

Analysis of SAGE Data from Next Generation Sequencers with NextGENe Software

Megan Manion, Kevin LeVan, Shouyong Ni, Jin Yu, Sean Liu, Jacie Wu and Changsheng Jonathan Liu.

Introduction

Next Generation Sequencing systems such as the AB SOLiD™ System, Roche Applied Science's Genome Sequencer FLX System (454 Sequencing) and the Illumina® Genome Analyzer utilizing Solexa Sequencing Technology promise to provide breakthroughs in Digital Gene Expression studies like SAGE (1). SAGE technology measures the counts of sequence tags relative to the genes of interest. The SAGE tags are produced by the restriction enzymes which cut the cDNA while the poly(A) end is bound to the biotin-labeled dT primer. The biotin is captured on beads so that biotin-labeled cDNA strands are bound and all other DNA strands are washed away. This method, in addition to similar techniques of MicroSAGE, LongSAGE, RL-SAGE and SuperSAGE, offers effective tools to analyze absolute expression numbers by counting tags.


All 2nd generation DNA sequencing technologies generate millions to hundreds of millions of the short sequence reads per instrument run, providing powerful solutions for analyzing gene expression. However, the high inherent error rates of these systems, as well as the sheer volume of data produced pose significant challenges for analysis.

SoftGenetics has developed NextGENe software to address these issues, providing researchers and clinicians with a “biologist friendly” analysis tool that provides 99%+ accuracy by statistically polishing high throughput sequencing data as well as unique algorithms to quickly analyze data in an easy-to-use Windows environment.

SAGE Libraries are available that contain lists of sequence tags associated with particular genes. NextGENe can load these libraries as a reference and align the sequence reads to the appropriate sequence tags. The alignment to the tag library is only performed in the forward orientation of the sequences, no reverse complementation is implemented. New gene tags that are not in the library are also reported. These tags can be added to the reference library as novel tags.

Figure 1: The Sequence Alignment Tool has a Whole Genome View at the top of the screen, which shows each sequence of the library. Placing the mouse over the library activates a yellow box containing the biological information for the tag that is currently at the cursor. The bottom of the screen contains all reads as they have been aligned to the library.

Procedure

1. Open NextGENe's Run Wizard by clicking on the  in the main toolbar.
2. Select Instrument Type.
3. Select SAGE under Application Type.
 - a. Sequence Alignment is automatically selected under Steps.
4. Click Next to input sample and reference files.
5. Browse to load sample data.
 - a. If sample data is not in fasta format use the Format Conversion Tool to convert to fasta.
6. In Reference File field, browse to input reference file in GBK or fasta format.
7. Browse to specify output file location and name
8. Click Next to determine settings for alignment process
9. Choose appropriate Alignment Settings and click Finish.
 - a. Under File Type select Load SAGE Expression Data
10. Select Run NextGENe to begin processing the project.

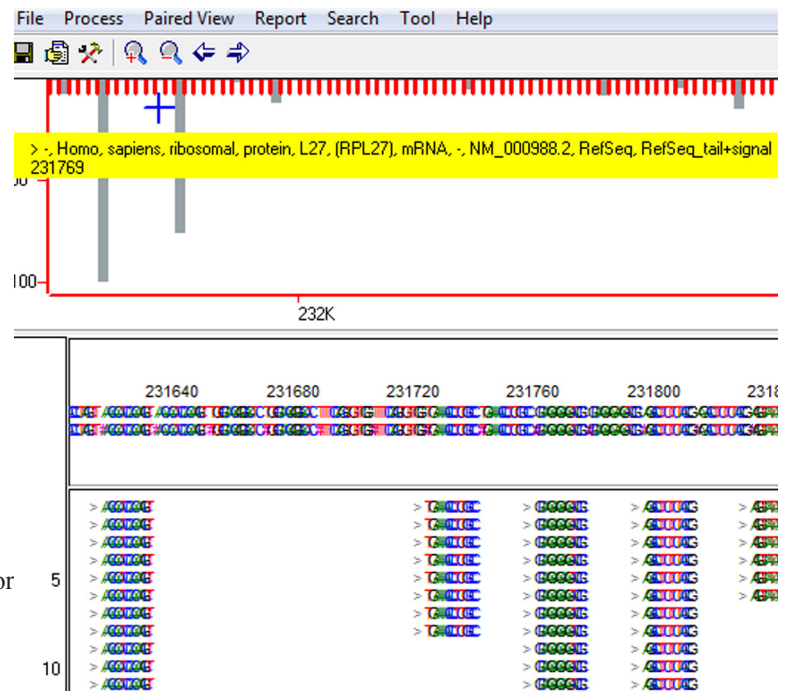


Figure 1

Results

Once the project is completed, the results will be shown automatically in the Sequence Alignment Tool. The display of the Sequence Alignment Tool is divided into two major sections, the Whole Genome View and the Alignment View. The Whole Genome View shows the reference library range with coverage for each sequence indicated by the gray bars, and biological information can be shown for each sequence. The Alignment View shows all sequence reads as they align to a specific sequence tag contained in the library.

An Expression Report can be generated showing the gene name, the sequence tag, the number of times the sequence was observed, the number of gene ambiguities, and more. Multiple genes within this SAGE library are identified by the same sequence tag. This information is tracked within the report by displaying the number of gene ambiguities.

Figure 2: The expression report shows information such as the index number, position, gene name, sequence, counts, number of ambiguities and expression for each tag.

Index	Position	Gene Name	Chromosome	Sequence	Occuring Counts	# of Gene Ambiguities	Expression
457	12q23-q	Uracil-DNA		GGCCCCA	0	1(456)	18
458	12q23-q	Uracil-DNA		AGATATAT	72	3(459,32778,32779)	18
459	3a	Splicing		AGATATAT	0	3(458,32778,32779)	18
460	3a	Splicing		CCCCAATG	67	1(461)	33

Figure 2

Novel sequences are contained in this dataset. The sample reads that match sequences contained in the library are aligned appropriately, but novel sequences that are not contained in the library are also recorded. By setting a minimum threshold for new sequences, all sequences found at a frequency above the threshold are added to the end of the reference library, sample reads are aligned to them, and the Expression report shows these sequences as New Genes.

Additional information, including reports comparing similar projects, and lists of unmatched reads can also be produced.

Discussion

NextGENE is a versatile software package designed for 2nd generation genome sequencers, supporting the Illumina Genome Analyzer, the Applied Biosystems SOLiD System and the Genome Sequencer FLX System from Roche Applied Science (454 Sequencing). NextGENE is able to efficiently analyze SAGE data by producing accurate tag counts and providing useful reports to display results.

In addition to SAGE analysis, NextGENE can be used for other expression studies like transcriptome analysis, ChIP-Seq and small RNA analysis, as well as SNP and Indel detection and de novo assembly. NextGENE's Condensation Tool is used polish and lengthen the short reads into longer reads with superior quality. Condensation increases the reliability of SNP and Indel detection, and also assists with the accuracy of de novo assembly. NextGENE software is able to assemble 36 bps short reads to produce large contigs up to 100kp.

Acknowledgements

We would like to thank Victor Velculescu from the Kimmel Cancer Center of Johns Hopkins University, Oleg Evgrafov and James Knowles from the Keck School of Medicine of University of Southern California, for their collaboration with the development of this software.

References

1. V. E. Velculescu et al. 1995. Serial Analysis of Gene Expression. *Science*. 270: 484-7.
2. M. Schena et al. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 270: 467-70.

Trademarks are property of their respective owners.