



ENCODE DATA AND METADATA THROUGH THE ENCODE PORTAL

J. Seth Strattan, ENCODE DCC, Stanford

Keystone Symposia Epigenomics & Methylation

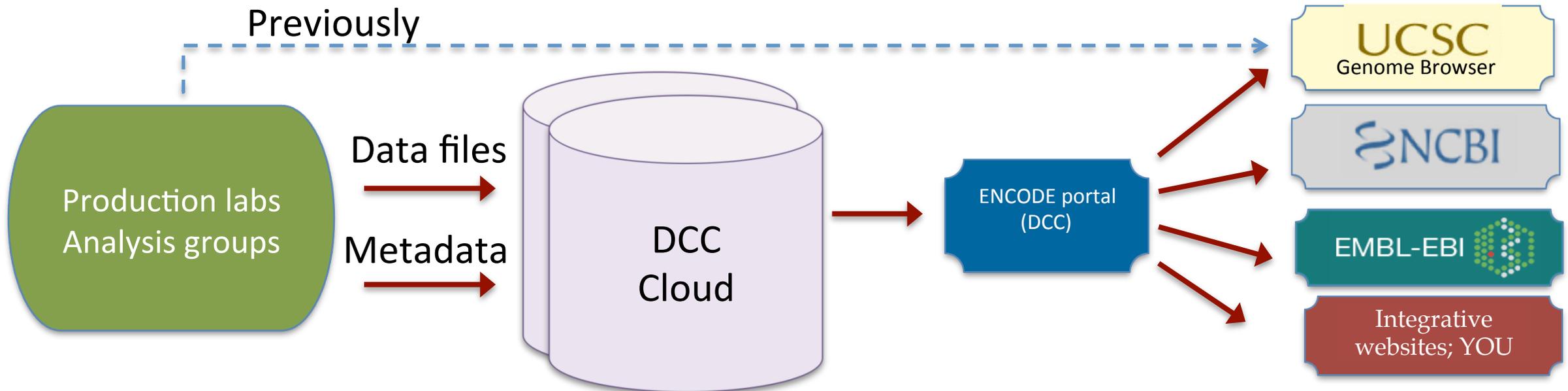
ENCODE Workshop

April, 2015



ENCODE Data Coordinating Center (DCC)

ENCODE data flow



Role: Data generation
Tasks: Perform assays & analyses
 Validate data
 Submit data and metadata

Data organization
 Data processing & validation
 Data file storage
 Metadata curation

Data access
 Web-based searches
 Data downloads



What would you like to learn?

How many of you:

1. ... have downloaded ENCODE data and intersected it with other data?
2. ... know where to go for a comprehensive catalog of all assays done by ENCODE?
3. ... could repeat an ENCODE analysis (from fastq's) to generate IDR-thresholded sets of peaks?
4. ... want to repeat one of the ENCODE analysis pipelines on your data?
5. ... need to access ENCODE data but found it difficult or don't know where to begin?



ENCODE Data Coordinating Center (DCC)

Challenges posed by large, diverse sets of data like ENCODE:

- Data can be **difficult to find**.
 - Different datatypes are in different places.
 - Browser tracks, ftp sites, web sites.
- Metadata describing experiments and analyses can be **hard to search**.
- Where is the “**official**” **list of everything** ENCODE has done?
- Can I just **download** all the ChIP data on ZNF143?



encodeproject.org: Why are we building it?

- encodeproject.org is the ENCODE Data Coordinating Center (DCC)'s new portal application.
 - Curated source for all ENCODE metadata and data.
 - Updated continuously.
 - Easy to find the data you care about ...
 - ... and maybe discover data you didn't even know about.

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 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.

News

Sept 12, 2014: Data release: 23 human and 5 mouse datasets. [\[read more\]](#)

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

June 16, 2014: Visualize tracks on the UCSC Genome Browser via trackhubs-on-the-fly. [\[read more\]](#)

March 17, 2014: Antibody characterization standard updated to include antibodies against chromatin regulators and RNA binding-

encodeproject.org: Features of the Portal

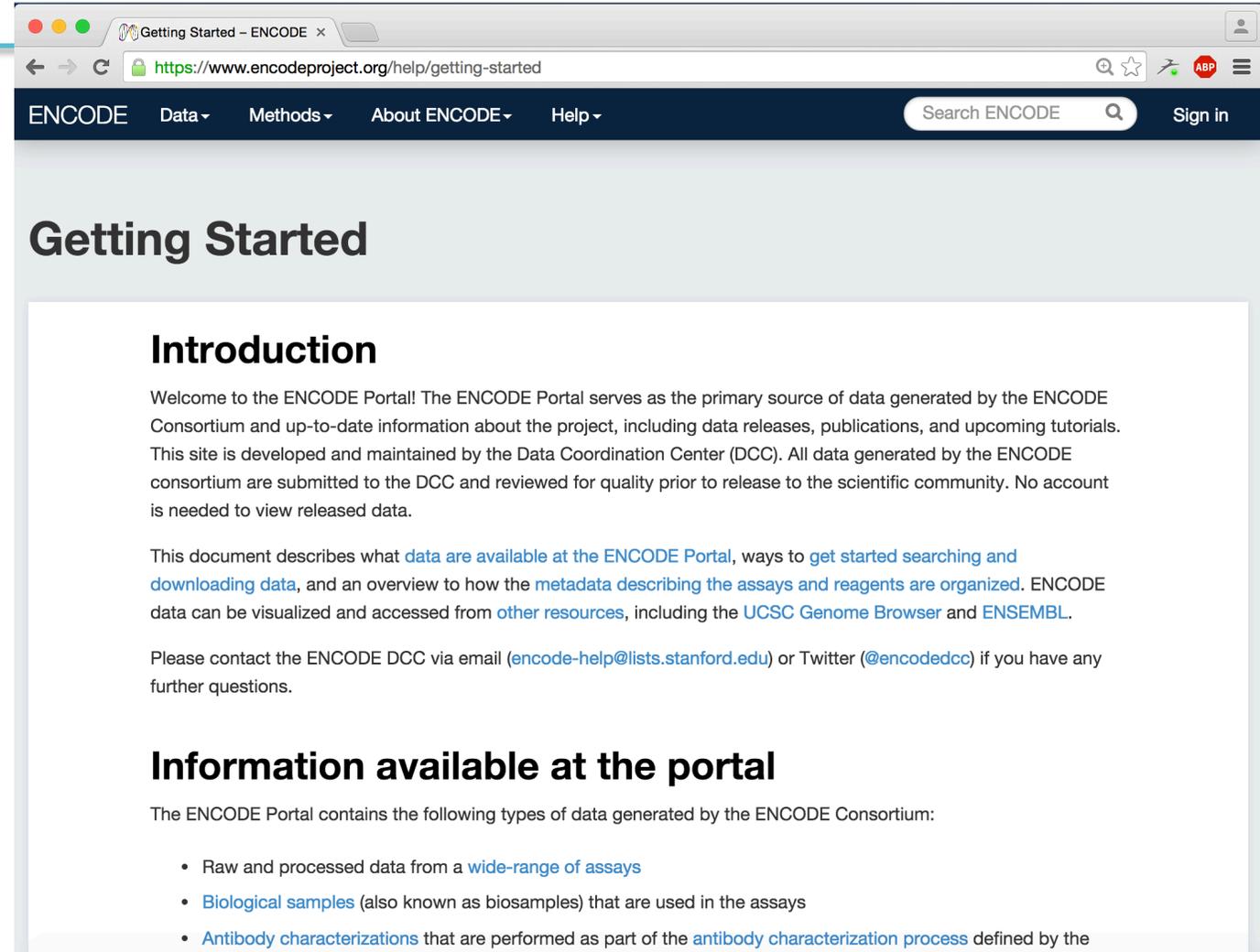
Topics for this workshop

1. **Organization** of the ENCODE Portal at <https://www.encodeproject.org/>
2. **Faceted browsing** of ENCODE, including search
3. **Batch download** of search results
4. **ENCODE annotations** through the Portal
5. **Preview** of ENCODE **analysis pipelines** on the cloud
6. **Programmatic access** using the ENCODE REST API



Getting Started

- <https://www.encodeproject.org/>
- <https://www.encodeproject.org/help/getting-started>
 - A good place to start
- <https://www.encodeproject.org/tutorials>
 - Links to slides from today's workshop
- Help ... Contact
 - encode-help@lists.stanford.edu
 - [@encodedcc](https://twitter.com/encodedcc)



The screenshot shows a web browser window with the URL <https://www.encodeproject.org/help/getting-started>. The page has a dark blue header with the ENCODE logo and navigation menus for Data, Methods, About ENCODE, and Help. A search bar and a 'Sign in' link are also present. The main content area is titled 'Getting Started' and contains an 'Introduction' section. The introduction text reads: 'Welcome to the ENCODE Portal! The ENCODE Portal serves as the primary source of data generated by the ENCODE Consortium and up-to-date information about the project, including data releases, publications, and upcoming tutorials. This site is developed and maintained by the Data Coordination Center (DCC). All data generated by the ENCODE consortium are submitted to the DCC and reviewed for quality prior to release to the scientific community. No account is needed to view released data.' Below this, it states: 'This document describes what [data are available at the ENCODE Portal](#), ways to [get started searching and downloading data](#), and an overview to how the [metadata describing the assays and reagents are organized](#). ENCODE data can be visualized and accessed from [other resources](#), including the [UCSC Genome Browser](#) and [ENSEMBL](#).' It concludes with contact information: 'Please contact the ENCODE DCC via email (encode-help@lists.stanford.edu) or Twitter ([@encodedcc](https://twitter.com/encodedcc)) if you have any further questions.' The 'Information available at the portal' section follows, stating: 'The ENCODE Portal contains the following types of data generated by the ENCODE Consortium:' and lists three bullet points: 'Raw and processed data from a [wide-range of assays](#)', '[Biological samples](#) (also known as biosamples) that are used in the assays', and '[Antibody characterizations](#) that are performed as part of the [antibody characterization process](#) defined by the'.

Faceted Browsing

Faceted Browsing of Assays

- Click on Data ... Assays
- <https://www.encodeproject.org/search/?type=experiment>
- Facets on the left are “filters”
- Limits to items of interest
- Facets from the same category can be stacked to combine results (logical OR)
- Facets from different categories further restrict results (logical AND)

The screenshot shows the ENCODE project search interface. The browser address bar displays the URL: https://www.encodeproject.org/search/?type=experiment&assay_term_name=ChIP-seq&replicates.library.biosample.dono.... The page title is "Search - ENCODE". The navigation bar includes "ENCODE", "Data", "Methods", "About ENCODE", and "Help". A search bar contains "Search ENCODE" and a "Sign in" button.

The main content area is divided into two columns. The left column displays facets for filtering results:

- Assay**: ChIP-seq (67)
- Experiment status**: released (67)
- Genome assembly (visualization)**: hg19 (64)
- Organism**: *Homo sapiens* (67), *Mus musculus* (6)
- Target of assay**: histone (153), histone modification (143), transcription factor (67), control (53), other context (1)
- Biosample type**: immortalized cell line (857), stem cell (71), primary cell (67), in vitro differentiated cells (11), tissue (4)

The right column displays the search results, showing 25 of 67 items. Each result includes the assay name, target, lab, and project. The results are:

- ChIP-seq of astrocyte of the spinal cord (*Homo sapiens*)**: Target: CTCF, Lab: John Stamatoyannopoulos, UW, Project: ENCODE, Experiment ENCSR000DSU released.
- ChIP-seq of choroid plexus epithelial cell (*Homo sapiens*)**: Target: CTCF, Lab: John Stamatoyannopoulos, UW, Project: ENCODE, Experiment ENCSR000DTL released.
- ChIP-seq of erythroblast (*Homo sapiens*)**: Target: POLR2A, Lab: Peggy Farnham, USC, Project: ENCODE, Experiment ENCSR000EXO released.
- ChIP-seq of erythroblast (*Homo sapiens*)**: Target: GATA1, Lab: Peggy Farnham, USC, Project: ENCODE, Experiment ENCSR000EXP released.
- ChIP-seq of mammary epithelial cell (*Homo sapiens*)**: Target: CTCF, Lab: John Stamatoyannopoulos, UW, Project: ENCODE, Experiment ENCSR000DUS released.

Buttons for "Visualize", "Download", and "View All" are visible at the top right of the results section.

Search

Search across ENCODE metadata

- In search box type “skin”
<https://www.encodeproject.org/search/?searchTerm=skin>
- Results include experiments, biosamples, etc
- Select experiments – see several assays
- Facet further by RNA-seq
- Note this is not just a pure text search
- “skin of body” comes up, but so does “keratinocyte”
- Search traverses ontology relationships

The screenshot shows a web browser window with the URL <https://www.encodeproject.org/search/?searchTerm=skin&type=experiment>. The page displays search results for the term "skin".

Assay

ChIP-seq	53
RNA-seq	26
DNase-seq	25
RNA profiling by array assay	24
CAGE	9

+ See more...

Experiment status

released	170
----------	-----

Genome assembly (visualization)

hg19	147
mm9	2

Organism

<i>Homo sapiens</i>	166
<i>Mus musculus</i>	2

Target of assay

histone	31
histone modification	29
transcription factor	13
control	9

Biosample type

primary cell	163
tissue	5

Showing 25 of 170 [Visualize] [Download] [View All]

RRBS of zone of skin (*Homo sapiens*, adult 83 year) Experiment
Lab: Richard Myers, HAIB
Project: ENCODE
ENCSR000DEF released

RNA-seq of skin of body (*Homo sapiens*, fetal) Experiment
Lab: Thomas Gingeras, CSHL
Project: ENCODE
ENCSR000AFG released

RNA-seq of skin of body (*Homo sapiens*, fetal) Experiment
Lab: Thomas Gingeras, CSHL
Project: ENCODE
ENCSR000AGA released

RAMPAGE of skin of body (*Homo sapiens*, fetal) Experiment
Lab: Thomas Gingeras, CSHL
Project: ENCODE
ENCSR000AGU released

DNA methylation profiling by array assay of zone of skin (*Homo sapiens*, adult 83 year) Experiment
Lab: Richard Myers, HAIB
Project: ENCODE
ENCSR000BWZ released

CAGE of melanocyte of skin (*Homo sapiens*, adult) Experiment
Lab: Piero Carninci, RIKEN
Project: ENCODE
ENCSR000CKZ released

Metadata to Data

Use metadata to find data

- Select Data ... Assays
- Facet on RNAseq; mouse; mm10 assembly
- Select an experiment, for example
<https://www.encodeproject.org/experiments/ENCSR236EGS/>
- Note metadata on protocols, replicates
- Graph: files are related by processing steps
- Download from the graph or a list
- Not all experiments have the graph yet, but files are always available from the list

The screenshot shows a web browser window displaying the ENCODE project website. The URL in the address bar is <https://www.encodeproject.org/experiments/ENCSR236EGS/>. The page title is "Experiment summary for ENCSR236EGS". The status is "released" (green) and "Validation: pending" (orange). The assay is "RNA-seq". The accession is "ENCSR236EGS". The biosample summary is "cerebral cortex, layer 5 (*Mus musculus*, adult 8 month)". The type is "tissue". The description is "RNA-seq on a dissected area of layer V from an 8 month old male wild type C57Bl6 mouse". The lab is "Barbara Wold, Caltech". The project is "ENCODE". The date released is "2014-12-17". The assay details section shows "Nucleic acid type: RNA", "Fragmentation method: Illumina/Nextera tagmentation", and "Size range: >200".

Assay:	RNA-seq
Accession:	ENCSR236EGS
Biosample summary:	cerebral cortex, layer 5 (<i>Mus musculus</i> , adult 8 month)
Type:	tissue
Description:	RNA-seq on a dissected area of layer V from an 8 month old male wild type C57Bl6 mouse
Lab:	Barbara Wold, Caltech
Project:	ENCODE
Date released:	2014-12-17

Nucleic acid type:	RNA
Fragmentation method:	Illumina/Nextera tagmentation
Size range:	>200

Batch Download

Maybe you care about several experiments

- Select Data ... Assays
- Facet on ChIP-seq; human; transcription factor; in vitro differentiated cells
- Click Download; then Download again
- Open files.txt: a list of links to all the files for those experiments
- Transfer this file to your server and use `xargs -n 1 curl -O -L < files.txt`

The screenshot shows a web browser window with the URL `https://www.encodeproject.org/search/?type=experiment&assay_term_name=RNA-seq&replicates.library.biosample.donor...`. The ENCODE website navigation bar is visible at the top. A modal dialog box titled "Using batch download" is open, providing instructions on how to download a "files.txt" file. The dialog box includes a terminal command: `xargs -n 1 curl -O -L < files.txt`. Below the dialog box, a table of search results is visible, showing details for several RNA-seq experiments.

Assay	Count	Experiment Name	Lab	Project	Status
primary cell	5	RNA-seq of Purkinje cell (<i>Mus musculus</i> , adult 8 month)	Barbara Wold, Caltech	ENCODE	released
immortalized cell line	4				
Organ					
brain	16				
liver	5	RNA-seq of liver (<i>Mus musculus</i> , embryonic 11.5 day)	Barbara Wold, Caltech	ENCODE	released
heart	4				
bone element	1				
kidney	1	RNA-seq of forebrain (<i>Mus musculus</i> , embryonic 11.5 day)	Barbara Wold, Caltech	ENCODE	released
Life stage					
embryonic	22				
postnatal	13	RNA-seq of G1E-ER4 (<i>Mus musculus</i> , postnatal 0 day)			
adult	8				

Batch Download – files.txt

```
1 |https://www.encodeproject.org/metadata/type=experiment&assay_term_name=ChIP-seq&target.
   |investigated_as=transcription%20factor&replicates.library.biosample.donor.organism.
   |scientific_name=Homo%20sapiens&replicates.library.biosample.
   |biosample_type=in%20vitro%20differentiated%20cells/metadata.tsv
2 |https://www.encodeproject.org/files/ENCFF002ELS/@download/ENCFF002ELS.fastq.gz
3 |https://www.encodeproject.org/files/ENCFF002ELT/@download/ENCFF002ELT.fastq.gz
4 |https://www.encodeproject.org/files/ENCFF002ELU/@download/ENCFF002ELU.fastq.gz
5 |https://www.encodeproject.org/files/ENCFF002ELV/@download/ENCFF002ELV.fastq.gz
6 |https://www.encodeproject.org/files/ENCFF002DTV/@download/ENCFF002DTV.fastq.gz
7 |https://www.encodeproject.org/files/ENCFF002EHT/@download/ENCFF002EHT.fastq.gz
8 |https://www.encodeproject.org/files/ENCFF002EHV/@download/ENCFF002EHV.fastq.gz
```

- Design goal is to allow easy web-based faceted browsing and search and download of large datasets to a server
- files.txt is meant to be used with something like `xargs -n 1 curl -O -L < files.txt`
- The first file is always the metadata that tells what each file is
- Documented here: <https://www.encodeproject.org/help/batch-download/>
- Want speed? Forget wget. Get wgot. <https://github.com/ENCODE-DCC/wgot>



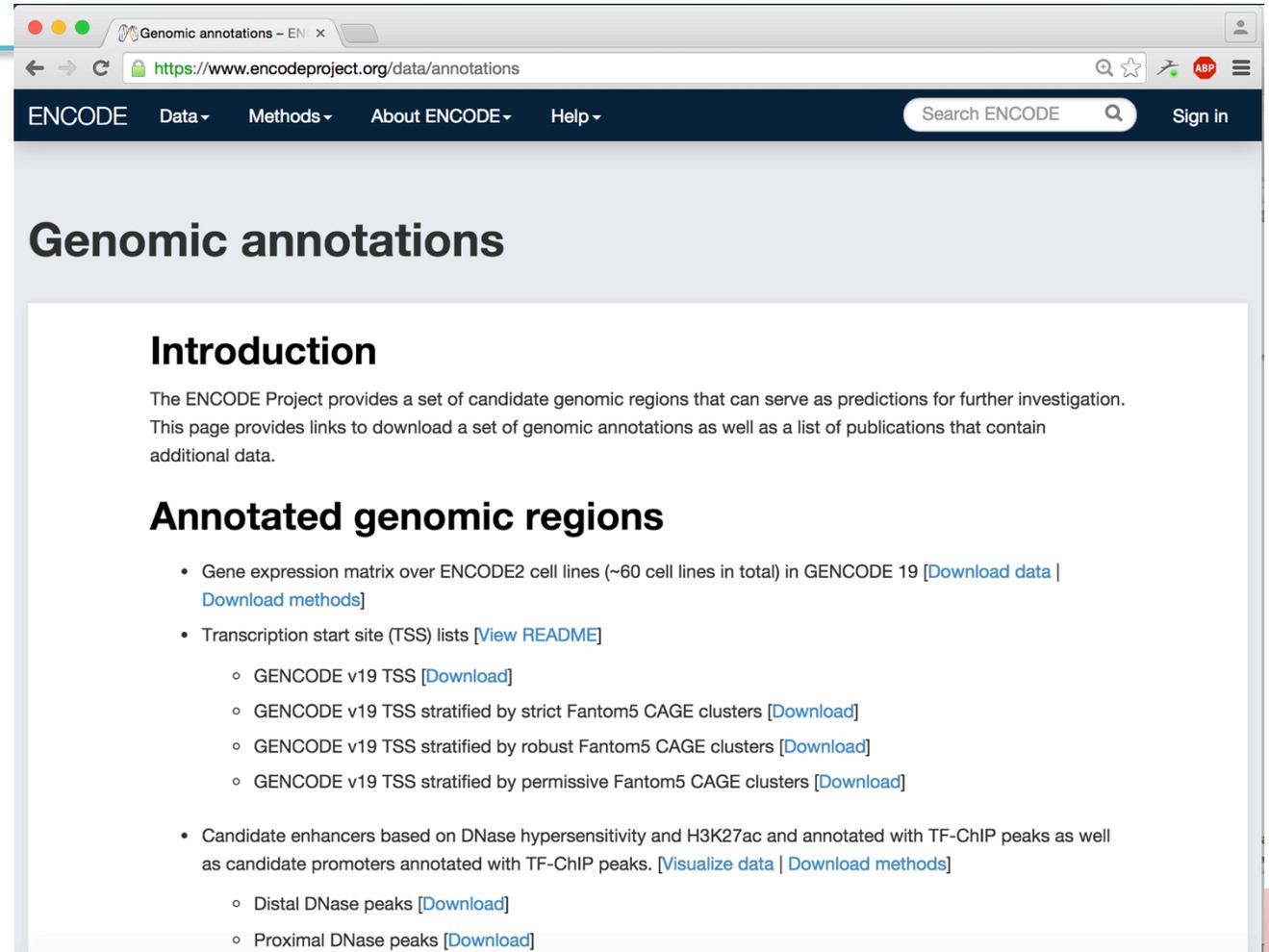
Annotations

Move on from experiments to annotations

- Select Data ... Annotations

<https://www.encodeproject.org/data/annotations>

- Links to region-based annotations (e.g. TSS's)
- Links to enhancer predictions and methods
- Links to integrative ENCODE publications that feature annotation datasets



The screenshot shows a web browser window with the URL <https://www.encodeproject.org/data/annotations>. The page title is "Genomic annotations". The navigation bar includes "ENCODE", "Data", "Methods", "About ENCODE", and "Help", along with a search bar and a "Sign in" button. The main content area has a heading "Genomic annotations" and an "Introduction" section. The introduction states: "The ENCODE Project provides a set of candidate genomic regions that can serve as predictions for further investigation. This page provides links to download a set of genomic annotations as well as a list of publications that contain additional data." Below this is a section titled "Annotated genomic regions" with a bulleted list of links to various datasets.

Genomic annotations

Introduction

The ENCODE Project provides a set of candidate genomic regions that can serve as predictions for further investigation. This page provides links to download a set of genomic annotations as well as a list of publications that contain additional data.

Annotated genomic regions

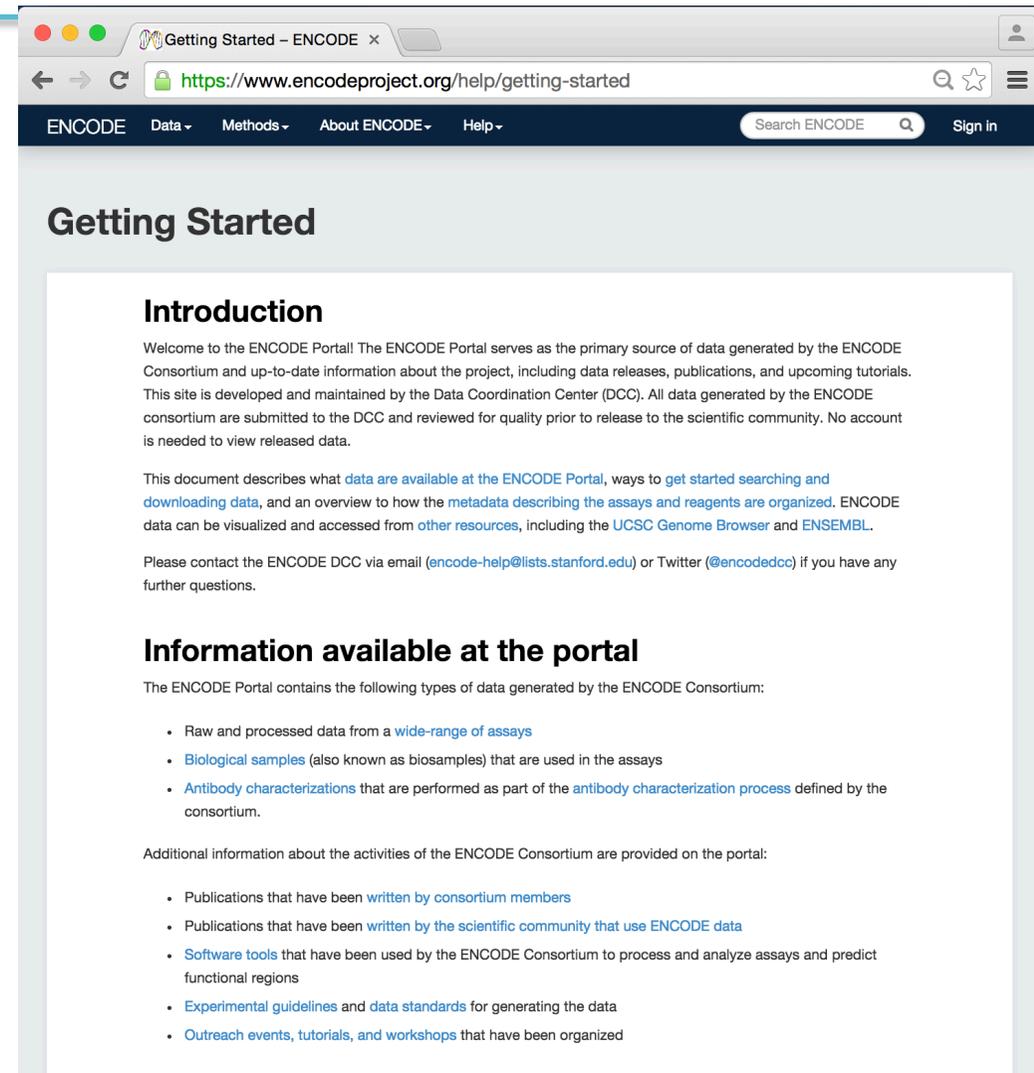
- Gene expression matrix over ENCODE2 cell lines (~60 cell lines in total) in GENCODE 19 [[Download data](#) | [Download methods](#)]
- Transcription start site (TSS) lists [[View README](#)]
 - GENCODE v19 TSS [[Download](#)]
 - GENCODE v19 TSS stratified by strict Fantom5 CAGE clusters [[Download](#)]
 - GENCODE v19 TSS stratified by robust Fantom5 CAGE clusters [[Download](#)]
 - GENCODE v19 TSS stratified by permissive Fantom5 CAGE clusters [[Download](#)]
- Candidate enhancers based on DNase hypersensitivity and H3K27ac and annotated with TF-ChIP peaks as well as candidate promoters annotated with TF-ChIP peaks. [[Visualize data](#) | [Download methods](#)]
 - Distal DNase peaks [[Download](#)]
 - Proximal DNase peaks [[Download](#)]



Summary of Interactive Features of the Portal

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

- Documentation and links to tutorials
- Faceted browsing: experiments and samples
- Detailed experiment metadata
- Visualize file relationships
- Interactive download of selected files
- Batch download of selected experiments
- Access to ENCODE annotations



The screenshot shows a web browser window with the URL <https://www.encodeproject.org/help/getting-started>. The page title is "Getting Started" and it features a navigation menu with "ENCODE", "Data", "Methods", "About ENCODE", and "Help". A search bar and a "Sign in" link are also visible. The main content area is titled "Introduction" and contains the following text:

Welcome to the ENCODE Portal! The ENCODE Portal serves as the primary source of data generated by the ENCODE Consortium and up-to-date information about the project, including data releases, publications, and upcoming tutorials. This site is developed and maintained by the Data Coordination Center (DCC). All data generated by the ENCODE consortium are submitted to the DCC and reviewed for quality prior to release to the scientific community. No account is needed to view released data.

This document describes what [data are available at the ENCODE Portal](#), ways to [get started searching and downloading data](#), and an overview to how the [metadata describing the assays and reagents are organized](#). ENCODE data can be visualized and accessed from [other resources](#), including the [UCSC Genome Browser](#) and [ENSEMBL](#).

Please contact the ENCODE DCC via email (encode-help@lists.stanford.edu) or Twitter ([@encodedcc](https://twitter.com/encodedcc)) if you have any further questions.

Information available at the portal

The ENCODE Portal contains the following types of data generated by the ENCODE Consortium:

- Raw and processed data from a [wide-range of assays](#)
- [Biological samples](#) (also known as biosamples) that are used in the assays
- [Antibody characterizations](#) that are performed as part of the [antibody characterization process](#) defined by the consortium.

Additional information about the activities of the ENCODE Consortium are provided on the portal:

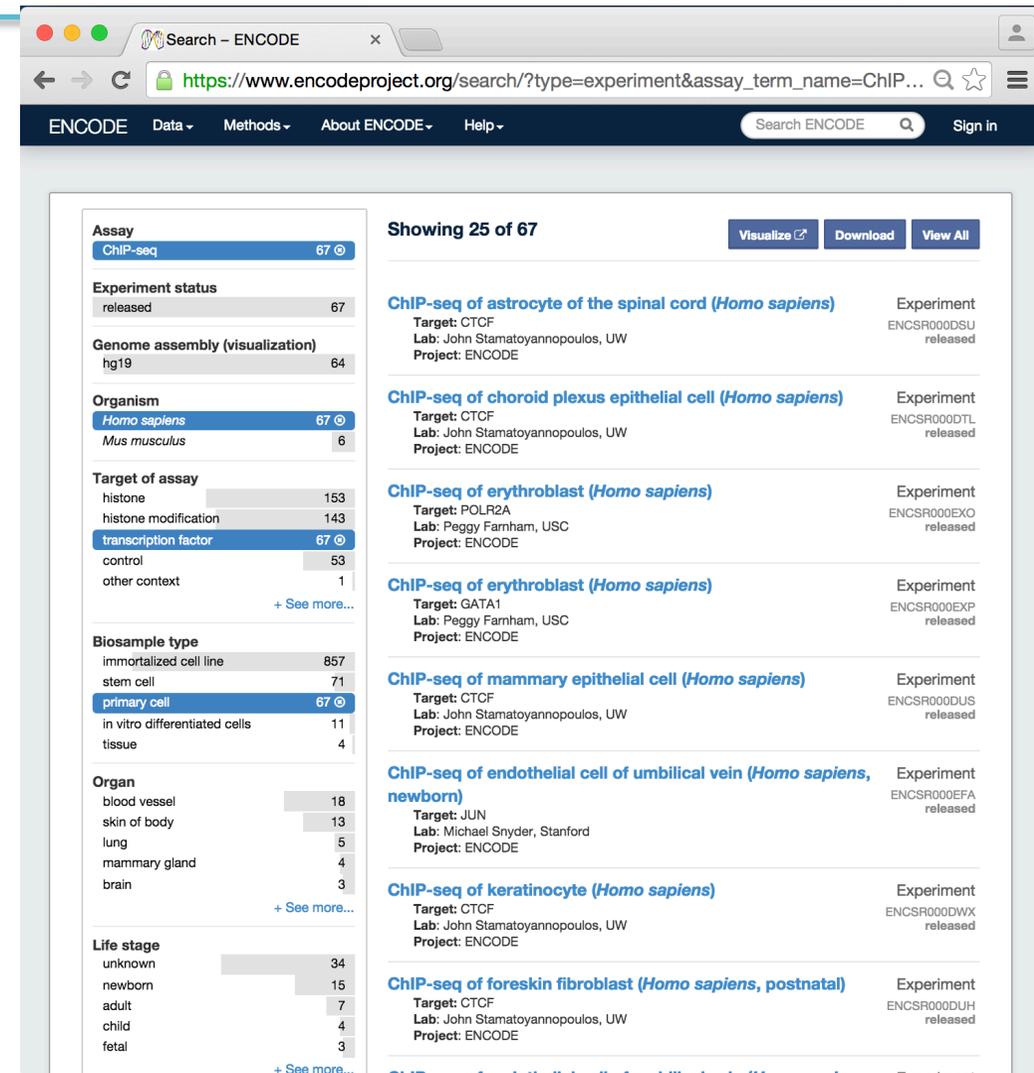
- Publications that have been [written by consortium members](#)
- Publications that have been [written by the scientific community that use ENCODE data](#)
- [Software tools](#) that have been used by the ENCODE Consortium to process and analyze assays and predict functional regions
- [Experimental guidelines](#) and [data standards](#) for generating the data
- [Outreach events, tutorials, and workshops](#) that have been organized



Summary of Interactive Features of the Portal

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

- Documentation and links to tutorials
- Faceted browsing: experiments and samples
- Detailed experiment metadata
- Visualize file relationships
- Interactive download of selected files
- Batch download of selected experiments
- Access to ENCODE annotations



The screenshot shows the ENCODE project search results page for ChIP-seq experiments. The page is titled "Search - ENCODE" and displays a list of 67 experiments. The left sidebar contains faceted browsing options for Assay, Experiment status, Genome assembly, Organism, Target of assay, Biosample type, Organ, and Life stage. The main content area shows a list of experiments with details such as Target, Lab, and Project. The experiments listed include:

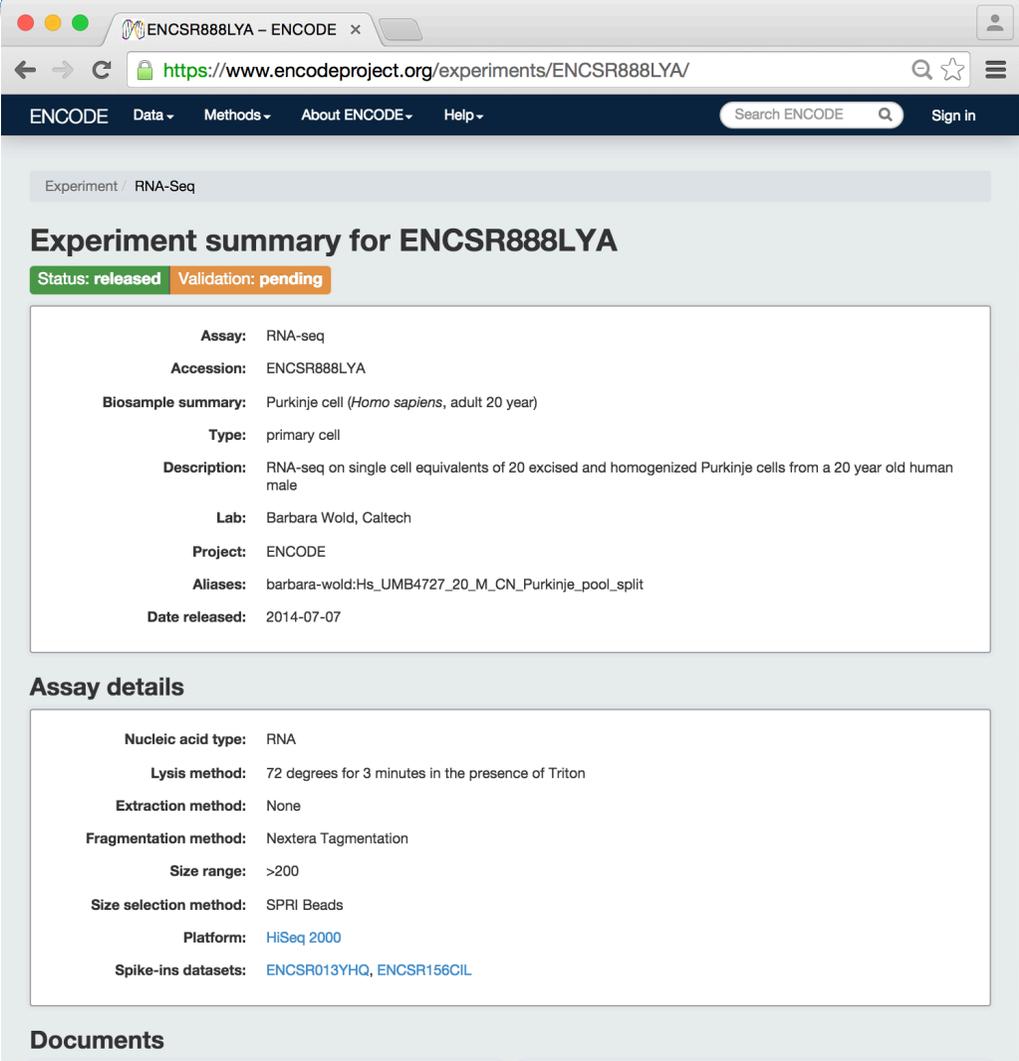
- ChIP-seq of astrocyte of the spinal cord (*Homo sapiens*)
- ChIP-seq of choroid plexus epithelial cell (*Homo sapiens*)
- ChIP-seq of erythroblast (*Homo sapiens*)
- ChIP-seq of erythroblast (*Homo sapiens*)
- ChIP-seq of mammary epithelial cell (*Homo sapiens*)
- ChIP-seq of endothelial cell of umbilical vein (*Homo sapiens, newborn*)
- ChIP-seq of keratinocyte (*Homo sapiens*)
- ChIP-seq of foreskin fibroblast (*Homo sapiens, postnatal*)



Summary of Interactive Features of the Portal

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

- Documentation and links to tutorials
- Faceted browsing: experiments and samples
- Detailed experiment metadata
- Visualize file relationships
- Interactive download of selected files
- Batch download of selected experiments
- Access to ENCODE annotations



The screenshot displays the ENCODE portal interface for the experiment ENCSR888LYA. The browser address bar shows the URL <https://www.encodeproject.org/experiments/ENCSR888LYA/>. The page header includes the ENCODE logo, navigation menus for Data, Methods, About ENCODE, and Help, along with a search bar and a Sign in link. The main content area is titled "Experiment summary for ENCSR888LYA" and features a status bar with "Status: released" and "Validation: pending". Below this, a detailed metadata table is presented:

Assay:	RNA-seq
Accession:	ENCSR888LYA
Biosample summary:	Purkinje cell (<i>Homo sapiens</i> , adult 20 year)
Type:	primary cell
Description:	RNA-seq on single cell equivalents of 20 excised and homogenized Purkinje cells from a 20 year old human male
Lab:	Barbara Wold, Caltech
Project:	ENCODE
Aliases:	barbara-wold:Hs_UMB4727_20_M_CN_Purkinje_pool_split
Date released:	2014-07-07

Below the metadata table, the "Assay details" section provides further information:

Nucleic acid type:	RNA
Lysis method:	72 degrees for 3 minutes in the presence of Triton
Extraction method:	None
Fragmentation method:	Nextera Tagmentation
Size range:	>200
Size selection method:	SPRI Beads
Platform:	HiSeq 2000
Spike-ins datasets:	ENCSR013YHQ, ENCSR156CIL

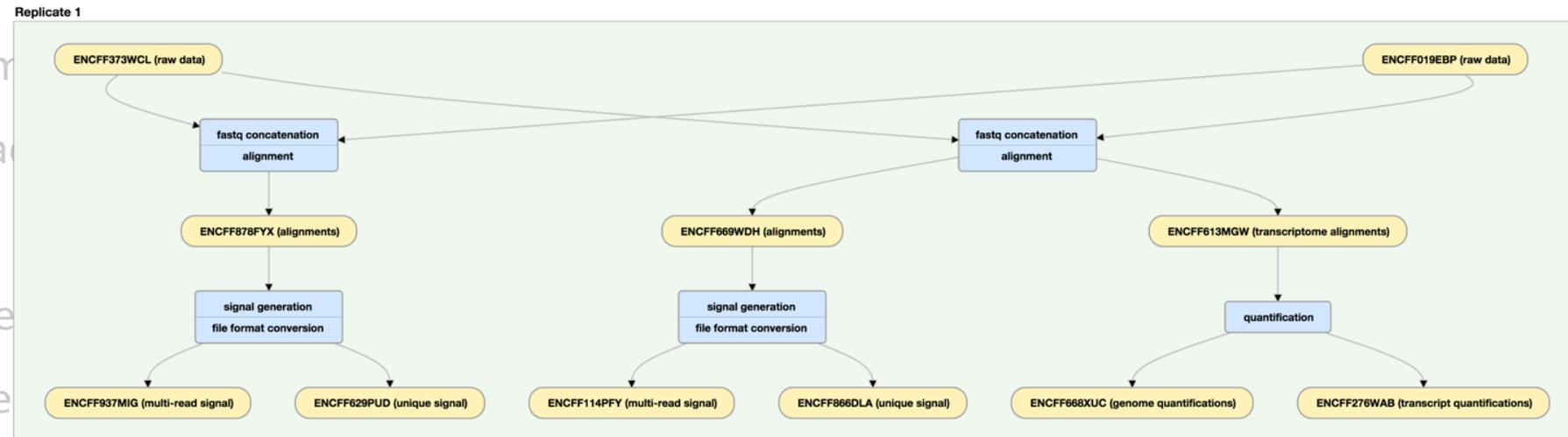
The bottom of the page shows a "Documents" section, which is currently empty.



Summary of Interactive Features of the Portal

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

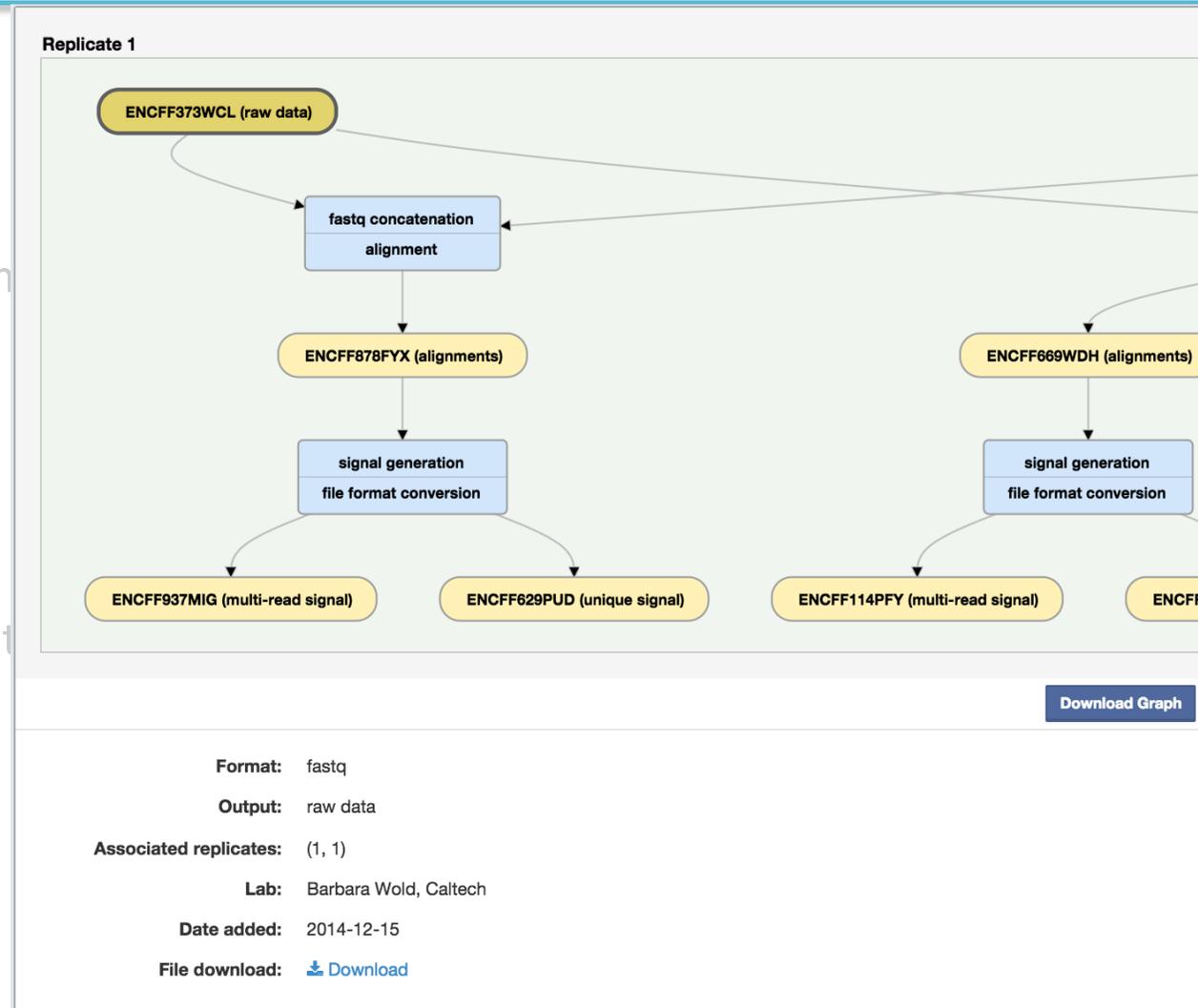
- Documentation and links to tutorials
- Faceted browsing: experiment
- Detailed experiment meta
- **Visualize file relationships**
- Interactive download of se
- Batch download of selecte
- Access to ENCODE annotations



Summary of Interactive Features of the Portal

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

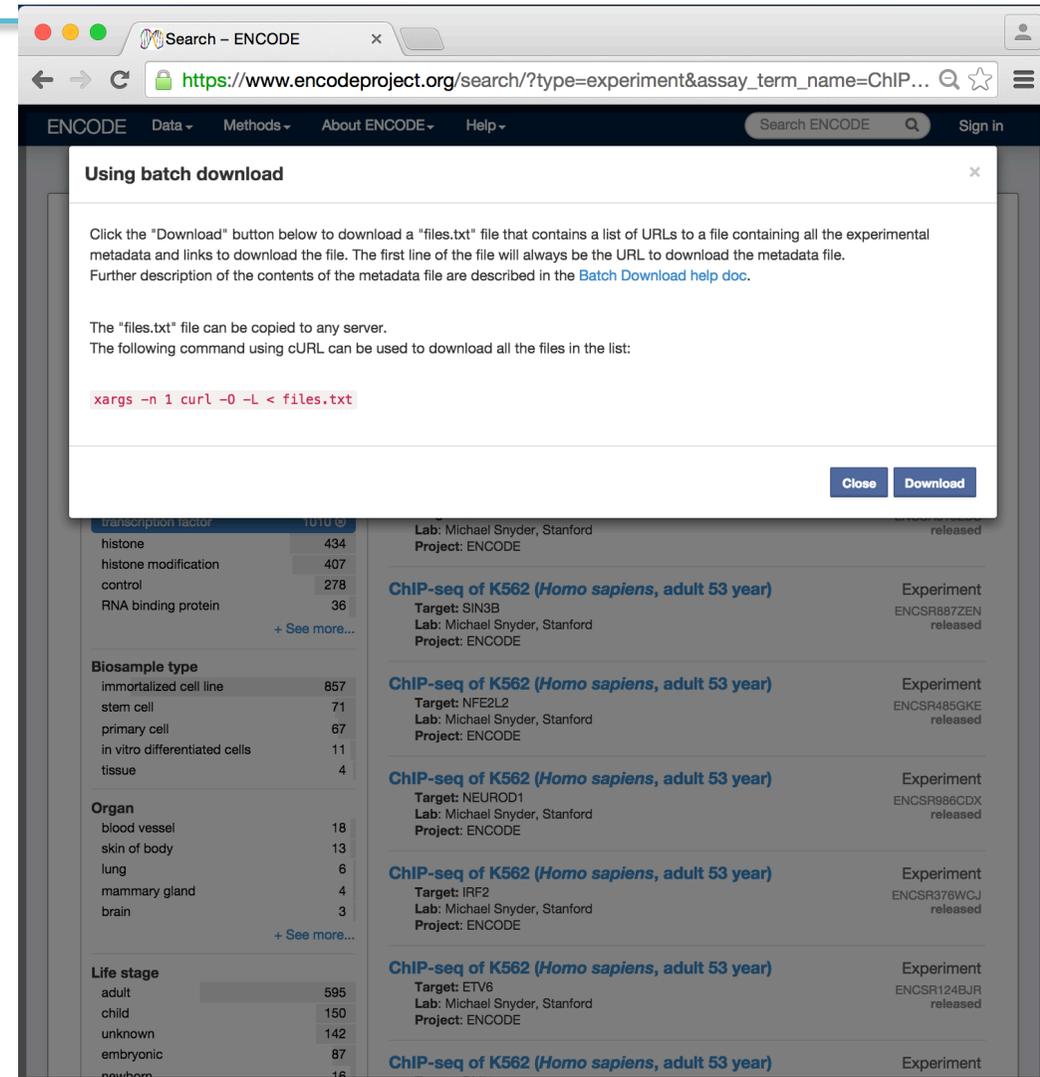
- Documentation and links to tutorials
- Faceted browsing: experiments and samples
- Detailed experiment metadata
- Visualize file relationships
- Interactive download of selected files
- Batch download of selected experiments
- Access to ENCODE annotations



Summary of Interactive Features of the Portal

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

- Documentation and links to tutorials
- Faceted browsing: experiments and samples
- Detailed experiment metadata
- Visualize file relationships
- Interactive download of selected files
- Batch download of selected experiments
- Access to ENCODE annotations



The screenshot shows a web browser window displaying the ENCODE portal search results. A modal window titled "Using batch download" is open, providing instructions on how to download a "files.txt" file containing a list of URLs for experimental metadata. The modal includes a code block with the command: `xargs -n 1 curl -O -L < files.txt`. Below the modal, the search results are displayed in a table format, showing various experiments such as "ChIP-seq of K562 (Homo sapiens, adult 53 year)" with details on the target, lab, and project.

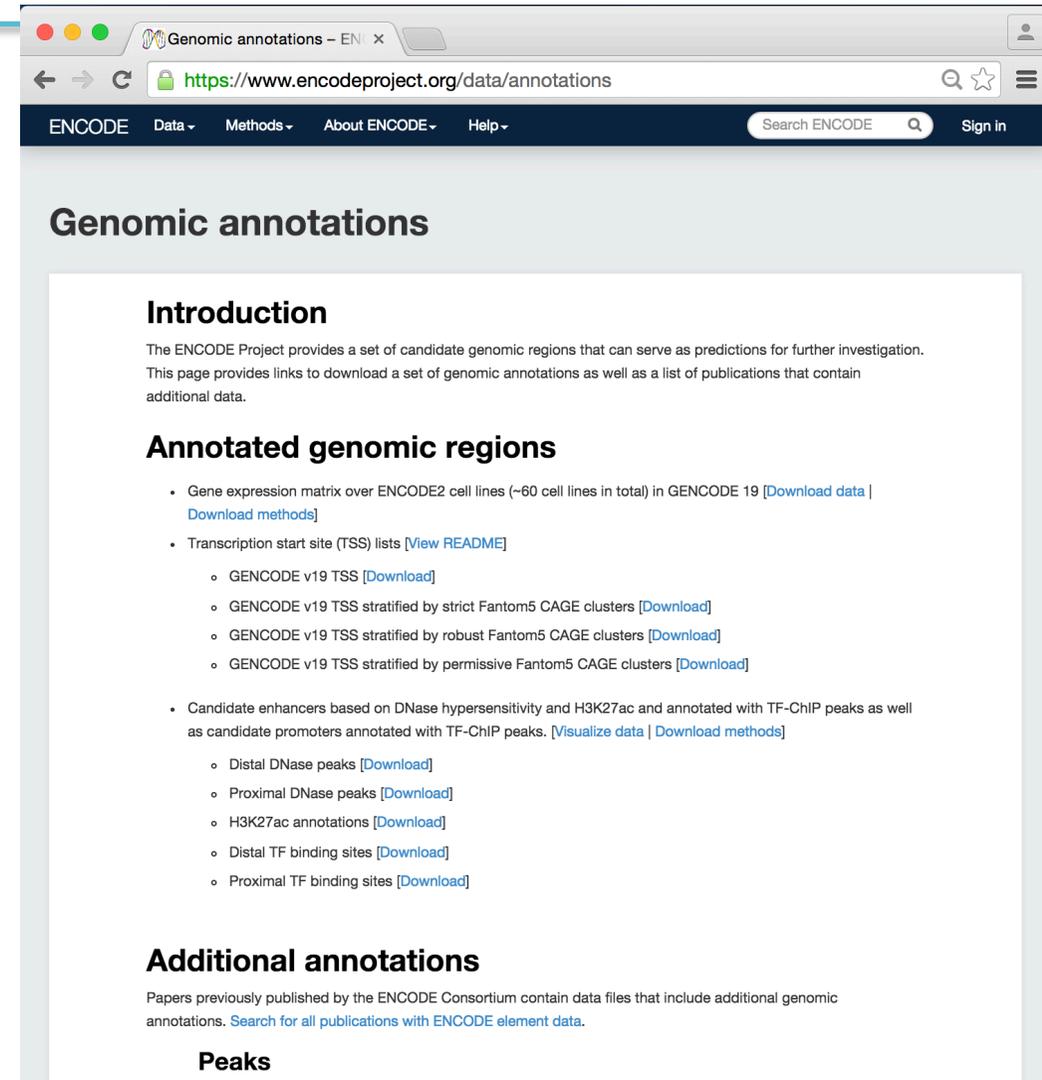
Transcription factor	Count	Lab	Project	Experiment
histone	434	Michael Snyder, Stanford	ENCODE	released
histone modification	407			
control	278			
RNA binding protein	36			
+ See more...				
Biosample type				
immortalized cell line	857			
stem cell	71			
primary cell	67			
in vitro differentiated cells	11			
tissue	4			
+ See more...				
Organ				
blood vessel	18			
skin of body	13			
lung	6			
mammary gland	4			
brain	3			
+ See more...				
Life stage				
adult	595			
child	150			
unknown	142			
embryonic	87			
newborn	16			



Summary of Interactive Features of the Portal

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

- Documentation and links to tutorials
- Faceted browsing: experiments and samples
- Detailed experiment metadata
- Visualize file relationships
- Interactive download of selected files
- Batch download of selected experiments
- Access to ENCODE annotations



The screenshot shows a web browser window displaying the ENCODE Genomic annotations portal. The browser's address bar shows the URL <https://www.encodeproject.org/data/annotations>. The page header includes the ENCODE logo and navigation links for Data, Methods, About ENCODE, and Help. A search bar and a 'Sign in' button are also present. The main content area is titled 'Genomic annotations' and contains an 'Introduction' section, an 'Annotated genomic regions' section with a bulleted list of data types and download links, and an 'Additional annotations' section. The 'Peaks' section is partially visible at the bottom.



ENCODE Bismark DNA-ME pipeline: (paired- 4 apps unconfigured Output folder... Readme Run Analysis...

Inputs	App	Outputs
*.gz A genome	WGBS-genome-... set inputs	Converted Genome Index
*.gz The reads (pair one) that ought to be methylated	WGBS-mott-rea... set inputs	Mott Trimmed Reads for input into Bismark (1st pair)
*.gz The reads (pair two) that ought to be methylated		Mott Trimmed Reads for input into Bismark (2nd pair)
via WGBS-genome-index Converted Genome Index	WGBS-bismark-... set inputs	tgz file of mapped bismark outputs
via WGBS-mott-read-trimmer-se Mott Trimmed Reads f		
via WGBS-mott-read-trimmer-se Mott Trimmed Reads f		
*.gz A genome		
*.gz A genome	WGBS-extract-r... set inputs	CG methylation BED file
via WGBS-bismark-map-pe tgz file of mapped bismark c		CHG methylation BED file
		CHH methylation BED file
		the sam file



Inputs

App

Outputs

Input Files

*.gz A genome

WGBS-genome-...
set inputs

Converted Genome Index

*.gz The reads (pair one) that ought to be methylated

WGBS-mott-rea...
set inputs

Mott Trimmed Reads for input into Bismark (1st pair)

*.gz The reads (pair two) that ought to be methylated

Mott Trimmed Reads for input into Bismark (2nd pair)

via WGBS-genome-index Converted Genome Index

WGBS-bismark-...
set inputs

tgz file of mapped bismark outputs

via WGBS-mott-read-trimmer-se Mott Trimmed Reads f

via WGBS-mott-read-trimmer-se Mott Trimmed Reads f

*.gz A genome

Output Files

*.gz A genome

WGBS-extract-r...
set inputs

CG methylation BED file

CHG methylation BED file

CHH methylation BED file

the sam file

via WGBS-bismark-map-pe tgz file of mapped bismark c

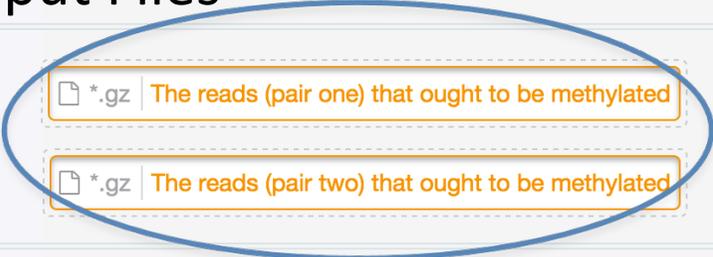


Inputs

App

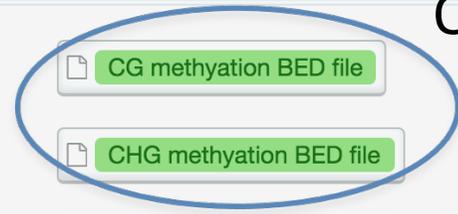
Outputs

Input Files



Outputs plumbed to inputs

Output Files



Deploy Analysis Pipelines to the Cloud

Provenance
Accessioned inputs

Ease of Use
Drop in your files

ENCODE Bismark DNA-ME pipeline: (paired-
4 apps unconfigured | Output folder... | Readme | Run Analysis...

Inputs	App	Outputs
*.gz A genome	WGBS-genome-... set inputs	Converted Genome Index
*.gz The reads (pair one) that ought to be methylated *.gz The reads (pair two) that ought to be methylated	WGBS-mott-rea... set inputs	Mott Trimmed Reads for input into Bismark (1st pair) Mott Trimmed Reads for input into Bismark (2nd pair)
via WGBS-genome-index Converted Genome Index	WGBS-bismark-... set inputs	tgz file of mapped bismark outputs
via WGBS-mott-read-trimmer-se Mott Trimmed Reads f		
via WGBS-mott-read-trimmer-se Mott Trimmed Reads f		
*.gz A genome		
*.gz A genome	WGBS-extract-r... set inputs	CG methylation BED file CHG methylation BED file CHH methylation BED file the sam file
via WGBS-bismark-map-pe tgz file of mapped bismark c		

Replicable
On the web to re-run.
Pipelines will be modeled in the metadata database

Scalable
1000's of runs will be populated from the metadata database.
Re-run on the web for a few datasets.



Deploy Analysis Pipelines to the Cloud

Provenance
Accessioned inputs

Ease of Use
Drop in your files

ENCODE Bismark DNA-ME pipeline: (paired-
4 apps unconfigured
Output folder...
Readme
Run Analysis...

Inputs	App	Outputs
*.gz A genome	WGBS-genome-... set inputs	Converted Genome Index
*.gz The reads (pair one) that ought to be methylated *.gz The reads (pair two) that ought to be methylated	WGBS-mott-rea... set inputs	Mott Trimmed Reads for input into Bismark (1st pair) Mott Trimmed Reads for input into Bismark (2nd pair)
via WGBS-genome-index Converted Genome Index	WGBS-bismark-... set inputs	tgz file of mapped bismark outputs
via WGBS-mott-read-trimmer-se Mott Trimmed Reads f		
via WGBS-mott-read-trimmer-se Mott Trimmed Reads f		
*.gz A genome		
*.gz A genome	WGBS-extract-r... set inputs	CG methylation BED file CHG methylation BED file CHH methylation BED file the sam file
via WGBS-bismark-map-pe tgz file of mapped bismark c		

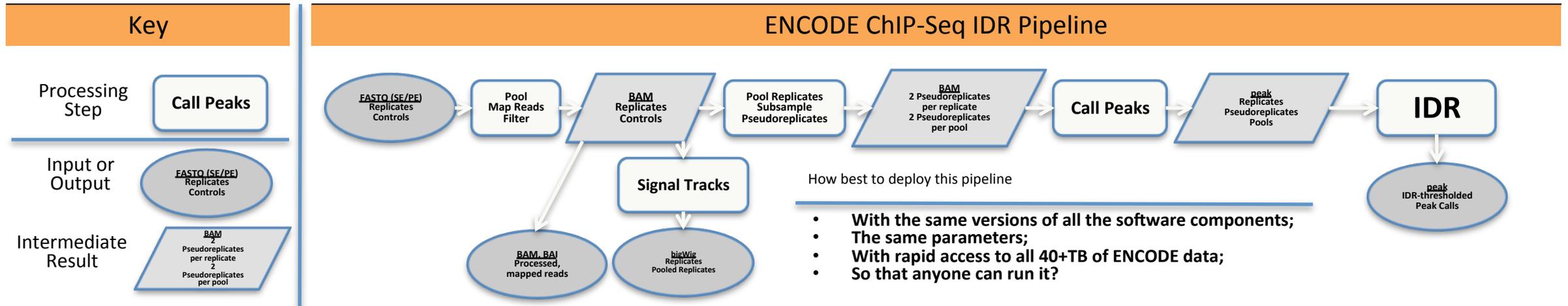
Replicable
On the web to re-run.
Pipelines will be modeled in the metadata database

Scalable
1000's of runs will be populated from the metadata database.
Re-run on the web for a few datasets.

Either way it's *exactly the same pipeline.*



Summary: ENCODE Analysis on the Cloud



- The ENCODE Data Analysis Center is defining standard ENCODE pipelines for ChIP-seq, Dnase-seq, RNA-seq, and WGBS
- The ENCODE DCC is implementing these pipelines on a web-accessible cloud computing platform
- Design goals: Transparency, replicability, provenance, accessibility



Programmatic access via the ENCODE REST API

```
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
```

- All Portal content is accessible via URL's; just add ?format=json
- The database record is returned in JSON format
- JSON can be parsed in your language of choice



Programmatic access via the ENCODE REST API

```
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
```

```
jseth:Keystone Epigenomics 2015 jseth$ ./GET_object.py
ENCSR236EGS
RNA-seq on a dissected area of layer V from an 8 month old male wild type C57Bl6 mouse
jseth:Keystone Epigenomics 2015 jseth$ █
```



Programmatic access via the ENCODE REST API

```
GET_search.py *
1  #!/usr/bin/env python
2
3  import requests
4
5  URL = ('https://www.encodeproject.org/search/?'
6        'type=experiment&'
7        'assay_term_name=ChIP-seq&'
8        'replicates.library.biosample.donor.organism.scientific_name=Homo sapiens&'
9        'target.investigated_as=transcription factor&'
10       'replicates.library.biosample.biosample_type=in vitro differentiated cells&'
11       'format=json')
12
13  response = requests.get(URL)
14
15  search_result = response.json()['@graph']
16
17  #extract and print the target for each experiment
18  print '\n'.join([experiment['target']['label'] for experiment in search_result])
19
```



Programmatic access via the ENCODE REST API

```
GET_search.py *
1  #!/usr/bin/env python
2
3  import requests
4
5  URL = ('http://www.encodeproject.org/experiments?query=Homo+sapiens&
6         'type=ChIP-seq&assay=EZH2&replid=CTCF&target=POLR2AphosphoS5
7         'assay=EZH2&replid=CTCF&target=POLR2AphosphoS5&format=json')
8
9  #extract and print the target for each experiment
10
11  #extract and print the target for each experiment
12
13  response = requests.get(URL)
14
15  search_result = response.json()['@graph']
16
17  #extract and print the target for each experiment
18  print '\n'.join([experiment['target']['label'] for experiment in search_result])
19
```

jseth:Keystone Epigenomics 2015 jseth\$./GET_search.py

SMC3
RAD21
MXI1
EP300
CTCF
EZH2
EZH2
CTCF
POLR2AphosphoS5
REST
TAF1

jseth:Keystone Epigenomics 2015 jseth\$ █



The ENCODE Portal: Recap

- Interactive access to ENCODE metadata via faceted browsing and search
- Interactive retrieval of ENCODE data one file at a time
- Batch download of ENCODE metadata and data files
- Access to ENCODE annotations
- Preview of ENCODE pipelines
- Programmatic access using the ENCODE REST API



The ENCODE Users Meeting

<https://www.encode2015.org> June 29 – July 1, 2015 – The Bolger Center in Potomac, MD

Key topics

- Navigate ENCODE data:
 - Learn to use resources for viewing, querying, and downloading ENCODE data
- Analyze ENCODE data:
 - Learn to use ENCODE web-based and command-line analysis tools
 - Run ENCODE processing pipelines on your own data (ChIP-seq, RNA-seq, DNase-seq, DNA methylation)
- Use ENCODE data to:
 - Interpret human variation and personal genomes
 - Interpret cancer genomes
 - Connect genes to their controlling regulatory elements to target genes across the genome
 - Identify likely cell types and pathways underlying non-coding disease associations



The ENCODE DCC



Eurie Hong, Mike Cherry (PI), Jim Kent (co-PI), Ben Hitz



Esther Chan, Jean Davidson, Venkat Malladi, Cricket Sloan, Seth Strattan



Nikhil Poddaturi, Laurence Rowe, Forrest Tanaka, Brian Lee, Stuart Miyasato, Matt Simison, Zhenhua Wang, Marcus Ho



@encodedcc



encode-help@lists.stanford.edu



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



ENCODE DATA AND METADATA THROUGH THE ENCODE PORTAL

J. Seth Strattan, ENCODE DCC, Stanford

Keystone Symposia Epigenomics & Methylation

ENCODE Workshop

April, 2015



@encodedcc



encode-help@lists.stanford.edu



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

