ENCODE Data and Metadata through the ENCODE Portal

J. Seth Strattan, ENCODE DCC, Stanford

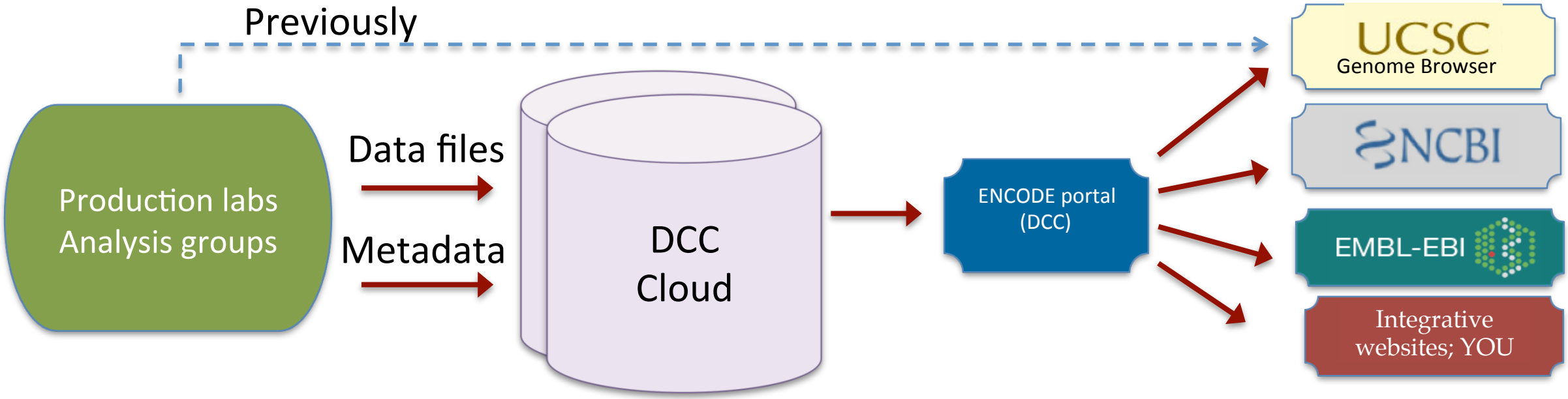
Keystone Symposia Epigenomics & Methylation

ENCODE Workshop

April, 2015

ENCODE Data Coordinating Center (DCC)

ENCODE data flow



Role:	**Data generation**	**Data organization**	**Data access**
Tasks:	Perform assays & analyses	Data processing & validation	Web-based searches
	Validate data	Data file storage	Data downloads
	Submit data and metadata	Metadata curation	

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.

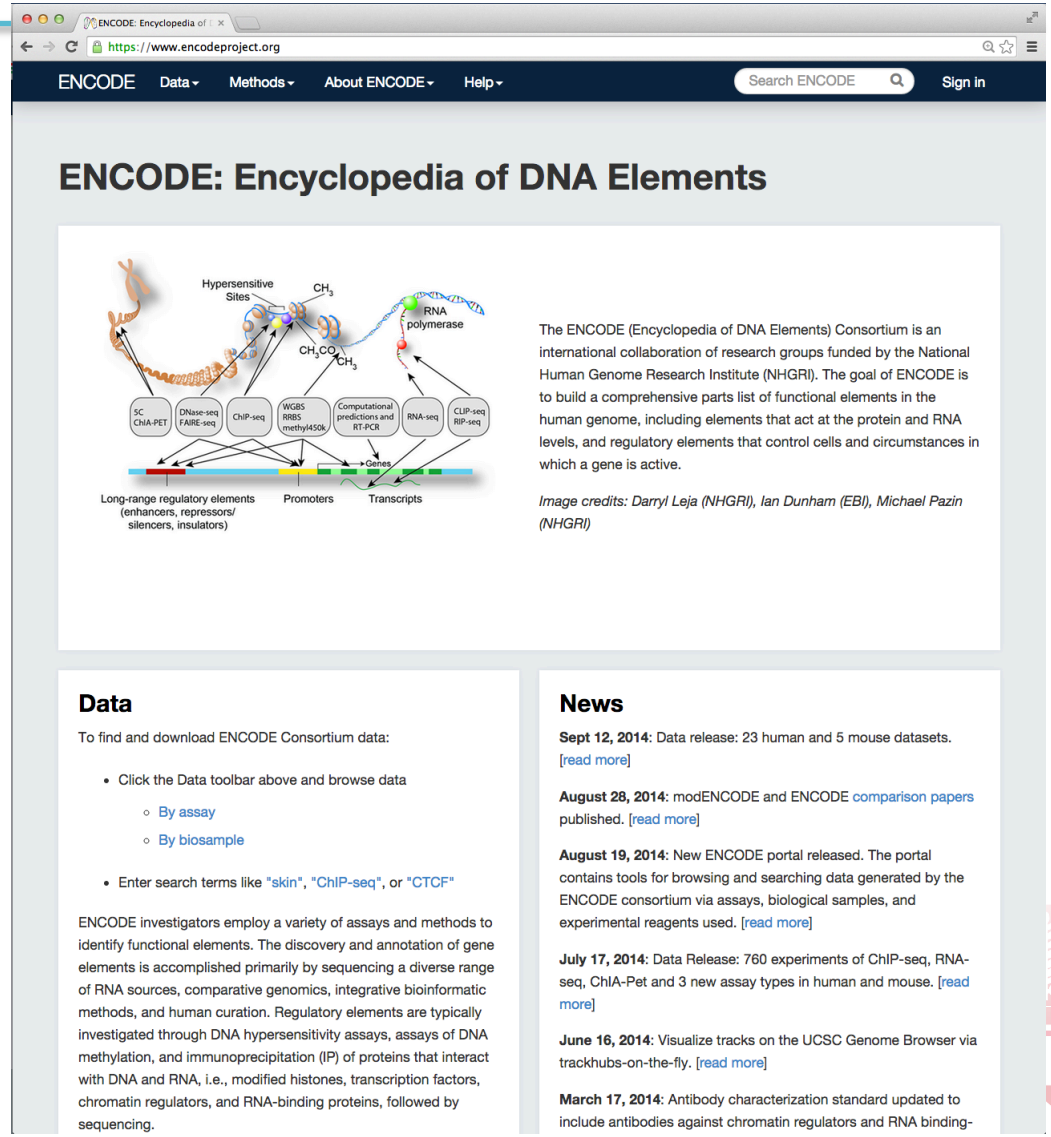
- Metadata describing experiments and analyses can be **hard to search**.

- 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.



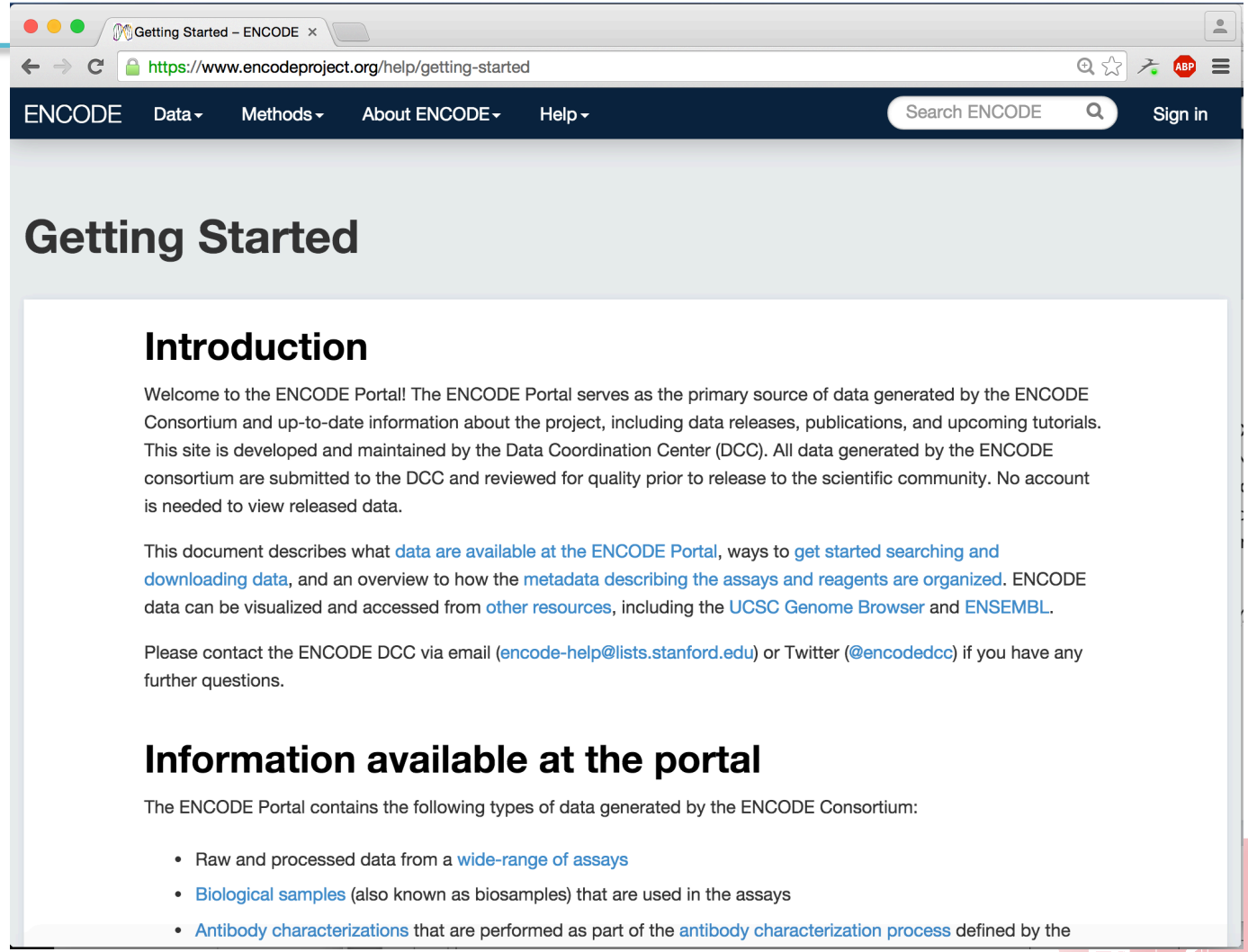
encodeproject.org: Features of the Portal

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

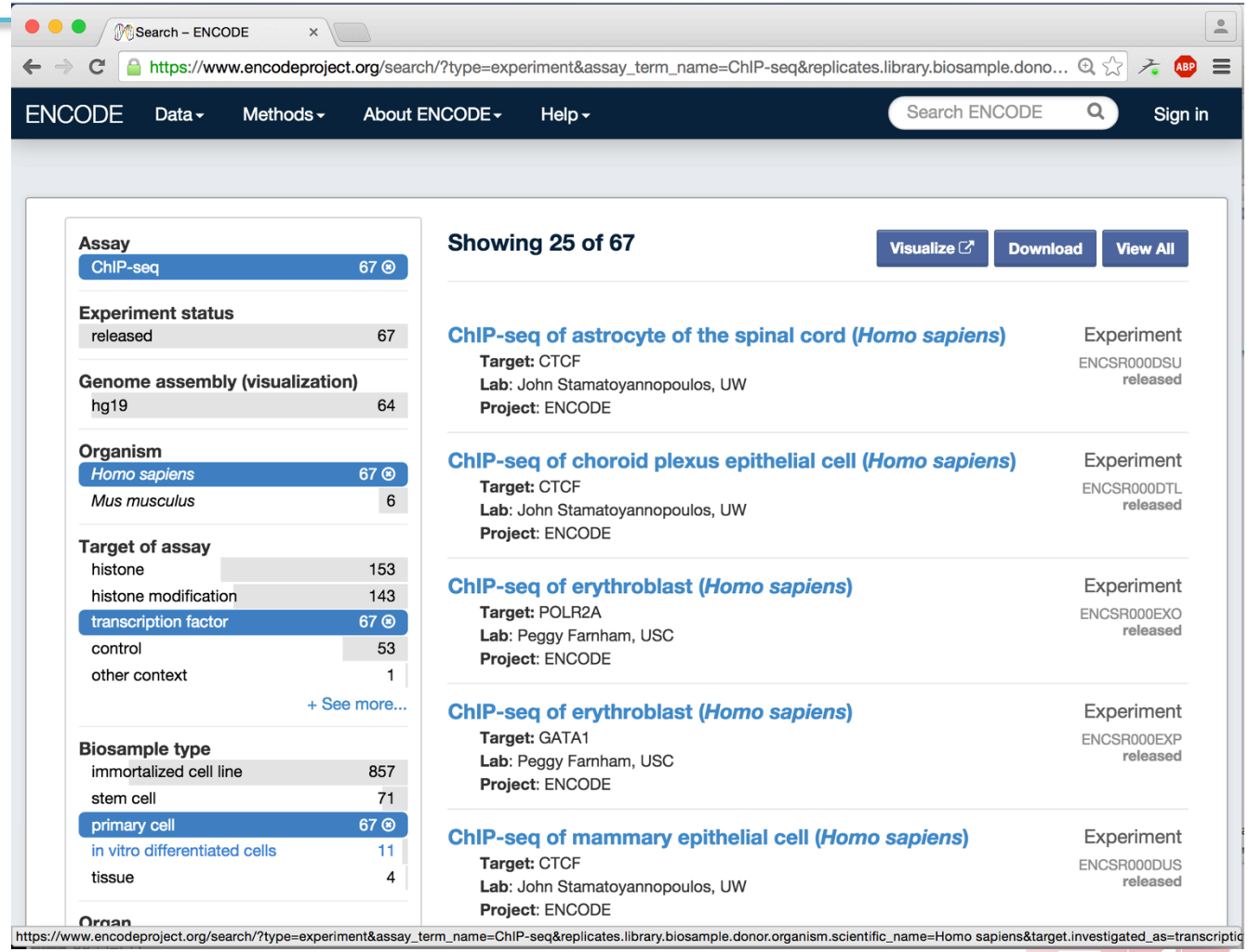
- https://www.encodeproject.org/
- https://www.encodeproject.org/help/getting-started
  - A good place to start
- https://www.encodeproject.org/tutorials
  - Links to slides from today's workshop
- Help … Contact
  - encode-help@lists.stanford.edu
  - @encodedcc



Faceted Browsing

Faceted Browsing of Assays

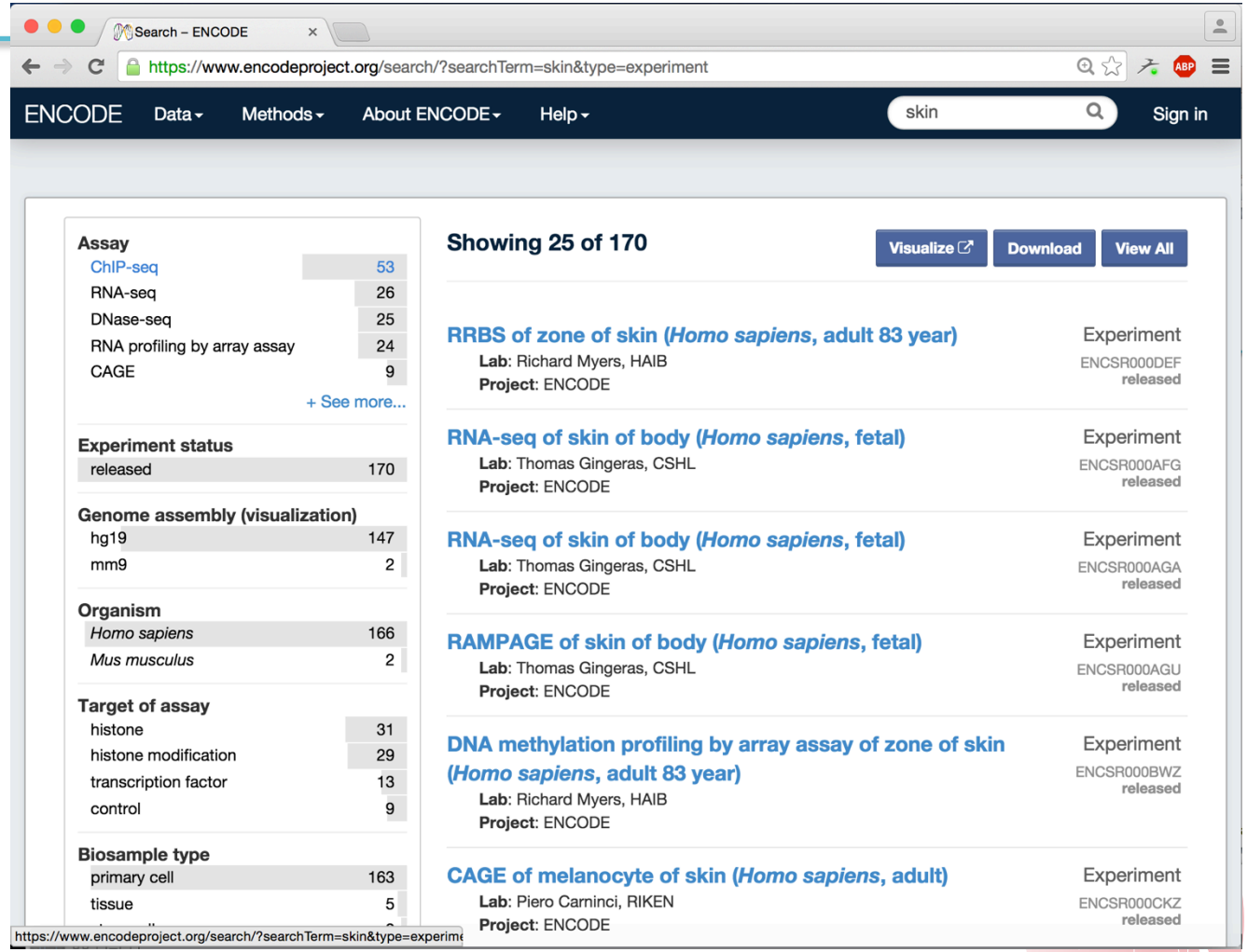
- Click on Data … Assays
- https://www.encodeproject.org/search/?type=experiment
- Facets on the left are “filters”
- Limits to items of interest
- Facets from the same category can be stacked to combine results (logical OR)
- Facets from different categories further restrict results (logical AND)



Search

Search across ENCODE metadata

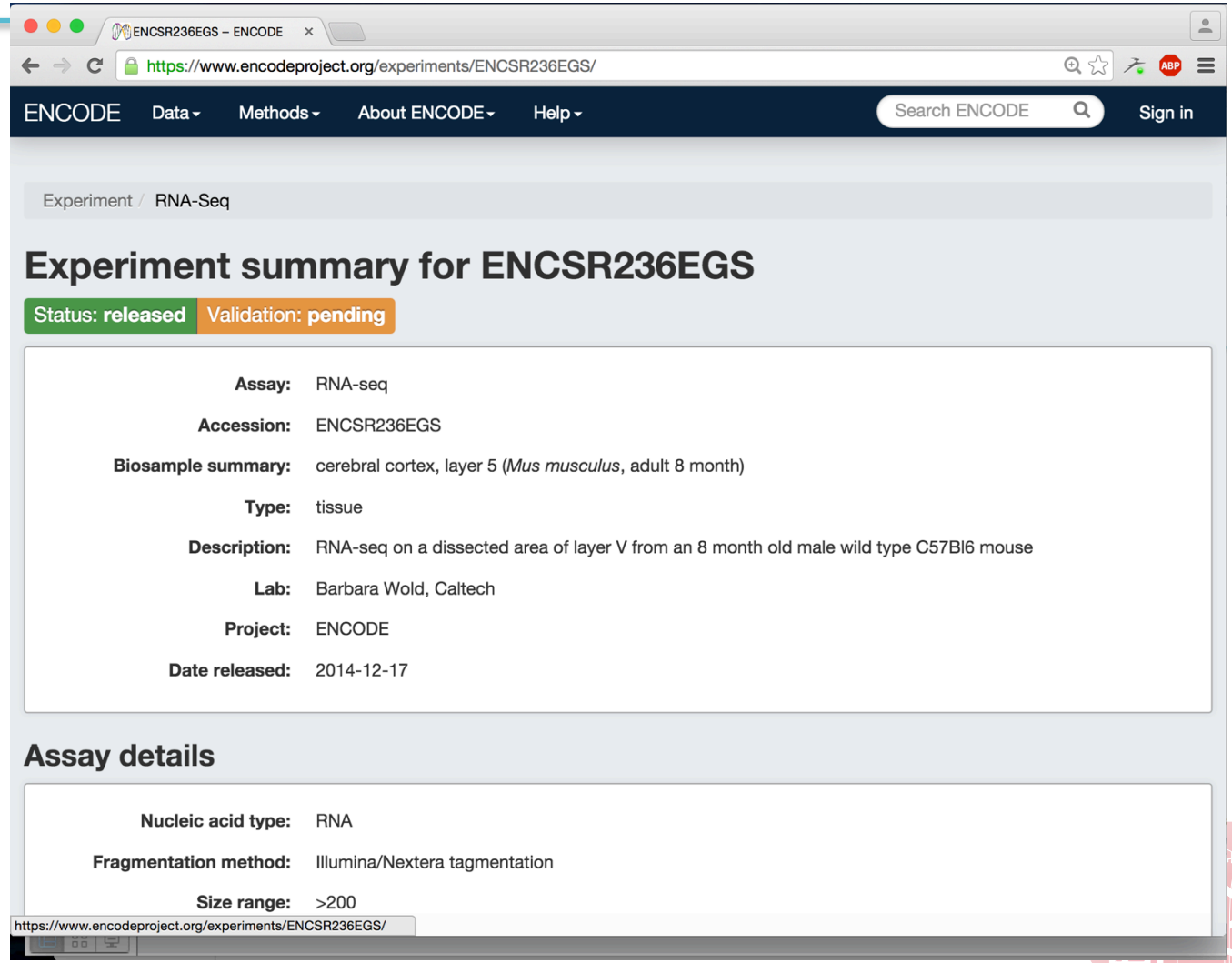
- In search box type “skin” https://www.encodeproject.org/search/?searchTerm=skin
- Results include experiments, biosamples, etc
- Select experiments – see several assays
- Facet further by RNA-seq
- Note this is not just a pure text search
- “skin of body” comes up, but so does “keratinocyte”
- Search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



Metadata to Data

Use metadata to find data

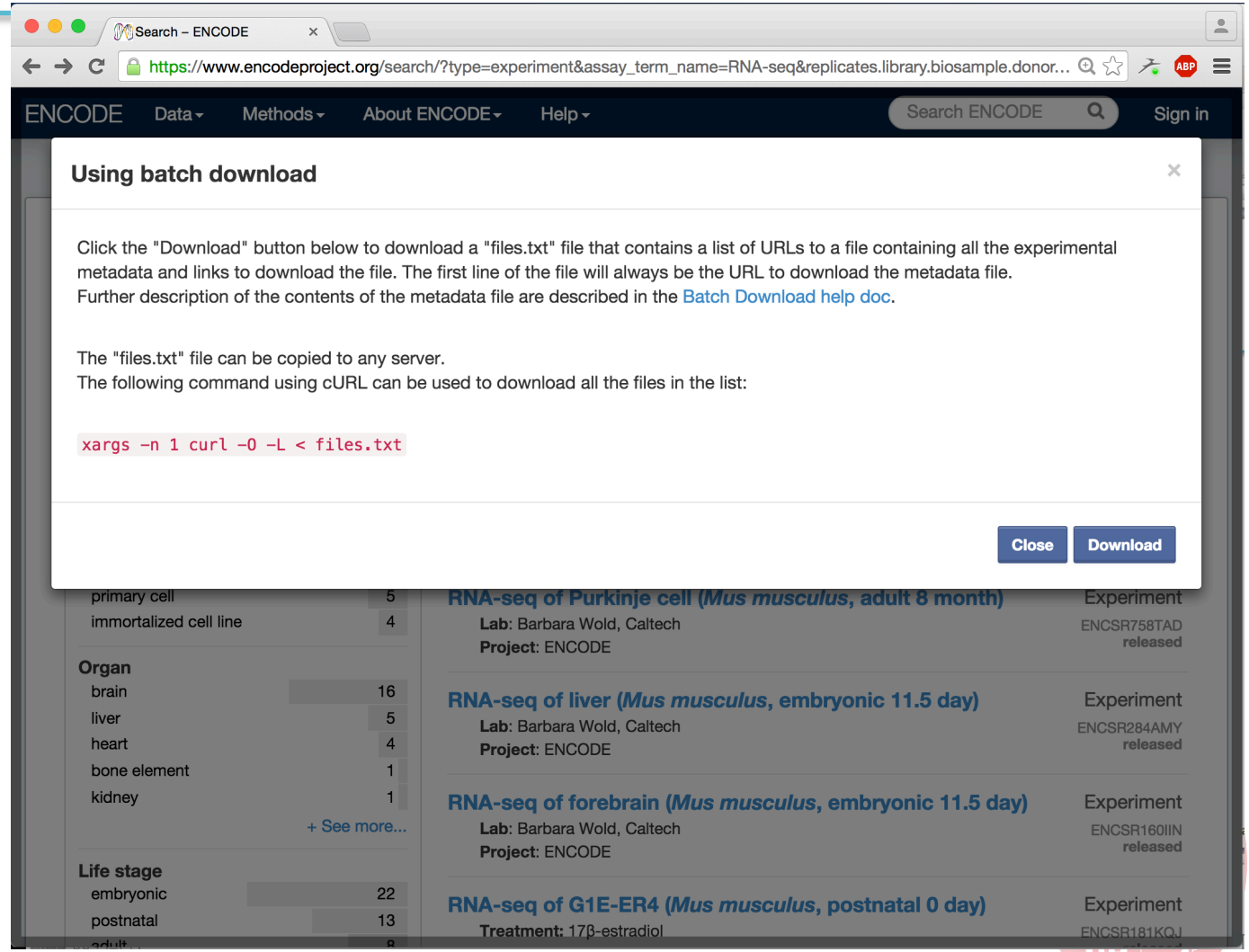
- Select Data … Assays
- Facet on RNAseq; mouse; mm10 assembly
- Select an experiment, for example https://www.encodeproject.org/experiments/ENCSR236EGS/
- Note metadata on protocols, replicates
- Graph: files are related by processing steps
- Download from the graph or a list
- Not all experiments have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



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`



Batch Download – files.txt

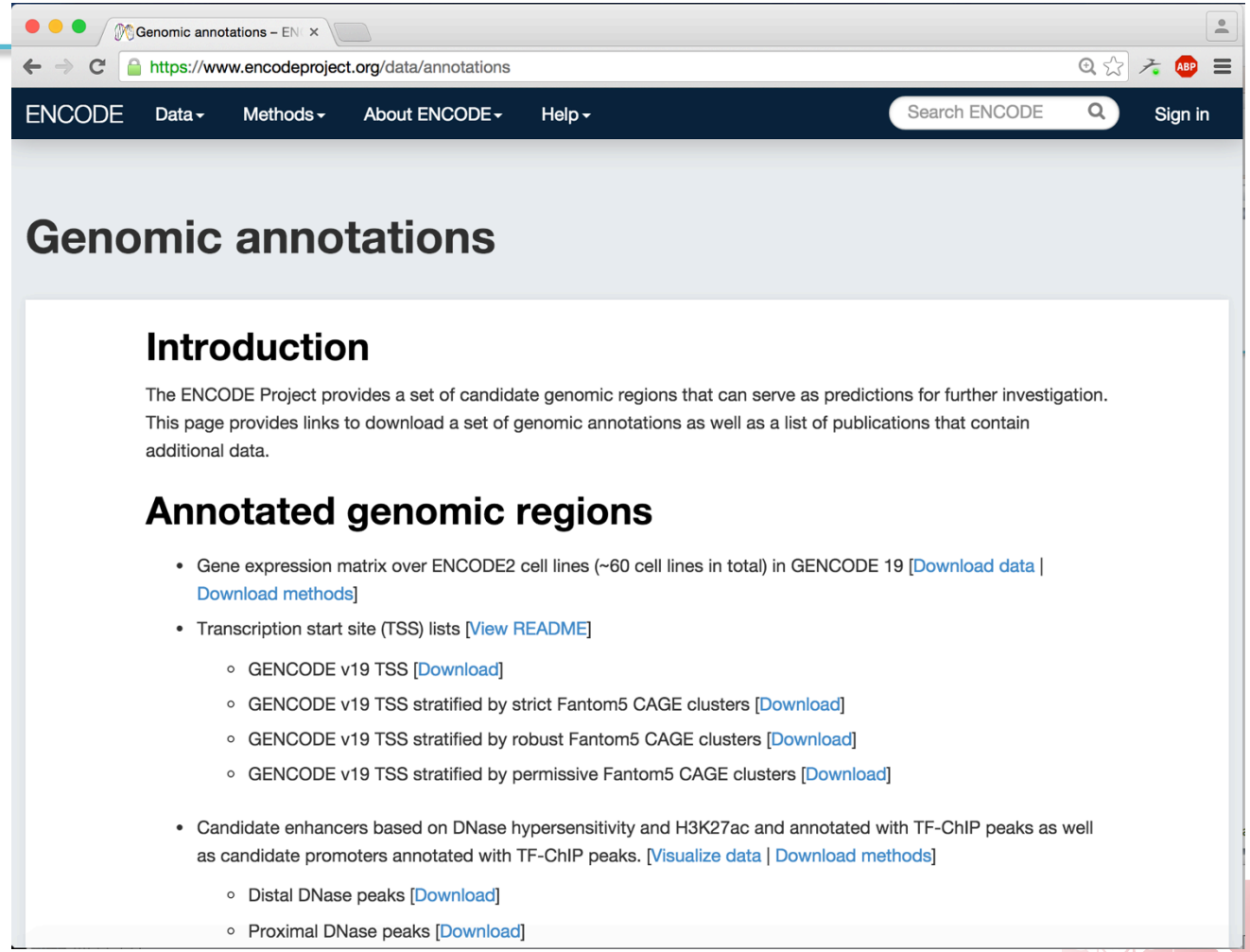
```
1  https://www.encodeproject.org/metadata/type=experiment&assay_term_name=ChIP-seq&target.
   investigated_as=transcription%20factor&replicates.library.biosample.donor.organism.
   scientific_name=Homo%20sapiens&replicates.library.biosample.
   biosample_type=in%20vitro%20differentiated%20cells/metadata.tsv
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

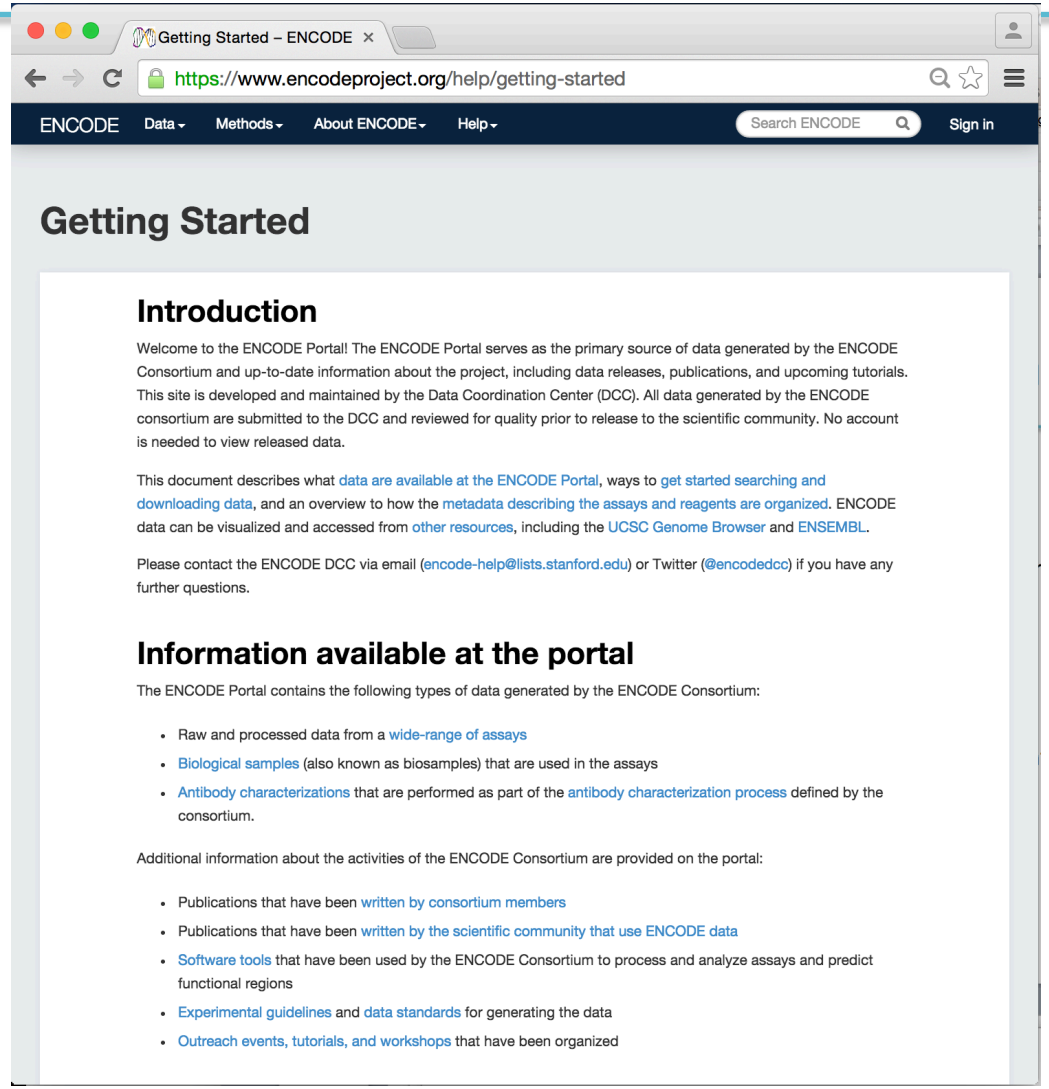
- Select Data … Annotations https://www.encodeproject.org/data/annotations
- Links to region-based annotations (e.g. TSS's)
- Links to enhancer predictions and methods
- Links to integrative ENCODE publications that feature annotation datasets



Summary of Interactive Features of the Portal

https://www.encodeproject.org/

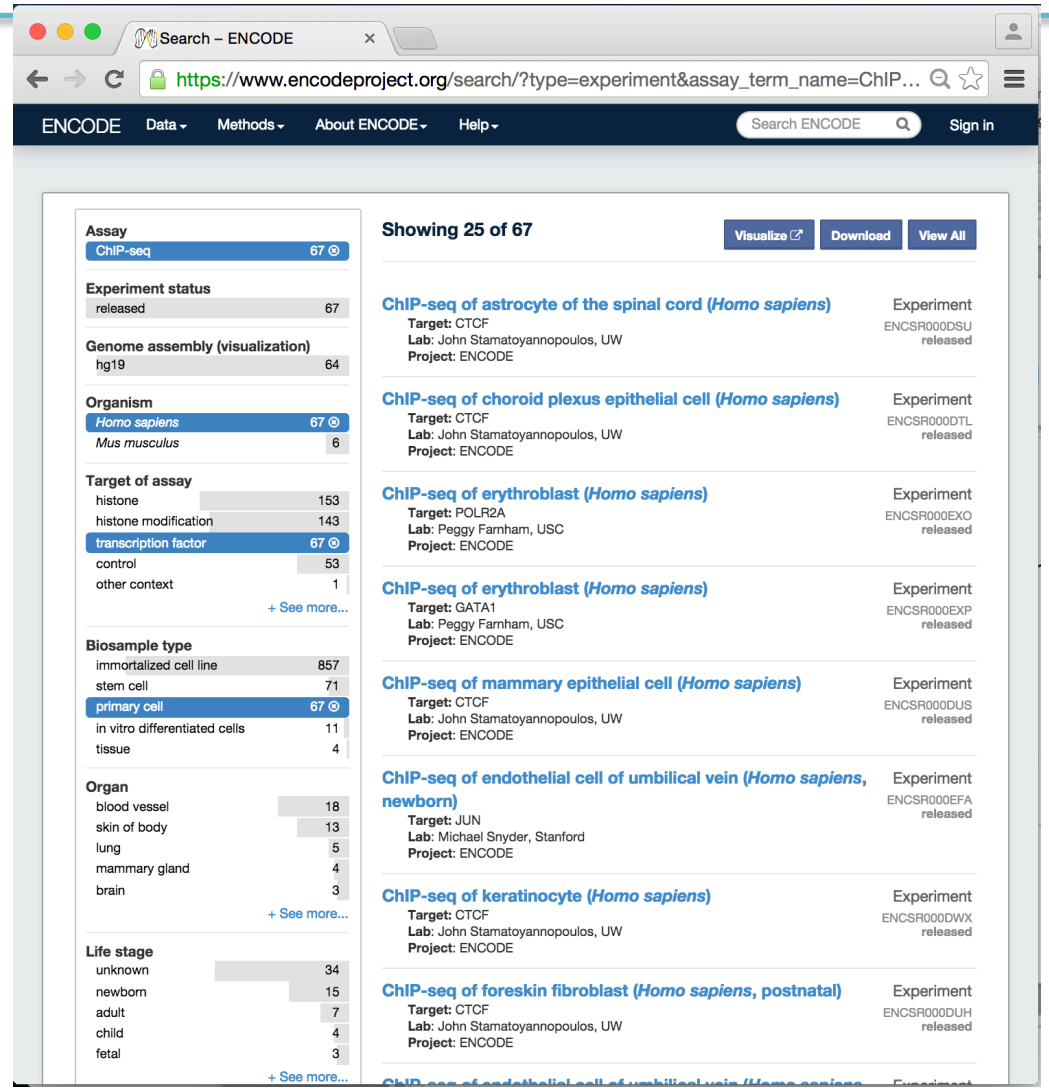
- Documentation and links to tutorials
- Faceted browsing: experiments and samples
- Detailed experiment metadata
- Visualize file relationships
- Interactive download of selected files
- Batch download of selected experiments
- Access to ENCODE annotations



Summary of Interactive Features of the Portal

https://www.encodeproject.org/

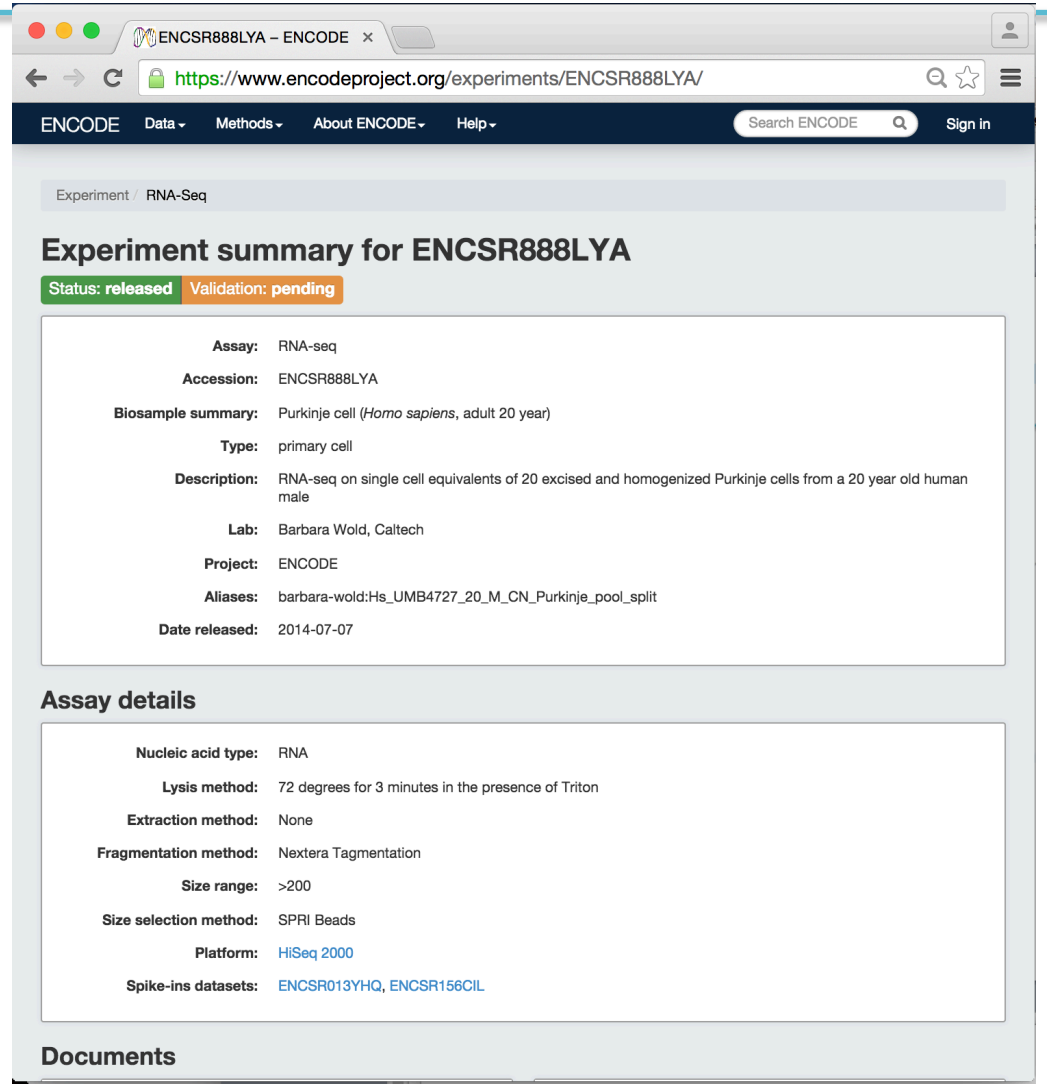
- Documentation and links to tutorials
- Faceted browsing: experiments and samples
- Detailed experiment metadata
- Visualize file relationships
- Interactive download of selected files
- Batch download of selected experiments
- Access to ENCODE annotations



Summary of Interactive Features of the Portal

https://www.encodeproject.org/

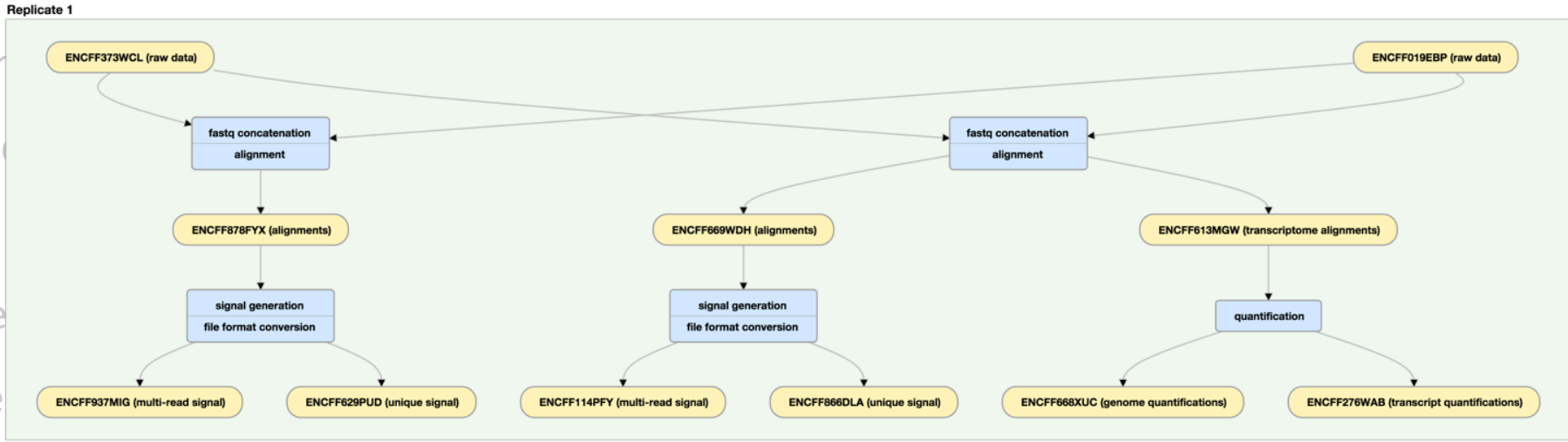
- Documentation and links to tutorials
- Faceted browsing: experiments and samples
- Detailed experiment metadata
- Visualize file relationships
- Interactive download of selected files
- Batch download of selected experiments
- Access to ENCODE annotations



Summary of Interactive Features of the Portal

https://www.encodeproject.org/

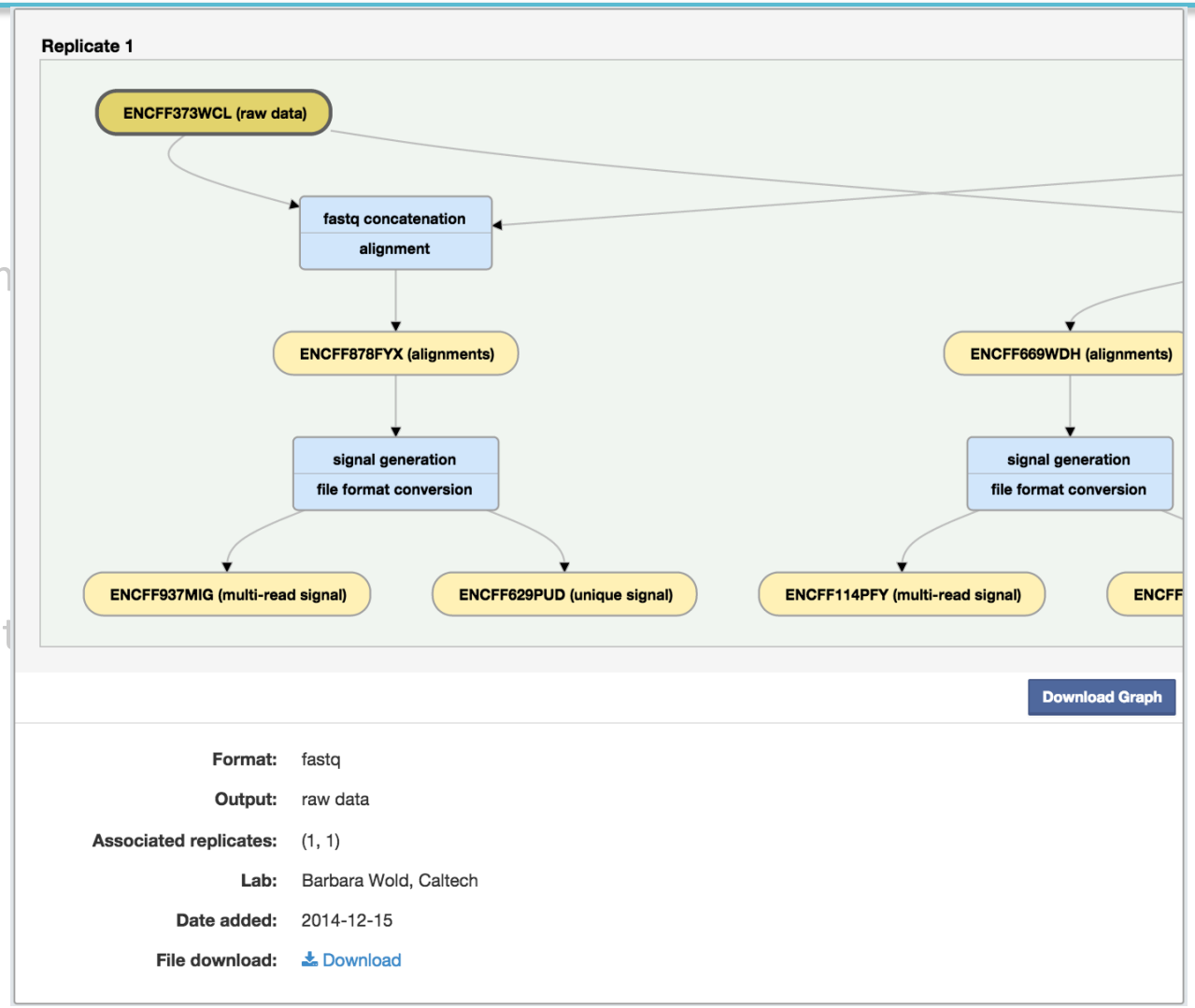
- Documentation and links to tutorials
- Faceted browsing: experim
- Detailed experiment meta
- Visualize file relationships
- Interactive download of se
- Batch download of selecte
- Access to ENCODE annotations



Summary of Interactive Features of the Portal

https://www.encodeproject.org/

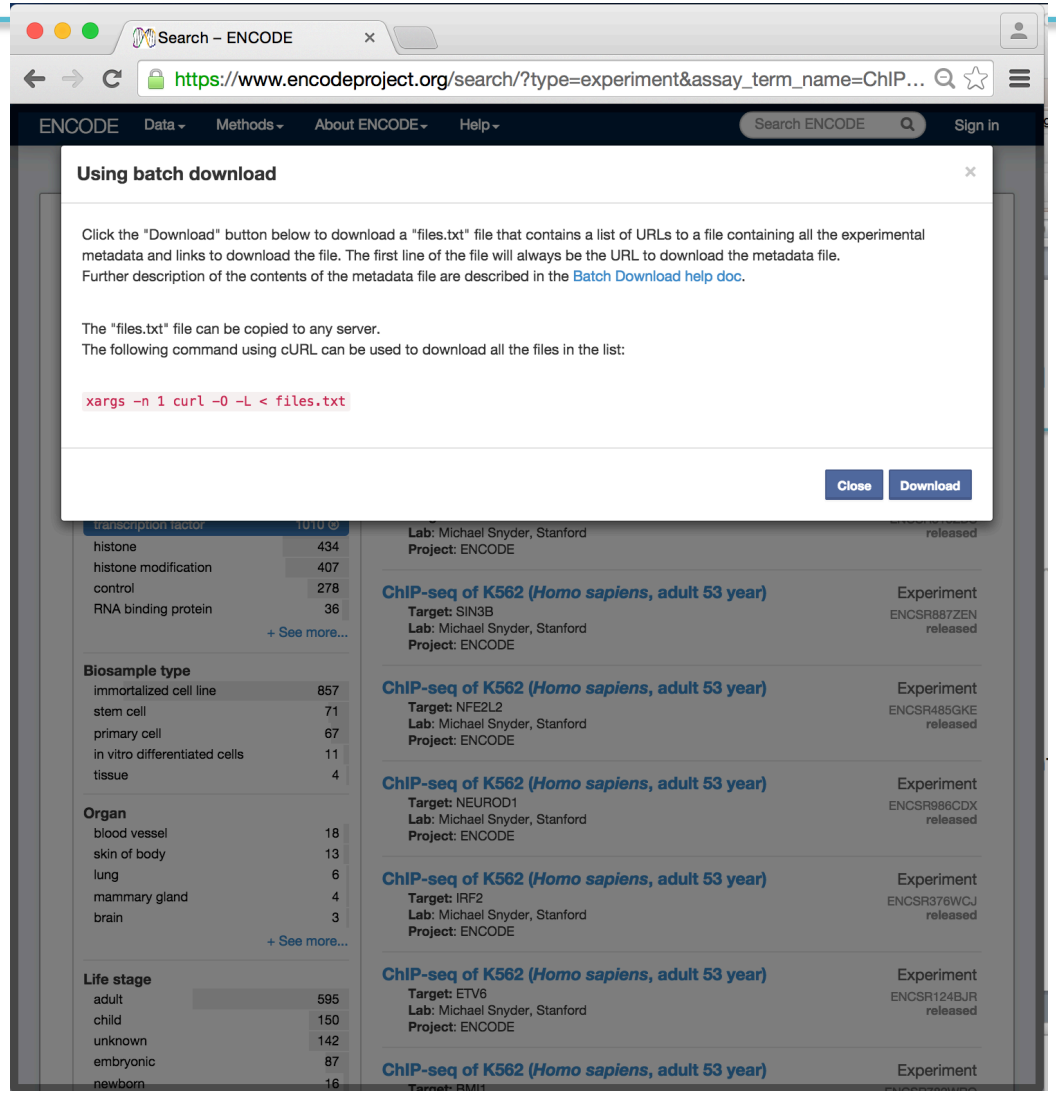
- Documentation and links to tutorials
- Faceted browsing: experiments and san
- Detailed experiment metadata
- Visualize file relationships
- Interactive download of selected files
- Batch download of selected experiment
- Access to ENCODE annotations



Summary of Interactive Features of the Portal

https://www.encodeproject.org/

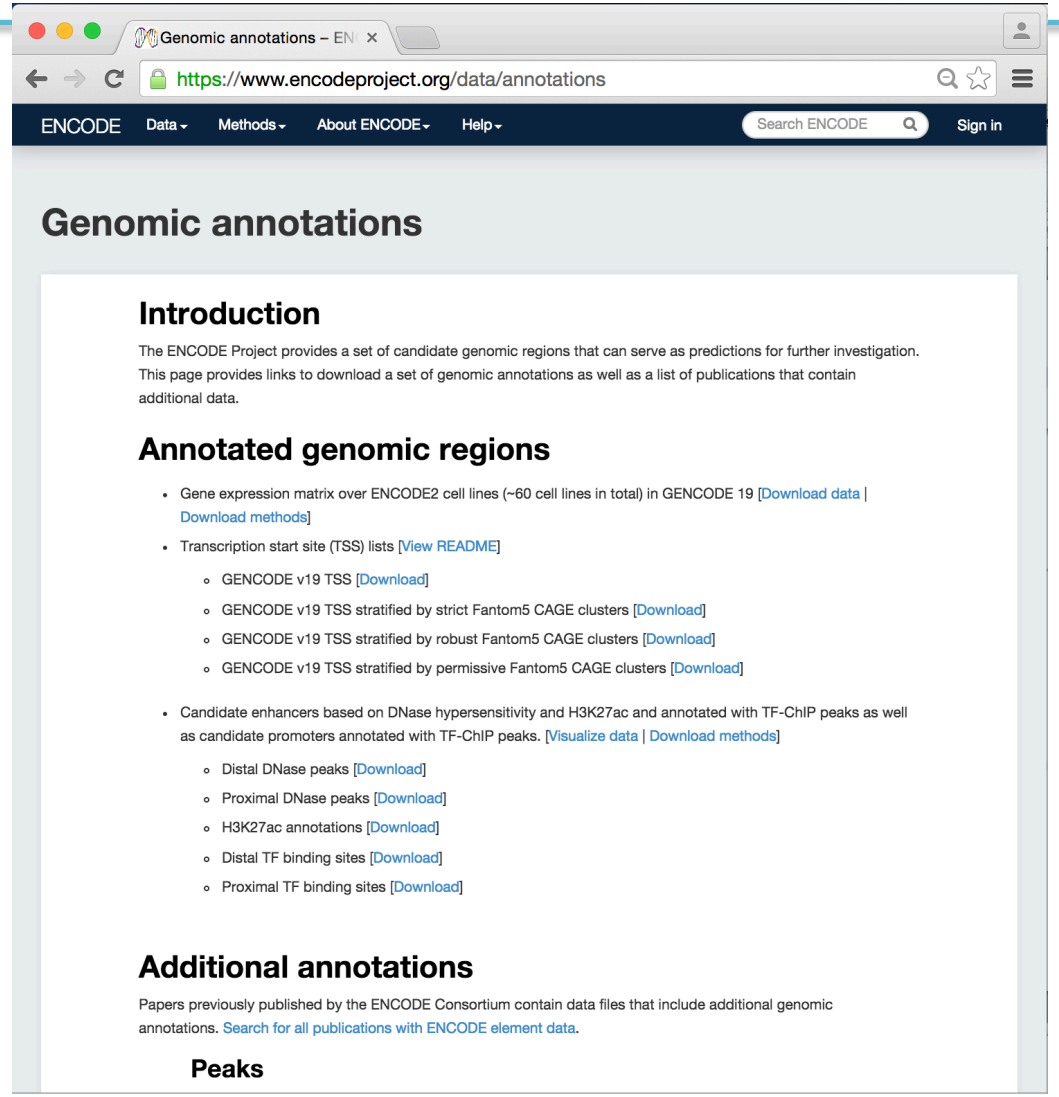
- Documentation and links to tutorials
- Faceted browsing: experiments and samples
- Detailed experiment metadata
- Visualize file relationships
- Interactive download of selected files
- Batch download of selected experiments
- Access to ENCODE annotations



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

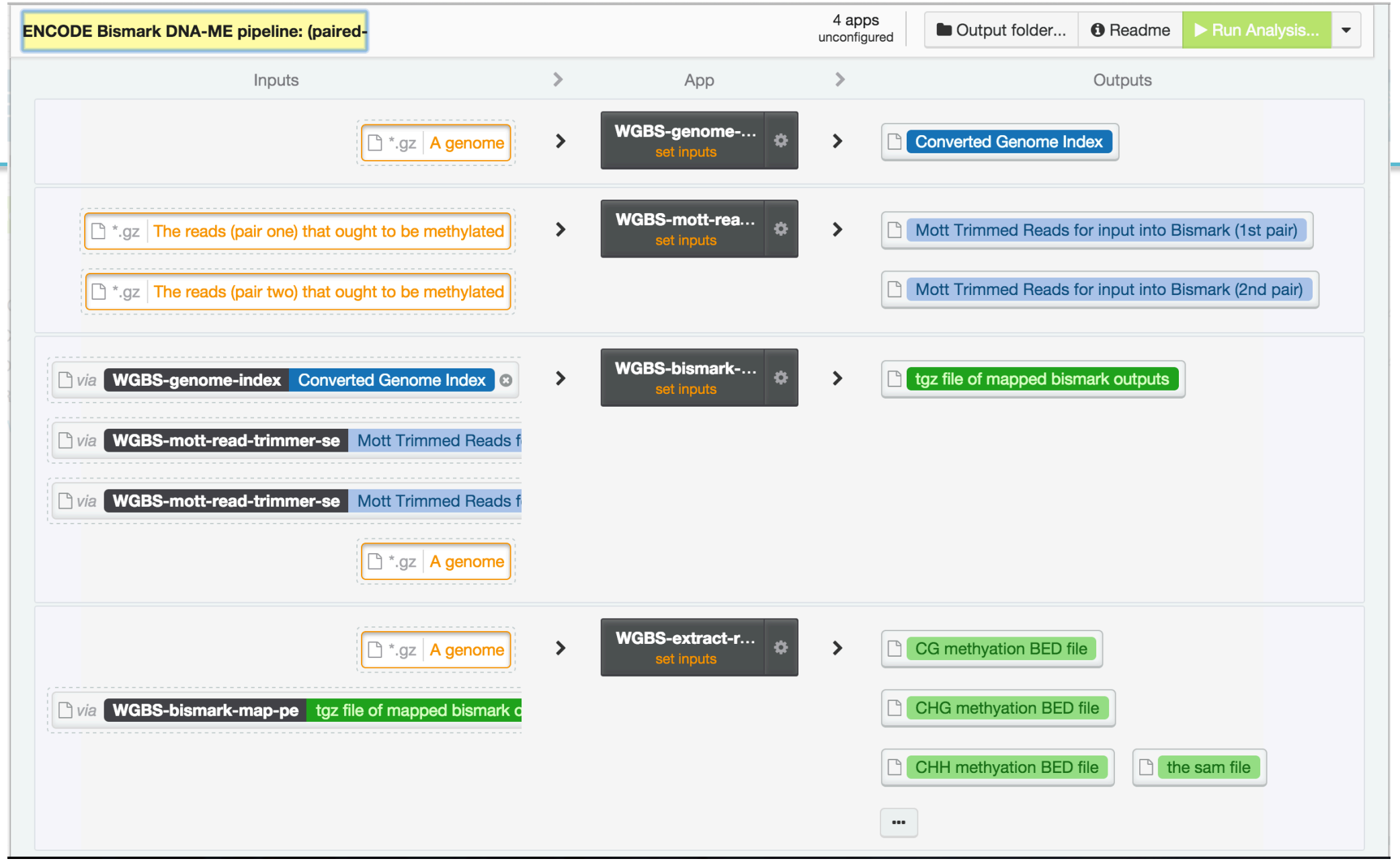


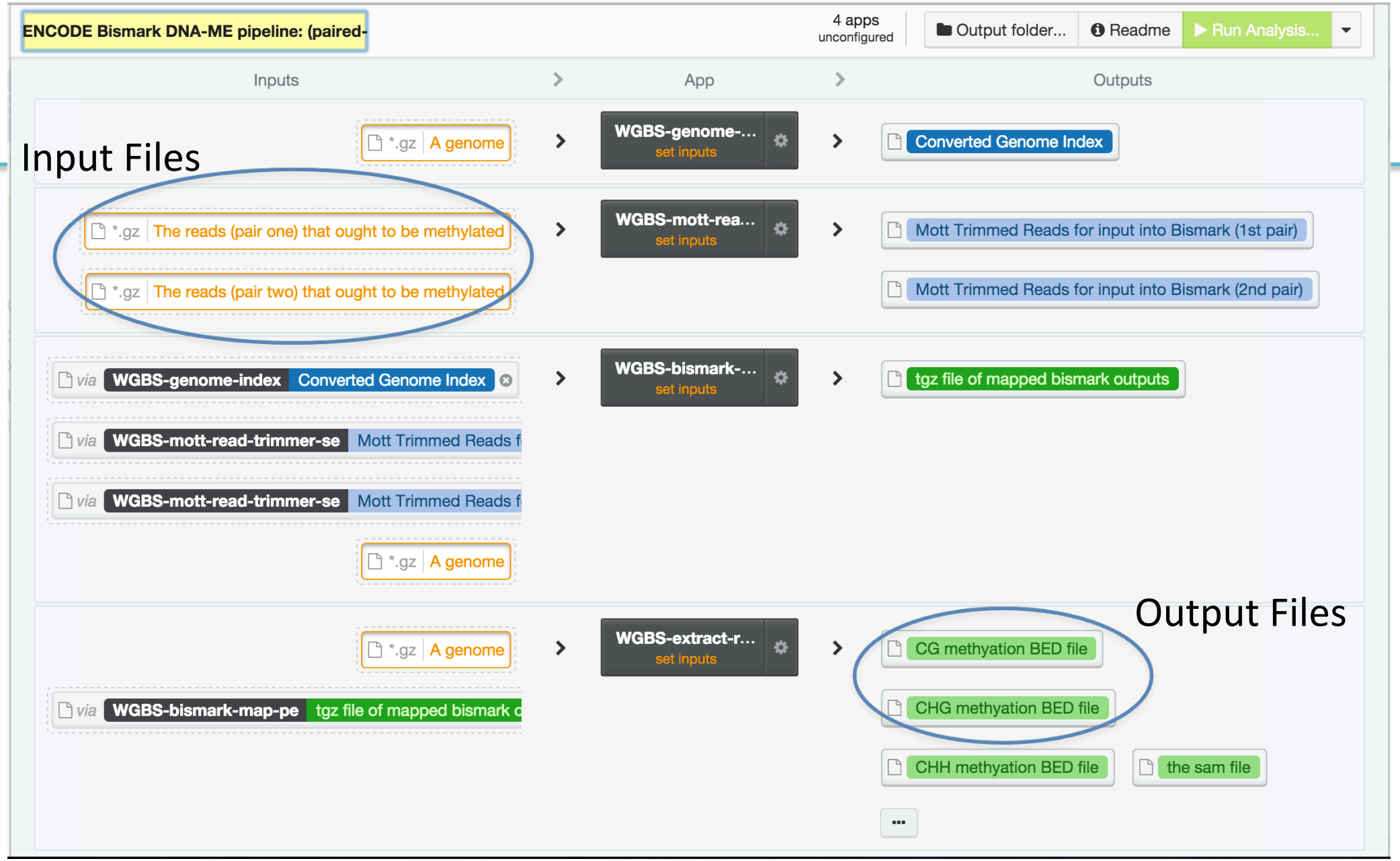
ENCODE Analysis Pipelines as Deliverables

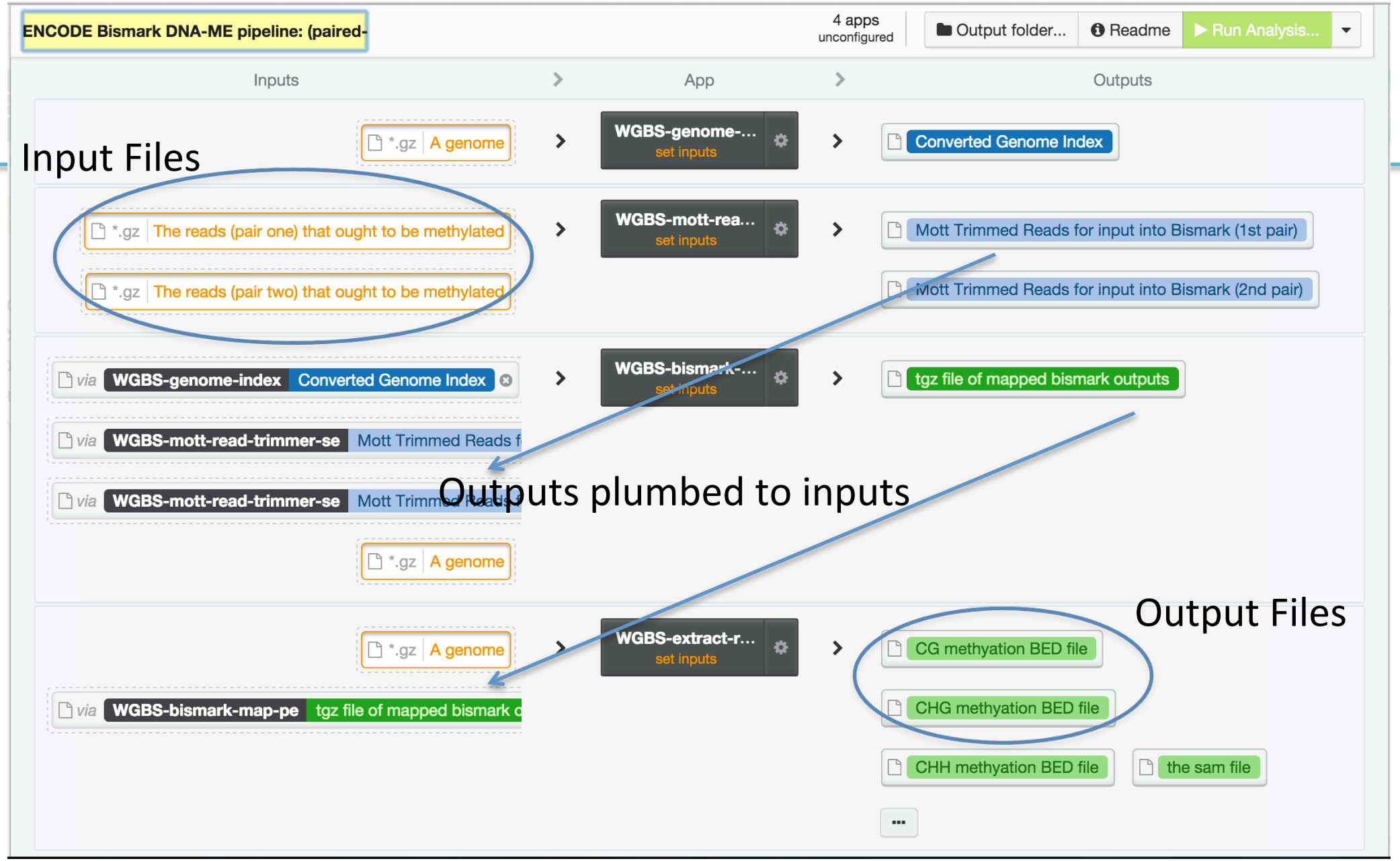
Goals:

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
We Hope


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 methyation BED file
via
WGBS-bismark-map-pe
tgz file of mapped bismark o
CHG methyation BED file
CHH methyation BED file
the sam file
...

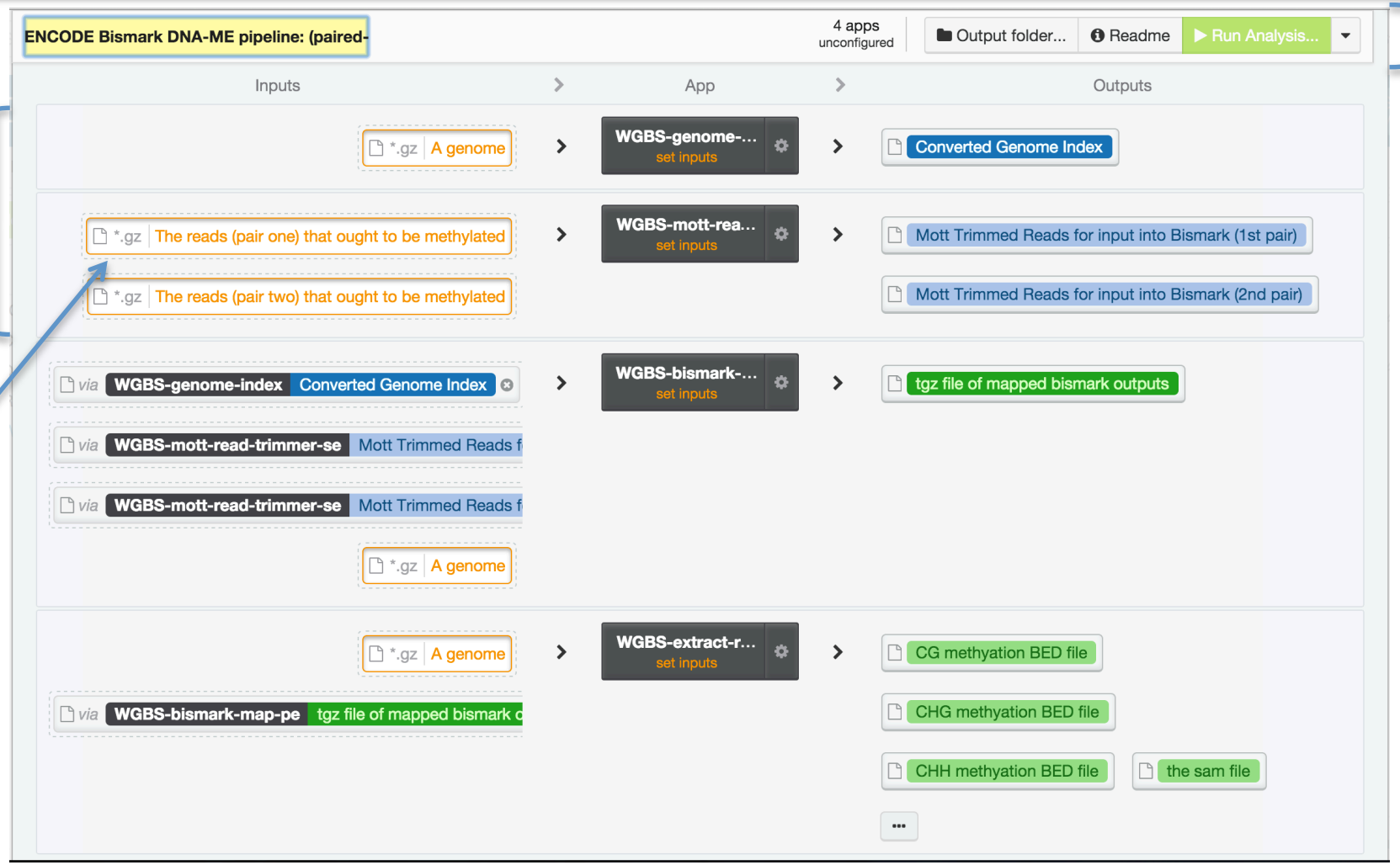

ENCODE Bismark DNA-ME pipeline: (paired-
4 apps unconfigured
Output folder...
Readme
Run Analysis...
Inputs
App
Outputs
Input Files
*.gz A genome
WGBS-genome-...
set inputs
Converted Genome Index
*.gz The reads (pair one) that ought to be methylated
*.gz The reads (pair two) that ought to be methylated
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
via WGBS-mott-read-trimmer-se Mott Trimmed Reads f
via WGBS-mott-read-trimmer-se Mott Trimmed Reads f
*.gz A genome
WGBS-bismark-...
set inputs
tgz file of mapped bismark outputs
Output Files
*.gz A genome
via WGBS-bismark-map-pe tgz file of mapped bismark o
WGBS-extract-r...
set inputs
CG methyation BED file
CHG methyation BED file
CHH methyation BED file
the sam file
...


ENCODE Bismark DNA-ME pipeline: (paired-
4 apps unconfigured
Output folder...
Readme
Run Analysis...
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
*.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
via WGBS-mott-read-trimmer-se Mott Trimmed Reads f
via WGBS-mott-read-trimmer-se Mott Trimmed Reads
WGBS-bismark-...
set inputs
tgz file of mapped bismark outputs
*.gz A genome
Outputs plumbed to inputs
Output Files
*.gz A genome
via WGBS-bismark-map-pe tgz file of mapped bismark o
WGBS-extract-r...
set inputs
CG methyation BED file
CHG methyation BED file
CHH methyation BED file
the sam file
...

Deploy Analysis Pipelines to the Cloud

Provenance
Accessioned inputs

Ease of Use
Drop in your files



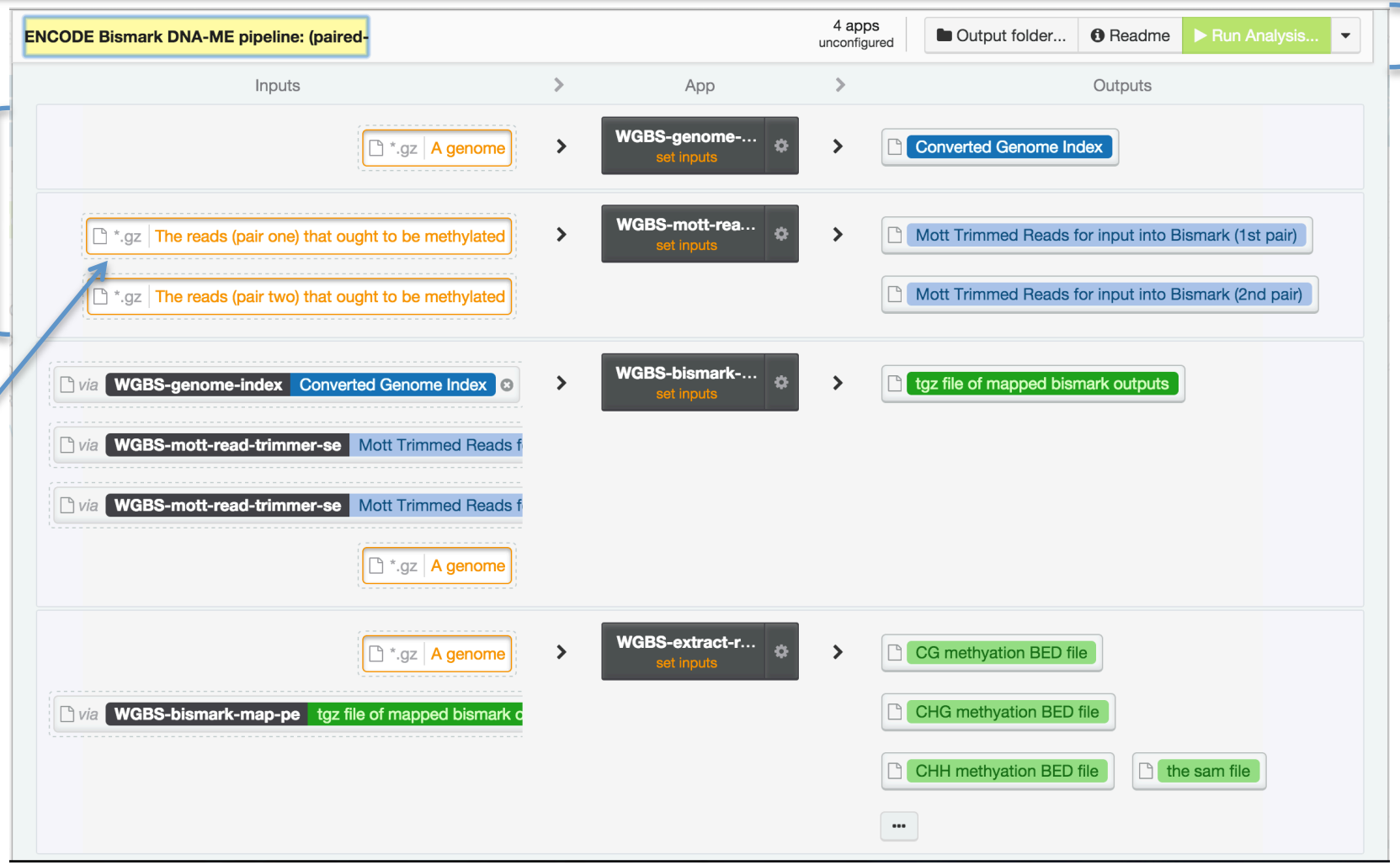
Replicable
On the web to re-run. Pipelines will be modeled in the metadata database

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

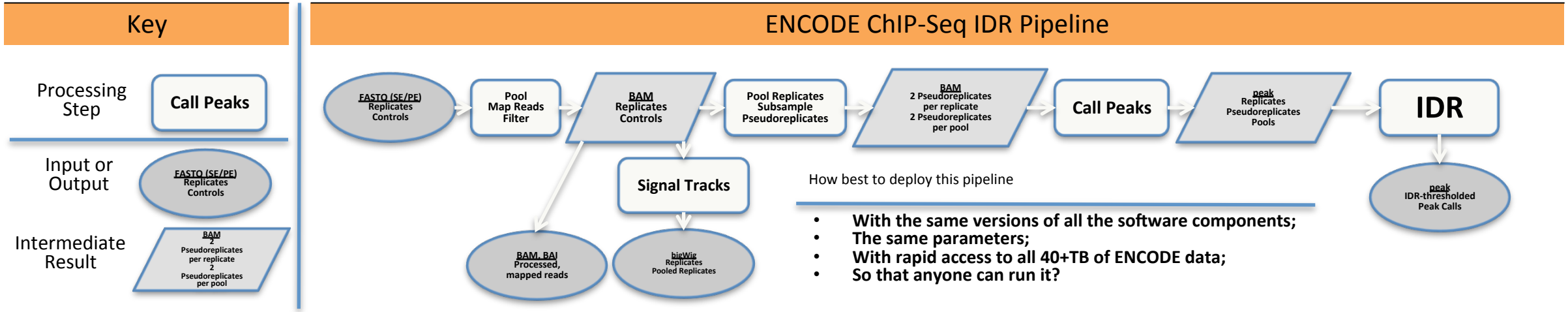


Replicable
On the web to re-run. Pipelines will be modeled in the metadata database

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

```
#!/usr/bin/env python

import requests

URL = 'https://www.encodeproject.org/experiments/ENCSR236EGS/?format=json'

response = requests.get(URL)

experiment = response.json()

print experiment['accession']
print experiment['description']
```

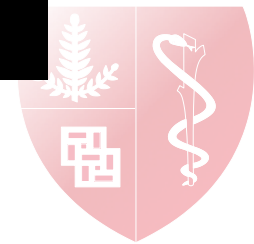
- All Portal content is accessible via URL's; just add `?format=json`
- The database record is returned in JSON format
- JSON can be parsed in your language of choice

Programmatic access via the ENCODE REST API

GET_object.py

```python
#!/usr/bin/env python

import requests

URL = 'https://www.encodeproject.org/experiments/ENCSR236EGS/?format=json'

response = requests.get(URL)

experiment = response.json()

print experiment['accession']
print experiment['description']
```

```
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

```
#!/usr/bin/env python

import requests

URL = ('https://www.encodeproject.org/search/?'
       'type=experiment&'
       'assay_term_name=ChIP-seq&'
       'replicates.library.biosample.donor.organism.scientific_name=Homo sapiens&'
       'target.investigated_as=transcription factor&'
       'replicates.library.biosample.biosample_type=in vitro differentiated cells&'
       'format=json')

response = requests.get(URL)

search_result = response.json()['@graph']

#extract and print the target for each experiment
print '\n'.join([experiment['target']['label'] for experiment in search_result])
```


Programmatic access via the ENCODE REST API

GET_search.py

```
jseth:Keystone Epigenomics 2015 jseth$ ./GET_search.py
SMC3
RAD21
MXI1
EP300
CTCF
EZH2
EZH2
CTCF
POLR2AphosphoS5
REST
TAF1
jseth:Keystone Epigenomics 2015 jseth$
```

```
1  #!/usr/bin/e
2
3  import reque
4
5  URL = ('http
6          'type=
7          'assay
8          'repli                                    omo sapiens&'
9          'targe
10         'repli                                    tiated cells&'
11         'forma
12
13 response = r
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
```

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 Podduturi, 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/