

ENCODE 2015

Where we are and will be?

April 2, 2015

Bing Ren

Keystone Symposium



Goals of ENCODE

- Catalog the functional elements in human and mouse genomes
- 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



Three Phases

I) Pilot Phase -1% of Genome (2003-2007)

II) Scale Up Phase I (2007-2012)

III) Current Production Phase (2012-2016)

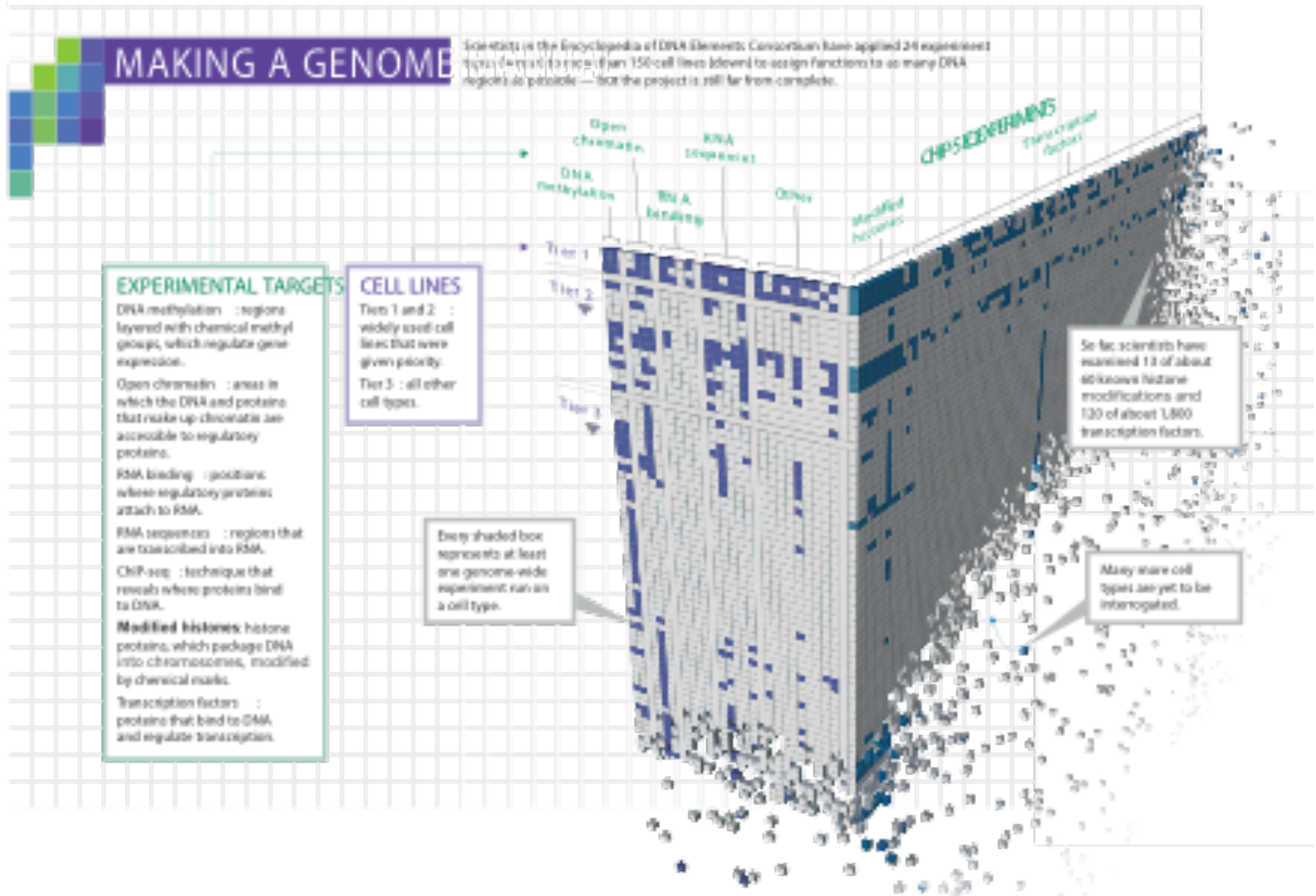
Related Projects:

Mouse ENCODE (2009-2012)

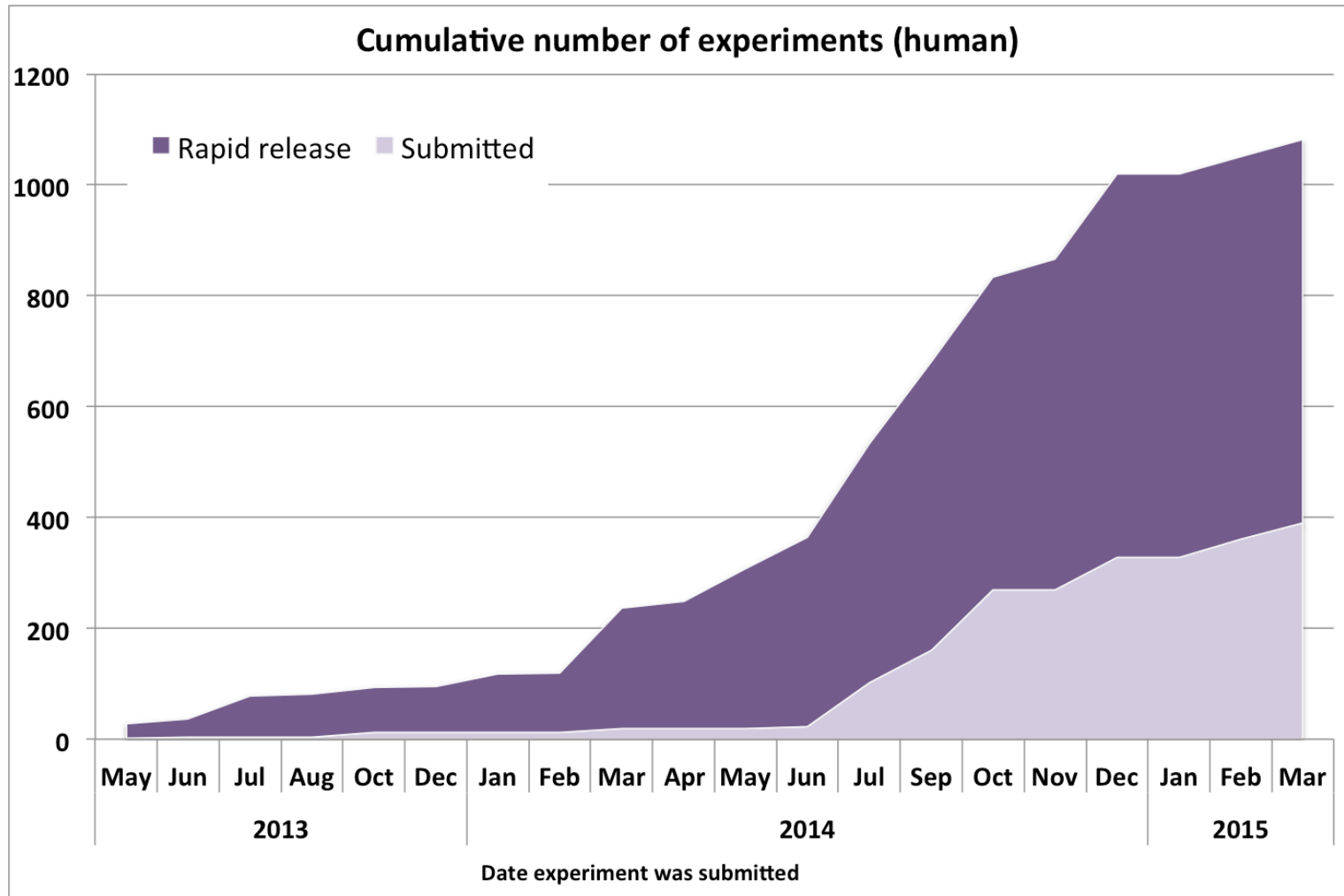
modENCODE (2007-2012)



The ENCODE Dataset has many dimensions

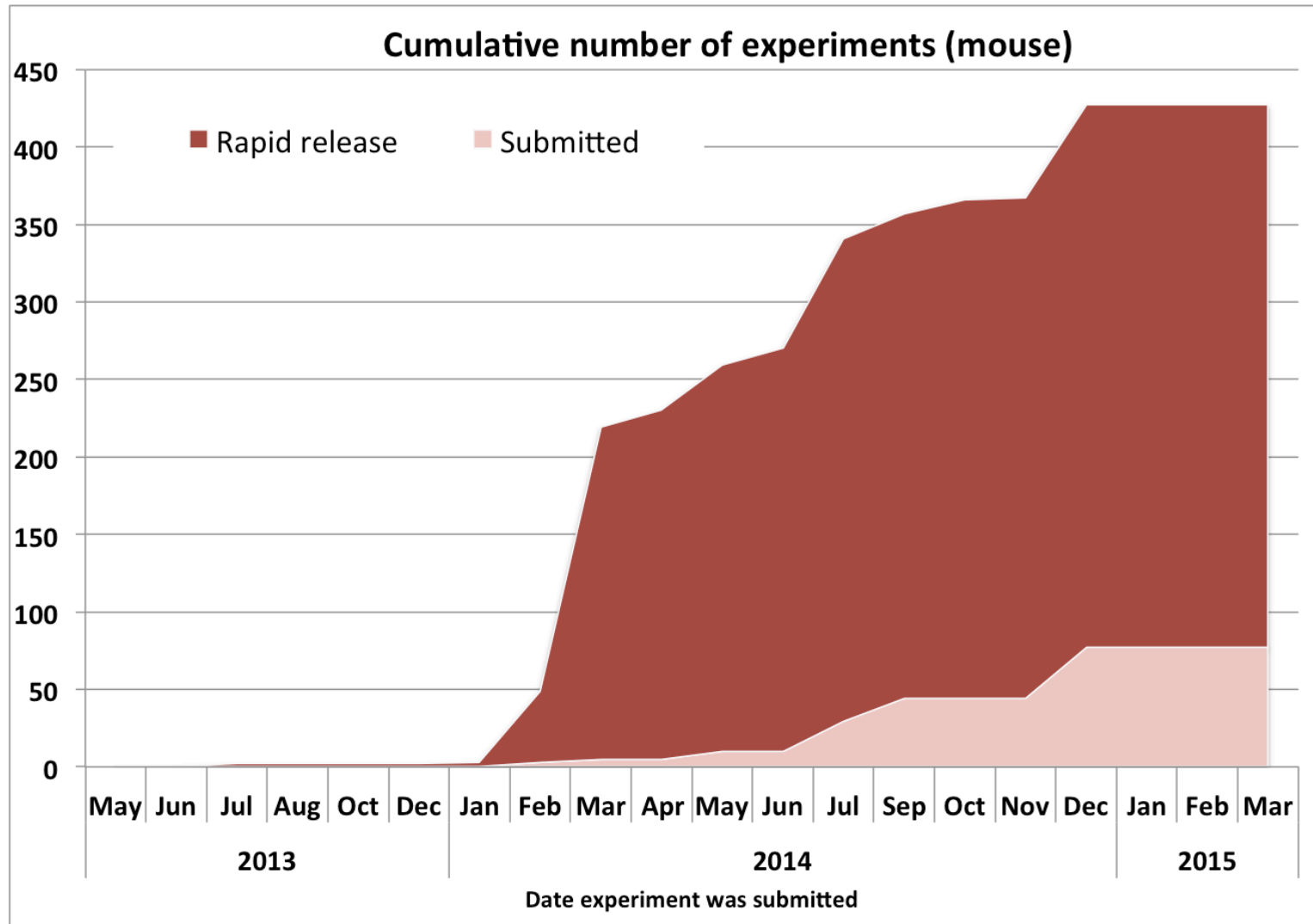


Overview of Human Datasets



Human: 3,331 datasets submitted/released; 5,501 proposed; 8,832 total

Overview of Mouse Datasets

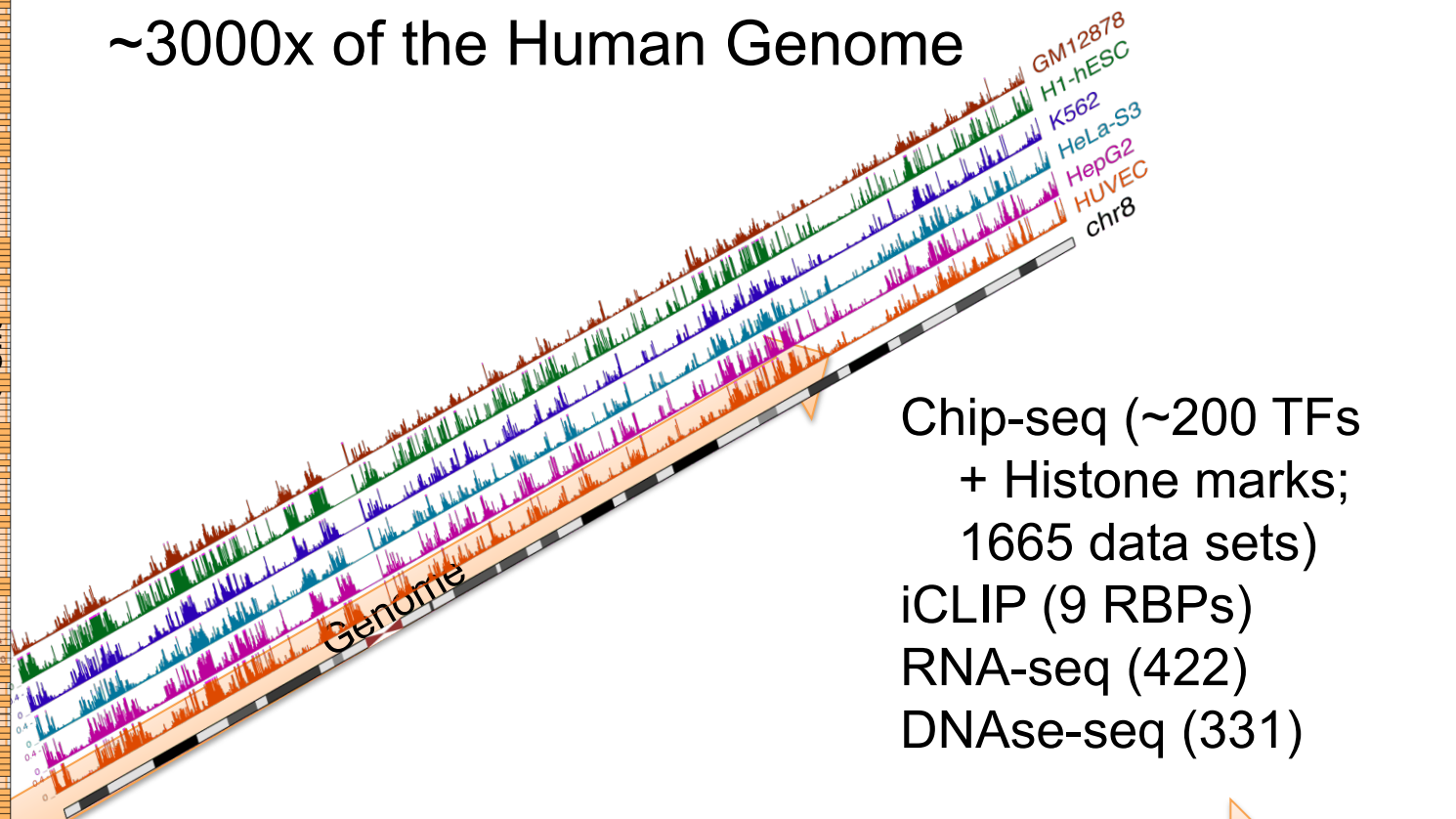
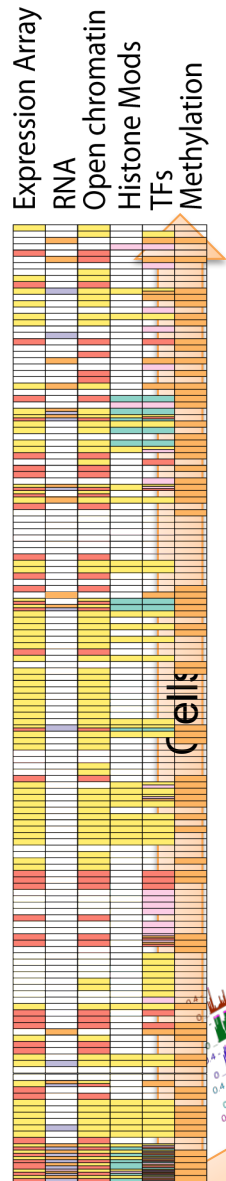


Mouse: 963 datasets submitted/released; 720 proposed; 1,683 total

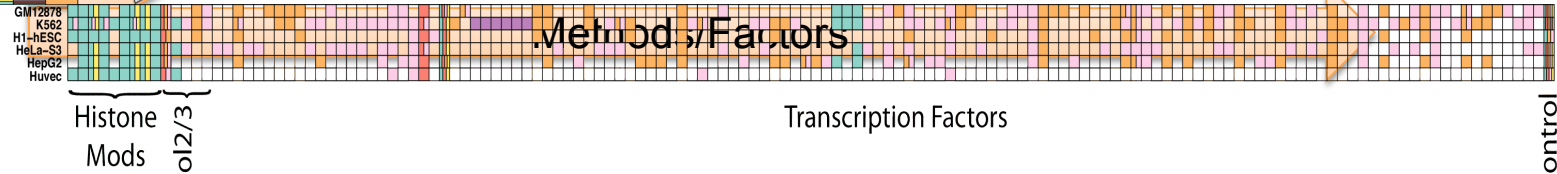
ENCODE Dimensions (Current)

3,331 Experiments
 >5 Tb TeraBases
 ~3000x of the Human Genome

282 Cell Lines/ Tissues



Chip-seq (~200 TFs + Histone marks; 1665 data sets)
 iCLIP (9 RBPs)
 RNA-seq (422)
 DNase-seq (331)



K562: 513 Assays (Epigenome/196 TFs/7 RBPs)

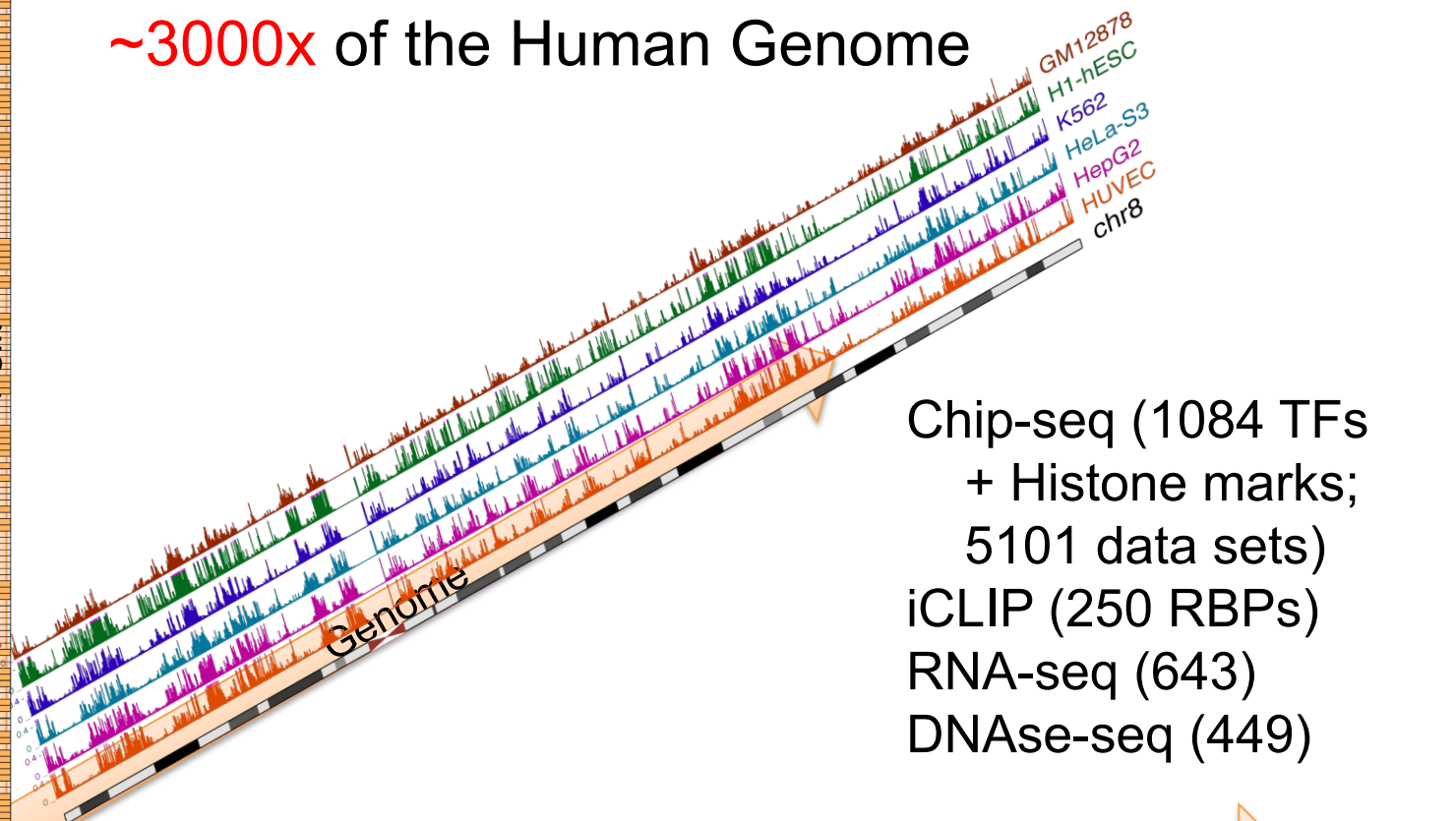
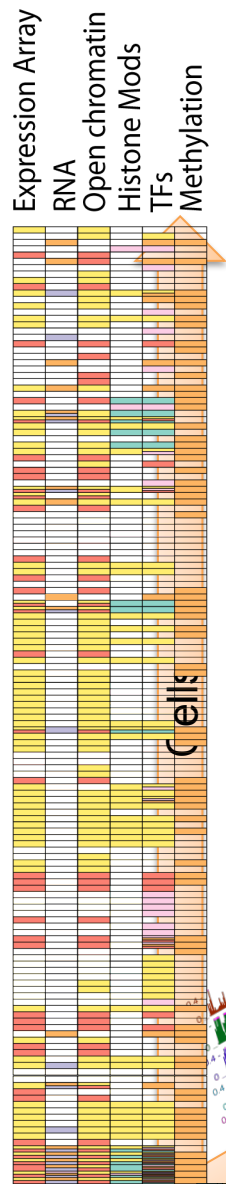
ENCODE Dimensions (2016?)

8,832 Experiments

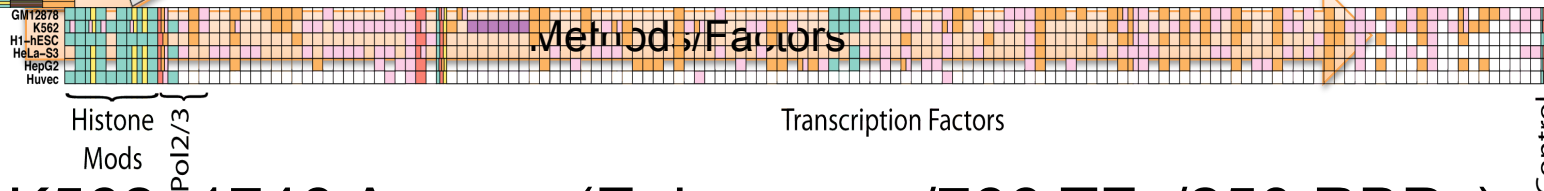
>5 Tb TeraBases

~3000x of the Human Genome

387 Cell Lines/ Tissues

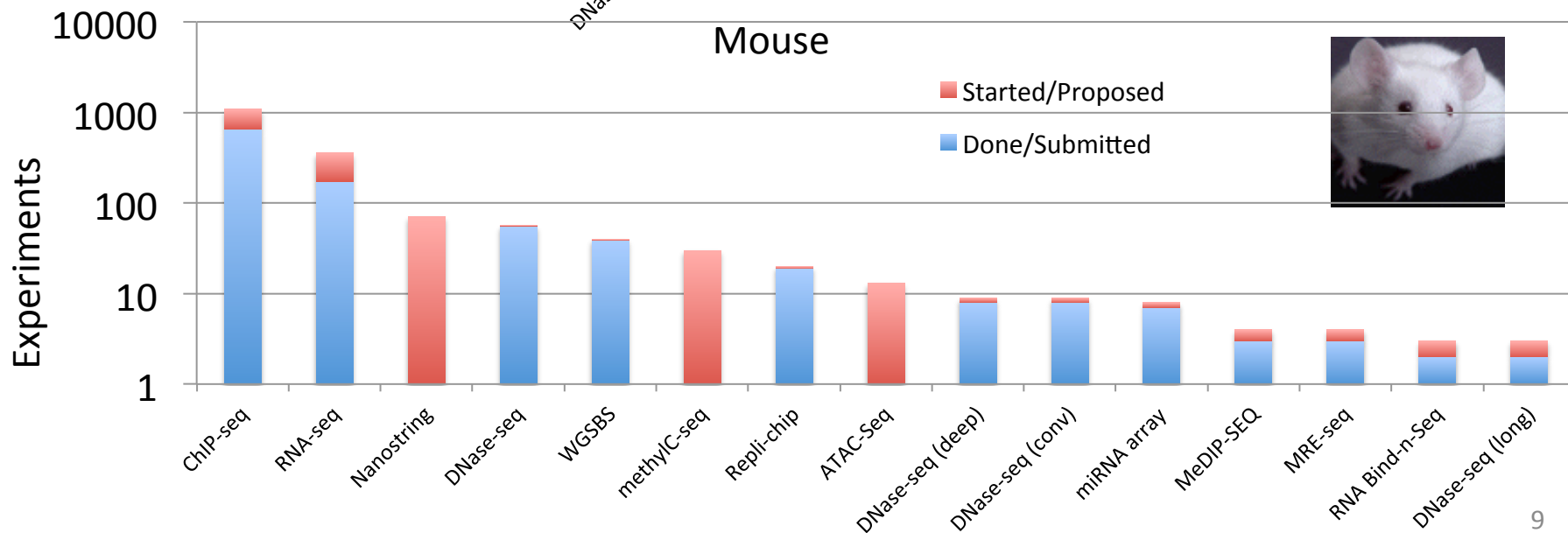
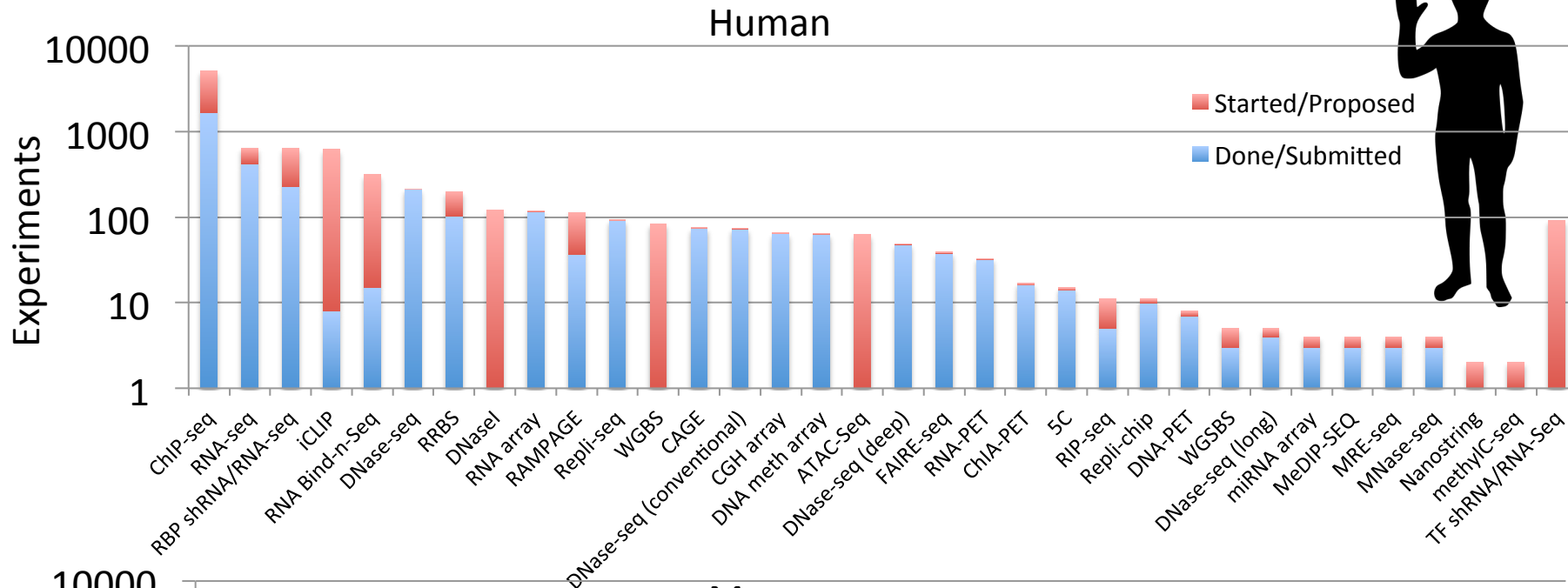


Chip-seq (1084 TFs + Histone marks; 5101 data sets)
iCLIP (250 RBPs)
RNA-seq (643)
DNase-seq (449)

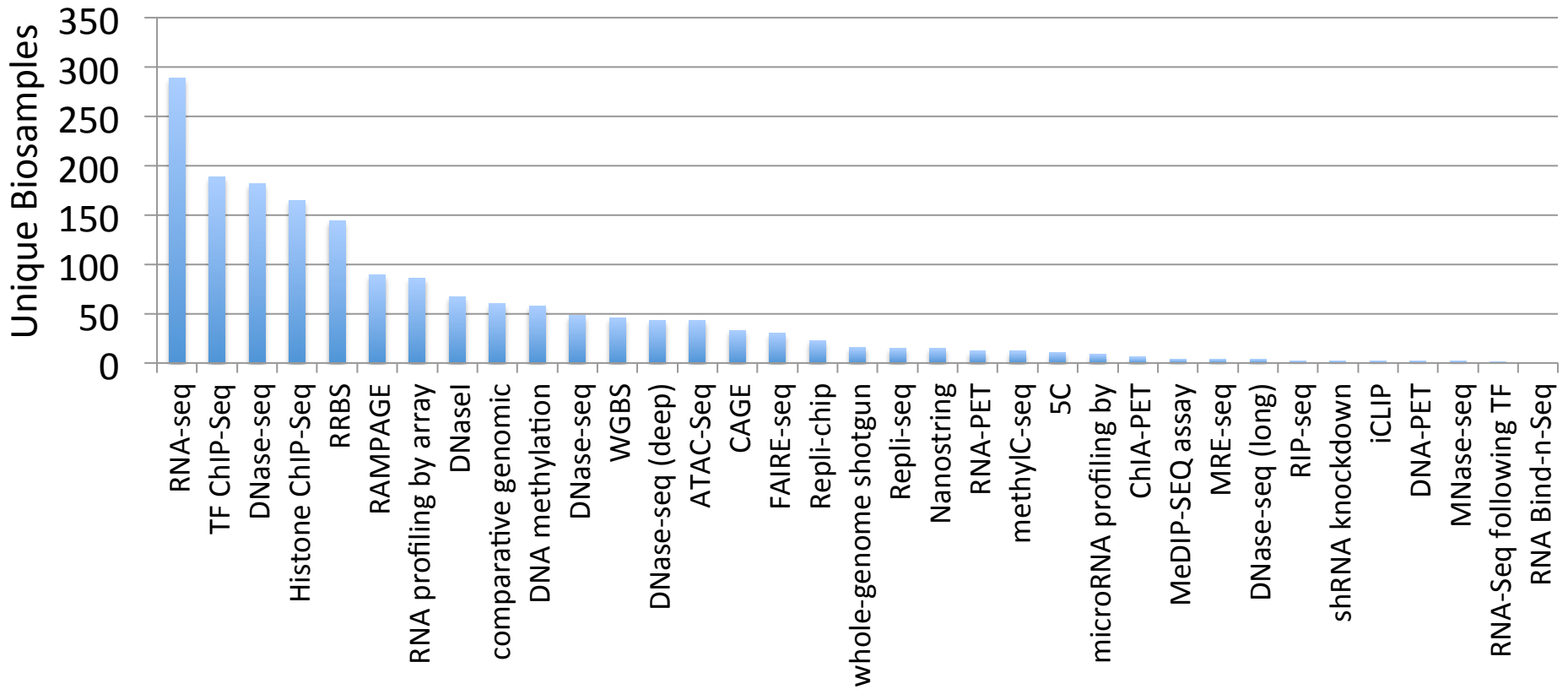


K562: 1746 Assays (Epigenome/766 TFs/250 RBPs)

Data Types

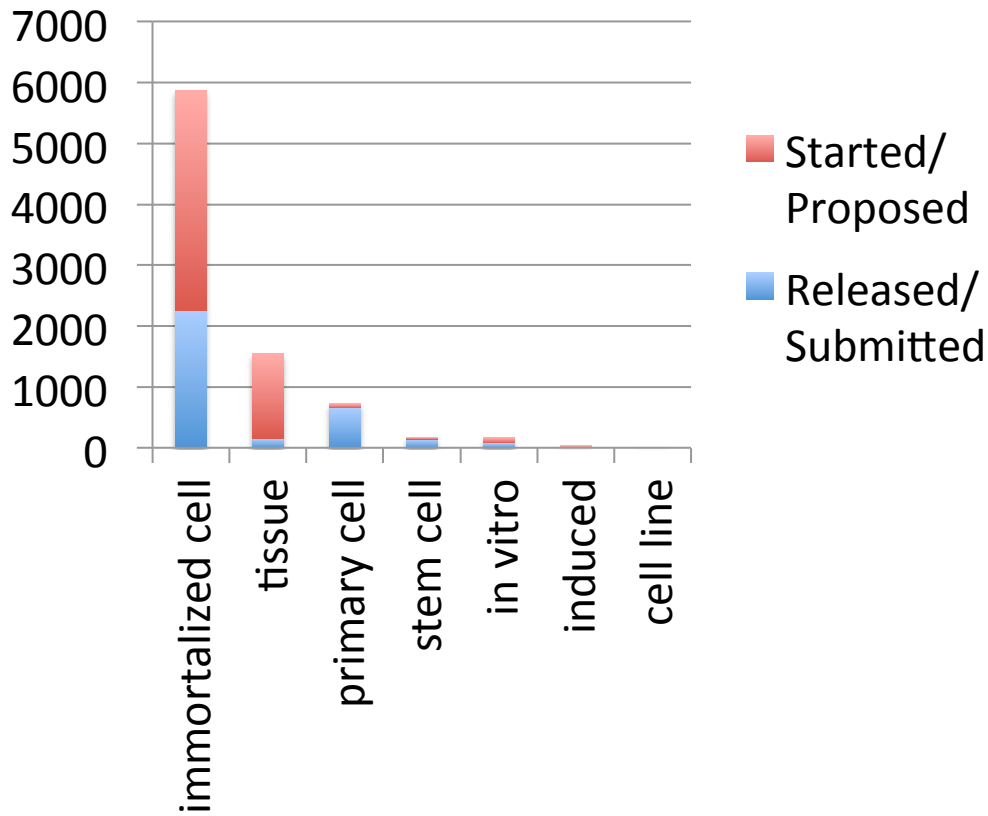


Some Assays Were Conducted Across a Broad Range of Biosamples

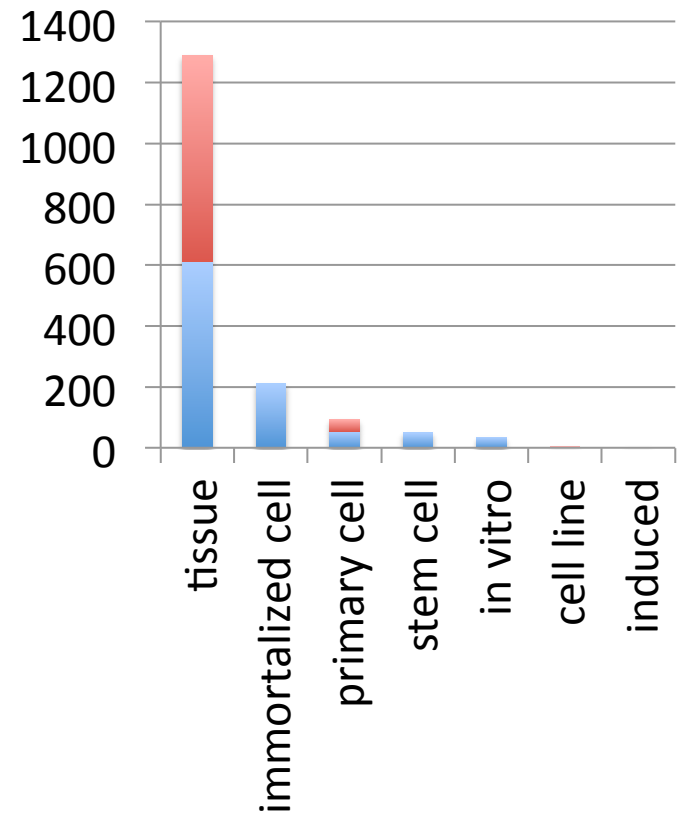


Number of Data Set Per Biosample Type

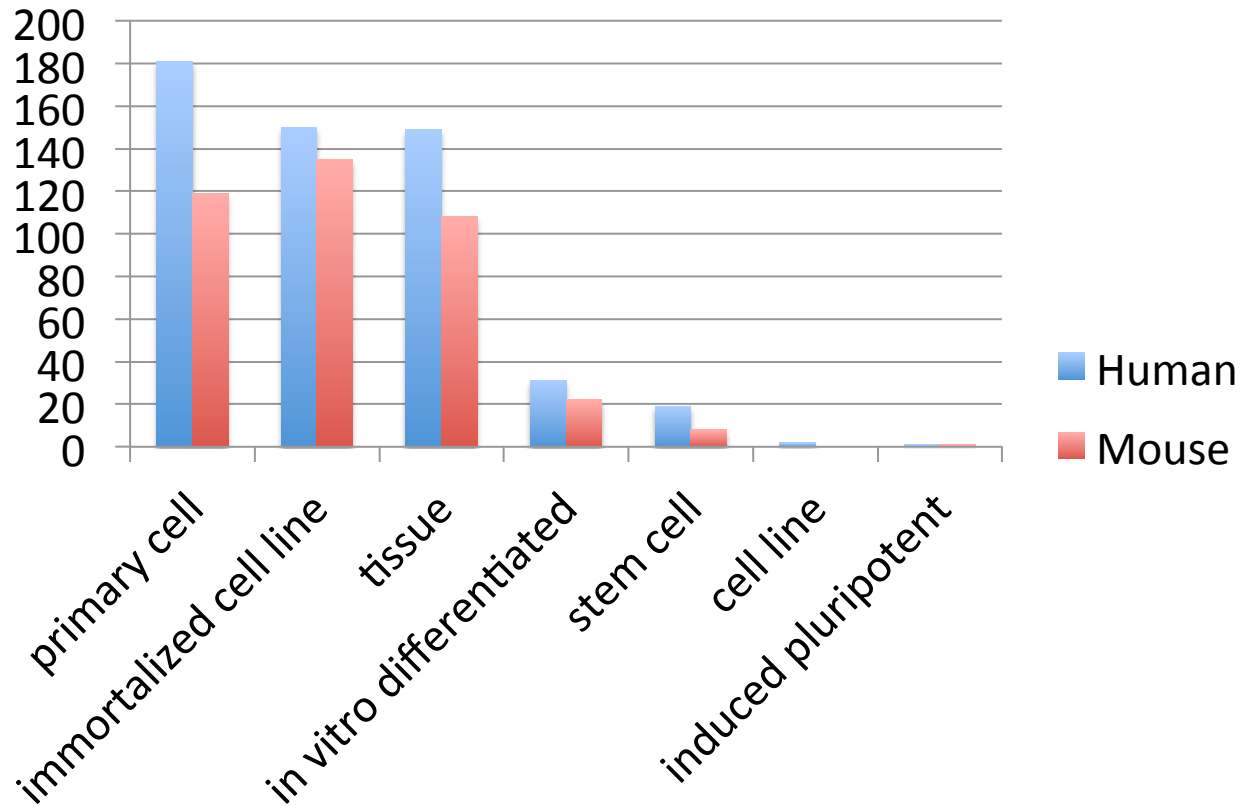
Human



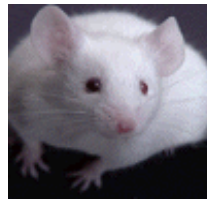
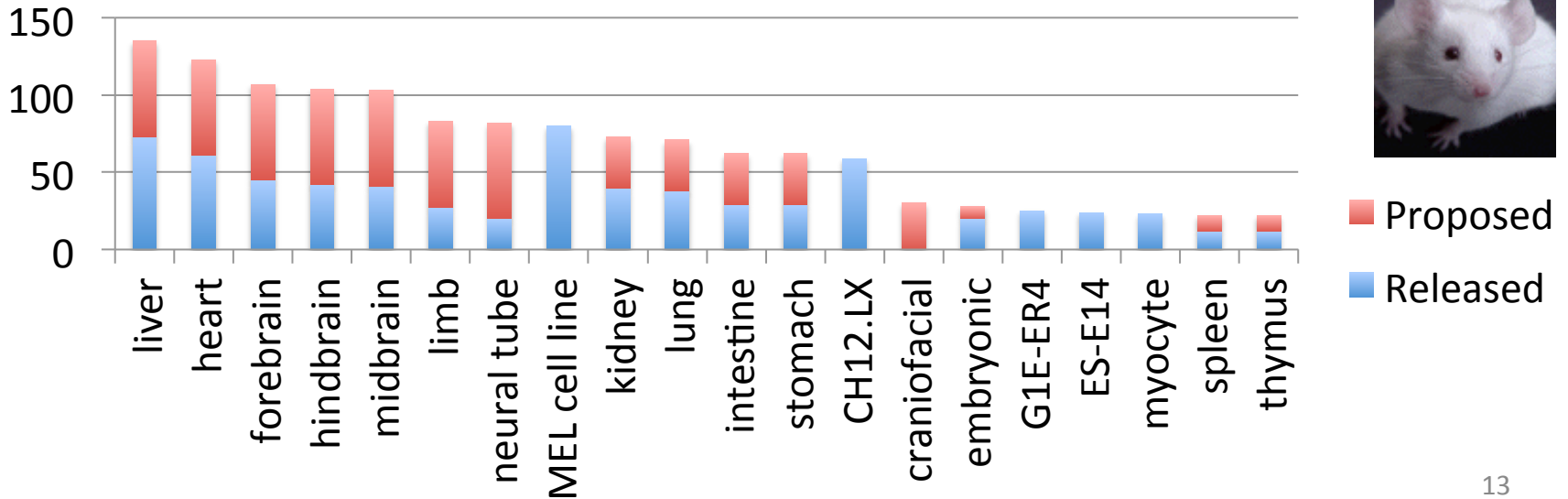
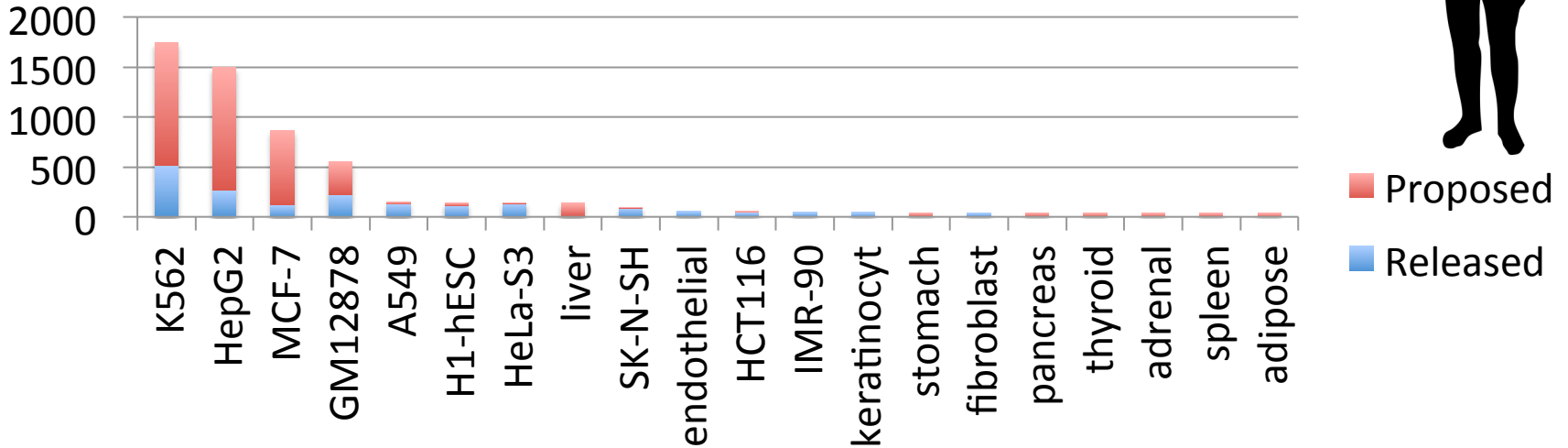
Mouse



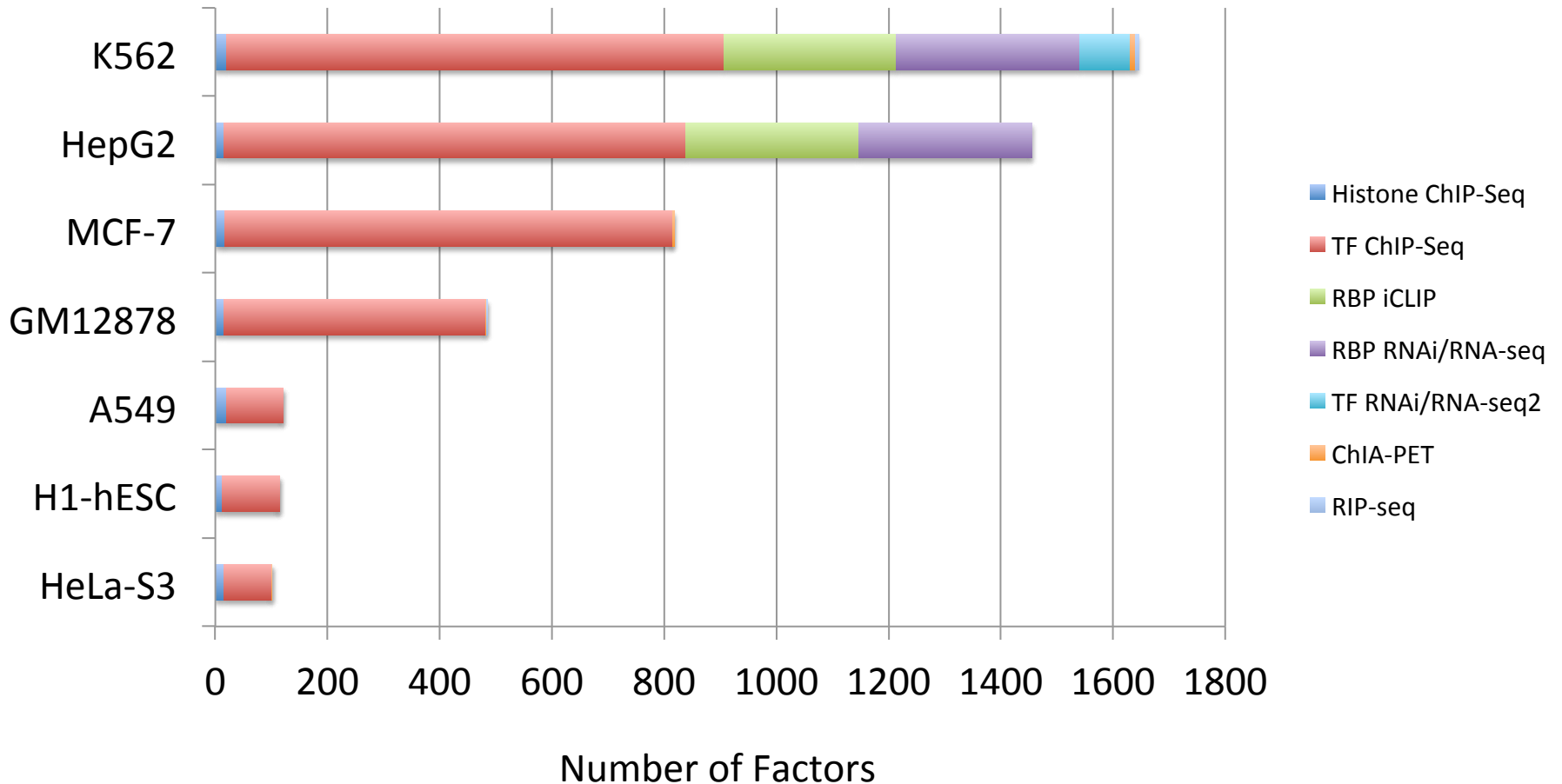
Unique Biosample Types



Assays Per Biosample



Deep Exploration of Factors



Established Standards For Community

- ChIP-Seq
- DNAaseHS
- RNA-Seq

Antibody characterizaiton, Biological replicates,
QC measures

Resource

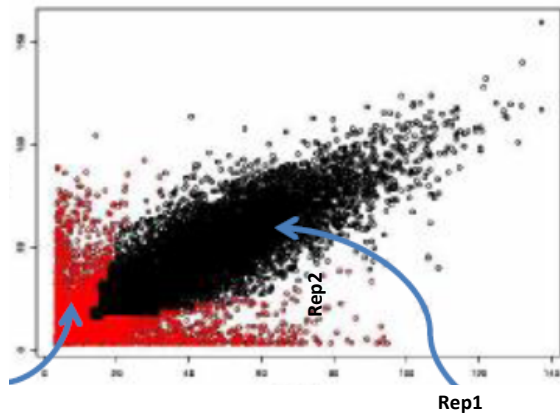
ChIP-seq guidelines and practices of the ENCODE
and modENCODE consortia

Stephen G. Landt,^{1,26} Georgi K. Marinov,^{2,26} Anshul Kundaje,^{3,26} Pouya Kheradpour,⁴
Florescia Pauli,⁵ Serafim Batzoglou,³ Bradley E. Bernstein,⁶ Peter Bickel,⁷ James B. Brown,⁷
Philip Cayting,¹ Yiwen Chen,⁸ Gilberto DeSalvo,² Charles Epstein,⁶
Katherine I. Fisher-Aylor,² Ghia Euskirchen,¹ Mark Gerstein,⁹ Jason Gertz,⁵

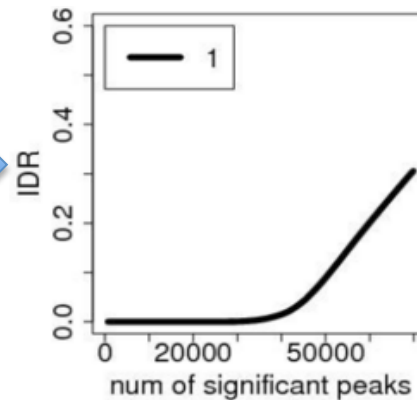
Genome Res.
2012

High Quality Data

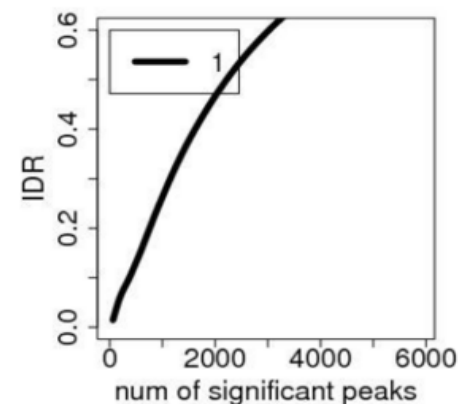
- \geq Two biological replicates
- Multiple quality control measures



Good
reproducibility



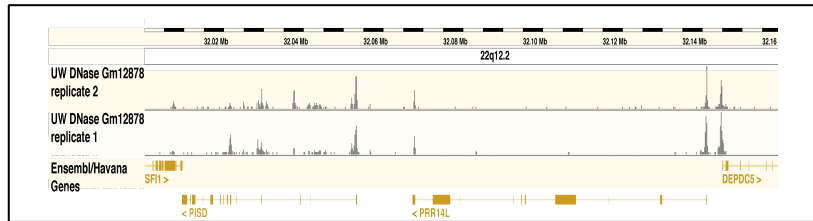
Poor
reproducibility



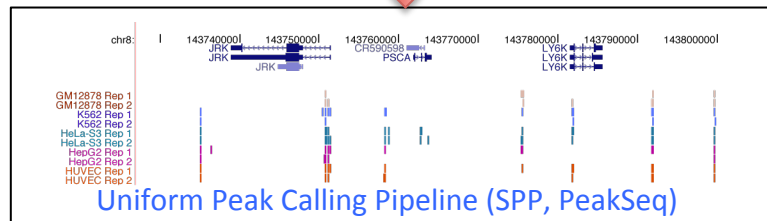
IDR Processing, QC
and Blacklist Filtering

ENCODE Uniform Analysis Pipeline

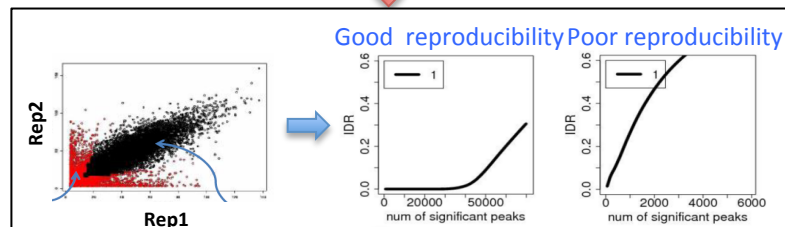
Anshul Kundaje



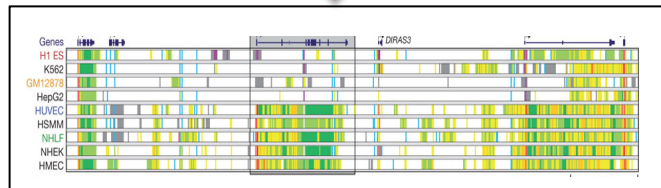
Mapped reads



Uniform peak Calling
(SPP, PeakSeq)



Quality Control



Derived Data
(Chromosome Segments,
Expression)

**Processing & Element Calling Compatible with Other Projects:
GTEx, REMC**

ENCODE Data

Cloud Storage and Computing

Data available at Amazon Web Services (AWS)

Uniform processing pipelines will be available at DNAnexus related projects

High Searchable

The screenshot shows the ENCODE Data website interface. At the top, there is a navigation bar with 'ENCODE' and menu items for 'Data', 'Methods', 'About ENCODE', and 'Help'. A search bar contains 'Search ENCODE' and a 'Sign in' link. On the left side, there are several filter panels:

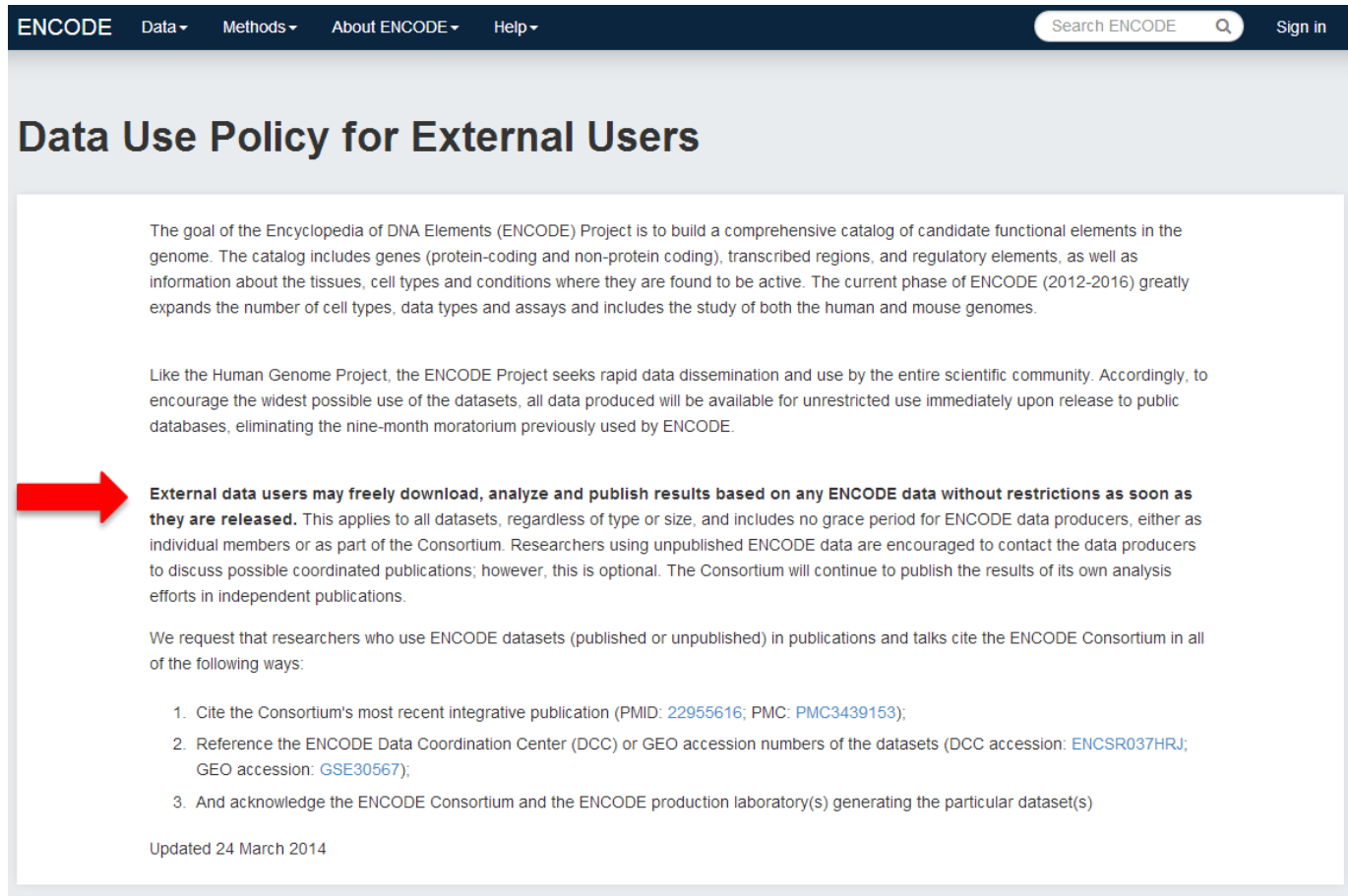
- Assay:** ChIP-seq (28), RNA-seq (6), Shotgun Bisulfite-seq (3), DNase-seq (3)
- Experiment status:** released (28)
- Organism:** *Mus musculus* (28)
- Biosample type:** tissue (28)
- Organ:** brain (57), liver (20), heart (20), bone element (9), stomach (8), lung (8), kidney (8)
- Life stage:** embryonic (28), adult (20), postnatal (16)
- Available data:** (empty)

The main content area shows 'Showing 25 of 28' results. A 'View All' button is in the top right. The first five results are:

- 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

The sixth result is partially visible: 'ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day) Experiment'.

ENCODE Data Open Access



ENCODE Data Methods About ENCODE Help Search ENCODE Sign in

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

New ENCODE Portal

<https://www.encodeproject.org>

The screenshot displays the ENCODE portal interface. At the top is a dark navigation bar with links for 'ENCODE', 'Data', 'Methods', 'About ENCODE', and 'Help'. A search bar and 'Sign in' button are on the right. Below the navigation bar is the title 'ENCODE: Encyclopedia of DNA Elements'. The main content area features a diagram illustrating the relationship between various assays and genomic features. The diagram shows a DNA strand with 'Long-range regulatory elements (enhancers, repressors/silencers, insulators)', 'Promoters', and 'Transcripts' (labeled as 'Genes'). Above the DNA, 'Hypersensitive Sites' are marked with blue dots, and 'CH₃' and 'CH₂CO' methyl groups are shown. An 'RNA polymerase' is depicted transcribing a gene into RNA. Below the diagram, several assay types are listed in boxes: '5C ChIA-PET', 'DNase-seq FAIRE-seq', 'ChIP-seq', 'WGBS RRBS methyl450k', 'Computational predictions and RT-PCR', 'RNA-seq', and 'CLIP-seq RIP-seq'. Arrows indicate the mapping from these assays to the genomic features. To the right of the diagram is a text block describing the ENCODE Consortium and its goals, followed by image credits.

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. [\[read more\]](#)

Software Tools

- >30 Different algorithms
- Wide array of areas. Examples:
 - Segmentation
 - Allele calling
 - 3D nuclear analysis
 - Data processing and peak calling
 - Data quality control

ENCODE Publications

