UB Next-Generation Sequencing and Expression Analysis Core Illumina ChIP-Seq Pipeline Document

Document Overview

An illumina sequencing run generates raw data in the form of a series of images and base pairs. The UB Next-Generation Sequencing and Expression Analysis Core Illumina ChIP-Seq Data Analysis Pipeline begins by transferring the data to our off-rig server maintained by the UB Computational Computer Research Center (CCR). Sequence analysis is performed, and the processed data is made available to the researcher. This document describes the on-instrument and off-instrument pipelines, as well as the data files generated and provided to the researcher. Finally, the document will provide some basis suggestions on how a researcher can further process the data.

TABLE OF CONTENTS

1) Overview of Illumina Pipeline

The illumina sequencing run generates raw data in the form of a series of images, which are analyzed in three different steps: image analysis, base calling, and sequence analysis. The UB Next-Generation Sequencing and Expression Analysis Core uses the illumina Sequencing Control Software (SCS), which runs on the instrument, to perform real-time image analysis and base calling. This analysis generates a file containing the sequencing reads and quality scores called s_N_sequence.txt. This data is then automatically transferred to our off-rig server, where sequence analysis occurs manually. The core uses the illumina Generation of Recursive Analyses Linked by Dependency (GERALD) module for sequence alignment and metrics visualization. The GERALD module employs the Efficient Large-Scale Alignment of Nucleotide Databases (ELAND) program as an alignment tool. ELAND works by aligning up to two errors from a reference for the first 32 bases, or more bases using the ELAND extended format. The program generates the data file s_N_sorted.txt. This file is used for downstream analysis in the core's custom ChIP-Seq Pipeline.

2) Custom ChIP-Seq Pipeline

The UB Next-Generation Sequencing and Expression Analysis Core custom ChIp-Seq Pipeline provides researchers with assay specific data analysis. The pipeline uses several common third-party software packages to provide the researcher with initial ChIP-Seq data analysis. The four major components of the pipeline are BOWTIE, MACS, SPP, and FSEQ software packages.

a) Bowtie – The Bowtie software package is a secondary alignment tool that uses the s_N_sequence.txt file generated by the illumina pipeline. The UB Next-Gen Sequencing and Expression Analysis Core uses the appropriate reference for each project as requested by the researcher and the default program parameters. Additional information about the Bowtie alignment tool can be found at http://bowtie-bio.sourceforge.net/index.shtml.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology **10**: R25.

b) MACS – Model-Based Analysis for ChIP-Seq (MACS) is a software program that empirically models the length of ChIP-Seq fragments and uses that information to predict binding sites. The UB Next-Gen Sequencing and Expression Analysis Core uses the output file from the Bowtie alignment software and the default parameters for MACS analysis. Additional information about MACS can be found at http://liulab.dfci.harvard.edu/MACS/.

Zhang, Y., and T. Liu. 2008. Model-based Analysis of ChIP-Seq (MACS). Genome Biology **9**: R137.

c) SPP – The SPP software has several different functions, but is largely used for assessing overall DNA-binding signals and selecting appropriate tag alignments based on quality. The software can discard or restrict positions with an unusually high number of tags, calculate genome-wide profiles of smoothed tag density and fold enrichment ratios, determine statistically significant tip binding positions, and assess the depth of point binding positions to determine if saturation criteria has been met. A major advantage of the SPP software is that the data output can be generated as ".wig" files, which can be viewed directly by the UCSC Genome Browser and similar data browsers. Additional information about the SPP software package can be found at http://compbio.med.harvard.edu/Supplements/ChIP-seq/.

Kharchenko, P. K., M. Y. Tolstorukov, and P. J. Park. 2008. Design and analysis of ChIP-Seq experiments for DNA-binding proteins. Nature Biotechnology **26**: 1351-1359.

d) FSEQ – FSEQ is used to identify specific sequence features such as transcription factor binding sites for ChIP-Seq projects. The FSEQ software package creates a continuous tag sequence density estimation, which allows for the identification of biologically significant sites. The data output files from FSEQ analysis can conveniently be viewed directly in the UCSC Genome Browser and other similar browsers. Additional information about the FSEQ software package can be found at http://www.genome.duke.edu/labs/furey/software/fseq/.

Boyle, A. P., J. Guinney, G. E. Crawford, and T. S. Furey. 2008. F-Seq: A feature density estimator for high-throughput sequence tags. Bioinformatics **24**: 2537-2538.

3) Data Directory Structure

The UB Next-Generation Sequencing and Expression Analysis Core will provide the researcher with data in a zipped directory via an ftp site. The zip directory (main directory) will follow the naming convention: Organism_SampleID. For example, if your sample name is Tup1 and is derived from yeast, the main directory name would be Yeast_Tup1. Inside the main directory there are five additional directories: Bowtie_Alignment, FSEQ_Output, MACS_Output, SPP_Output, and Illumina_Output. Each of the five additional directories will contain analysis files as seen in the directory structure diagram below. Descriptions of all of the analysis files can be found in section 4 of this document.

Main Directory - Organism_SampleID

- **Illumina_Output**
  - s_N_sequence.txt
  - s_N_sorted.txt
- **Bowtie_Output**
  - s_N_bowtie.txt
- **MACS_Output**
  - MACS.peaks.xls
  - MACS.peaks.bed
  - MACS.model.pdf
  - MACS.model.r
- **SPP_Output**
  - SPP.binding.positions.txt
  - SPP.crosscorrelation.pdf
  - SPP.density.wig
- **FSEQ_Output**
  - FSEQ.chrN.npf
  - FSEQ.chrN.wig
  - FSEQ.chrN.bed

4) File Descriptions

This section of the document describes all of the illumina and custom ChIP pipeline data files that are provided to a researcher after the sequencing run and data analysis are completed.

a) Illumina_Output –

The researcher will receive two files from the illumina pipeline. The s_N_sequence.txt file is generated on-instrument by the Sequencing Control Software. The s_N_sorted.txt file is created manually, off-instrument using the ELAND alignment tool.

i) s_N_sequence.txt –

contains all of the sequence reads and quality scores from one flow cell lane. The file is in FASTQ format and the "N" in the file name will be the flow cell lane number. For example, if your sample was run in lane number seven of the flow cell the file you receive will be called s_7_sequence.txt.

ii) s_N_sorted.txt –

only contains reads that have passed a purity filter and have a unique alignment to the reference genome. The reads are sorted by alignment position in regards to the reference sequence. The "N" in the file name will be the flow cell lane number. For example, if your sample was run in lane number three of the flow cell the file you receive will be called s_3_sequence.txt.

b) Bowtie_Output –

The researcher will receive the s_N_bowtie.txt file from the Bowtie alignment tool. The file is compatible with all of the third-party software programs in the core's ChIP-Seq pipeline, as well as a variety of other analysis programs and tools.

i) Bowtie_alignment.txt –

The file reports the Bowtie alignment analysis as one row/line for each aligned read. There are eight data containing columns/fields that are separated by tabs and each aligned read has the following information:

1) Read identification.

2) Orientation of reference strand aligned to, + for forward strand, - for reverse.

3) Name of reference sequence where alignment occurs or numeric ID if no name was provided.

4) 0-based offset into the forward reference strand where leftmost character of the alignment occurs.

5) Read sequence (reverse-complemented if orientation is -).

6) ASCII-encoded read qualities (reversed if orientation is -). The encoded quality values are on the Phred scale and the encoding is ASCII-offset by 33 (ASCII char !).

7) Number of additional instances where the same sequence aligned against the same reference characters as it was aligned against in the reported alignment. This is *not* the number of other places the read aligns with the same number of mismatches. The number in this column is generally not a good proxy for that number (e.g., the number in this column may be '0' while the number of other alignments with the same number of mismatches might be large).

8) Comma-separated list of mismatch descriptors. If there are no mismatches in the alignment, this field is empty. A single descriptor has the format offset:reference-base>read-base. The offset is expressed as a 0-based offset from the high-quality (5') end of the read.

c) MACS_Output – The researcher will receive four files from the MACS software package.

i) MACS.peaks.xls – tabular file containing information about called peaks that can be opened in Microsoft Excel. Coordinates in this format are 1-based, which is different than the .bed format. The following information is included in the file:

1) Chromosome name.

2) Start position of peak.

3) End position of peak.

4) Length of peak region.

5) Peak summit position related to the start position of peak region.

6) Number of tags in peak region.

7) -10*log10 (pvalue) for the peak region (e.g. pvalue = 1e-10, then the value should be 100).

8) Fold enrichment for this region against random Poisson distribution with local lambda.

9) False discovery rate (FDR) in percentage.

ii) MACS.peaks.bed – file contains information about peak locations and can be viewed in the UCSC Genome Browser. More information about this file format can be found at http://genome.ucsc.edu/FAQ/FAQformat - format1.

iii) MACS.model.pdf – PDF file containing an image model based on the data.

iv) MACS.model.r – contains and R script that can be used to produce a PDF image of the model based on your data. To load the script into R use the following command: $ R —vanilla < NAME_model.r. Running this script in R results in the MACS.model.pdf output file. The UB Next-Gen Sequencing and Expression Analysis Core runs this R script and provides the researcher with the file (see above section 4ciii).

d) SPP_Output – The researcher will receive three files from the SPP data analysis.

i) SPP.binding.positions.txt – file contains information about binding positions based on the window tag density (WTD) method using a false discovery rate of 1%. The WTD method scores positions based on the strand-specific tag counts upstream and downstream of the examined position. The following information is included in the file:

Field	Type	Description
Chr	String	Chromosome description
Pos	Integer	Location on chromosome
Score	Double	Score
Evalue	Double	E-value
FDR	Double	False discovery rate. For peaks higher than the maximum control peak the highest dataset FDR is reported.
Enrichment.lb	Double	Lower bound of the fold-enrichment ratio confidence interval. Estimate determined using scale of 1. Estimates corresponding to higher scales are returned in other enr columns with scale appearing in the name.
Enrichment.mle	Double	enrichment ratio maximum likely estimate

ii) SPP.crosscorrelation.pdf – This file is a genome-wide cross correlation plot.

iii) SPP.density.wig – This file contains smoothed tag density in wiggle format. This file can be read with most common genome browsers such as the UCSC Genome Browser and the Integrated Genome Browser (IGB). More information about wiggle format can be found at http://genome.ucsc.edu/goldenPath/help/wiggle.html.

e) FSEQ_Output – The researcher will receive three files from the FSEQ analysis.

i) FSEQ.chrN.npf – The Narrow Peaks Format (.npf) is used to provide called peaks of signal enrichment based on pooled, normalized data. The following information is included in the file:

Field	Type	Description
chrom	String	Name of the chromosome
chromStart	Integer	The starting position of the feature in the chromosome. The first base in a chromosome is numbered 0.
chromEnd	Integer	The ending position of the feature in the chromosome or scaffold. The chromEnd base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as chromStart=0, chromEnd=100.
name	String	Name given to a region. '.' used if no name is assigned.
score	Integer	Indicates how dark the peak will be displayed in the browser (1-1000). If '0', the DCC will assign this based on signal value. Ideally average signalValue per base spread between 100-1000.
strand	Character	+/- denotes strand or orientation. '.' used if no orientation is assigned
signalValue	Float	Measurement of overall enrichment for the region.
pValue	Float	Measurement of statistical significance (-log10). -1 used if no pValue is assigned.
qValue	Float	Measurement of statistical significance using false discovery rate. -1 used if no qValue is assigned.
peak	Integer	Point-source called for this peak; 0-based offset from chromStart. -1 used if no point-source called.

ii) FSEQ.chrN.wig – This file contains data such as GC percent, probability scores, and transcriptome data. The file is in wiggle format, which can be viewed with most common genome browsers such as the UCSC Genome Browser and the Integrated Genome Browser (IGB). More information about wiggle format can be found at http://genome.ucsc.edu/goldenPath/help/wiggle.html.

ii) FSEQ.chrN.bed – contains data in ".bed" format, which can be viewed in the UCSC Genome Browser. More information about this file format can be found at http://genome.ucsc.edu/FAQ/FAQformat - format1.

5) Additional Analysis

The UB Next-Generation Sequencing and Expression Analysis Core recommends that researchers use the UCSC Genome Browser to view files from our illumina ChIP-Seq pipeline. The UCSC Genome Browser is able to stack annotation tracks beneath genome coordinate position, which allows for quick and easy visual correlation of different types of information. The browser allows the user to look at a whole chromosome to get an idea of gene density; open specific cytogenetic bands to see positional mapped disease gene candidates; and zoom in to particular genes to view spliced ESTs and alternative splicing regions. This browser is a great tool for ChIP-Seq researchers because it allows all types of relevant data to be viewed and interpreted in one location. More information about the UCSC Genome Browser can be found at http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html.

Contact Information

UB Next-Generation Sequencing and Expression Analysis Core Facility (B3-123)
State University of New York at Buffalo
New York State Center of Excellence in Bioinformatics and Life Sciences
701 Ellicott Street
Buffalo, NY 14203
phone: (716) 881-7514
email: cbi-ubnextgencore@buffalo.edu