

## Transcriptome Analysis of SOLiD™ System Data

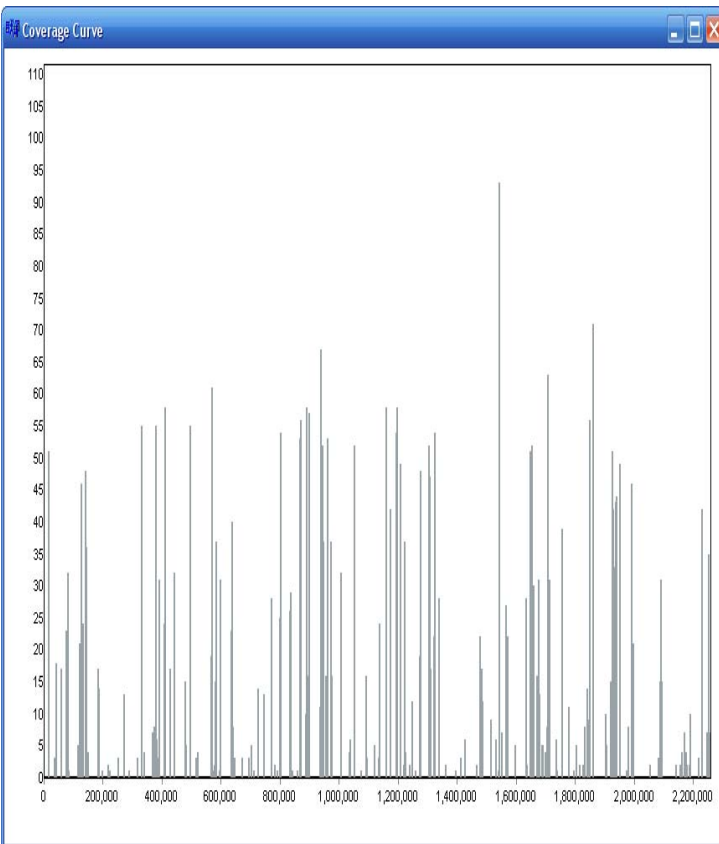
Kevin LeVan, Shouyong Ni, Jin Yu, Sean Liu, Jacie Wu and ChangSheng Jonathan Liu

### 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 of 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).

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.

The Applied Biosystems SOLiD™ System uses emulsion PCR and sequencing by ligation to generate millions to hundreds of millions of the short sequence reads. By using the Condensation Tool, short reads are statistically polished and nearly doubled in length, allowing for noise and error to more reliably be filtered out. When using the Alignment Tool, the highly expressed color-space 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 113 million base transcriptome reference that was used in the analysis.

### Procedure

NextGENe statistically polishes regions of a genome within a dataset containing adequate coverage to remove random sequencing errors and increase read lengths with the Condensation Tool. Once the dataset has been cleaned to remove low quality reads and ends, the remainder of the process is fully automated via use of a Run Wizard that guides you through the project configuration.

The first step is utilizing the Condensation Tool to generate statistically polished reads. 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 and there is 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%.

### Sample File Preparation

NextGENe can remove low quality reads and trim low quality ends from reads. 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 System 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 *Transcriptome* for Application Type. This enables the Condensation and Alignment 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. Click the Load button next to the Reference Files field to add the reference file.
7. Set an Output Path location to save the assembled project files and click Next.
8. Configure Condensation cycles. Set the number of cycles to 1 and click Set. Default settings are ideal for 100X coverage.
9. Click finish and choose to Run NextGENE.
10. The Running Log shows when the project has completed. The resultant project contains several files, including a statistics file. Results are opened in the Sequence Alignment Tool.

#### Results

**One cycle of the Condensation Tool generates elongated consensus sequences with most of the random and systematic errors removed from the analysis without rejecting SNPs/Indels. Color calls towards the ends show a higher error rate than calls toward the beginning of the reads. 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 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 113 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.



**Figure 2:** The top graph shows a view of the entire genome currently zoomed in on a 15 kbps region. Location of consensus sequence is represented by the + in the Whole Genome View at top. Transcriptome sequences are located across this 9K region of the reference sequence in addition to others in genome.

#### Discussion

NextGENE is a powerful tool for the analysis of transcriptome data produced by Applied Biosystems SOLiD System. Errors within the reads are removed and reads are elongated with the Condensation 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 200 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 SOLiD systems is another important feature of NextGENE. Use of the Condensation Tool removes random systematic instrument error. Increasing the number of cycles run by the Condensation can lengthen the reads to 400 bps. These elongated reads can then be assembled into much larger contigs.

NextGENE is a versatile software package designed to analyze data from the Applied Biosystem SOLiD™ System. It can be used for other expression studies and SNP/Indel discovery. 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 (ChIP-Seq).

#### 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.