ENCODE Project Update

Brenton R. Graveley, Ph.D.

Professor and Chair, Department of Genetics and Genome Sciences

Associate Director, Institute for Systems Genomics

UConn and UConn Health

NHGRI Council Meeting: May 20, 2019

Understanding the Human Genome @ 2003

- Limited ability to decipher genomic sequence

- Evolutionary conservation helps to identify functionally important regions

    ~5% conserved/ ~1.5% protein coding

- Using the genetic code - moderately good at identifying protein-coding regions, but fine structures difficult to predict from sequence

- No similar "regulatory code" for non-coding DNA
    - Regulatory (functional DNA) elements can be very far away from genes so hard to predict gene target

Importance of Non-coding DNA

Carries out regulatory functions of genome and is important for disease and gene regulation:

- Genes are regulated throughout time (e.g., development) and space (e.g., different tissues), health and disease, and in response to environmental exposures

- Many regulatory elements are only active in specific cells/tissues

- Majority of common disease-associated genetic variants lie outside of protein-coding regions

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

- RNA regulatory elements are encoded in the genome but only act after being transcribed.

ENCODE: Encyclopedia of DNA Elements

- Catalog all functional elements in the human and select model organism genomes
- Develop freely available resource for research community to enhance understanding of
    - Regulation of gene expression
    - Genetic basis of disease
- Project components:
    - Data generation
    - Data analysis
    - Data repository
    - Outreach

ENCODE Timeline

2003-2007	2008-2012	2013-2016	2017-2021
Pilot Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
Pilot Phase (1%) Data Production (Human Only)	**Data Production (DCC & DAC)** (Human Only)	**Data Production (DCC & DAC)** (Human & Mouse)	**Data Production (DCC & DAC)** (Human & Mouse)
	modENCODE (Fly and Worm)	Computational Analysis 1	Computational Analysis 2
	Mouse ENCODE		Functional Characterization
Technology Development 1 & 2	Technology Development 3	Technology Development 4	
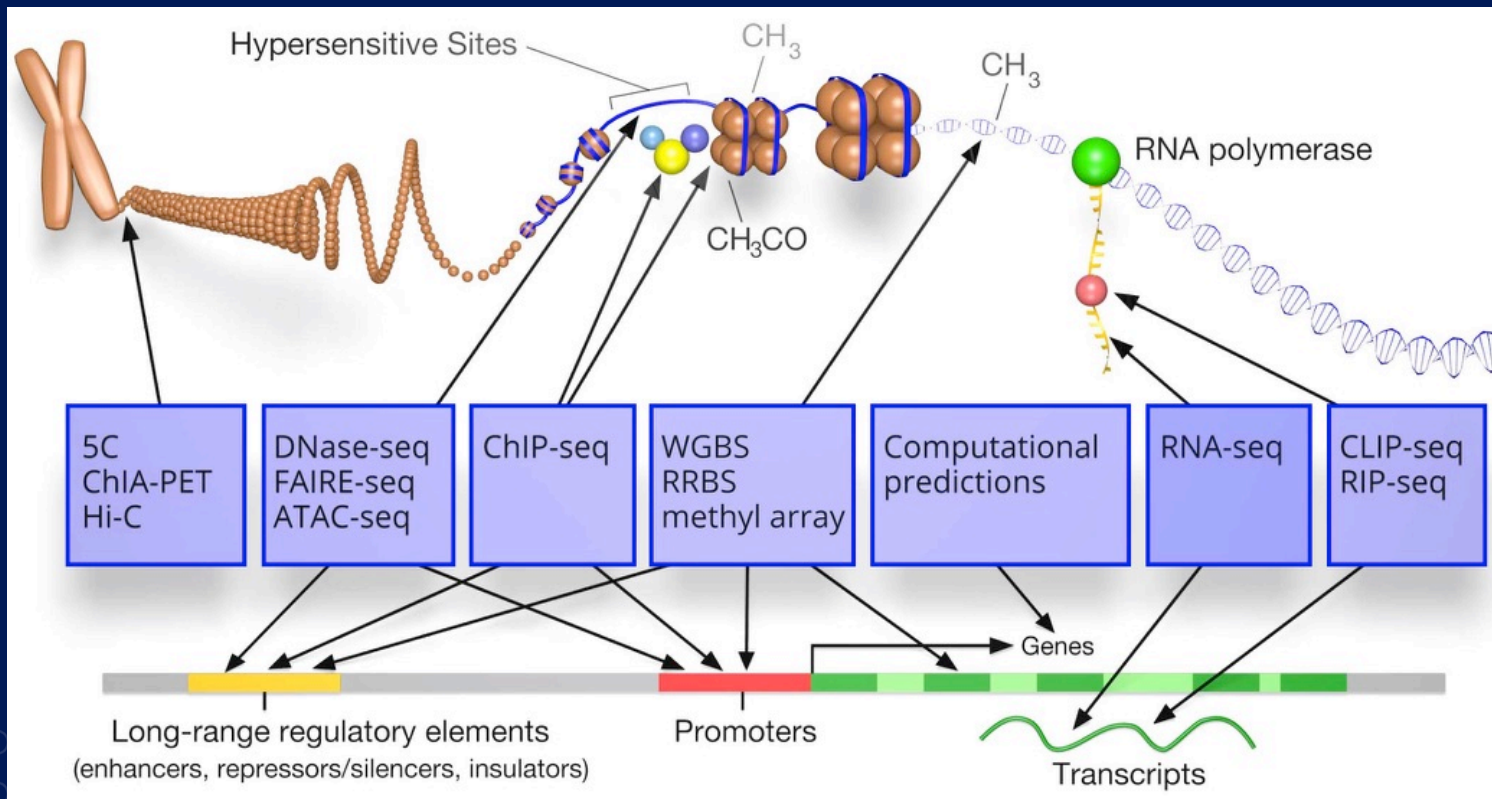
ENCODE Assays
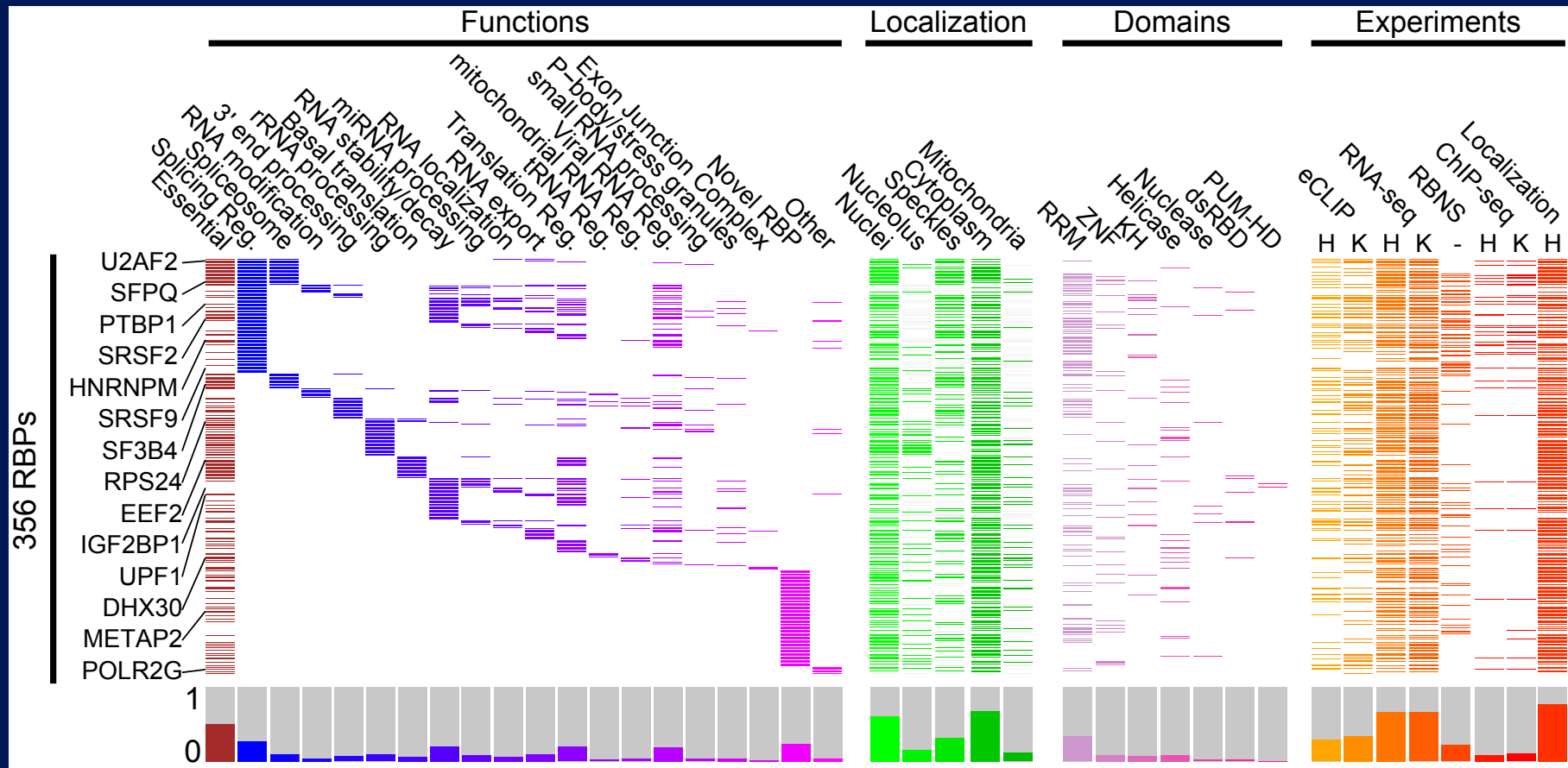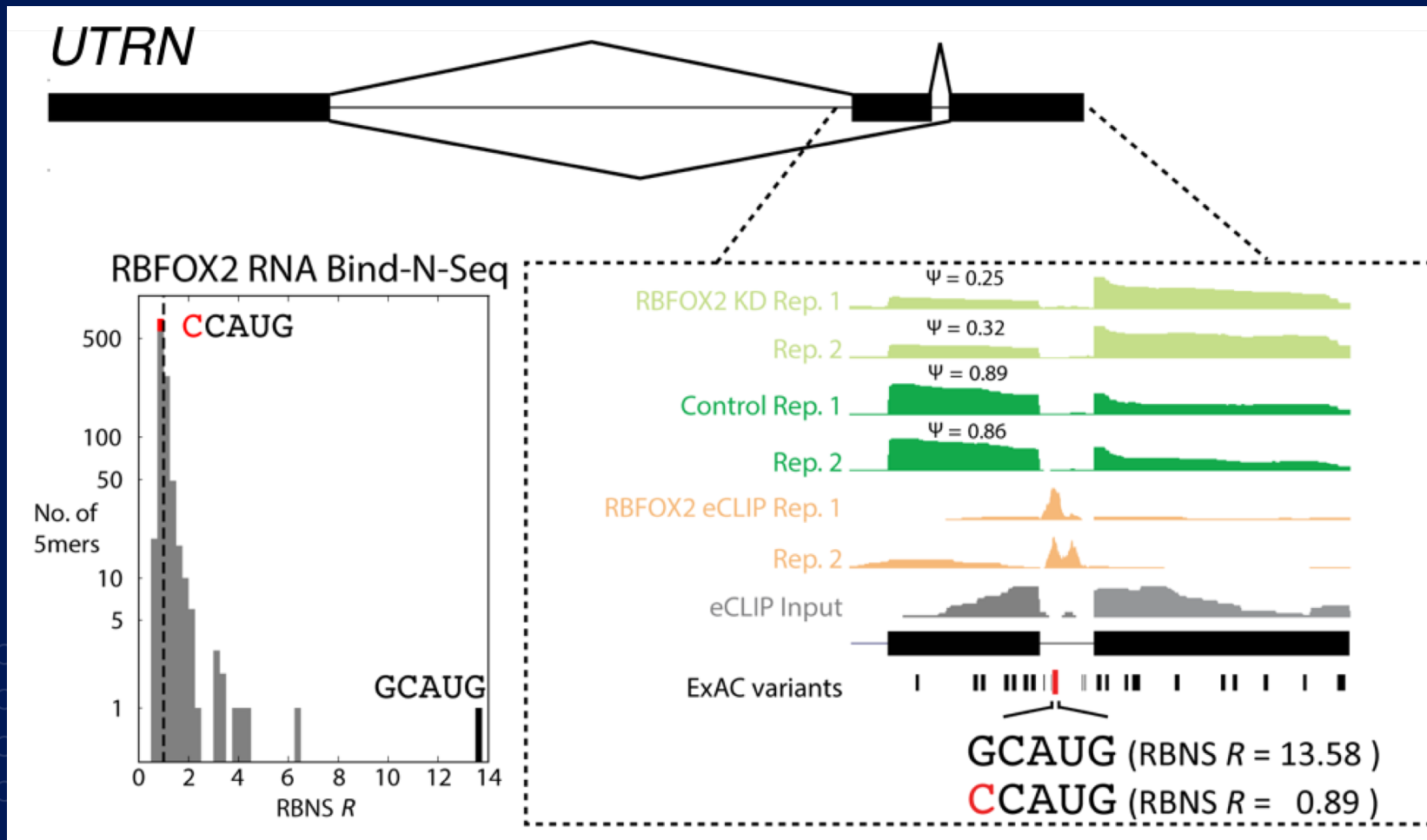
ENCODE Data Generation
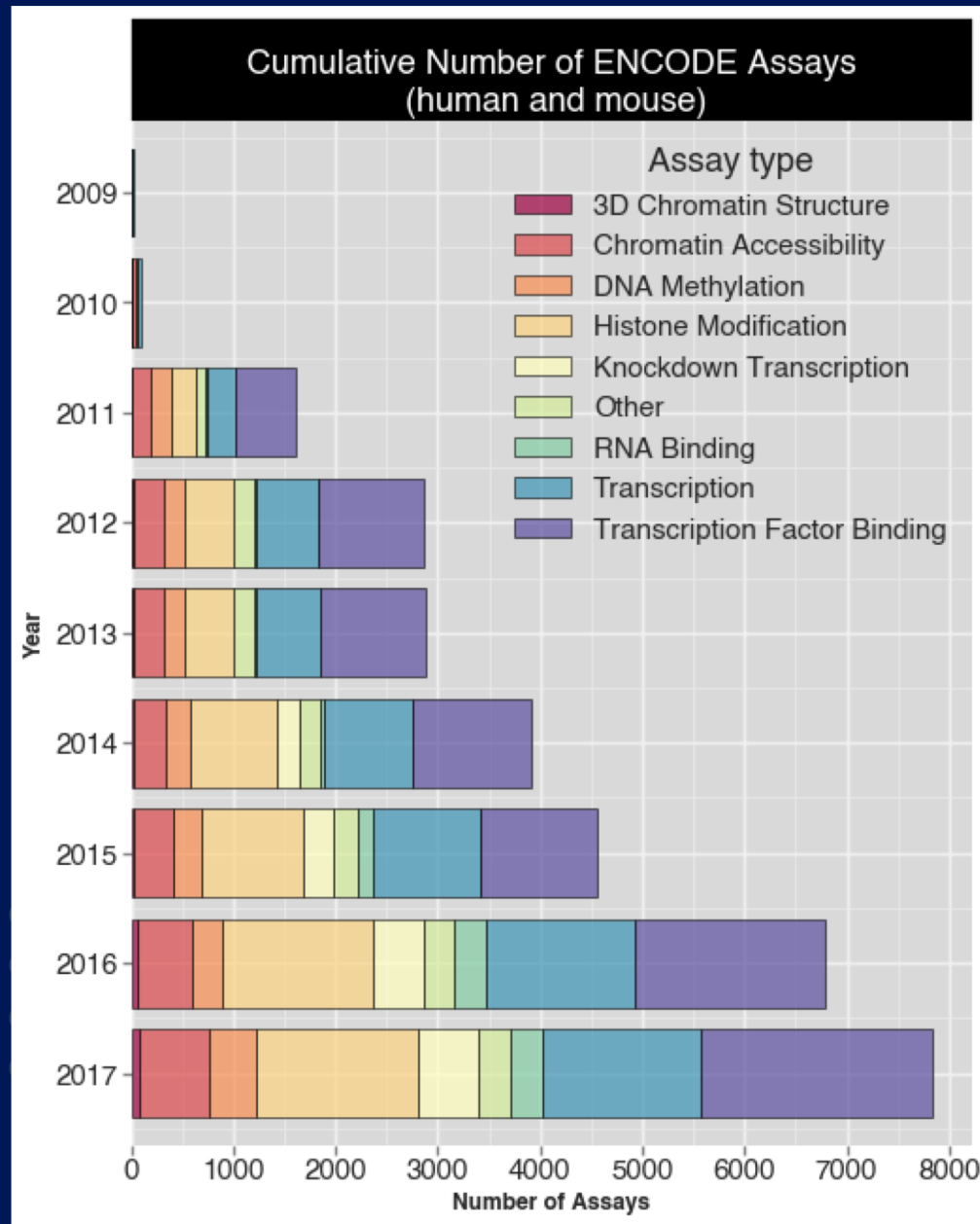
New ENCODE 3 Data Type: RBPs

1,223 Replicated Datasets – 356 RBPs

Integrating ENCODE RBP Data with Genomic Variants

The ENCODE Encyclopedia

https://www.encodeproject.org/data/annotations/
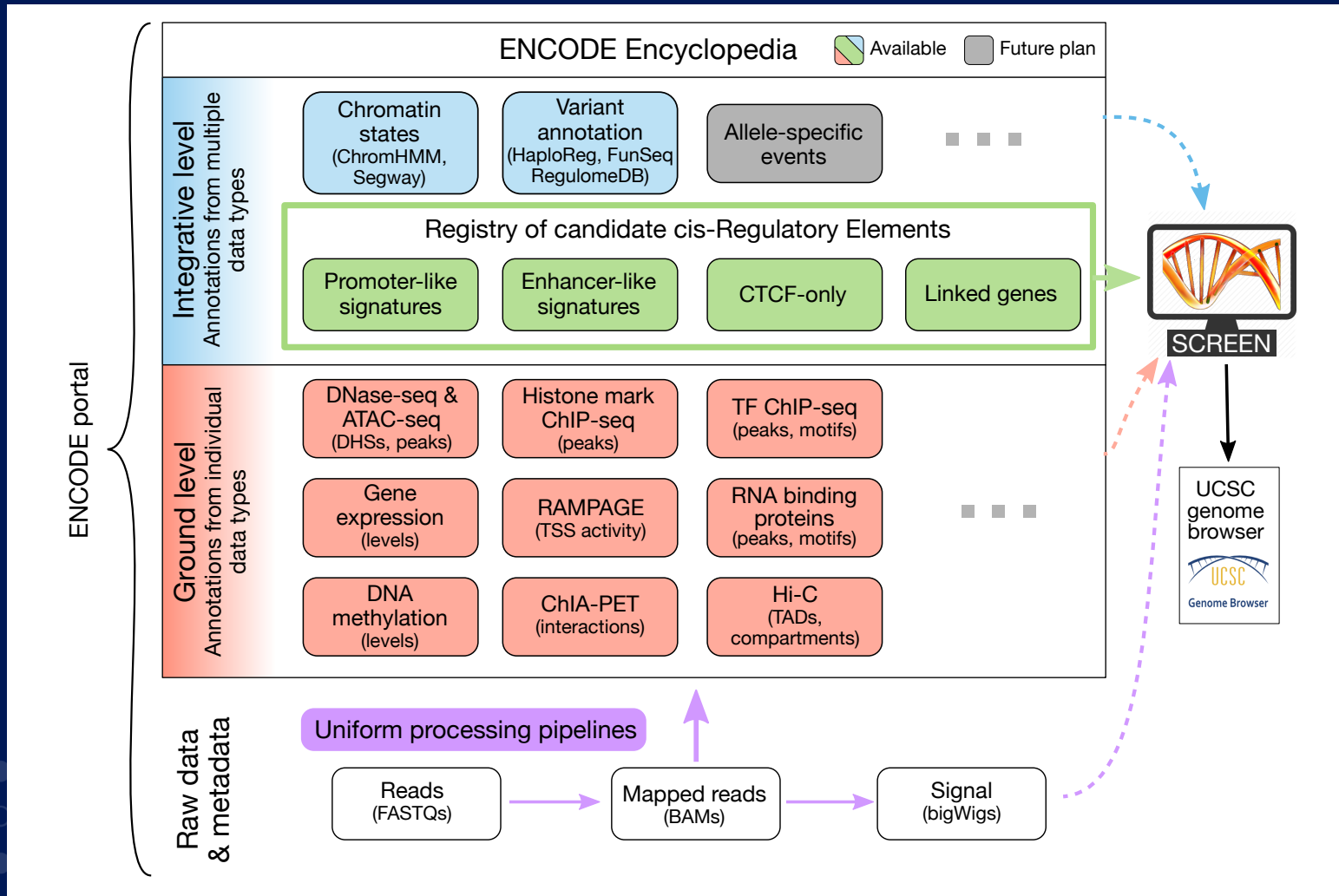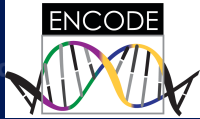
ENCODE 4 Participants

Mapping Center PIs

Brad Bernstein	Yijun Ruan
Erez Lieberman-Aiden	Mike Snyder
Mats Ljungman	John Stamatoyannopoulos
Rick Myers Eric Mendenhall	Barbara Wold Ali Mortazavi

Functional Characterization PIs

Nadav Ahituv Jay Shendure	Pardis Sabeti
Will Greenleaf Mike Bassik	Yin Shen Bing Ren
John Lis Haiyuan Yu	Tim Reddy Maria Ciofani Greg Crawford Charlie Gersbach
Len Pennacchio Axel Visel	Kevin White

Computational Analysis PIs

Mike Beer	Johnathan Pritchard
Christina Leslie	Ting Wang Barak Cohen Cedric Feschotte
Alkes Price Soumya Raychaudhury	Grace Xiao

Data Analysis Center

Zhiping Weng	Mark Gerstein

Data Coordination Center

Mike Cherry

ENCODE 4 Participants

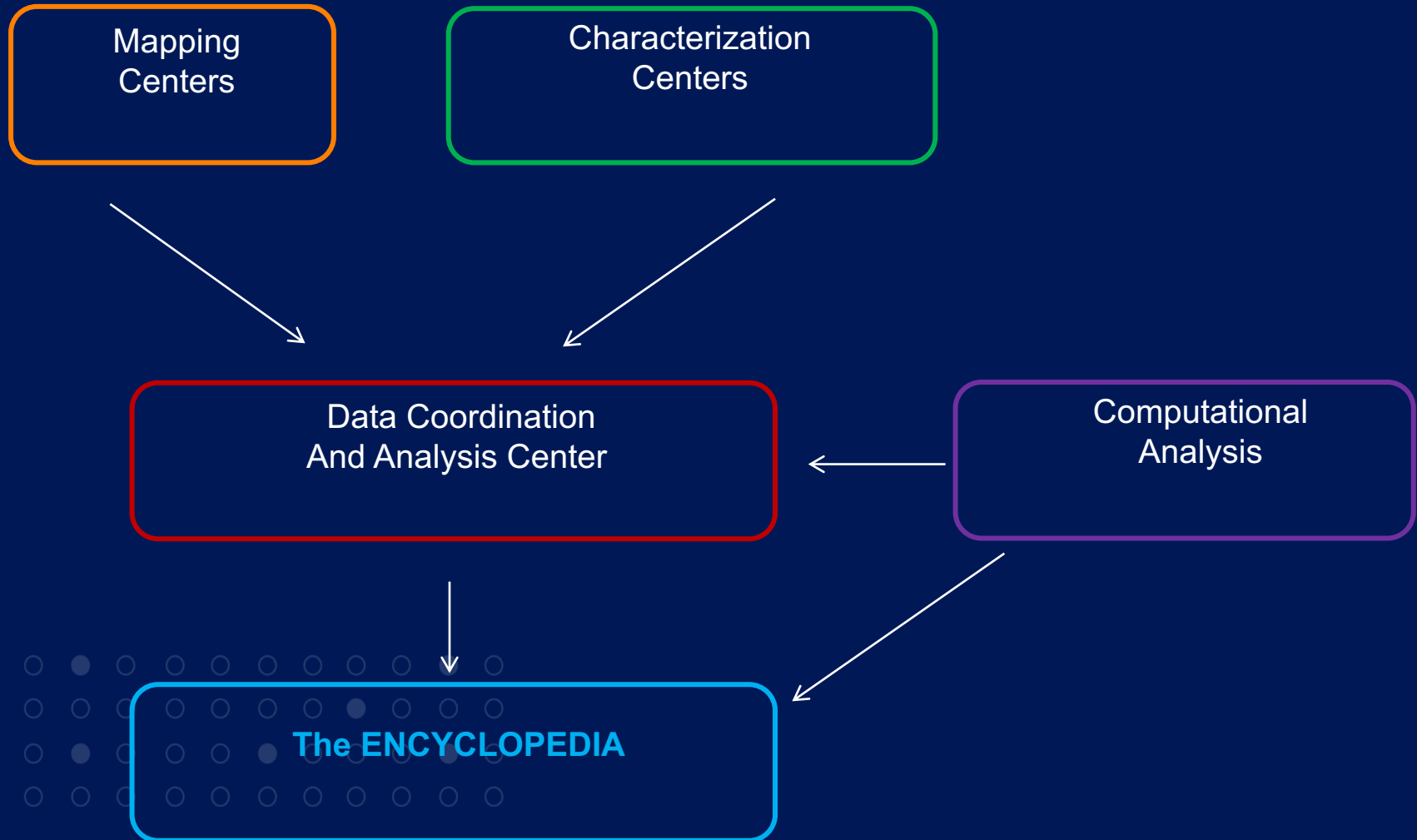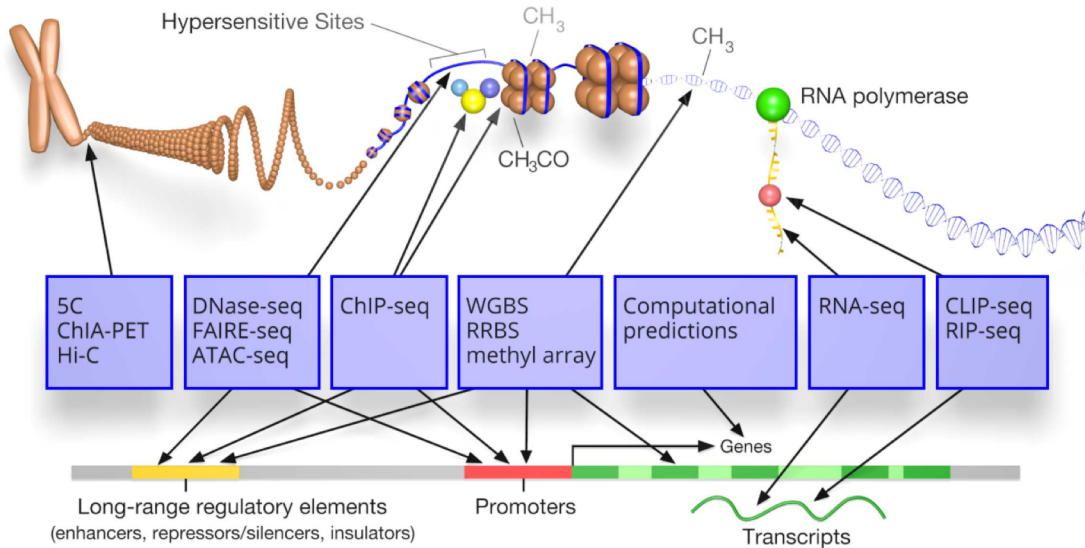| Affiliate Members |
|---|
| Alex Dobin |
| Jesse Engreitz |
| Melissa Fullwood |
| Brent Graveley Gene Yeo |
| Doug Phanstiel |
| Nir Yosef |
| Guo-Cheng Yuan |
| Feng Yue |

ENCODE 4 Organization

Community Resource

- Rapid pre-publication data release
- Unrestricted access sharing of uniformly processed data, software tools and analysis pipelines
- Analysis requires development of:
  - Common data reporting formats
  - Rich, ontology-based metadata
  - Data standards
  - Uniform Processing Pipelines
  - Analytical tools
- Consortium publications
- Outreach (Online Tutorials, Users Meetings)

ENCODE Data Portal

https://www.encodeproject.org

ENCODE Data Portal

https://www.encodeproject.org

SCREEN

Publications Using ENCODE Data

https://www.encodeproject.org/publications/
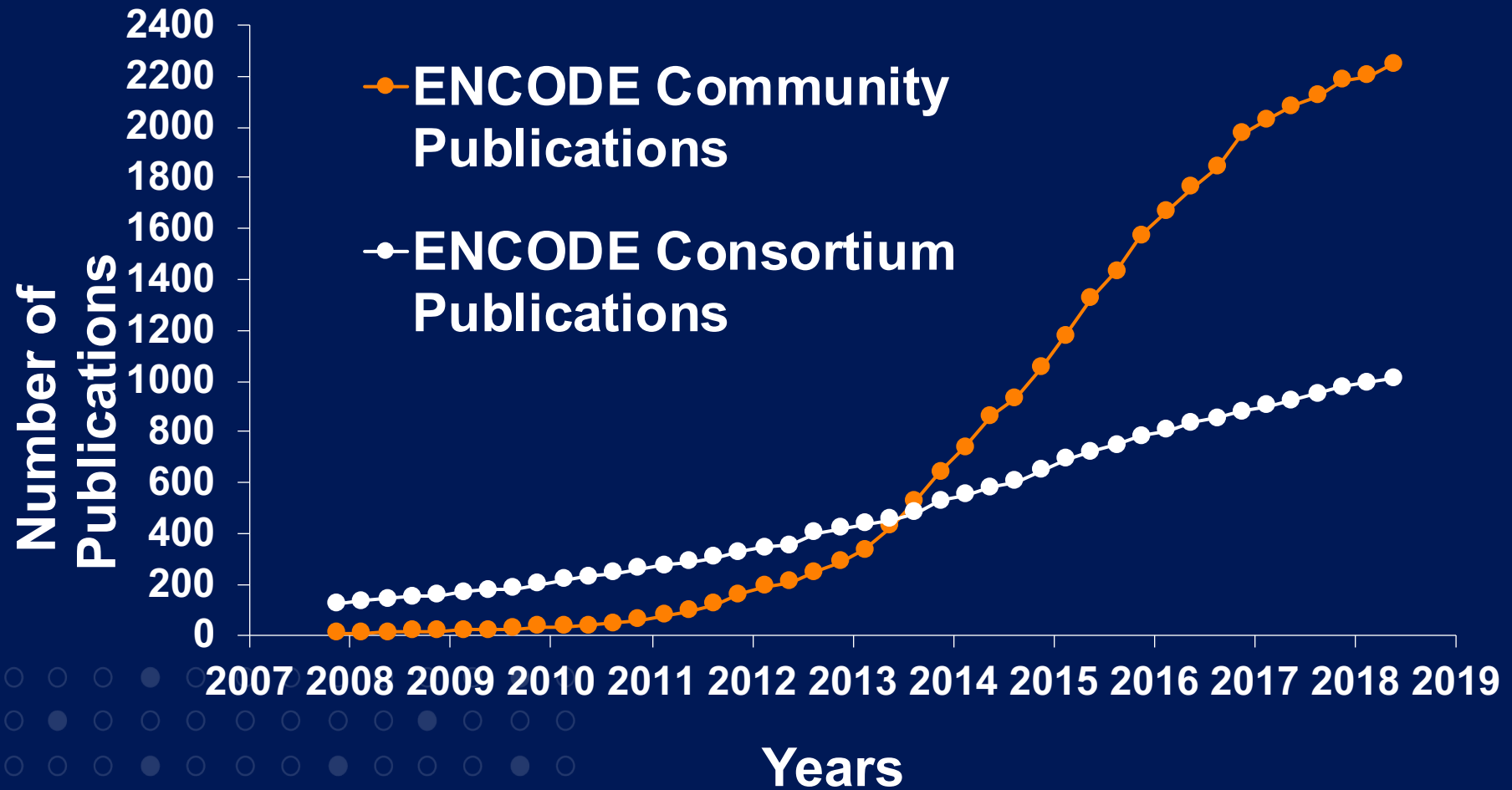
ENCODE Community Publications, By Disease Category

Cancer, 36%

Autoimmunity, Allergy, 12%

Human Genetics, 12%

Neurological/ Psychiatric, 10%

Cardiovascular, 7%
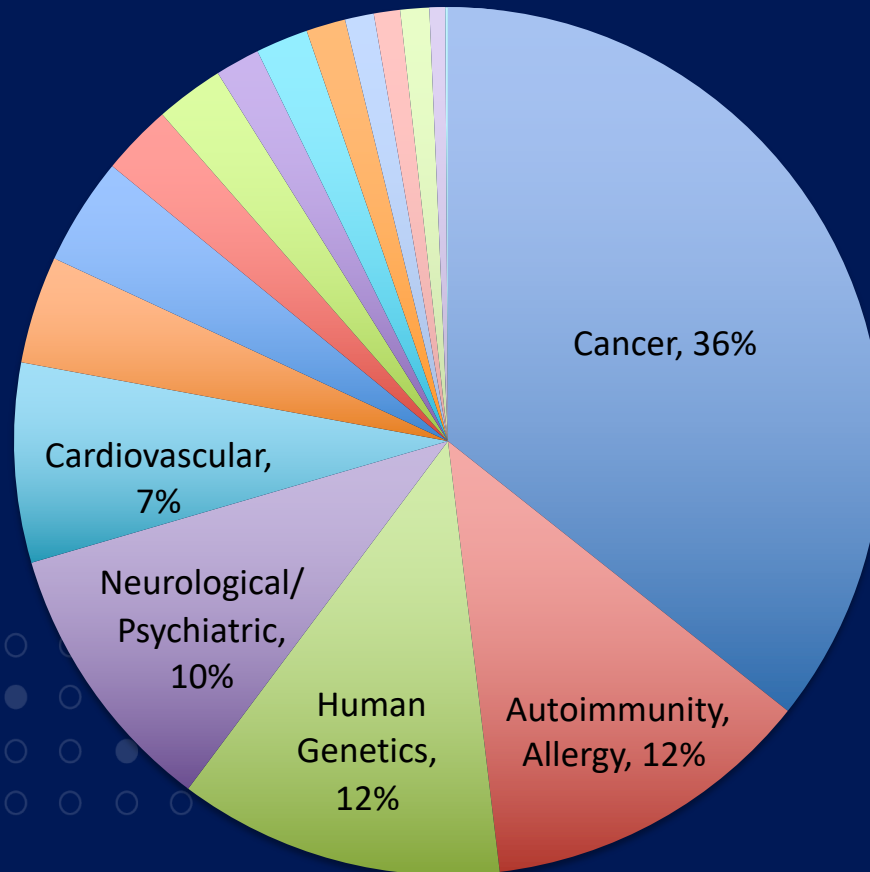
ENCODE Outreach

ENCODE Tutorials

These tutorials were prepared for biologists using human and/or mouse gene biology. They explain what data are available, what they mean, how they can in genetic research on human disease.

- **ENCODE: Getting Started** 📄 New

- **2016 Society of Toxicology Annual Meeting Workshop**

- **ENCODE/Roadmap Epigenomics Tutorial, October 2015, ASHG**

- **ENCODE 2015: Research Applications and Users Meeting**

- **ENCODE Workshop, April 2015, Keystone Symposia**

Workshop materials

- NEW: Uniform Processing Pipelines/DNAnexus Tutorial from the ASHG 2016 Workshop in Vancouver, BC O
- Presentations, tutorials, and video from the ENCODE Users Meeting 2016 in Palo Alto, CA June 8 - 10, 2016
  - Video and workshop materials from hands-on tutorial sessions on accessing, processing, analyzing, and ut resources, along with presentations from leading experts in disease, biology, and computational fields expl ENCODE resources in their work.
  - Pictures by Forrest Tanaka can be viewed using #EUM16 on twitter
- Keystone Symposium 2016, Chromatin and Epigenomics, Whistler, British Columbia, March 21, 2016.
  - J. Seth Strattan, Find and Visualize ENCODE Data
- Asia Pacific Bioinformatics Conference, San Francisco, January 10, 2016.
  - J. Michael Cherry, Introduction to ENCODE and the ENCODE DCC
  - Aditi Narayanan, The ENCODE Portal: Searching for Metadata and Data
  - J. Seth Strattan, ENCODE Data Availability and Standardized Processing

https://www.genome.gov/27553900/encode-tutorials/   https://www.encodeproject.org/tutorials/

ENCODE Users Meeting

ENCODE 2019: Research Applications & Users Meeting

📅 Jul 8, 2019 8:00 AM - Jul 10, 2019 1:00 PM

📍 Bell Harbor International Conference Center
2211 Alaskan Way, Seattle, Washington 98121, United States

Registration Fee   **$175.00**

BUY TICKETS

ENCODE Data on AWS

Registry of Open Data on AWS

Encyclopedia of DNA Elements (ENCODE)

`biology` `bioinformatics` `genetic` `genomic` `life sciences` `machine learning` `deep learning`

Description

The Encyclopedia of DNA Elements (ENCODE) 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. ENCODE investigators employ a variety of assays and methods to identify functional elements. The discovery and annotation of gene elements is accomplished primarily by sequencing a diverse range of RNA sources, comparative genomics, integrative bioinformatic methods, and human curation. Regulatory elements are typically investigated through DNA hypersensitivity assays, assays of DNA methylation, and immunoprecipitation (IP) of proteins that interact with DNA and RNA, i.e., modified histones, transcription factors, chromatin regulators, and RNA-binding proteins, followed by sequencing.

Update Frequency

Daily

License

External data users may freely download, analyze, and publish results based on any ENCODE data without restrictions.

Documentation

https://www.encodeproject.org

Managed By

ENCODE Data Coordinating Center

See all datasets managed by ENCODE Data Coordinating Center.

Contact

encode-help@lists.stanford.edu

Usage Examples

- ENCODE CTCF ChIP-seq data correlation across different cell types by Paul Sud
- Exploring ENCODE data from EC2 with Jupyter notebook by Keenan Graham
- Ingesting ENCODE data into TileDB with S3 backend by Otto Jolanki

Resources on AWS

Description
Released and archived ENCODE data

Resource type
S3 Bucket

Amazon Resource Name (ARN)
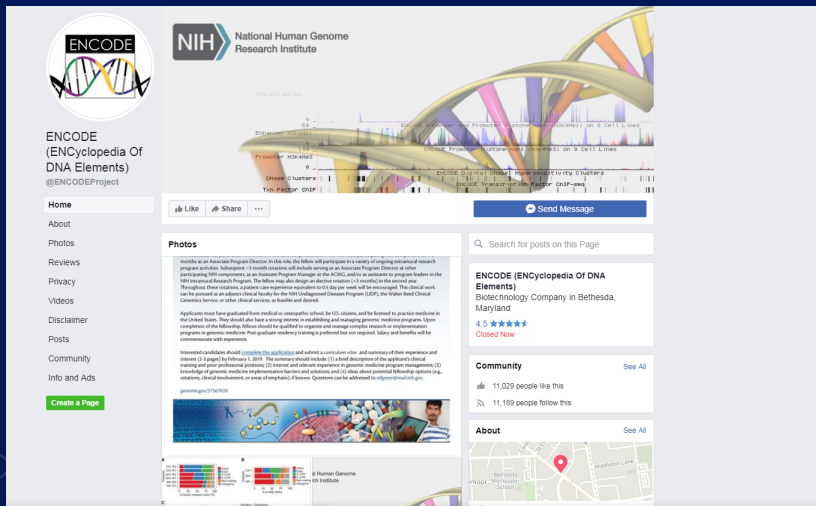`arn:aws:s3:::encode-public`

AWS Region
`us-west-2`

Follow ENCODE

facebook

ENCODE (ENCyclopedia Of DNA Elements)

Twitter

@ENCODE_NIH

ENCODE Consortium

NHGRI Staff

Program Directors

Elise Feingold

Dan Gilchrist

Mike Pazin

Program Analysts

Michael Pagan

Eileen Cahill

Division Director
Carolyn Hutter