Using ENCODE Data to Interpret Disease-associated Genetic Variation

Dan Gilchrist
National Human Genome Research Institute, NIH
Society of Toxicology Satellite Meeting
March 17, 2016
ENCODE 2016: Research Applications and Users Meeting
June 8-10, 2016 at Stanford University

- Hands-on workshops on how to navigate, analyze, and integrate ENCODE and mouseENCODE data into your research
- Leading-edge research applications from distinguished invited speakers from across the research community
- Tutorials on newly-available informatics pipelines that greatly facilitate working with ENCODE data
- Short talks selected from abstracts
- For details and registration see:
  
  https://www.encode2016.org
Overview

• The ENCODE Resource
• Use Cases of ENCODE to illuminate the role of genetic variation in human disease
• Accessing ENCODE materials
Overview

• The ENCODE Resource
• Use Cases of ENCODE to illuminate the role of genetic variation in human disease
• Accessing ENCODE materials
ENCODE:
Encyclopedia Of DNA Elements

• Identify all candidate functional elements in the genome
  – Promoters, enhancers, transcribed regions

• Make resource freely available to community for use in studies of:
  – genetic basis of disease
  – gene regulation
  – whatever you want

Research Parasites Welcome!
ENCODE: Encyclopedia Of DNA Elements

- Thousands of data sets
- Hundreds of different bio-samples
- Data sets related to exposure
ENCODE: Encyclopedia Of DNA Elements

ENCODE ‘Phase 4’ RFAs –
Applications Due March 21st

• Expanding the Encyclopedia of DNA Elements in Human and Mouse
• Characterizing ENCODE Functional Elements
• Computational Analysis of ENCODE Data
• Data Coordination Center
• Data Analysis Center
Understanding Non-coding DNA Is Central To NHGRI Goals

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.
Many GWAS Associations Lie In Regions Annotated By ENCODE And Epigenomics Data
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 candidate functional elements (but not when or where they act)
- Need unbiased experimental investigation
1,500 Letters Of Our 3 Billion Letter Genome

agccaagcagcaaaagtggttctgctttttatattttgtagctttactatattcactttttaccattgaaatatattgaggaagtatttatattttctattttlatattatatatttatattttatgattattaatatattactattacacataattatttttatatatgtaagtaaccaatgacttcctttttccagagcaataaagtgaatttccagcagttgagtagctttgtctacaaaattgtgttcccttttctcaagctgtttttctaagattttctactatatctgagcaacactttgttagcaagaaaatgttaaagggctgucagagcctaacaggactaattgactgcagtaggcaataatctgctgaaatttagctgatgacataatgggtctgtctatagttagataggagaatagccaaatgcaataagaaaaacatatccagaggaactttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
Maps And Annotation Help Us To Understand The Sequence
Richer Maps Provide More Information
ENCODE Data: Built on Decades of Research Into Gene Regulation

Science 306:636, 2004
RNA Data

- RNA amount (mRNA and short RNA)
- Cell Specificity
- Transcript modeling (GENCODE)

https://encodeproject.org
ENCODE Chromatin Structure Data

- DNase
- Histone modifications
- DNA methylation
  - Enhancers
  - Promoters
  - Transcription factor footprints
  - Transcribed regions
  - Active and repressed regions
  - Cell specificity

Richmond, PDB 1ZBB

Richmond, PDB 1KX5
ENCODE Nucleic Acid Binding Data

• DNA binding proteins (Transcription factors)
  – Activators
  – Repressors
  – Remodelers
  – RNA Polymerases
  – Cell specificity

• RNA binding proteins
  – RNA Splicing
  – Translation
  – RNA Stability
  – RNA Localization
  – Cell specificity

Harrison, PDB_1A02
ENCODE Consortium Structure

Data Production Groups
- RNA
- Histone Mods
- DNase
- DNAm
- TF Binding
- RBP Binding

Data Coordination Center

Data Analysis Center
- Analysis Working Group
  - Gene Models
  - Chromatin States
  - Element ID

Technology Development Groups

Computational Analysis Groups

The ENCYCLOPEDIA
Summary- ENCODE Resource

- Freely shared catalog of 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
Using ENCODE Data: Access for Scientists with Varying Expertise

• Many tools for deriving biological insight and hypotheses from ENCODE data
• Many do not require informatics experience
• Many do not require deep functional genomics background
• Encouraging broad use of ENCODE resources
Standard ENCODE Use Cases: Hypothesis Generation

• Prediction of causal variants/regulatory elements
• Prediction of target genes
• Prediction of target cell types

• Applied to germline genetic variation, but many other applications (e.g., differentially methylated regions)
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
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.

Query SNP: rs6575353 and variants with $r^2 >= 0.8$

www.broadinstitute.org/mammals/haploreg/
ENCODE Data From RegulomeDB

http://regulomedb.org/
Cherry, Snyder, Genome Research 22-1790, 2012
RegulomeDB GWAS Database

Data supporting chr8:117630682 (rs16892766)

Score: 2b
Likely to affect binding

Human Feb. 2009 (GRCh37) chr0:117,630,462-117,630,032 (401 bp)
RefSeq Genes
Publications: Sequences in scientific articles

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

Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE

Transcription Factor ChIP-seq from ENCODE

Placental Mammal Basewise Conservation by PhyloP

Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples

Repetitive Elements by RepeatMasker

http://www.regulomedb.org/GWAS/
Cherry, Snyder, Genome Research 22-1790, 2012
### 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>chr8:128402460-128402630</td>
<td>HSMM</td>
</tr>
<tr>
<td>chr8:128405680-128405730</td>
<td>HMEC, osteoblast, HSMM, NHDF-Ad, HSMMtube, NH-A, HeLa-S3, NHEK, 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>chr8:128405300-128405550</td>
<td>HSMM</td>
</tr>
<tr>
<td>chr8:128405420-128405570</td>
<td>HeLa-S3</td>
</tr>
<tr>
<td>chr8:128407530-128407630</td>
<td>HMEC</td>
</tr>
<tr>
<td>chr8:128408305-128408310</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(K562), USF1(SK-N-SH_RA)</td>
</tr>
<tr>
<td>chr8:128399500-128399650</td>
<td>RUNX3</td>
<td>RUNX3(GM12878), SPI1(GM12878), SPI1(GM12891)</td>
</tr>
</tbody>
</table>

---

[https://www.encodeproject.org/data/annotations/](https://www.encodeproject.org/data/annotations/)
ENCODE Browser
Viewing Locus Of Interest

https://genome.ucsc.edu/encode/
Prediction of Target Genes

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

Many GWAS Associations Lie In Regions Linked To Distal Genes

Stamatoyannopoulos, Science 337:1190, 2012
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**

Cis-elements linked to your queried gene.

**Human (hg19)**
Gene **IL13** [NM_002188, ENSG00000169194, ENST00000304506]

### 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>
<tr>
<td>chr5</td>
<td>131993580</td>
<td>131993730</td>
<td>IL13</td>
<td>chr5</td>
<td>131917860</td>
<td>131918010</td>
<td>0.797702</td>
</tr>
<tr>
<td>chr5</td>
<td>131993580</td>
<td>131993730</td>
<td>IL13</td>
<td>chr5</td>
<td>131921920</td>
<td>131922070</td>
<td>0.800141</td>
</tr>
<tr>
<td>chr5</td>
<td>131993580</td>
<td>131993730</td>
<td>IL13</td>
<td>chr5</td>
<td>131970980</td>
<td>131971130</td>
<td>0.772113</td>
</tr>
<tr>
<td>chr5</td>
<td>131993580</td>
<td>131993730</td>
<td>IL13</td>
<td>chr5</td>
<td>131971640</td>
<td>131971790</td>
<td>0.763557</td>
</tr>
<tr>
<td>chr5</td>
<td>131993580</td>
<td>131993730</td>
<td>IL13</td>
<td>chr5</td>
<td>131972960</td>
<td>131973110</td>
<td>0.797839</td>
</tr>
<tr>
<td>chr5</td>
<td>131993580</td>
<td>131993730</td>
<td>IL13</td>
<td>chr5</td>
<td>131977060</td>
<td>131977210</td>
<td>0.848905</td>
</tr>
<tr>
<td>chr5</td>
<td>131993580</td>
<td>131993730</td>
<td>IL13</td>
<td>chr5</td>
<td>131990420</td>
<td>131990570</td>
<td>0.855445</td>
</tr>
<tr>
<td>chr5</td>
<td>131993580</td>
<td>131993730</td>
<td>IL13</td>
<td>chr5</td>
<td>131993880</td>
<td>131994030</td>
<td>0.769354</td>
</tr>
<tr>
<td>chr5</td>
<td>131993580</td>
<td>131993730</td>
<td>IL13</td>
<td>chr5</td>
<td>132011780</td>
<td>132011930</td>
<td>0.756074</td>
</tr>
<tr>
<td>chr5</td>
<td>131993580</td>
<td>131993730</td>
<td>IL13</td>
<td>chr5</td>
<td>132077740</td>
<td>132077890</td>
<td>0.820222</td>
</tr>
<tr>
<td>chr5</td>
<td>131993580</td>
<td>131993730</td>
<td>IL13</td>
<td>chr5</td>
<td>132083520</td>
<td>132083670</td>
<td>0.770558</td>
</tr>
<tr>
<td>chr5</td>
<td>131993580</td>
<td>131993730</td>
<td>IL13</td>
<td>chr5</td>
<td>132164240</td>
<td>132164390</td>
<td>0.837496</td>
</tr>
</tbody>
</table>

[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 genes 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

www.broadinstitute.org/mammals/haploreg/

http://regulomedb.org/
Cherry, Snyder, Genome Research 22-1790, 2012
ENCODE cis-element Browser

Gene **IL10** (mGI1-2645) [NM_010548, ENSMUSG00000016529, ENSMUST00000016673]

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Value</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>
<tr>
<td>CNS_D2</td>
<td>0.01</td>
</tr>
</tbody>
</table>

https://www.encodeproject.org/data/annotations/
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
Overview

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

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 analyses 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 assay, 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

Quality metrics
The ENCODE consortium analyzes the quality of the data produced using a variety of metrics. Those generated for datasets published as part of

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

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.

- **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
Data Use Policy for External Users

The goal of the Encyclopedia of DNA Elements (ENCODEx 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.

**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:

https://www.encodeproject.org
Accessing ENCODE

Downloading and Visualizing

https://www.encodeproject.org
ENCODE Encyclopedia Prototype

Genomic annotations

- Distal DNase peaks [Download]
- Proximal DNase peaks [Download]
- Distal H3K27ac annotations (cell type specific) [Download]
- Distal H3K4me1 annotations (cell type specific) [Download]
- Distal H3K4me3 annotations (cell type specific) [Download]
- Distal H3K9 ac annotations (cell type specific) [Download]
- Proximal H3K27ac annotations (cell type specific) [Download]
- Proximal H3K4me1 annotations (cell type specific) [Download]
- Proximal H3K4me3 annotations (cell type specific) [Download]
- Proximal H3K9 ac annotations (cell type specific) [Download]
- Distal TF binding sites [Download]
- Proximal TF binding sites [Download]

- Gene expression matrix over ~60 cell types with genes annotated by GENCODE 19 [Download data | Download methods]
- Transcription start site (TSS) lists [View README]
  - GENCODE v19 TSS [Download]
  - GENCODE v19 TSS stratified by strict Fantom5 CAGE clusters [Download]
  - GENCODE v19 TSS stratified by robust Fantom5 CAGE clusters [Download]
  - GENCODE v19 TSS stratified by permissive Fantom5 CAGE clusters [Download]

https://www.encodeproject.org
ENCODE Encyclopedia: Coming Soon!
Publications

Cancer 35%
Autoimmunity/Allergy 15%
Neuro/Psych 13%
CVD 7%

https://www.encodeproject.org
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
ENCODE Accomplishments

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

• Data interoperability

• Sharing software
Publications Using ENCODE Data

Hundreds of Consortium publications
~1500 community publications using ENCODE data:

~340 Human Disease
~500 Basic Biology
~170 Methods/Software Development

- Cancer 38%
- Allergy, Autoimmunity 13%
- Human Genetics 10%
- Neurologic, Psychiatric 9%
- Cardiovascular 6%
- Metabolic 6%
ENCODE Consortium
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
  – http://www.genome.gov/27553900
  – https://www.encodeproject.org/tutorials/

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

• IHEC resources
  – IHEC Home Page http://ihec-epigenomes.org
  – IHEC Data Portal http://epigenomesportal.ca/ihec/
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 7 consortia
A Toxicology User’s Guide to the Roadmap Epigenomics and ENCODE Data Resources

Thursday, March 17, 1:00 PM–6:00 PM

Room 205

Hosted by: Ivan Rusyn, Texas A&M University, College Station, TX; and Lisa Chadwick, NIEHS, Research Triangle Park, NC.

Purpose of the Meeting: Improvements in DNA sequencing technologies have resulted in an exponential increase in the amount of genomic and epigenomic data available. Some of these data have been generated as part of large-scale, focused mapping efforts aimed at understanding how genes are regulated, such as the NIH Roadmap Epigenomics Program, and ENCODE (Encyclopedia of DNA Elements). Efforts such as these can be extremely valuable for hypothesis generation and data mining, but can only be useful if one knows what is available and how to use it. This SOT satellite meeting will provide toxicology researchers with an overview of these two NIH-funded programs, introduce attendees to the informatics tools that have been developed to help navigate these large datasets, and walk through several use cases. The meeting will be of broad interest to researchers interested in learning more about how environmental exposure might impact gene regulation.

Registration: Open registration. No fee to register and attend.

Lectures followed by Q&A and a poster session.

For more information on this Satellite Meeting, contact Lisa Chadwick.
ENCODE And Epigenomics Data Can Be Used To Predict Cell Types

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

Data from Table S7, Stamatoyannopoulos, Crawford, Nature 489:75, 2012