Gene regulation: ENCODE project and the Human Epigenome Atlas

Joshua W. K. Ho, PhD

Head, Bioinformatics and Systems Medicine Laboratory
Victor Chang Cardiac Research Institute

NHMRC/National Heart Foundation Career Development Fellow
Senior Lecturer (Conjoint), St Vincent’s Clinical School, UNSW

http://bioinformatics.victorchang.edu.au

j.ho@victorchang.edu.au

UNSW, 29th Jan 2016
Human genome has 23 pairs of chromosomes

Human somatic cells (i.e., not gametes) have a diploid genome – one haploid genome from each parent

Each chromosome is a double helix

A haploid genome contains ~20,000 protein-coding genes

22 pairs of autosomes
1 pair of sex chromosomes

Complete* human genome sequenced by 2003
Beyond DNA sequence: chromatin organisation is important for epigenetic regulation

Nucleosome core = 146bp of DNA + histone octamer

Linker DNA = ~10-80 bp in length

Density nucleosome is an important determinant of chromatin accessibility
Transcription factor (TF), RNA polymerase, histone modification, chromatin accessibility, and DNA methylation

Wang et al. (2016) *Computational Biology & Bioinformatics: Gene Regulation*
Histone modification is associated with genomic regulatory elements

Naming of histone modification:
H3K4me3 = Trimethylation of histone H3 at lysine 4

Ho (2012) Biophysical Reviews
The ENCODE Project

The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the National Human genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which gene is active.

https://www.encodeproject.org
ENCODE – past and present

• Launched in September 2003 to identify all functional elements in the human genome sequence – as a follow-up to the Human Genome Project.

• 2003-2007: The pilot phase and technology development phase. Test and compare existing method to rigorously analyse a defined portion (1%) of the human genome sequence.

• 2007-now: genome-wide data production phase

• Major finding: showing that ~80% of the genome participate in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type (ENCODE 2012, Nature)
Who are the ENCODE team?

https://www.encodeproject.org/about/contributors/
List of current ENCODE research groups: http://www.genome.gov/26525220
ENCODE 2015: Research Applications and Users Meeting

ENCODE 2015: Research Applications and Users Meeting
June 29 - July 1, 2015

Bolger Center
Potomac, Md.

On June 29 - July 1, 2015, the National Human Genome Research Institute (NHGRI) sponsored the ENCODE 2015: Research Applications and Users Meeting at the Bolger Center in Potomac, Md.

The meeting featured:
• Hands-on workshops on learning to navigate, analyze, and integrate ENCODE and ENCODE data into your research
• Leading-edge research applications from distinguished invited speakers
• Tutorials on newly-available informatics pipelines that greatly facilitate working with ENCODE data
• Short talks selected from abstracts

View videos and slides from:

- June 29
- June 30
- July 1

Monday, June 29, 2015

<table>
<thead>
<tr>
<th></th>
<th>Topic</th>
<th>Moderator/Presenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Using ENCODE Data to Interpret Disease-associated Genetic Variation</td>
<td>Mike Pazin, NHGRI</td>
</tr>
<tr>
<td></td>
<td><a href="https://www.youtube.com/watch?v=example">YouTube Video</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="example">Slides</a></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Data Integration: Genome x Transcriptome x EMR</td>
<td>Nancy Cox, Vanderbilt University</td>
</tr>
<tr>
<td></td>
<td><a href="https://www.youtube.com/watch?v=example">YouTube Video</a></td>
<td></td>
</tr>
</tbody>
</table>

Scientific Session 1: Common Disease

<table>
<thead>
<tr>
<th></th>
<th>Topic</th>
<th>Presenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>An Epigenomic and Transcriptional Basis for Insulin Resistance</td>
<td>Evan Rosen, Harvard Medical School</td>
</tr>
<tr>
<td></td>
<td><a href="https://www.youtube.com/watch?v=example">YouTube Video</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="example">Slides</a></td>
<td></td>
</tr>
</tbody>
</table>
Roadmap Epigenomics

Defining the human epigenome

http://www.roadmapepigenomics.org/
GTEx – Genotype-tissue expression

Discovering association between genotype (e.g., SNP) and gene expression in humans

http://www.gtexportal.org/
FANTOM – another international effort

Discovery of functional elements in *H. sapiens* (human) and *M. musculus* (mouse)
Related project: mouse ENCODE

ENCODE for *M. musculus* (mouse)

http://www.mouseencode.org/
Related project - modENCODE

ENCODE for *D. melanogaster* (fruit fly) and *C. elegans* (roundworm)

http://www.modencode.org/
What can we learn from ENCODE?

- Genome-wide data
  - RNA-seq, ChIP-seq, DNase-seq,…
- Genome annotation
  - Gene annotation, chromatin state,…
- Experimental protocols and best practices
- Bioinformatics analysis methods and software

Resource

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Stephen G. Landt, 1,26 Georgi K. Marinov, 2,26 Anshul Kundaje, 3,26 Pouya Kheradpour, 4
Florecia Pauli, 5 Serafim Batzoglou, 3 Bradley E. Bernstein, 6 Peter Bickel, 7 James B. Brown, 7

Landt et al. (2012) Genome Research

https://www.encodeproject.org/
ENCODE human cell lines

- ENCODE data were mostly generated from human cell lines. Tier 1 cell lines are more widely assayed than tier 2 cell lines, and in turn are more widely assayed than tier 3 cell lines.

- Tier 1: **GM12878** (lymphoblastoid cell line, from a female HapMap individual); **H1-hESC** (embryonic stem cells); **K562** (leukemia cell line)

- Tier 2: 15 cell lines, including **IMR90** (fetal lung fibroblasts), **HUVEC** (umbilical endothelial cells), **HeLa-S3** (cervical carcinoma), **MCF-7** (mammary gland), etc.

- Tier 3: 338 cell lines

https://genome.ucsc.edu/ENCODE/cellTypes.html
Experimental assays

- DNA methylation
  - Methyl Array, RRBS

- Chromatin accessibility
  - DNase-seq, FAIRE-seq

- RNA binding proteins
  - RIP tiling array, RIP-seq

- Chromatin modification
  - ChIP-seq

- TF binding
  - ChIP-seq

- Higher order chromatin structure
  - 5C, ChIA-PET

- Replication
  - Repli-chip, Repli-seq

https://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataMatrixHuman.html
RNA-seq

Align reads to genome

Assemble transcripts from spliced alignments

More abundant

Less abundant

Assemble transcripts de novo

Align transcripts to genome
RNA-seq reads in a browser

Applications:
1) Discovery of novel transcript or alternative promoter
2) Look at alternative splicing
3) Profiling of gene expression
ChIP-seq maps genome-wide DNA-protein interactions

- ChIP = Chromatin-immunoprecipitation
- Enrich for sequence fragments that are bound by a specific protein
  - Transcription factors
  - Chromatin-associated proteins
  - Histone proteins

ChIP-seq

Histone with a modified tail

DNA

nucleosome

ChIP-seq profiles

Transcription factor

Histone modification

Bulk nucleosome

Ho et al. (2011) Tag-based Approaches for Next Generation Sequencing
Identifying chromatin landscape in human cell lines

Chromatin state = a combination of histone modifications

Ernst et al. (2011) Nature
Chromatin state can differ between organisms

Ho et al. (2014) Nature
Accessing ENCODE data from UCSC Genome Browser

Search for gene, or genomic interval

Species

Ref. genome assembly

Human (Homo sapiens) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz. Software Copyright © The Regents of the University of California. All rights reserved.

human genome for gene or genomic interval

Human Genome Browser – hg19 assembly (sequences)

The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User’s Guide for more information.
Many genomic ‘tracks’ in UCSC GB

Select track

Click “Refresh”
ENCODE data and annotation on UCSC GB

ENCODE track’s naming format: cell type and experiment: e.g., K562 H3K4me3
You can upload custom tracks into UCSC Genome Browser.

Custom track data
(BED format in this example)
Using encodeproject.org

ENCODE: Encyclopedia of DNA Elements

The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Quick Start
To find and download ENCODE Consortium data:
- Click the Data toolbar above and browse data
- By assay

News
Oct 6, 2015: We are pleased to announce that the metadata from 3127 Roadmap experiments have been released on the ENCODE Portal.

Aug 24, 2015: The presentation video and tutorial for the...
Showing 25 of 156 results

**cardiac muscle cell (Homo sapiens)**
- Type: primary cell
- Source: ScienCell

**cardiac muscle cell (Homo sapiens)**
- Type: primary cell
- Source: ScienCell

**cardiac myocyte (Homo sapiens, fetal)**
- Type: in vitro differentiated cells
- Source: John Stamatoyannopoulos

**cardiac myocyte (Homo sapiens, fetal)**
- Type: in vitro differentiated cells
- Source: John Stamatoyannopoulos

**regular cardiac myocyte (Homo sapiens, adult 51 year)**
- Type: primary cell
- Source: PromoCell

**regular cardiac myocyte (Homo sapiens, adult 48 year)**
- Type: primary cell
- Source: PromoCell
<table>
<thead>
<tr>
<th>Accession</th>
<th>Assay</th>
<th>Biosample term name</th>
<th>Target</th>
<th>Description</th>
<th>Lab</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENCSR000ABO</td>
<td>DNA methylation profiling by array assay</td>
<td>cardiac muscle cell</td>
<td>cardiac muscle cell</td>
<td>UW HCM</td>
<td>Richard Myers, HAIB</td>
</tr>
<tr>
<td>ENCSR000DTI</td>
<td>ChIP-seq</td>
<td>cardiac muscle cell</td>
<td>CTCF</td>
<td>CTCF ChIP-seq on human HCM</td>
<td>John Stamatoyannopoulos, UW</td>
</tr>
<tr>
<td>ENCSR000DTJ</td>
<td>ChIP-seq</td>
<td>cardiac muscle cell</td>
<td>Control</td>
<td>Control ChIP-seq on human HCM</td>
<td>John Stamatoyannopoulos, UW</td>
</tr>
<tr>
<td>ENCSR000DTK</td>
<td>ChIP-seq</td>
<td>cardiac muscle cell</td>
<td>H3K4me3</td>
<td>H3K4me3 ChIP-seq on human HCM</td>
<td>John Stamatoyannopoulos, UW</td>
</tr>
<tr>
<td>ENCSR000ENJ</td>
<td>DNase-seq</td>
<td>cardiac muscle cell</td>
<td></td>
<td>DNase-seq on human cardiac muscle</td>
<td>John Stamatoyannopoulos, UW</td>
</tr>
</tbody>
</table>

Displaying 5 of 8
<table>
<thead>
<tr>
<th>Feature</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>1</td>
</tr>
<tr>
<td>histone</td>
<td>1</td>
</tr>
<tr>
<td>histone modification</td>
<td>1</td>
</tr>
<tr>
<td>narrow histone mark</td>
<td>1</td>
</tr>
<tr>
<td>transcription factor</td>
<td>1</td>
</tr>
<tr>
<td>Biosample type</td>
<td></td>
</tr>
<tr>
<td>primary cell</td>
<td>5</td>
</tr>
<tr>
<td>Life stage</td>
<td></td>
</tr>
<tr>
<td>unknown</td>
<td>5</td>
</tr>
<tr>
<td>Available data</td>
<td></td>
</tr>
<tr>
<td>fastq</td>
<td></td>
</tr>
<tr>
<td>bam</td>
<td>4</td>
</tr>
<tr>
<td>bed broadPeak</td>
<td>4</td>
</tr>
<tr>
<td>bigBed broadPeak</td>
<td>4</td>
</tr>
<tr>
<td>bigWig</td>
<td>4</td>
</tr>
<tr>
<td>Run type</td>
<td></td>
</tr>
<tr>
<td>single-ended</td>
<td>4</td>
</tr>
<tr>
<td>Read length (nt)</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>4</td>
</tr>
<tr>
<td>Library made from</td>
<td></td>
</tr>
<tr>
<td>DNA</td>
<td>4</td>
</tr>
<tr>
<td>RNA</td>
<td>1</td>
</tr>
<tr>
<td>Date released</td>
<td></td>
</tr>
<tr>
<td>February, 2011</td>
<td>2</td>
</tr>
<tr>
<td>April, 2011</td>
<td>1</td>
</tr>
<tr>
<td>December, 2010</td>
<td>1</td>
</tr>
<tr>
<td>January, 2012</td>
<td>1</td>
</tr>
</tbody>
</table>

**ChiP-seq of cardiac muscle cell (Homo sapiens)**

Target: H3K4me3  
Lab: John Stamatoyannopoulos, UW  
Project: ENCODE
Using batch download

Click the “Download” button below to download a “files.txt” file that contains a list of URLs to a file containing all the experimental metadata and links to download the file. The first line of the file will always be the URL to download the metadata file. Further description of the contents of the metadata file are described in the Batch Download help doc.

The “files.txt” file can be copied to any server.
The following command using cURL can be used to download all the files in the list:

```
xargs -n 1 curl -0 -L < files.txt
```
|   | File accession | File format | Output type | Experiment | Assay | Biosample type | Biosample id | Biosample | Biosample | Biosample | Biosample | Experiment | Antigen |
|---|----------------|-------------|-------------|------------|-------|----------------|--------------|------------|------------|------------|------------|------------|----------|---------|
| 1 | ENCF001HK        | bigWig      | raw signal  | ENCSR00001T | ChIP-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | CTCF-human | ENCAE     |
| 2 | ENCF0021HK       | bigWig      | raw signal  | ENCSR00004T | ChIP-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | CTCF-human | ENCAE     |
| 3 | ENCF003HKX       | bed broadPe hotspots | ENCSR00001T | ChIP-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | CTCF-human | ENCAE     |
| 4 | ENCF003XK        | bed broadPe hotspots | ENCSR00001T | ChIP-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | CTCF-human | ENCAE     |
| 5 | ENCF0021VK       | bed broadPe signal | ENCSR00004T | ChIP-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | CTCF-human | ENCAE     |
| 6 | ENCF0021V        | bed broadPe signal | ENCSR00004T | ChIP-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | CTCF-human | ENCAE     |
| 7 | ENCF0021BA       | bigWig      | raw signal  | ENCSR00006N | DNase-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | CTCF-human | ENCAE     |
| 8 | ENCF0021BB       | bigWig      | signal      | ENCSR00006N | DNase-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | CTCF-human | ENCAE     |
| 9 | ENCF0021DA       | bigWig      | raw signal  | ENCSR00006N | DNase-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | CTCF-human | ENCAE     |
| 10| ENCF0021VZ       | bed broadPe signal | ENCSR00006N | DNase-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | CTCF-human | ENCAE     |
| 11| ENCF0031W        | bed broadPe hotspots | ENCSR00006N | DNase-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | CTCF-human | ENCAE     |
| 12| ENCF0021DA       | bigWig      | raw signal  | ENCSR00006N | DNase-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | CTCF-human | ENCAE     |
| 13| ENCF0031W        | bed broadPe hotspots | ENCSR00006N | DNase-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | CTCF-human | ENCAE     |
| 14| ENCF001FH        | bigWig      | raw signal  | ENCSR00007T | ChIP-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | H3K4me3-hu | ENCAE     |
| 15| ENCF0021F        | bigWig      | raw signal  | ENCSR00007T | ChIP-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | H3K4me3-hu | ENCAE     |
| 16| ENCF0021XK       | bed broadPe hotspots | ENCSR00007T | ChIP-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | H3K4me3-hu | ENCAE     |
| 17| ENCF0021XK       | bed broadPe hotspots | ENCSR00007T | ChIP-seq | CL0000746      | cardiac muscle primary cell | unknown    | unknown    | Homo sapiens |            | H3K4me3-hu | ENCAE     |
Visualise data

The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User’s Guide for more information.

Request:  | Genome Browser Response:
----------|------------------------------------------------------
chr7       | Displays all of chromosome 7
chrUn_gl000212 | Displays all of the unplaced contig gl000212
20p13     | Displays region for band p13 on chr 20
chr3:1-1000000 | Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000 | Displays a region of chr3 that spans 2000 bases, starting with position 1000000
Use RegulomeDB to identify all SNPs around NKX2-5, and predict if they are potential regulatory elements (promoter/enhancers)

Try chr5:172,644,666-172,676,755 at http://regulomedb.org/
### Summary of SNP analysis

<table>
<thead>
<tr>
<th>Coordinate (0-based)</th>
<th>dbSNP ID</th>
<th>Regulome DB Score</th>
<th>Other Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr5:172665804</td>
<td>rs148505573</td>
<td>2a</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172665811</td>
<td>rs190123272</td>
<td>2a</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172644773</td>
<td>rs77787950</td>
<td>2b</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172645094</td>
<td>rs4410656</td>
<td>2b</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172645163</td>
<td>rs7448948</td>
<td>2b</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172655699</td>
<td>rs77646165</td>
<td>2b</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172655732</td>
<td>rs3132139</td>
<td>2b</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172656216</td>
<td>rs3132140</td>
<td>2b</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172656809</td>
<td>rs13361978</td>
<td>2b</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172657154</td>
<td>rs3131916</td>
<td>2b</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172658279</td>
<td>rs75673847</td>
<td>2b</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172660514</td>
<td>rs200467566</td>
<td>2b</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172682023</td>
<td>rs2277923</td>
<td>2b</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172682191</td>
<td>rs77083308</td>
<td>2b</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172682576</td>
<td>rs73317175</td>
<td>2b</td>
<td>UCSC</td>
</tr>
<tr>
<td>chr5:172683306</td>
<td>rs73317176</td>
<td>2b</td>
<td>UCSC</td>
</tr>
</tbody>
</table>
Data supporting chr5:172665804 (rs148505573)

Score: 2a
Likely to affect binding

H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE

DNAse1 Hypersensitivity Clusters in 125 cell types from ENCODE (V3)

Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs

100 vertebrates Basewise Conservation by PhyloP

Truncation Alignments of 100 Vertebrates

Simple Nucleotides Polymorphisms (dbSNP 144) Found in >1% of Samples

Protein Binding

<table>
<thead>
<tr>
<th>Method</th>
<th>Location</th>
<th>Bound Protein</th>
<th>Cell Type</th>
<th>Additional Info</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Filter:
**What does the RegulomeDB score represent?**

The scoring scheme refers to the following available datatypes for a single coordinate.

<table>
<thead>
<tr>
<th>Score</th>
<th>Supporting data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak</td>
</tr>
<tr>
<td>1b</td>
<td>eQTL + TF binding + any motif + DNase Footprint + DNase peak</td>
</tr>
<tr>
<td>1c</td>
<td>eQTL + TF binding + matched TF motif + DNase peak</td>
</tr>
<tr>
<td>1d</td>
<td>eQTL + TF binding + any motif + DNase peak</td>
</tr>
<tr>
<td>1e</td>
<td>eQTL + TF binding + matched TF motif</td>
</tr>
<tr>
<td>1f</td>
<td>eQTL + TF binding / DNase peak</td>
</tr>
<tr>
<td>2a</td>
<td>TF binding + matched TF motif + matched DNase Footprint + DNase peak</td>
</tr>
<tr>
<td>2b</td>
<td>TF binding + any motif + DNase Footprint + DNase peak</td>
</tr>
<tr>
<td>2c</td>
<td>TF binding + matched TF motif + DNase peak</td>
</tr>
<tr>
<td>3a</td>
<td>TF binding + any motif + DNase peak</td>
</tr>
<tr>
<td>3b</td>
<td>TF binding + matched TF motif</td>
</tr>
<tr>
<td>4</td>
<td>TF binding + DNase peak</td>
</tr>
<tr>
<td>5</td>
<td>TF binding or DNase peak</td>
</tr>
<tr>
<td>6</td>
<td>other</td>
</tr>
</tbody>
</table>
FactorBook: The ENCODE knowledgebase of transcription factors

http://www.factorbook.org/

Welcome to Factorbook!

The Encyclopedia of DNA Elements (ENCODE) consortium aims to identify all functional genome sequence. These elements include genomic regions bound by transcription factors, occupied by nucleosomes with modified histones, hypersensitive to chromatin immunoprecipitation (ChIP-seq) and ChIP-seq peaks. The transcription factor positions on the genomic DNA bound by TFs and tend to be located around the sites of TF binding. Click the individual pages to explore.

Citation: J Wang, J Zhuang, S Lyer, XY Lin, et al. Sequence features and chromatin regions bound by 119 human transcription factors. Genome Research 22 (9), 1791
Average Profiles of Modified Histones around the Summit of ChIP-seq Peaks

Average histone modification profiles are shown for the [-2 kb, +2 kb] window around the summits of TF ChIP-seq peaks, separately for peaks that are proximal to annotated transcripts (dashed lines) start sites and for peaks that are distal (more than 1kb) to all annotated transcripts (solid lines). Proximal profiles are arranged such that the transcriptional direction of the nearest transcript is toward the right. Histone modification data were generated from the Broad team, using antibodies to pull down modified histones followed by deep sequencing of the genomic DNA associated with the modified histones. C

Mouse over a curve to reveal its identity. Mouse over a histone modification in the legend to show its curves and gray out other histone modifications in all figures. Click a histone modification in the legend to toggle on/off its curve in all figures. Click the "Proximal" or "Distal" button in the legend to show only average histone modification profiles anchored around ChIP-seq peaks that are proximal or distal to annotated transcripts.

Broad - Hepg2
M1
500 / 500
5.0e-738
GCCACCAGGGGC

M2
154 / 500
7.7e-072
CGGCAACCAGGGGCAGCA

M3
31 / 500
1.4e-017
AGGCCCCATCTGAGGAGCAAA

M4
18 / 500
2.1e-002
TGCAGGTACTGCAGGCC

M5
18 / 500
5.1e+000
AAACAAGAAAAGAA