

# ENCODE Standards, Data and Access

Advanced workshop on integrative analysis using  
ENCODE and Roadmap Epigenomics data

J. Michael Cherry  
Stanford University, Department of Genetics  
October 10, 2015



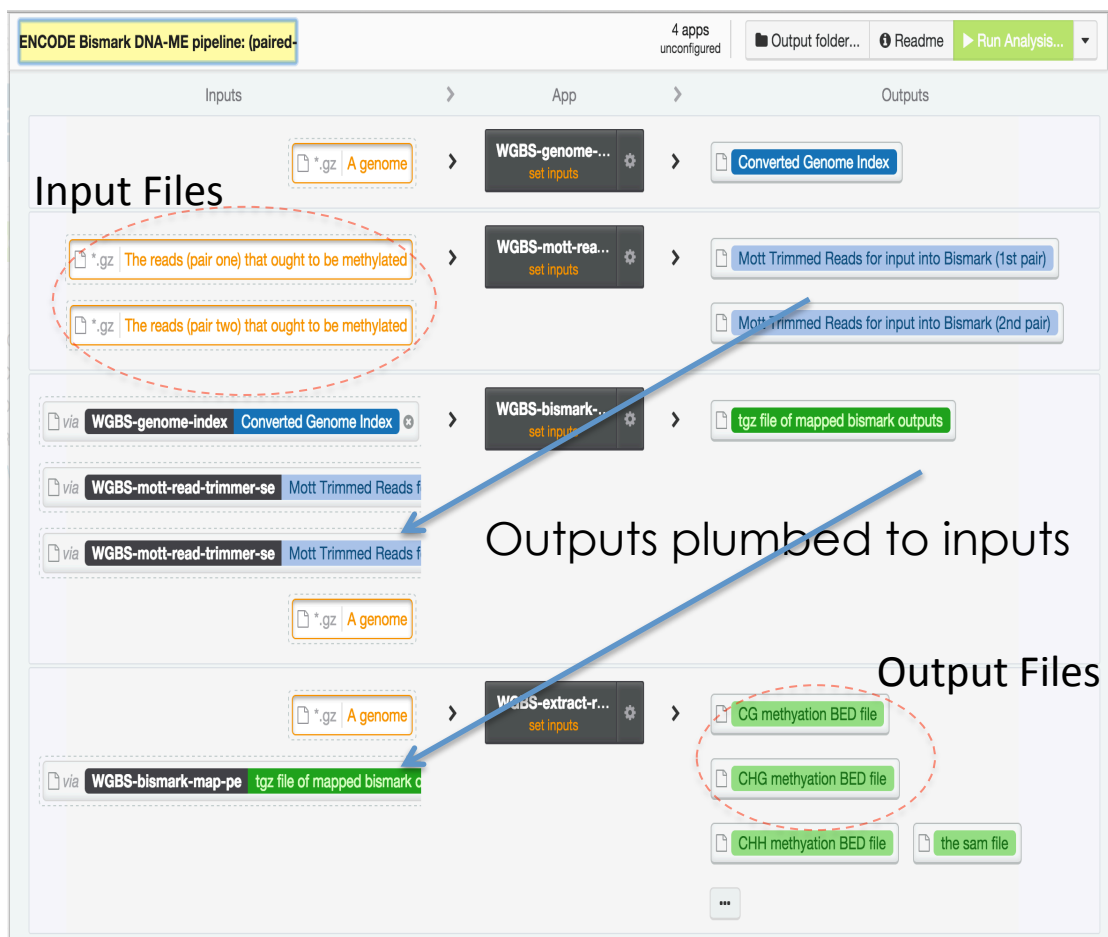
# Workshop outline

Data Access & processing → Interpretation → Visualization → Advanced Analysis

## ENCODE Portal

The screenshot shows the ENCODE Portal interface. On the left, there are filters for Assay (ChIP-seq), Experiment status (released, submitted), Organism (Mus musculus), Target of assay (histone, histone modification, control), Biosample type (tissue), Organ (brain, heart, liver), Life stage (embryonic), Available data (fastq), Read length (nt), Library insert size (nt), Library made from (DNA), Date released (June, 2015), Lab (Bing Ren, UCSF), and Project (ENCODE). The main area displays a list of 25 experiments, including ChIP-seq of neural tube (Mus musculus, embryonic 13.5 day), ChIP-seq of heart (Mus musculus, embryonic 13.5 day), and ChIP-seq of neural tube (Mus musculus, embryonic 13.5 day). Each experiment entry includes details like Target, Lab, Project, and a download button.

## ENCODE Processing Pipelines



# Workshop outline

Data Access & processing → Interpretation → Visualization → Advanced Analysis

Jill Moore: HaploReg and RegulomeDB

## rs17293632 is Associated with IBD and Crohn's Disease

Date Added to Catalog (since 11/25/08)	First Author/Date/Journal/Study	Disease/Trait	Initial Sample Description	Replication Sample Description	Region	Reported Gene(s)	Mapped Gene(s)
02/12/13	Justine L. November 01, 2012 Nature Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease.	Inflammatory bowel disease	12,924 European ancestry cases, 21,442 European ancestry controls	25,683 European ancestry cases, 17,015 European ancestry controls	15q22.33	SMAD3	SMAD3
10/19/12	Frankie A. November 21, 2010 Nat Genet Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci.	Crohn's disease	6,333 European ancestry cases, 15,056 European ancestry controls	15,694 European ancestry cases, 14,026 European ancestry controls, 414 European ancestry trios	15q22.33	SMAD3	SMAD3

chr	pos (hg18)	LD (r <sup>2</sup> )	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SIPhy cons	Promoter histone marks	Enhancer histone marks	DNase	Proteins bound	eQTL tissues changed	Motifs changed	GENCODE genes	dbSNP func annot
2	29422	0.82	rs4263140	A	G	0.48	0.13	0.20	0.09		NHEK, HMEC	4 cell types	CEBPB	7 altered motifs	9.4kb 3' of FAM110C			
2	29443	1	rs4637157	T	C	0.39	0.12	0.17	0.08		NHEK, HMEC	4 cell types	CEBPB	8 altered motifs	9.4kb 3' of FAM110C			
2	30091	0.8	rs28446791	C	G	0.47	0.13	0.20	0.09						8.7kb 3' of FAM110C			
2	31318	0.96	rs6732811	G	C	0.40	0.12	0.16	0.08						6 altered motifs	7.5kb 3' of FAM110C		
2	31324	0.96	rs6706828	C	T	0.40	0.12	0.16	0.08						Ets,ZNF263	7.5kb 3' of FAM110C		
2	31791	0.98	rs28433318	C	T	0.52	0.13	0.20	0.08		NHEK				BAF155,CHD2	7kb 3' of FAM110C		
2	38733	0.8	rs112074103	GA	G	0.47	0.13	0.20	0.09		NHEK, HMEC	Fibrobl			TATA	80bp 3' of FAM110C		
2	38340	0.8	rs4530399	A	G	0.47	0.13	0.20	0.09		HMEC, NHEK				GCNFI,Nr2f2,Zbtb3	FAM110C	3'-UTR	
2	40569	0.8	rs6731388	T	C	0.52	0.14	0.20	0.09		HMEC, NHEK	Chorion,HeLa-S3	4 bound proteins		Pou2f2,Pou8f1,Rbox11	FAM110C	3'-UTR	
2	41404	0.8	rs10173732	G	A	0.36	0.13	0.20	0.09		NHEK	HES			Spz1	FAM110C	3'-UTR	
2	50092	0.96	rs6749595	T	C	0.54	0.13	0.20	0.08						4 altered motifs	3.2kb 5' of FAM110C		
2	53652	0.96	rs4438516	G	A	0.47	0.13	0.20	0.08						7 altered motifs	6.8kb 5' of FAM110C		
2	55007	0.96	rs112988427	CAG	C	0.47	0.13	0.20	0.08						GR,NF- $\kappa$ B,TLX1,NFIC	8.1kb 5' of FAM110C		
2	55237	0.95	rs10188860	T	C	0.47	0.14	0.20	0.08						4 altered motifs	8.4kb 5' of FAM110C		
2	61687	0.98	rs10197241	A	T	0.44	0.13	0.20	0.08						4 altered motifs	15kb 5' of FAM110C		
2	66839	0.96	rs10200968	C	T	0.56	0.13	0.20	0.08		NHEK				GR	20kb 5' of FAM110C		
2	67321	0.96	rs11680031	G	A	0.56	0.13	0.20	0.08		K562	HMEC, NHEK			Ets,GR	20kb 5' of FAM110C		
2	70074	0.95	rs300761	A	G	0.56	0.14	0.20	0.08		NHEK, HMEC	Jurkat,PreC	STAT1		Myc,Sox	23kb 5' of FAM110C		

# Workshop outline

Data Access & processing → Interpretation → Visualization → Advanced Analysis

Yanli: Element and 3D Browser

HOME MOUSE HUMAN DOWNLOAD LINKS CONTACT

Query human ENCODE data!

**Option 1: Search gene expression across ~ 60 human cell types (total 108 datasets)**

Human (hg19) : Gene name(Sox2, Nanog ...)

**Option 2: Search cis-elements in a given genomic region**

Human (hg19) : chr1 : start:  end:

**Option 3: search cis-elements surrounding a gene**

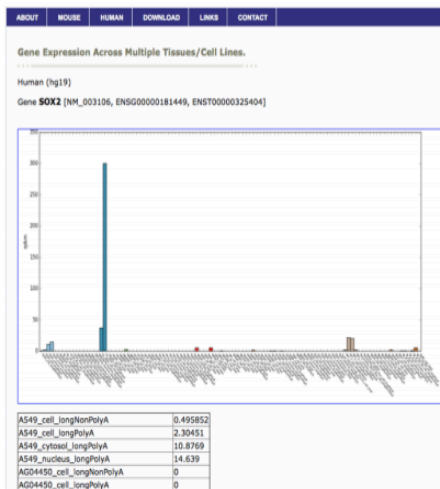
Human (hg19) : Gene name(Sox2, Nanog ...)

Extended region (default +/- 100kb)  kb

**Option 4: search cis-elements LINKED to a gene based on DNaseI H3S specificity**

Human (hg19) : Gene name(Sox2, Nanog ...)

You will be re-directed to the following result page.

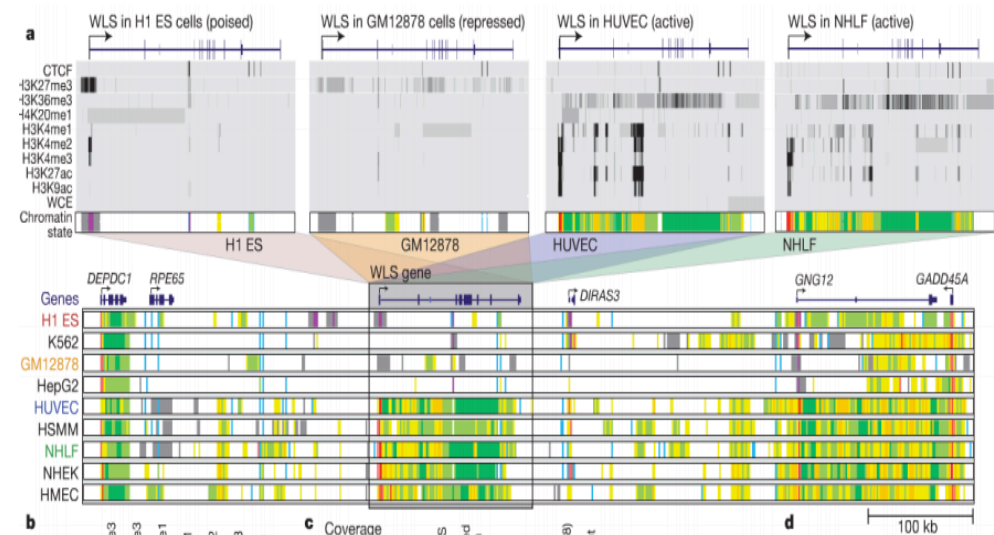


Gene ID

Gene expression  
in bar graph

Gene expression  
(in RKPM)

Jason Ernst: Genome annotation ChromHMM





# Goals for this section

- Equip you with the knowledge and tools to explore ENCODE on your own
  - Learn how to find and download data
  - Learn about the features of the portal
  - Learn about the rich metadata available that describes and contextualizes data
  - Learn about the ENCODE uniform processing pipelines

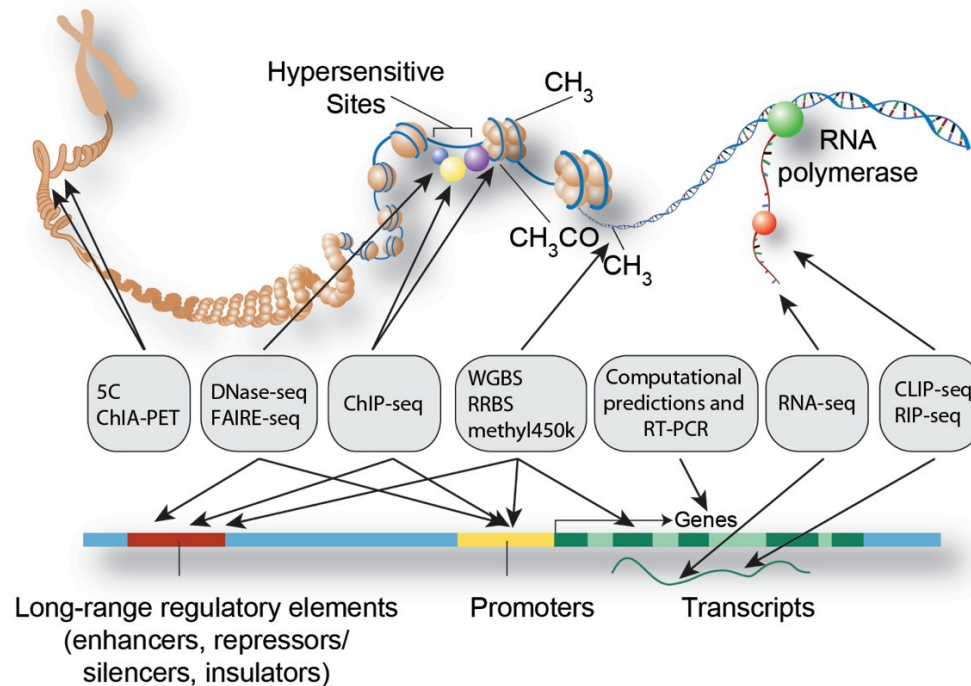
# ENCODE has produced a lot of data

## ENCODE

From 2003 to present  
>5000 experiments  
>3500 biosamples  
Hundreds of terabytes of files

## Roadmap Epigenomics

>3127 experiments (ChIP-seq,  
Dnase-seq, RNA-seq and  
bisulfite-seq  
>1200 biosamples



# Nature feature: Challenges in irreproducible research



# Nature feature: Challenges in irreproducible research

## FEATURES

---



### **Reproducibility crisis: Blame it on the antibodies**

Antibodies are the workhorses of biological experiments, but they are littering the field with false findings. A few evangelists are pushing for change.

*Nature* (19 May 2015)

antibodies



### **Statistical errors**

*P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

*Nature* (12 February 2014)

quality control &  
measures of confidence



### **Replication studies: Bad copy**

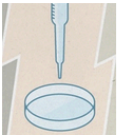
In the wake of high-profile controversies, psychologists are facing up to problems with replication.

*Nature* (16 May 2012)

replication in experimental  
design

## NEWS AND ANALYSIS

---



### **Irreproducible biology research costs put at \$28 billion per year**

Study calculates cost of flawed biomedical research in the United States.

*Nature* (09 June 2015)

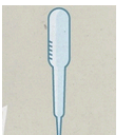


### **Sluggish data sharing hampers the reproducibility effort**

Initiative trying to validate 50 cancer papers finds difficulty in accessing original study data.

*Nature* (03 June 2015)

data not readily accessible



### **Researchers argue for standard format to cite lab resources**

Research Resource Identifier (RRID) aims to clean up poorly referenced data.

*Nature* (29 May 2015)

Need for standard and  
unique identifiers

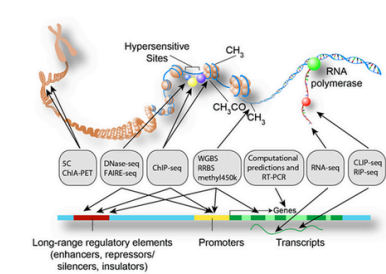
# ENCODE portal

## www.encodeproject.org

- Central source for ENCODE data: experimental and analysis data
- Hub for project information: data standards & publications
- High-quality metadata: data provenance & transparency

**ENCODE** Data - Methods - About ENCODE - Help -

### 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)*

#### Quick Start

To find and download ENCODE Consortium data:

- Click the Data toolbar above and browse data
  - [By assay](#)
  - [By biosample](#)
  - [By genomic annotations](#)
- Enter search terms like "skin", "ChIP-seq", or "CTCF"

Additional help about the ENCODE Portal:

- [Getting Started](#)
- [Batch Download](#)

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.

All ENCODE data is freely available for download and analysis. Please refer to the [ENCODE Data Release Policy](#)

#### News

**UPDATED:** The presentation video and tutorials from the [ENCODE 2015: Research Applications and Users Meeting](#) have been posted for public use.

**June 29 to July 1, 2015:** The ENCODE 2015: Research Applications and Users Meeting was held at the [Bolger Center](#) in Potomac, MD

**June 23, 2015:** Data release: 3 human and 91 mouse datasets. [\[read more\]](#)

**May 28, 2015:** Changelog released for metadata schema updates. [\[read more\]](#)

**May 18, 2015:** Data release: 12 human datasets. [\[read more\]](#)

**May 13, 2015:** 519 publications that use ENCODE data, published by authors not funded by ENCODE, added to the ENCODE Portal [\[read more\]](#)

**April 13, 2015:** Data release: 28 human datasets. [\[read more\]](#)

**March 31, 2015:** Data release: 4 human datasets. [\[read more\]](#)

**March 11, 2015:** Experiment pages have been updated to show a graphical display of the pipeline used to generate the processed files associated with that experiment. [\[read more\]](#)

**March 9, 2015:** Batch Download of files released. [\[read more\]](#)

**February 12, 2015:** Data release: 1 human and 3 mouse datasets. [\[read more\]](#)

See [news archive](#) for additional news and updates.

# The ENCODE portal at a glance

The image shows the ENCODE portal homepage with several red arrows pointing to specific features and text labels. The portal has a top navigation bar with 'ENCODE', 'Data', 'Methods', 'About ENCODE', and 'Help' menus. A search bar is in the top right. The main content area is divided into sections: 'Assays', 'Data standards', 'Project overview', 'Getting started', and a large text block about the consortium. At the bottom, there are 'Quick Start' and 'News' sections. A central diagram illustrates the genomic context of the data, showing regulatory elements, promoters, and transcripts.

**ENCODE: Encyclopedia of DNA Elements**

**Navigation Menu:**

- Assays
  - Biosamples
  - Antibodies
  - Annotations
  - Release policy
- Data
  - Data standards
  - Software tools
  - Pipelines
  - Experiment guidelines
- Methods
  - Project overview
  - News
  - Publications
  - Release policy
  - Data access
- About ENCODE
  - Getting started
  - REST API
  - File formats
  - Tutorials
  - Contact
- Help

**Search**

**Tutorials & guides**

**Project information & publications**

**Find data**

**Software, pipelines & experiment guidelines**

**Quick help**

**Recent news**

**Quick Start**

To find and download ENCODE Consortium data:

**News**

**UPDATED:** The agenda has been posted for the [First ENCODE Users Meeting](#) will be held at the [Baylor Center](#) in Potomoc, MD from June 29 - July 1, 2015.

The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NIH). The goal of ENCODE is to build a comprehensive map 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.

Diagram illustrating the genomic context of the data, showing regulatory elements (Long-range regulatory elements, enhancers, repressors/silencers, insulators), Promoters, Transcripts, and Genes.

# Metadata-driven faceted browsing: by assay

Select “Assay” under the “Data” menu:

Facets for filtering

ENCODE

Data

Methods

About ENCODE

Help

Search ENCODE

Assay

ChIP-seq2546

RNA-seq698

DNase-seq357

shRNA knockdown followed by RNA-seq333

transcription profiling by array assay180

+ See more...

Experiment status

released5169

revoked25

Genome assembly (visualization)

hg193114

mm9571

mm10236

dm3108

Organism

Homo sapiens3827

Mus musculus1096

Drosophila melanogaster197

Target of assay

transcription factor1219

histone927

histone modification900

control520

RNA binding protein516

+ See more...

Biosample type

immortalized cell line2924

tissue941

primary cell821

stem cell211

in vitro differentiated cells167

Showing 25 of 5194 experiments

Filter to 500 to visualize

Download

View All

RNA Bind-n-Seq

Target: Input library control

Lab: Chris Burge, MIT

Project: ENCODE

Experiment

ENCSR944GMN

released

shRNA knockdown followed by RNA-seq of K562 (Homo sapiens, adult 53 year)

Target: DDX55

Lab: Brenton Graveley, UConn

Project: ENCODE

Experiment

ENCSR856CJK

released

shRNA knockdown followed by RNA-seq of K562 (Homo sapiens, adult 53 year)

Target: Non-specific target control

Lab: Brenton Graveley, UConn

Project: ENCODE

Experiment

ENCSR667PLJ

released

shRNA knockdown followed by RNA-seq of K562 (Homo sapiens, adult 53 year)

Target: Non-specific target control

Lab: Brenton Graveley, UConn

Project: ENCODE

Experiment

ENCSR572FFX

released

shRNA knockdown followed by RNA-seq of HepG2 (Homo sapiens, child 15 year)

Target: Non-specific target control

Lab: Brenton Graveley, UConn

Project: ENCODE

Experiment

ENCSR376RJN

released

shRNA knockdown followed by RNA-seq of HepG2 (Homo sapiens, child 15 year)

Target: SRSF5

Lab: Brenton Graveley, UConn

Project: ENCODE

Experiment

ENCSR781YNI

released

shRNA knockdown followed by RNA-seq of HepG2 (Homo sapiens, child 15 year)

Target: DAZAP1

Lab: Brenton Graveley, UConn

Project: ENCODE

Experiment

ENCSR220TBR

released



# Metadata-driven faceted browsing: by assay

ENCODE

Data ▾

Methods ▾

About ENCODE ▾

Help ▾

Search ENCODE 🔍

Assay

ChIP-seq2546

RNA-seq698

DNase-seq357

shRNA knockdown followed by RNA-seq333

transcription profiling by array assay180

+ See more...

Experiment status

released5169

revoked25

Genome assembly (visualization)

hg193114

mm9571

mm10236

dm3108

Organism

Homo sapiens3827

Mus musculus1096

Drosophila melanogaster197

Target of assay

transcription factor1219

histone927

histone modification900

control520

RNA binding protein516

+ See more...

Biosample type

immortalized cell line2924

tissue941

primary cell821

stem cell211

in vitro differentiated cells167

Showing 25 of 5194 experiments

Filter to 500 to visualize 🔗

Download

View All

RNA Bind-n-Seq

Target: Input library control

Lab: Chris Burge, MIT

Project: ENCODE

Experiment ENCSR944GMN released

shRNA knockdown followed by RNA-seq of K562 (Homo sapiens, adult 53 year)

Target: D/

Lab: Bren

Project: E

Experiment

Assay

ChIP-seq6 🔍

Experiment status

released6

Genome assembly (visualization)

hg196

Organism

Homo sapiens6 🔍

Mus musculus2

Target of assay

histone17

histone modification16

transcription factor6 🔍

control5

Biosample type

primary cell5

tissue1

Organ

blood vessel18

skin of body13

lung6 🔍

mammary gland4

brain3

Showing 6 of 6 experiments

Visualize 🔗

Download

ChIP-seq of fibroblast of lung (Homo sapiens)

Target: CTCF

Lab: John Stamatoyannopoulos, UW

Project: ENCODE

Experiment ENCSR000DWY released

ChIP-seq of fibroblast of lung (Homo sapiens, fetal 12 week)

Target: CTCF

Lab: John Stamatoyannopoulos, UW

Project: ENCODE

Experiment ENCSR000DPM released

ChIP-seq of fibroblast of lung (Homo sapiens)

Target: CTCF

Lab: John Stamatoyannopoulos, UW

Project: ENCODE

Experiment ENCSR000DVA released

ChIP-seq of fibroblast of lung (Homo sapiens)

Target: CTCF

Lab: Bradley Bernstein, Broad

Project: ENCODE

Experiment ENCSR000ANO released

ChIP-seq of fibroblast of lung (Homo sapiens)

Target: EZH2

Lab: Bradley Bernstein, Broad

Project: ENCODE

Experiment ENCSR000ARO released



# Main units of an experimental analysis are uniquely accessioned

Experiments – ENCSR###XXX  
Biosamples – ENCBS###XXX  
Donors/strains – ENCDO###XXX  
Libraries – ENCLB###XXX  
Antibody lots – ENCAB###XXX  
Files – ENCFF###XXX

**ENCODE** Data Methods About ENCODE Help Search ENCODE

## Experiment summary for ENCSR000ANO

Status: released

Assay:	ChIP-seq
Accession:	ENCSR000ANO
Biosample summary:	fibroblast of lung ( <i>Homo sapiens</i> )
Type:	primary cell
Target:	CTCF
Antibody:	ENCAB000AXY
Controls:	ENCSR000AMZ
Description:	CTCF ChIP-seq on human NHLF
Lab:	Bradley Bernstein, Broad
Award PI:	Bradley Bernstein, Broad
Project:	ENCODE
External resources:	<a href="#">UCSC-ENCODE-hg19:wgEncodeEH000120</a> <a href="#">GEO:GSM733695</a>
Date released:	2011-02-10

### Assay details

Nucleic acid type:	DNA
Lysis method:	see document
Extraction method:	see document
Fragmentation method:	see document
Size selection method:	see document
Platform:	Illumina Genome Analyzer IIe

# Nature feature: Challenges in irreproducible research

## FEATURES



### **Reproducibility crisis: Blame it on the antibodies**

Antibodies are the workhorses of biological experiments, but they are littering the field with false findings. A few evangelists are pushing for change.

*Nature* (19 May 2015)

antibodies



### **Statistical errors**

*P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

*Nature* (12 February 2014)

quality control &  
measures of confidence



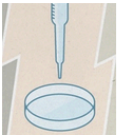
### **Replication studies: Bad copy**

In the wake of high-profile controversies, psychologists are facing up to problems with replication.

*Nature* (16 May 2012)

replication in experimental  
design

## NEWS AND ANALYSIS



### **Irreproducible biology research costs put at \$28 billion per year**

Study calculates cost of flawed biomedical research in the United States.

*Nature* (09 June 2015)

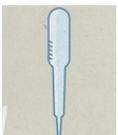


### **Sluggish data sharing hampers the reproducibility effort**

Initiative trying to validate 50 cancer papers finds difficulty in accessing original study data.

*Nature* (03 June 2015)

✓ data not readily accessible



### **Researchers argue for standard format to cite lab resources**

Research Resource Identifier (RRID) aims to clean up poorly referenced data.

*Nature* (29 May 2015)

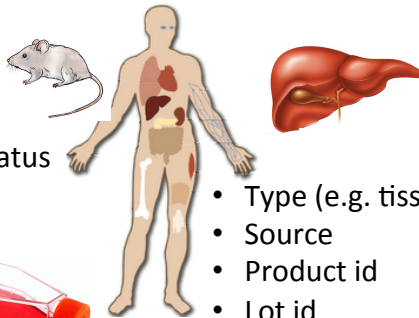
Need for standard and  
unique identifiers

# Rich experimental metadata is collected and presented for clarity and context

For example:

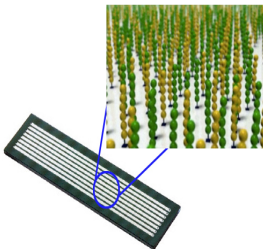
## Donor & biosample

- Species
- Age
- Sex
- Health status
- Ethnicity
- Strain



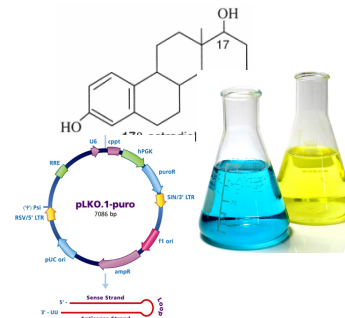
- Type (e.g. tissue, cell line)
- Source
- Product id
- Lot id
- Dates (e.g. growth, harvest, procurement)
- Passage number
- Starting amount
- Lab assigned IDs

## Platform



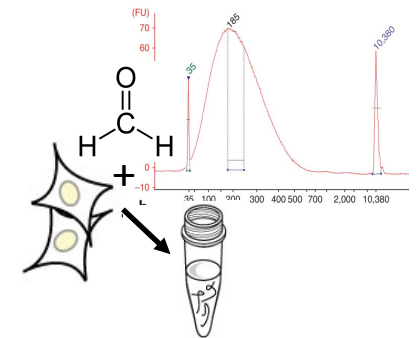
- Instrument
- Read length
- Single or Paired end
- Lane number
- Sequencing depth

## Treatment & genetic modifications



- Agent (chemical, biological)
- Concentration
- Duration
- Construct type
- Tag
- Tag location
- Insert sequence
- Target
- Transfection type
- Protocol

## Library preparation



- Lysis method
- Sonication method
- Extraction method
- Nucleic acid type
- Nucleic acid size range
- Library preparation protocol
- Strand specificity
- Size selection method
- Validation document

# Replication and transparency of methods

<https://www.encodeproject.org/experiments/ENCSR000ANO/>

ENCODE ChIP-seq pipeline ENCODE Interesting stuff Mortgage stuff Celniker

ENCODE Data Methods About ENCODE Help Search ENCODE

## Biological replicate - 1

Technical replicate: 1  
Library: ENCLB695AMN  
Biosample: [ENCBS339AAA](#) - fibroblast of lung

Most ENCODE experiments are designed to minimally have two replicates.

## Biological replicate - 2

Technical replicate: 1  
Library: ENCLB695AMO  
Biosample: [ENCBS339AAA](#) - fibroblast of lung

## Documents

**General protocol**

**Description excerpt:**  
The whole cell extract library preparation protocol used by the Bernstein lab in their ChIP-seq...

[whole\\_cell\\_extract\\_library\\_construction\\_v1.0.pdf](#)

[Visualize Data](#)

## Files linked to ENCSR000ANO

### Raw data

Accession	File type	Biological replicate	Technical replicate	Read length	Run type	Lab	Date added
<a href="#">ENCFF000CQU</a> <a href="#">Download</a> 636 MB	fastq	2	1		single-ended	Bradley Bernstein, Broad	2010-11-16
<a href="#">ENCFF000CQV</a> <a href="#">Download</a>	fastq	1	1		single-ended	Bradley Bernstein, Broad	2010-11-16

# Nature feature: Challenges in irreproducible research

## FEATURES



### Reproducibility crisis: Blame it on the antibodies

Antibodies are the workhorses of biological experiments, but they are littering the field with false findings. A few evangelists are pushing for change.

*Nature* (19 May 2015)

antibodies



### Statistical errors

*P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

*Nature* (12 February 2014)

quality control &  
measures of confidence



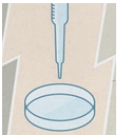
### Replication studies: Bad copy

In the wake of high-profile controversies, psychologists are facing up to problems with replication.

*Nature* (16 May 2012)

✓ replication in experimental design

## NEWS AND ANALYSIS



### Irreproducible biology research costs put at \$28 billion per year

Study calculates cost of flawed biomedical research in the United States.

*Nature* (09 June 2015)

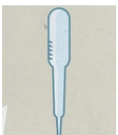


### Sluggish data sharing hampers the reproducibility effort

Initiative trying to validate 50 cancer papers finds difficulty in accessing original study data.

*Nature* (03 June 2015)

✓ data not readily accessible



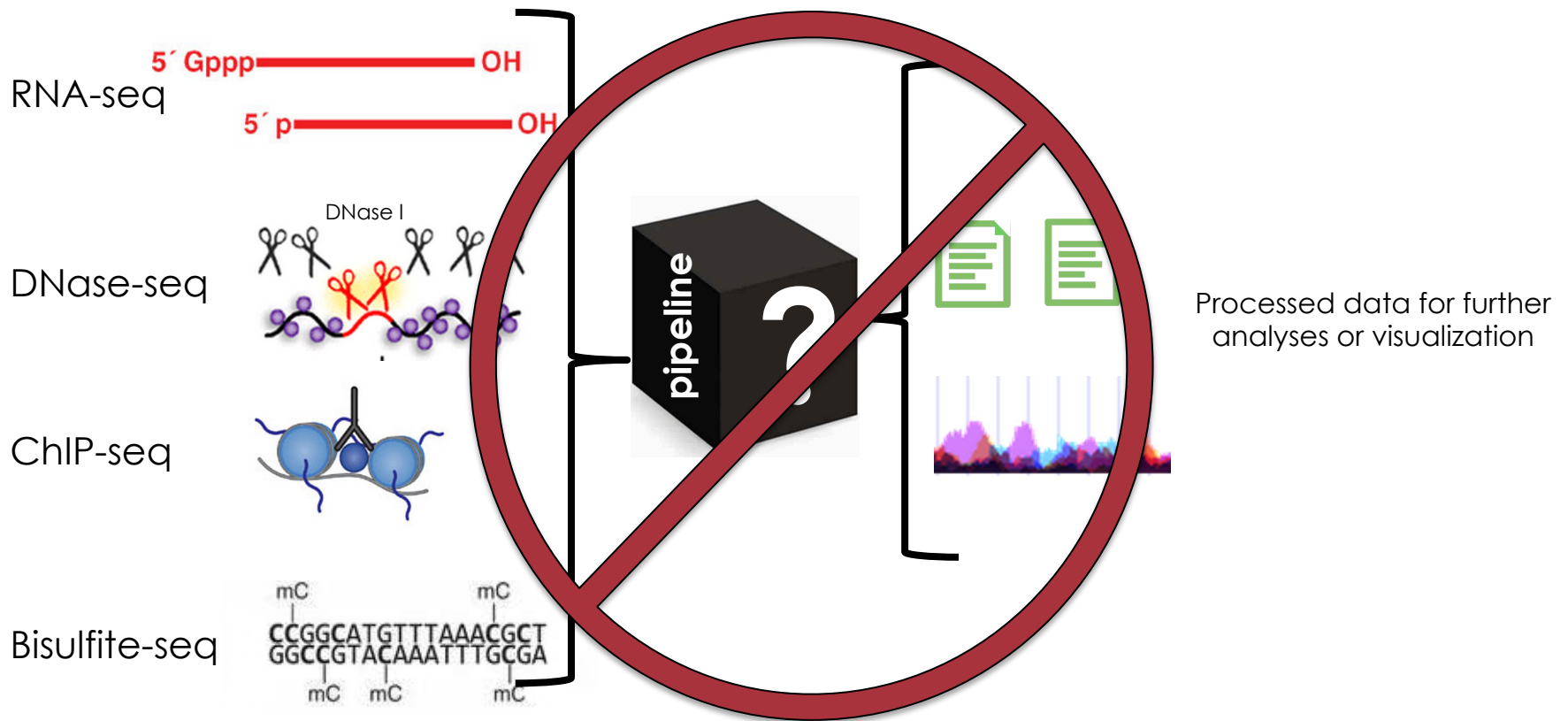
### Researchers argue for standard format to cite lab resources

Research Resource Identifier (RRID) aims to clean up poorly referenced data.

*Nature* (29 May 2015)

✓ Need for standard and unique identifiers

# Data provenance & process transparency



Avoid the pipeline blackbox

# ENCODE runs uniform processing pipelines to enhance data comparability & replicability

## RNA-seq of long RNAs (single-end, unstranded)

Status: active

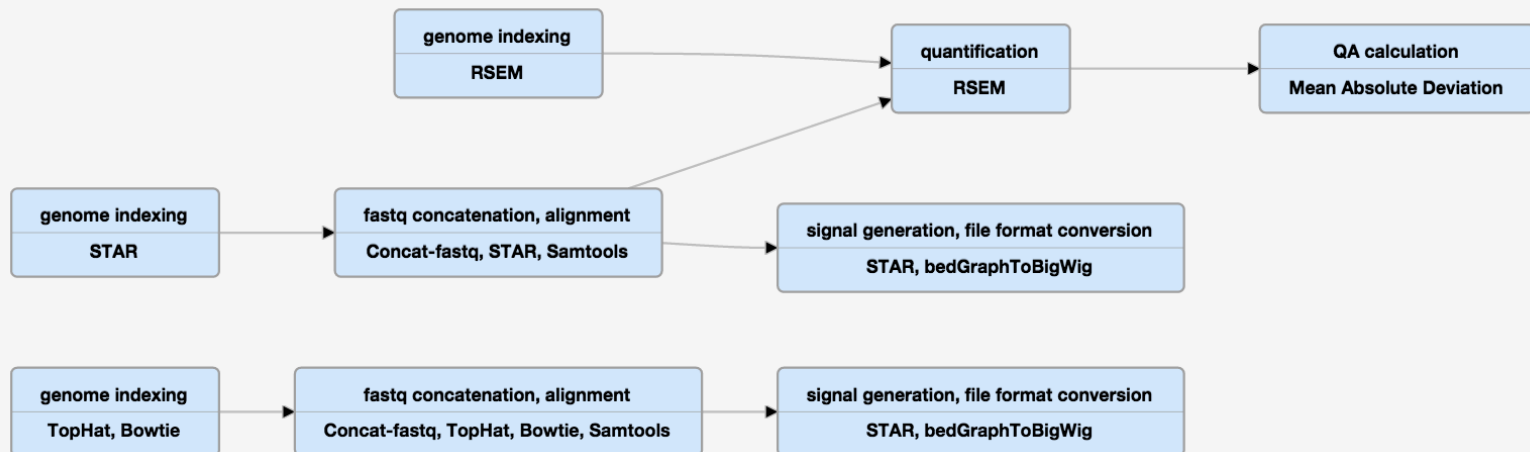
**Title:** [RNA-seq of long RNAs \(single-end, unstranded\)](#) 

**Assay:** RNA-seq

**Description:** Pipeline for single-ended unstranded data long RNA-seq data as developed by the ENCODE RNA Working Group

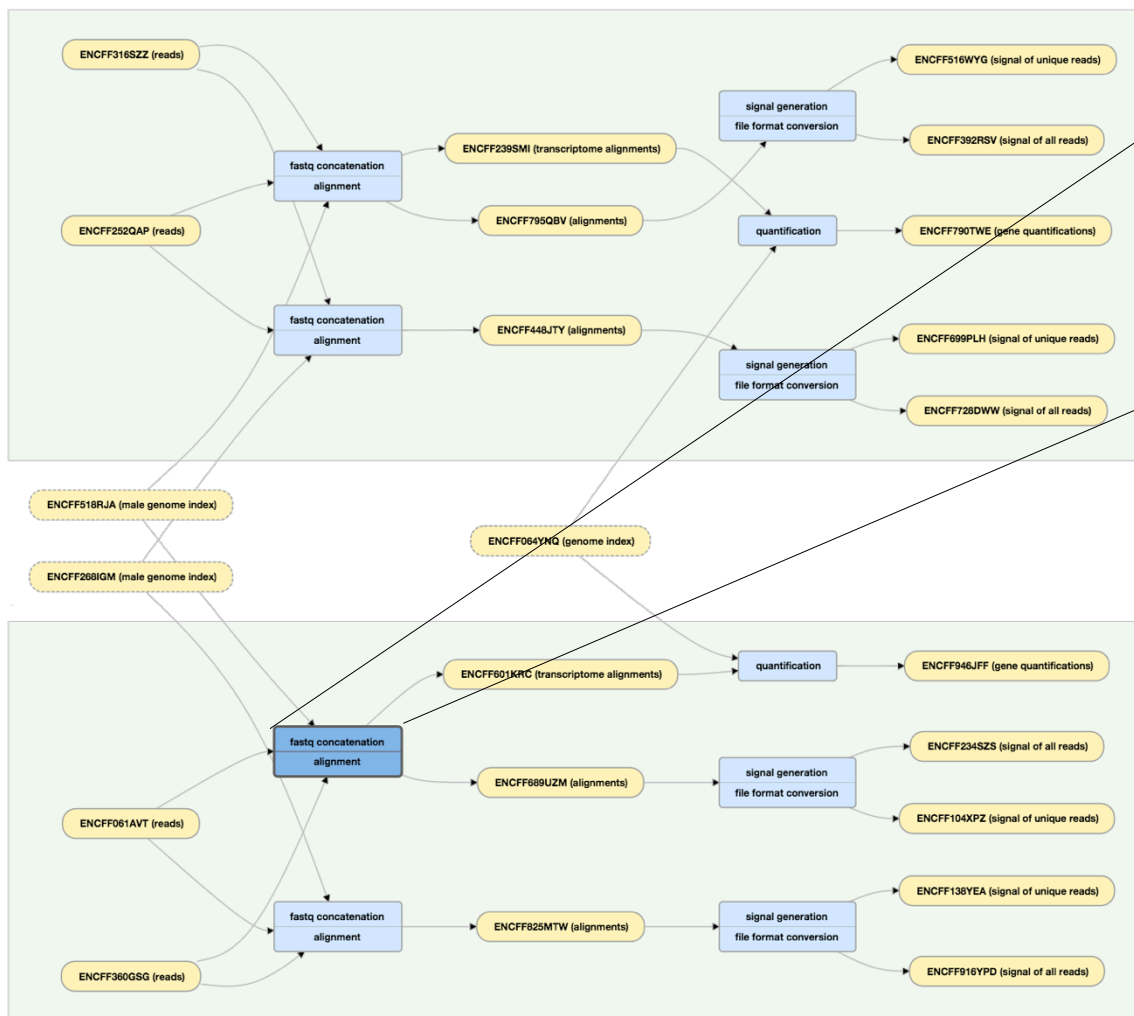
**Lab:** ENCODE Processing Pipeline

### Pipeline schematic



Download Graph

# Each pipeline run is tracked and viewable on the portal



Name:	Long RNA-seq STAR single-ended alignment
Step type:	fastq concatenation, alignment
Step aliases:	dnanexus:align-star-se-v-1
Input:	reads
Output:	alignments
Pipeline:	RNA-seq of long RNAs (single-end, unstranded)
Software:	<div>Concat-fastq 1.0.2</div> <div>STAR 2.4.0h</div> <div>Samtools</div>

- ENCODE pipelines are deployed on the cloud via **DNAnexus**
- Users can access and use the same pipelines on their own data to compare with ENCODE data
- Coming soon: display of calculated QC metrics

Single-end long RNA-seq pipeline run for experiment ENCSR823VEE



# Nature feature: Challenges in irreproducible research

## FEATURES



### Reproducibility crisis: Blame it on the antibodies

Antibodies are the workhorses of biological experiments, but they are littering the field with false findings. A few evangelists are pushing for change.

*Nature* (19 May 2015)

antibodies



### Statistical errors

*P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

*Nature* (12 February 2014)

✓ quality control & measures of confidence



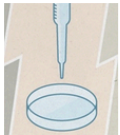
### Replication studies: Bad copy

In the wake of high-profile controversies, psychologists are facing up to problems with replication.

*Nature* (16 May 2012)

✓ replication in experimental design

## NEWS AND ANALYSIS



### Irreproducible biology research costs put at \$28 billion per year

Study calculates cost of flawed biomedical research in the United States.

*Nature* (09 June 2015)

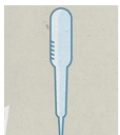


### Sluggish data sharing hampers the reproducibility effort

Initiative trying to validate 50 cancer papers finds difficulty in accessing original study data.

*Nature* (03 June 2015)

✓ data not readily accessible



### Researchers argue for standard format to cite lab resources

Research Resource Identifier (RRID) aims to clean up poorly referenced data.

*Nature* (29 May 2015)

✓ Need for standard and unique identifiers

# Each antibody lot is characterized & accessioned

## Experiment summary for ENCSR993OLA

Status: released Validation: pending

Assay: eCLIP

Accession: ENCSR993OLA

Biosample summary: HepG2 (*Homo sa*

Type: immortalized cell

Target: IGF2BP3

Antibody: ENCAB934MDN

Controls: ENCSR077KVG

Description: eCLIP experiment

Lab: Gene Yeo, UCSD

Award PI: Brenton Graveley

Project: ENCODE

Aliases: gene-yeo:211

Date released: 2015-07-15

### ENCAB934MDN

Antibody against *Homo sapiens* IGF2BP3

*Homo sapiens*

K562

Eligible for new data 

Source (vendor): MBL 

Product ID: RN009 

Lot ID: 002

Targets: IGF2BP3 (*Homo sapiens*)

Host: Rabbit

Clonality: Polyclonal

Isotype: IgA

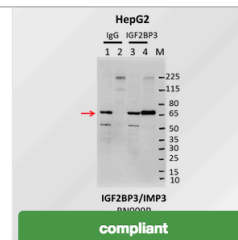
Antigen description: KLH-conjugated synthetic peptide HQQQKALQSGPPQSRRK (563-579 aa)

#### IGF2BP3 (*Homo sapiens*)

Method:  
immunoprecipitation

##### Caption excerpt:

IP-Western Blot analysis of HepG2 whole cell lysate using IGF2BP3 specific antibody. Lane 1 is 1% of twenty million whole cell lysate input and lane 2 is 25% of IP enrichment using rabbit normal IgG...



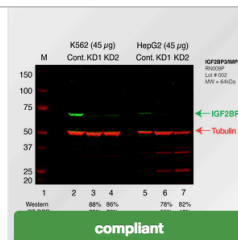
More

#### IGF2BP3 (*Homo sapiens*)

Method:  
knockdown or knockout

##### Caption excerpt:

Western blot following shRNA against IGF2BP3 in K562 and HepG2 whole cell lysate using IGF2BP3 specific antibody. Lane 1 is a ladder, lane 2 is K562 non-targeting control knockdown, lane 3 and 4 are...



More

## Assay details

Nucleic acid type: RNA

Lysis method: see document

Extraction method: see document

Fragmentation method: see document

Size range: 175-300

Size selection method: agarose gel extraction

# Nature feature: Challenges in irreproducible research

## FEATURES

---

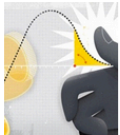


### Reproducibility crisis: Blame it on the antibodies

Antibodies are the workhorses of biological experiments, but they are littering the field with false findings. A few evangelists are pushing for change.

*Nature* (19 May 2015)

✓ antibodies



### Statistical errors

*P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

*Nature* (12 February 2014)

✓ quality control & measures of confidence



### Replication studies: Bad copy

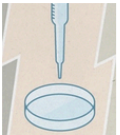
In the wake of high-profile controversies, psychologists are facing up to problems with replication.

*Nature* (16 May 2012)

✓ replication in experimental design

## NEWS AND ANALYSIS

---



### Irreproducible biology research costs put at \$28 billion per year

Study calculates cost of flawed biomedical research in the United States.

*Nature* (09 June 2015)

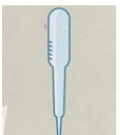


### Sluggish data sharing hampers the reproducibility effort

Initiative trying to validate 50 cancer papers finds difficulty in accessing original study data.

*Nature* (03 June 2015)

✓ data not readily accessible



### Researchers argue for standard format to cite lab resources

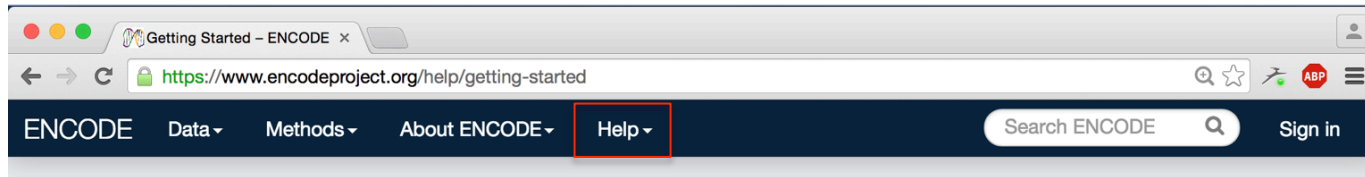
Research Resource Identifier (RRID) aims to clean up poorly referenced data.

*Nature* (29 May 2015)

✓ Need for standard and unique identifiers

# For more details, there's help pages, tutorials, exercises and more

Learn about  
programmatic  
access via the  
REST API



## Getting Started

### Introduction

Welcome to the ENCODE Consortium and up  
This site is develop  
consortium are sub  
is needed to view r

This document des  
[downloading data](#),  
data can be visuali

Please contact the  
further questions.

### Informa

The ENCODE Port

- Raw and pro
- [Biological sa](#)
- [Antibody cha](#)

ENCODE Data Methods About ENCODE Help

## Tutorials & Workshops

### Upcoming tutorials and workshops

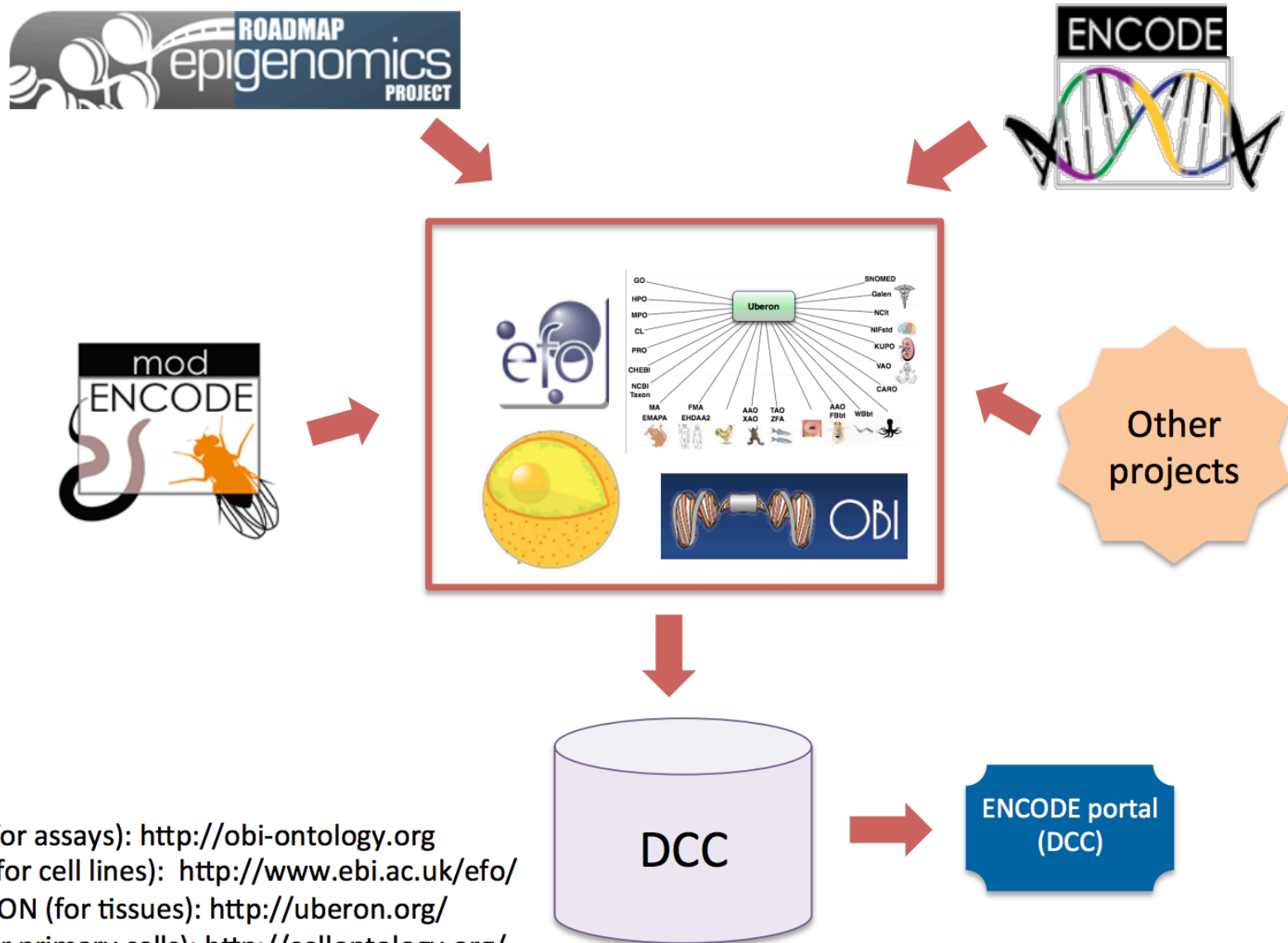
- ENCODE will be at the American Society of Human Genetics' 2015 Annual Meeting from October 6th to October 10th in Baltimore, MD, presenting the [Advanced Workshop on Integrative Analysis using ENCODE and Roadmap Epigenomics Data](#).

### Workshop materials

- [Tutorials and video](#) from the ENCODE 2015: Research Applications and Users Meeting at the [Bolger Center](#) in Potomoc, MD, June 29 - July 1, 2015
  - Video and workshop materials from hands-on tutorial sessions on accessing, processing, analyzing, and utilizing ENCODE data and resources, along with presentations from leading experts in disease, biology, and computational fields explaining how they employ ENCODE resources in their work.

```
GET_object.py
1  #!/usr/bin/env python
2
3  import requests
4
5  URL = 'https://www.encodeproject.org/experiments/ENCSR236EGS/?format=json'
6
7  response = requests.get(URL)
8
9  experiment = response.json()
10
11 print experiment['accession']
12 print experiment['description']
13
```

# Metadata integration using ontologies



# The ENCODE DCC



Mike Cherry (PI)



Ben Hitz



Cricket Sloan



STANFORD  
SCHOOL OF MEDICINE



Esther Chan



Jean Davidson



Idan Gabdank



Seth Strattan



Marcus Ho



Aditi Narayanan



Tim Dreszer



Marissa Melen



Nikhil Podduturi



Laurence Rowe



Forrest Tanaka



Stuart Miyasato



Matt Simison



Zhenhua Wang



@encodedcc



encode-help@lists.stanford.edu



<https://github.com/ENCODE-DCC/>