NextGENe®

December 2008

by SoftGenetics

Small RNA Analysis of Short Sequence Reads With NextGENe® Software

Megan Manion, Kevin LeVan, Ni Shouyong, CS Jonathan Liu

Introduction

Next Generation sequencing technologies such as the Applied Biosystems SOLiD™ System Illumina® Genome Analyzer (Solexa), and the Genome Sequencer FLX System from Roche Applied Science (454 Life Sciences) present promising opportunities for evaluating the expression of known small RNAs as well as revealing novel small RNAs (5). However, the volume of data and high error rates of these systems require efficient and effective software for analysis.

RNA are generally single-stranded nucleic acid molecules that are used to code for amino acids. However, not all RNAs are involved in protein coding. Several types of small RNAs have been identified including small interfering RNAs (siRNAs) and microRNAs (miRNAs). These are short RNA molecules (~21-26 nucleotides) that function as regulators of gene expression. These small regulatory RNAs target specific mRNA sequences by complementary base pair bonding, and are aided by interaction with a multi-protein complex known as an RNA-induced silencing complex (RISC). Other types of small RNA include small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA) which are involved in mRNA splicing and ribosomal RNA modification, respectively.

Short interfering RNAs (siRNAs) function to silence gene expression by bonding with target mRNA strands. They have been observed to be triggered by transgenes, microinjected RNA, transposons and viruses. For this reason, they are considered to be a host defense mechanism against foreign nucleic acids. siRNAs offer a novel method for disease treatment by selectively disabling genes that cause disease (1). MicroRNAs (miRNAs) are more recently discovered short (~21 bases) non-coding RNA strands that are involved in gene regulation of functions such as cell growth, differentiation and developmental timing. They have been shown to be correlated with various disease conditions such as several forms of cancer as well as autism spectrum disorders (ASD) (2), and diabetes (3). Since miRNAs are involved in the regulation of cell growth and division, which is a key element to cancer, they are a major research target for cancer studies (4). Because of this, miRNAs could have valuable implications for disease diagnosis and treatment options.

NextGENe utilizes massively parallel sequencing reads to determine expression levels by producing accurate hit counts. Reads are aligned to a whole reference genome to determine transcript locations. Regions of high coverage are used to indicate transcript regions. These regions of the genome can be saved and used as a reference transcript sequence. Samples are then aligned to the transcript reference and coverage counts are made for each transcript. Prior to each alignment, reads can be processed by NextGENe's exclusive, patent pending Condensation Tool to reduce errors and merge identical reads.

Procedure

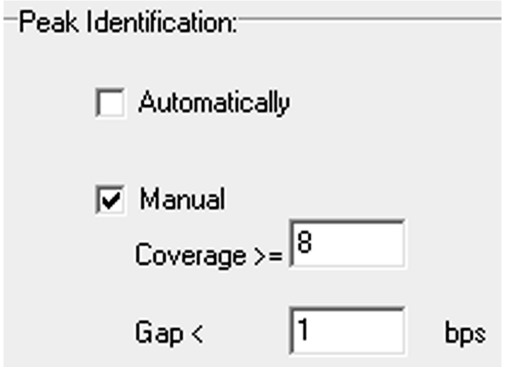
Align Sample Reads to Reference Genome

1. Open NextGENe's Run Wizard.
2. Select Instrument Type.
3. Select "Transcriptome" for Application Type.
4. Click "Next" to Load Sample and Reference Files
   a. If sample file is not in fasta format, click on "Format Conversion" button to convert file.
5. Load Reference File in GBK format with genome annotation or in fasta format for sequence only.
6. Specify Output Field and File Name.
7. Click Next to proceed to Condensation and Alignment Settings.
   a. Alignment settings include an option to "Allow Ambiguous." For reads that align perfectly at multiple locations, selecting this option allows software to align these reads to each location. Otherwise, reads will be aligned arbitrarily to the first perfect match.
   b. It is recommended to set matching base percentage near 90% to allow for two base pair errors for reads that are 25 base pairs in length.
8. Choose appropriate settings and click "Finish."
9. Choose "Run NextGENe" to begin processing project.

Identify Transcript Locations

Once NextGENe completes the alignment, results will be shown automatically in the Sequence Alignment window. Then NextGENe's Peak Identification Tool can be used to identify transcript locations.

1. In Sequence Alignment window, select "Peak Identification" from the Tools menu.
2. Choose option to manually input Peak Identification Settings.



Peak Identification:

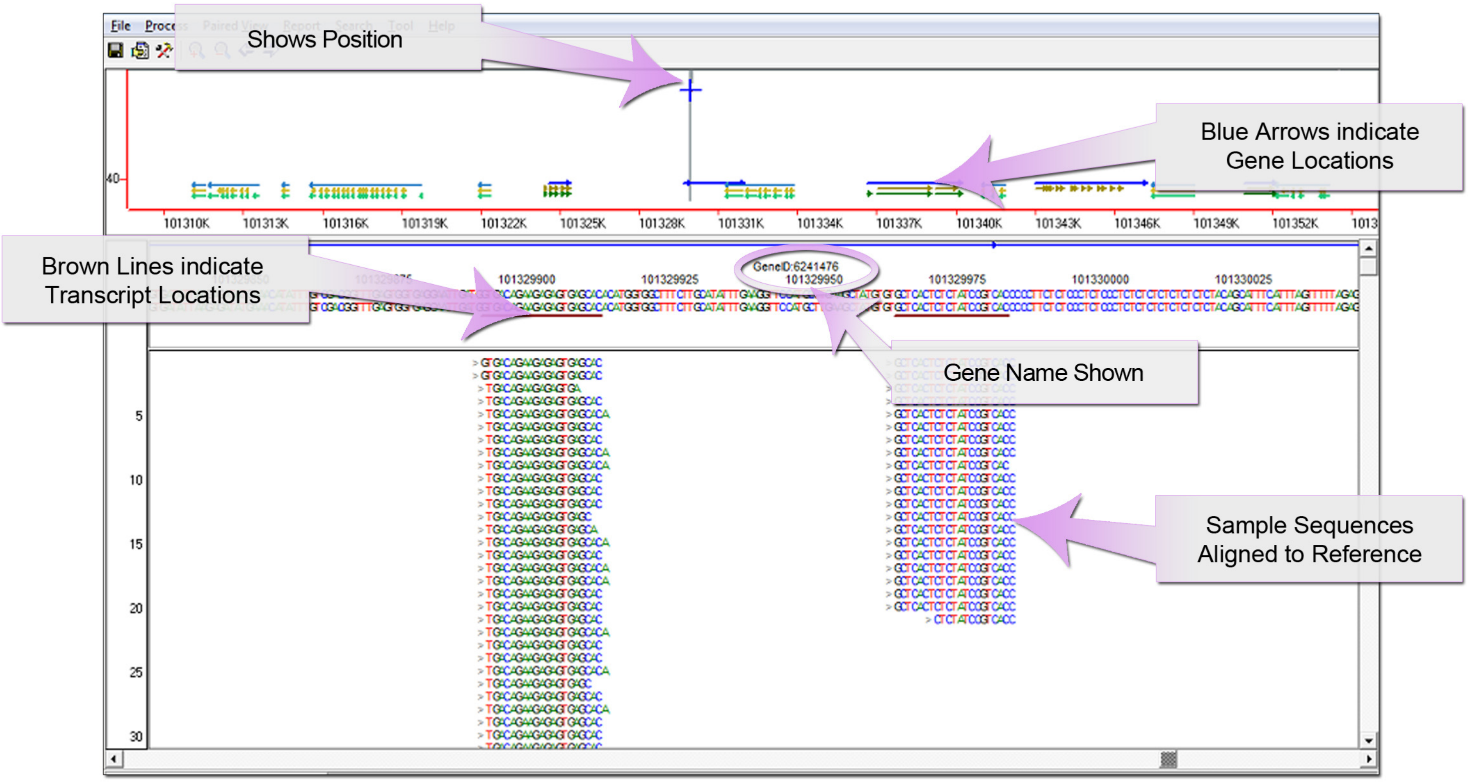
☐ Automatically

☑ Manual

Coverage >= 8

Gap < 1 bps

Figure 1: Manually input Peak Identification Settings for coverage and gap size thresholds or allow software to use automatic settings. The coverage setting refers to the amount of coverage required to be considered a transcript. A coverage threshold of 8 is recommended to detect small RNA peaks for data sets with average coverage. The starting position and ending position are contained within a continuous sequence. The gap size refers to the maximum distance (in base pairs) between two regions that meet the coverage requirement to be considered as one transcript. For microRNA, gap size should be set to zero or one. Gap size can be set higher to locate other, relatively longer, small RNAs.



Shows Position

Blue Arrows indicate Gene Locations

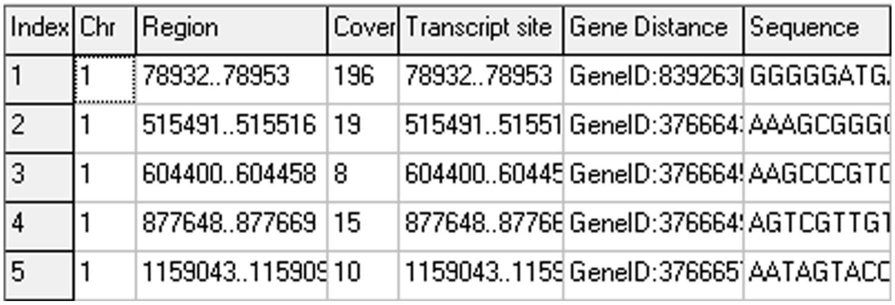
Brown Lines indicate Transcript Locations

Gene Name Shown

Sample Sequences Aligned to Reference

Figure 2: After using Peak Detection Tool, Sequence Alignment window displays brown ticks to indicate regions that meet transcript requirements. This figure shows two small RNA transcripts located within Gene 6241476. Blue arrows are used to indicate gene locations in the reference file. The green and gold arrows below gene indicator identify the mRNA and coding sequence respectively.

Select "Peak Identification Report" to view a list of regions identified by the software as transcripts. Click on the Save icon (💾) in the Peak Identification report window to save these regions as a new file that can be used for aligning sample data to the transcripts.



Index	Chr	Region	Cover	Transcript site	Gene Distance	Sequence
1	1	78932..78953	196	78932..78953	GeneID:839263	GGGGGATG.
2	1	515491..515516	19	515491..51551	GeneID:376664	AAAGCGGG
3	1	604400..604458	8	604400..60445	GeneID:376664	AAGCCCGTC
4	1	877648..877669	15	877648..87766	GeneID:376664	AGTCGTTGT
5	1	1159043..115909	10	1159043..1159	GeneID:376665	AATAGTACC

Figure 3: The Peak Identification Report contains information about all regions of the reference that meet the requirements to be considered transcripts. Click on the Save icon (💾) to save these regions as a new file which will be used as the transcript reference.

Align Samples to Transcripts

To align samples to the transcript file, the same general procedure is used as when aligning to the whole genome reference. However, the file created from the Peak Identification report, which contains only the transcripts, should be used as the reference file.

Results

After the samples are aligned with the transcript reference file, the results will automatically be shown in the sequence alignment window to provide a graphic view of the number of sample reads aligned to each transcript. To view numerical information about expressions levels for each transcript, select "Expression Report" from the Reports section of the tool bar.

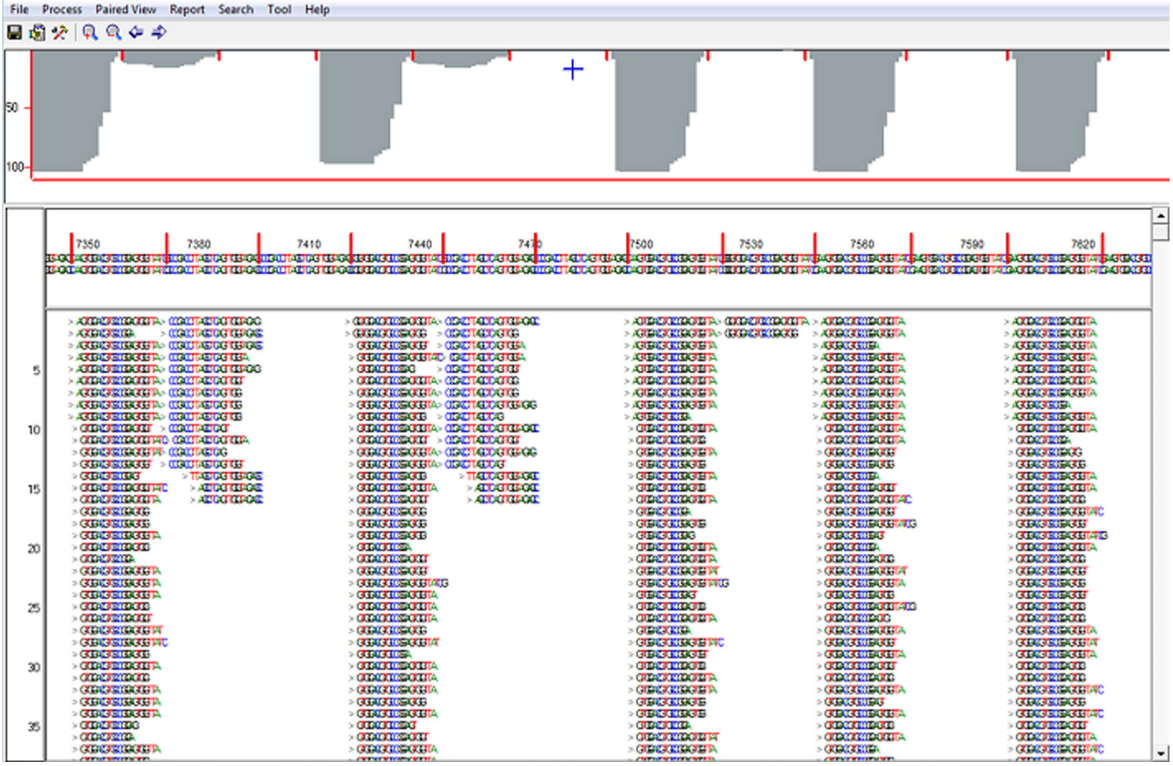
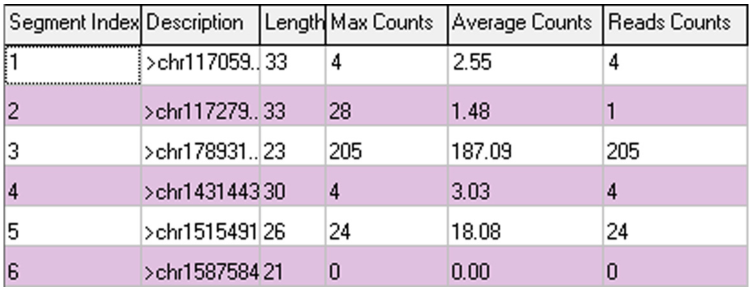


Figure 4: Once NextGENe completes aligning sample file(s) to the transcript reference file, the results are shown in the sequence alignment window which provides a graphic representation of expression levels for each transcript. Red lines indicate transcript boundaries. Sequence reads that align with each transcript are shown beneath where they align. Gray bars indicate coverage (expression level).



Segment Index	Description	Length	Max Counts	Average Counts	Reads Counts
1	>chr117059..	33	4	2.55	4
2	>chr117279..	33	28	1.48	1
3	>chr178931..	23	205	187.09	205
4	>chr1431443	30	4	3.03	4
5	>chr1515491	26	24	18.08	24
6	>chr1587584	21	0	0.00	0

Figure 5: The Expression report displays quantitative information about each segment (transcript) including its length, the maximum and average count numbers and the read counts.

The expression levels can be normalized to the total read numbers aligned to the transcripts. P-values can also be determined from the normalized expression ratio. The null hypothesis (samples are drawn from the same population) is postulated when the two expression ratios are the same. Error estimates of the read number can be produced using sqrt(N). A t-test can be used to determine the false discovery rate.

In addition to providing coverage information, NextGENe's Sequence Alignment Tool also identifies SNPs and indels. Information about these variations is displayed in the Mutation Report which can be accessed by selecting "Mutation Report" from the Reports section of the Sequence Alignment tool bar.

Discussion

SoftGenetics' NextGENe software is a data analysis tool designed to handle the unique problems of Next Generation sequencing platforms such as the Illumina® Genome Analyzer (Solexa), the Genome Sequencer FLX System from Roche Applied Science (454 Life Sciences) and the Applied Biosystems SOLiD™ System. These high-throughput platforms are capable of producing several hundred million reads per sequencing run, creating great opportunity for large scale sequencing projects. However, data from these platforms also presents challenges for analysis, such as processing the overwhelming volume of data, short read lengths and high error rates. NextGENe software is specifically tailored to address these challenges.

NextGENe's small RNA analysis tools are able to process small RNA data from Next Generation sequencing platforms to produce accurate analysis of expression levels. Sample data are aligned to a reference genome to determine transcript locations and then realigned to a reference file containing only transcript locations to determine coverage for each transcript.

NextGENe also includes software applications for a variety of expression studies (Digital Gene Expression, transcriptome studies and SAGE) as well as de novo assembly, SNP and indel detection, and ChIP-Seq.

Acknowledgements

We would like to thank Dr. Csaba Koncz from the Max Planck Institute in Munich, Germany for providing data and software requests.

References

1. H Groβhans, W Filipowicz. 2008. Microbiology: The expanding world of small RNAs. Nature 451: 414-416.
2. K Abu-Elneel et al. 2008. Heterogeneous dysregulation of microRNAs acrossthe autism spectrum. Neurogenetics. 9: 153-161.
3. M Poy et al. 2004. A Pancreatic islet-specific microRNA regulates insulin secretion. Nature. 432: 226-230.
4. J Lu et al. 2005. MicroRNA expression profiles classify human cancers. Nature. 435: 834-838.
5. R D Morin et al. 2008. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Research. 18: 610-621.