

ChIP-Seq Analysis with NextGENe Software

Megan Manion, Kevin LeVan, Shouyong Ni, Yiqiong Jacie Wu and CS Jonathan Liu

Introduction

ChIP-Seq studies are used to analyze the interaction between DNA and proteins. ChIP-Seq methodology uses the same chromatin immunoprecipitation technique as ChIP-chip but uses massively parallel sequencing methods as opposed to microarray methods for sequencing (1). The chromatin immunoprecipitation technique involves several steps. DNA is crosslinked with the protein of interest, then sheared to produce small DNA fragments bound to the protein. An antibody against the protein of interest is used to filter out the DNA fragments that are enriched with the protein. The protein-DNA complex is reverse-crosslinked to produce a sample of the DNA regions where the protein was attached. This DNA is sequenced using Next Generation sequencers.

High throughput sequencing technologies such as the Genome Sequencer FLX from Roche Applied Science (454 Sequencing), the SOLiD™ System from Applied Biosystems and the Illumina® Genome Analyzer utilizing Solexa Sequencing Technology have made analyzing DNA binding sites for an entire genome possible at just a fraction of the cost and time of microarray techniques. The large number of reads produced per run leads to greater specificity and an improved signal-to-noise ratio as well.

This technique focuses on identifying the location of protein binding sites within DNA and can be used for purposes such as the detection of transcription factor binding sites and methylation sites (2). Information from ChIP-Seq analysis provides an enhanced understanding of the interactions between DNA and proteins, including their involvement in gene regulation, which has important implications for the study of many diseases.

NextGENe provides a software module specifically designed for ChIP-Seq analysis. This application utilizes coverage information for the detection of peaks to indicate protein binding sites as well as a specialized report to provide information about each peak region.

Procedure

To use the ChIP-Seq application of NextGENe, sample reads are aligned to a reference. Coverage thresholds are used to identify peak regions and counts are produced for each.

1. Choose Instrument Type.
2. Choose ChIP-Seq for Application Type.
 - a. Sequence Condensation and Sequence Alignment are automatically selected.
 - b. Sequence Condensation can be deselected to align raw reads.
3. Click “Next” to input data files.
4. Browse to select sample data file(s) in fasta format, or csfasta for SOLiD System data.
 - a. If sample file is not in fasta format, use the Format Conversion Tool to convert.
5. Browse to select reference file(s) in fasta or gbk format.
6. Specify output location and file name.
7. Click “Next” to continue to Sequence Condensation settings (when applicable).
 - a. For projects not utilizing the condensation function, clicking “Next” at this step will open Sequence Alignment settings.
8. Choose appropriate settings and click “Next” to continue to Sequence Alignment settings.
9. Choose appropriate settings and click “Finish” to continue.
10. Choose “Run NextGENe” to begin processing project.

Once NextGENe completes alignment of the samples to the reference, the results are automatically displayed in the Sequence Alignment window. When “ChIP-Seq” is selected as the Application Type, automatic peak detection is applied during the initial processing and peak regions are indicated in the Sequence Alignment Window upon project completion.

Peak Identification Tool

The Peak Identification Tool can be used to detect peak regions for any alignment project when “ChIP-Seq” is not selected as the Application Type. This tool can also be used to manually set Peak Identification Settings. Manual coverage and gap size thresholds can be set or software can identify peaks automatically.

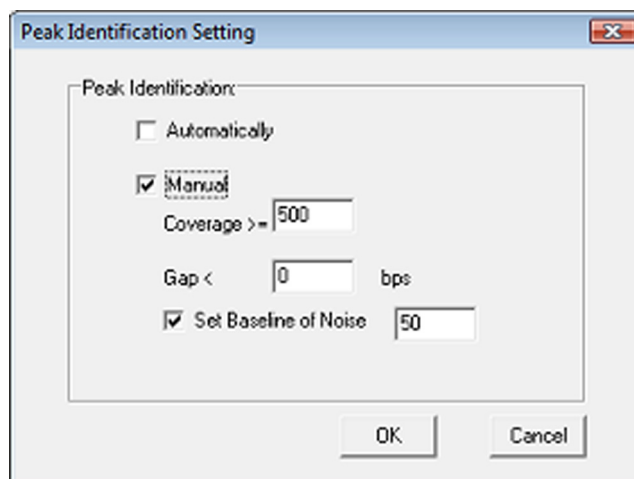


Figure 1: Choose to manually set guidelines for peak identification or allow software to automatically identify peaks. When manual settings are used, coverage refers to the coverage threshold for a position to be considered part of a peak. The gap setting refers to the maximum number of bases between regions that meet the coverage threshold to be considered as one continuous peak. A baseline for the noise level can also be set.

Results

When a ChIP-Seq project has been completed, the results are shown in the Sequence Alignment window. Brown bars are shown to graphically illustrate regions where peaks have been identified.

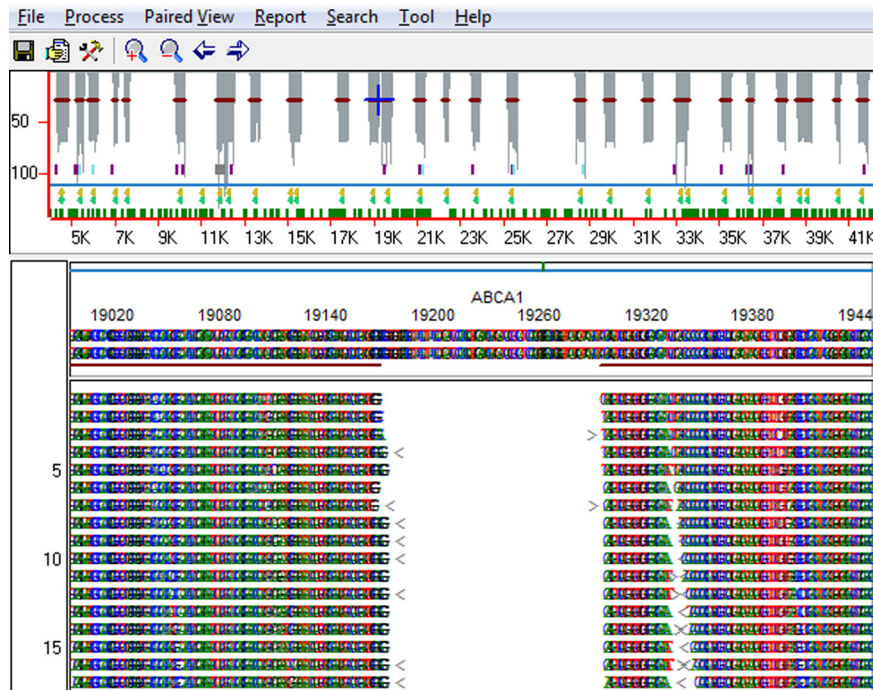


Figure 2: After using Peak Identification, or after NextGENe completes processing of a ChIP-Seq project, the Sequence Alignment window shows brown bars across regions identified as peaks.

The Peak Identification Report lists all regions identified as peaks along with the length of the entire region, the chromosome location and gene location, the coverage, the central transcript site and the sequence of the region. The report can be sorted by any of the columns by double-clicking on the column heading and can be saved by clicking on the save () icon.

Peak Identification Report							
Index	Chr	Region	Length	Coverage	Transcript site	Gene Distance	Sequence
1	12	2461..2708	248	50	2534..2634	ABCC9(+2461)	ATTAATCAGAAAG
2	12	10576..10985	410	55	10730..10830	ABCC9(+10576)	AATAAGAAATTAT
3	12	30194..30729	536	50	30411..30511	ABCC9(+30194)	GGTGGAAAGAAAT
4	12	42226..42746	521	55	42436..42535	ABCC9(+42226)	GTTCTCCATCCCT
5	12	48584..49019	436	74	48751..48851	ABCC9(+48584)	CCTCAGCAATTTA

Figure 3: The Peak Identification Report provides information about each region that has been designated as a peak.

Discussion

NextGENe provides accurate analysis of ChIP-Seq data. Depth of coverage is evaluated to identify high coverage regions as peaks which indicate protein binding sites. NextGENe's unique Condensation Tool™ improves read quality by lengthening reads and statistically removing instrument errors.

NextGENe also includes software modules for de novo assembly, SNP and Indel Detection, Transcriptome analysis, small RNA discovery and quantification and SAGE analysis. Both *de novo* assembly and SNP/Indel Detection can be analyzed with or without the use of paired reads.

References

1. Jothi R, et al. 2008. Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data. *Nucleic Acids Research*. 36(16): 5221–5231.
2. Barski A, et al. 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*. 129(4): 823-837.

Trademarks are property of their respective owners.