NextGENe™
by SoftGenetics
September 2008

ChIP-Seq Analysis of SOLiD™ Sequence Reads with NextGENe™ Software

Kevin LeVan, Jacie Wu, Haitao Ren, Shouyong Ni, Teresa Snyder-Leiby and ChangSheng Jonathan Liu

Introduction

Vital cell functions influenced by DNA binding proteins include: gene transcription, DNA replication and recombination, repair, segregation, chromosomal stability, cell cycle progression, and epigenetic silencing. ChIP-Seq identifies protein binding sites in DNA by combining chromatin immunoprecipitation of Protein-DNA complexes with high throughput sequencing. Binding affinities of a protein to different DNA sites can also be compared by quantifying the copy number of a given sequence1,2.

Next generation DNA sequencing strategies have lowered the costs of sequencing, increasing the speed and amount of information gathered. These instruments generate billions of bases in just a few days for just a few thousand dollars. The Applied Biosystems SOLiD™ System generates more than 9 gigabases per run with reading lengths between 25-35 bps.

One of the limitations to the ChIP-Seq technique is the ability to sort and condense the millions of short reads generated from sequencing the enriched DNA and mapping the Protein/DNA interaction sites in genomes. NextGENe software was developed to solve these current problems. The software is able to use a Condensation technique to correct calling errors and elongate reads while maintaining allele frequency information. The Alignment tool is designed to match the sequence reads to a reference; mapping the location of DNA binding sites in the reference sequence.

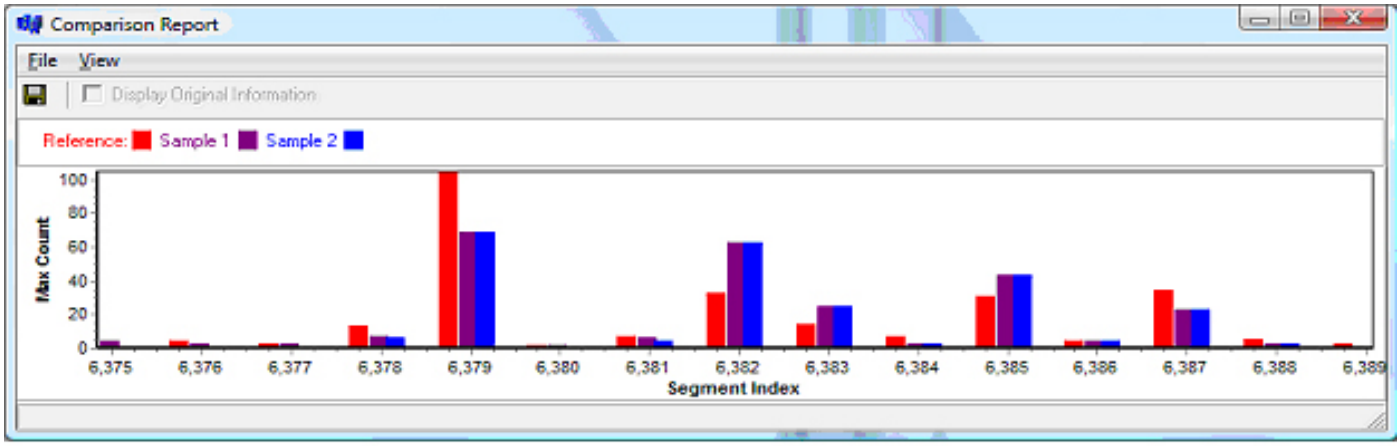

Figure 1: The comparison report can open multiple projects simultaneously to compare expression levels. In this example, the genome is divided into multiple 10 kbps segments. For a reference that is divided into multiple genes, expression would be shown for each gene for applications such as transcriptome analysis.

Condensation Methodology to Produce Contigs

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 or 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%.

Procedure

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 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 ChIP-Seq for Application Type. This enables the Condensation and Assembly 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

Coverage is displayed in the gray histogram at the top of the screen. Clicking on the area of interest in the histogram links to that portion of the reference sequence. The consensus sequence is directly below the reference sequence. The sequence reads used to produce the consensus sequence are beneath the solid line.

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. Protein binding sequences are located across this 9K of the reference sequence in addition to others in genome.

Segment Index	Description	Max Count	Average Count	Reads Count
1505	15040001-15050000	5	0	20
1506	15050001-15060000	1	0	1
1507	15060001-15070000	2	0	3
1508	15070001-15080000	7	0	41
1509	15080001-15090000	1587	17	5182
1510	15090001-15100000	8	0	58
1511	15100001-15110000	5	0	8
1512	15110001-15120000	1	0	2
1513	15120001-15130000	0	0	0
1514	15130001-15140000	2	0	2
1515	15140001-15150000	1	0	1
1516	15150001-15160000	0	0	0
1517	15160001-15170000	2	0	4
1518	15170001-15180000	0	0	0
1519	15180001-15190000	14	0	62

Figure 3: NextGENe produces an expression report that shows the level of coverage for genes or segments of the reference. Max Count is the depth of coverage within each segment and Reads Count is the total number of reads in each segment.

Discussion

NextGENe is a versatile software package capable of loading an entire human chromosome for the detection of regions of Protein-DNA binding. The color-space sequence reads can be condensed to simultaneously correct many of the color-call errors and elongate the reads to increase the accuracy of alignment. The software shows the aligned reads in both base-space and color-space as shown in Figure 2. In addition, reports can be generated showing the coverage within each region of DNA-protein interaction.

NextGENe also contains tools for applications including SNP discovery, assembly and other expression analyses.

References

1. T. Goldfarb and E. Alani. 2004. Chromatin immunoprecipitation to investigate protein-DNA interactions during genetic recombination. Methods Mol Biol 262:223-227

2. David S. Jonson, Ali Mortazavi, Richard M. Myers, Barbara Wold. 2007. Genome-Wide Mapping of in vivo Protein-DNA Interactions. Science 316:1497