Using ENCODE Data To Interpret Disease-associated Genetic Variation

Mike Pazin
National Human Genome Research Institute, NIH
ENCODE Users Meeting
June 8, 2016
Welcome

• Objectives
  – We want to tell the community about the ENCODE resource
  – We want to hear community experiences and suggestions
Overview

• The ENCODE Resource
• Use of ENCODE to illuminate the role of genetic variation in human disease
• Accessing ENCODE materials
Reading The Human Genome Is Difficult

- Genetic code very powerful for 1% of the human genome
- No correspondingly powerful regulatory code
- Sequence conservation can identify some candidate functional elements (but not when or where they act)
- Regulatory regions aren’t always in the same order as gene targets

Need unbiased experimental investigation

![Gene expression graph showing IL10 and IL19]

- 99% Non-coding
- 1% Coding
Non-coding DNA Is Important For Disease And Gene Regulation

• Vast majority of common disease associations and heritability lie outside of protein-coding regions

• Non-coding DNA variants are known to cause human diseases and alter human traits (FXS, ALS)

Functional information is needed to interpret the role of genetic variation in human disease, and to apply genomics in the clinic.

PMID: 22955828, PMID: 25439723, PMID: 23128226
PMID: 17477822, PMID: 25679767
1,500 Letters Of Our
3 Billion Letter Genome
Maps And Annotation Help Us To Understand The Sequence

agccaagcagcaagtttctgctgtttatatttttagctcttactatatttctacttttaccattgaaatattgaggaagttatttatattttctatttttatatttatattattattattttatattatatattatatatatgtatatatatataattactattacacataattattttatatatatgagtaacctagtacctcttttccagagcaataatgaaatttcacagtatgaaaatggaagaaatcaataaaaattatatagctgtacctgtggcgaagtacctcttcgtggacaggtactcatatcgctggacaaggtgagtacctatgtggtatcacaaatgctctttccaaagccccctctcgcagctctttccccctatagactttcctatcatgcagcagcattacctcctcctggcccccctttctaagcattgccttgagattttctaagaattttctttctgttgccagcagactatcatatttttagctgatgacataataggtttctgtctagtgtagataggagataagcagaaatgcaataaaagaaaaccatccagagggaaactctttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
Richer Maps Provide More Information
Richer Maps Provide More Information
Richer Maps Provide More Information
Richer Maps Provide More Information
ENCODE: Encyclopedia Of DNA Elements

- Identify all candidate functional elements in the genome
- Make resource freely available to community
  - genetic basis of disease
  - gene regulation
ENCODE Data Types

Science 306:636, 2004
ENCODE Data Are Cell-Type Specific

Clustering based on DHS patterns

- Paraxial Mesoderm
- Primitive Mesoderm
- Hematopoietic
- Endothelia
- Endoderm/Ectoderm
- ESCs

Size and composition of DHS landscape

- ESCs
- Hematopoietic Progenitors
  - Th2
    - 30% (17% 54%)
  - B-cell (rest.)
    - 34% (21% 46%)
  - NK-cell
    - 27% (15% 57%)

DHSs originating in ES cells
DHSs arising during CD34+ development
DHSs arising during lymphoid development

Stamatoyannopoulos, Cell 154:888, 2013
ENCODE Accomplishments

• Sharing 1000s of datasets
  – No embargo
  – Unrestricted access
  – High quality
  – Uniformly processed

• Sharing software

• Data interoperability
Hundreds of Consortium publications

~1500 community publications using ENCODE data:

~675 Human Disease
~600 Basic Biology
~225 Methods/Software Development

Cancer 38%
Allergy, Autoimmunity 13%
Human Genetics 10%
Neurologic, Psychiatric 9%
Cardiovascular 6%
Metabolic 6%
Summary- ENCODE Resource

• Freely shared catalog of genomic data and candidate genomic functional elements

• ENCODE is built upon established techniques and interpretations developed for the study of gene regulation

• ENCODE maps can be used to make predictions about genome function
Overview

- The ENCODE Resource
- Use of ENCODE to illuminate the role of genetic variation in human disease
- Accessing ENCODE materials
Standard ENCODE Use Cases: Hypothesis Generation

Major use: Hypothesis generation and refinement

• Prediction of causal variants/regulatory elements
• Prediction of target genes
• Prediction of target cell types
• Prediction of upstream regulators
Prediction of Causal Variants

• Multiple variants may be in linkage disequilibrium

• The causal variant may not have been tested during data collection

• Multiple variants may be causal

Snyder, Genome Research 22:1748, 2012
Many GWAS Associations Lie In Regions Annotated By ENCODE And CF Epigenomics Data

Stamatoyannopoulos, Science 337:1190, 2012
ENCODE/Epigenomics Data From HaploReg

HaploReg v4.1

HaploReg is a tool for exploring annotations of the noncoding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease-associated loci. Using LD information from the 1000 Genomes Project, linked SNPs and small indels can be visualized along with chromatin state and protein binding annotation from the Roadmap Epigenomics and ENCODE projects, sequence conservation across mammals, the effect of SNPs on regulatory motifs, and the effect of SNPs on expression from eQTL studies. HaploReg is designed for researchers developing mechanistic hypotheses of the impact of non-coding variants on clinical phenotypes and normal variation.

Update 2015.11.05: Version 4.1 GWAS and eQTL have been updated; a simpler pruning strategy is applied when combining GWAS; and links out to other NHGRI/EBI GWAS hits and GRASP QTL hits are provided.

Update 2015.09.15: Version 4.0 now includes many recent eQTL results including the GTEx pilot, four different options for defining enhancers using Roadmap Epigenomics data, and a complete set of source files for download and local analysis. Older versions available: v3, v2, v1.

Build Query Set Options Documentation

Use one of the three methods below to enter a set of variants. If an r² threshold is specified (see the Set Options tab), results for each variant will be shown in a separate table along with other variants in LD. If r² is set to NA, only queried variants will be shown, together in one table.

Query (comma-delimited list of rsIDs OR a single region as chrN:start-end): 

Query SNP: rs16892766 and variants with r² >= 0.8

www.broadinstitute.org/mammals/haploreg/

Workshop Session 3

ENCODE Data From RegulomeDB

http://regulomedb.org/

Workshop Session 3
Cherry, Snyder, Genome Research 22-1790, 2012
ENCODE cis-element Browser

Welcome to YUE Lab
Computational and Functional Genomics/Epigenomics

Query human ENCODE data!

Option 1: Search gene expression across ~60 human cell types (total 108 datasets)

human (hg19) ➔ Gene name (Sox2, Nanog ...) ➔ submit!

Option 2: Search cis-elements in a given genomic region

human (hg19) ➔ chr8 ➔ start: 128390000 end: 128410000 ➔ submit!

Option 3: Search cis-elements surrounding a gene

human (hg19) ➔ Gene name (Sox2, Nanog ...) ➔ submit!

Extended region (default +/- 100kb) ➔ submit!

Workshop Session 3

https://www.encodeproject.org/data/annotations/
ENCODE cis-element Browser

Candidate cis-elements in your queried region.

- **DNaseI Hypersensitive Sites:**

<table>
<thead>
<tr>
<th>Coordinate</th>
<th>Tissue/cell type</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr8:128394860-128395010</td>
<td>NHDF-Ad</td>
</tr>
<tr>
<td>chr8:128395580-128395730</td>
<td>HSMMtube, HSMM</td>
</tr>
<tr>
<td>chr8:128398205-128398355</td>
<td>Osteoblast</td>
</tr>
<tr>
<td>chr8:128398585-128398735</td>
<td>Osteoblast</td>
</tr>
<tr>
<td>chr8:128399500-128399650</td>
<td>GM12876, NHDF-Ad, HSMM</td>
</tr>
<tr>
<td>chr8:128400960-128401110</td>
<td>HSMM, HSMMtube</td>
</tr>
<tr>
<td>chr8:128402490-128402630</td>
<td>HSMM</td>
</tr>
<tr>
<td>chr8:128403560-128403730</td>
<td>HMEC, Osteoblast, HSMM, NHDF-Ad, HSMMtube, NH-A, HeLa-S3, HEK, NHLF</td>
</tr>
<tr>
<td>chr8:128404560-128404710</td>
<td>HMEC</td>
</tr>
<tr>
<td>chr8:128404720-128404870</td>
<td>HSMM</td>
</tr>
<tr>
<td>chr8:128405400-128405550</td>
<td>HSMM</td>
</tr>
<tr>
<td>chr8:128407420-128407570</td>
<td>HeLa-S3</td>
</tr>
<tr>
<td>chr8:128407985-128408035</td>
<td>HUVEC, Osteoblast, NHDF-Ad</td>
</tr>
<tr>
<td>chr8:128408160-128408310</td>
<td>HMEC</td>
</tr>
</tbody>
</table>

- **TF binding Site:**

<table>
<thead>
<tr>
<th>Coordinate</th>
<th>TF</th>
<th>tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr8:128398585-128398735</td>
<td>USF1</td>
<td>USF1(KS62), USF1(SK-N-SH_RA)</td>
</tr>
<tr>
<td>chr8:128399500-128399650</td>
<td>RUNX3, SPI1</td>
<td>RUNX3(GM12876), SPI1(GM12878), SPI1(GM12891)</td>
</tr>
<tr>
<td>chr8:128403560-128403730</td>
<td>multiple</td>
<td>CEBP8(HeLa-S3), CEBP8(IMR90), EP300(HeLa-S3), FOS(MCF10A-Er-Src), FOXA1(A549), GATA3(T-47D), JUN(HeLa-S3), JUND(HeLa-S3), MAX(HeLa-S3), MYC(MCF10A-Er-Src), NR3C1(A549), POLR2A(HeLa-S3), POLR2A(MCF10A-Er-Src), RCOR1(HeLa-S3), SMC3(HeLa-S3), STAT3(HeLa-S3), STAT3(MCF10A-Er-Src)</td>
</tr>
</tbody>
</table>

25. [https://www.encodeproject.org/data/annotations/](https://www.encodeproject.org/data/annotations/)

Workshop Session 3
Prediction of Target Genes

- Regulatory regions can operate on multiple, distal genes
- The target gene could be a non-coding RNA

Many GWAS Associations Are Predicted To Be Linked To Distal Genes

Stamatoyannopoulos, Science 337:1190, 2012
Many GWAS Associations Are Predicted To Be Linked To Distal Genes

Stamatoyannopoulos, Science 337:1190, 2012
Prediction of Linkage Between Regulatory Elements and Genes

Regulatory Elements Database

This database provides a user interface to the results of the analysis presented in Sheffield et al. (2013). This is the first DNaseI signal in 112 human samples. Questions? contact Nathan Sheffield. See also: Supplemental Files

There are 5 ways to explore the data:
- By CELLTYPE - Select cell-types to include or exclude.
- By CLUSTER - View promoter, CpG-island, and conserved element overlap for all clusters, and select individual clusters:
  - Muscle-specific cluster: cluster 1520
  - Prostate and hepatocyte cluster: cluster 910
  - Prostate-only cluster: cluster 2483
  - Hematopoietic cluster: cluster 25
  - Pluripotency cluster: cluster 104
- By GENE - Search by gene of interest
  - MyoG
  - RBFOX1
  - LIN28A
  - HBG1
- By COORDINATE: Give chr, start, stop to find all regulatory elements in a region.
  - IRF2 regulator: chr4: 185240845-185240995
  - MyoD1 regulator: chr11: 17828545-17828695
  - Blood regulator: chr3: 128166420-128166570
- By FACTOR: find a specific TF of interest
  - CTCF
  - Myf
  - AP1
  - Pou5f1 (Oct-4)

http://dnase.genome.duke.edu
Furey, Crawford, Stamatoyannopoulos, Genome Res. 23:777, 2013
Prediction of Linkage Between Regulatory Elements and Genes

http://dnase.genome.duke.edu
Furey, Crawford, Stamatoyannopoulos, Genome Res. 23:777, 2013
Prediction of Linkage Between Regulatory Elements and Genes

http://dnase.genome.duke.edu

Furey, Crawford, Stamatoyannopoulos, Genome Res. 23:777, 2013
ENCODE cis-element Browser

Welcome to YUE Lab
Computational and Functional Genomics/Epigenomics

Query human ENCODE data!

Option 1: Search gene expression across ~ 60 human cell types (total 108 datasets)

human (hg19) Gene name(Sox2, Nanog ...)

Option 2: Search cis-elements in a given genomic region

human (hg19) chr1 start: end:

Option 3: search cis-elements surrounding a gene

human (hg19) Gene name(Sox2, Nanog ...)

Extended region (default +/- 100kb) kb

Option 4: search cis-elements LINKED to a gene based on DNaseI HSS specificity

human (hg19) Gene name(Sox2, Nanog ...) IL13

Workshop Session 3  
https://www.encodeproject.org/data/annotations/
## ENCODE cis-element Browser

**Human (hg19)**

**Gene** [IL13](https://www.encodeproject.org/data/annotations/) [NM_0002188, ENSG00000169194, ENST00000304506]

### Cis-element linked to your queried gene.

### Cis-element lined by DNaseI Hypersensitive Sites Linkage:

<table>
<thead>
<tr>
<th>Proximal DHS (TSS)</th>
<th>start</th>
<th>end</th>
<th>Gene</th>
<th>Distal DHS</th>
<th>start</th>
<th>end</th>
<th>correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr5</td>
<td>131992140</td>
<td>131992290</td>
<td>IL13</td>
<td>chr5</td>
<td>131512800</td>
<td>131512950</td>
<td>0.743283</td>
</tr>
<tr>
<td>chr5</td>
<td>131992140</td>
<td>131992290</td>
<td>IL13</td>
<td>chr5</td>
<td>131558440</td>
<td>131558590</td>
<td>0.761866</td>
</tr>
<tr>
<td>chr5</td>
<td>131992140</td>
<td>131992290</td>
<td>IL13</td>
<td>chr5</td>
<td>131571820</td>
<td>131571970</td>
<td>0.782866</td>
</tr>
<tr>
<td>chr5</td>
<td>131992140</td>
<td>131992290</td>
<td>IL13</td>
<td>chr5</td>
<td>131720440</td>
<td>131720590</td>
<td>0.766176</td>
</tr>
<tr>
<td>chr5</td>
<td>131992140</td>
<td>131992290</td>
<td>IL13</td>
<td>chr5</td>
<td>131732540</td>
<td>131732690</td>
<td>0.739405</td>
</tr>
<tr>
<td>chr5</td>
<td>131992140</td>
<td>131992290</td>
<td>IL13</td>
<td>chr5</td>
<td>131745200</td>
<td>131745350</td>
<td>0.765629</td>
</tr>
<tr>
<td>chr5</td>
<td>131992140</td>
<td>131992290</td>
<td>IL13</td>
<td>chr5</td>
<td>131747860</td>
<td>131748010</td>
<td>0.749684</td>
</tr>
</tbody>
</table>

---

**Workshop Session 3**

[https://www.encodeproject.org/data/annotations/](https://www.encodeproject.org/data/annotations/)
Prediction of Target Cell Types

- Some diseases are known to affect multiple cell types
- The defect may not be intrinsic to the cell type with obvious pathology
- The disease etiology may not be completely known
Prediction of Linkage Between Regulatory Elements and Cell Type

DHS: #2174550
chr5: 131972960-131973110
Belongs to SOM cluster: 2072
Site Hypersensitivity Profile

Cluster Profile:

RESOURCES
Correlated Genes:
p-values indicate significant higher or lower correlation; 1 gene found
Gene P-value
IL13 0.009

External Databases
UCSC
Ensembl

http://dnase.genome.duke.edu
Furey, Crawford, Stamatoyannopoulos, Genome Res. 23:777, 2013
Prediction of Linkage Between Regulatory Elements and Cell Type

Workshop Session 3

www.broadinstitute.org/mammals/haploreg/
http://regulomedb.org/

Cherry, Snyder, Genome Research 22-1790, 2012
ENCODE cis-element Browser

Welcome to YUE Lab
Computational and Functional Genomics/Epigenomics

Query mouse ENCODE data!

Option 1: Search gene expression across 32 mouse tissue/cell types

mouse (mm9)  Gene name (Sox2, Nanog ...)

Submit!
ENCODE cis-element Browser

Gene **II10** (mCG02645) [NM_010548, ENSMUSG00000016529, ENSMUST00000016673]

<table>
<thead>
<tr>
<th>Tissue</th>
<th>pkm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adrenal</td>
<td>0</td>
</tr>
<tr>
<td>Bladder</td>
<td>0.07</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>0</td>
</tr>
<tr>
<td>CH12</td>
<td>348.42</td>
</tr>
<tr>
<td>CNS_D1</td>
<td>0</td>
</tr>
</tbody>
</table>

Workshop Session 3  [https://www.encodeproject.org/data/annotations/](https://www.encodeproject.org/data/annotations/)
ENCODE And Epigenomics Data Can Be Used To Predict Cell Types

https://github.com/mauranolab/GWAS_plots
Stamatoyannopoulos, Science 337:1190, 2012
Summary- ENCODE Use Cases

Major use: Hypothesis generation and refinement
- Prediction of causal variants/regulatory elements
- Prediction of target genes
- Prediction of target cell types
- Prediction of upstream regulators

- Genetic v. epigenetic
- Germline v. somatic
Overview

• The ENCODE Resource
• Use of ENCODE by the research community
• Accessing ENCODE materials

Workshop Session 2, 1, 5
The ENCODE (Encyclopedia of DNA Elements) Consortium is a collaboration of research groups funded by the National Institutes of Health. The goal of ENCODE is to build a catalog of functional elements in the human genome, including elements that affect DNA and RNA levels, and regulatory elements that control if a gene is active.

Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI),
ENCODE Data

Experiment Matrix

- Assay
  - ChiP-seq: 3105
  - shRNA RNA-seq: 445
  - DNase-seq: 421
  - RNA-seq: 307
  - eCLIP: 270
  - Target of assay
    - DNA binding: 3105
    - Transcription: 1823
    - DNA accessibility: 470
    - RNA binding: 460
    - DNA methylation: 310

- Organism
  - Homo sapiens: 4732
  - Mus musculus: 1609

- Biosample type
  - immortalized cell line: 3405
  - tissue: 1548
  - primary cell: 913
  - in vitro differentiated cells: 227
  - stem cell: 223

- Organ
  - brain: 459
  - skin of body: 186
  - liver: 183
  - heart: 155
  - lung: 141

- Project
  - ENCODE: 6452
  - Roadmap: 3127
  - modENCODE: 883
  - modERN: 198

6452 results

- Clear Filters

- ASSAY

- BIOSAMPLE
  - immortalized cell line
    - K562: 374227
    - HepG2: 150218
    - GM12878: 174
    - HeLa-S3: 103
    - A549: 110
    - ...and 141 more

- tissue
  - liver: 102
  - heart: 84
  - forebrain: 65
  - hindbrain: 65
  - midbrain: 65
  - ...and 101 more

- primary cell
  - endothelial cell of umbilical vein: 35
  - Purkinje cell: 1
  - ...and 61 more

https://www.encodeproject.org

Workshop Session 2
The goal of the Encyclopedia of DNA Elements (ENCODE) Project is to build a comprehensive catalog of candidate functional elements in the genome. The catalog includes genes (protein-coding and non-protein coding), transcribed regions, and regulatory elements, as well as information about the tissues, cell types and conditions where they are found to be active. The current phase of ENCODE (2012-2016) greatly expands the number of cell types, data types and assays and includes the study of both the human and mouse genomes.

Like the Human Genome Project, the ENCODE Project seeks rapid data dissemination and use by the entire scientific community. Accordingly, to encourage the widest possible use of the datasets, all data produced will be available for unrestricted use immediately upon release to public databases, eliminating the nine-month moratorium previously used by ENCODE.

Data Use Policy for External Users

External data users may freely download, analyze and publish results based on any ENCODE data without restrictions as soon as they are released. This applies to all datasets, regardless of type or size, and includes no grace period for ENCODE data producers, either as individual members or as part of the Consortium. Researchers using unpublished ENCODE data are encouraged to contact the data producers to discuss possible coordinated publications; however, this is optional. The Consortium will continue to publish the results of its own analysis efforts in independent publications.

We request that researchers who use ENCODE datasets (published or unpublished) in publications and talks cite the ENCODE Consortium in all of the following ways:

1. Cite the Consortium’s most recent integrative publication (PMID: 22955616; PMC: PMC3439153);
2. Reference the ENCODE Data Coordination Center (DCC) or GEO accession numbers of the datasets (DCC accession: ENCSR037HRJ; GEO accession: GSE30567);
3. And acknowledge the ENCODE Consortium and the ENCODE production laboratory(s) generating the particular dataset(s)
ENCODE Data

The ENCODE (Encyclopedia of DNA Elements) Consortium is a collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive inventory of functional elements in the human genome, including elements that control DNA and RNA levels, and regulatory elements that control the expression of a gene in a tissue-specific manner.

Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI),
ENCODE Data

Region search

Please enter valid coordinates

Long-range regulatory elements (enhancers, repressors/silencers, insulators) Promoters Transcripts

https://www.encodeproject.org
ENCODE Data

https://www.encodeproject.org
ENCODEx Data

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr8:117,630,483-117,630,883  401 bp. enter position, gene symbol or search terms

Click to alter the display density of ENCF001DFV and similar substracks

 Workshop Session 2

https://www.encodeproject.org
The ENCODE (Encyclopedia of DNA Elements) Consortium, a collaboration of research groups funded by the National Human Genome Research Institute (NHGRI), is a resource dedicated to the annotation of functional elements in the human genome, including genomic DNA and RNA levels, and regulatory elements that control when a gene is active.

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

[https://www.encodeproject.org](https://www.encodeproject.org)
Introduction

The ENCODE Consortium not only produces data, but also analyzes the data in an integrative fashion. The ENCODE Encyclopedia organizes the most salient analysis products into annotations, and provides tools to search and visualize them. The Encyclopedia has three levels of annotations:

- Ground level annotations are typically derived directly from the experimental data.
- Middle level annotations integrate multiple types of experimental data and multiple ground level annotations.
- Top level annotations integrate a broad range of experimental data and ground and middle level annotations.
Ground Level Annotations

Gene expression (RNA-seq)
The expression levels of genes annotated by GENCODE 19 in over 100 human cell types and 70 mouse cell types.
[ Long RNA-seq Data | Query | Download | Method ]

Transcription factor binding (TF ChIP-seq)
Peaks (enriched genomic regions) of TFs computed from ~900 human and mouse ChIP-seq experiments.
[ Raw Data | Peaks ]
Visualize sequence motifs and other information [ Factorbook ]

Histone mark enrichment (ChIP-seq)
Peaks of a variety of histone marks computed from ~600 ChIP-seq experiments.
[ Raw Data | Peaks ]

Open chromatin (DNase-seq)
DNase I hypersensitive sites (also known as DNase-seq peaks) computed from ~300 human and mouse experiments.
[ Raw Data | Peaks ]

Topologically associating domains (TADs) and compartments (Hi-C)
TADs and A and B compartments computed from 12 human cell lines.
[ Raw Data | Visualize ]
ENCODEx Encyclopedia

Middle Level Annotations

Promoter-like regions
DNase hypersensitivity and histone modification H3K4me3 are well-known indicators of promoter function. We have developed an unsupervised method that combines DNase and H3K4me3 signals in the same cell type to predict promoter-like regions. When used to predict ranked gene expression from RNA-seq data, our method shows higher accuracy than DNase and H3K4me3 individually. We have applied this method to 107 human cell types and 14 mouse cell types with both DNase and H3K4me3 data generated by the ENCODE and Roadmap Epigenomic consortia. For cell and tissue types with only H3K4me3 data, we centered predictions on H3K4me3 peaks and ranked them by H3K4me3 signals. You can query these promoter-like regions by genomic locations, nearby genes, or SNPs, and visualize them in the UCSC and WashU genome browsers. [Visualize | Method]

Enhancer-like regions
DNase hypersensitivity and histone modification H3K27ac are well-known indicators of enhancer function. We have developed an unsupervised method that combines DNase and H3K27ac signals in the same cell type to predict enhancer-like regions. When tested on mouse transgenic assays, our method shows higher accuracy than DNase and H3K27ac individually. We have applied this method to 47 human cell types and 14 mouse cell types with both DNase and H3K27ac data generated by the ENCODE and Roadmap Epigenomic consortia. For cell and tissue types with only H3K27ac or DNase data, we rank the peaks using the available data and make predictions of enhancer-like regions. You can query these enhancers by genomic locations, nearby genes, or SNPs, and visualize them in the UCSC and WashU genome browsers. [Visualize | Method]

Enhancer-like genomic regions were tested on VISTA experimentally-validated enhancer elements: [VISTA]

https://www.encodeproject.org
Top Level Annotations

Chromatin states

Semi-automated genomic annotation methods such as ChromHMM and Segway take as input a panel of epigenomic data (including histone mark ChIP-seq and DNase-seq) in a particular cell type and use machine learning methods to simultaneously partition the genome into segments and assign chromatin states to these segments; the states are assigned such that two segments with the same state exhibit similar epigenomic patterns. The procedure is "semi-automated" because states are then manually compared with known biological information in order to designate each state as an enhancer-like, promoter-like, gene body, etc.

[Search]

Variant Annotation

Over the past decade, Genome Wide Association Studies (GWAS) have provided insights into how genetic variations contribute to human diseases. However, over 80% of the variants reported by GWAS are in noncoding regions of the genome and the mechanism of how they contribute to disease onset is unknown. By integrating data from the ENCODE project and other public sources, RegulomeDB and HaploReg are two resources developed by ENCODE labs to aid the research community in annotating GWAS variants. FunSeq2 is another ENCODE resource for annotating both germline and somatic variants, particularly in the noncoding regions of cancer genomes.

[RegulomeDB] [HaploReg] [FunSeq2]
The ENCODE (Encyclopedia of DNA Elements) Consortium is a collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive catalog of functional elements in the human genome, including elements that regulate DNA and RNA levels, and regulatory elements that control or activate a gene.

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

ENCODE Encyclopedia of DNA Elements

Publications using ENCODE data

ENCODE-funded publications

These are publications by members of the ENCODE, mouse ENCODE, and modENCODE consortia.

- Key integrative publications by consortia members
- Human ENCODE publications
- Mouse ENCODE publications
- modENCODE publications
- Technology development publications
- ENCODE pilot project publications

Community publications

These are publications that use ENCODE, published by authors not funded by ENCODE, as well as papers that use modENCODE data, published by authors not funded by modENCODE. The ENCODE project tracks these papers to assess impact of the resource and to provide examples of how the resource can be used. Please contact Mike Pazin at NHGRI to suggest publications to add to these lists.

- Human disease publications
- Basic biology publications
- Software tools
- modENCODE community publications

https://www.encodeproject.org
ENCODE Data Standards

The ENCODE (Encyclopedia of DNA Elements) Consortium is a collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive map of functional elements in the human genome, including DNA, chromatin structure, and RNA levels, and regulatory elements that control when and where a gene is active.

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

https://www.encodeproject.org
ENCODE Data Standards

Overview

The ENCODE consortium analyzes the quality of the data produced using a variety of metrics. This page describes the data standards and metrics that are used to evaluate the data and what they appear to measure. These quality metrics will be updated on occasion to include analysis of more recent data.

It is important to note that quality metrics for evaluating epigenomic assays is an area of research, so standards are emerging as more metrics are used with more datasets and types of experiments. The typical values for a quality metric can be quite different with different assays, or even comparing different features in the same assays, such as different antibodies used in ChIP-seq experiments. Currently there is no single measurement that identifies all high-quality or low-quality samples. As with quality control for other types of experiments, multiple assessments (including manual inspection of tracks) are useful because they may capture different concerns. Comparisons within an experimental method (e.g., comparing replicates to each other, or comparing values for one antibody in several cell types, or the same antibody and cell type in different labs) can help identify possible stochastic error.

Experimental guidelines

The ENCODE Consortium has adopted uniform guidelines for the most common ENCODE experiments. The guidelines have evolved over time as technologies have changed. The current guidelines are informed by results gathered during the project. Previous versions of the standards are also available for reference.

- Current experiment guidelines
- Antibody characterizations guidelines
- Terms and definitions

Quality metrics

https://www.encodeproject.org
ENCODE Software Tools

ENCODE: Encyclopedia of DNA Elements

The ENCODE (Encyclopedia of DNA Elements) is a collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive collection of functional elements in the human genome, including elements that control gene expression, and RNA levels, and regulatory elements that control the activity of a gene.

Image credits: Daryl Leja (NHGRI), Ian Dunham (EBI).

Quick Start

News Follow @EncodeDCC

https://www.encodeproject.org
The goal of the ENCODE project is to generate a comprehensive catalog of all functional elements. To facilitate this task, members of the consortium have developed and refined software tools. View all software used or developed by the ENCODE Consortium or select one of the following:

- **Software tools used to identify ENCODE elements**: On this page are brief descriptions of some of the software used to identify ENCODE elements. Software for identification of functional elements, for integrated analysis of multiple data types, and for quality measurement of the data are described.

- **Software tools used to generate ENCODE quality metrics**: On this page are brief descriptions of some of the software used to generate quality metrics for ENCODE datasets.

- **External software tools used to create the ENCODE resource**: On this page are brief descriptions of some of the software used to create the ENCODE resource. This software was not funded by ENCODE, or developed by the consortium.

- **Software tools and resources for applying and analyzing ENCODE data**: On this page are brief descriptions of software and resources that others might find useful for analyzing and using ENCODE data in their own research.

https://www.encodeproject.org
International Human Epigenome Consortium (IHEC)

- Data Portal: [http://epigenomesportal.ca/ihec/](http://epigenomesportal.ca/ihec/)
- Goal: Coordinate production of 1000 human epigenome maps for cellular states relevant to health and disease [http://ihec-epigenomes.org](http://ihec-epigenomes.org)
- Can view by consortium, by assay, by cell type
- Data from 8 consortia
Summary- Accessing ENCODE Resources

• ENCODE portal [https://www.encodeproject.org]
  – Display/download ENCODE and Roadmap Epigenomics data
  – Data Standards
  – Software tools
  – Publications
  – Encyclopedia prototype

• ENCODE Analysis Tools
  – RegulomeDB [http://regulomedb.org/]
  – HaploReg [http://www.broadinstitute.org/mammals/haploreg/]
  – Regulatory Elements Database [http://dnase.genome.duke.edu]
  – RegulomeDB GWAS Database [http://www.regulomedb.org/GWAS/]

• ENCODE Tutorials
  – [https://www.encodeproject.org/tutorials/]

• ENCODE mailing list:
  – [https://mailman.stanford.edu/mailman/listinfo/encode-announce]

• IHEC resources
  – IHEC Data Portal [http://epigenomesportal.ca/ihec]
Goals Of ENCODE

• Catalog all functional elements in the genome
• Develop freely available resource for research community

ENCODE data are being used in the study of human disease and basic biology