Description

This track displays binding sites of the specified transcription factors in the given cell types as identified by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq — see Johnson *et al*., 2007 and Fields, 2007).

ChIP-seq was used to assay chromatin fragments bound by specific or general transcription factors as described below. DNA enriched by chromatin immunoprecipitation was sequenced and short sequence reads of 25-36 nt were mapped to the human reference genome. Enriched regions (peaks) of high sequence read density relative to input chromatin control sequence reads were identified with a peak calling algorithm.

The sequence reads with quality scores (fastq files) and alignment coordinates (BAM files) from these experiments are available for download.

Display Conventions and Configuration

This track is a multi-view composite track that contains multiple data types (*views*). For each view, there are multiple subtracks that display individually on the browser. Instructions for configuring multi-view tracks are here. The subtracks in this track are grouped by transcription factor targeted antibody and by cell type. For each experiment (cell type vs. antibody), the following views are included:

Peaks
: Sites with the greatest evidence of transcription factor binding, calculated using the MACS peak caller (Zhang *et al*., 2008), as enriched regions of high read density in the ChIP experiment relative to total input chromatin control reads.

Raw Signal
: A continuous signal which indicates density of aligned reads. The sequence reads were extended to the size-selected length (225 bp), and the read density computed as reads per million.

Metadata for a particular subtrack can be found by clicking the down arrow in the list of subtracks.

Methods

Cells were grown according to the approved ENCODE cell culture protocols. Cross-linked chromatin was immunoprecipitated with an antibody, the protein-DNA crosslinks were reversed and the DNA fragments were recovered and sequenced. Please see protocol notes below and check here for the most current version of the protocol. Biological replicates from each experiment were completed.

Libraries were sequenced with an Illumina Genome Analyzer I or IIx according to the manufacturer's recommendations. Sequence data produced by the Illumina data pipeline software were quality-filtered and then mapped to NCBI GRCh37 (hg19) using the integrated Eland

software; 32 nt of the sequence reads were used for alignment. Up to two mismatches were tolerated; reads that mapped to multiple sites in the genome were discarded.

To identify likely transcription factor occupancy sites, peak calling was applied to the aligned sequence data sets using MACS (Zhang *et al*., 2008). The MACS method models the shift size of ChIP-seq tags empirically, and uses the shift to improve the spatial resolution of predicted binding sites. The MACS method also uses a dynamic Poisson distribution to capture local biases in the genome, allowing for more robust predictions (Zhang *et al.*, 2008).

Protocol Notes

Several changes and improvements were made to the original ChIP-seq protocol (Jonshon *et al*., 2008). The major differences between protocols are the number of cells and magnetic beads used for IP, the method of sonication used to fragment DNA, the method used for fragment size selection, and the number of cycles of PCR used to amplify the sequencing library. The protocol field for each file denotes the version of the protocol used as being PCR1x, PCR2x or a version number (e.g., v041610.1).

The sequencing libraries labeled as PCR2x were made with two rounds of amplification (25 and 15 cycles) and those labeled as PCR1x were made with one 15-cycle round of amplification. Experiments that were completed prior to January 2010 were originally aligned to NCBI36 (hg18). They have been re-aligned to NCBI GRCh37 (hg19) with the Bowtie software (Langmead *et al*., 2009) for this data release. The libraries labeled with a protocol version number were competed after January 2010 and were only aligned to NCBI GRCh37 (hg19).

Please refer to the Myers Lab website for details on each protocol version and the most current protocol in use.

Verification

The MACS peak caller was used to call significant peaks on the individual replicates of a ChIP-seq experiment. Next, the irreproducible discovery rate (IDR) method developed by Li *et al*. (2011), was used to quantify the consistency between pairs of ranked peaks lists from replicates. The IDR methods uses a model that assumes that the ranked lists of peaks in a pair of replicates consist of two groups: a reproducible group and an irreproducible group. In general, the signals in the reproducible group are more consistent (i.e. with a larger rank correlation coefficient) and are ranked higher than the irreproducible group. The proportion of peaks that belong to the irreproducible component and the correlation of the reproducible component are estimated adaptively from the data. The model also provides an IDR score for each peak, which reflects the posterior probability of the peak belonging to the irreproducible group. The aligned reads were pooled from all replicates and the MACS peak caller was used to call significant peaks on the pooled data. Only datasets containing at least 100 peaks passing the IDR threshold were considered valid and submitted for release.

As part of the validation of ChIP-seq antibodies and to study the downstream targets of several transcription factors, inducible short hairpin RNA (shRNA) cell lines were generated to knock

down the expression of these factors. K562 cells (non-adherent, human erythromyeloblastoid leukemia cell line; ENCODE Tier 1) were transduced with lentiviral vectors carrying an inducible shRNA to a specific transcription factor as described in this protocol. Expression of shRNA was induced with doxycycline in the growth media. Only cell lines that exhibited at least 70% reduction in expression of the targeted transcription factor (determined by qPCR) were used. The cell lines were designated K562-shX, where X is the transcription factor targeted by shRNA and K562 denotes the parent cell line. For example, K562-shATF3 cells are K562 cells selected for stable integration of shRNA targeting the ATF3 gene. Gene expression in doxycycline-induced and uninduced cells were measured and profiled using RNA-seq. The RNA-seq data were submitted to GEO (Accession:GSE33816).

Release Notes

- This is Release 3 (Sept 2012). It contains 110 new experiments including 3 new cell lines and 1 new antibodies.
- The entire HepG2/HEY1 (Accession: wgEncodeEH001502) and K562/HEY1 (Accession: wgEncodeEH001481) datasets have been revoked due to problems with the quality of the antibody.
- All experiments with the U87 cell line were remapped. Previously, the sex of the cell was unknown and was mapped to the male genome. It was discovered that the cell line is female.
- Other files from the previous releases also contained errors. They have been corrected with a version number appended to the name (e.g., V2).
- shRNA validation data have been included in previous releases. The Verification section above provides a more in-depth explanation of the method.

Credits

These data were provided by the Myers Lab at the HudsonAlpha Institute for Biotechnology.

Contact: Flo Pauli

References

Fields S. Molecular biology. Site-seeing by sequencing. *Science*. 2007 Jun 8;316(5830):1441-2.

Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions.*Science*. 2007 Jun 8;316(5830):1497-502.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.

Li Q, Brown JB, Huang H, Bickel PJ. Measuring Reproducibility of High-throughput experiments. *Ann*. *Appl*. *Stat*.Volume 5, Number 3 (2011), 1752-1779.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137.

Data Release Policy

Data users may freely use ENCODE data, but may not, without prior consent, submit publications that use an unpublished ENCODE dataset until nine months following the release of the dataset. This date is listed in the*Restricted Until* column, above. The full data release policy for ENCODE is available here.