ChIP-Seq Data Analysis

ChIP-Seq is a powerful method to identify genome-wide DNA binding sites for a protein of interest. This technical note describes a simple approach to building annotated tag and count tables from ChIP-Seq data sets from the Illumina Genome Analyzer.

INTRODUCTION

ChIP-Seq data is less complex than other types of massively parallel sequencing data since analysis consists of determining a census count of tags from a relatively purified DNA sample. There are, however, some informatics steps that must be followed to extract meaningful data from the raw sequence reads.

The procedures described in this technical note are not intended to be complete and rigorous, but are meant to provide a starting point for researchers to successfully generate counts of sequence tags at various positions in the genome. This starting point is sufficiently flexible to facilitate novel or previously described techniques for the analysis of output data.

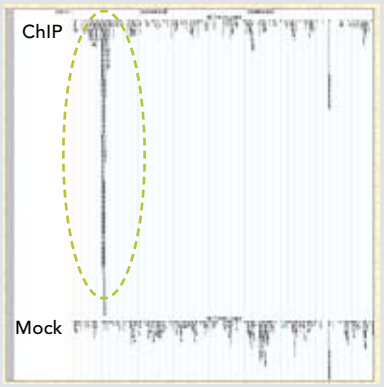
In addition to descriptions of how data are handled by Illumina Genome Analyzer Pipeline Software, several publicly available analysis algorithms for ChIP-Seq data analysis are discussed.

STANDARD ANALYSIS PROCESS

Illumina ChIP-Seq data produced from the Genome Analyzer are transitioned through several phases to prepare them for thorough analysis. The approach outlined below provides data display amenable for visual analysis only, which means that a researcher may have to be familiar with, or have prior expectations of a certain region of the genome to be enriched for the ChIP sequences.

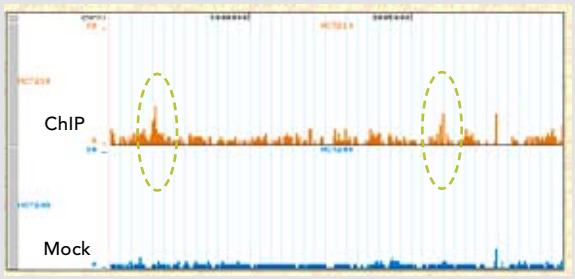
1. First, the data enter the Image Analysis and Base Calling phases. Here, the actual sequence data are generated from the images acquired during sequencing by synthesis chemistry on the Genome Analyzer.
2. The short sequence reads are then aligned to the genome using ELAND. The ELAND output sequentially lists all of the sequence reads with their respective genomic coordinates. ELAND is described fully in the Genome Analyzer Pipeline Software User Guide.
3. Read data that are uniquely aligned to a genome can be viewed as a custom track in the UCSC genome browser. The track can be submitted in either a BED *(Figure 1)* or WIG *(Figure 2)* format. More information on UCSC custom tracks is available at http://genome.ucsc.edu/goldenPath/help/customTrack.html.
   a. The BED format is a simple text file that contains the chromosomal start and end positions. A track in the

FIGURE 1: CHIP-SEQ TAGS IN BED FORMAT, DISPLAYED IN UCSC BROWSER



Example of custom tracks submitted in BED format (upper track is from ChIP sample and lower track is from mock control sample). The peak on the left in the ChIP sample (green circle) is significant. However, the peak on the right side is detected in both the ChIP and mock samples and is not significant.

FIGURE 2: CHIP-SEQ TAGS IN WIG FORMAT



Shown is the UCSC Browser display of WIG format tracks for the same region shown in Figure 1. In this format, peaks are less obvious (green circles).

BED format can be easily generated from GERALD *_realign.txt files. GERALD is described fully in the Genome Analyzer Pipeline Software User Guide. The following is example code that would create a BED format text file from the sequence tags in lane 3, assuming sequence reads that are 25 nucleotides in length:

```
cat s_3_????_realign.txt|egrep -v '^#'| \
perl -ane 'if (@F>3){$_=~/(chr.+):(\d+)\
s([F|R])/;print $1,"\t",$2,"\t",($2+25),"\n"}'\
> s3_customTrack.txt
```

b. The wiggle (WIG) format allows display of continuous-valued data in a track format. This display type is most useful for examining GC percent, probability scores, and transcriptome data. More information on wiggle format is available at http://genome.ucsc.edu/google/goldenPath/help/wiggle.html.

The BED format provides enhanced visual presentation of the data compared to the WIG format. However, there are limitations on the file size that can be uploaded to the UCSC Browser. Since the WIG format is much more compact than the BED format, the WIG format may be useful in some cases where a larger dataset is examined.

ANALYZING CHIP-SEQ CONTROLS

Data from Illumina ChIP and mock control (also referred to as input or IgG control) reads should be analyzed in combination for each individual experiment. These negative controls are important to identify regions showing a biologically relevant over-representation of tags in the ChIP sample versus the control sample. Some investigators have used a ratiometric comparison of experimental to control samples to identify true binding sites.

SOFTWARE REQUIREMENTS FOR CHIP-SEQ ANALYSIS

Illumina Genome Analyzer Pipeline release 0.2.2.6 or greater is required for ChIP-Seq data analysis as described in this document. This includes three essential scripts, described further in the Software User Guide and in the `<pathTopipeline>/docs` directory:

1. Image analysis script
   `<pathTopipeline>/Goat/goat_pipeline.py`
2. Base calling script
   `<pathTopipeline>/Goat/bustard.py`
3. Alignment to reference genome
   `<pathTopipeline>/Gerald/GERALD.pl`

ADDITIONAL RESOURCES FOR CHIP-SEQ DATA ANALYSIS

There are several resources available for more detailed information on ChIP-Seq data analysis. The Wold lab's *ChIP-Seq PeakFinder* is freely available from http://woldlab.caltech.edu/html/chipseq_peak_finder/ (described in Johnson, 2007). This and other available algorithms and papers describing applications of the ChIP-Seq method and downstream analysis published before October 2007 are listed below. Please visit www.illumina.com for the latest list of ChIP-Seq resources, including application protocols, current software solutions, and scientific publications.

- Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502.
- Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature methods 4: 651-657.
- Barski A, Cuddapah S, Cui K, Roh TY, Schones DE et al. (2007) High resolution profiling of histone methylations in the human genome. Cell 129: 823-837.
- Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

ADDITIONAL INFORMATION

Visit our website or contact us at the address below to learn more about Illumina Sequencing Applications and Software Solutions.

Illumina, Inc.
Customer Solutions
9885 Towne Centre Drive
San Diego, CA 92121-1975
1.800.809.4566 (toll free)
1.858.202.4566 (outside the U.S.)
techsupport@illumina.com
www.illumina.com

FOR RESEARCH USE ONLY

Pub. No. 770-2007-009 Current as of 26 November 2007

illumina®