# ENCODE Encyclopedia

**Goal**: Use a genome browser to show functional features in a genomic region of interest.

The ENCODE Consortium has thus far produced hundreds of DNase-seq, TF ChIP-seq, and histone ChIP-seq datasets. How should we combine these data into an encyclopedia that functionally annotates the genome? There are many different ways to build these annotations. Here, we present an annotation that combines ENCODE and Roadmap datasets to directly annotate candidate enhancers and promoters in the genome.

In total, 177 ENCODE and ROADMAP cell types are annotated in this release; among them, 94 cell types have both DNase-seq data and ChIP-seq data for one or more of the following histone marks: H3K4me1, H3K4me3, H3K9ac, and H3K27ac. These histone marks provide useful information when attempting to annotate promoters and enhancers:

- H3K4me1 is enriched at enhancers (both active and poised)
- H3K4me3 is enriched at actively transcribed promoters
- H3K9ac is enriched at promoters and enhancers
- H3K27ac is enriched at active enhancers

For a unified view of chromatin accessibility, the Stam lab merged all DNase-seq experiments in 177 cell types into a "master" track; this track includes all biological and technical replicates from the Stam and Crawford labs. These "master peaks" cover approximately 20% of the genome. Master peaks within 2KB of a TSS (GENCODE TSS V19) are defined as TSS-proximal, while the remaining peaks are TSS-distal.

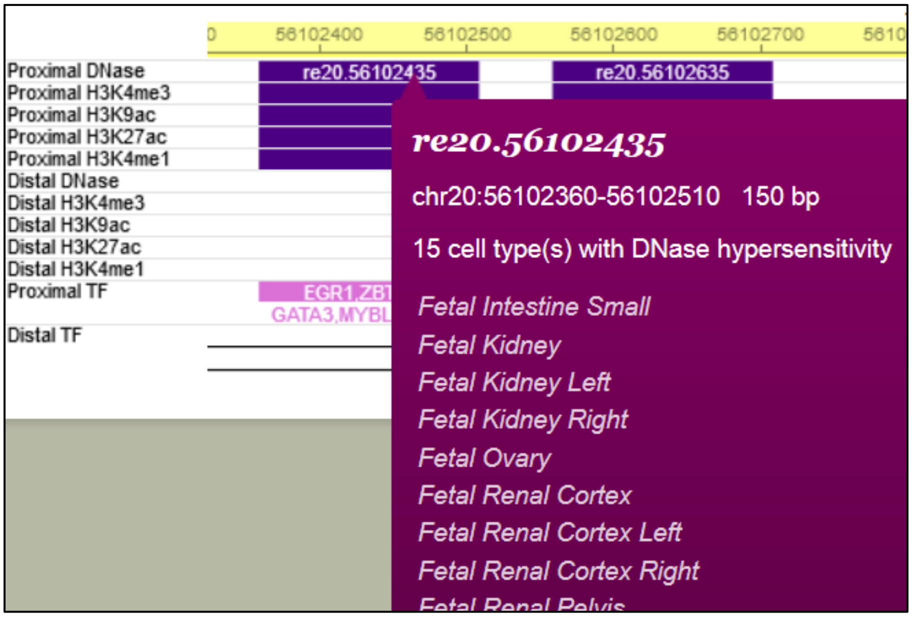
Candidate promoters are located by intersecting DNase-seq, H3K4me3 ChIP-seq, and TF ChIP-seq datasets, while candidate enhancers are located by intersecting DNase-seq, H3K27ac ChIP-seq, and TF ChIP-seq datasets.

Visualization:

- UCSC Genome Browser
- WashU browser

## Functional Annotations:

**Proximal and Distal DNase**: 177 different DNase-seq datasets from the Stam (UW) and Crawford (Duke) labs were merged together to form one unified, "master" DNase dataset. Overlapping peaks are merged into non-overlapping DNase- hypersensitive regions. The Stam lab then identified the "master" peak in each region, defined as the single peak in the region with highest peak height (i.e. the "strongest" peak in the region). The master DNase peaks were separated into TSS -proximal and TSS- distal groups based on whether or not they intersected a 2000-bp window centered on any GENCODE TSS. Each peak includes the cell types present in the merged region.



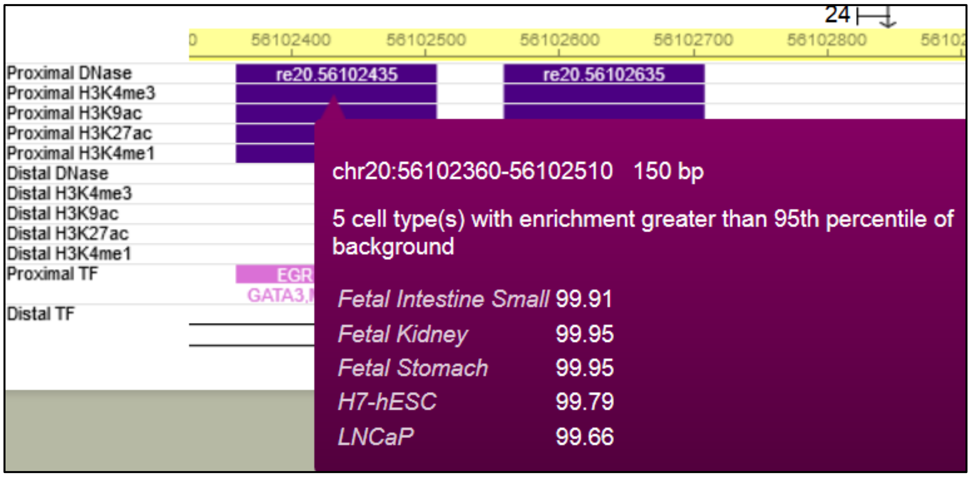
**re20.56102435**

chr20:56102360-56102510 150 bp

15 cell type(s) with DNase hypersensitivity

*Fetal Intestine Small*
*Fetal Kidney*
*Fetal Kidney Left*
*Fetal Kidney Right*
*Fetal Ovary*
*Fetal Renal Cortex*
*Fetal Renal Cortex Left*
*Fetal Renal Cortex Right*
*Fetal Renal Pelvis*

**Proximal and Distal Histone**: Histone data for H3K4me1, H3K4me3, H3K9ac, and H3K27ac were downloaded from both ENCODE and Roadmap projects. For each DNase master peak, the average histone signal in the matching cell type was calculated in a 1000-bp window around the center of the peak. This signal was converted to a percentile using the background distribution of histone signal in the matching cell type in randomly chosen 1000-bp genomic regions (regions outside all DNase peaks and ENCODE blacklisted regions). DNase master peaks that have at least one cell type with a histone signal >95th percentile of background are reported in the track. If there are multiple cell types that fulfil the 95th percentile criteria, they are displayed as separate lines in the track, with the actual percentile over background also displayed.



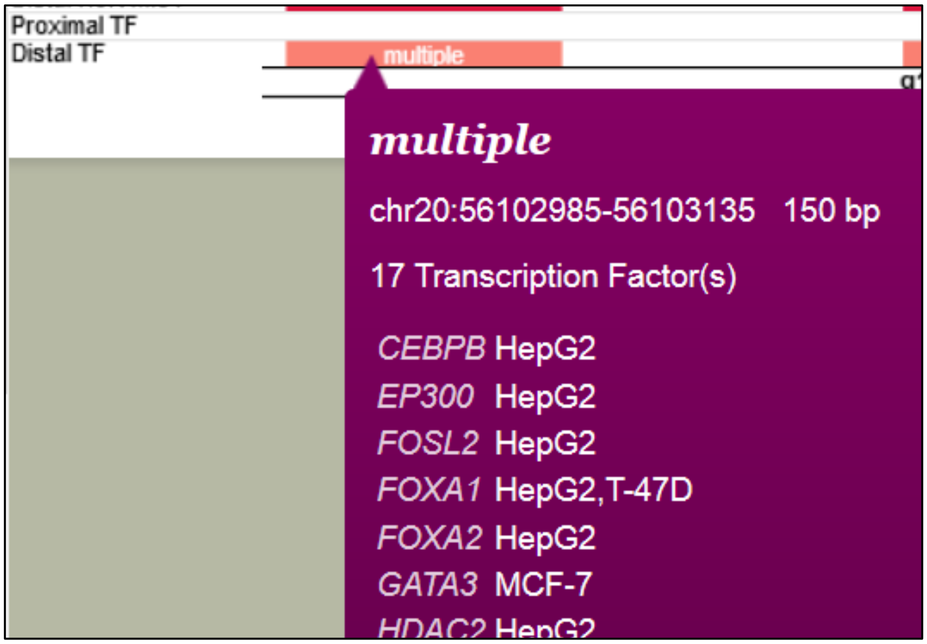
chr20:56102360-56102510 150 bp

5 cell type(s) with enrichment greater than 95th percentile of background

| | |
|---|---|
| *Fetal Intestine Small* | 99.91 |
| *Fetal Kidney* | 99.95 |
| *Fetal Stomach* | 99.95 |
| *H7-hESC* | 99.79 |
| *LNCaP* | 99.66 |

**Proximal and Distal TFs**:

For each of the distal and proximal DNase master peaks, overlapping TF ChIP-seq peaks across all cell types available were identified. The TF peak with the maximum score in each master DNase peak is displayed. Track details include all names (with cell type information) of TFs whose peaks overlapped with the DNase master peak.



**multiple**

chr20:56102985-56103135 150 bp

17 Transcription Factor(s)

| | |
|---|---|
| *CEBPB* | HepG2 |
| *EP300* | HepG2 |
| *FOSL2* | HepG2 |
| *FOXA1* | HepG2,T-47D |
| *FOXA2* | HepG2 |
| *GATA3* | MCF-7 |
| *HDAC2* | HepG2 |

**Acknowledgements**:

Weng Lab

- Zhiping Weng
- Michael Purcaro
- Sowmya Iyer
- Arjan van der Velde

Stam Lab

- John Stamatoyannopoulos
- Bob Thurman
- Richard Sandstorm

ENCODE

- Brad Bernstein
- Ross Hardison
- Mark Gerstein
- Data producers

# factorbook.org

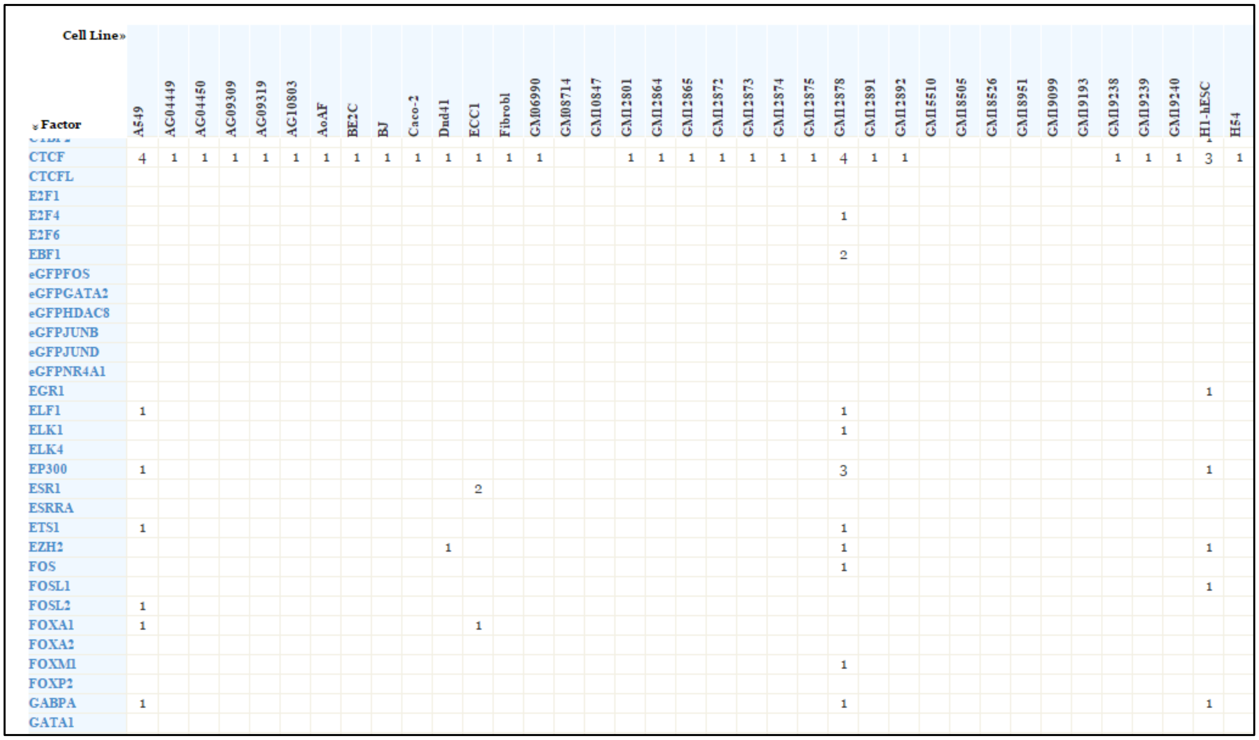
Created by Zhiping Weng's lab at UMass Med, factorbook is a transcription factor (TF)-centric repository of all ENCODE ChIP-seq datasets on TF-binding regions. It includes a number of useful analyses and statistical information for these datasets. In the first release, factorbook contained 457 ChIP-seq datasets on 119 TFs in a number of human cell types. The analyses included average profiles of histone modifications and nucleosome positioning around the TF-binding regions; sequence motifs enriched in the regions; and the distance and orientation preferences between motif sites. The second release (in beta) increases the number of ChIP-seq datasets to 678 on 167 TFs in 90 cell types, and also adds all available ENCODE mouse ChIP-seq data.

*Citation*:

- Wang J, Zhuang J, Iyer S, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*. 2012;22(9):1798-1812. doi:10.1101/gr.139105.112.
- Wang J, Zhuang J, Iyer S, et al. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Research*. 2013;41(Database issue):D171-D176. doi:10.1093/nar/gks1221.

## Features:

**Matrix:** The main page features an alphabetized matrix of TFs (rows) and cell types (columns). Each non-empty cell in the matrix identifies the number of ChIP-seq experiments available for that transcription factor for that particular cell type. Clicking on the name of the transcription factor opens the factorbook page for that TF.

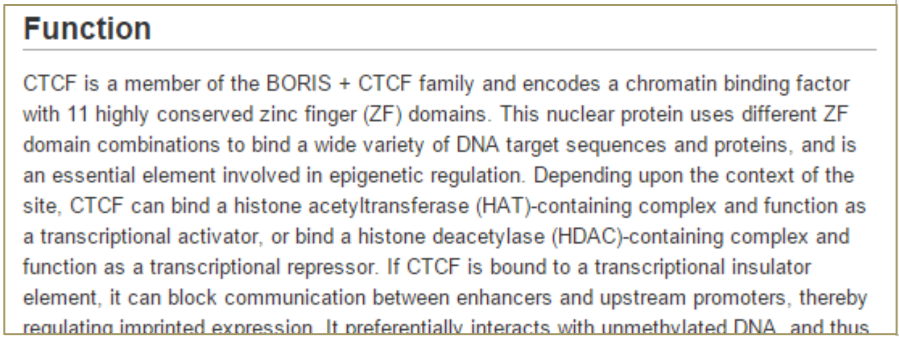


Cell Line»

| Factor | A549 | AG04449 | AG04450 | AG09309 | AG09319 | AG10803 | AoAF | BE2C | BJ | Caco-2 | Dnd41 | ECC1 | Fibrobl | GM06990 | GM08714 | GM10847 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CTCF | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |  |  |
| CTCFL |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| E2F1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| E2F4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| E2F6 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| EBF1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| eGFPFOS |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| eGFPGATA2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| eGFPHDAC8 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| eGFPJUNB |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| eGFPJUND |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| eGFPNR4A1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| EGR1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ELF1 | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ELK1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ELK4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| EP300 | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ESR1 |  |  |  |  |  |  |  |  |  |  |  | 2 |  |  |  |  |
| ESRRA |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ETS1 | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| EZH2 |  |  |  |  |  |  |  |  |  |  | 1 |  |  |  |  |  |
| FOS |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FOSL1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FOSL2 | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FOXA1 | 1 |  |  |  |  |  |  |  |  |  |  | 1 |  |  |  |  |
| FOXA2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FOXM1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FOXP2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| GABPA | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| GATA1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

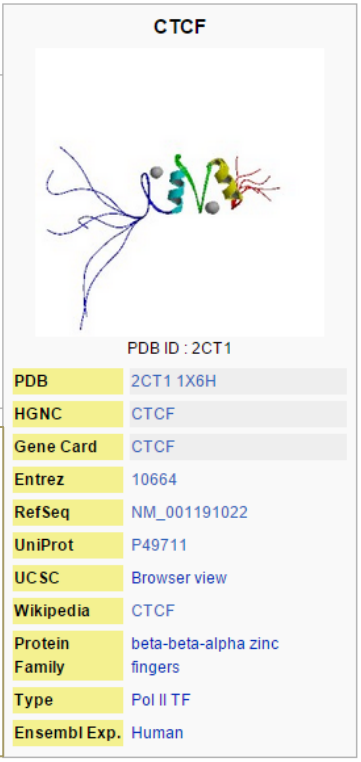
| Factor | GM12801 | GM12864 | GM12865 | GM12872 | GM12873 | GM12874 | GM12875 | GM12878 | GM12891 | GM12892 | GM15510 | GM18505 | GM18526 | GM18951 | GM19099 | GM19193 | GM19238 | GM19239 | GM19240 | H1-hESC | H54 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CTCF | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 | 1 | 1 |  |  |  |  |  |  | 1 | 1 | 1 | 3 | 1 |
| CTCFL |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| E2F1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| E2F4 |  |  |  |  |  |  |  | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| E2F6 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| EBF1 |  |  |  |  |  |  |  | 2 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| eGFPFOS |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| eGFPGATA2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| eGFPHDAC8 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| eGFPJUNB |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| eGFPJUND |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| eGFPNR4A1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| EGR1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 1 |  |
| ELF1 |  |  |  |  |  |  |  | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ELK1 |  |  |  |  |  |  |  | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ELK4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| EP300 |  |  |  |  |  |  |  | 3 |  |  |  |  |  |  |  |  |  |  |  | 1 |  |
| ESR1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ESRRA |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ETS1 |  |  |  |  |  |  |  | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| EZH2 |  |  |  |  |  |  |  | 1 |  |  |  |  |  |  |  |  |  |  |  | 1 |  |
| FOS |  |  |  |  |  |  |  | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FOSL1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 1 |  |
| FOSL2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FOXA1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FOXA2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FOXM1 |  |  |  |  |  |  |  | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FOXP2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| GABPA |  |  |  |  |  |  |  | 1 |  |  |  |  |  |  |  |  |  |  |  | 1 |  |
| GATA1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**Function:** This section contains a brief overview of the molecular function of the TF. If known, information about the TF's protein family, consensus-binding sequence, functional-binding partners, and disease phenotypes will be described. This information was distilled from UCSC TF annotations, RefSeq, and Gene Card. A table with the 3D protein structure of TF (if available) if also provided, along with links to resource outside of factorbook for the given TF including PDB, HGNC, Gene Card, Entrez, RefSeq, UCSC, UniProt, UCSC, ENCODE Project, and Wikipedia.



## Function

CTCF is a member of the BORIS + CTCF family and encodes a chromatin binding factor with 11 highly conserved zinc finger (ZF) domains. This nuclear protein uses different ZF domain combinations to bind a wide variety of DNA target sequences and proteins, and is an essential element involved in epigenetic regulation. Depending upon the context of the site, CTCF can bind a histone acetyltransferase (HAT)-containing complex and function as a transcriptional activator, or bind a histone deacetylase (HDAC)-containing complex and function as a transcriptional repressor. If CTCF is bound to a transcriptional insulator element, it can block communication between enhancers and upstream promoters, thereby regulating imprinted expression. It preferentially interacts with unmethylated DNA, and thus

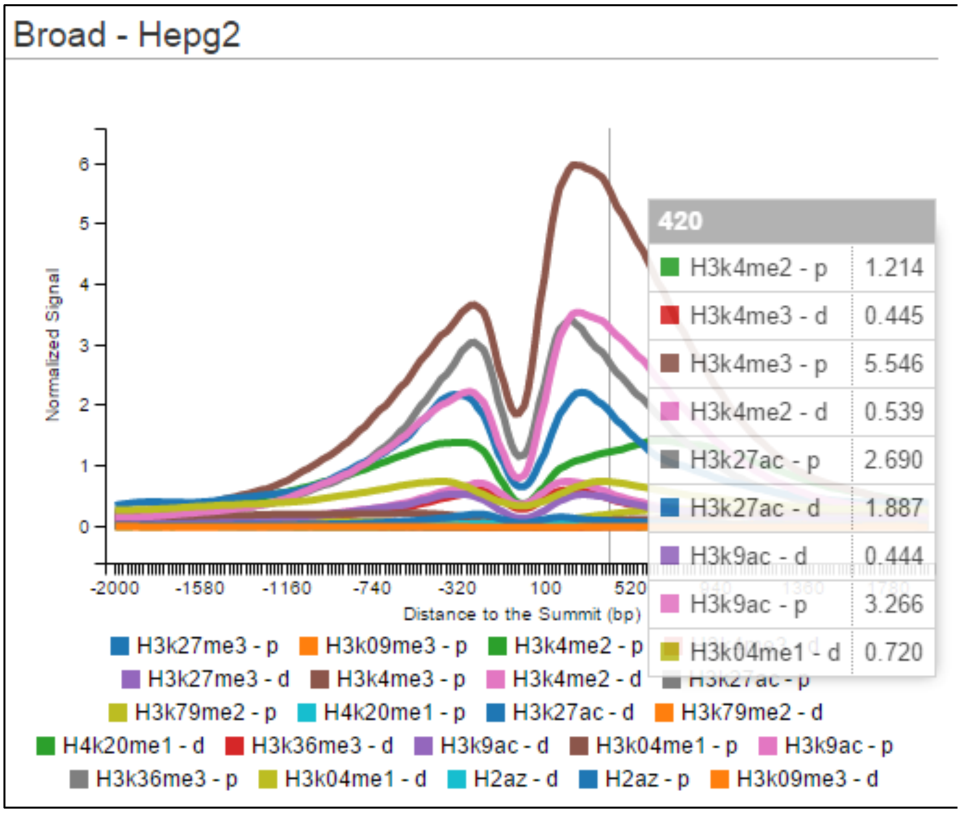


CTCF

PDB ID : 2CT1

| | |
|---|---|
| PDB | 2CT1 1X6H |
| HGNC | CTCF |
| Gene Card | CTCF |
| Entrez | 10664 |
| RefSeq | NM_001191022 |
| UniProt | P49711 |
| UCSC | Browser view |
| Wikipedia | CTCF |
| Protein Family | beta-beta-alpha zinc fingers |
| Type | Pol II TF |
| Ensembl Exp. | Human |

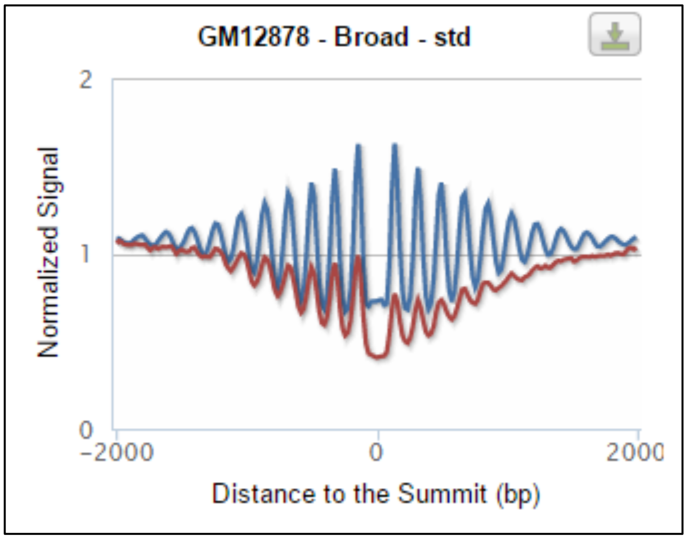
**Average Histone Profiles**: Average histone modification profiles are shown for a +/- 2kb (inclusive) window around the summits (the position with the most sequence reads) of TF ChIP-seq peaks. These profiles are separated by distance to the nearest annotated transcription start site: *proximal* profiles have peaks within 1 kb of a TSS, while *distal* profiles have all other peaks. Proximal profiles are arranged such that the transcriptional direction of the nearest transcript is toward the right. Only histone modification data from the same cell type as the TF ChIP-seq data are shown. The profile figures are interactive JavaScript objects; profiles for a particular histone mark can be show individually or hidden, and tables containing actual data values are shown by default when hovering over the figure.



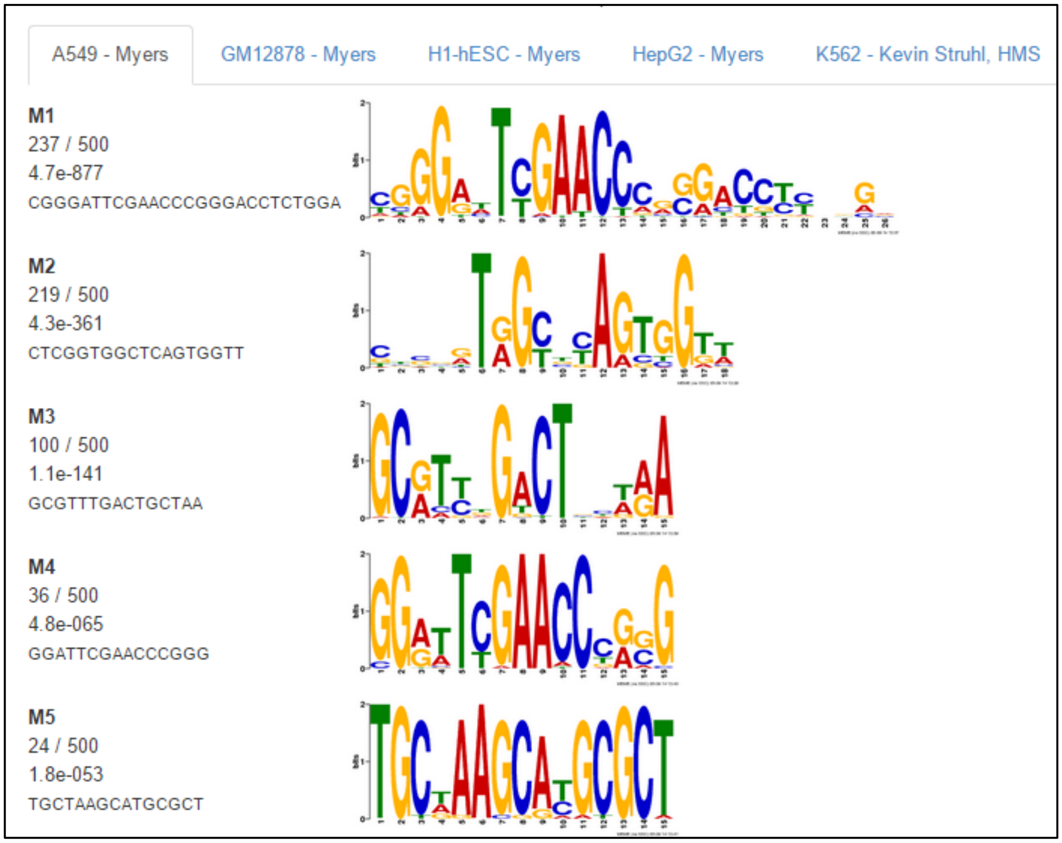
Broad - Hepg2

| 420 | |
|---|---|
| 🟩 H3k4me2 - p | 1.214 |
| 🟥 H3k4me3 - d | 0.445 |
| H3k4me3 - p | 5.546 |
| H3k4me2 - d | 0.539 |
| H3k27ac - p | 2.690 |
| 🟦 H3k27ac - d | 1.887 |
| H3k9ac - d | 0.444 |
| H3k9ac - p | 3.266 |
| H3k04me1 - d | 0.720 |

**Average Nucleosome Profiles**: These profiles show the effect of binding of transcription factors on the regional positioning of nucleosomes. The average nucleosome occupancy profiles are shown for a +/- 2kb (inclusive) window around the summits of TF ChIP-seq peaks. Red lines represent peaks that are proximal to annotated transcripts (i.e. within 1 kb of a TSS), while blue lines represent all other peaks. As for average histone profiles, proximal profiles of nucleosome occupancy are sorted so the transcriptional direction of the nearest transcript is towards the right. The nucleosome positioning data were generated in GM12878 and K562 cell types using MNase digestion of chromatin followed by deep sequencing of mononucleosomal DNA.



**Motif Enrichment**: The sequences of the top 500 TF ChIP-seq peaks were used to identify enriched motifs de novo via the MEME suite of tools. Five motifs are reported (M1 to M5), with motif name, sequence logo, number of peaks out of the top 500 peaks that contain a site for the motif, e-value, and consensus sequence shown.



**M1**
237 / 500
4.7e-877
CGGGATTCGAACCCGGGACCTCTGGA

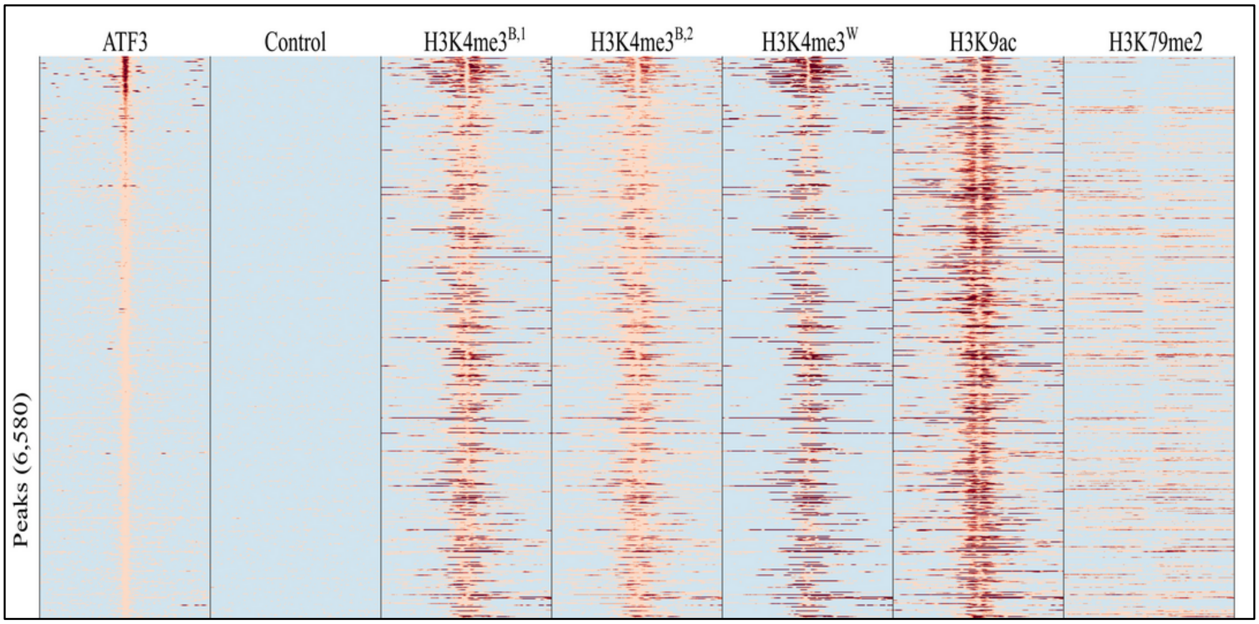
**M2**
219 / 500
4.3e-361
CTCGGTGGCTCAGTGGTT

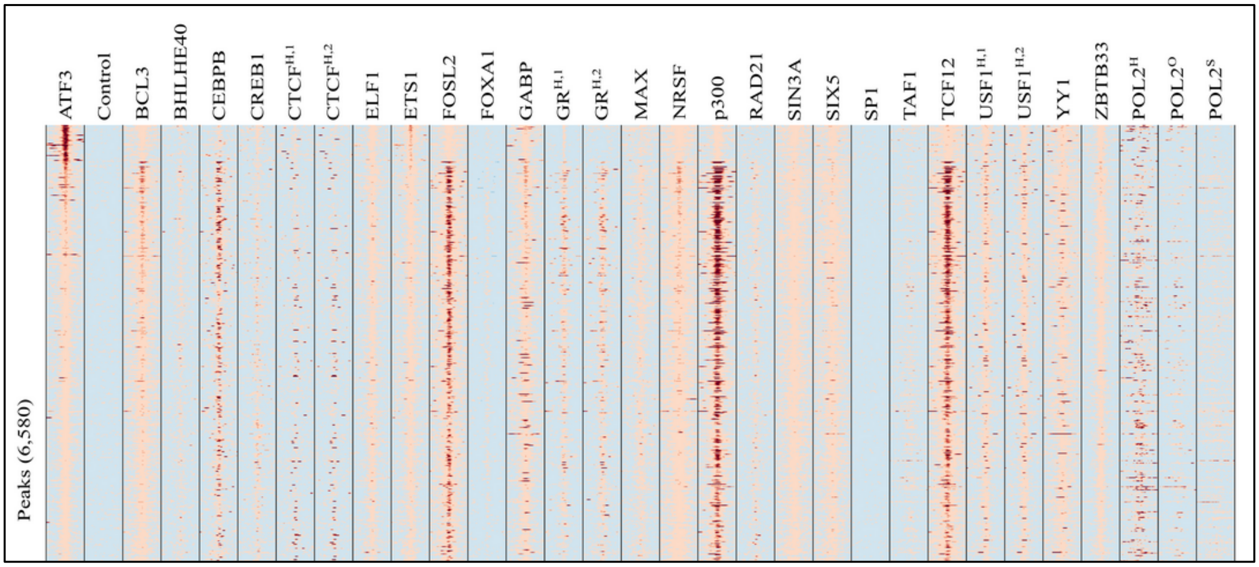
**M3**
100 / 500
1.1e-141
GCGTTTGACTGCTAA

**M4**
36 / 500
4.8e-065
GGATTCGAACCCGGG

**M5**
24 / 500
1.8e-053
TGCTAAGCATGCGCT

**Histone and TF Heatmaps**: Heatmaps are generated to compare a given TF in a specific cell type against the histone marks and other transcription factors with datasets in the same cell type. Each row in a heatmap column indicates a ChIP-seq peak of the "pivot" TF. If fewer than 10,000 peaks are available, all the peaks will be shown; otherwise, a random sampling of 10,000 peaks is made. Rows are sorted in descending order of ChIP-seq signal. For histone marks, enrichment is represented in a normalized scale over a 10kb window centered on the peak summit. For TF heatmaps, binding strengths are represented in a normalized scale over a 2kb window, also centered on the peak summit.


ATF3
Control
H3K4me3B,1
H3K4me3B,2
H3K4me3W
H3K9ac
H3K79me2
Peaks (6,580)


ATF3
Control
BCL3
BHLHE40
CEBPB
CREB1
CTCFH,1
CTCFH,2
ELF1
ETS1
FOSL2
FOXA1
GABP
GRH,1
GRH,2
MAX
NRSF
p300
RAD21
SIN3A
SIX5
SP1
TAF1
TCF12
USF1H,1
USF1H,2
YY1
ZBTB33
POL2H
POL2O
POL2S
Peaks (6,580)