ENCODE Encyclopedia: README

This revision 1 of the Encyclopedia directly annotates the genome (hg19) with candidate promoters and enhancers through integration of DNase-seq, histone mark ChIP-seq, and transcription factor (TF) ChIP-seq datasets. 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 (H3K27ac, H3K4me1, H3K4me3, H3K9ac):

cell types w/ H3K27ac	45
cell types w/ H3K4me1	48
cell types w/ H3K4me3	94
cell types w/ H3K9ac	27

numbers after cell type: 1 = has H3K4me1 datasets 3 = has H3K4me3 datasets
9 = has H3K9ac datasets 27 = has H3K27ac datasets

For each of these 94 cell types, we annotate the DNase peaks with the percentile of each histone mark signal in the matching cell type.

A549 - 1,3,9,27
AG04449 - 3
AG04450 - 3,27
AG09309 - 3
AG09319 - 3
AG10803 - 3
Adult_Th1_
AoAF - 3
BE2_C - 3
BJ - 3
Breast_vHMEC - 1,3
CD14
CD14 - 1,3,27
CD19 - 1,3,27
CD20+_RO01778 - 3
CD20
CD34
CD34+_Mobilized
CD3 - 1,3,27
CD4+_Naive_Wb11970640
CD4+_Naive_Wb78495824
CD4 - 3
CD56 - 1,3,27
CD8 - 3,27
CMK
Caco-2 - 3
Fetal_Adrenal_Gland - 1,3,27
Fetal_Brain - 1,3,9
Fetal_Heart - 1,3,9
Fetal_Intestine_Large - 1,3,27
Fetal_Intestine_Small - 1,3,27
Fetal_Kidney - 1,3,9
Fetal_Kidney_Left
Fetal_Kidney_Right
Fetal_Lung - 1,3,9
Fetal_Lung_Left
Fetal_Lung_Right
Fetal_Muscle_Arm
Fetal_Muscle_Back
Fetal_Muscle_Leg - 1,3,27
Fetal_Muscle_Lower_Limb_Skeletal
Fetal_Muscle_Trunk - 1,3,27
Fetal_Muscle_Upper_Limb_Skeletal
Fetal_Muscle_Upper_Trunk
Fetal_Ovary
Fetal_Placenta - 1,3,27
Fetal_Renal_Cortex
Fetal_Renal_Cortex_Left
Fetal_Renal_Cortex_Right
Fetal_Renal_Pelvis
Fetal_Renal_Pelvis_Left
Fetal_Renal_Pelvis_Right
Fetal_Skin
Fetal_Spinal_Cord
Fetal_Spleen
Fetal_Stomach - 1,3,27
Fetal_Testes
Fetal_Thymus - 1,3,27
Fibroblasts_Fetal_Skin_Abdomen
Fibroblasts_Fetal_Skin_Back
Fibroblasts_Fetal_Skin_Biceps_Left
Fibroblasts_Fetal_Skin_Biceps_Right
Fibroblasts_Fetal_Skin_Quadriceps_Left
Fibroblasts_Fetal_Skin_Quadriceps_Right
Fibroblasts_Fetal_Skin_Scalp
Fibroblasts_Fetal_Skin_Upper_Back
GM04503
GM04504
GM06990 - 3
GM12864 - 3
GM12865 - 3
GM12878 - 1,3,9,27
Gastric - 1,3,27
H1-hESC - 1,3,9,27
H1_BMP4_Derived_Mesendoderm - 1,3,9,27
H1_BMP4_Derived_Trophoblast - 1,3,9,27
H1_Derived_Mesenchymal_Stem_Cells - 1,3,9,27
H1_Derived_Neuronal_Progenitor - 1,3,9,27
H7-hESC - 3
HA-h
HA-sp - 3
HAEpiC
HAc - 3
HBMEC - 3
HBVP
HBVSMC
HCF - 3
HCFaa - 3
HCM - 3
HCPEpiC - 3
HCT-116 - 3,27
HConF
HEEpiC - 3
HEK293T
HFF - 3
HFF-Myc - 3
HGF
HIPEpiC
HL-60 - 3
HMEC - 1,3,9,27
HMF - 3
HMVEC-LBl
HMVEC-LLy
HMVEC-dAd
HMVEC-dBl-Ad
HMVEC-dBl-Neo
HMVEC-dLy-Ad
HMVEC-dLy-Neo
HMVEC-dNeo
HNPCEpiC
HPAEC
HPAF - 3
HPF - 3
HPdLF
HRCEpiC
HRE - 3
HRGEC
HRPEpiC - 3
HSMM - 1,3,9,27
HSMMtube - 1,3,9,27
HUVEC - 1,3,9,27
HVMF - 3
HeLa-S3 - 1,3,9,27
Heart - 1,3,9
HepG2 - 1,3,9,27
IMR90 - 1,3,9,27
Jurkat - 3
K562 - 1,3,9,27
LHCN-M2
LNCaP - 3
M059J
MCF-7 - 3,27
Mobilized_CD3
Mobilized_CD4
Mobilized_CD56
Mobilized_CD8
Monocytes-CD14+_RO01746 - 1,3,9,27
NB4 - 3
NH-A - 1,3,9,27
NHBE_RA
NHDF-Ad - 1,3,9,27
NHDF-neo - 3
NHEK - 1,3,9,27
NHLF - 1,3,9,27
NT2-D1 - 1,3,9
Ovary - 1,3,27
PANC-1 - 1,3,27
Pancreas - 1,3,27
Penis_Foreskin_Fibroblast - 1,3,27
Penis_Foreskin_Keratinocyte - 1,3,9,27
PrEC
Psoas_Muscle - 1,3,27
RPMI-7951
RPTEC - 3
SAEC - 3
SK-N-MC - 3
SK-N-SH_RA - 3
SKMC - 3
Small_Intestine - 1,3,27
T-47D
Th1
Th17
Th1_Wb33676984
Th1_Wb54553204
Th2
Th2_Wb33676984
Th2_Wb54553204
Treg_Wb78495824
Treg_Wb83319432
WERI-Rb-1 - 3
WI-38 - 3
bone_marrow_HS27a
bone_marrow_HS5
bone_marrow_MSC
iPS_DF_19_11 - 1,3,27
iPS_DF_19_7
iPS_DF_4_7

Procedure for generating annotations:

Metadata:

- Roadmap: data from Dr. Anshul Kundaje
- ENCODE: JSON from https://www.encodeproject.org

DNase master peaks:

- The Stam lab (UW) merged all DNase peak data from the Stam and Crawford (Duke U.) labs. This merging process formed one combined DNase-seq dataset with nonoverlapping DNase hypersensitive regions. The Stam lab then identified the "master" peak in each region, defined as the the peak in the region with highest peak height/z-score.
    - files:
        - multi-tissue.master.v2.hg19.bed
        - multi-tissue.master.v2.ntypes.simple.names.hg19.txt

Distal and proximal DHS peaks:

- The master DNase peaks were separated into TSS proximal and TSS distal groups based on whether or not they intersected a 2000bp window centered on any GENCODE TSS (gencode.v19.TSS_plusminus_2K.sorted.bed).
    - Track details include names and number of cell types in each merged DNase hypersensitive regions
    - files:
        - dnase_track_distal.bb
        - dnase_track_proximal.bb

Histone mark (H3K27ac, H3K4me1, H3K4me3, H3K9ac) tracks:

- Signal files downloaded from Roadmap and encodeproject.org
- Distal and proximal tracks
    - For each DNase master peak, the average histone signals in the matching cell type was calculated in a 1000bp 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 1000bp genomic regions that were outside all DNase peaks and ENCODE blacklisted regions.
    - DNase master peaks that have at least one cell type with 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
    - Track details include actual percentile over background.
    - files:
        - distal.H3K27ac.cellType_specific.bb
        - distal.H3K4me1.cellType_specific.bb

        - distal.H3K4me3.cellType_specific.bb
        - distal.H3K9ac.cellType_specific.bb
        - proximal.H3K27ac.cellType_specific.bb
        - proximal.H3K4me1.cellType_specific.bb
        - proximal.H3K4me3.cellType_specific.bb
        - proximal.H3K9ac.cellType_specific.bb

Transcription factor (TF) binding

- ChIPseq TF peaks were obtained from encodeproject.org
- For each of the distal and proximal DNase master peaks, overlapping TF ChIP-seq peaks across all cell types available were identified. The TF peaks with 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.
- files:
    - tf.distal.bb
    - tf.proximal.bb

Visualization:

- UCSC Genome Browser Track Hub File
- WashU browser