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
Sabeti, Cell 152:703, 2013
Fraser, Genome Research 23:1089, 2013
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

 Science 306:636, 2004
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/
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
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: 22955516; PMC: PMC3430153);
2. Reference the ENCODE Data Coordination Center (DCC) or GEO accession numbers of the datasets (DCC accession: ENCSR037HRJ, GEO accession: GSE30567);
3. 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
Metadata-driven searches

[ENCODE website screenshot showing search results for ChIP-seq experiments involving liver and kidney samples in Mus musculus, embryonic 11.5 day development.]
Data Standards and Software Tools

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

Number of Publications

- Community Publications
- ENCODE Consortium Publications
Categories of Disease-Related ENCODE Community Publications

- Cancer: 35%
- Autoimmunity, allergy: 15%
- Neurological/psychiatric: 13%
- Cardiovascular: 7%
- Developmental
- Metabolism
- Infectious disease
- Musculo-skeletal
- Reproductive/dimorphism
- Ophthalmology
- Dermatological
- Hematology
- Human Genetics
- Multiple diseases
- Respiratory
- Aging
- Trauma
Outreach Activities

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

• CHARGE-ENCODE workshop

• User’s meeting in 2015
Social Media

Facebook

ENCODE (ENCyclopedia Of DNA Elements)

Twitter

@ENCODE_NIH
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