

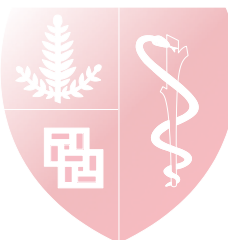
FROM ENCODE DATA TO ENCODE ANALYSES

J. Seth Strattan, PhD

ENCODE Data Coordinating Center (DCC)

Asia Pacific Bioinformatics Conference

January, 2016



ENCODE: Metadata, Data, and Analyses

So far, you have learned

- The ENCODE Portal is the **canonical source** for ENCODE metadata and data.
- The Portal also documents **ENCODE standards** like antibody standards, data release.
- The Portal links to **documentation and tutorials**.
- How to use the Portal to **browse and search** what ENCODE has done.

Focus for the rest of the course

- **Visualization** of ENCODE data.
- Programmatic **search and download** of ENCODE metadata and data.
- ENCODE **data analyses**, and how you can replicate them.



Find an experiment

Use metadata to find data:

- Search for “H3K9ac neural tube”
- Facet on ChIP-seq; mouse; mm10 assembly
- Select an experiment, for example

<https://www.encodeproject.org/experiments/ENCSR087PLZ/>

- Note metadata on protocols, replicates
- Graph: files are related by processing steps
- Download from the graph or a list
- Click on “Visualize Data” to visualize the results of this experiment.

Assay details

Nucleic acid type: DNA
Lysis method: SDS
Fragmentation method: sonication (generic)
Size range: 300-500
Size selection method: gel
Platform: HiSeq 2500

Documents

General protocol

Description:
The general library preparation protocol used by the Ren lab in their ChIP-seq experiments.



Ren_ChIP_Library_Preparation_v060614.pdf

More

General protocol

Description excerpt:
The general tissue fixation and sonication protocol used by the Ren lab in their ChIP-seq...



Ren_Tissue_Fixation_and_Sonication_v060614.pdf

More

General protocol

Description:
This document describes the Ren lab's Chromatin Immunoprecipitation protocol. June 10th, 2015



Ren_Chromatin_Immunoprecipitation.pdf

More

Biological replicate - 1

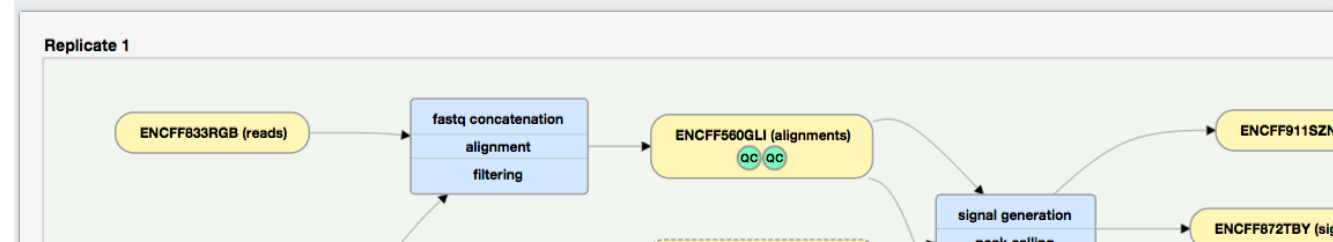
Technical replicate: 1
Library: ENCLB960XUI
Biosample: ENCBS956BQY - (*Mus musculus*, embryonic 13.5 day mixed)

Biological replicate - 2

Technical replicate: 1
Library: ENCLB045XUT
Biosample: ENCBS549OYR - (*Mus musculus*, embryonic 13.5 day mixed)

Files generated by pipeline

All Assemblies and Annotations



Visualize Data

Visualize the experiment

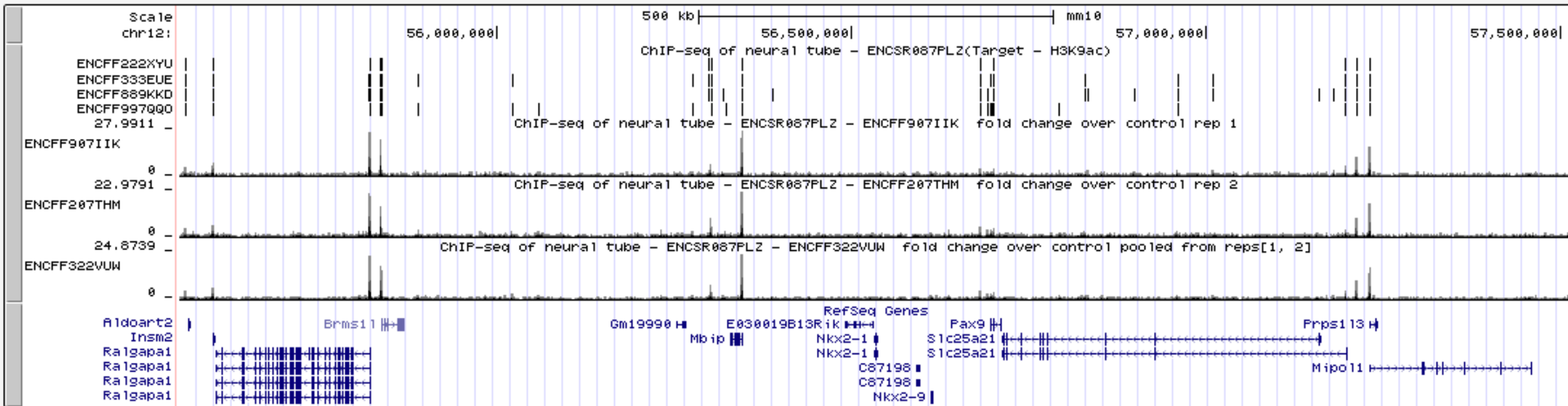
Adjust the browser settings to display fold-over-signal in "full"

UCSC Genome Browser on Mouse Dec. 2011 (GRCm38/mm10) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr12:55,552,346-57,515,345 1,963,000 bp.

chr12 (qC1) 12qA1.1 qA1.3 12qA2 12qA3 12qB1 12qB3 12qC1 12qC2 12qC3 qD1 qD2 12qD3 12qE 12qF1 12qF2



Find several experiments

Use metadata to find data:

- Search for “H3K9ac neural tube”
- Facet on ChIP-seq; mouse; mm10 assembly
- Get a list of several experiments
- Click on “Visualize Data” to visualize all the experiments matching this search.

ENCODE Data Methods About Help

H3K9ac neural tube

Showing 4 of 4 results

Visualize Download

ChIP-seq of neural tube (*Mus musculus*, embryonic 15.5 day) Experiment
Target: H3K9ac
Lab: Bing Ren, UCSD
Project: ENCODE
ENCSR571HOT released

ChIP-seq of neural tube (*Mus musculus*, embryonic 13.5 day) Experiment
Target: H3K9ac
Lab: Bing Ren, UCSD
Project: ENCODE
ENCSR087PLZ released

ChIP-seq of neural tube (*Mus musculus*, embryonic 11.5 day) Experiment
Target: H3K9ac
Lab: Bing Ren, UCSD
Project: ENCODE
ENCSR547PLI released

ChIP-seq of neural tube (*Mus musculus*, embryonic 14.5 day) Experiment
Target: H3K9ac
Lab: Bing Ren, UCSD
Project: ENCODE
ENCSR511LWL released

Assay
ChIP-seq 4

Project
ENCODE 4

Experiment status
released 4

Genome assembly (visualization)
mm10 4

Organism
Mus musculus 4

Target of assay
histone 4
histone modification 4

Biosample type
tissue 4

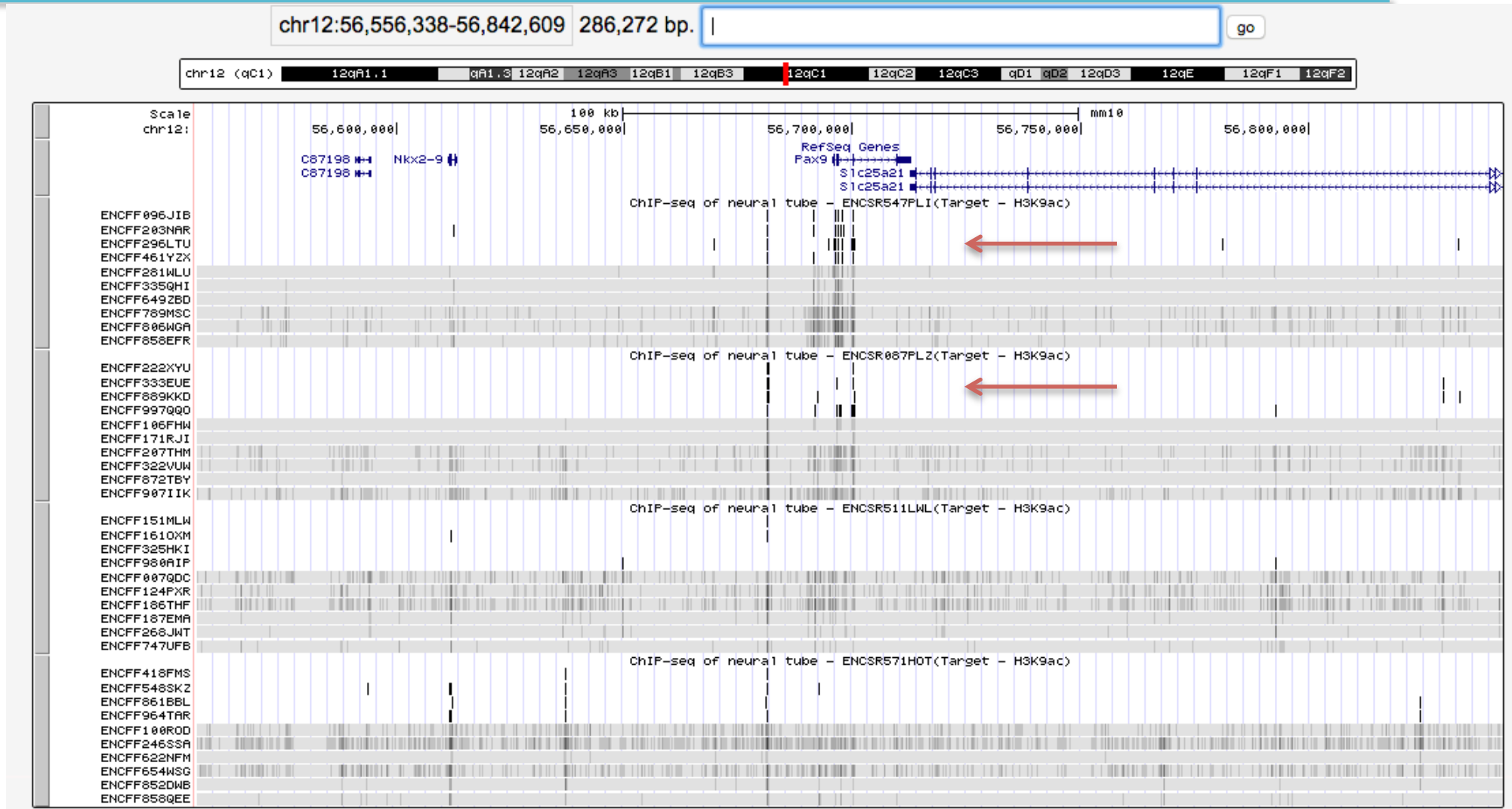
Life stage
embryonic 4

Pipeline
Histone ChIP-seq 4



Visualize several experiments

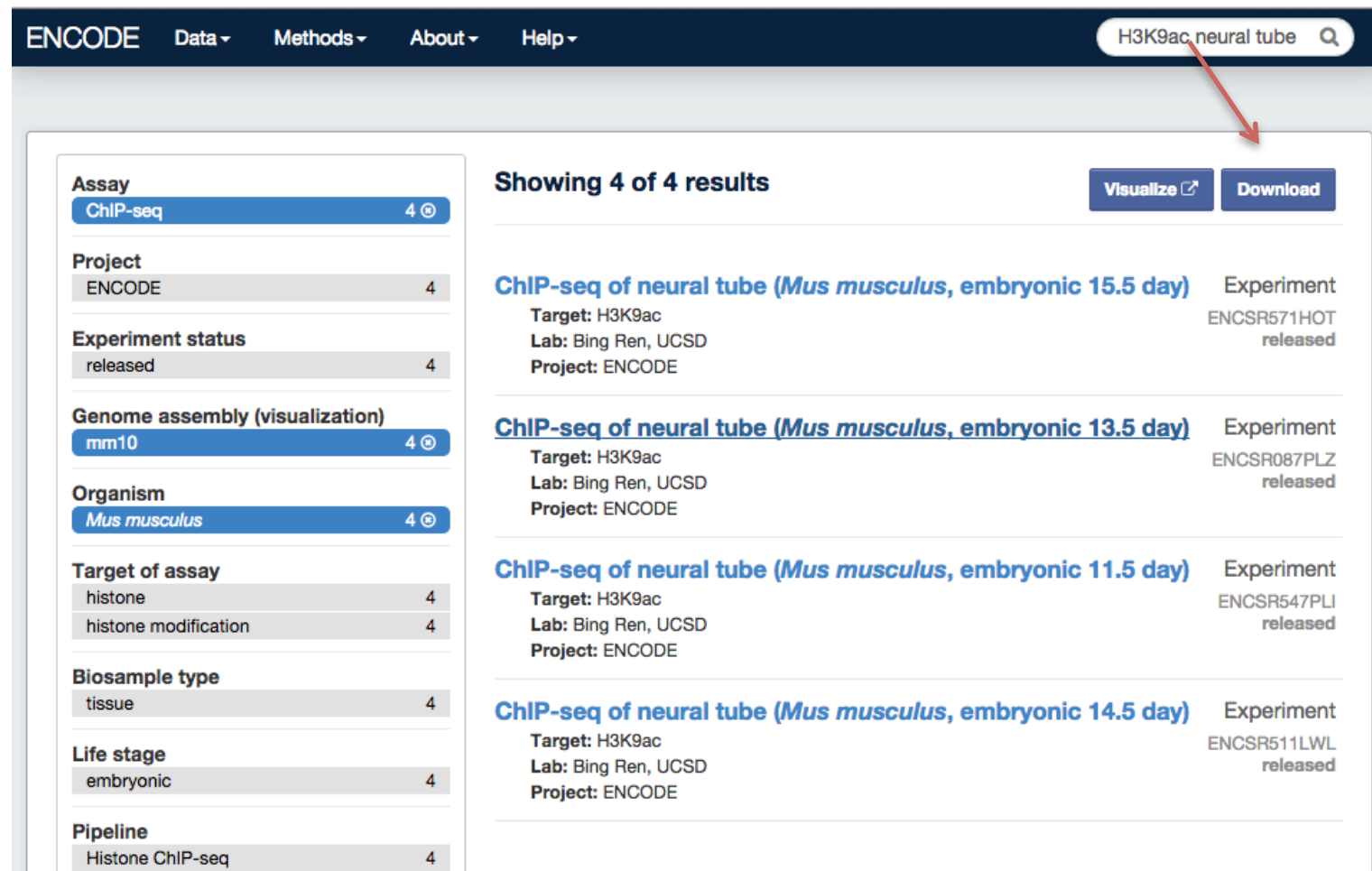
Stage-dependent
H3K9ac signal
present at Pax9 in
neural tube at
e11.5, e13.5.



Find & download several experiments

Use metadata to find data:

- Search for “H3K9ac neural tube”
- Facet on ChIP-seq; mouse; mm10 assembly
- Get a list of several experiments
- Click on “Download” to download selected metadata and complete links to data.



The screenshot shows the ENCODE website interface. At the top, there is a navigation bar with 'ENCODE' and dropdown menus for 'Data', 'Methods', 'About', and 'Help'. A search bar on the right contains the text 'H3K9ac neural tube' with a magnifying glass icon. Below the navigation bar, the main content area is divided into two columns. The left column contains a series of facet filters, each with a blue bar indicating the number of results: 'Assay' (ChIP-seq, 4), 'Project' (ENCODE, 4), 'Experiment status' (released, 4), 'Genome assembly (visualization)' (mm10, 4), 'Organism' (Mus musculus, 4), 'Target of assay' (histone, 4; histone modification, 4), 'Biosample type' (tissue, 4), 'Life stage' (embryonic, 4), and 'Pipeline' (Histone ChIP-seq, 4). The right column displays 'Showing 4 of 4 results' and a 'Download' button. Below this, four experiment entries are listed, each with a title, target, lab, project, and experiment ID. A red arrow points from the search bar to the first experiment entry.

Assay	Count
ChIP-seq	4

Project	Count
ENCODE	4

Experiment status	Count
released	4

Genome assembly (visualization)	Count
mm10	4

Organism	Count
Mus musculus	4

Target of assay	Count
histone	4
histone modification	4

Biosample type	Count
tissue	4

Life stage	Count
embryonic	4

Pipeline	Count
Histone ChIP-seq	4

Showing 4 of 4 results

Visualize Download

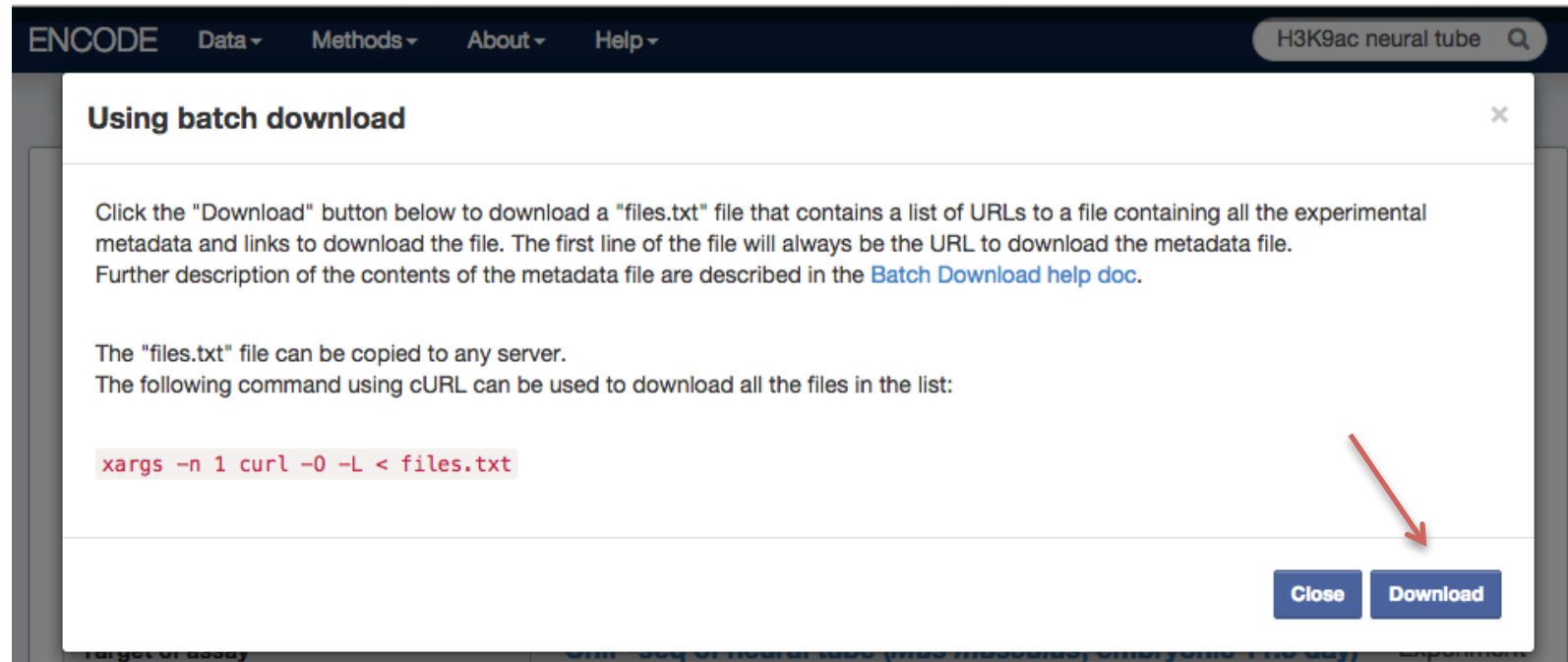
- ChIP-seq of neural tube (*Mus musculus*, embryonic 15.5 day)** Experiment ENCSR571HOT released
Target: H3K9ac
Lab: Bing Ren, UCSD
Project: ENCODE
- ChIP-seq of neural tube (*Mus musculus*, embryonic 13.5 day)** Experiment ENCSR087PLZ released
Target: H3K9ac
Lab: Bing Ren, UCSD
Project: ENCODE
- ChIP-seq of neural tube (*Mus musculus*, embryonic 11.5 day)** Experiment ENCSR547PLI released
Target: H3K9ac
Lab: Bing Ren, UCSD
Project: ENCODE
- ChIP-seq of neural tube (*Mus musculus*, embryonic 14.5 day)** Experiment ENCSR511LWL released
Target: H3K9ac
Lab: Bing Ren, UCSD
Project: ENCODE



Download several experiments

Use metadata to find data:

- Search for “H3K9ac neural tube”
- Facet on ChIP-seq; mouse; mm10 assembly
- Get a list of several experiments
- Click on “Download” to download selected metadata and complete links to data.



ENCODE Data Methods About Help H3K9ac neural tube

Using batch download

Click the "Download" button below to download a "files.txt" file that contains a list of URLs to a file containing all the experimental metadata and links to download the file. The first line of the file will always be the URL to download the metadata file. Further description of the contents of the metadata file are described in the [Batch Download help doc](#).

The "files.txt" file can be copied to any server.
The following command using cURL can be used to download all the files in the list:

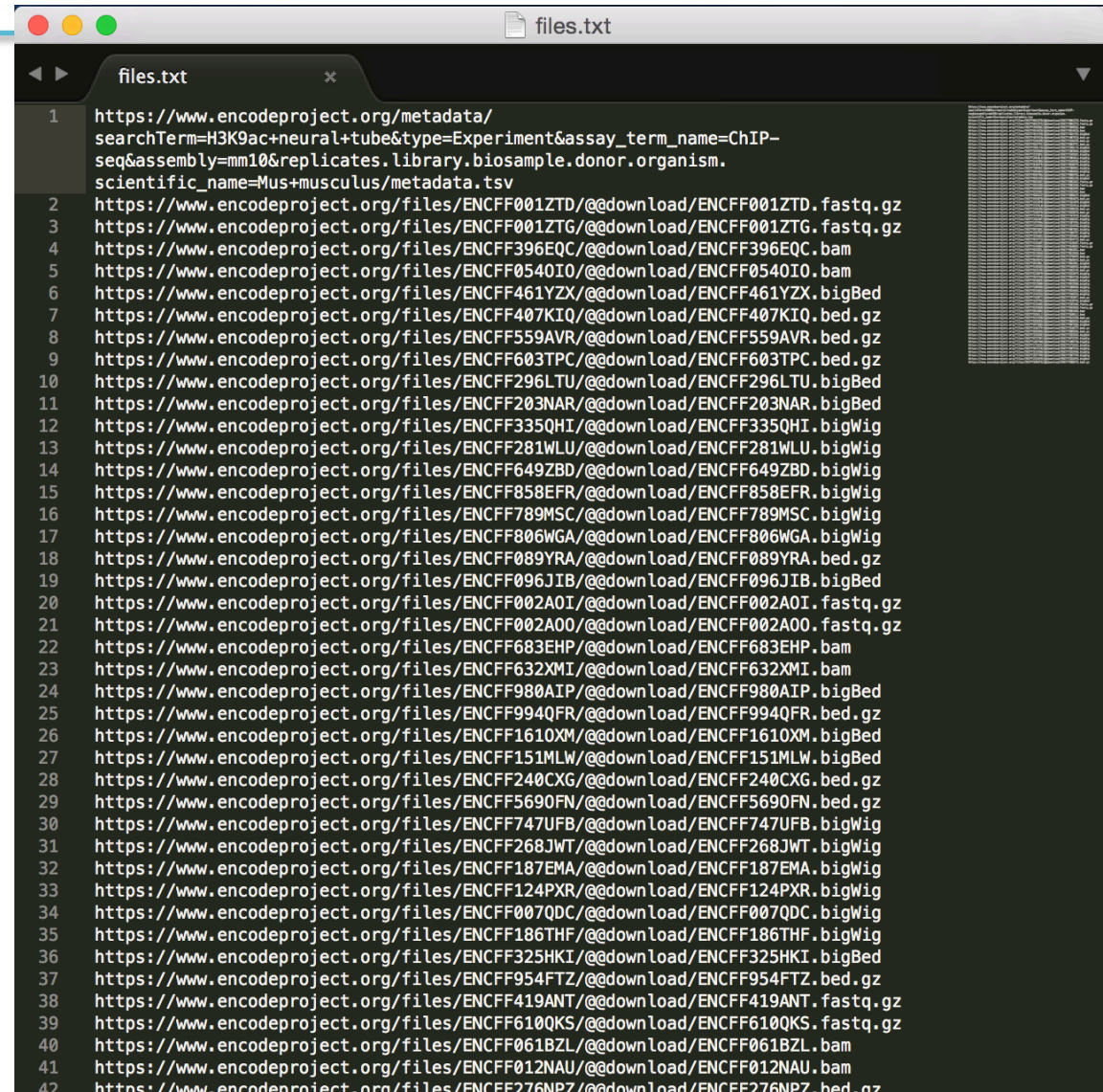
```
xargs -n 1 curl -O -L < files.txt
```

Close Download



Download several experiments

- “Download” produces a file with a list of links to all the files for all the experiments in your search.
- You can iterate through the list in your own script.
- Or:
`xargs -n 1 curl -O -L < files.txt`
- The first link is to a file called metadata.tsv that contains metadata you need to interpret what each file is.



```
files.txt
1 https://www.encodeproject.org/metadata/
  searchTerm=H3K9ac+neural+tube&type=Experiment&assay_term_name=ChIP-
  seq&assembly=mm10&replicates.library.biosample.donor.organism.
  scientific_name=Mus+musculus/metadata.tsv
2 https://www.encodeproject.org/files/ENCFF001ZTD/@download/ENCFF001ZTD.fastq.gz
3 https://www.encodeproject.org/files/ENCFF001ZTG/@download/ENCFF001ZTG.fastq.gz
4 https://www.encodeproject.org/files/ENCFF396EQC/@download/ENCFF396EQC.bam
5 https://www.encodeproject.org/files/ENCFF0540I0/@download/ENCFF0540I0.bam
6 https://www.encodeproject.org/files/ENCFF461YZX/@download/ENCFF461YZX.bigBed
7 https://www.encodeproject.org/files/ENCFF407KI0/@download/ENCFF407KI0.bed.gz
8 https://www.encodeproject.org/files/ENCFF559AVR/@download/ENCFF559AVR.bed.gz
9 https://www.encodeproject.org/files/ENCFF603TPC/@download/ENCFF603TPC.bed.gz
10 https://www.encodeproject.org/files/ENCFF296LTU/@download/ENCFF296LTU.bigBed
11 https://www.encodeproject.org/files/ENCFF203NAR/@download/ENCFF203NAR.bigBed
12 https://www.encodeproject.org/files/ENCFF335QHI/@download/ENCFF335QHI.bigWig
13 https://www.encodeproject.org/files/ENCFF281WLU/@download/ENCFF281WLU.bigWig
14 https://www.encodeproject.org/files/ENCFF649ZBD/@download/ENCFF649ZBD.bigWig
15 https://www.encodeproject.org/files/ENCFF858EFR/@download/ENCFF858EFR.bigWig
16 https://www.encodeproject.org/files/ENCFF789MSC/@download/ENCFF789MSC.bigWig
17 https://www.encodeproject.org/files/ENCFF806WGA/@download/ENCFF806WGA.bigWig
18 https://www.encodeproject.org/files/ENCFF089YRA/@download/ENCFF089YRA.bed.gz
19 https://www.encodeproject.org/files/ENCFF096JIB/@download/ENCFF096JIB.bigBed
20 https://www.encodeproject.org/files/ENCFF002A0I/@download/ENCFF002A0I.fastq.gz
21 https://www.encodeproject.org/files/ENCFF002A00/@download/ENCFF002A00.fastq.gz
22 https://www.encodeproject.org/files/ENCFF683EHP/@download/ENCFF683EHP.bam
23 https://www.encodeproject.org/files/ENCFF632XMI/@download/ENCFF632XMI.bam
24 https://www.encodeproject.org/files/ENCFF980AIP/@download/ENCFF980AIP.bigBed
25 https://www.encodeproject.org/files/ENCFF994QFR/@download/ENCFF994QFR.bed.gz
26 https://www.encodeproject.org/files/ENCFF1610XM/@download/ENCFF1610XM.bigBed
27 https://www.encodeproject.org/files/ENCFF151MLW/@download/ENCFF151MLW.bigBed
28 https://www.encodeproject.org/files/ENCFF240CXG/@download/ENCFF240CXG.bed.gz
29 https://www.encodeproject.org/files/ENCFF5690FN/@download/ENCFF5690FN.bed.gz
30 https://www.encodeproject.org/files/ENCFF747UFB/@download/ENCFF747UFB.bigWig
31 https://www.encodeproject.org/files/ENCFF268JWT/@download/ENCFF268JWT.bigWig
32 https://www.encodeproject.org/files/ENCFF187EMA/@download/ENCFF187EMA.bigWig
33 https://www.encodeproject.org/files/ENCFF124PXR/@download/ENCFF124PXR.bigWig
34 https://www.encodeproject.org/files/ENCFF007QDC/@download/ENCFF007QDC.bigWig
35 https://www.encodeproject.org/files/ENCFF186THF/@download/ENCFF186THF.bigWig
36 https://www.encodeproject.org/files/ENCFF325HKI/@download/ENCFF325HKI.bigBed
37 https://www.encodeproject.org/files/ENCFF954FTZ/@download/ENCFF954FTZ.bed.gz
38 https://www.encodeproject.org/files/ENCFF419ANT/@download/ENCFF419ANT.fastq.gz
39 https://www.encodeproject.org/files/ENCFF610QKS/@download/ENCFF610QKS.fastq.gz
40 https://www.encodeproject.org/files/ENCFF061BZL/@download/ENCFF061BZL.bam
41 https://www.encodeproject.org/files/ENCFF012NAU/@download/ENCFF012NAU.bam
42 https://www.encodeproject.org/files/ENCFF276NPZ/@download/ENCFF276NPZ.bed.gz
```



Download several experiments

- metadata.tsv: Each line contains metadata on a file from the download package.

	A	B	C	D	E	F	G	H	I	J	K
1	File accession	File format	Output type	Experiment accession	Assay	Biosample term id	Biosample term name	Biosample type	Biosample life stage	Biosample sex	Biosample organism
2	ENCFF001ZTD	fastq	reads	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
3	ENCFF001ZTG	fastq	reads	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
4	ENCFF396EQC	bam	alignments	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
5	ENCFF054OIO	bam	alignments	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
6	ENCFF461YZX	bigBed narrowPeak	peaks	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
7	ENCFF407KIQ	bed narrowPeak	peaks	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
8	ENCFF559AVR	bed narrowPeak	peaks	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
9	ENCFF603TPC	bed narrowPeak	peaks	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
10	ENCFF296LTU	bigBed narrowPeak	peaks	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
11	ENCFF203NAR	bigBed narrowPeak	peaks	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
12	ENCFF335QHI	bigWig	signal p-value	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
13	ENCFF281WLU	bigWig	signal p-value	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
14	ENCFF649ZBD	bigWig	signal p-value	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
15	ENCFF858EFR	bigWig	fold change over control	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
16	ENCFF789MSC	bigWig	fold change over control	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
17	ENCFF806WGA	bigWig	fold change over control	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
18	ENCFF089YRA	bed narrowPeak	replicated peaks	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
19	ENCFF096JIB	bigBed narrowPeak	replicated peaks	ENCSR547PLI	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
20	ENCFF002AOI	fastq	reads	ENCSR511LWL	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
21	ENCFF002A00	fastq	reads	ENCSR511LWL	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
22	ENCFF683EHP	bam	alignments	ENCSR511LWL	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
23	ENCFF632XMI	bam	alignments	ENCSR511LWL	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
24	ENCFF980AIP	bigBed narrowPeak	peaks	ENCSR511LWL	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
25	ENCFF994QFR	bed narrowPeak	peaks	ENCSR511LWL	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
26	ENCFF161OXM	bigBed narrowPeak	peaks	ENCSR511LWL	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
27	ENCFF151MLW	bigBed narrowPeak	peaks	ENCSR511LWL	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus
28	ENCFF240CYC	bed narrowPeak	peaks	ENCSR511LWL	ChIP-seq	UBERON:0001049	neural tube	tissue	embryonic	mixed	Mus musculus



Programmatic access via the ENCODE REST API

- 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

```
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
```



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 C57B16 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/assays/
6         'type=ChIP-seq&organism=Homo+sapiens&
7         'assay=ChIP-seq&target=REST&format=json
8         'replid=REST&format=json')
9         'target=POLR2AphosphoS5
10        'replid=REST
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
```

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
- Programmatic access using the ENCODE REST API

Next: ENCODE Data Analysis Pipelines

- What do they produce?
- How can they be run?



Pipelines Demonstration and Exercise

To set up an account:

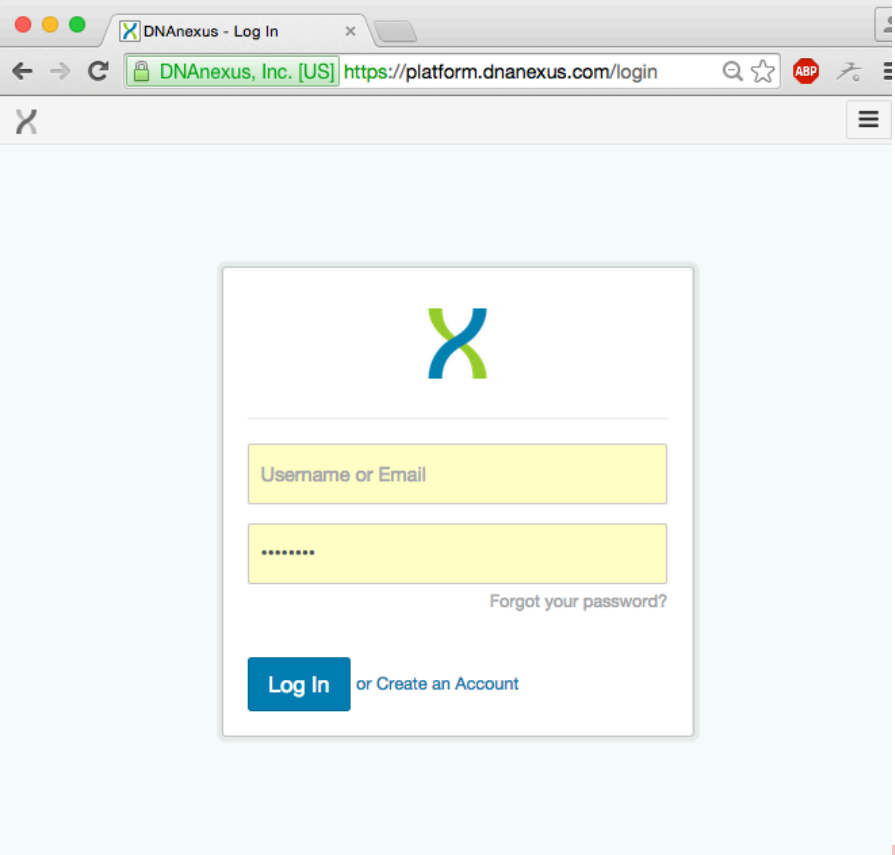
<https://www.encodeproject.org/tutorials/apbc-2016/>

Click “Prepare to run web-based pipelines”

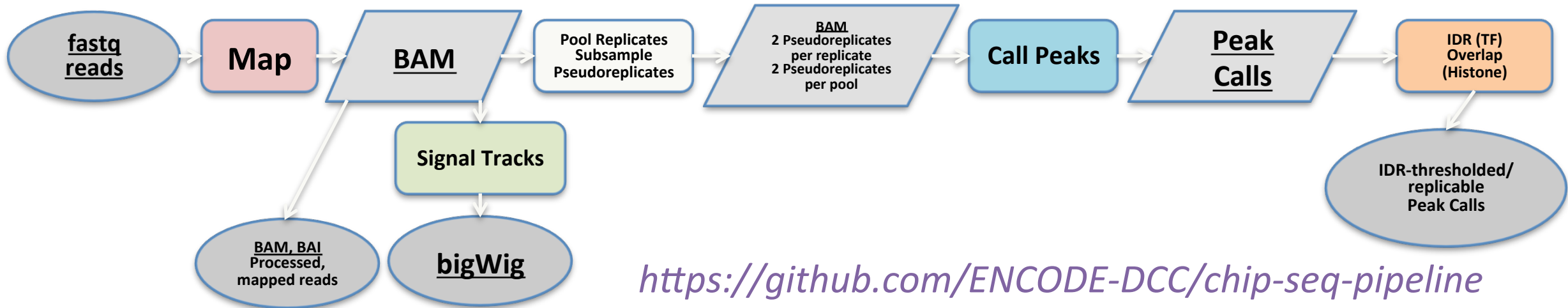
Instructions for setting up your DNAnexus environment to run ENCODE pipelines are [here](#):

[Prepare to run web-based pipelines](#)

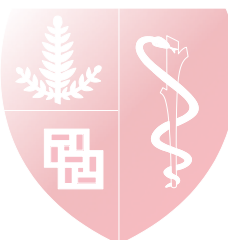
Log in ->

A screenshot of a web browser window showing the DNAnexus login page. The browser's address bar displays "DNAnexus, Inc. [US] https://platform.dnanexus.com/login". The page features the DNAnexus logo (a stylized 'X' in blue and green) at the top. Below the logo is a login form with two input fields: "Username or Email" and a password field with masked characters. A "Forgot your password?" link is positioned below the password field. At the bottom of the form, there is a blue "Log In" button followed by the text "or Create an Account".

Schema: ENCODE ChIP-seq IDR Pipeline



Target	Key Software	Input Files	Output Files	QA Metrics
TF's	bwa	fastq's (SE or PE) Two biological replicates Matched controls	One bam per replicate	NRF (Non-redundant fraction) PBC1 and 2 (PCR bottleneck coefficients) Number of distinct uniquely-mapping reads NSC/RSC (Strand cross-correlation) IDR Rescue Ratio IDR Self-Consistency Ratio IDR Reproducibility Test
	Picard markDuplicates		bigWig fold signal over control	
	samtools		bigWig p-value signal over control	
	MACS2 (Signal tracks)		bed/bigBed true replicates peaks	
SPP (PeakSeq, GEM future)	bed/bigBed pooled replicates peaks			
IDR2	bed/bigBed IDR thresholded peaks			
Histone Mods	MACS2 for peaks			
	Overlap thresholding			
	IDR2 (future)		bed/bigBed Replicated peaks	



Pipelines Demonstration and Exercise

To set up an account:

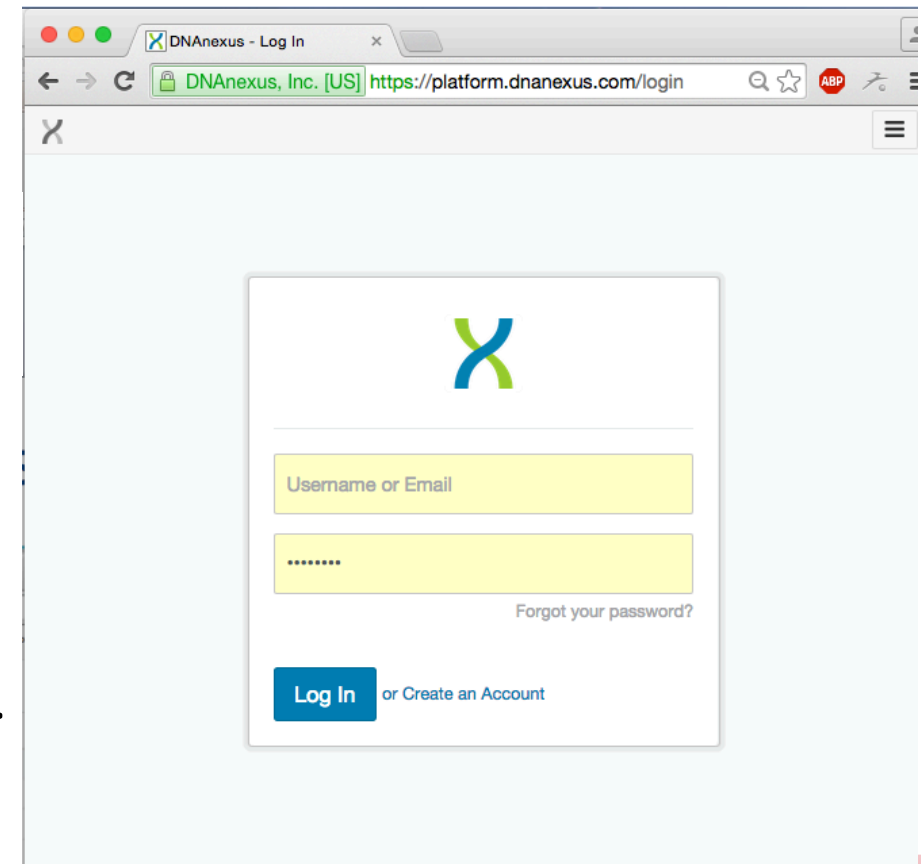
<https://www.encodeproject.org/tutorials/apbc-2016/>

Instructions for setting up your DNAnexus environment to run ENCODE pipelines are [here](#):

Prepare to run web-based pipelines

J. Seth Strattan	ENCODE DCC, Stanford University	Slides (coming soon) Prepare to run web-based pipelines ChIP-seq Pipeline How-to RNA-seq Pipeline How-to
------------------	---------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------

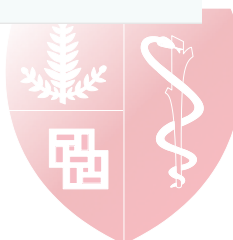
Log in ->



Exercises

[Histone ChIP-seq](#)

[RNA-seq](#)



Uniformly Processed Data On the ENCODE Portal

Histone ChIP-seq Example

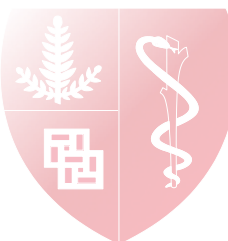
<https://www.encodeproject.org/experiments/ENCSR087PLZ/>

- Pipeline graph shows relationships between files
- Click on files to see more file metadata and download links
- Click on steps to see more software metadata and download links

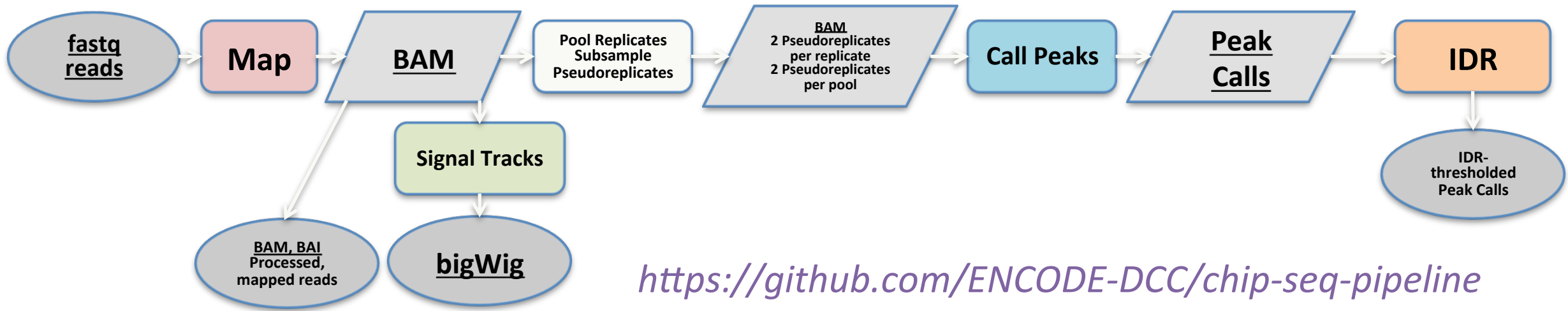
Transcription Factor ChIP-seq Example

<https://www.encodeproject.org/experiments/ENCSR077DKV/>

- Same mapping, signal tracks and peak calls
- Also have the IDR-thresholded peak calls
- “Conservative” set, based on “true” replicates; “optimal” set if peaks can be rescued by pseudo-replication.



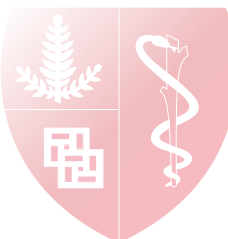
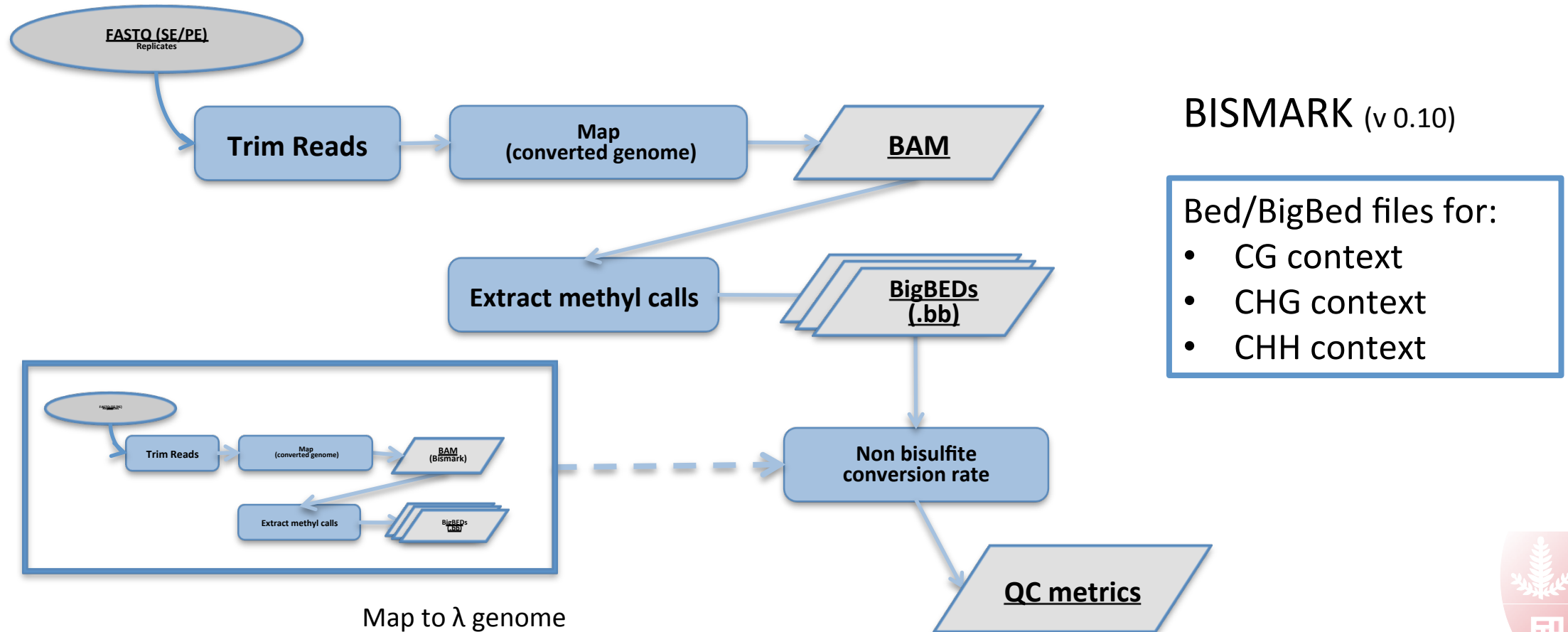
ENCODE ChIP-seq Quality Metrics: Resources



Estimates	Description	References
Depth	Number of uniquely mapping reads Number of distinct uniquely mapping reads	Jung YL, et al. Nucleic Acids Research. 2014;42(9):e74
Library Complexity	Non-Redundant Fraction PCR Bottleneck Coefficient	Landt S, et al. Genome Res. 2012. 22: 1813-1831
ChIP Quality	Normalized Strand Cross-Correlation Relative Strand Cross-Correlation	
Replicate Concordance	IDR Rescue Ratio IDR Self-Consistency Ratio IDR Reproducibility Test	Li Q, et al. Annals Applied Statistics. 2011, Vol. 5, No. 3, 1752–1779

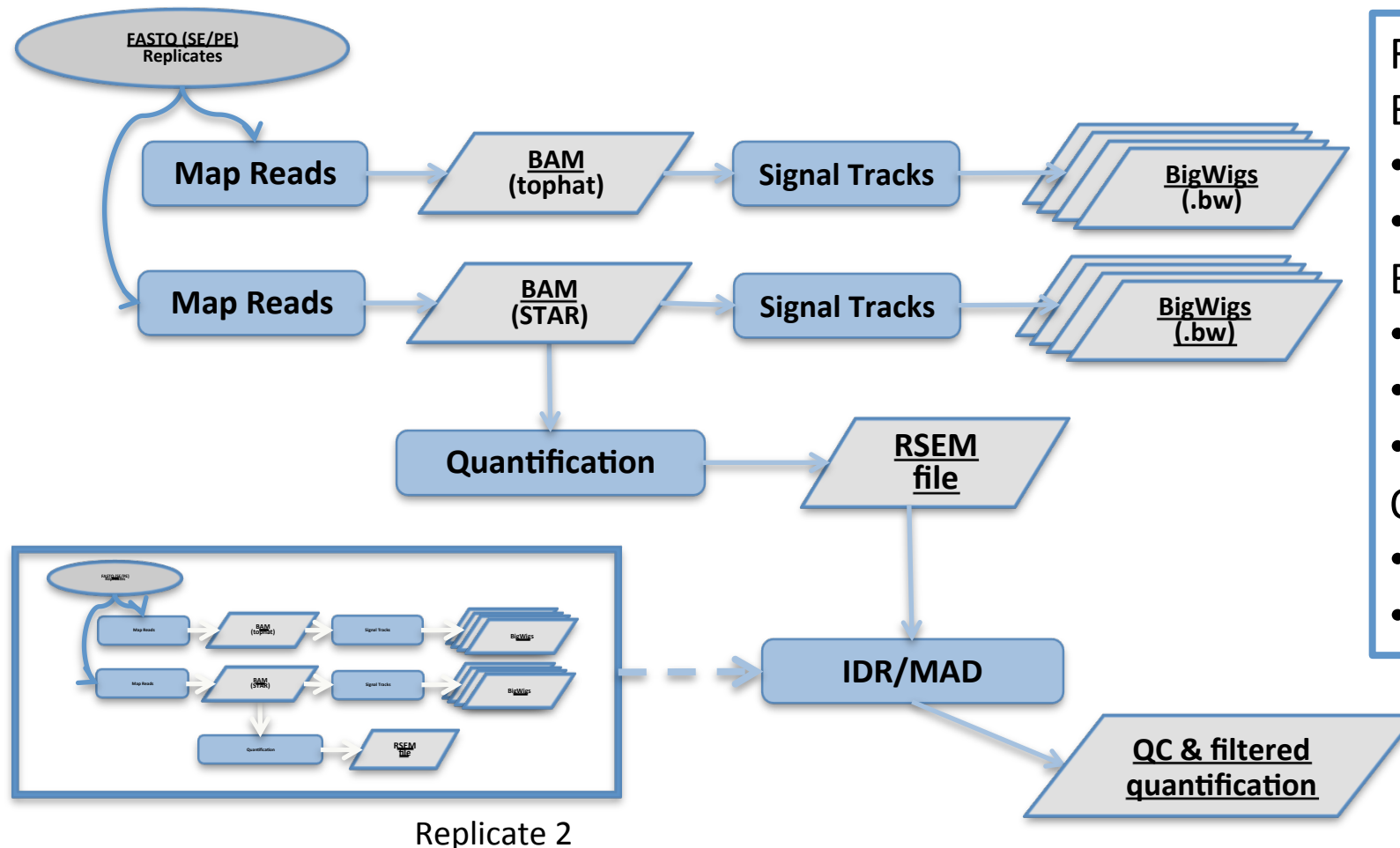
Schema: ENCODE WGBS Pipeline

<https://github.com/ENCODE-DCC/dna-me-pipeline>



Schema: ENCODE RNA-seq Pipeline

<https://github.com/ENCODE-DCC/long-rna-seq-pipeline>



For *each* Mapper (STAR, tophat) BAM files:

- mapped to genome
- mapped to transcriptome

BigWig files:

- plus/minus strand (paired)
- uniquely mapped
- multi+uniquely mapped

Quantifications (RSEM):

- genome
- transcriptome



Uniformly Processed Data On the ENCODE Portal

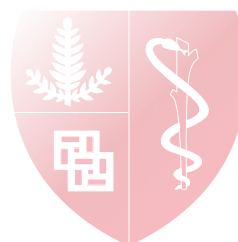
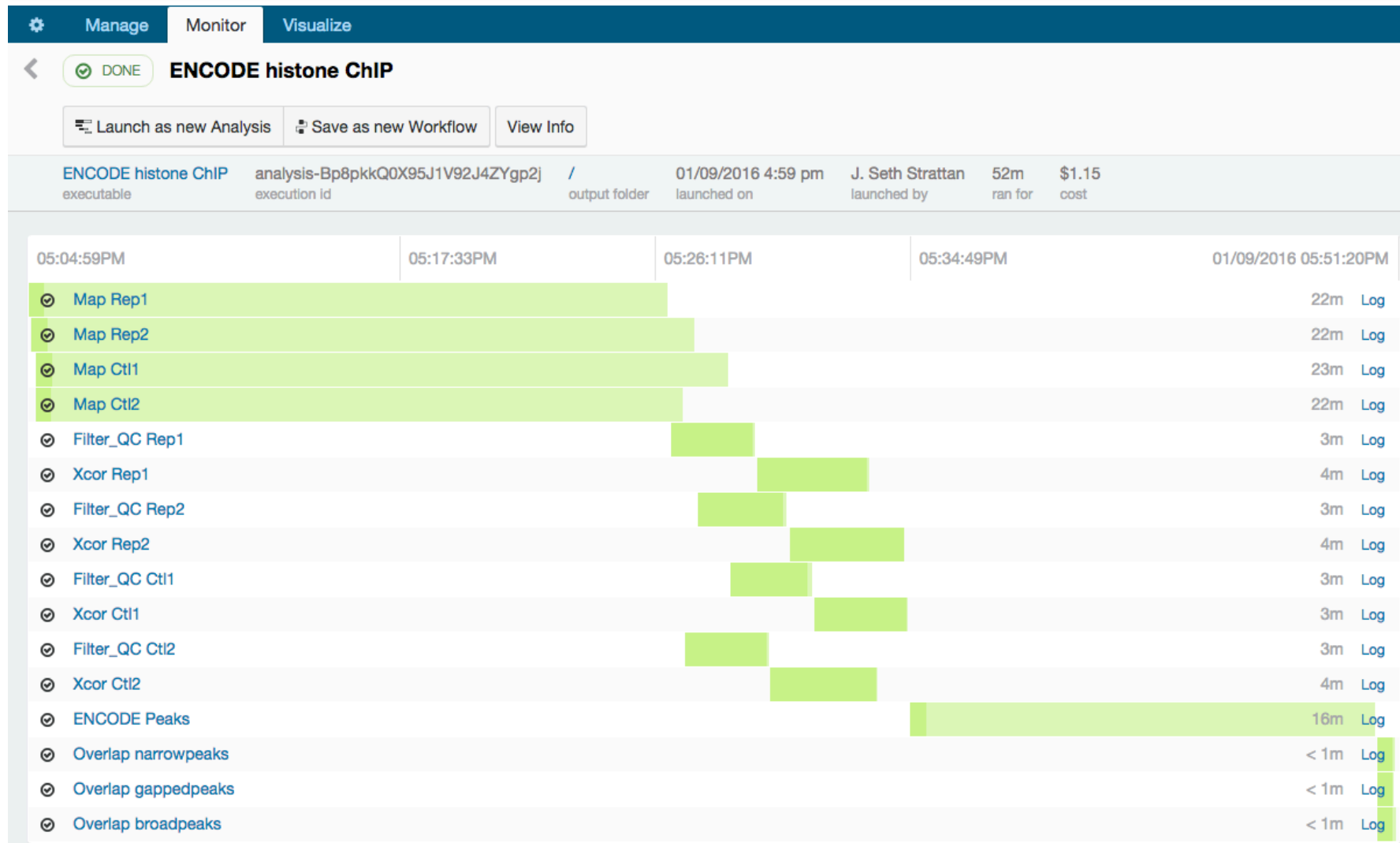
RNA-seq Example

<https://www.encodeproject.org/experiments/ENCSR368QPC/>

- Pipeline graph shows relationships between files
- Click on files to see more file metadata and download links
- Click on steps to see more software metadata and download links



Results from the ChIP-seq exercise



Results from the ChIP-seq exercise

05:34:46PM
05:39:02PM
05:41:57PM
05:44:53PM
01/09/2016 05:50:29PM

Manage
Monitor
Visualize

DONE
ENCODE histone ChIP

Launch as new Analysis
Save as new Workflow
View Info

ENCODE histone ChIP / analysis-Bp8pkkQ0X95J1V92J4ZYgp2] / ot
executable execution id

05:04:59PM 05:17:33PM

- Map Rep1
- Map Rep2
- Map Ctl1
- Map Ctl2
- Filter_QC Rep1
- Xcor Rep1
- Filter_QC Rep2
- Xcor Rep2
- Filter_QC Ctl1
- Xcor Ctl1
- Filter_QC Ctl2
- Xcor Ctl2
- ENCODE Peaks**
- Overlap narrowpeaks
- Overlap gappedpeaks
- Overlap broadpeaks

Task	Duration	Size	Log
<input checked="" type="checkbox"/> ENCODE Peaks		16m	Log
<input checked="" type="checkbox"/> ENCODE TF ChIP-seq Pooler		< 1m	Log
<input checked="" type="checkbox"/> ENCODE TF ChIP-seq Pooler		< 1m	Log
<input checked="" type="checkbox"/> ENCODE TF ChIP-seq Pseudoreplicator		< 1m	Log
<input checked="" type="checkbox"/> ENCODE TF ChIP-seq Pseudoreplicator		< 1m	Log
<input checked="" type="checkbox"/> ENCODE TF ChIP-seq Pooler		< 1m	Log
<input checked="" type="checkbox"/> ENCODE TF ChIP-seq Pooler		< 1m	Log
<input checked="" type="checkbox"/> ENCODE TF ChIP-seq Cross-Correlation Analysis		5m	Log
<input checked="" type="checkbox"/> ENCODE TF ChIP-seq Cross-Correlation Analysis		2m	Log
<input checked="" type="checkbox"/> ENCODE TF ChIP-seq Cross-Correlation Analysis		2m	Log
<input checked="" type="checkbox"/> ENCODE TF ChIP-seq Cross-Correlation Analysis		2m	Log
<input checked="" type="checkbox"/> ENCODE TF ChIP-seq Cross-Correlation Analysis		2m	Log
<input checked="" type="checkbox"/> ENCODE TF ChIP-seq Cross-Correlation Analysis		3m	Log
<input checked="" type="checkbox"/> ENCODE TF ChIP-seq Cross-Correlation Analysis		3m	Log
<input checked="" type="checkbox"/> MACS2		5m	Log
<input checked="" type="checkbox"/> MACS2		4m	Log
<input checked="" type="checkbox"/> MACS2		6m	Log
<input checked="" type="checkbox"/> MACS2		4m	Log
<input checked="" type="checkbox"/> MACS2		4m	Log
<input checked="" type="checkbox"/> MACS2		3m	Log
<input checked="" type="checkbox"/> MACS2		3m	Log
<input checked="" type="checkbox"/> MACS2		5m	Log
<input checked="" type="checkbox"/> MACS2		5m	Log

INPUTS

boolean (rep2_paired_end)
false

Rep 1 tagAlign (rep1_ta)
R1.raw.srt.filt.nodup.srt.SE.tagAlign.gz

Control 2 tagAlign (ctl2_ta)
C2.raw.srt.filt.nodup.srt.SE.tagAlign.gz

Rep 2 cross-correlation scores (rep2_xcor)
R2.raw.srt.filt.nodup.srt.filt.nodup.sample.15.SE.tagAlign.gz.cc.gz

Rep 2 tagAlign (rep2_ta)
R2.raw.srt.filt.nodup.srt.SE.tagAlign.gz

string for MACS2, hs for human, mm for mouse (genomesize)
hs

OUTPUTS

Narrowpeaks file (rep2pr1_narrowpeaks)
R2.raw.srt.filt.nodup.srt.SE.SE.pr1.tagAlign.narrowPeak.gz

Signal track fold-enrichment over control (rep2pr2_fc_signal)
R2.raw.srt.filt.nodup.srt.SE.tagAlign.gz.SE.pr2.tagAlign.fc_signal.bw

Narrowpeaks file (rep1pr1_narrowpeaks)
R1.raw.srt.filt.nodup.srt.SE.SE.pr1.tagAlign.narrowPeak.gz

Bigbed file (rep1_gappedpeaks_bb)
R1.raw.srt.filt.nodup.srt.SE.tagAlign.gappedPeak.bb

Broadpeaks file (pooledpr1_broadpeaks)
R1.raw.srt.filt.nodup.srt.SE.SE.pr1-
R2.raw.srt.filt.nodup.srt.SE.SE.pr1_pooled.tagAlign.broadPeak.gz

Signal track p-value (rep2pr2_pvalue_signal)



Results from the ChIP-seq exercise

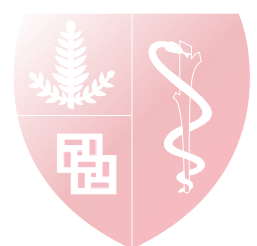
05:34:46PM 05:39:02PM 05:41:57PM 05:44:53PM 01/09/2016 05:50:29PM

Task	Progress	Time	Log
ENCODE Peaks	100%	16m	Log
ENCODE TF ChIP-seq Pooler	100%	< 1m	Log
ENCODE TF ChIP-seq Pooler	100%	< 1m	Log
ENCODE TF ChIP-seq Pseudoreplicator	100%		
ENCODE TF ChIP-seq Pseudoreplicator	100%		
ENCODE TF ChIP-seq Pooler	100%		
ENCODE TF ChIP-seq Pooler	100%		
ENCODE TF ChIP-seq Cross-Correlation Analysis	100%		
ENCODE TF ChIP-seq Cross-Correlation Analysis	100%		
ENCODE TF ChIP-seq Cross-Correlation Analysis	100%		
ENCODE TF ChIP-seq Cross-Correlation Analysis	100%		
ENCODE TF ChIP-seq Cross-Correlation Analysis	100%		
ENCODE TF ChIP-seq Cross-Correlation Analysis	100%		
ENCODE TF ChIP-seq Cross-Correlation Analysis	100%		
MACS2	100%		
MACS2	100%		
MACS2	100%		
MACS2	100%		
MACS2	100%		
MACS2	100%		
MACS2	100%		
MACS2	100%		
MACS2	100%		
MACS2	100%		



```
R2.raw.srt.filt.nodup.srt.SE.tagAlign.gz.SE.pr2_pooler.tagAlign.fc_si
gnal.bw
Narrowpeaks file (rep2_narrowpeaks)
R2.raw.srt.filt.nodup.srt.SE.tagAlign.narrowPeak.gz
Narrowpeaks file (pooled_narrowpeaks)
R1.raw.srt.filt.nodup.srt.SE-
R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.narrowPeak.gz
Gappedpeaks file (pooled_gappedpeaks)
R1.raw.srt.filt.nodup.srt.SE-
R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.gappedPeak.gz
Broadpeaks file (rep1pr1_broadpeaks)
R1.raw.srt.filt.nodup.srt.SE.SE.pr1.tagAlign.broadPeak.gz
Signal track fold-enrichment over control (rep1pr2_fc_signal)
R1.raw.srt.filt.nodup.srt.SE.tagAlign.gz.SE.pr2.tagAlign.fc_signal.bw
```

INPUTS	OUTPUTS
boolean (rep2_paired_end) false	Narrowpeaks file (rep2pr1_narrowpeaks) R2.raw.srt.filt.nodup.srt.SE.SE.pr1.tagAlign.narrowPeak.gz
Rep 1 tagAlign (rep1_ta) R1.raw.srt.filt.nodup.srt.SE.tagAlign.gz	Signal track fold-enrichment over control (rep2pr2_fc_signal) R2.raw.srt.filt.nodup.srt.SE.tagAlign.gz.SE.pr2.tagAlign.fc_signal.bw
Control 2 tagAlign (ctl2_ta) C2.raw.srt.filt.nodup.srt.SE.tagAlign.gz	Narrowpeaks file (rep1pr1_narrowpeaks) R1.raw.srt.filt.nodup.srt.SE.SE.pr1.tagAlign.narrowPeak.gz
Rep 2 cross-correlation scores (rep2_xcor) R2.raw.srt.filt.nodup.srt.filt.nodup.sample.15.SE.tagAlign.gz.cc.qc	Bigbed file (rep1_gappedpeaks_bb) R1.raw.srt.filt.nodup.srt.SE.tagAlign.gappedPeak.bb
Rep 2 tagAlign (rep2_ta) R2.raw.srt.filt.nodup.srt.SE.tagAlign.gz	Broadpeaks file (pooledpr1_broadpeaks) R1.raw.srt.filt.nodup.srt.SE.SE.pr1- R2.raw.srt.filt.nodup.srt.SE.SE.pr1_pooled.tagAlign.broadPeak.gz
string for MACS2, hs for human, mm for mouse (genomesize) hs	Signal track n-value (rep2pr2_nvalue_signal)



Results from the ChIP-seq exercise

Narrowpeaks file (rep1_narrowpeaks)
R1.raw.srt.filt.nodup.srt.SE.tagAlign.narrowPeak.gz

Gappedpeaks file (rep1_gappedpeaks)
R1.raw.srt.filt.nodup.srt.SE...

Signal track fold-enrichment over contr
R1.raw.srt.filt.nodup.srt.SE-

R2.raw.srt.filt.nodup.srt.SE...

Broadpeaks file (pooledpr2_broadpeaks)

R1.raw.srt.filt.nodup.srt.SE...

R2.raw.srt.filt.nodup.srt.SE...

Peak.gz

The screenshot shows a file management interface with a dark blue header containing 'Manage', 'Monitor', and 'Visualize' tabs. Below the header is a toolbar with buttons for 'Add Data', 'New Folder', 'New Workflow', 'Start Analysis...', 'Copy', 'Delete', 'Info', and 'Download'. A search bar is present with 'SEARCH SCOPE' set to 'Current folder' and 'NAME' set to '.fc_signal|narrowPeak.bb'. A file list is displayed with columns for 'Name', 'ID', 'CLASS', 'MODIFIED', and 'TAGS'. A red arrow points from the 'Name' column of the file list to the 'encode_macs2' folder in the left-hand navigation pane.

SEARCH SCOPE	NAME	ID	CLASS	MODIFIED	TAGS
Current folder	.fc_signal narrowPeak.bb	Any	Any (5)	Any	Any

Name
<input checked="" type="checkbox"/> R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.fc_signal.bw
<input type="checkbox"/> R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.narrowPeak.bb
<input type="checkbox"/> R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.rejected.narrowPeak.bb
<input checked="" type="checkbox"/> R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.replicated.narrowPeak.bb
<input type="checkbox"/> R1.raw.srt.filt.nodup.srt.SE.SE.pr1-R2.raw.srt.filt.nodup.srt.SE.SE.pr1_pooled.tagAlign.fc_signal...
<input type="checkbox"/> R1.raw.srt.filt.nodup.srt.SE.SE.pr1.tagAlign.fc_signal.bw
<input type="checkbox"/> R1.raw.srt.filt.nodup.srt.SE.tagAlign.fc_signal.bw
<input type="checkbox"/> R1.raw.srt.filt.nodup.srt.SE.tagAlign.gz.SE.pr2-R2.raw.srt.filt.nodup.srt.SE.tagAlign.gz.SE.pr2_p...



Results from the ChIP-seq exercise

“Download” to generate temporary URL’s to the selected files

The screenshot displays a web-based file management interface. At the top, there are tabs for 'Manage', 'Monitor', and 'Visualize'. Below these is a toolbar with buttons for 'Add Data', 'New Folder', 'New Workflow', 'Start Analysis...', 'Copy', 'Delete', 'Info', and 'Download'. A red arrow points to the 'Download' button. Below the toolbar, there are search filters for 'SEARCH SCOPE' (Current folder), 'NAME' (.fc_signal.bw|narrowPeak.bb), 'ID' (Any), 'CLASS' (Any 5), 'MODIFIED' (Any), and 'TAGS' (Any). On the left, a sidebar shows a folder structure under 'APBC 2016', including 'histone-chip-seq', 'encode_bwa', 'encode_macs2', 'test_data', and 'Reference Files'. The main area shows a table of files with columns for 'Name', 'T...', 'S...', and 'C...'. Two files are selected, indicated by blue checkmarks in the first column:

<input type="checkbox"/>	Name ^	T...	S...	C...
<input checked="" type="checkbox"/>	R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.fc_signal.bw	File	4...	J
<input type="checkbox"/>	R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.narrowPeak.bb	File	1...	J
<input type="checkbox"/>	R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.rejected.narrowPeak.bb	File	8...	J
<input checked="" type="checkbox"/>	R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.replicated.narrowPeak.bb	File	1...	J
<input type="checkbox"/>	R1.raw.srt.filt.nodup.srt.SE.SE.pr1-R2.raw.srt.filt.nodup.srt.SE.SE.pr1_pooled.tagAlign.fc_signal...	File	3...	J

Results from the ChIP-seq exercise


“Download” to generate temporary URL’s to the selected files


The screenshot displays a web-based file management interface. At the top, there are tabs for 'Manage', 'Monitor', and 'Visualize'. Below these is a toolbar with buttons for 'Add Data', 'New Folder', 'New Workflow', and 'Start Analysis...'. A secondary toolbar contains 'Copy', 'Delete', 'Info', and 'Download' buttons. A red arrow points to the 'Download' button. Below the toolbars, there are search filters for 'SEARCH SCOPE' (Current folder), 'NAME' (.fc_signal.bw|narrowPeak.bb), 'ID' (Any), 'CLASS' (Any, 5), 'MODIFIED' (Any), and 'TAGS' (Any). On the left, a sidebar shows a folder tree with 'APBC 2016' expanded to show 'histone-chip-seq', 'encode_bwa', 'encode_macs2', and 'test_data'. The main area shows a table of files with columns for 'Name', 'T...', 'S...', and 'C...'. Two files are selected with blue checkboxes:

<input type="checkbox"/>	Name ^	T...	S...	C...
<input checked="" type="checkbox"/>	R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.fc_signal.bw	File	4...	J
<input type="checkbox"/>	R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.narrowPeak.bb	File	1...	J
<input type="checkbox"/>	R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.rejected.narrowPeak.bb	File	8...	J
<input checked="" type="checkbox"/>	R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.replicated.narrowPeak.bb	File	1...	J
<input type="checkbox"/>	R1.raw.srt.filt.nodup.srt.SE.SE.pr1-R2.raw.srt.filt.nodup.srt.SE.SE.pr1_pooled.tagAlign.fc_signal...	File	3...	J

Visualize on the UCSC Genome Browser

Get Your Data

 Download files

 Get bulk URLs

These links will remain active for 24 hours. **Warning: anybody with these links can download your files without additional authentication, so please be careful when you share this list!**

You may also [download this list as a text file](#).

```
https://dl.dnanex.us/F/D/3ZQ061X6bP19y40P2PpFvz8qz9pPZK3B2XfyyxY3/R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.fc_signal.bw  
https://dl.dnanex.us/F/D/7ZB2KXZpxyQp5Y92k2G59bvY79VXQqj6pgpvP6pk/R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.replicated.narrowPeak.bb
```

Close

genome.ucsc.edu

UCSC Genome Bioinformatics

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

Genome About the UCSC Genome Bioinformatics Site

- Sessions
- Track Hubs
- Custom Tracks

Paste URLs or data: Or upload: No file chosen

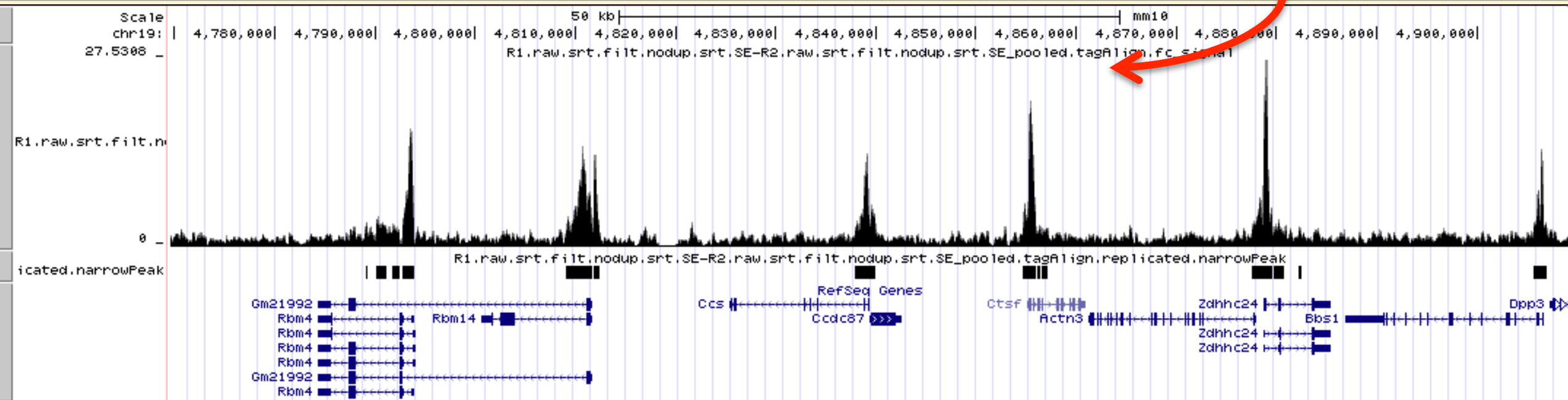
```
https://dl.dnanex.us/F/D/3ZQ061X6bP19y40P2PpFvz8qz9pPZK3B2XfyyxY3/R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.fc_signal.bw  
https://dl.dnanex.us/F/D/7ZB2KXZpxyQp5Y92k2G59bvY79VXQqj6pgpvP6pk/R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.replicated.narrowPeak.bb
```

Visualize on the UCSC Genome Browser

Manage Custom Tracks

genome Mouse assembly Dec. 2011 (GRCm38/mm10) [mm10]

Name	Description	Type	Doc	delete	view in
R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.fc_signal	R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.fc_signal	bigWig		<input type="checkbox"/>	Genome Browser go
R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.replicated.narrowPeak	R1.raw.srt.filt.nodup.srt.SE-R2.raw.srt.filt.nodup.srt.SE_pooled.tagAlign.replicated.narrowPeak	bigBed		<input type="checkbox"/>	add custom



Pipeline Workshop Summary

DCC Goals:

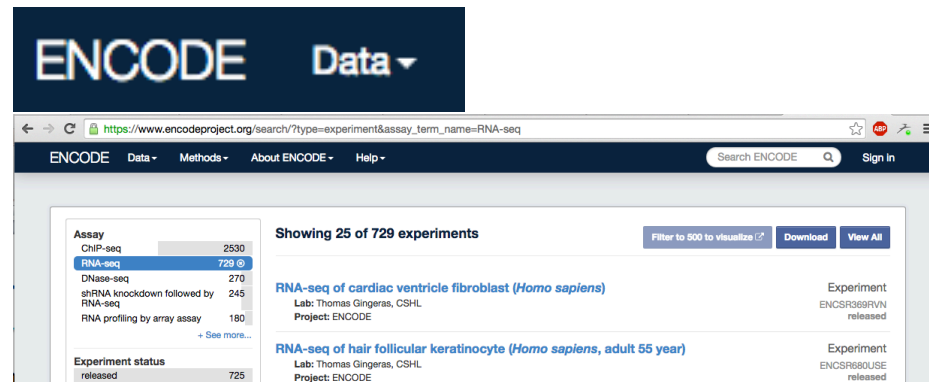
1. Deploy ENCODE-defined pipelines for ChIP-seq, RNA-seq, DNase-seq, methylation.
2. Use those pipelines to generate the standard ENCODE peaks, quantitations, CpG.
3. Capture metadata to make clear what software, versions, parameters, inputs were used.
4. Capture, accession, and distribute the output.
5. Deliver *exactly the same* pipelines in a form that *anyone can run* on their data or with ENCODE data – one experiment or 1000.

Replicability – Provenance – Ease of Use – Scalability

DNAneXus

Featured projects

- ENCOD Uniform Processing Pipelines
- Parliament



ENCOD Data

Showing 25 of 729 experiments

Assay	Count
ChIP-seq	2530
RNA-seq	729
DNase-seq	270
shRNA knockdown followed by RNA-seq	245
RNA profiling by array assay	180

Experiment status released 725

RNA-seq of cardiac ventricle fibroblast (*Homo sapiens*)
Lab: Thomas Gingeras, CSHL
Project: ENCODE
Experiment ENCSR368RVN released

RNA-seq of hair follicular keratinocyte (*Homo sapiens*, adult 55 year)
Lab: Thomas Gingeras, CSHL
Project: ENCODE
Experiment ENCSR680USE released



GitHub, Inc. [US] <https://github.com/ENCODE-DCC>

Search GitHub

ENCOD DCC

enocode-help@lists.stanford.edu

Contributors

ENCODE Data Coordinating Center

Mike Cherry, PI, Stanford
Jim Kent, co-PI, UCSC
Eurie Hong, Project Manager
Pipeline Developers
Ben Hitz, WGBS, Software Lead
Tim Dreszer, RNA-seq, DNase-seq
J. Seth Strattan, CHIP-seq

Portal Developers

Laurence Rowe
Nikhil Podduturi
Forrest Tanaka

Data Wranglers

Esther Chan
Jean Davidson
Venkat Malladi
Cricket Sloan
J. Seth Strattan

QA & Biocuration Assistance

Brian Lee
Marcus Ho
Aditi Narayanan
Support Staff
Stuart Miyasato
Matt Simison
Zhenhua Wang



@encodedcc



encode-help@lists.stanford.edu

ENCODE Data Analysis Center

Zhiping Weng, PI, University of Massachusetts
Mark Gerstein, co-PI, Yale

Methylation

Junko Tsuji, U Mass
Eric Mendenhall, U Alabama, HAIB

RNA-seq

Alex Dobin, CSHL
Carrie Davis, CSHL
Rafael Irizarry, Harvard
Xintao Wei, UConn
Brent Gravely, UConn
Colin Dewey, U Wisconsin
Roderic Guigó, CRG
Sarah Djebali, CRG

ChIP-seq

Anshul Kundaje, Stanford
Nathan Boley, Stanford
Jin Lee, Stanford

DNAnexus

Mike Lin
Andey Kislyuk
Singer Ma
Brett Hannigan
Ohad Rodeh
Joe Dale
George Asimenos

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

