NextGENe™ by SoftGenetics

April, 2008

Transcriptome Analysis Using NextGENe Software

Kevin LeVan, Shouyong Ni, Jin Yu, Sean Liu, Jacie Wu and ChangSheng Jonathan Liu

Introduction

A transcriptome is a collection of all transcribed elements within a genome. Differences in the RNA expression levels are observed between cell types, stages of development, as well as normal and disease tissue. Understanding these differences is valuable in areas such as the discovery for novel drugs **(1)**. Some of the common techniques for analyzing transcripts include Serial Analysis of Gene Expression, or SAGE **(2)** and microarrays **(3)**. Additionally, transcriptomes are studied using genome analyzers **(4)**.

The next generation DNA sequence technologies generate millions to hundreds of millions of the short sequence reads. Illumina® Genome Analyzer utilizing the Solexa sequencing technology uses PCR on a surface, the Applied Biosystem SOLiD™ System uses emulsion PCR and sequencing by ligation and the Helicos™ Genetic Analysis System from Helicos Biosciences Corporation employed the technology of true single molecule sequencing (tSMS™).

Analyzing an organism's transcriptome with the Next Generation Sequencing technology presents several challenges, including a high level of sequence variation to the reference genome due to SNPs/Indels, a single analysis often including multiple transcripts for each gene and high variability in expression rates. Short reads (25 to 35 bases) are not always unique, causing ambiguities between the various isoforms. In addition, high expression of some genes can mask genes of low expression levels when misinterpreted as noise. For example, the imbalance of gene expression from maternal and paternal alleles is difficult to measure because the data may contain SNPs and Indels that are often discarded.

By using the Condensation Tool, short reads are statistically polished and nearly doubled in length, allowing for noise and error to more reliably be filter out. When using the Alignment Tool, the highly expressed sequences are matched to the reference. The low level reads, often mistaken for sequencing errors, are rescanned and matched to the reference allowing for more accurate detection of genes expressed at lower rates. The expression ratio of multiple alleles that differ by SNPs/Indels can more accurately be evaluated.

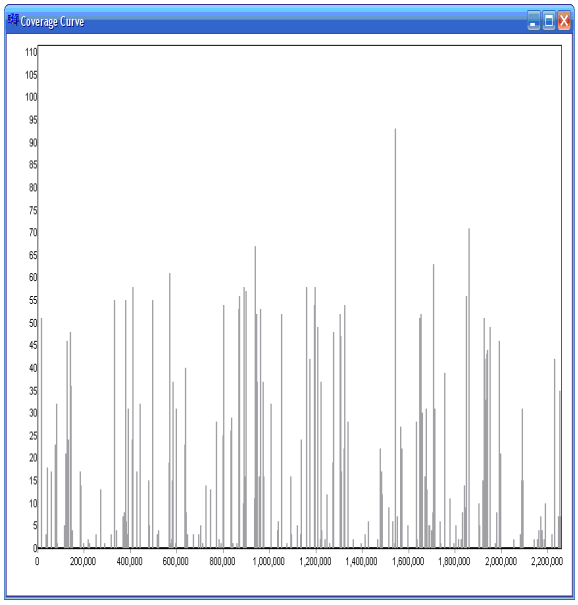

Figure 1: Coverage for the entire genome can be shown in Coverage Curve. Reference base position and base count are plotted on the x- and y-axes respectively. Currently in view are the first 2.2 million bases of the 13 million base transcriptome reference that was used in the analysis.

Procedure

The first step is utilizing the Condensation Assembly Tool to generate the first assembly. All of the reads with the same anchor sequence of 12 bps are collected into a cluster. The two shoulder sequences of 10 bps are used to sort the short reads into multiple groups. The consensus sequence in each group is obtained from the short reads. The ending bases are ignored from the consensus when the base has covered only one sequence read or inconsistency between multiple reads. The 5' sequence has higher weight than that of 3' end because of quality. With 50x coverage, confidence of the condensed sequence is about 99.8%. Then all of the possible anchor sequences with 16.7 million possibilities are calculated. A short sequence read may be used multiple times in the Condensation Assembly. The second step is to further assemble the condensed results using a similar process while tolerating a 1 bp error rate.

Once the short sequence reads have been statistically polished with the Condensation Assembly Tool, the file can be loaded into the Sequence Alignment Tool for determination of expression.

First Condensation

1. Open the Condensation Assembly Tool and click on the Open Folder button.
2. Click the Add button and choose the sample file.
3. Set the Load Sample Section value to the number of reads analyzed simultaneously.

NOTE: Data input is limited to 3 million reads or 200 megabytes with a 32-bit Windows® system. Input size increases to 10 million short reads with a 64-bit Windows system with 8GB RAM.

4. Click the Options button and set options according to Figure 2.
5. Click the Save button.
6. When condensation is complete, a message will appear showing the start and end time of analysis. Click OK to view results.

NOTE: The Condensation Assembly Tool generates a condensed file that contains the elongated reads and an uncondensed file that contains all reads that were not used for elongation (often the reads containing errors and repeat sequences). Other files may also be created. Depending on the number of reads simultaneously analyzed, the condensed reads may be parsed into multiple files.

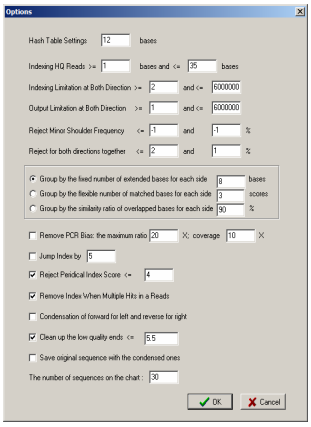

Figure 2: Options for first cycle of Condensation Assembly.

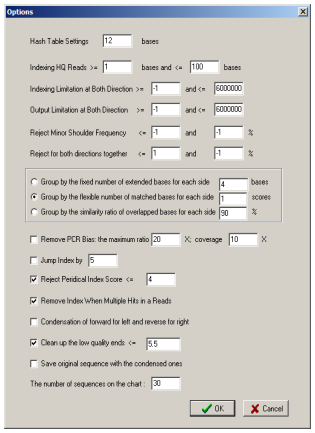

Figure 3: Options for second Condensation cycle.

Second Condensation

7. In the Condensation Assembly Tool, click the Add button and choose the Condensed sample file produced by the first cycle.
8. Click the Options button and set options according to Figure 3.
9. Click the Save button to start.

Align Reads

10. Open the Sequence Alignment Tool and select Load Data ().
11. In GBK File field, select Open and choose the annotated sequence file(s) to use as the reference.
12. In Sample File field, select Open and choose the sample file(s) to be analyzed.
13. Choose Settings from the Process drop-down menu () and adjust accordingly.
14. Close Settings and click the Run button. When analysis is complete, a message will appear showing the start and end time of analysis.
15. View alignments, open reports and export results.

Results

Two cycles of the Condensation Assembly Tool generates elongated consensus sequences with most of the random and systematic errors removed from the analysis without rejecting SNPs/Indels. The sequence reads of 35 bps within this Solexa run contain only a 1% error rate in calls for the first 25 bases. Base calls towards the ends show an error rate closer to 5%. Therefore, the software assumes the accuracy at 5' end of reads is more reliable. Reads that are oriented in the forward direction for a particular anchor sequence are more reliable upstream of the anchor (left side), and reads that are reverse complemented for the anchor are more reliable downstream of the anchor (right side). Utilizing this information, the reads were initially lengthened to an average of 55 bps. After a second condensation cycle, systematic error is removed and reads are elongated further, as shown in Figure 4.

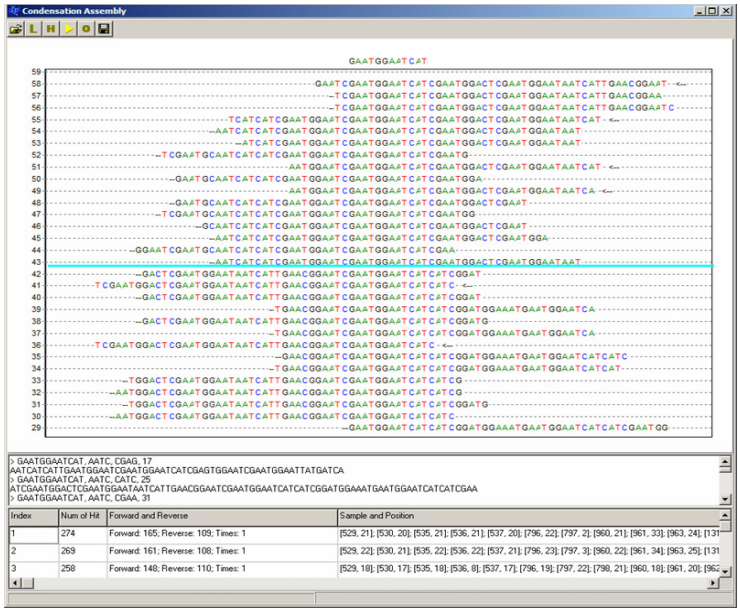

Figure 4: After two condensation cycles, the cluster of short sequence reads originally 35 bp in length containing the anchor sequence GAATGGAATCAT were elongated into groups as much as 84 bps. The blue line is highlighting the border between two to the groups containing this common anchor sequence.

After the reads were statistically polished – many of the errors have been removed and reads were lengthened – the Sequence Alignment Tool was used to align reads to the transcriptome in order to determine coverage. Figure 1 is the Coverage plot for this 13 million base portion of the transcriptome. The region in view is from 0 to 2.2 million bases with transcripts of low or no coverage and those of 100 times.

Expression results can often be skewed for new alleles that may not be included in the transcriptome reference. NextGENe is able to tolerate these SNPs and Indels, align these reads to the closest reference, and detect these variations as observed in Figure 5.

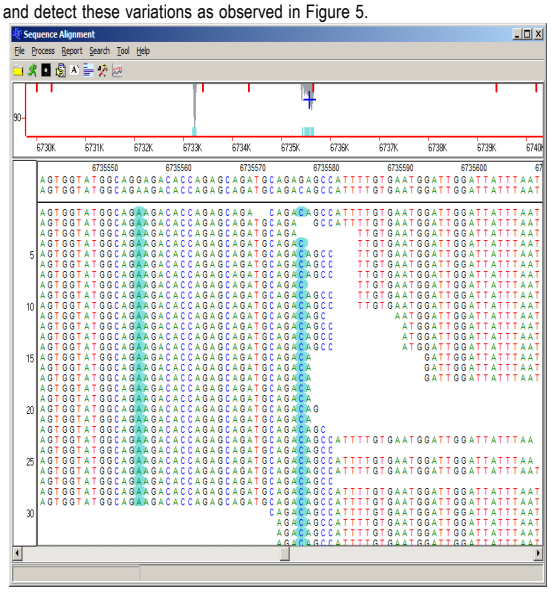

Figure 5: High frequency variations between the transcriptome and the sample reads, automatically highlighted in blue, can easily be aligned to the reference. Reports are available for viewing, editing and exporting this information.

Discussion

NextGENe is a powerful tool for the analysis of transcriptome data produced by short sequence genome analyzers. Errors within the reads are removed and reads are elongated with the Condensation Assembly Tool, and then aligned to the reference transcriptome. A 32 bit computer system can accept over a 10 million base reference, and a 64 bit system can extend this to as much as 100 Mbps.

In addition to expression analyses, NextGENe can be used for SNP/Indel detection. A Mutation Output report can be generated for each project, showing a list of all variations marked as mutation calls. Calls can be manually reviewed, and this report allows for calls to be edited, deleted or added. Options are available for customizing the view of this information, in addition to further filtering. The calls within this report are organized by reference position, and each line contains this position number, segment description, segment position, the reference nucleotide, coverage, relative gain/loss for each allele, percentages of reads containing Indels and any additional comments. Graphical representation of these results is also available.

de novo assembly of short reads from the Solexa and SOLiD systems is another important feature of NextGENe. Use of the Condensation Assembly Tool removes random systematic instrument error. Increasing the number of cycles run on the Condensation Assembly can lengthen the reads to 0.5 to 1 kb. These elongated reads can then be assembled into much larger contigs that end in repeat sequences.

NextGENe is a versatile software package designed for the new era of genome sequencing, supporting the Illumina Genome Analyzer, the Applied Biosystem SOLiD™ System and the Genome Sequencer FLX System from Roche Applied Science. It can be used for expression studies including Transcriptome, SAGE, and microRNA analyses. The results of the analysis can be saved as a reference file, allowing for direct comparison to the results from another analysis. This is a useful feature for comparison studies such as Chromatin Immunoprecipitation (ChIPSeq).

Acknowledgements

We would like to thank Professor Hong Ma of Pennsylvania State University, for providing data and collaborating with the development of this software.

References

1. C. Freiberg et al. 2004. The impact of transcriptome and proteome analyses on antibiotic drug discovery. Current Opinion in Microbiology. 7: 451–459.
2. V. E. Velculescu et al. 1995. Serial Analysis of Gene Expression. Science. 270: 484-7.
3. M. Schena et al. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 270: 467-70.
4. A. Toth et al. 2007. Wasp Gene Expression Supports an Evolutionary Link Between Maternal Behavior and Eusociality. Science. 318: 441-444.

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