# 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)

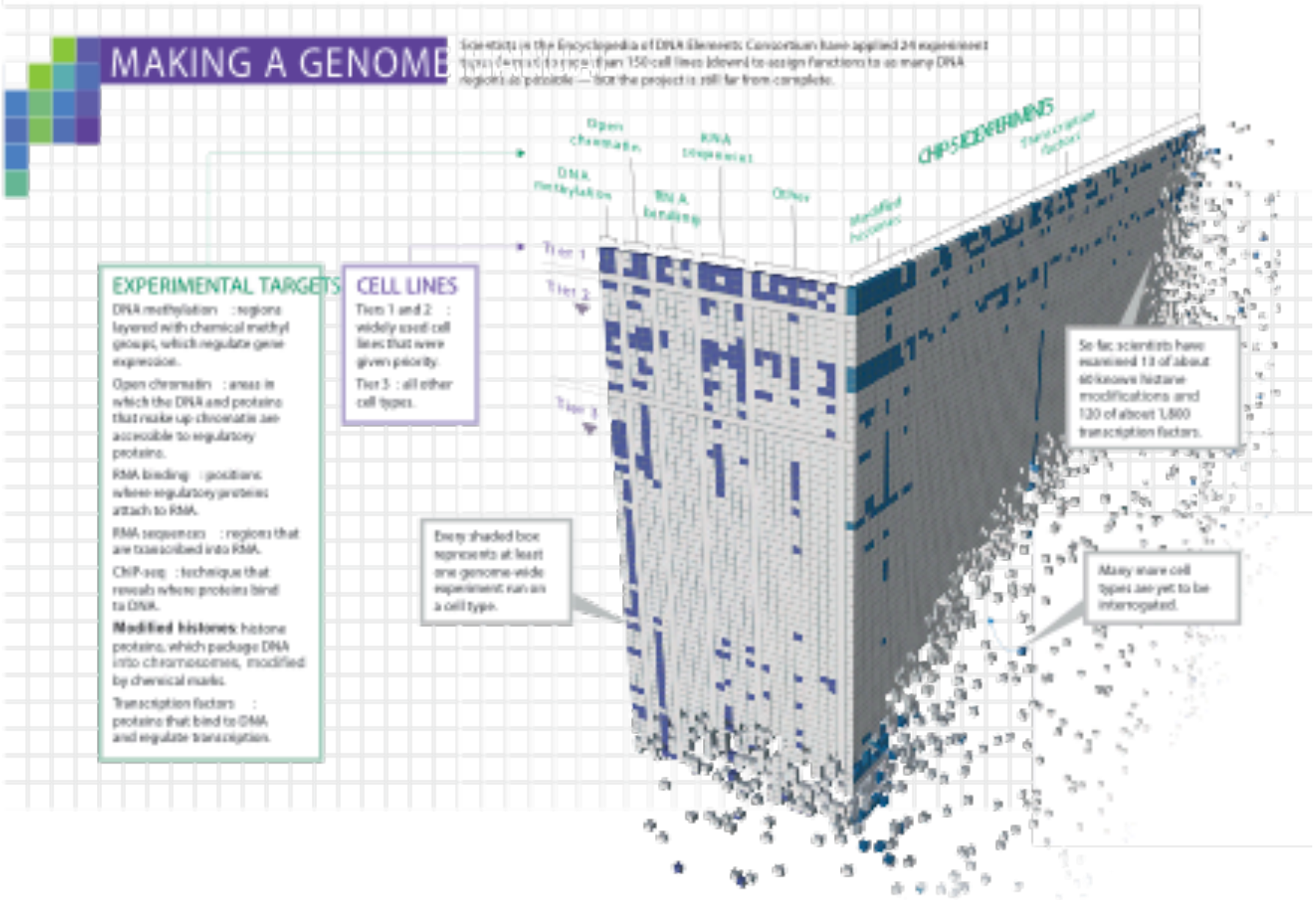
III) Current Production Phase (2012-2016)

**Related Projects**:

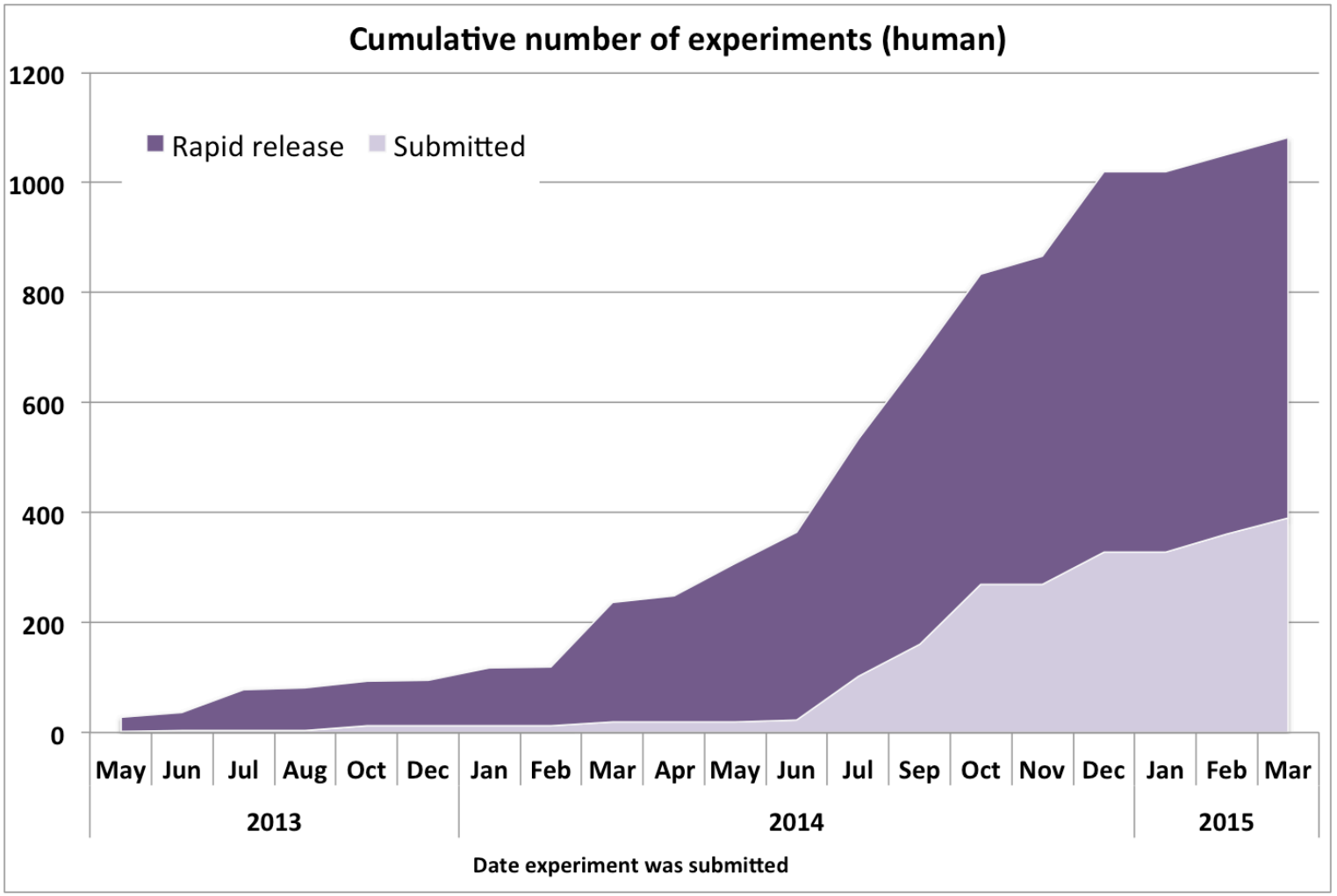
Mouse ENCODE (2009-2012)

modENCODE (2007-2012)

# The ENCODE Dataset has many dimensions

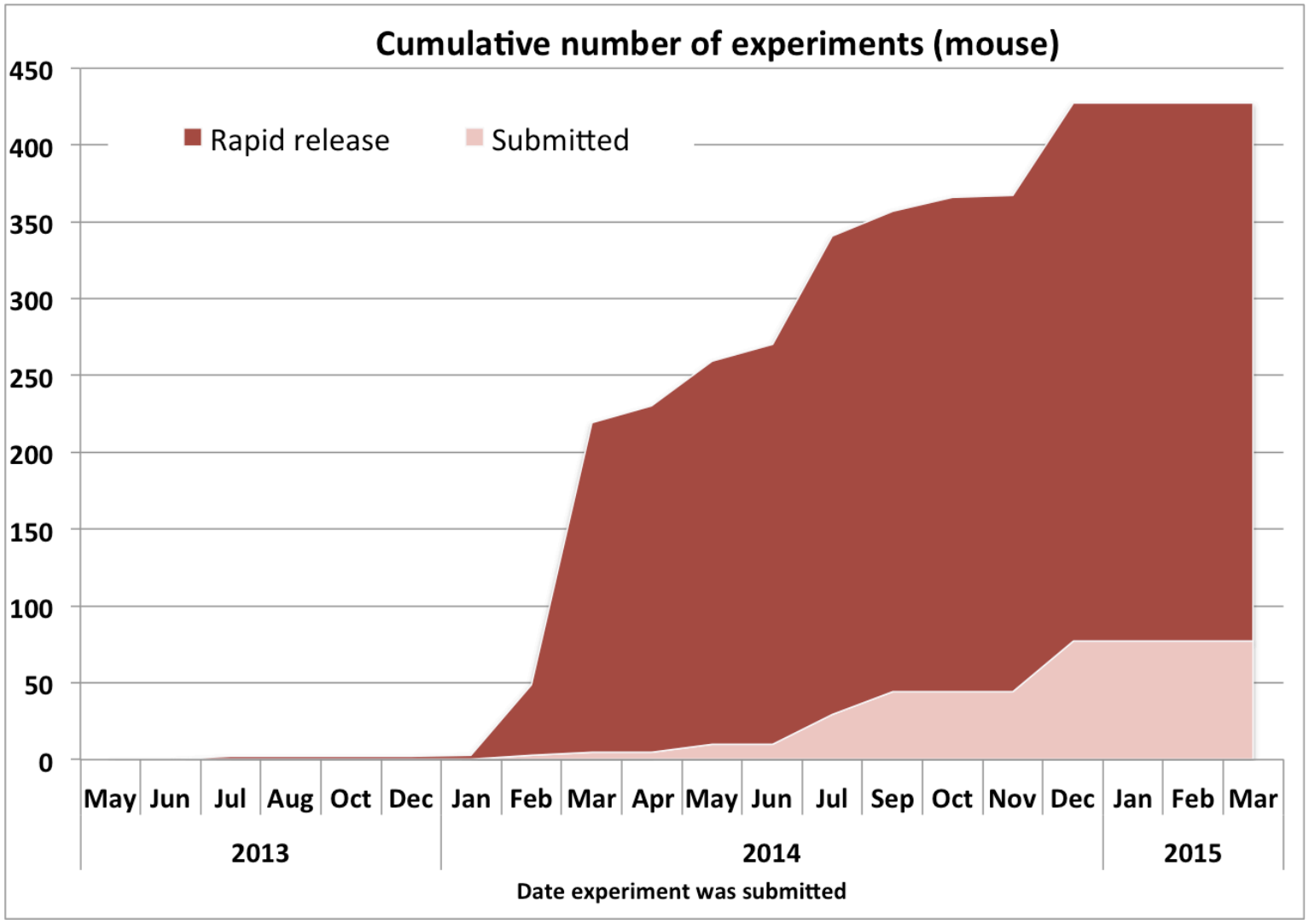


# Overview of Human Datasets




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

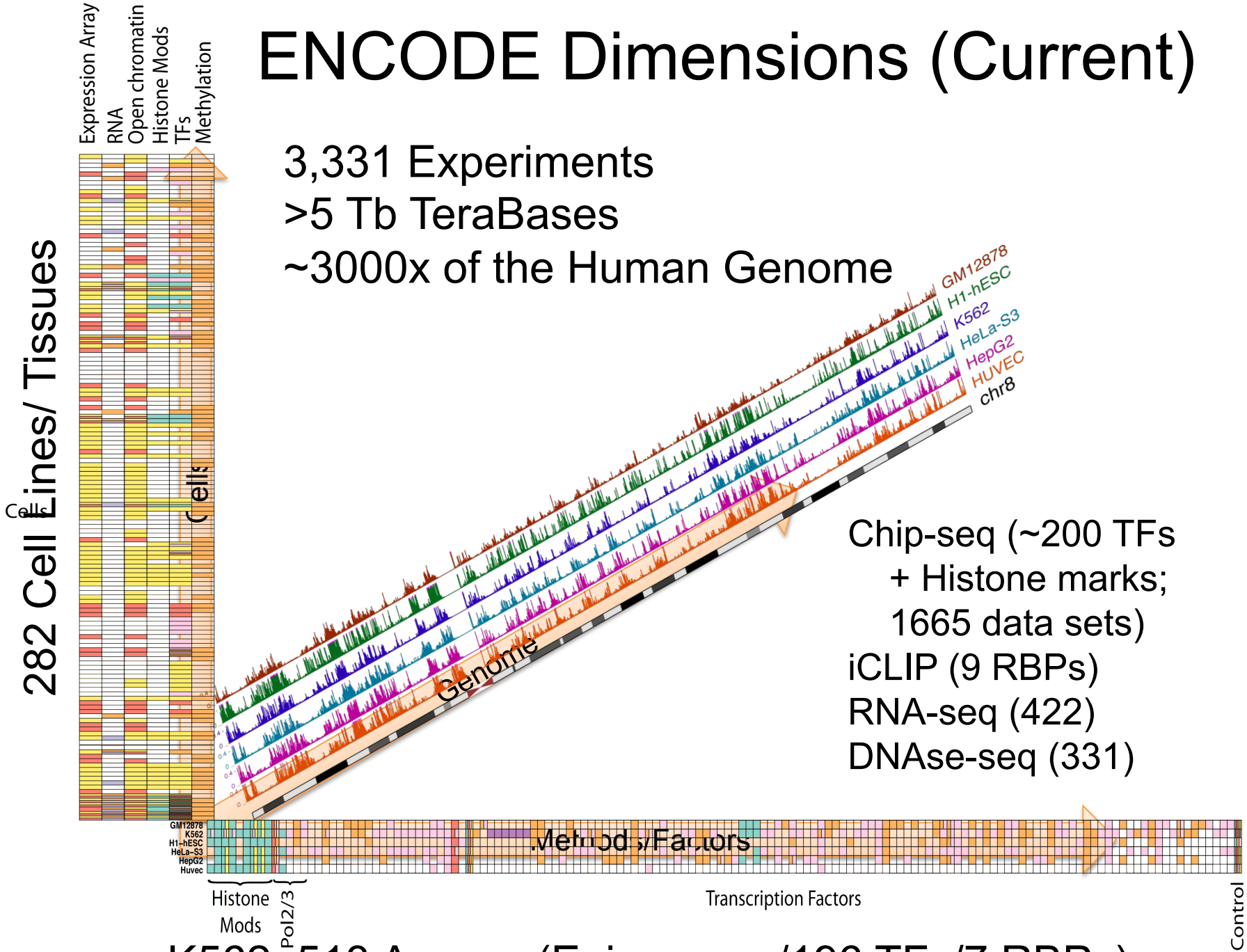
# Overview of Mouse Datasets




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

# ENCODE Dimensions (Current)

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




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

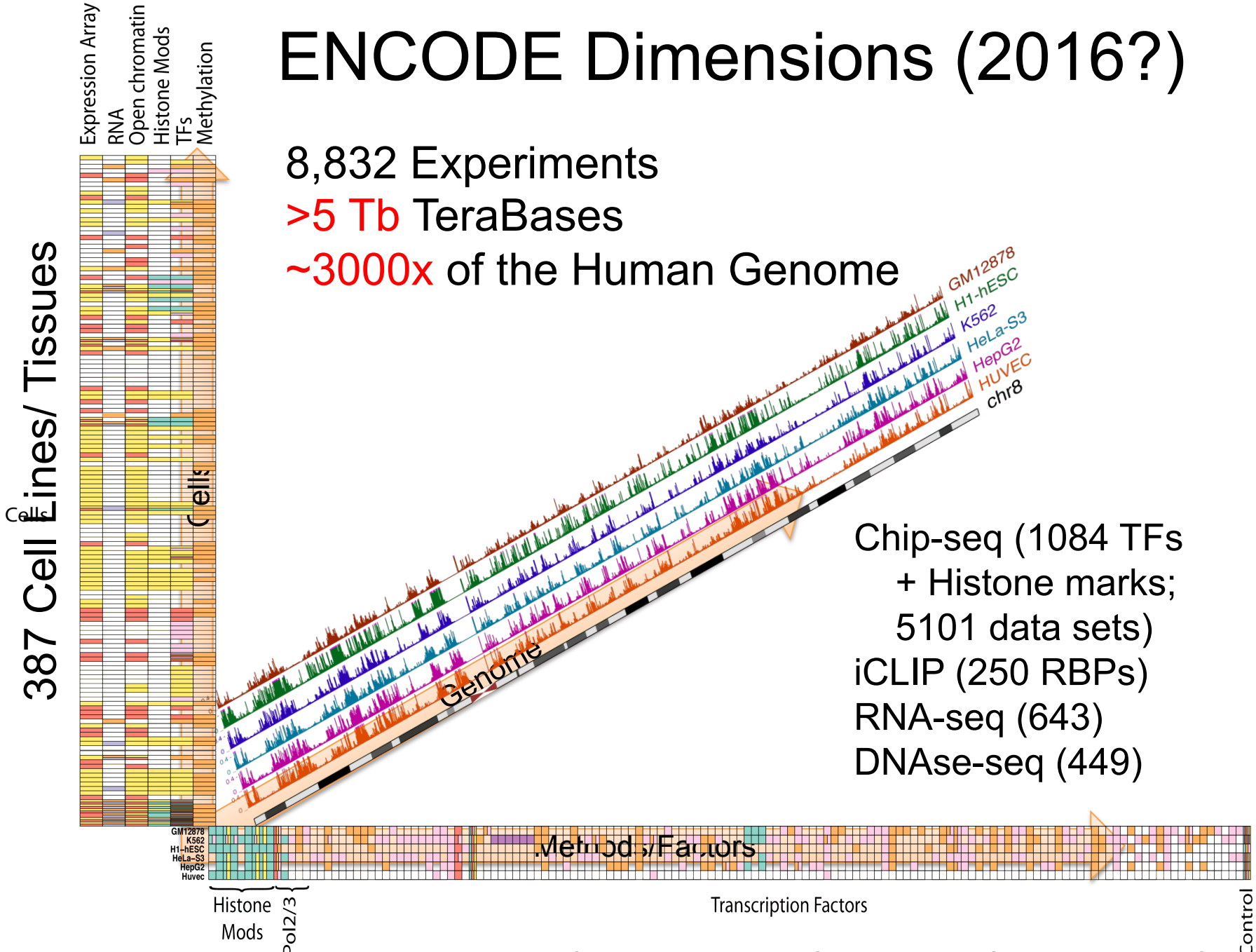
K562: 513 Assays (Epigenome/196 TFs/7 RBPs)

# ENCODE Dimensions (2016?)

8,832 Experiments

>5 Tb TeraBases

~3000x of the Human Genome




Chip-seq (1084 TFs + Histone marks; 5101 data sets)

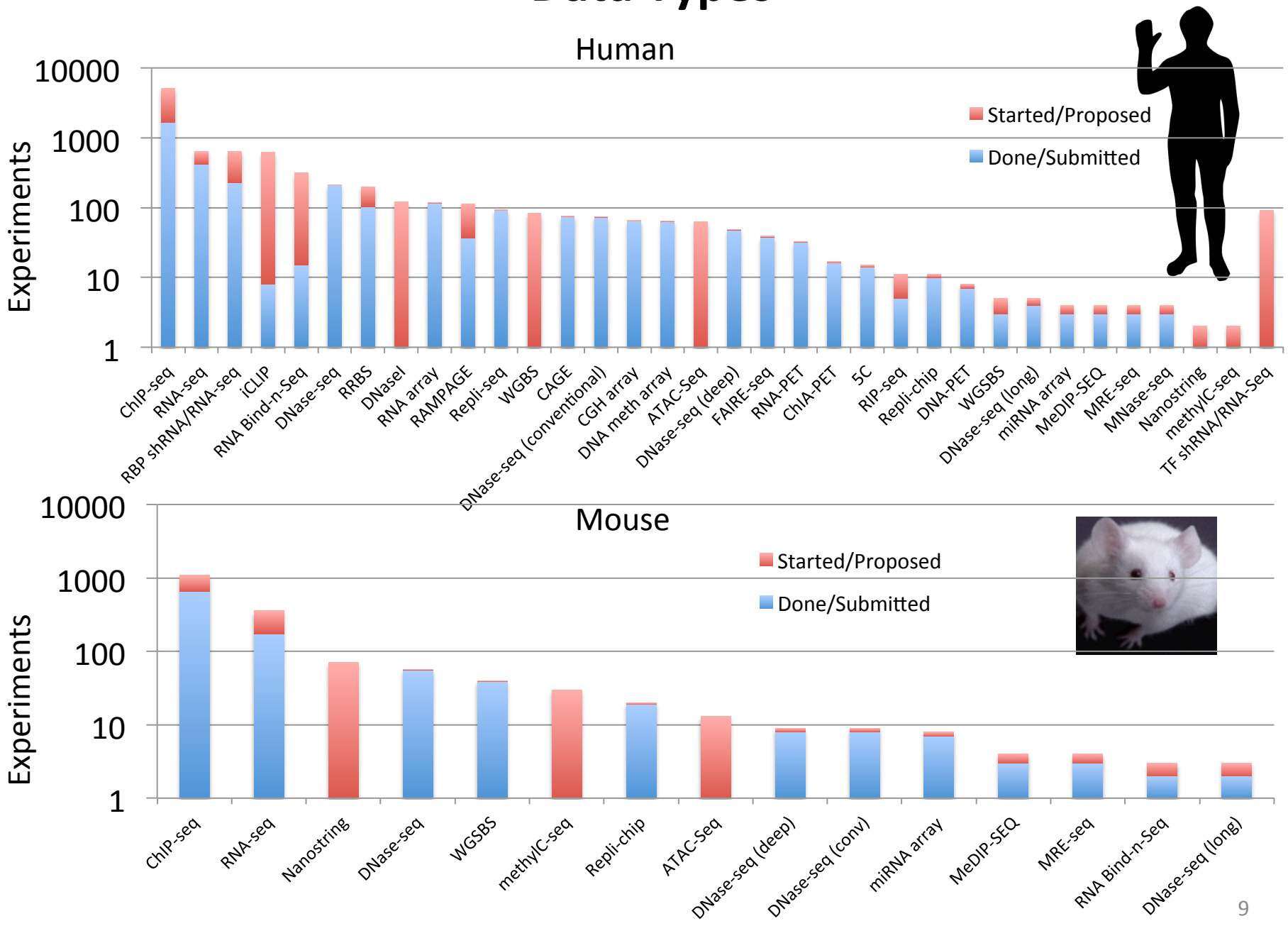
iCLIP (250 RBPs)

RNA-seq (643)

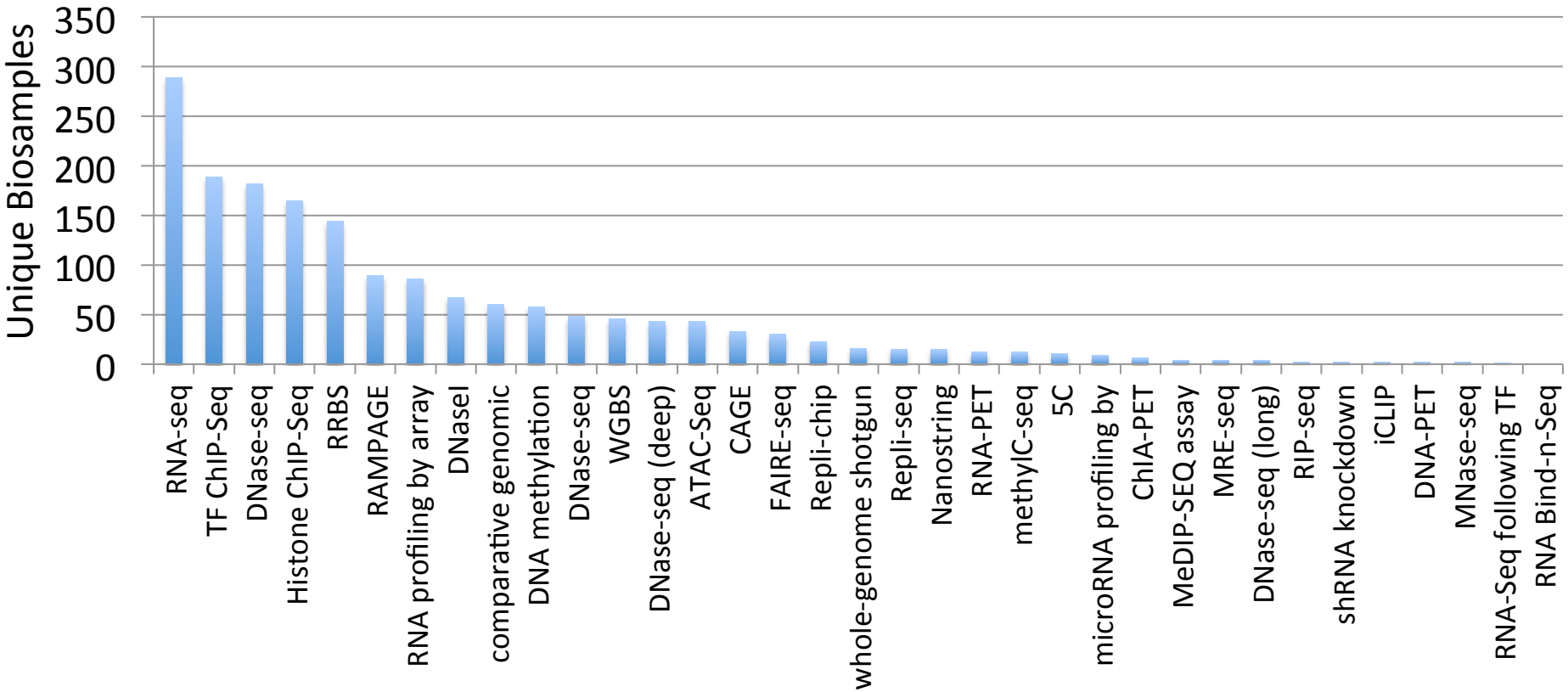
DNAse-seq (449)

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

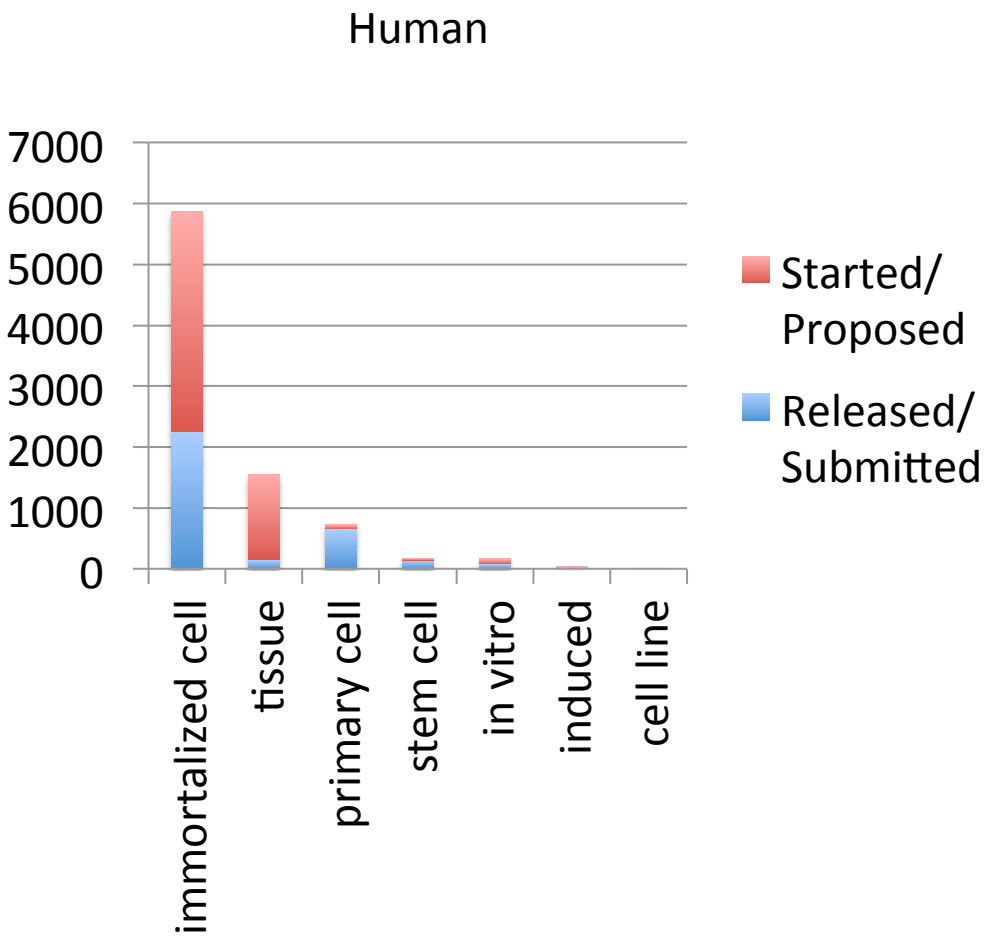
# Data Types

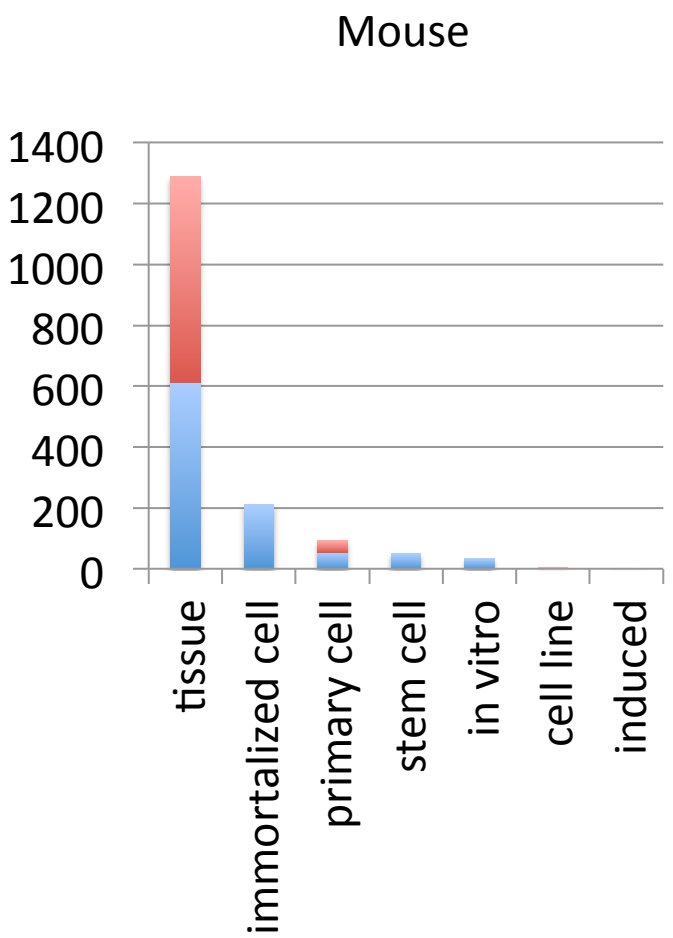

Human
Experiments
10000
1000
100
10
1
Started/Proposed
Done/Submitted
ChIP-seq
RNA-seq
RBP shRNA/RNA-seq
iCLIP
RNA Bind-n-Seq
DNase-seq
RRBS
DNaseI
RNA array
RAMPAGE
Repli-seq
WGBS
CAGE
DNase-seq (conventional)
CGH array
DNA meth array
ATAC-Seq
DNase-seq (deep)
FAIRE-seq
RNA-PET
ChIA-PET
5C
RIP-seq
Repli-chip
DNA-PET
WGSBS
DNase-seq (long)
miRNA array
MeDIP-SEQ
MRE-seq
MNase-seq
Nanostring
methylC-seq
TF shRNA/RNA-Seq
Mouse
Experiments
10000
1000
100
10
1
Started/Proposed
Done/Submitted
ChIP-seq
RNA-seq
Nanostring
DNase-seq
WGSBS
methylC-seq
Repli-chip
ATAC-Seq
DNase-seq (deep)
DNase-seq (conv)
miRNA array
MeDIP-SEQ
MRE-seq
RNA Bind-n-Seq
DNase-seq (long)

# Some Assays Were Conducted Across a Broad Range of Biosamples

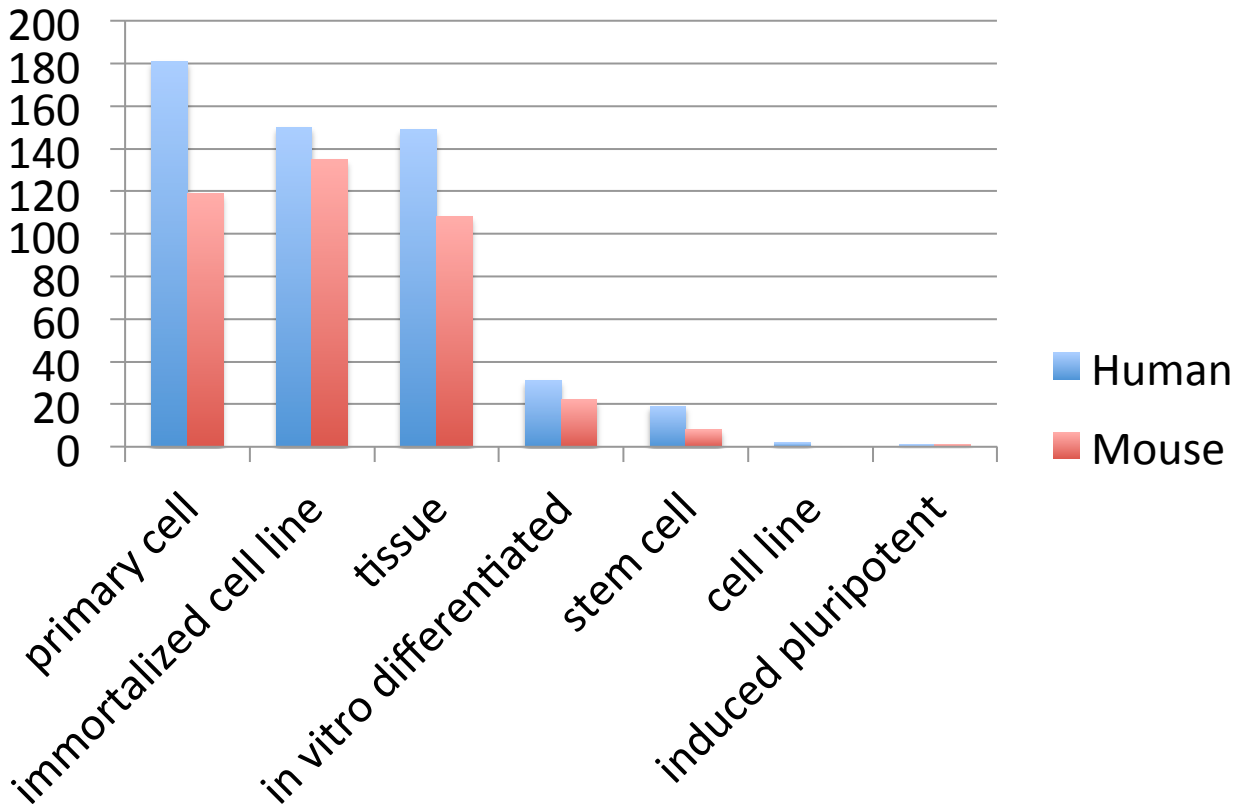


# Number of Data Set Per Biosample Type

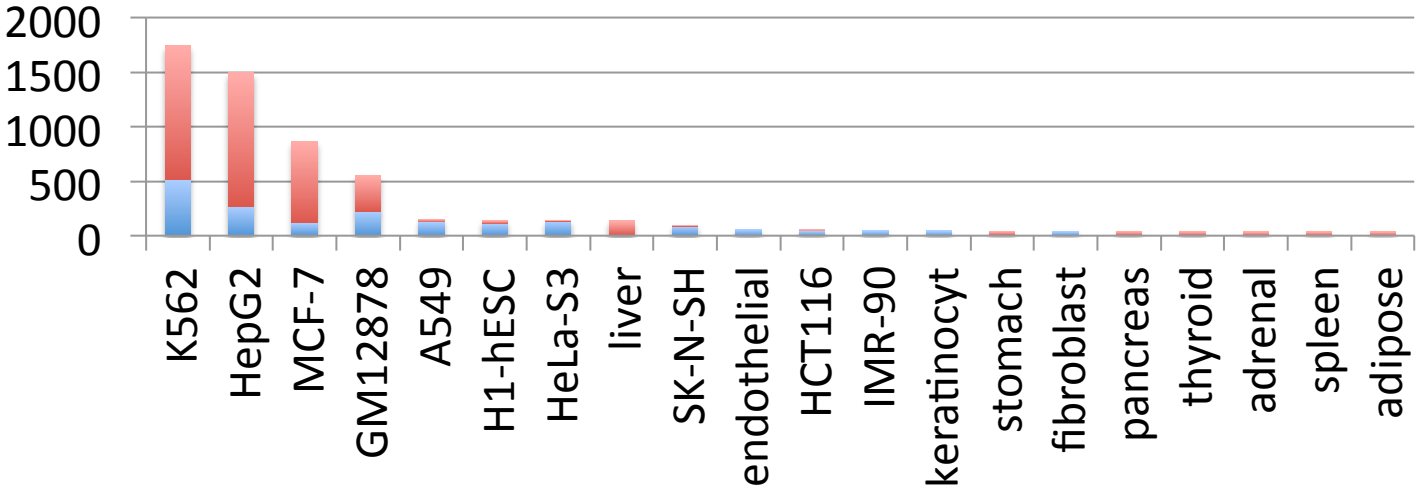


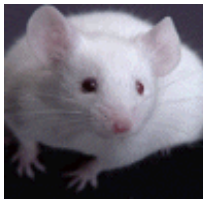



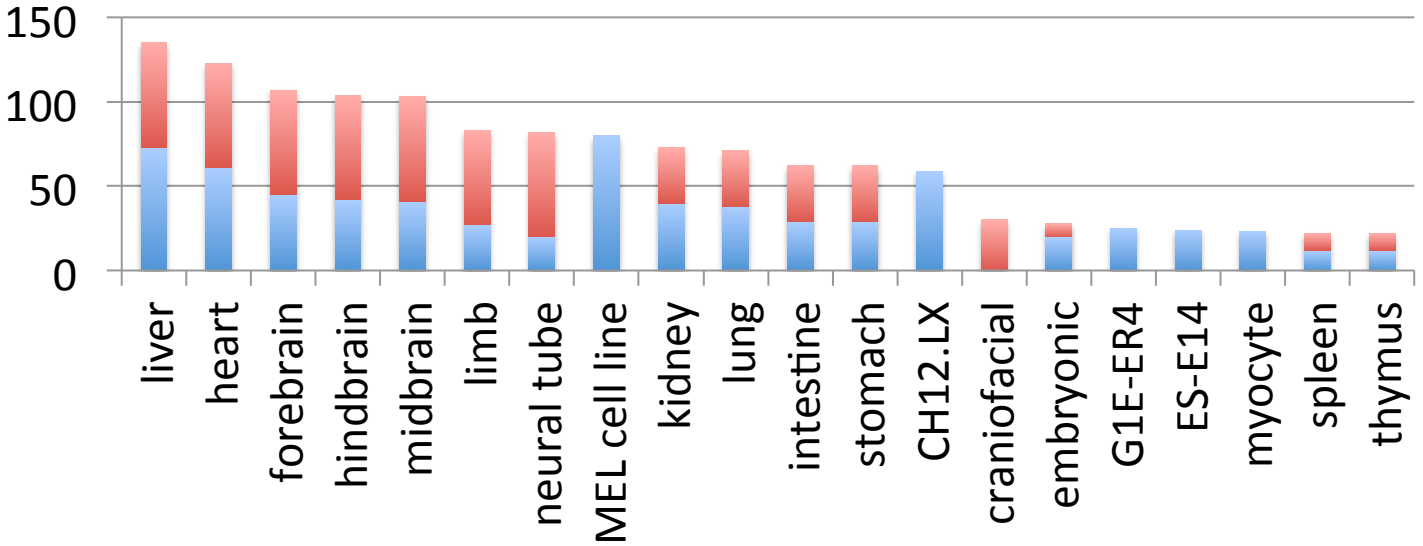
# Unique Biosample Types



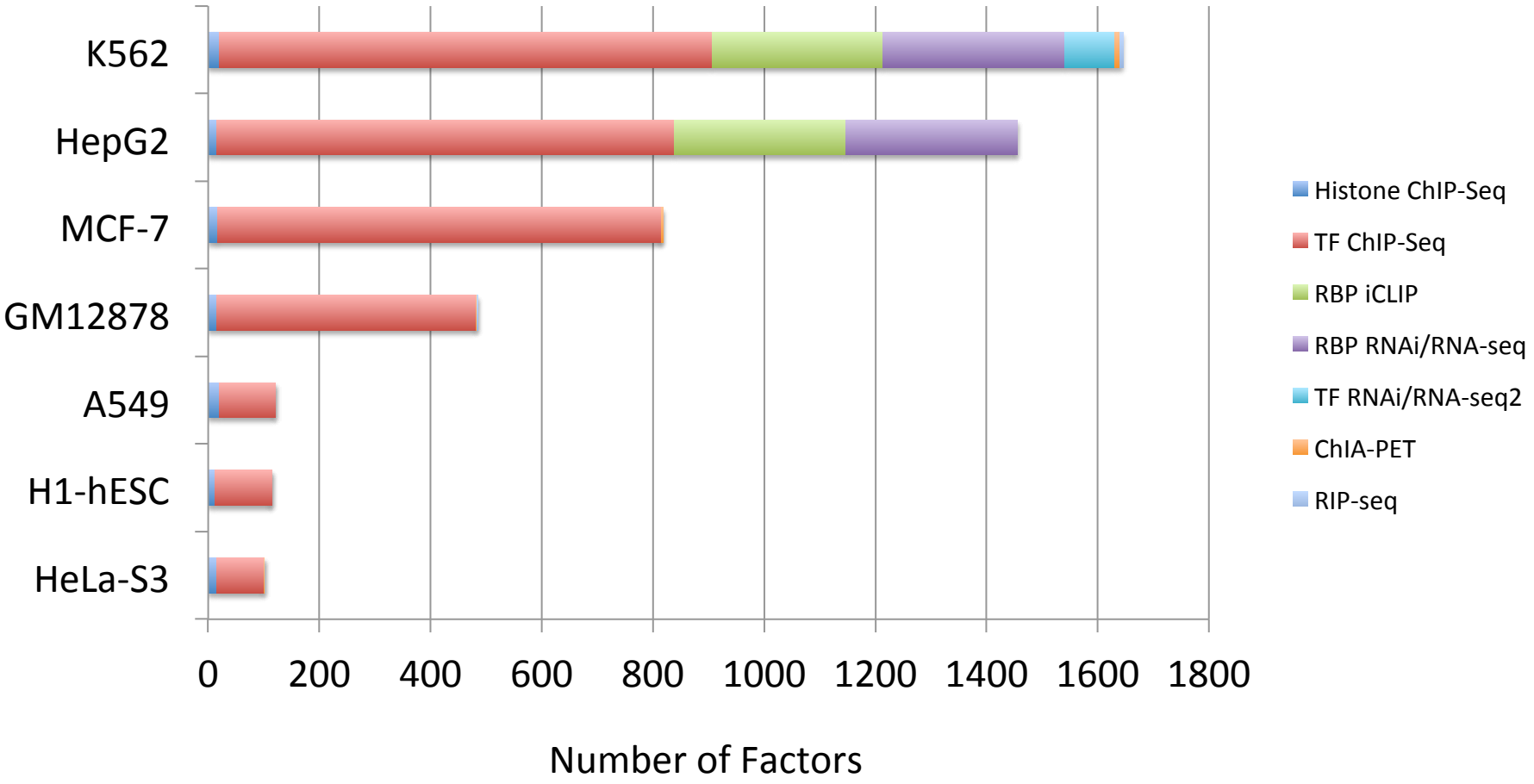
# Assays Per Biosample





# Deep Exploration of Factors



# Established Standards For Community

- ChIP-Seq
- DNAaseHS
- RNA-Seq

Antibody characterizaiton, Biological replicates, QC measures

**Resource**

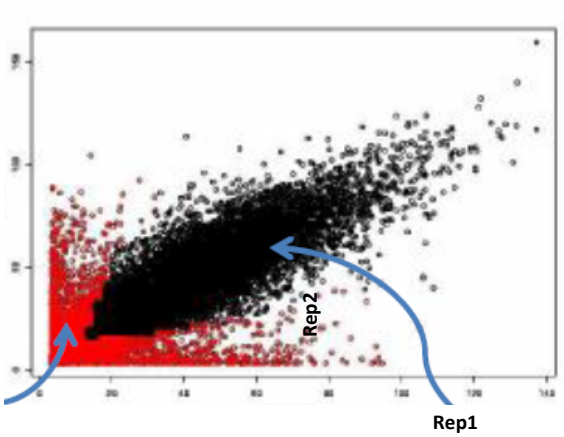
## ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Stephen G. Landt,[1,26] Georgi K. Marinov,[2,26] Anshul Kundaje,[3,26] Pouya Kheradpour,[4] Florencia 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

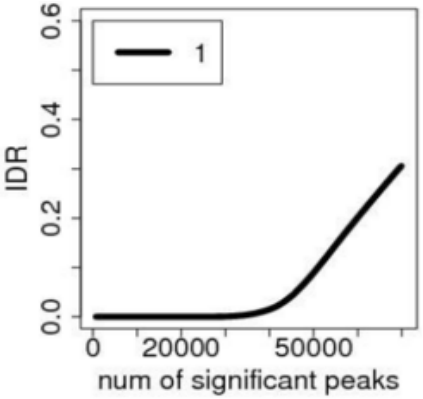
# High Quality Data

- ≥ Two biological replicates
- Multiple quality control measures

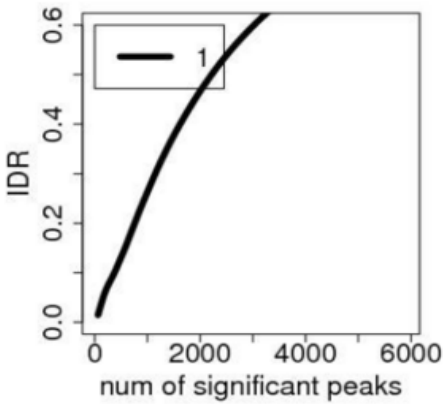



IDR Processing, QC and Blacklist Filtering

Good reproducibility

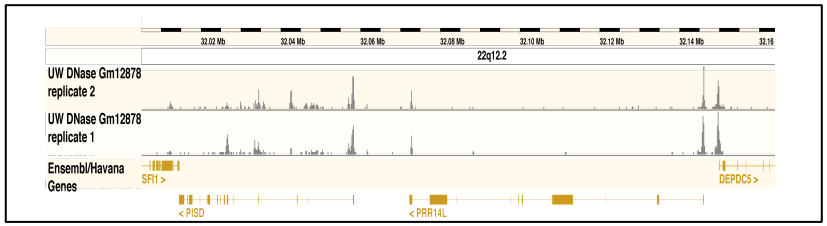



Poor reproducibility

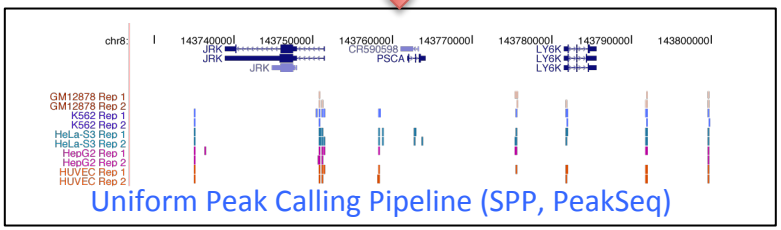


# ENCODE Uniform Analysis Pipeline

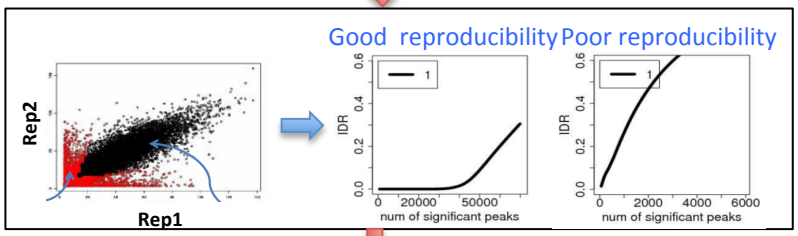
*Anshul Kundaje*



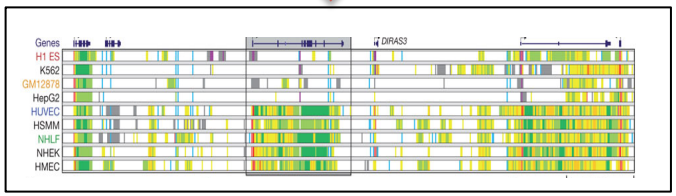

Mapped reads




Uniform peak Calling (SPP, PeakSeq)




Quality Control




Derived Data (Chromosome Segments, Expression)

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

# ENCODE Data

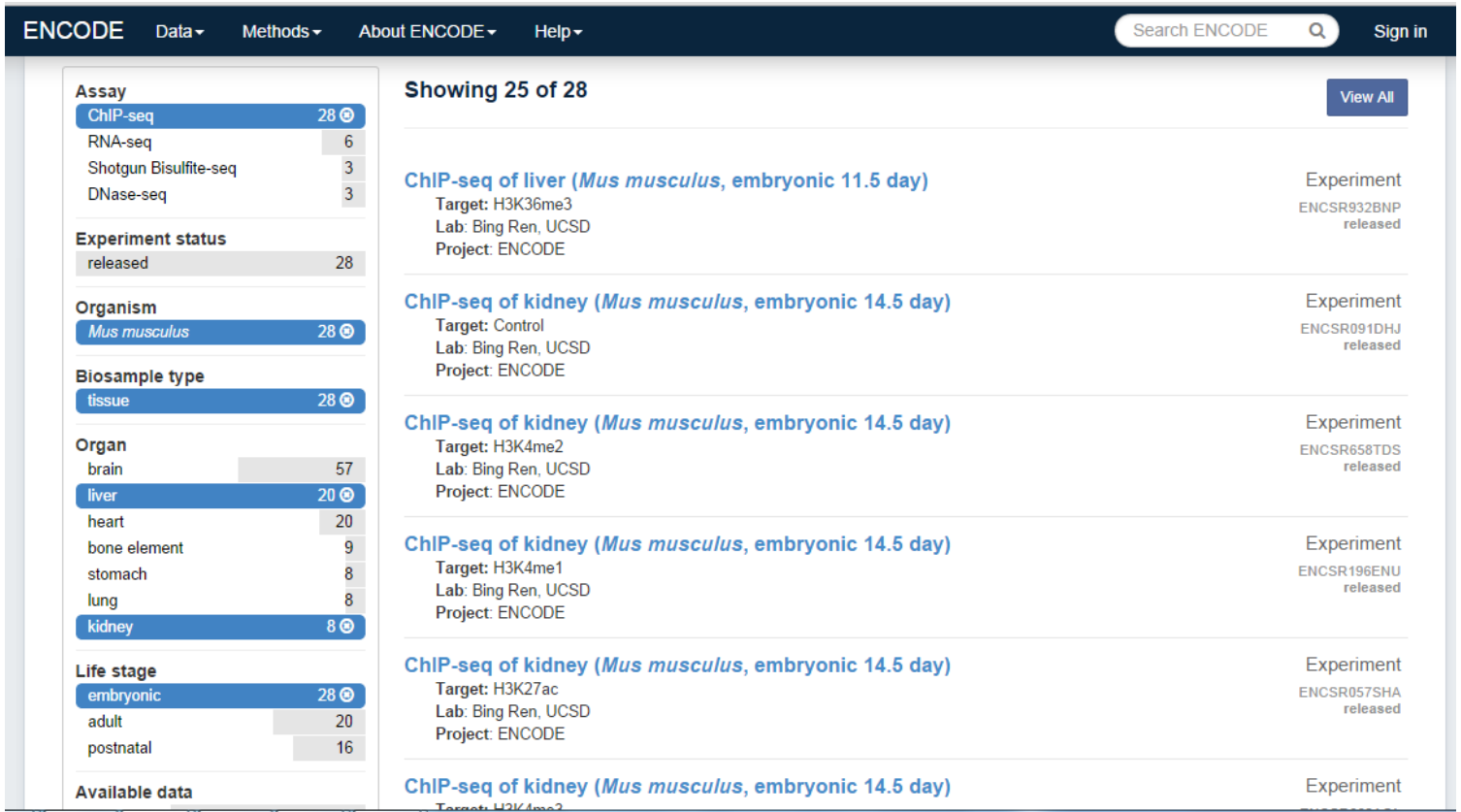
## Cloud Storage and Computing

Data available at Amazon Web Services (AWS)

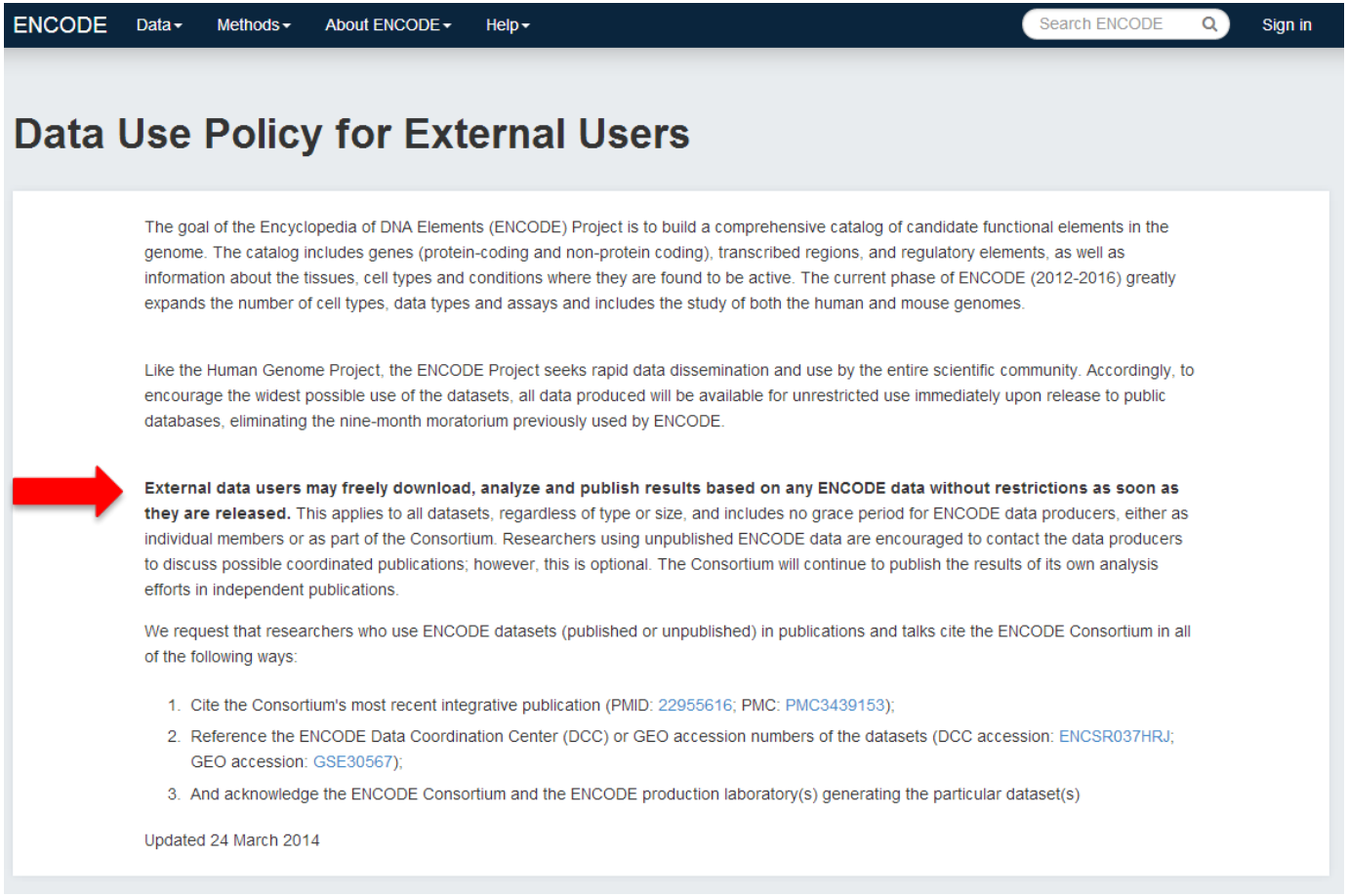
Uniform processing pipelines will be available at DNAnexus

related projects

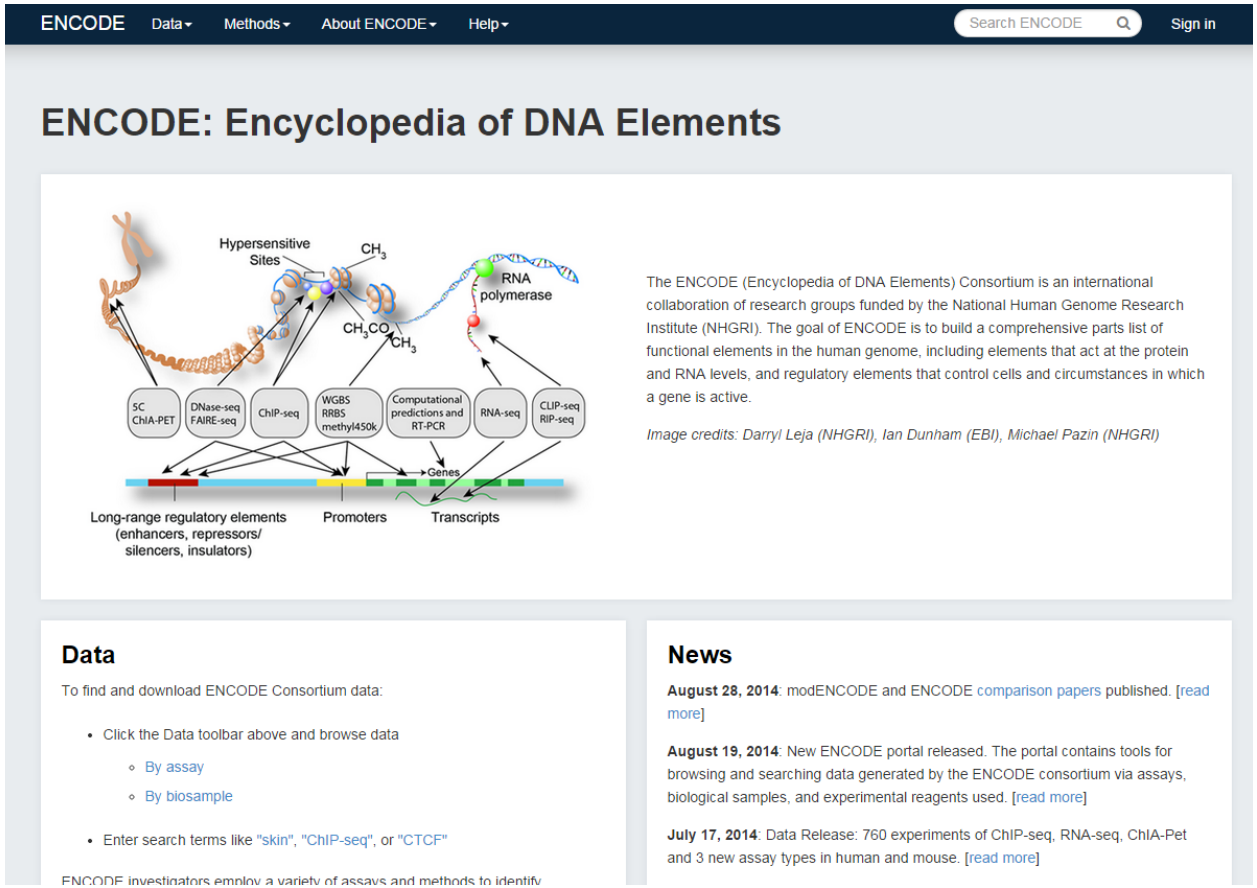
## High Searchable



# ENCODE Data Open Access



# New ENCODE Portal
# https://www.encodeproject.org



# 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