illumina®

Data Analysis: ChIP-Seq

An overview of data processing using the BaseSpace® Correlation Engine for ChIP-Seq analysis.

Introduction

The BaseSpace Correlation Engine contains chromatin-immunoprecipitation sequencing (ChIP-Seq) studies obtained by systematically mining publicly available next-generation sequencing data through a methodical screening, curation, and data analysis process. This technical note provides an overview of the ChIP-Seq data processing pipeline.

Raw Data Processing

Raw data are obtained from the following sources:

1. Gene Expression Omnibus (GEO)[1]
2. ENCODE - Encyclopedia of DNA Elements (ENCODE Project Consortium, 2004)[2]
3. Individual scientific articles
4. Other data sources, as they become available

GEO data are downloaded as RAW files when available. Otherwise, sequence reads are derived from FASTQ files. These reads are mapped to a current reference genome using ELAND (Cox A, unpublished data). Finally, coordinate files are converted into a QuEST-friendly format. Coordinates lacking strand information are assigned a random strand. ENCODE data are downloaded as FASTQ files. For individual articles, coordinates mapping methods are assessed and, if acceptable, coordinates are used as provided. Otherwise, sequence mapping is performed, providing that sequences are available. This workflow is summarized in Figure 1.

Curation

As part of the curation process, the experiment type is annotated as one of 3 categories:

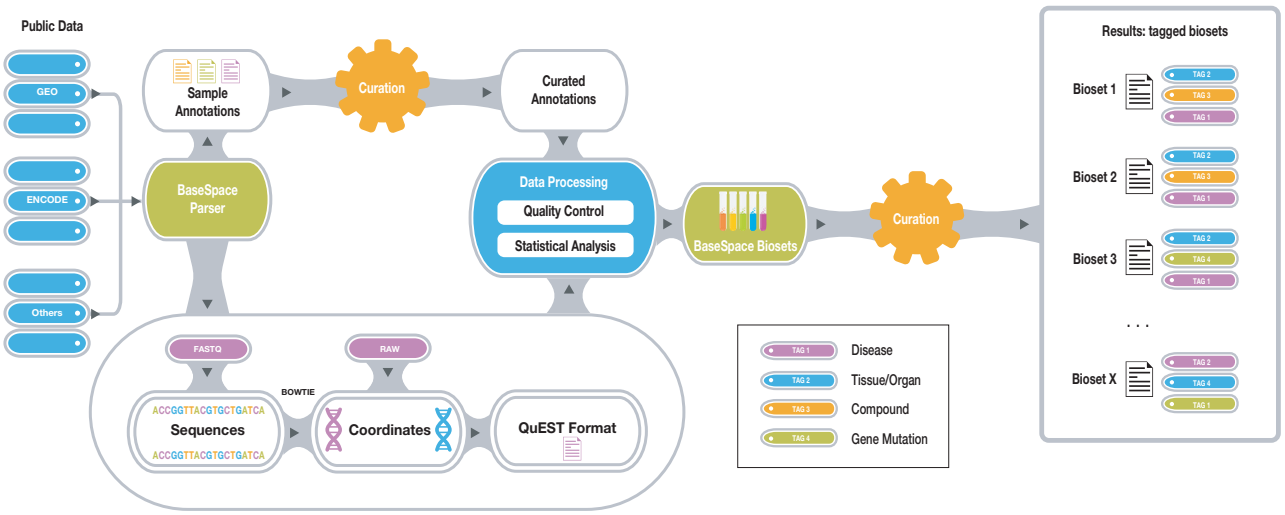
1. Transcription factor binding site
2. Pol II
3. Histone modification

The experiment type controls 2 parameters: bandwidth and region size (both measured in bp). If settings from none of the 3 experiment types fit a specific study, these parameters can be tuned specifically to match the typically observed value. Replicates are united, and samples and controls are identified, titled, and described. When the experimental control is not provided, a species-specific universal control signal is used.

Enriched-region calling parameters:

- Sample data format
- Sample files
- Control files
- Reference genome (as RefSeq or genome table)
- Experiment type category
- Threshold stringency (tuned based on the results of test studies)
- False discovery rate (FDR) analysis

Figure 1: ChIP-Seq Raw Data Processing Pipeline



Sample annotation comprises:

- Sample identifier
- Bioset name
- Species
- Bioset description
- Platform identifier
- Platform name

Additional information is curated based on parsed GEO family.soft files.

Enriched-Region Calling

Enriched regions are genomic positions that are enriched in ChIP data compared to control data. QuEST, a statistical package for the analysis of ChIP-Seq data,[3] is used to identify enriched regions.

The QuEST pipeline is composed of 5 main modules:

Peak Shift Estimation

A sliding window identifies candidate regions having both a high tag count and enrichment compared to the control. The top 200 candidate regions are used to calculate a distance, called peak shift, for shifting the forward and reverse tag density profiles toward one another to maximize correlation between these profiles. The 200 local peak shifts are used to estimate a global peak shift, which is later applied to all regions.

CDP Calculation

The probability density function is estimated for each strand based on tag distribution by kernel density. These probabilities are adjusted by the estimated peak shift and summed to give the Combined Density Profile (CDP).

Peak Calling

The peak calling stage identifies enriched regions and peaks within them. Region seeds are calculated based on CDP and ChIP-to-background fold enrichment. Region seeds are extended in both directions until the CDP value is lower than the ChIP extension threshold. Peaks are identified within enriched regions as local CDP maxima. When more than a single peak is found, a “dip” is expected between 2 adjacent peaks within the same region.

Metrics

For each identified region and peak, a local peak shift is estimated. Q values, fold enrichment, and tag fold enrichment are calculated. Regions and peaks are ranked by Q value.

Filter

Long local peak shifts (> 28 bp) are filtered to remove repeat mismappings. Regions without any peaks are removed.

If less than 100 enriched regions are identified for a single sample, and if enough control reads are provided, an FDR analysis is performed for the choice of QuEST parameters (Figure 2).

Figure 2: QuEST Analysis

Tag tracks calculated with QuEST for control (upper track) or ChIP (lower track) displayed by UCSC browser. Data are shown for GSE13322 study, GSM336333 sample, chr14:331,101–332,667.

Bioset Generation

Enriched regions identified by QuEST create a bioset. Regions are ranked based on fold ChIP tag count.

For each region, the following data are collected from QuEST output files:

- Chromosome
- Start
- End
- Max position
- Peaks
- ChIP signal
- Control signal
- ChIP tag count
- Control tag count
- Tag enrichment fold
- Q value
- P-value

Post-Processing Analysis

At the postprocessing stage, studies with older genome versions are converted to the current genomic build. Chromosome identifiers are modified to naming conventions for BaseSpace Correlation Engine.

References

1. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002;30(1):207–210.
2. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science.* 2004;306(5696):636–640.
3. Valouev A, Johnson DS, Sundquist A, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Meth.* 2008;5(9):829–834.

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY. NOT FOR USE IN DIAGNOSTIC PROCEDURES.

Pub. No. 970-2014-017 Current as of 14 April 2016

illumina®