ENCODE Project Update

National Advisory Council for Human Genome Research

September 8, 2014

How can we “read” the human genome sequence?

- Genetic code
  - Moderately good at identifying protein-coding regions, but fine structures difficult to predict from sequence
- No powerful regulatory code
- Evolutionary conservation helps to identify functionally important regions

~5% conserved/ ~1.5% protein coding

- **Function of non-coding conserved sequences?**
- **Function of non-conserved sequences?**

Non-coding DNA Is Important For Disease And Evolution

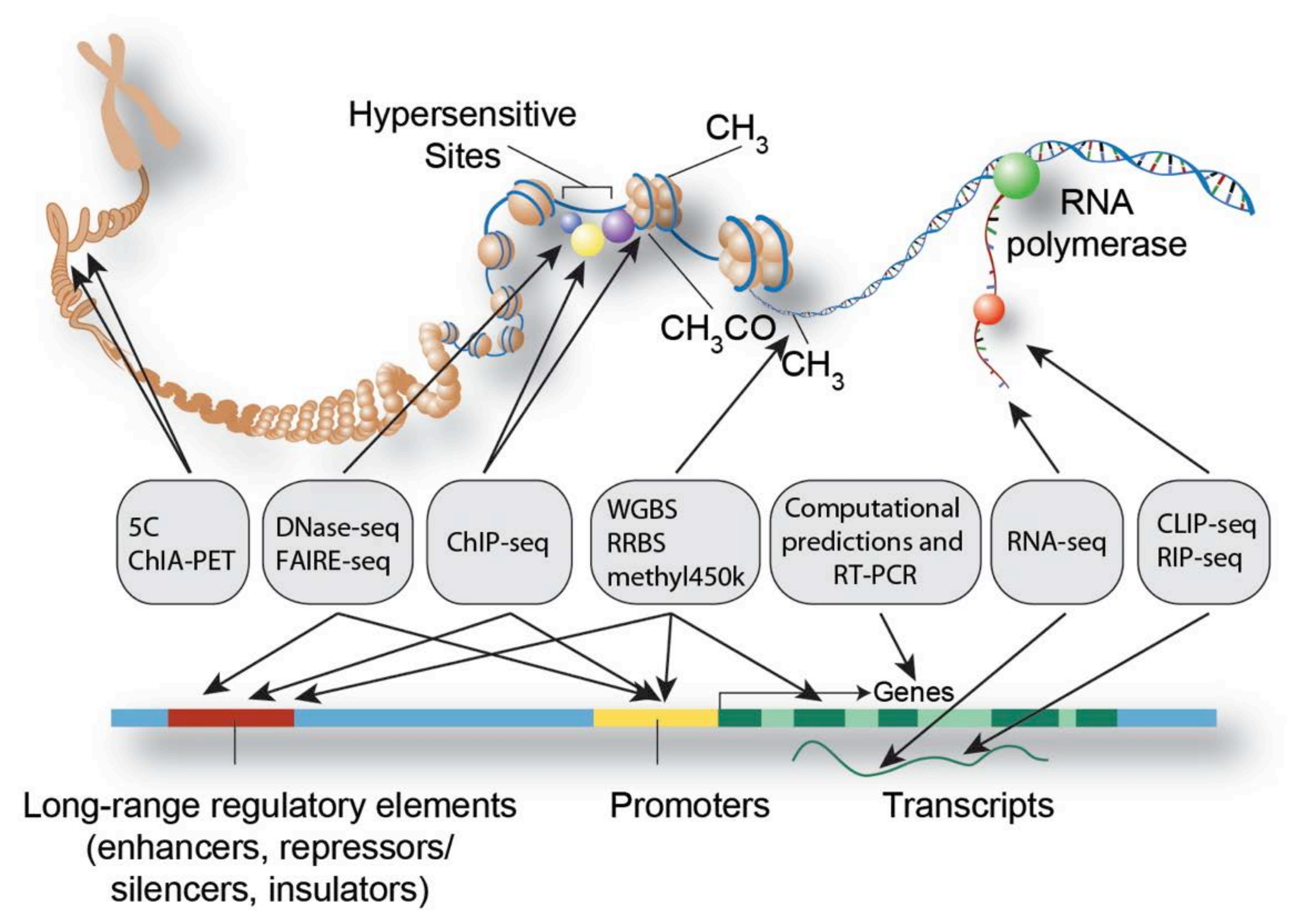
- About 90% of GWAS findings lie outside of protein-coding regions
- Non-coding DNA variants are known to cause human diseases and alter human traits
  - Fragile X Syndrome
  - ALS (amyotrophic lateral sclerosis)
- About 80% of recent adaptation signatures lie outside of protein-coding regions

Stamatoyannopoulos, Science 337:1190, 2012
Kingsley, Nature 484:55, 2012
Sabeti, Cell 152:703, 2013
Fraser, Genome Research 23:1089, 2013
Annu Rev Genomics Hum Genet. 8:109, 2007
Herdewyn,Hum Mol Genet. 21:2412, 2012

ENCODE:
Encyclopedia of DNA Elements

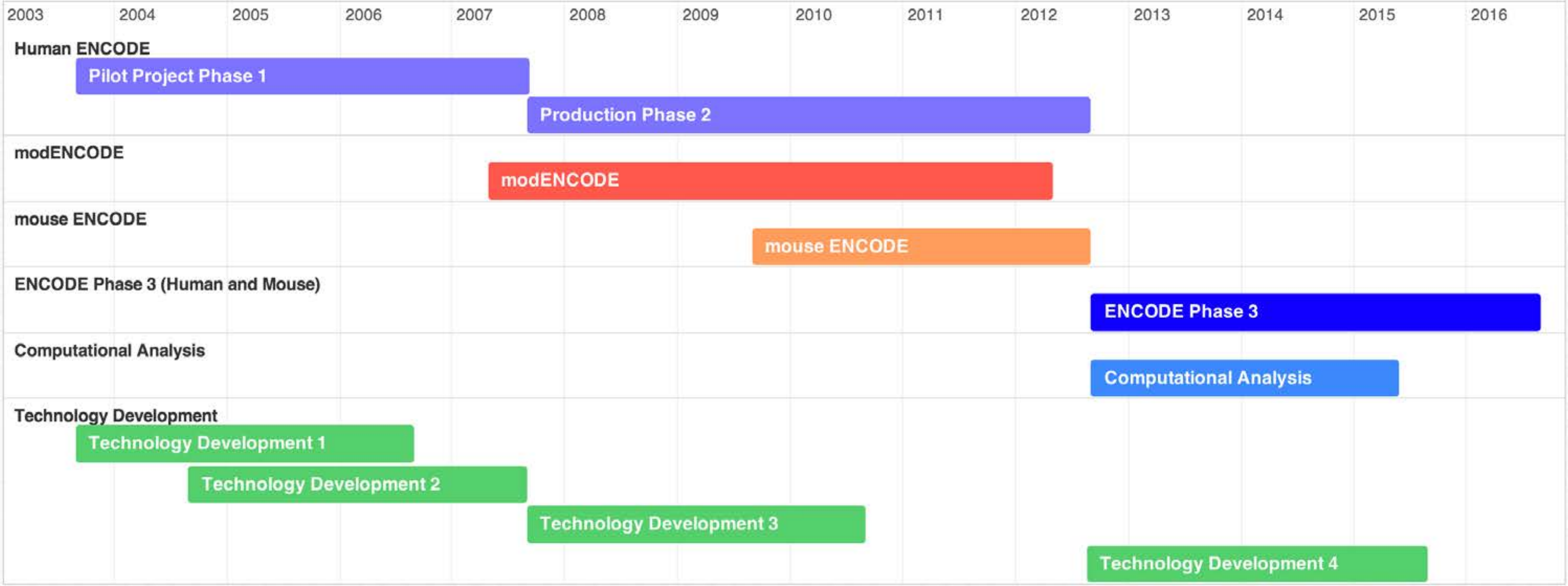
- Compile a comprehensive encyclopedia of all sequence features in the human genome and in the genomes of select model organisms.
  - Rapid pre-publication data release
- Make resource freely available to community to enhance understanding of:
  - regulation of gene expression on a spatial, temporal and quantitative level
  - genetic basis of disease

ENCODE Data Types



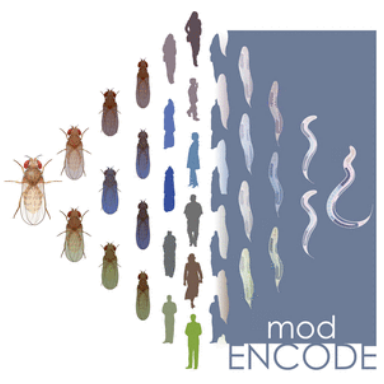
Modified from PLoS Biol 9:e1001046, 2011
Science 306:636, 2004

ENCODE Timeline



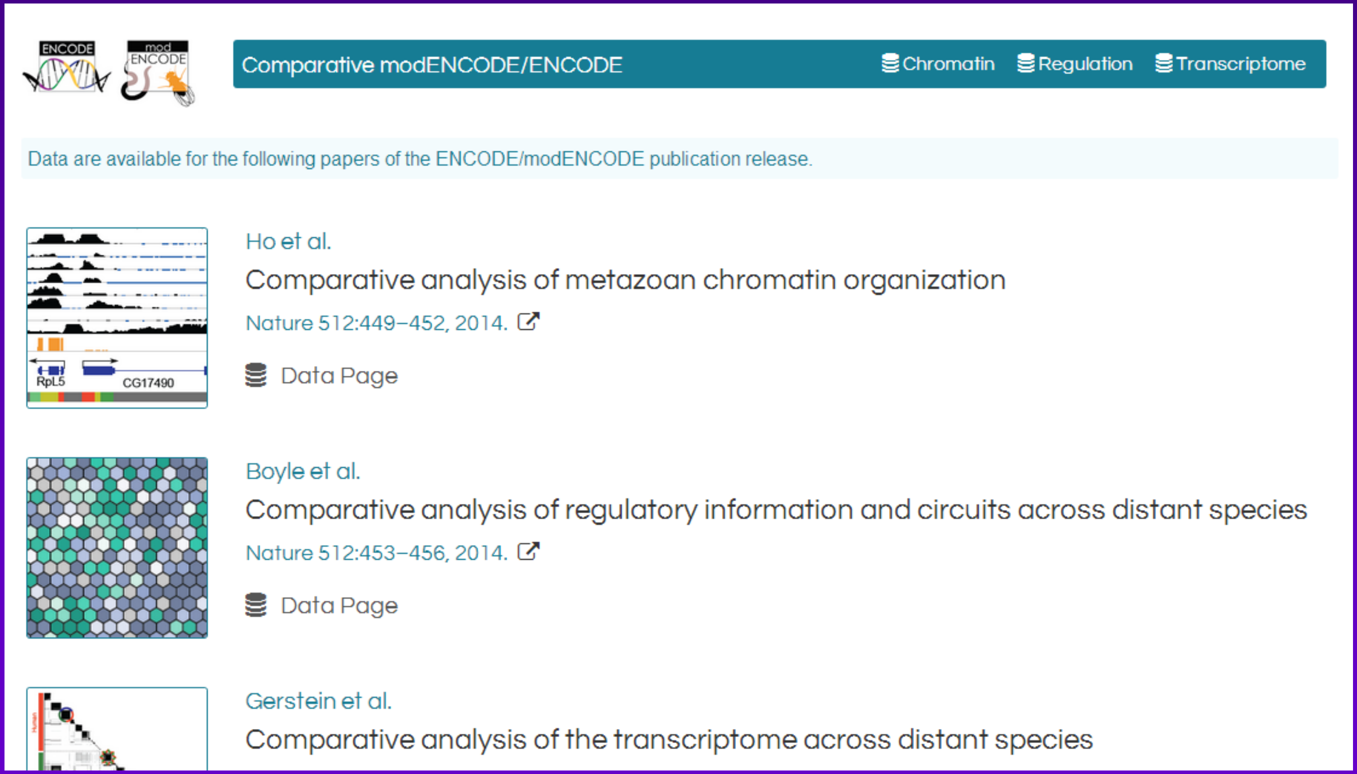
modENCODE Publications

- Worm/fly/human comparative analyses
  - Three main integrated papers published in Nature on August 28, 2014 focused on:
    - Transcription
    - Chromatin
    - Regulation
- More than 15 additional companion papers published in Nature, Genome Research, Genome Biology and other journals



modENCODE Publications

- Data and links to papers can be found at: https://www.encodeproject.org/comparative/



Data are available for the following papers of the ENCODE/modENCODE publication release.

Ho et al.
Comparative analysis of metazoan chromatin organization
Nature 512:449–452, 2014.

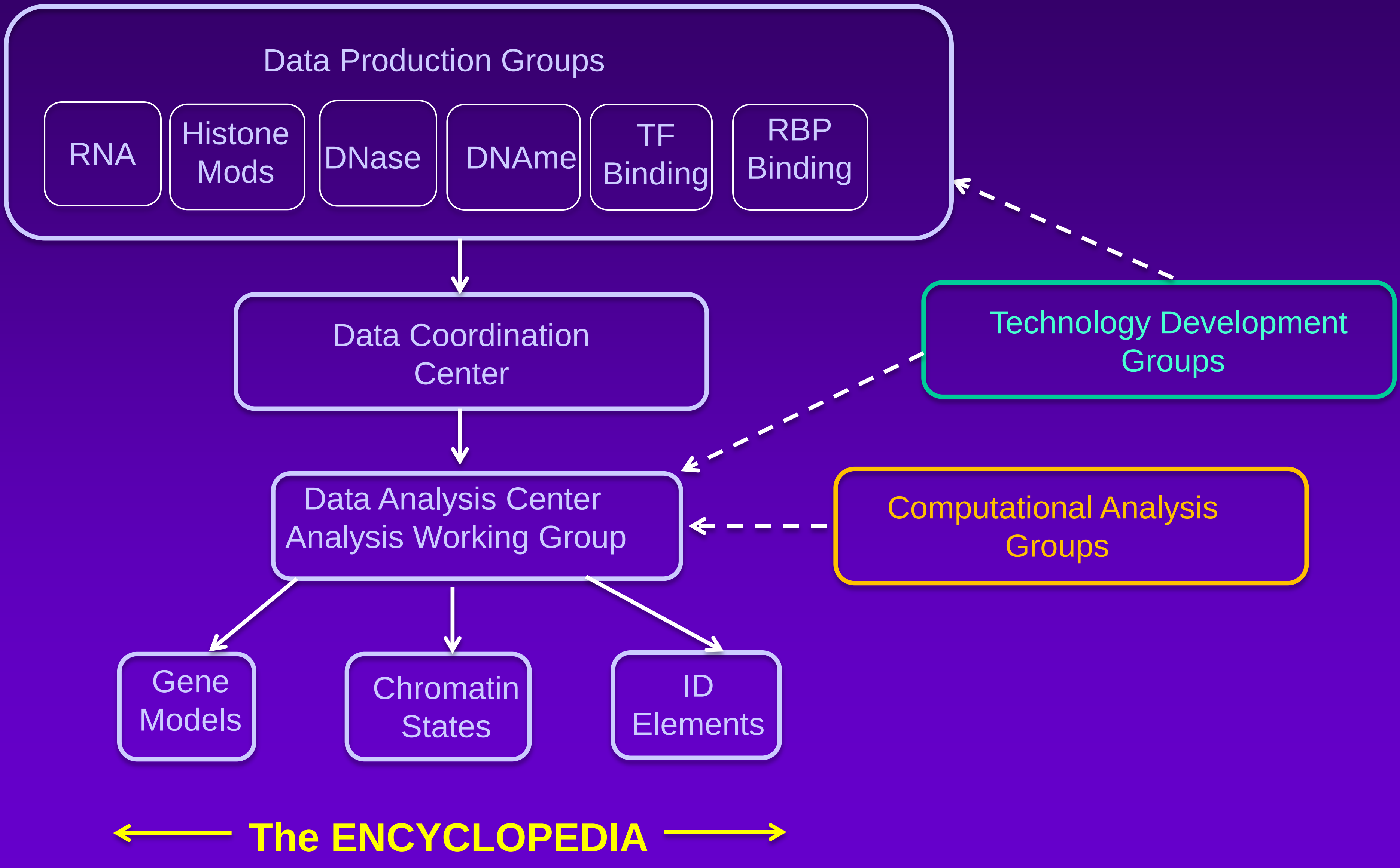
Boyle et al.
Comparative analysis of regulatory information and circuits across distant species
Nature 512:453–456, 2014.

Gerstein et al.
Comparative analysis of the transcriptome across distant species

Goals of ENCODE 3

- Expand towards completion the catalog of functional elements in human and mouse genomes
  - Additional cell types, more primary sources
  - Additional data types, e.g., RBPs
- Generate high quality data using high throughput pipelines
- Develop new technologies and analytical tools to generate, analyze and validate data
- Provide data and tools to the community in as useful form as possible
  - Provide easy and rapid access
  - Provide sufficient metadata
  - Describe ways to use it
  - Recognize needs of different types of users

ENCODE 3 Structure



Data Production Groups

RNA

Histone Mods

DNase

DNAme

TF Binding

RBP Binding

Data Coordination Center

Technology Development Groups

Data Analysis Center
Analysis Working Group

Computational Analysis Groups

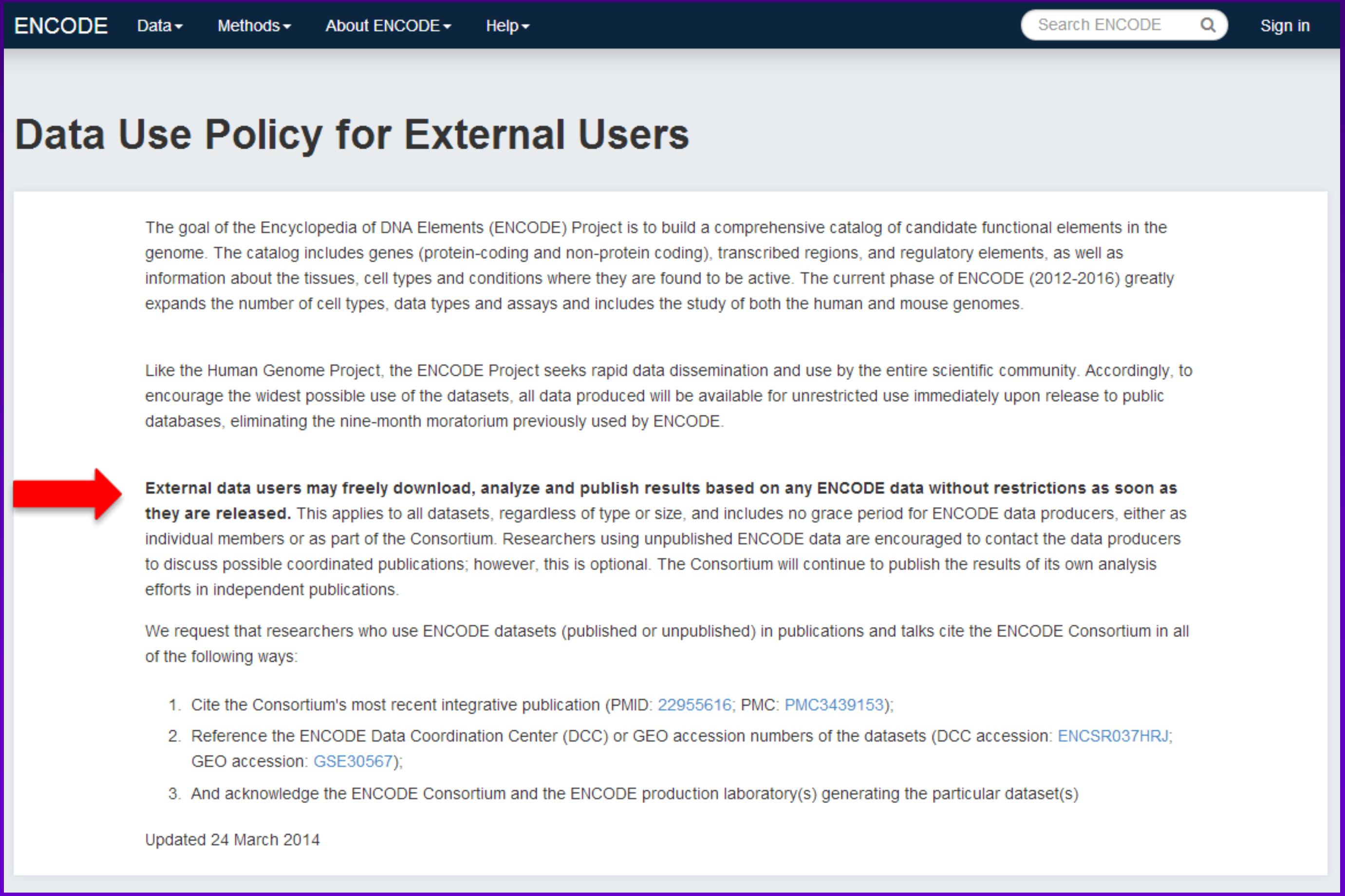
Gene Models

Chromatin States

ID Elements

The ENCYCLOPEDIA

Revised ENCODE Data Release Policy



Data Use Policy for External Users

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.

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)

Updated 24 March 2014

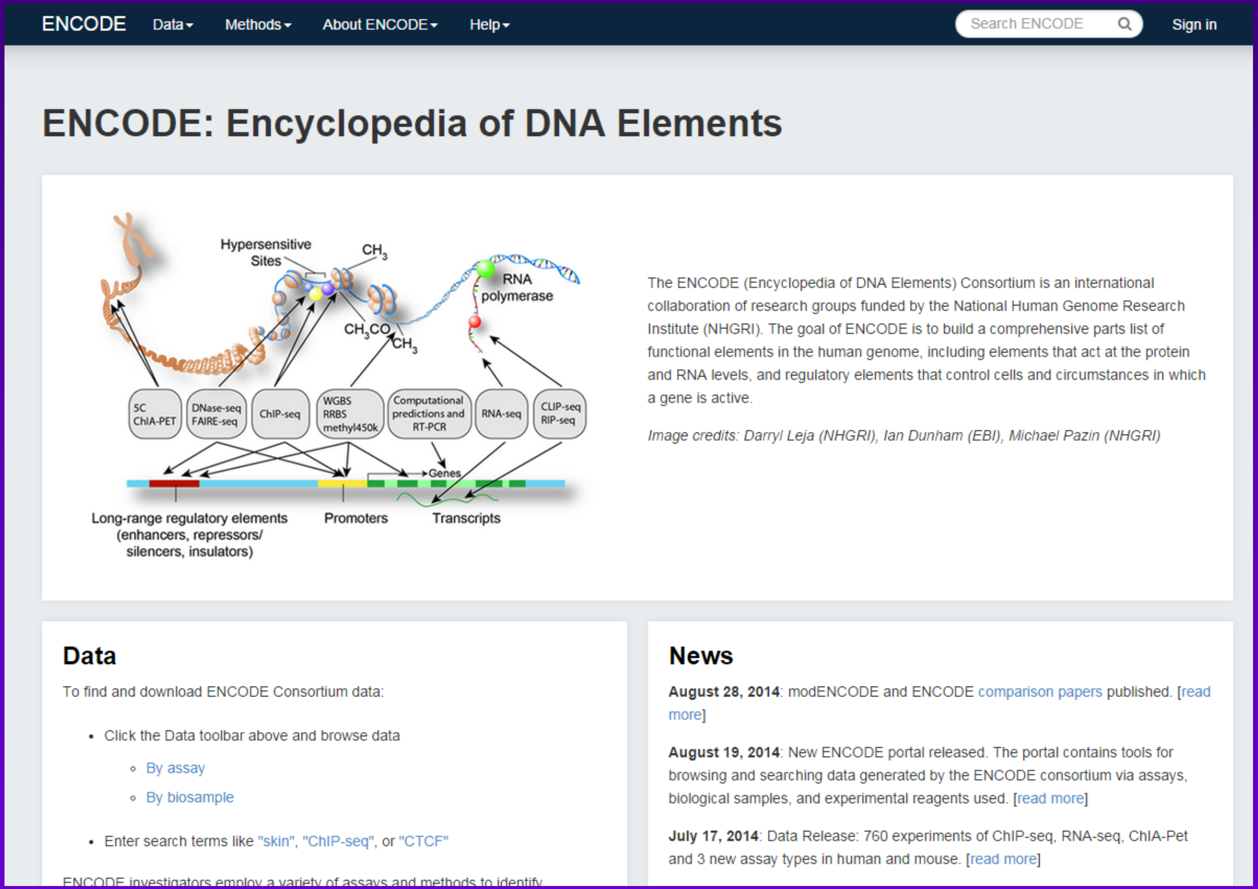
ENCODE 3 Features

- Cloud computing
  - Data available at Amazon Web Services (AWS)
  - Uniform processing pipelines will be available at DNAnexus
    - Provides transparency about ENCODE data processing
    - Available for broader community to use on own data

- Data Interoperability
  - Coordinating ontology selection and metadata standards with related projects

- Genomics Data Sharing Policy
  - ENCODE working ahead of new GDS policy
    - Developed sample consent language for open access to genomic data
    - Currently working to obtain wide range of samples using new consent language

New ENCODE Portal
https://www.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)

Data

To find and download ENCODE Consortium data:

- Click the Data toolbar above and browse data
    - By assay
    - By biosample
- Enter search terms like "skin", "ChIP-seq", or "CTCF"

ENCODE investigators employ a variety of assays and methods to identify

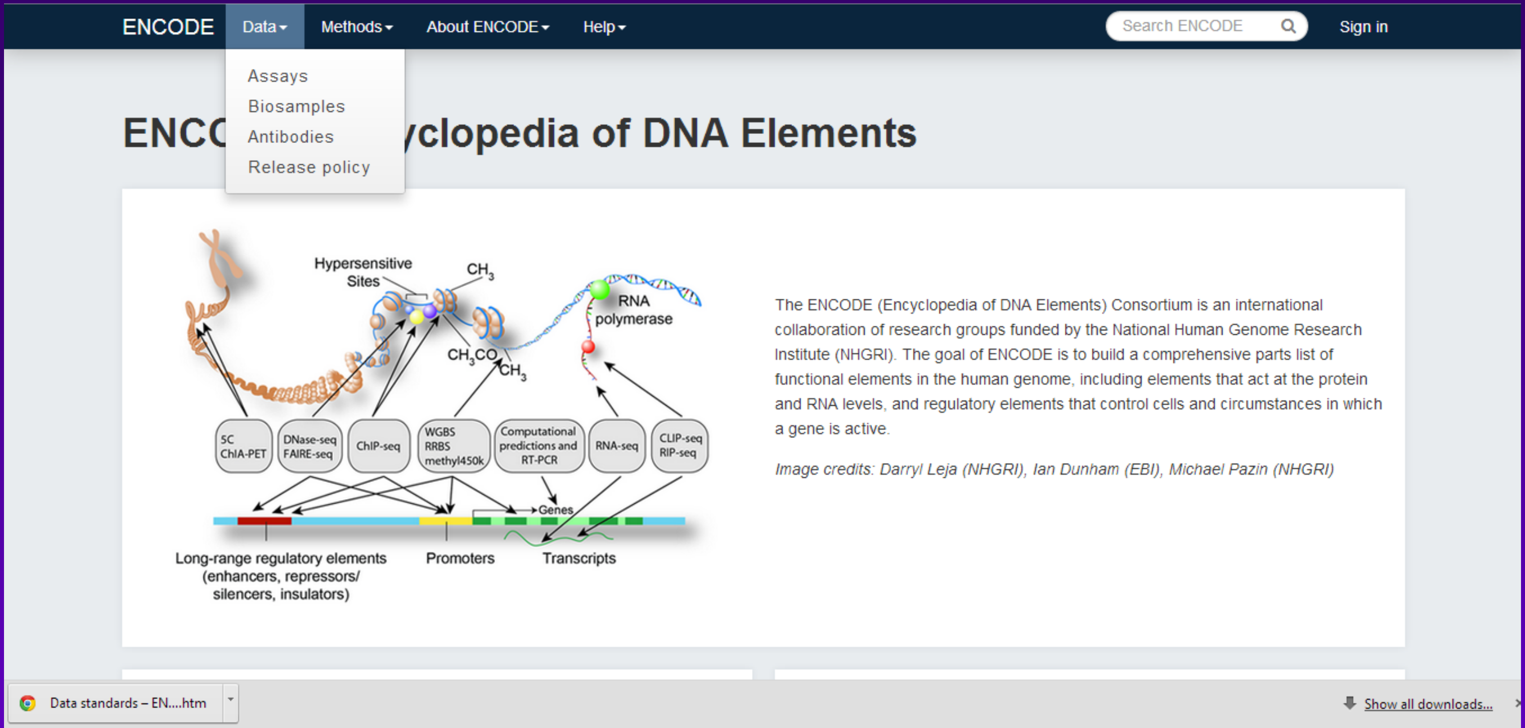
News

August 28, 2014: modENCODE and ENCODE comparison papers published. [read more]

August 19, 2014: New ENCODE portal released. The portal contains tools for browsing and searching data generated by the ENCODE consortium via assays, biological samples, and experimental reagents used. [read more]

July 17, 2014: Data Release: 760 experiments of ChIP-seq, RNA-seq, ChIA-Pet and 3 new assay types in human and mouse. [read more]

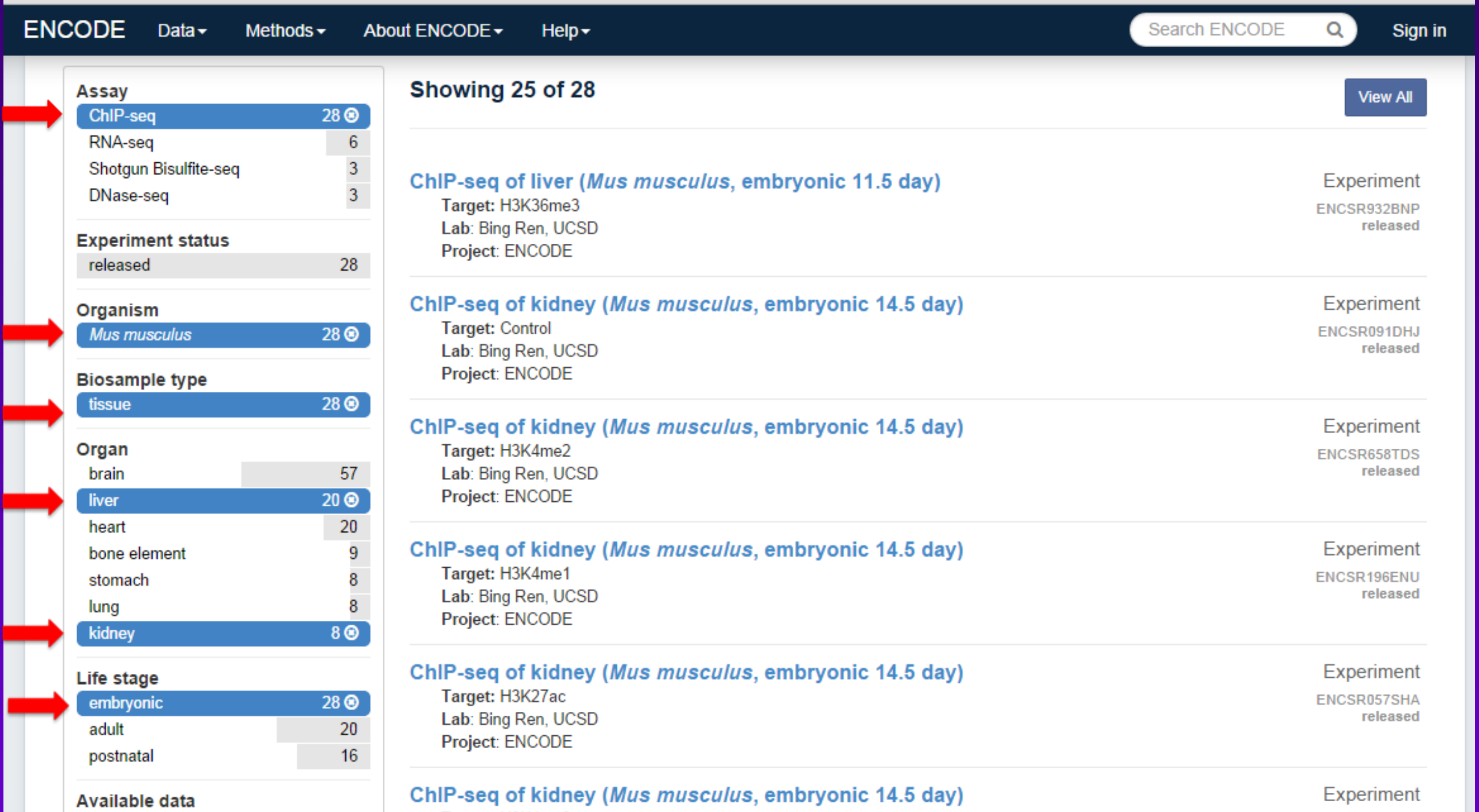
Data Access



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)

Metadata-driven searches



Showing 25 of 28

Assay

ChIP-seq	28
RNA-seq	6
Shotgun Bisulfite-seq	3
DNase-seq	3

Experiment status

released	28

Organism

Mus musculus	28

Biosample type

tissue	28

Organ

brain	57
liver	20
heart	20
bone element	9
stomach	8
lung	8
kidney	8

Life stage

embryonic	28
adult	20
postnatal	16

Available data

ChIP-seq of liver (*Mus musculus*, embryonic 11.5 day)
Target: H3K36me3
Lab: Bing Ren, UCSD
Project: ENCODE
Experiment
ENCSR932BNP
released

ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day)
Target: Control
Lab: Bing Ren, UCSD
Project: ENCODE
Experiment
ENCSR091DHJ
released

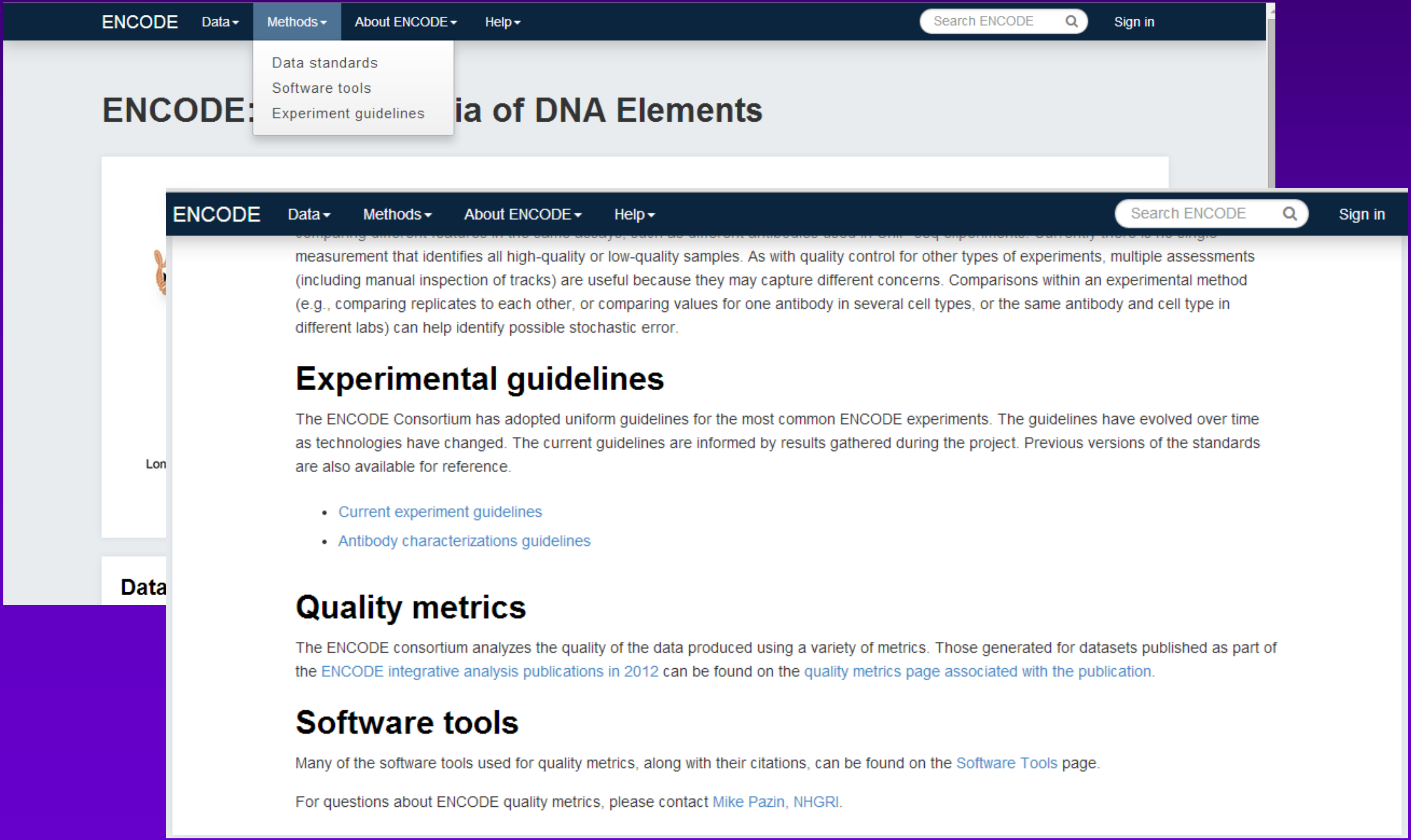
ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day)
Target: H3K4me2
Lab: Bing Ren, UCSD
Project: ENCODE
Experiment
ENCSR658TDS
released

ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day)
Target: H3K4me1
Lab: Bing Ren, UCSD
Project: ENCODE
Experiment
ENCSR196ENU
released

ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day)
Target: H3K27ac
Lab: Bing Ren, UCSD
Project: ENCODE
Experiment
ENCSR057SHA
released

ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day)
Target: H3K4me3
Experiment

Data Standards and Software Tools



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

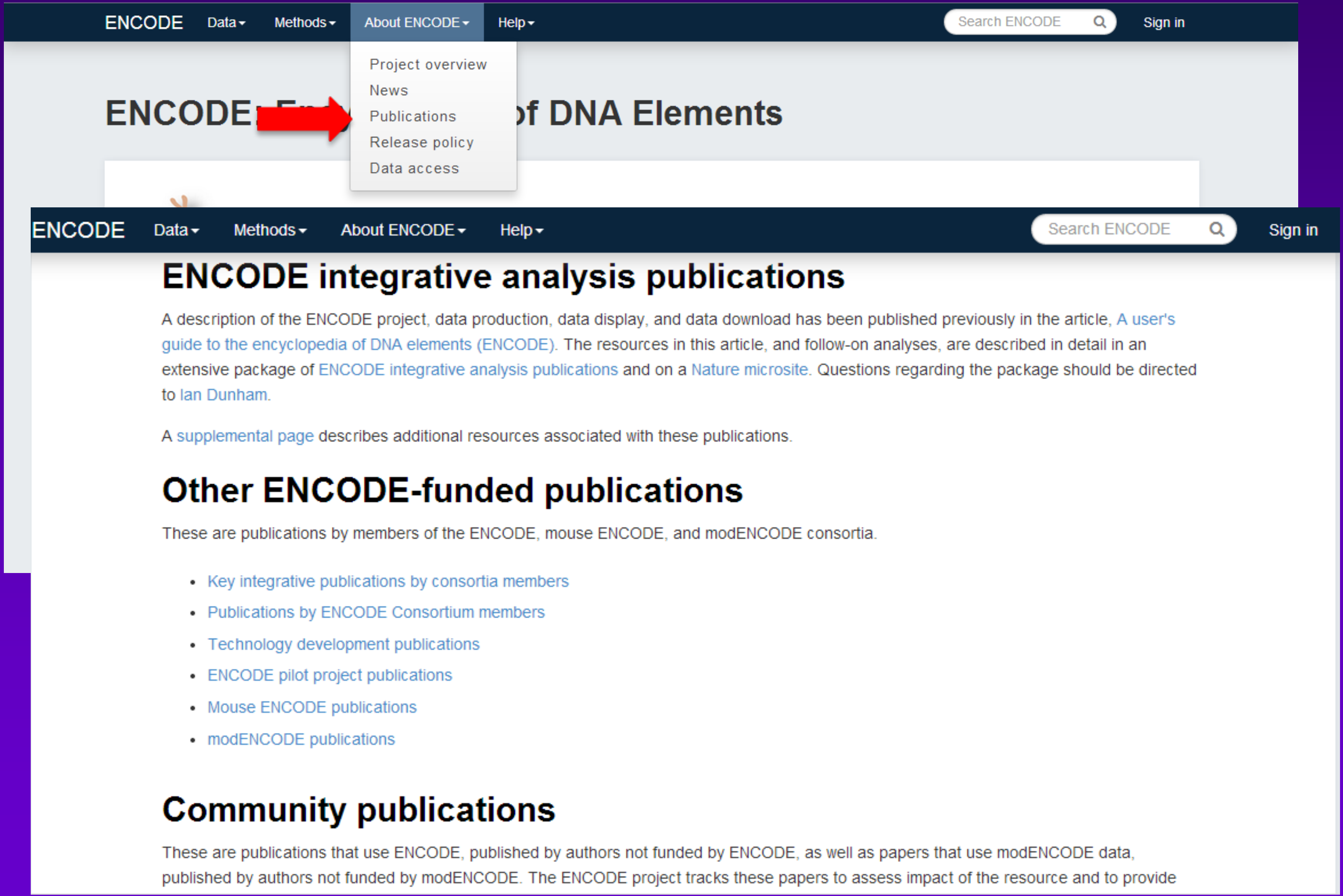
The ENCODE consortium analyzes the quality of the data produced using a variety of metrics. Those generated for datasets published as part of the ENCODE integrative analysis publications in 2012 can be found on the quality metrics page associated with the publication.

Software tools

Many of the software tools used for quality metrics, along with their citations, can be found on the Software Tools page.

For questions about ENCODE quality metrics, please contact Mike Pazin, NHGRI.

ENCODE Publications



ENCODE integrative analysis publications

A description of the ENCODE project, data production, data display, and data download has been published previously in the article, A user's guide to the encyclopedia of DNA elements (ENCODE). The resources in this article, and follow-on analyses, are described in detail in an extensive package of ENCODE integrative analysis publications and on a Nature microsite. Questions regarding the package should be directed to Ian Dunham.

A supplemental page describes additional resources associated with these publications.

Other ENCODE-funded publications

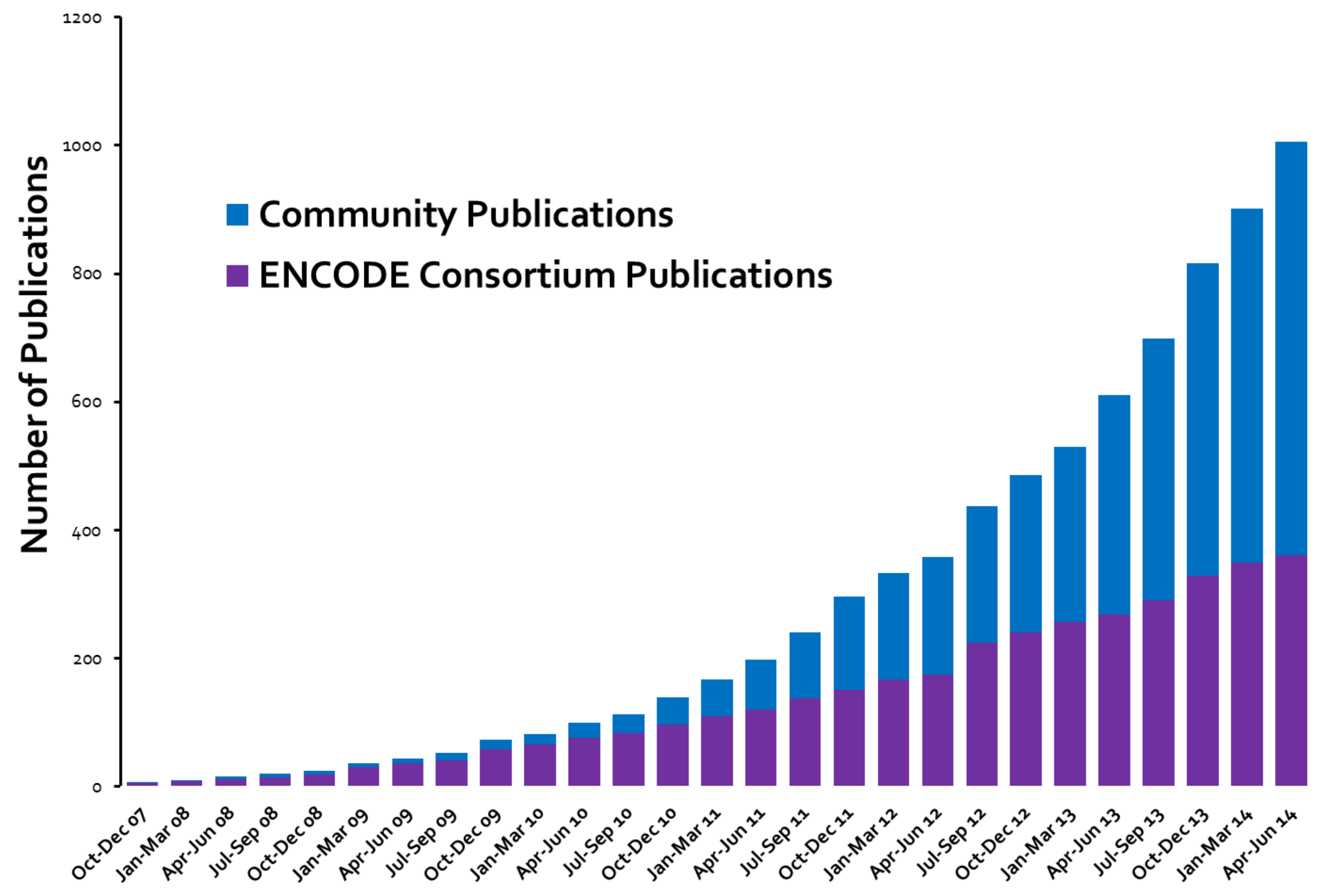
These are publications by members of the ENCODE, mouse ENCODE, and modENCODE consortia.

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

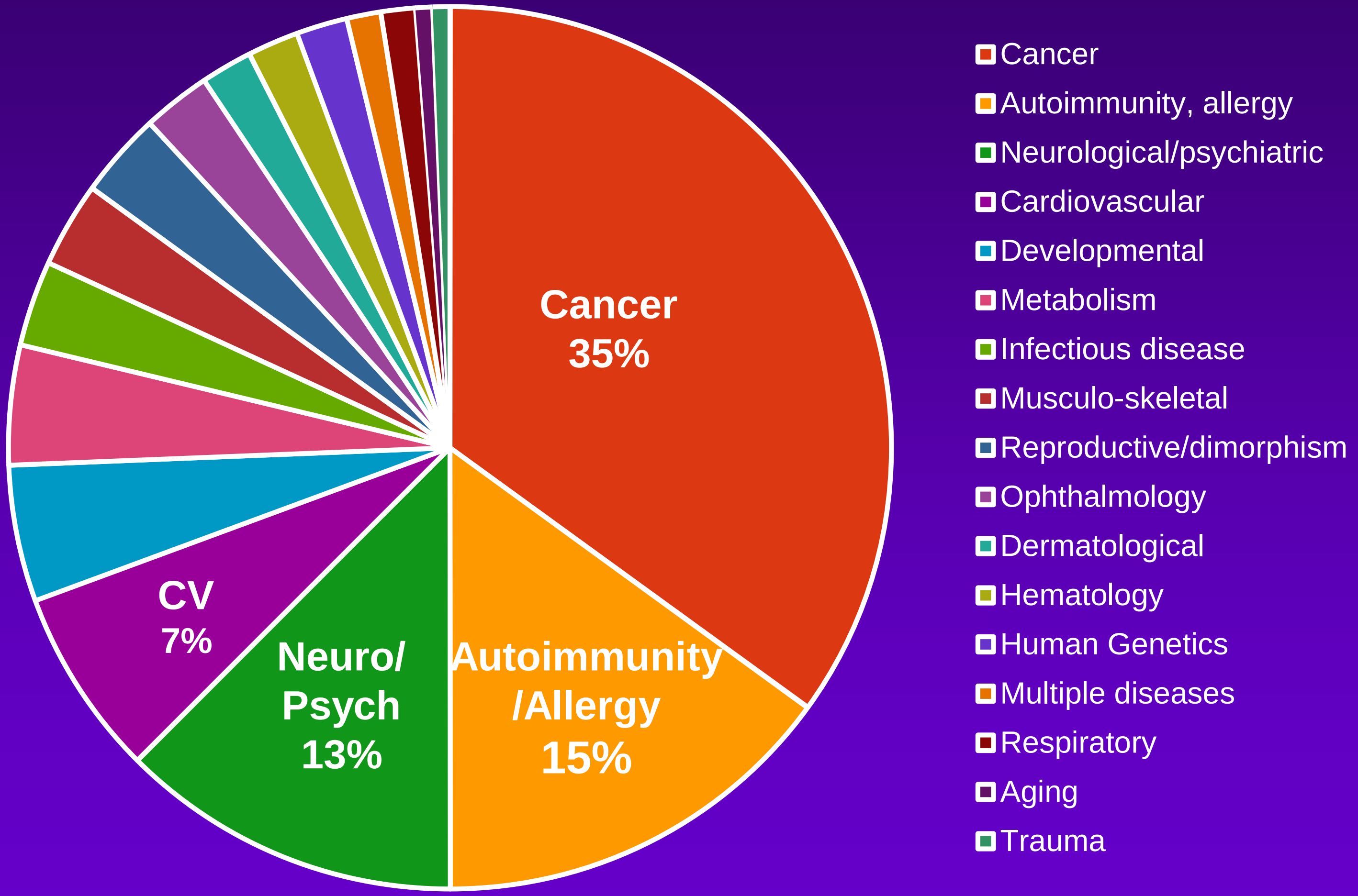
Community publications

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

ENCODE Publications

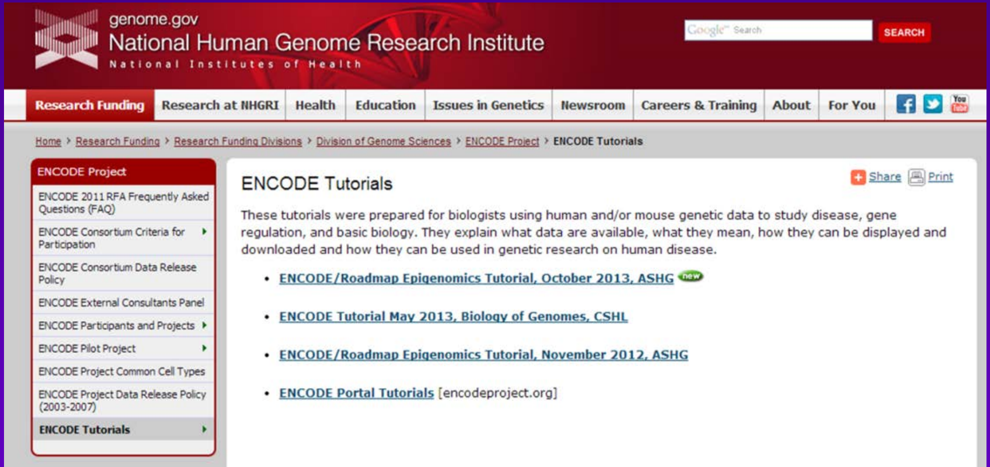


Categories of Disease-Related ENCODE Community Publications



Outreach Activities

- Tutorials (http://www.genome.gov/27553900)



ENCODE Tutorials

These tutorials were prepared for biologists using human and/or mouse genetic data to study disease, gene regulation, and basic biology. They explain what data are available, what they mean, how they can be displayed and downloaded and how they can be used in genetic research on human disease.

- ENCODE/Roadmap Epigenomics Tutorial, October 2013, ASHG
- ENCODE Tutorial May 2013, Biology of Genomes, CSHL
- ENCODE/Roadmap Epigenomics Tutorial, November 2012, ASHG
- ENCODE Portal Tutorials [encodeproject.org]

- CHARGE-ENCODE workshop
- User's meeting in 2015

Social Media

Facebook

ENCODE (ENCyclopedia Of DNA Elements)

ENCODE (ENCyclopedia Of DNA Elements)

63 likes · 13 talking about this · 0 were here

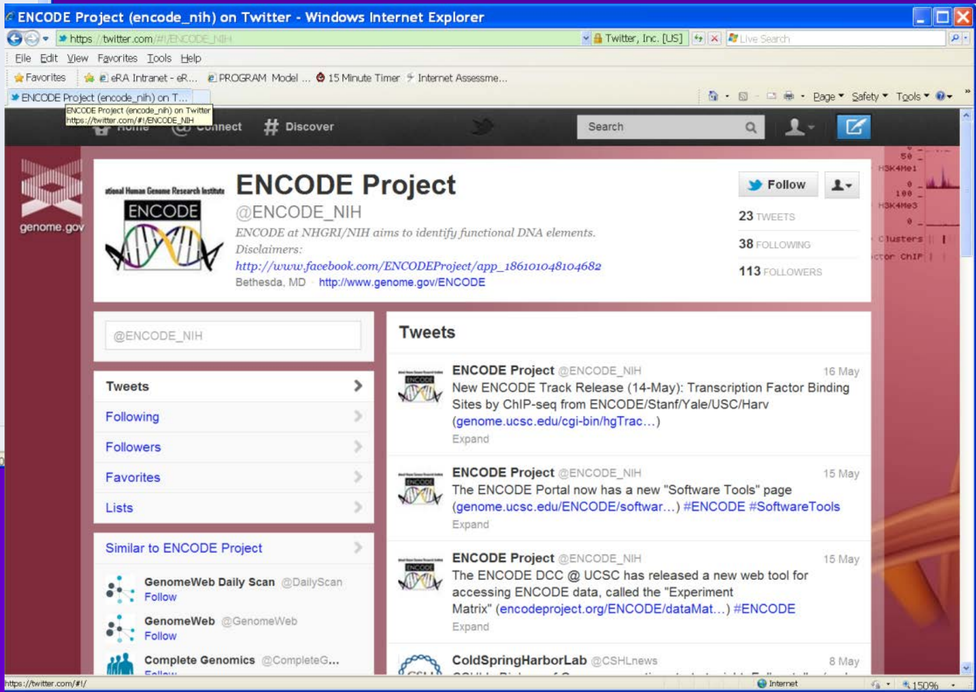
Government Organization
ENCyclopedia Of DNA Elements (ENCODE) Project

About – Suggest an edit

63

Twitter

@ENCODE_NIH



ENCODE Project
@ENCODE_NIH
ENCODE at NHGRI/NIH aims to identify functional DNA elements. Disclaimers: http://www.facebook.com/ENCODEProject/app_186101048104682
Bethesda, MD · http://www.genome.gov/ENCODE

23 TWEETS
38 FOLLOWING
113 FOLLOWERS

Tweets

ENCODE Project @ENCODE_NIH 16 May
New ENCODE Track Release (14-May): Transcription Factor Binding Sites by ChIP-seq from ENCODE/Stanf/Yale/USC/Harv (genome.ucsc.edu/cgi-bin/hgTrac…)

ENCODE Project @ENCODE_NIH 15 May
The ENCODE Portal now has a new "Software Tools" page (genome.ucsc.edu/ENCODE/softwar…) #ENCODE #SoftwareTools

ENCODE Project @ENCODE_NIH 15 May
The ENCODE DCC @ UCSC has released a new web tool for accessing ENCODE data, called the "Experiment Matrix" (encodeproject.org/ENCODE/dataMat…) #ENCODE

ColdSpringHarborLab @CSHLnews 8 May

GenomeWeb Daily Scan @DailyScan Follow
GenomeWeb @GenomeWeb Follow
Complete Genomics @CompleteG…

Future of ENCODE

- Current project period 2012-2016
- Planning process in 2015
  - Planning Workshop
    - Relevant discussion at July Sequencing Meeting
  - Council Concept Clearance(s) in May 2015
  - Potential RFA(s) to be released Summer 2015
- Review and Funding 2016

The ENCODE 3 Consortium

http://www.genome.gov/26525220

NHGRI Staff

Program Directors

Elise Feingold
Dan Gilchrist
Peter Good
Michael Pazin

Program Analysts

Julie Coursen
Hannah Naughton

Division Director

Jeff Schloss