

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 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).

Next generation DNA sequence technologies generate millions to hundreds of millions of short sequence reads. The Illumina® Genome Analyzer utilizing the Solexa sequencing technology uses sequencing by synthesis, the Applied Biosystems SOLiD™ System uses emulsion PCR and sequencing by ligation and the Genome Sequencer FLX System from Roche Applied Sciences (454 Sequencing) utilizes pyrosequencing technology.

Analyzing an organism's transcriptome with the Next Generation Sequencing technology presents several challenges, including a high level of sequence variation to the reference due to SNPs and Indels, a single analysis often including multiple transcripts for each gene and high variability in expression rates. Short reads such as those produced by next generation sequencers are not always unique within a genome, causing ambiguities between the various isoforms and making accurate alignment to a reference difficult. In addition, high expression of some genes can mask genes of low expression levels which can be 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.

NextGENe Software solves many of these inherent problems with its unique function for error reduction. By using NextGENe's Condensation Tool™, short reads are statistically polished and nearly doubled in length, allowing for noise and error to be reliably filtered out. The Alignment Tool matches the highly expressed sample sequences 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 be accurately evaluated.

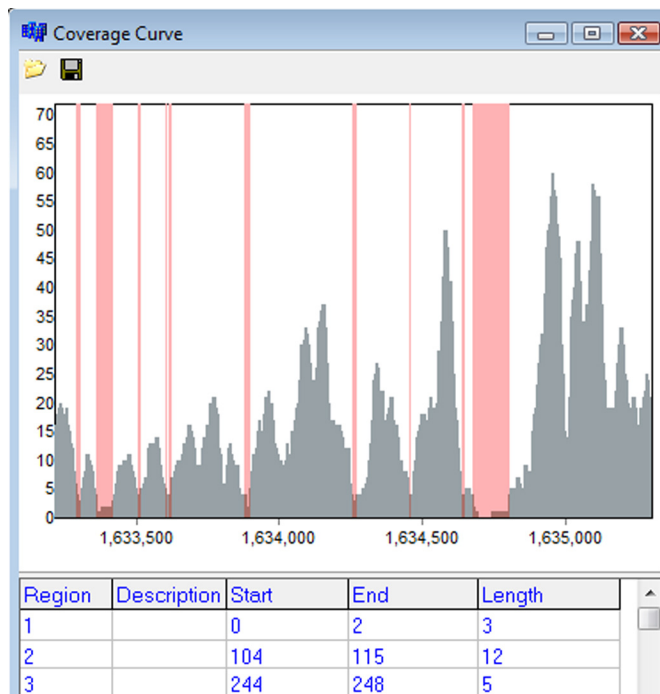




Figure 1

Figure 1: Coverage across the entire genome is shown in the Coverage Curve. Reference base position and aligned base count are plotted on the x- and y-axes respectively. Regions with coverage that falls below a user-set threshold are highlighted in red. The chart beneath the curve contains information about these low coverage regions. Sequences of the low coverage regions can be saved by clicking on the Save () icon. These regions can be useful for indicating regions where PCR may have failed or where large indels have been detected.

Procedure

1. Open NextGENe's Run Wizard by clicking on the  icon in the main toolbar.
2. Select Instrument Type.
3. Select "Transcriptome" under Application Type
 - a. Sequence Condensation and Sequence Alignment are automatically selected.
4. Click Next to upload Sample and Reference files.
5. Browse to select sample file(s).
 - a. If sample file is not in fasta format, use the Format Conversion Tool to convert to fasta.
 - b. 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.
6. Browse to select reference file(s).
7. Specify output location and file name.
8. Click Next to continue to Condensation Settings.
 - a. For projects not using Condensation, clicking Next will open Sequence Alignment settings automatically.
9. Choose appropriate Condensation Settings and click next to continue to Sequence Alignment settings.
10. Click Finish and select Run NextGENe to begin processing project, or Create More Projects to set up more projects, or select Exit Wizard to close the Run Wizard.

Results

The Condensation Tool generates elongated consensus sequences with most of the random and systematic errors removed from the analysis without rejecting SNPs/Indels. After one cycle of condensation, read lengths can be nearly doubled. Additional cycles further elongate the reads.

After the reads were statistically polished, the Sequence Alignment Tool was used to align reads to the transcriptome in order to determine coverage. When aligning transcriptome sequence reads to the transcriptome reference, expression results are often 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 3.

Discussion

NextGENE is a powerful tool for the analysis of transcriptome data produced by massively parallel genome analyzers. Errors within the reads are removed and the reads are elongated with the Condensation Tool prior to alignment to the reference transcriptome.

NextGENE is a versatile software package designed for the new era of high through-put genome sequencing, supporting the Illumina Genome Analyzer, the Applied Biosystems SOLiD™ System and the Genome Sequencer FLX System from Roche Applied Science.

NextGENE also includes software applications for a variety of other application types including other expression studies like Digital Gene Expression, microRNA studies and SAGE, as well as de novo assembly, SNP and indel detection, and ChIP-Seq.

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.

Trademarks are property of their respective owners.

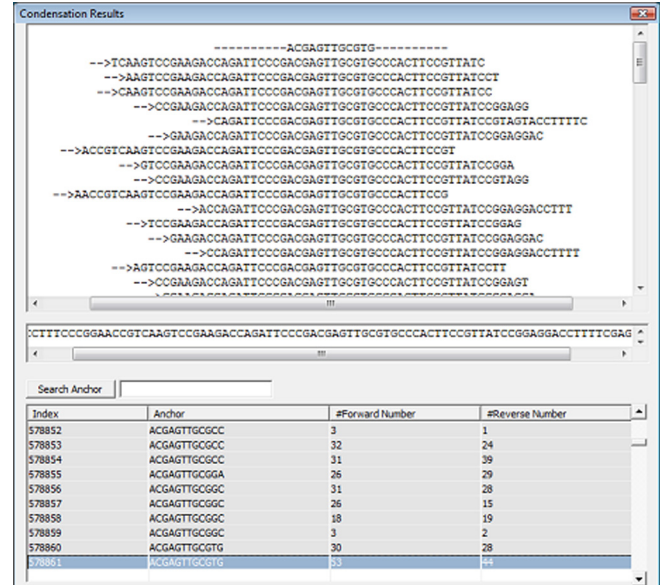


Figure 2

Figure 2: 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 long as 84 bps.

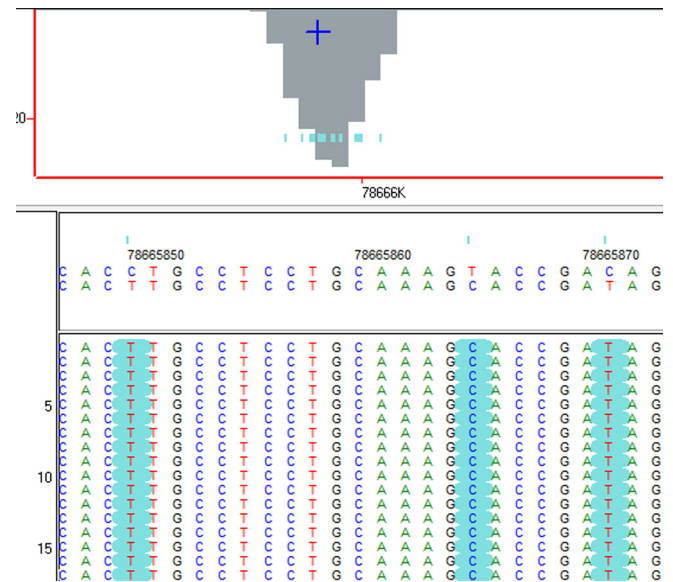


Figure 3

Figure 3: 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.