

Analysis of Digital Gene Expression Using Short Sequence Reads With NextGENe Software

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

Introduction

Gene expression studies are often currently analyzed using the technologies of microarray and DNA sequencing such as Serial Analysis of Gene Expression, or SAGE (1). In the microarray experiment, cDNA probes are hybridized to the sequence targets of the gene of interest on the microarray, where many probes of interests are located in different spots (2). The cDNA is labeled with a chromophore, and fluorescence intensity is proportional to the cDNA concentration of the probes. SAGE technology measures the counts of the sequence tags relative to the genes of interest. The SAGE tags are produced from the restriction enzymes cut to the cDNA with the poly-A end bounding to the biotin-labeled dT primer. The portion bound to the solid surface will be kept. The NlaIII restriction enzyme of SAGE targeting CATG, in addition to the techniques such as MicroSAGE, LongSAGE, RL-SAGE, SuperSAGE and more offer powerful solutions to read the absolute expression number by counting the tags.

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 and the Applied Biosystem SOLiD™ System uses emulsion PCR and sequencing by ligation. Both of these systems can produce the short reads ideal for analyzing gene expression. The NextGENe software package takes full advantage of the short sequencing reads and has tools for analyzing the SAGE tags.

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. Digital gene expression reports are created to show the sequence of each tag, the coverage, gene names, and the location in the genome. New gene tags that are not in the library are also reported.

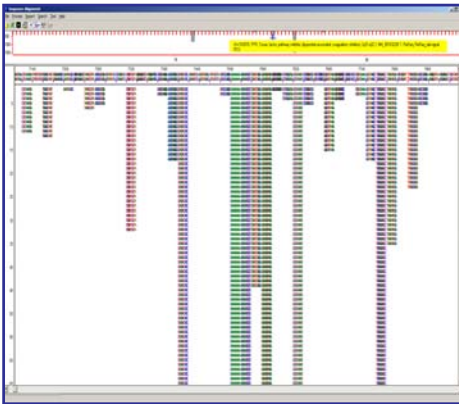


Figure 1: The Sequence Alignment Tool has a Whole Genome View at the top of the screen, which shows each sequence of the library. Mousing over the library activates a yellow box containing the biological information for the tag that is currently at the cursor. The bottom of the screen contains all reads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

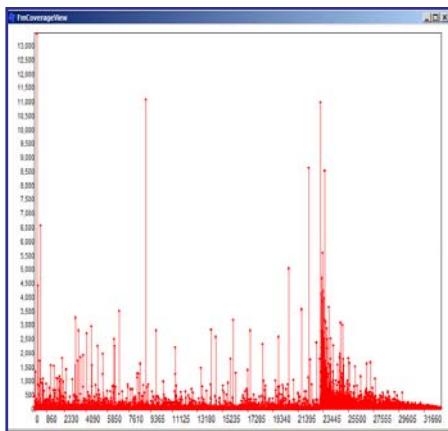


Figure 2: NextGENe produces a chart with the sequence tag number on the x-axis and coverage of each tag on the y-axis. Most tags are expressed less than 500 times, but several genes show very high expression levels. Positions on this chart after 23K are new genes that have been added to the reference file because the sequence was found many times. Several of these new sequences were found in the project with expression levels above 4000.

Procedure

1. Open the Sequence Alignment Tool and select Load Data.
2. In GBK File field, select Open and choose the Gene Expression library to use as reference.
3. In Sample File field, select Open and choose the sample file to be analyzed.
4. Choose Settings from the Process drop-down menu.
 - a. Set "Matching Base Number >" field to the length of the tag in library.
 - b. Select Load Expression Data.
 - c. Set Extract Bases range to start and end position of tag within the sequence reads. (By having the length of bases that are extracted equal to the matching number of bases, only reads that are perfect matches will align with a given sequence from the library.)
5. Close Settings and click the Run button. When analysis is complete, a message will appear showing the start and end time of analysis.
6. Select Report and click on Expression Report.

Results

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 all possible sequences that are contained in the library in sequential order, coverage for each sequence is 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.

Index	Gene Name	Sequence	Occuring Count	# of Gene Ambiguities	Expression
0	TRIP11	TATTTGTGGCAATTAT	4	0	4
3	XRCC4	CCTATAATTACATAAAT	4	2	1
4	PANK1	ATGAATCTTCTTTTCTC	11	2	3
5	PLAGL1	ATCATAATGTTAACTAA	19	1	9
6	ASPM	GAAGAAATCACAAATCC	45	0	45
7	DCUN1D1	ATTTTGAATAGACTAG	1	0	1
8	RNASET2	GGACCTGGCCGCCCAG	2	0	2
9	SEPT6	TCAATTTTCAATAAAT	2	1	1
10	KLHL5	ATGTTTTGCATTAAAT	72	2	24
11	CALD1	TTCTGTGAATCTGCCAT	460	4	92
12	ACSL3	TCGACTAGTTACTTTG	6	1	3
13	TCP1	CTACCCCTTTCAAACTCA	20	1	10
14	FEZ1	GTGGGGGATGTTTTTA	1	0	1
18	MAP3K3	GTATTGTGGAACTGTG	22	1	11
20	ANK1	CTCTGTATGGGAGAGA	1	7	0
23	THEM5	CCTGTAATCCAGCATT	57	1	28
24	MITF	GTTAAGCAACCATATAG	17	5	2
28	PAK4	AATGTCCGAAGAGTGCC	13	5	2

Figure 3: A report is automatically created showing the Expression Results. The table includes the gene name for each tag found in the library, in addition to the number of occurrence for each sequence, number of ambiguities due to sequence duplication within the library and more.

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 a Mutation Output report, a chart showing coverage for each gene, and lists of unmatched reads, can also be produced.

Discussion

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 SAGE, microRNA and Transcriptome analyses.

In addition, NextGENe can be used for SNP and Indel detection as well as *de novo* assembly. Tools are present within NextGENe to condense, polish and lengthen the short reads into longer reads. The sequence errors of short reads are polished with the statistical approach through the Condensation Assembly Tool. The reads of 36 bps are condensed to 60 bps reads called unitigs. The unitigs increase the reliability of SNP and Indel detection, and also assists with the accuracy of *de novo* assembly. We are able to assembly 36 bps short reads to about 1KB fragments ending with repeat sequences.

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 collaborating 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.

Additional Application Notes for the Analysis of “Next Generation Data” with NextGENe software:

- SNP and Micro Indel Detection with NextGENe Software.
- *de novo* Assembly of Short Sequence Reads with NextGENe Software.
- Transcriptome Analysis Using NextGENe Software.
- NextGENe Software Tools for Analysis of Protein-DNA Interactions by ChIPSeq.

Request a copy today: info@biogene.com