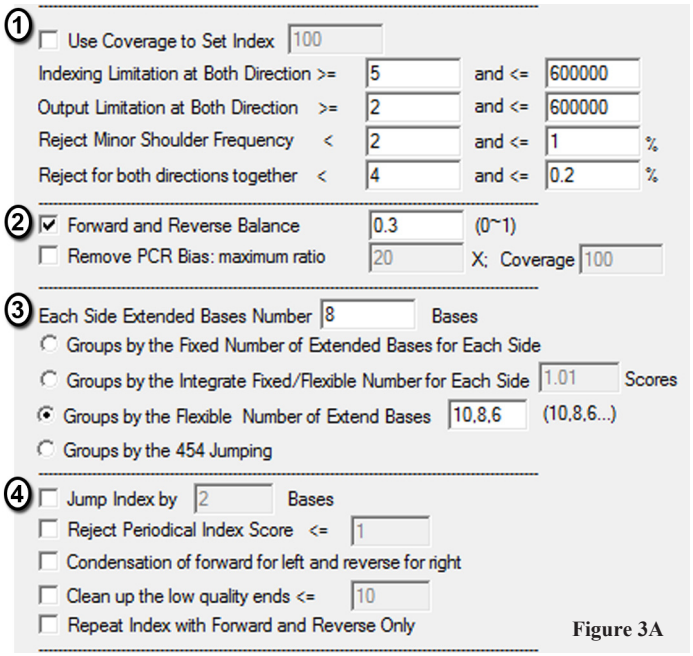

Figure 3A: NextGENe's Condensation Tool provides options to optimize results for each project. Recommended settings for projects with two-directional coverage are shown. Section **1** includes settings for count requirements for indexing anchor sequences, section **2** includes settings for balancing forward and reverse read counts, section **3** includes settings for indexing flanking shoulder sequences, and section **4** includes settings for scoring the 5' and 3' ends of reads.

Figure 3B

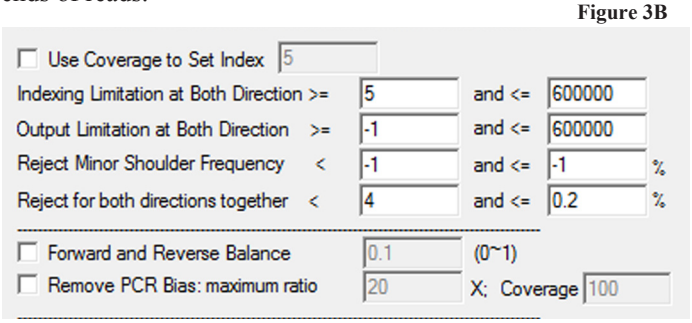

Figure 3B: Different condensation settings should be used for one-directional data such as ChIP-Seq analysis data.

For Consolidation, the consensus reads are produced by merging all reads within a cluster and are saved in place of the original reads. The total read count is reduced while the average sequence length is increased. This feature is ideal for datasets with high depth of coverage.

For Elongation, the consensus sequence is used to correct low frequency errors and lengthen reads within a cluster, but the original reads are kept. Total read count does not change while average sequence length is increased. This feature is important for paired read (mate pair) analysis so that both reads within a pair are always kept. Merging or removing reads in paired read analyses results in the loss of read pair information. This feature is also ideal for low coverage datasets where reducing read count would be detrimental or applications such as expression studies where accurate read count information is required.

Error Correction

NextGENe's Condensation Tool includes an application that is specifically designed to correct low frequency homopolymer errors of Roche/454 pyrosequencing reads. Error Correction works by parsing sequencing reads into shorter keywords and comparing the keywords between the reads to help determine the correct bases at the ends of each keyword. Keywords are produced by dividing the reads where a nucleotide is repeated three or more times (homopolymers of 3 or more bases) and there are at least 16 bases between the homopolymers. Reads that include variations that are found at low frequencies are corrected. Users can set relative and absolute frequencies for acceptable variations.

Results

Following use of the Condensation Tool, read lengths can be increased up to twice the original length and low frequency errors are removed while true variations are maintained.

Figure 4: Average Read length was increased from 34bp to 53bp, more than a 1.56 -fold increase. The consensus sequence errors are reduced significantly, far below 0.1%.

Figure 5: On the left, raw reads are aligned to the reference. Low frequency errors are highlighted in gray while mutation calls are highlighted in blue. On the right, condensed reads are aligned to the reference. Low frequency errors, most likely instrument errors, were eliminated while the true SNP was maintained.

Indel Detection

In addition to the accurate detection of SNPs, NextGENe's Condensation Tool allows for the detection of Indels by increasing read length.

Figure 6: By increasing read length, NextGENe's Condensation Tool allows for the detection of Indels up to about 30 bps in length. In this figure a 13 base pair deletion of "TGACCATACACCA" was detected at position 12243-12255.

Results

NextGENe's exclusive Condensation Tool is designed to facilitate accurate and streamlined analysis of massively parallel sequencer data. By reducing errors and lengthening reads, the Condensation Tool makes reads more accurate and unique within the genome, enhancing the potential for accurate genome assembly and alignment to reference genomes. The tool is designed with flexibility in mind, containing multiple methods and options to suit a variety of data and application types. Reads of low coverage may not be able to be condensed; however, these reads are saved in a separate uncondensed reads file and can be used for later analysis.

NextGENe is an easy-to-use, biologist friendly software tool that can be used to analyze Second Generation Sequencer data for a variety of applications. In addition to its custom Condensation Tool, NextGENe includes software applications for expression studies like transcriptome studies, SAGE and small RNA analysis, as well as de novo assembly, SNP and Indel detection, and ChIP-Seq.

References

1. J Shendure and H Ji. 2008. Next-Generation DNA sequencing. Nature Biotechnology. 26: 1135-1145.

Trademarks are property of their respective owners.

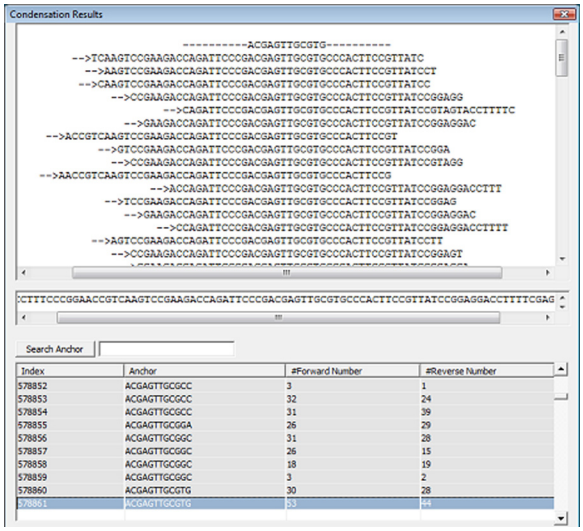

Figure 4

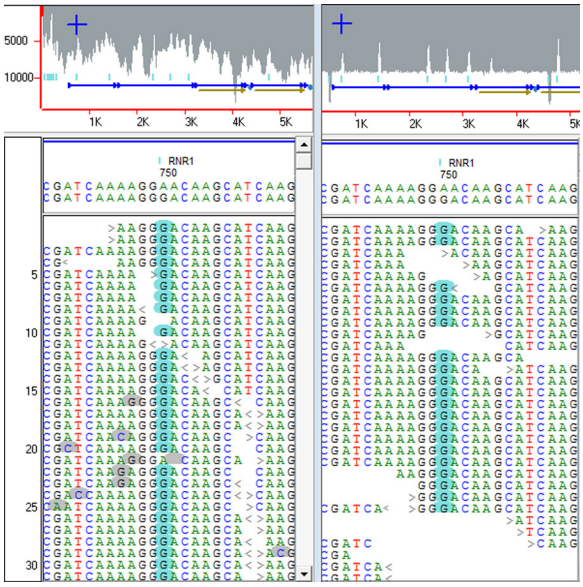

Figure 5

Figure 6

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