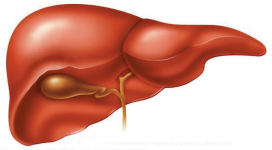
# ENCODE ChIP Pipelines

J. Seth Strattan, ENCODE DCC

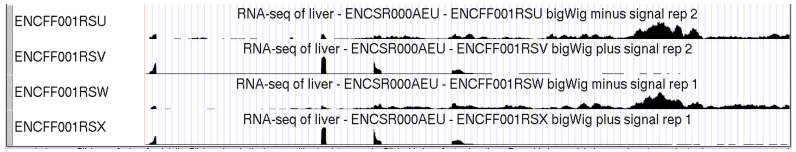
ENCODE Binding Working Group

November, 2015

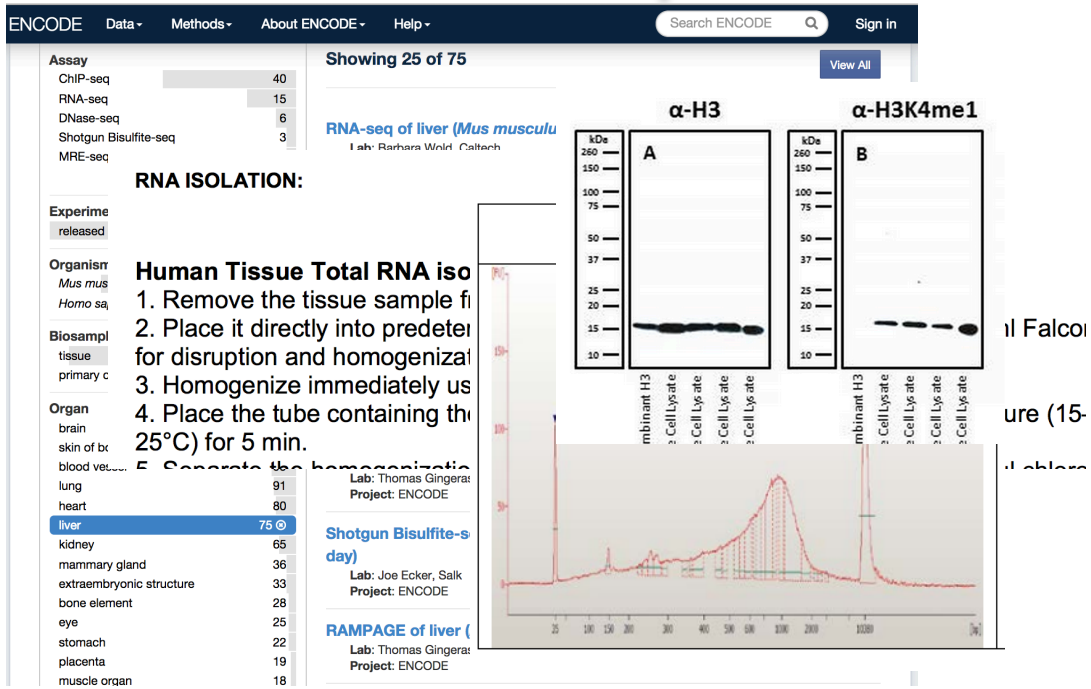
# ENCODE DCC Delivers ENCODE Metadata



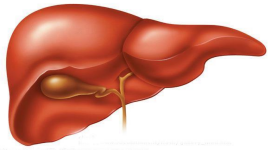
+
CCCFFFFFHHHHGIJJIGGHEIIEGGEGGIJJBHIG
@BI:SL-HAB:D0RRAACXX:8:2309:21201:7829 1:X:0:GCCGTCGA
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
+
CCCFFFFFHHHHHJJJJJJJGJJJJIIJJJJGGIGJ
@BI:SL-HAB:D0RRAACXX:8:2113:4623:40045 1:X:0:GCCGTCGA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
+
??@ADDBDH:CDHHI+AEFHI?GGHII:EFIII?F=
@BI:SL-HAB:D0RRAACXX:8:2206:11680:21762 1:X:0:GCCGTCGA
AGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTT
+



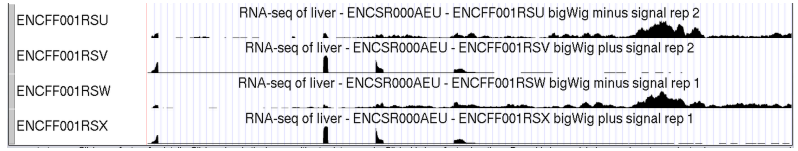

Sample → Library → Primary Data → Processed Data



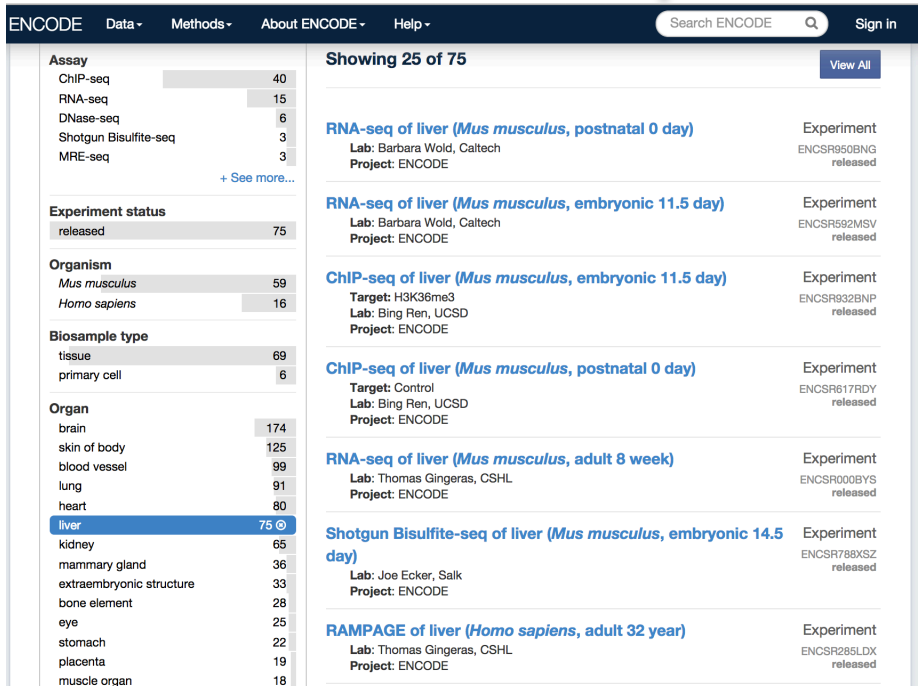
# DCC Delivers ENCODE Data



```
+
CCCFFFFFHHHHHGIJJIGGHEIIEGGEGGIJJBHIG
@BI:SL-HAB:D0RRAACXX:8:2309:21201:7829 1:X:0:GCCGTCGA
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
+
CCCFFFFFHHHHHHJJJJJJJGJJJJIIJJJJGGIGJ
@BI:SL-HAB:D0RRAACXX:8:2113:4623:40045 1:X:0:GCCGTCGA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
+
??@ADDBDH:CDHHI+AEFHI?GGHII:EFIII?F=
@BI:SL-HAB:D0RRAACXX:8:2206:11680:21762 1:X:0:GCCGTCGA
AGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTT
+
```



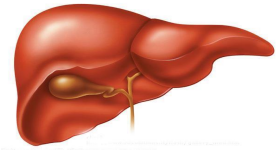

Sample → Library → Primary Data → Processed Data

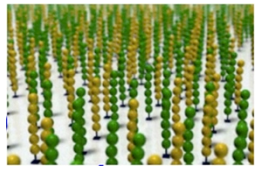



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ENCFF001RTZ | fastq | reads | 2 | (1, 1) | Esther Chan | 2013-08-15 | Download pending |
| ENCFF001RSV | bigWig | plus signal | | (2, 1) | Esther Chan | 2013-07-26 | Download pending |
| ENCFF001RSU | bigWig | minus signal | | (2, 1) | Esther Chan | 2013-07-26 | Download pending |

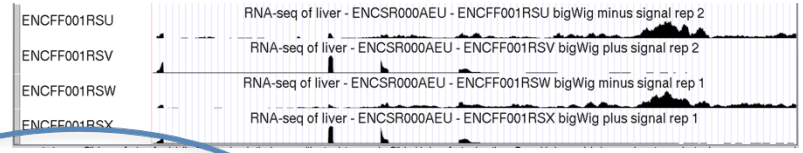
**AWS S3 Bucket**
**ENCODE Files**

# ENCODE Analysis Pipelines as Deliverables





```
+
CCCFFFFFHHHHHGIJJIGGHEIIEGGEGGIJJBHIG
@BI:SL-HAB:D0RRAACXX:8:2309:21201:7829 1:X:0:GCCGTCGA
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
+
CCCFFFFFHHHHHHJJJJJJJGJJJJIIJJJJGGIGJ
@BI:SL-HAB:D0RRAACXX:8:2113:4623:40045 1:X:0:GCCGTCGA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
+
??@ADDBDH:CDHHI+AEFHI?GGHII:EFIII?F=
@BI:SL-HAB:D0RRAACXX:8:2206:11680:21762 1:X:0:GCCGTCGA
AGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTT
+
```




Sample → Library → Primary Data → Processed Data

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

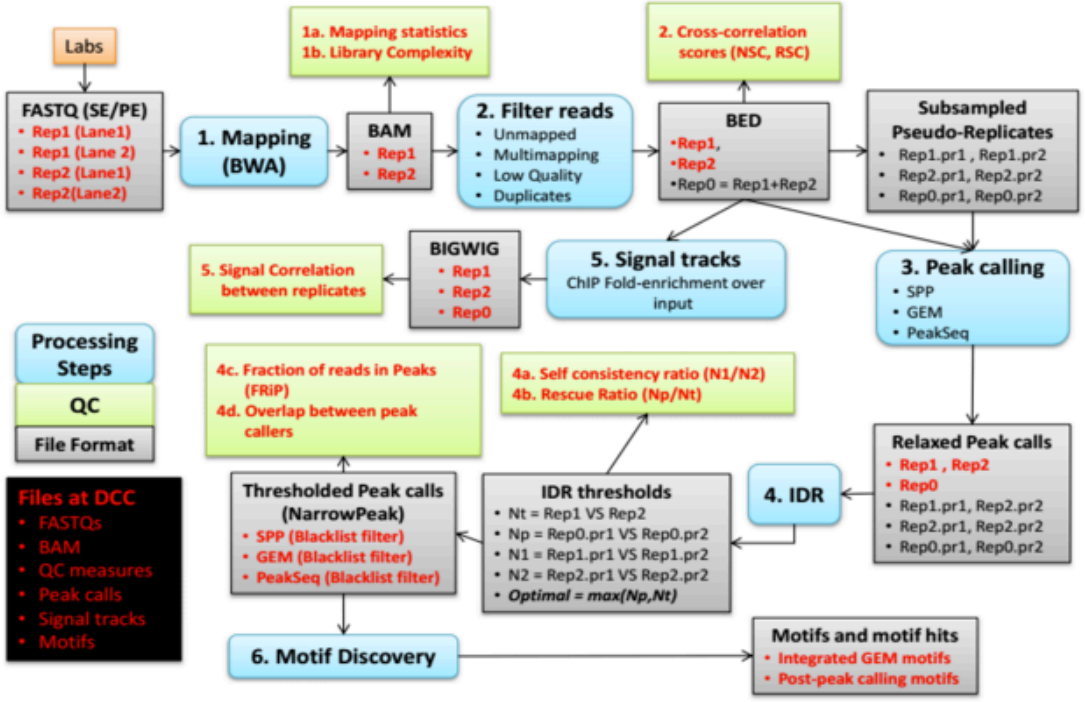
# ENCODE DAC Specifies, DCC Implements

ENCODE3 pipeline v1 specifications

Pipeline Overview

https://docs.google.com/document/d/1lG_Rd7fnYgRpSIqrIfuVlAz2dW1VaSQThzk836Db99c/edit?pli=1#heading=h.9ecc41kilcvq

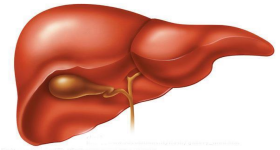
TF ChIP-seq processing pipeline

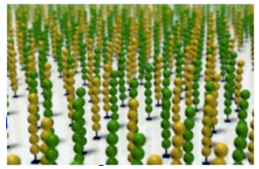




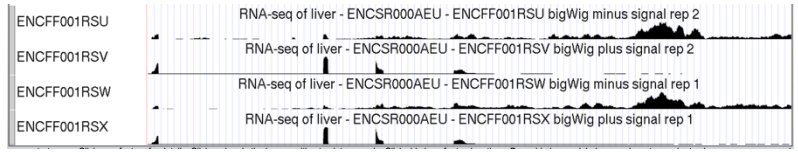

J. Seth Strattan, PhD ENCODE DCC

# Deployment Platform Considerations





```
+
CCCFFFFFHHHHHGIJJIGGHEIIEGGEGGIJJBHIG
@BI:SL-HAB:D0RRAACXX:8:2309:21201:7829 1:X:0:GCCGTCGA
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
+
CCCFFFFFHHHHHHJJJJJJJGJJJJIIJJJJGGIGJ
@BI:SL-HAB:D0RRAACXX:8:2113:4623:40045 1:X:0:GCCGTCGA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
+
??@ADDBDH:CDHHI+AEFHI?GGHII:EFIII?F=
@BI:SL-HAB:D0RRAACXX:8:2206:11680:21762 1:X:0:GCCGTCGA
AGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTT
+
```




Sample → Library → Primary Data → Processed Data

| | Develop | Share | Run | Elastic | Provenance | Cost |
|---|---|---|---|---|---|---|
| **HPC Cluster (Scripts)** | Hard | Hard | Hard | Cluster-Dependent | Moderate | Obscure/Subsidized |
| **HPC Container** | Hard | Moderate | Moderate | Cluster-Dependent | Good | Obscure/Subsidized |
| **Web/Cloud** | Moderate | Easy | Easy | Highly | Excellent | Apparent but Low |

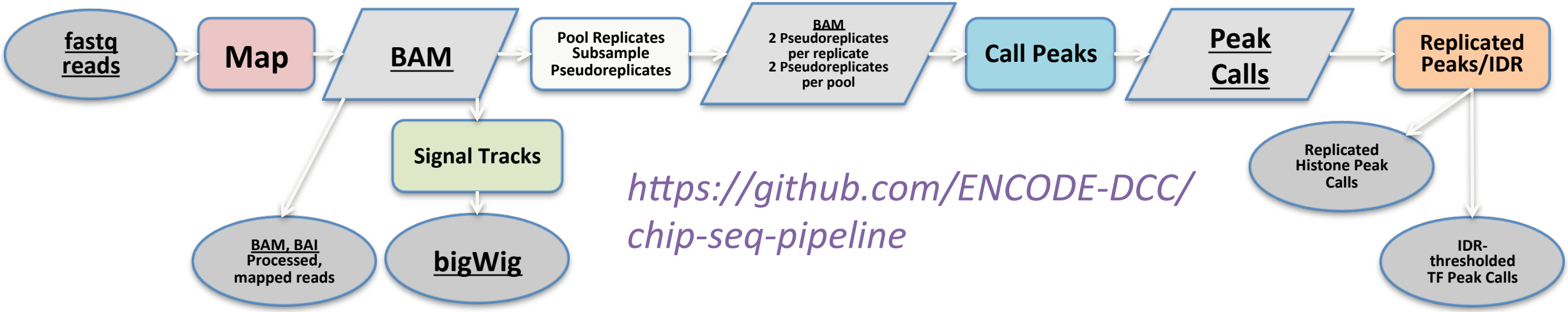
Replicability – Provenance – Ease of Use – Scalability

We chose to deploy first to a web/cloud-based platform, DNAnexus

Code is open source now and will be deployed toadditional platforms

https://github.com/ENCODE-DCC

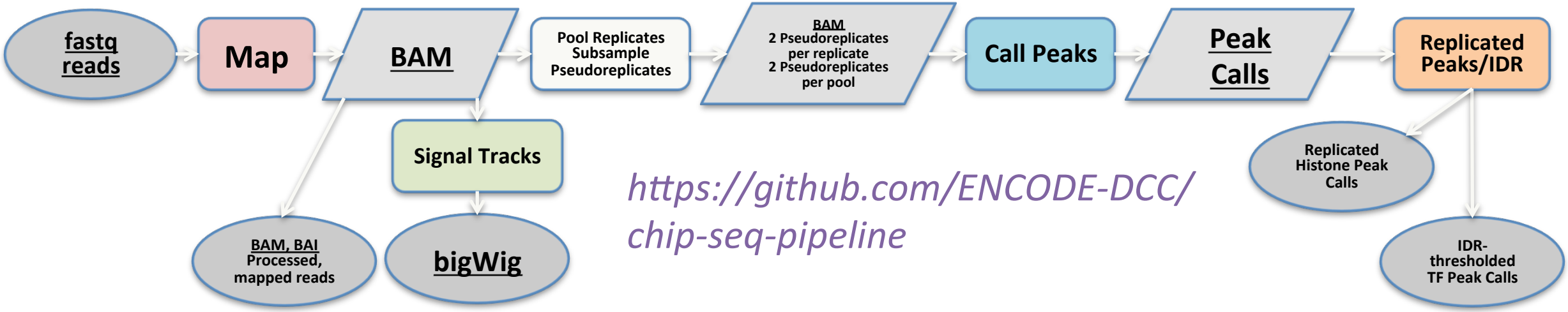
# Schema: ENCODE ChIP-seq Pipeline




*https://github.com/ENCODE-DCC/chip-seq-pipeline*

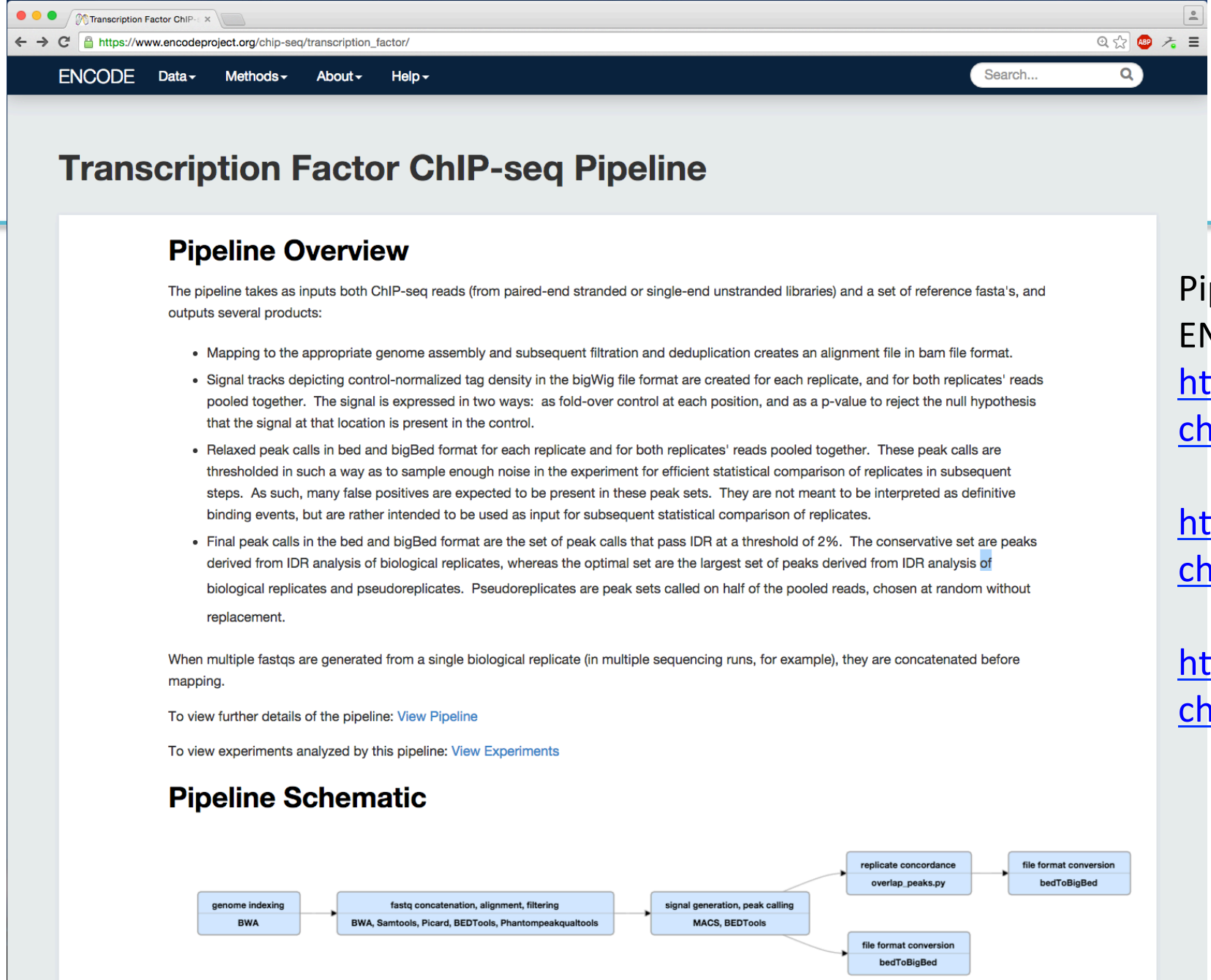
| Target | Key Software | Input Files | Output Files | QA Metrics |
|---|---|---|---|---|
| TF's | bwa<br>Picard markDuplicates<br>samtools<br>MACS2 (Signal tracks)<br>SPP (PeakSeq, GEM future)<br>IDR2 | fastq's (SE or PE)<br>Two biological replicates<br>Matched, replicated controls | One bam per replicate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

# ENCODE ChIP-seq Quality Metrics: Resources




*https://github.com/ENCODE-DCC/chip-seq-pipeline*

| Estimates | Description | References |
|---|---|---|
| **Depth** | Number of mappable reads<br>Useable fragments (=SE filtered reads, PE/2) | Jung YL, et al. Nucleic Acids Research. 2014;42(9):e74 |
| **Library Complexity**<br>**ChIP Quality** | Non-Redundant Fraction<br>PCR Bottleneck Coefficient<br>Normalized Strand Cross-Correlation<br>Relative Strand Cross-Correlation | Landt S, et al. Genome Res. 2012. 22: 1813-1831 |
| **Replicate Concordance** | IDR Rescue Ratio<br>IDR Self-Consistency Ratio<br>IDR Reproducibility Test | Li Q, et al. Annals Applied Statistics. 2011, Vol. 5, No. 3, 1752–1779 |




Pipeline overview pages on the ENCODE Portal:
https://www.encodeproject.org/chip-seq

https://www.encodeproject.org/chip-seq/transcription_factor/

https://www.encodeproject.org/chip-seq/histone/

https://www.encodeproject.org/pipelines/ENCPL272XAE/

ENCODE Data Methods About Help Search...

# Histone ChIP-seq

Status: **active**

**Title:** Histone ChIP-seq

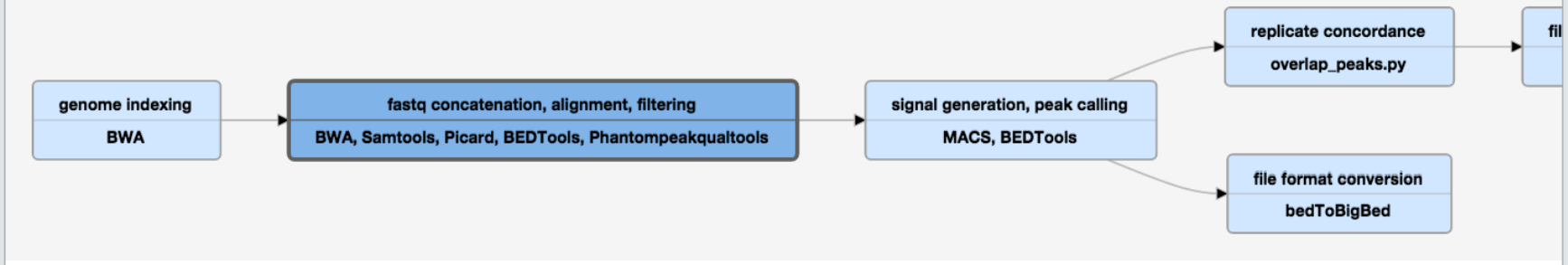
**Assay:** ChIP-seq

**Description:** ENCODE defined analysis pipeline for histone ChIP-seq

**Lab:** ENCODE Processing Pipeline

**Award PI:** J. Michael Cherry, Stanford

## Pipeline schematic




Download Graph

**Name:** BWA alignment and filtration step - Version 1

**Step type:** fastq concatenation, alignment, filtering

**Step aliases:** encode:bwa-alignment-step-v-1

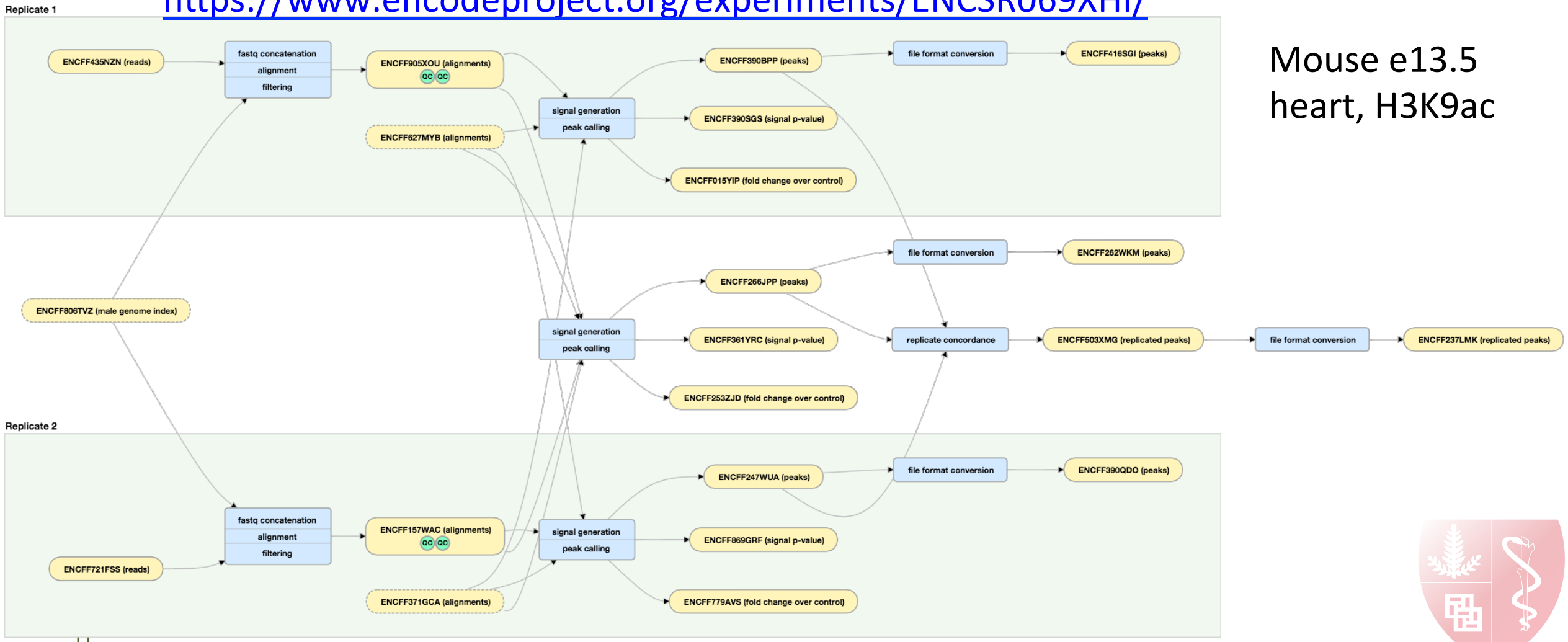
**Input:** genome index, reads

**Output:** alignments

**Version 1 — software:** BWA 0.7.10 | Samtools 1.0 | Picard 1.92 | BEDTools 2.17.0 | Phantompeakqualtools 1.1

# Uniformly Processed Data On the ENCODE Portal

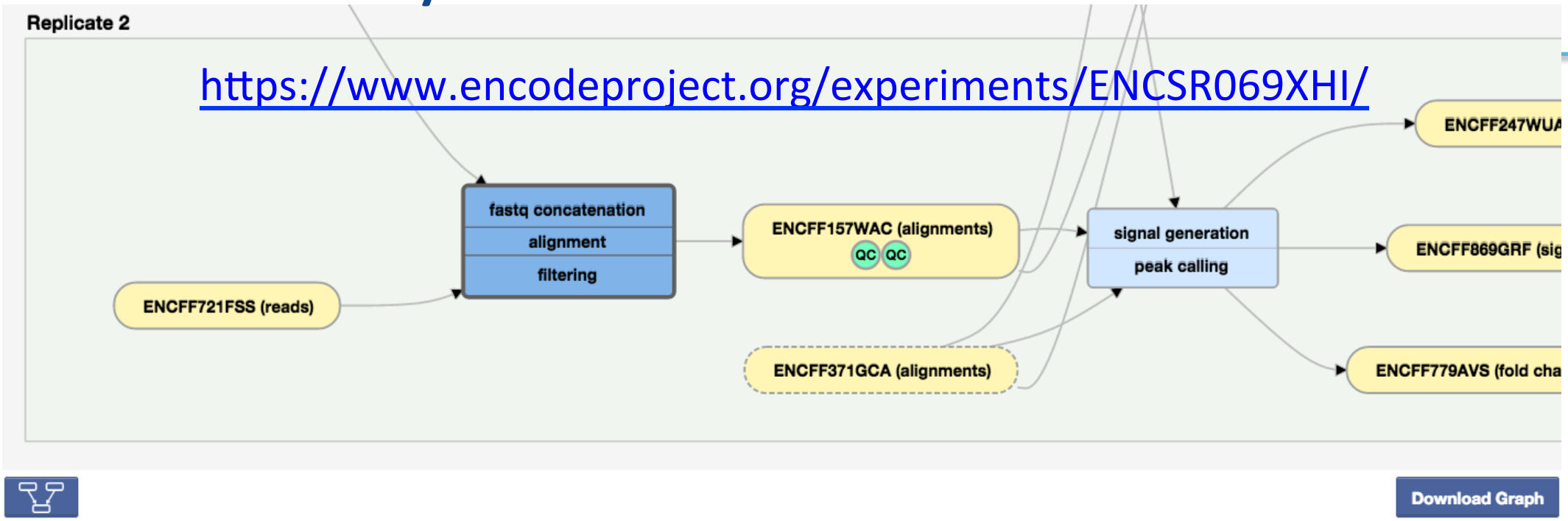
https://www.encodeproject.org/experiments/ENCSR069XHI/




Mouse e13.5
heart, H3K9ac

# Uniformly Processed Data On the ENCODE Portal

https://www.encodeproject.org/experiments/ENCSR069XHI/




**Name:** BWA alignment and filtration step - Version 1— Version 1

**Step type:** fastq concatenation, alignment, filtering

**Step aliases:** encode:bwa-alignment-step-v-1

**Input:** genome index, reads

**Output:** alignments

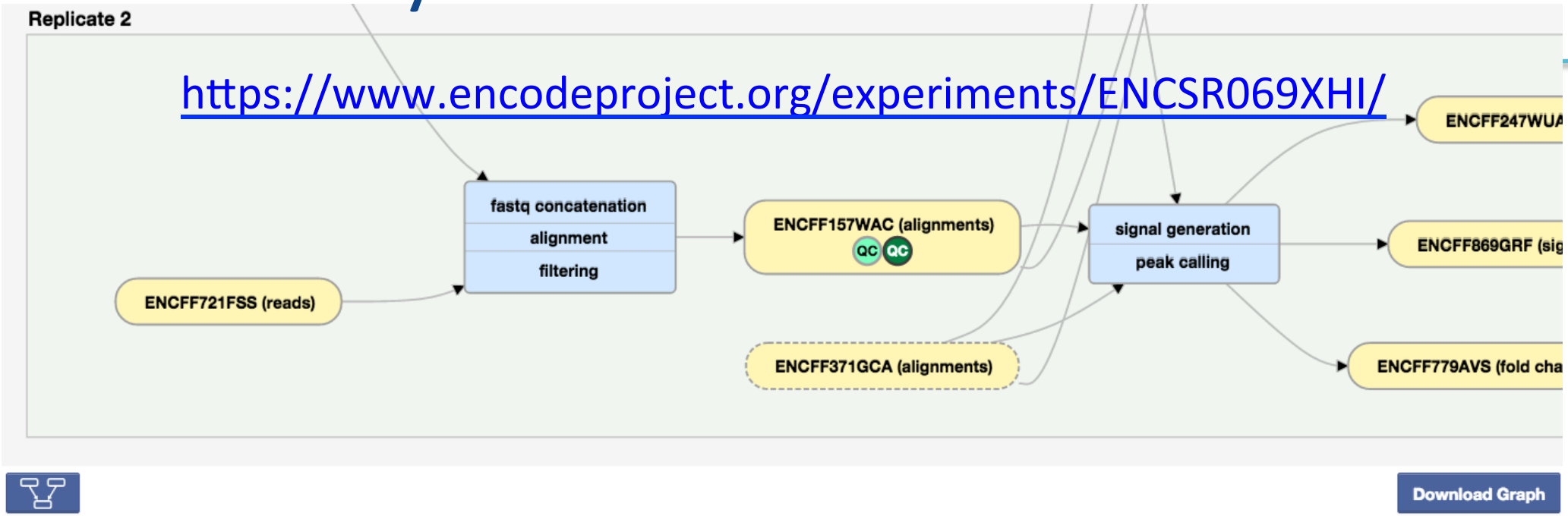
**Pipeline:** Histone ChIP-seq, Transcription factor ChIP-seq

**Software:** BWA 0.7.10 | Samtools 1.0 | Picard 1.92 | BEDTools 2.17.0 | Phantompeakqualtools 1.1

Analysis step metadata

# Uniformly Processed Data On the ENCODE Portal

https://www.encodeproject.org/experiments/ENCSR069XHI/




**Quality metrics of ENCFF157WAC**

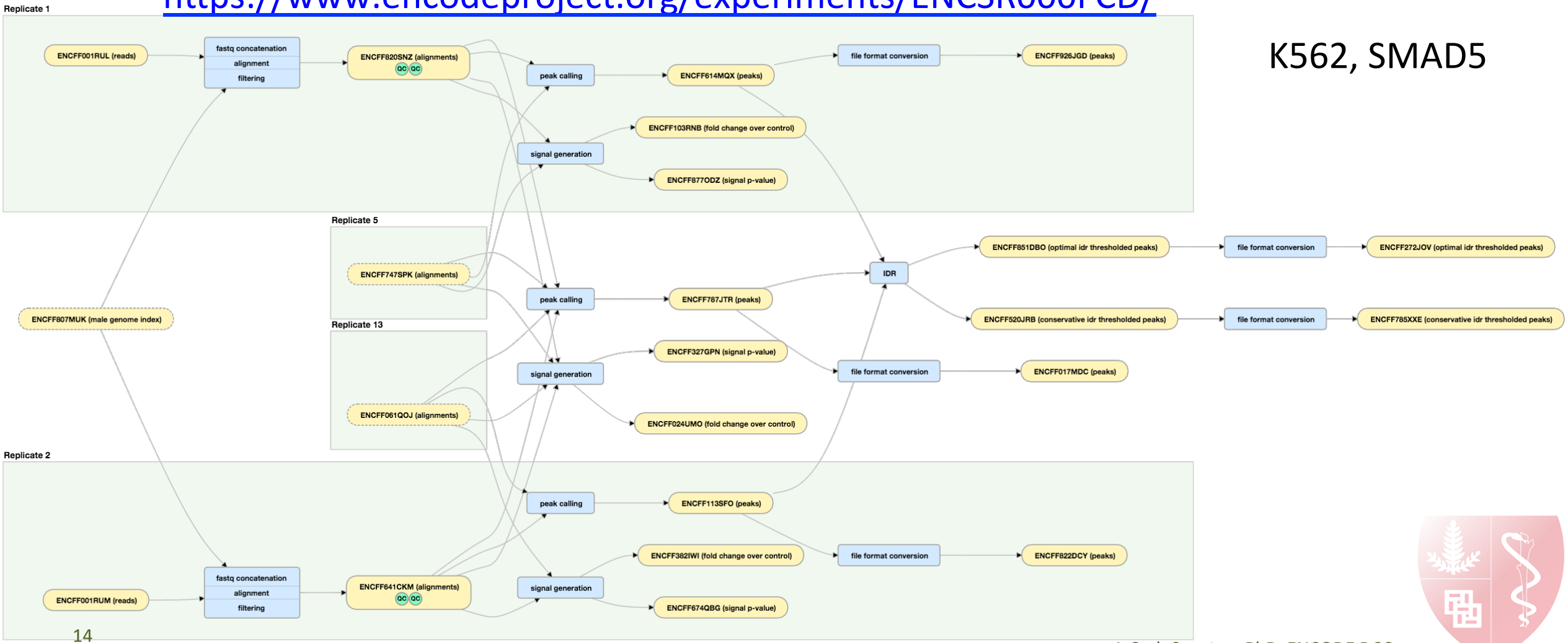
| | |
|---|---|
| duplicates: | 0 |
| duplicates_qc_failed: | 0 |
| mapped: | 53623474 |
| mapped_pct: | 100.00% |
| mapped_qc_failed: | 0 |
| total: | 53623474 |
| total_qc_failed: | 0 |

Analysis step metadata
QC metrics associated with files

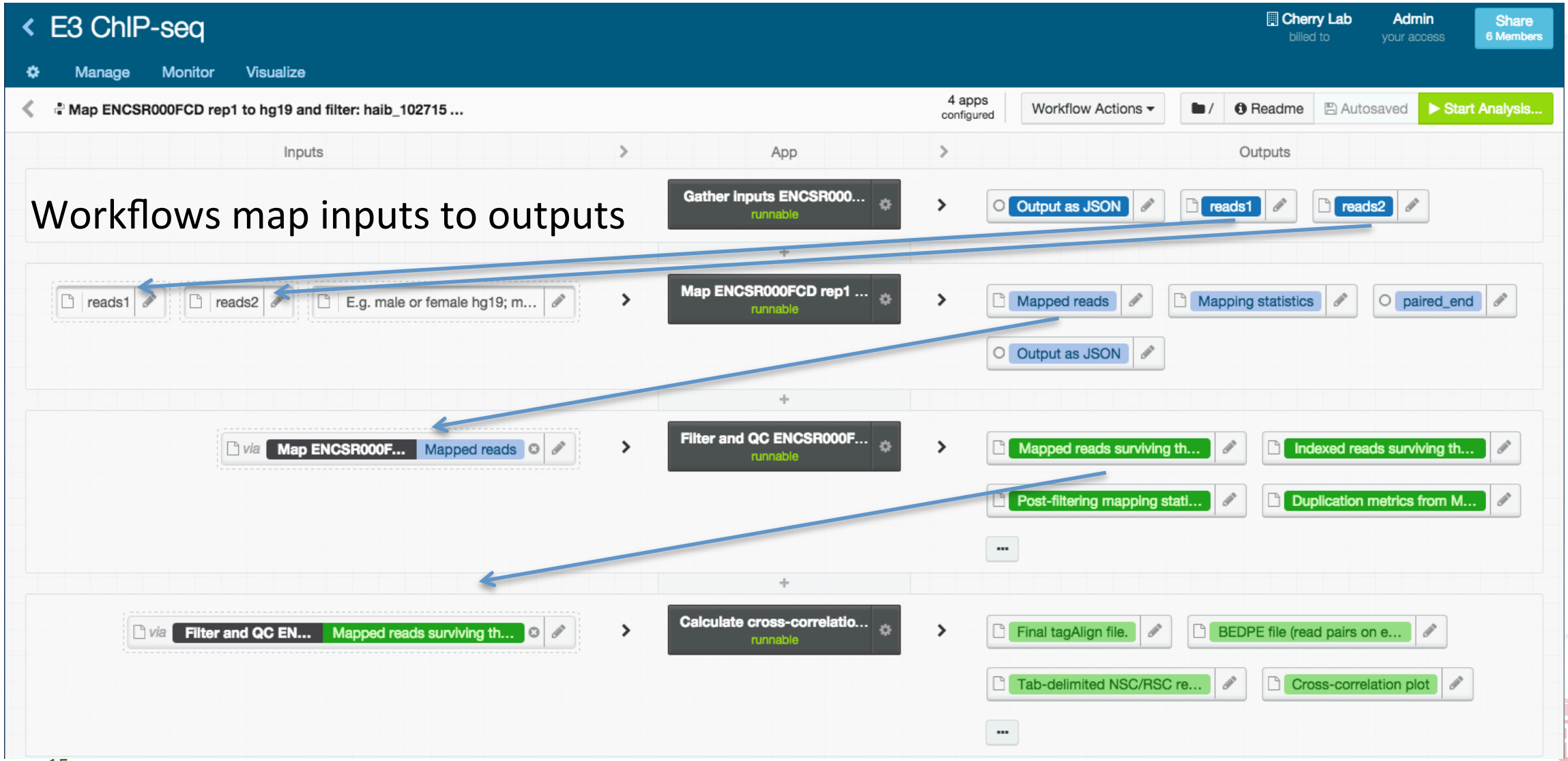
# Uniformly Processed Data On the ENCODE Portal

https://www.encodeproject.org/experiments/ENCSR000FCD/

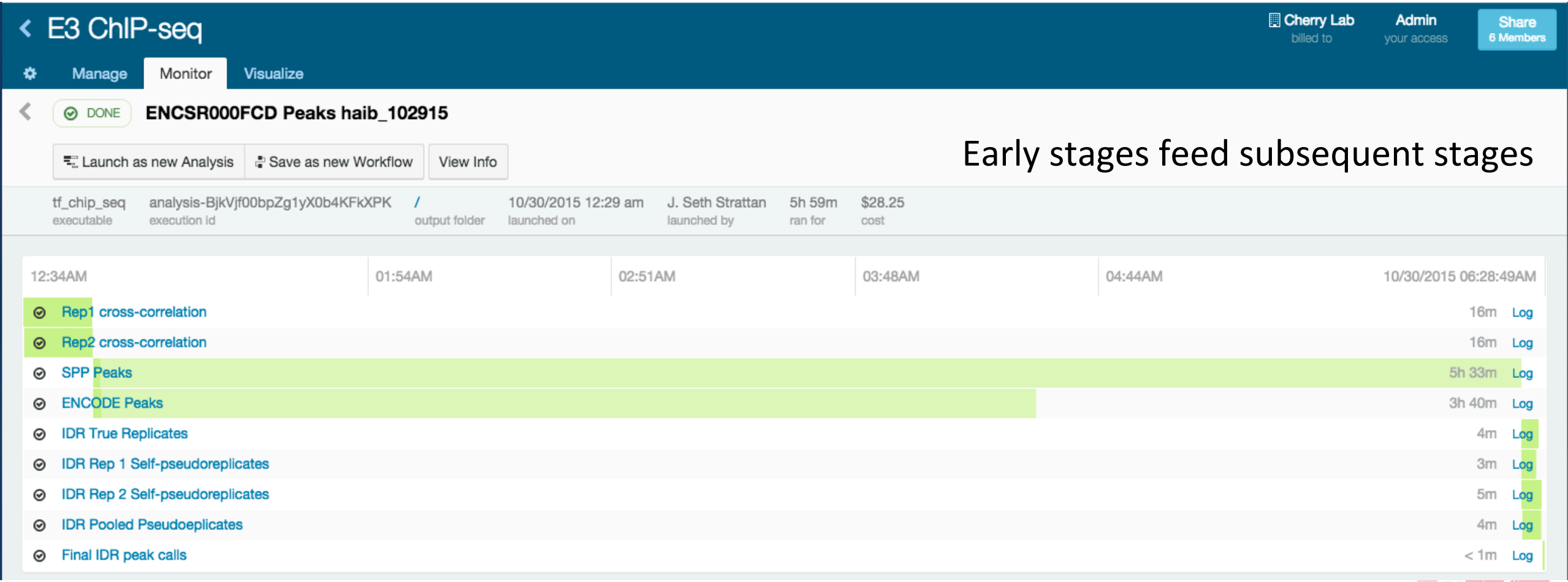
K562, SMAD5



# ENCODE ChIP-seq on the Cloud



# ENCODE ChIP-seq on the Cloud




Early stages feed subsequent stages

# ENCODE ChIP-seq on the Cloud

06:28:01AM | 06:28:10AM | 06:28:16AM | 06:28:23AM | 06:28:29AM | 10/30/2015 06:28:41AM

**Final IDR peak calls** < 1m Log

**INPUTS**

IDR peaks from replicate 2 self-pseudoreplicates (r2pr_peaks)
`ENCFF001RUMvENCFF001RUM.IDRv2.IDR0.02.narrowPeak.gz`

IDR peaks from replicate 1 self-pseudoreplicates (r1pr_peaks)
`ENCFF001RULvENCFF001RUL.IDRv2.IDR0.02.narrowPeak.gz`

IDR peaks from pooled self-pseudoreplicates (pooledpr_peaks)
`ENCFF001RULvENCFF001RUL.IDRv2.IDR0.02.narrowPeak.gz`

chrom.sizes for bedToBigBed (chrom_sizes)
`ENCODE - Production runs ▸ male.hg19.chrom.sizes`

Blacklist (gzipped BED file) (blacklist)
`ENCODE Reference Files ▸`
`wgEncodeDacMapabilityConsensusExcludable.bed.gz`

IDR peaks from true replicates (reps_peaks)
`ENCFF001RULvENCFF001RUM.IDRv2.IDR0.02.narrowPeak.gz`

.as file for bedToBigBed (as_file)
`ENCODE Reference Files ▸ narrowPeak.as`

**OUTPUTS**

Final peak calls - conservative set (conservative_set)
`IDR_final_conservative.narrowPeak.gz`

Final peak calls - optimal set (optimal_set)
`IDR_final_optimal.narrowPeak.gz`

Final peak calls - conservative set bigBed (conservative_set_bb)
`IDR_final_conservative.narrowPeak.bb`

Final peak calls - optimal set bigBed (optimal_set_bb)
`IDR_final_optimal.narrowPeak.bb`

Result of the reproducibility test (reproducibility_test)
`pass`

Rescue ratio (rescue_ratio)
`1.4036567105366853`

Number of peaks from pooled pseudoreplicates (Np)
`17811`

Number of peaks from replicate 1 self-pseudoreplicates (N1)
`9167`

Number of peaks from replicate 2 self-pseudoreplicates (N2)
`17815`

Number of peaks from true replicates (Nt)
`12689`

Self-consistency ratio (self_consistency_ratio)
`1.9433838769499292`

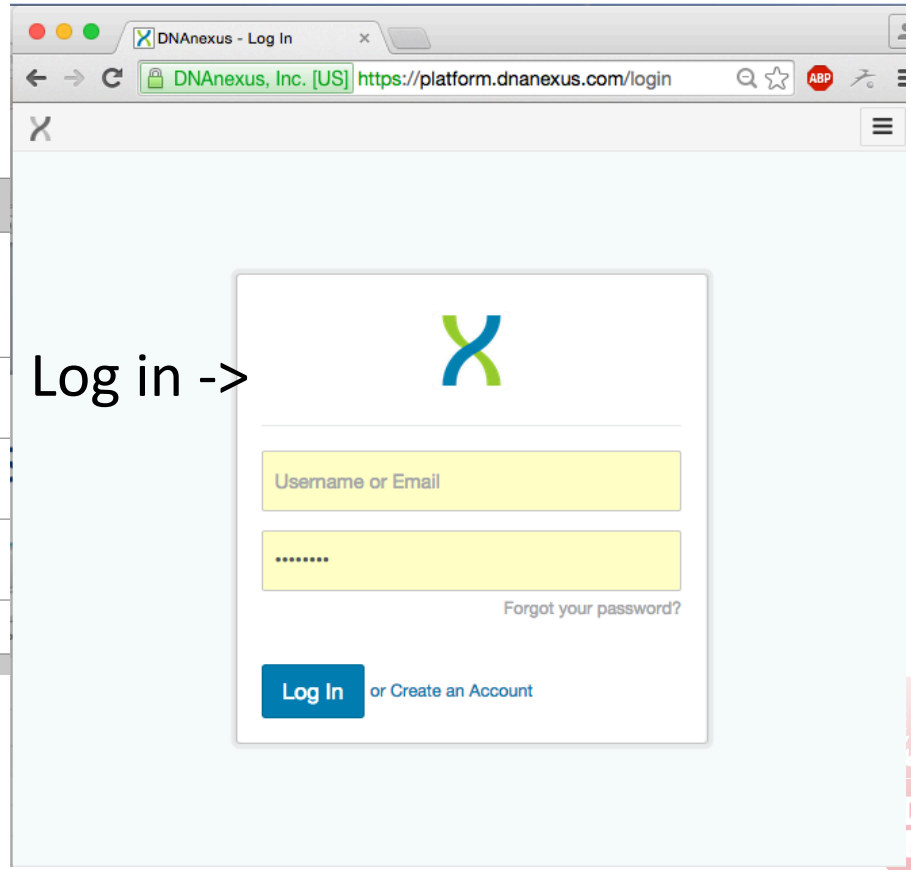
Outputs that are accessioned at the ENCODE Portal

# Running Pipelines on DNAnexus

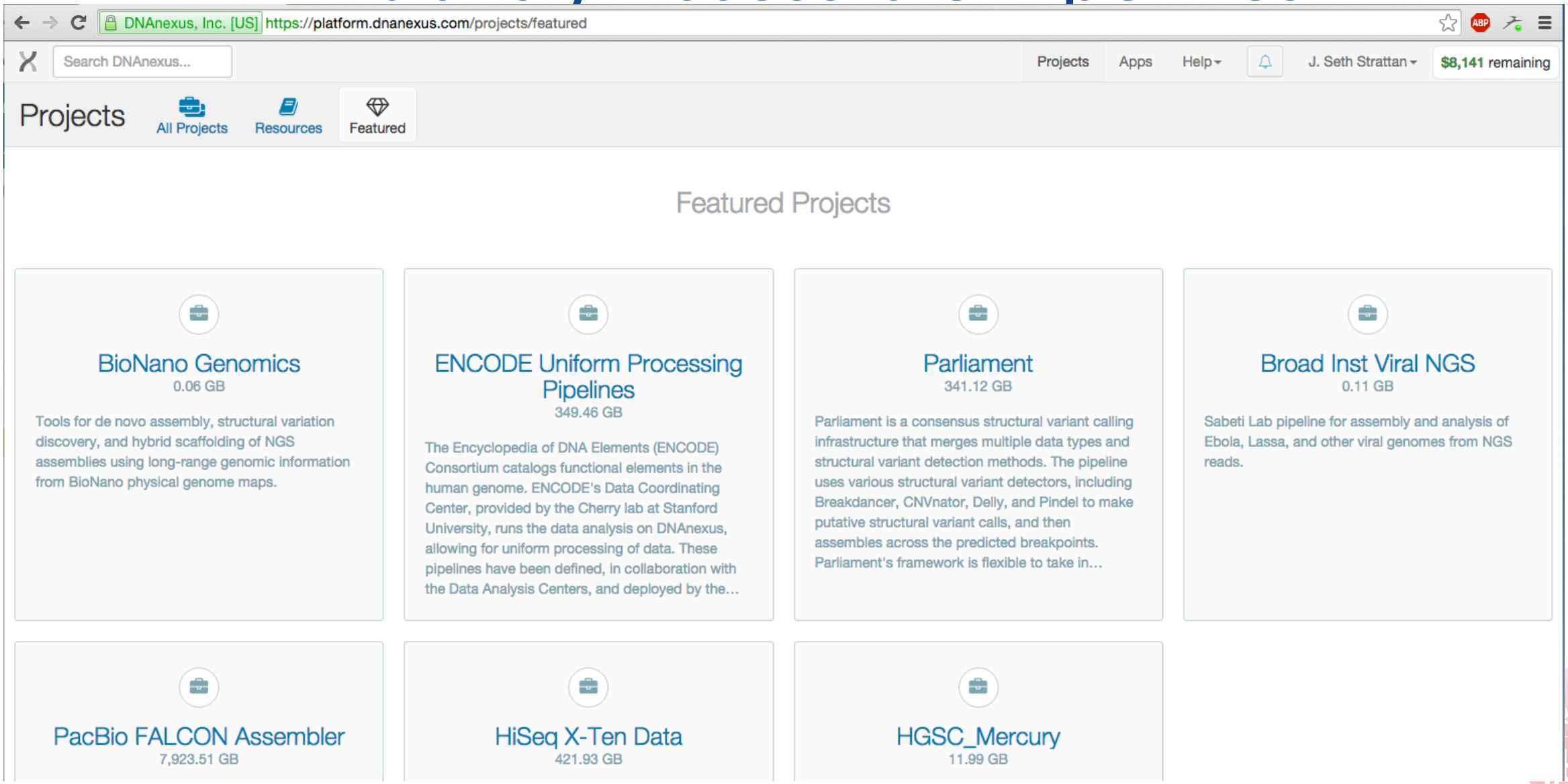
To set up an account:
https://www.encodeproject.org/tutorials/encode-users-meeting-2015/

Click “Prepare to run web-based pipelines”

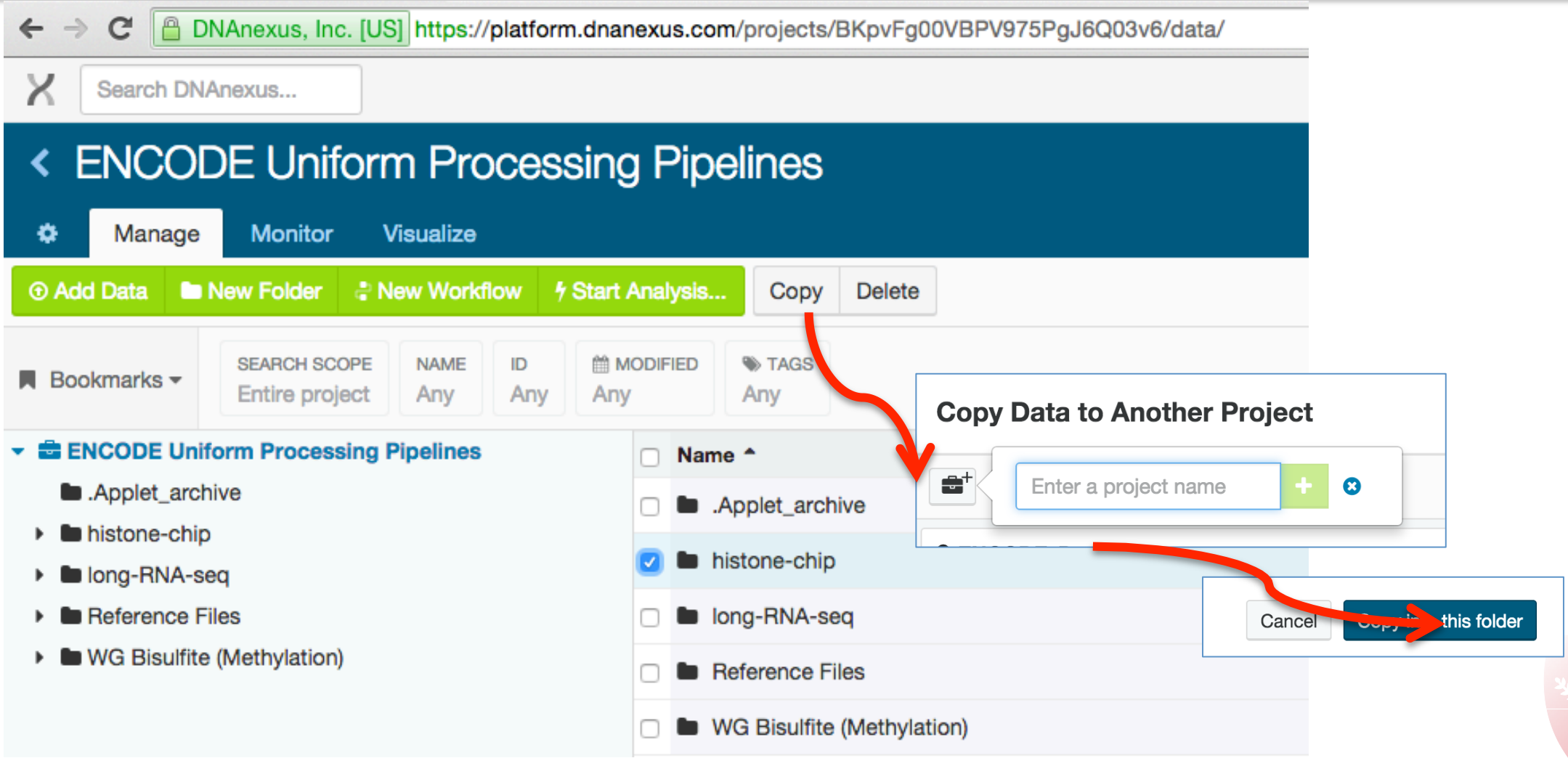
| Workshop Session 3: ENCODE Uniform Processing Pipelines | | | | |
|---|---|---|---|---|
| 18 | ENCODE Uniform Data Processing Pipelines: Introduction | J. Seth Strattan | Stanford University | Slides \| Prepare to run web-based pipelines |
| 19 | ENCODE Uniform Data Processing Pipelines: Demo and Hands-on Tutorial | Ben Hitz | Stanford University | Hands-on Tutorial |
| 20 | ENCODE Uniform Data Processing Pipelines: ChIP-seq IDR Architecture | J. Seth Strattan | Stanford University | |
| 21 | ENCODE Uniform Data Processing Pipelines: Demo and Hands-on Tutorial (continued) | Ben Hitz | Stanford University | |
| 22 | ENCODE Uniform Data Processing Pipelines: Wrap-up | J. Seth Strattan | Stanford University | |



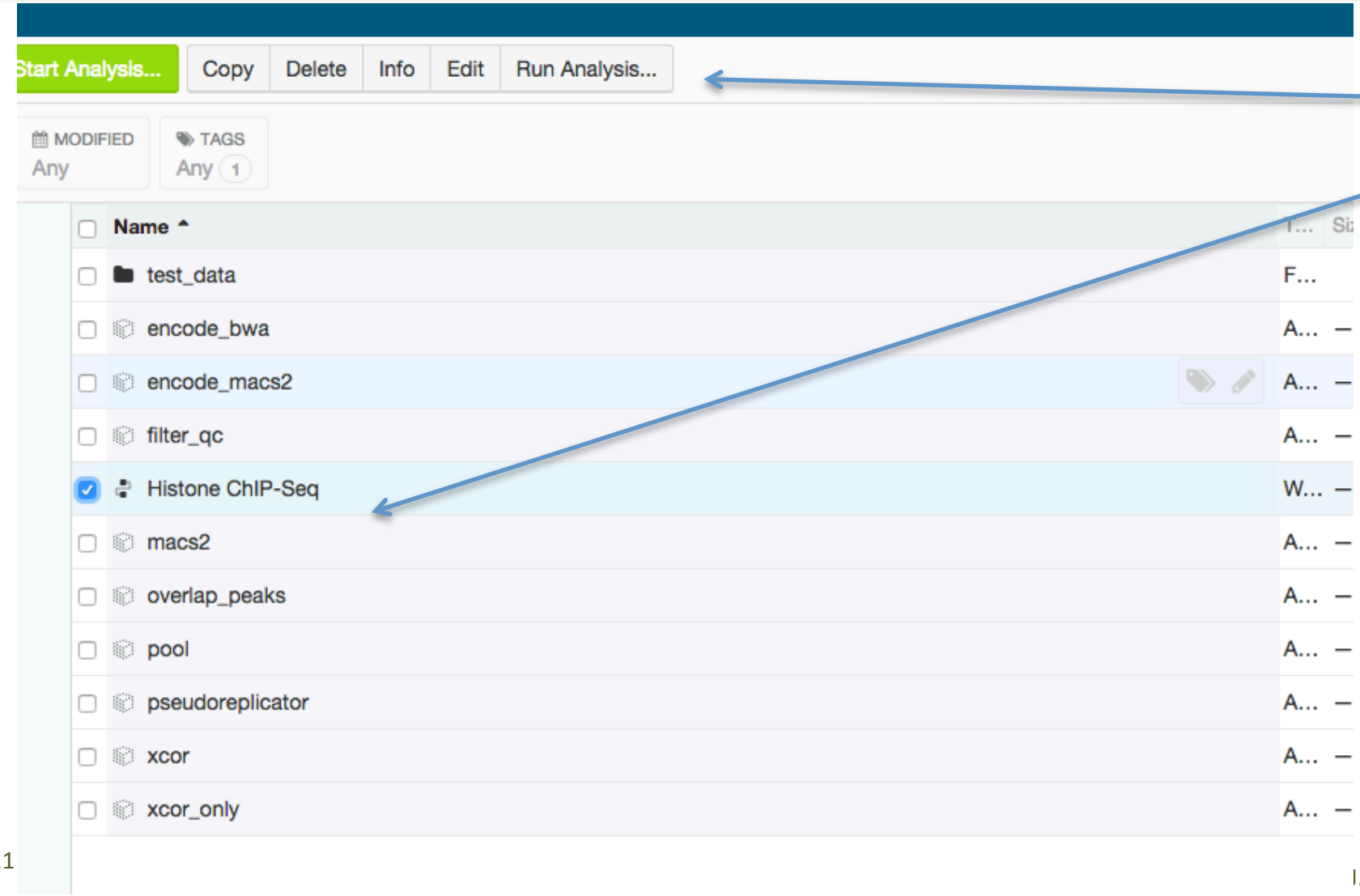
# Publicly Accessible Pipelines



# Copy Pipeline to Your Project

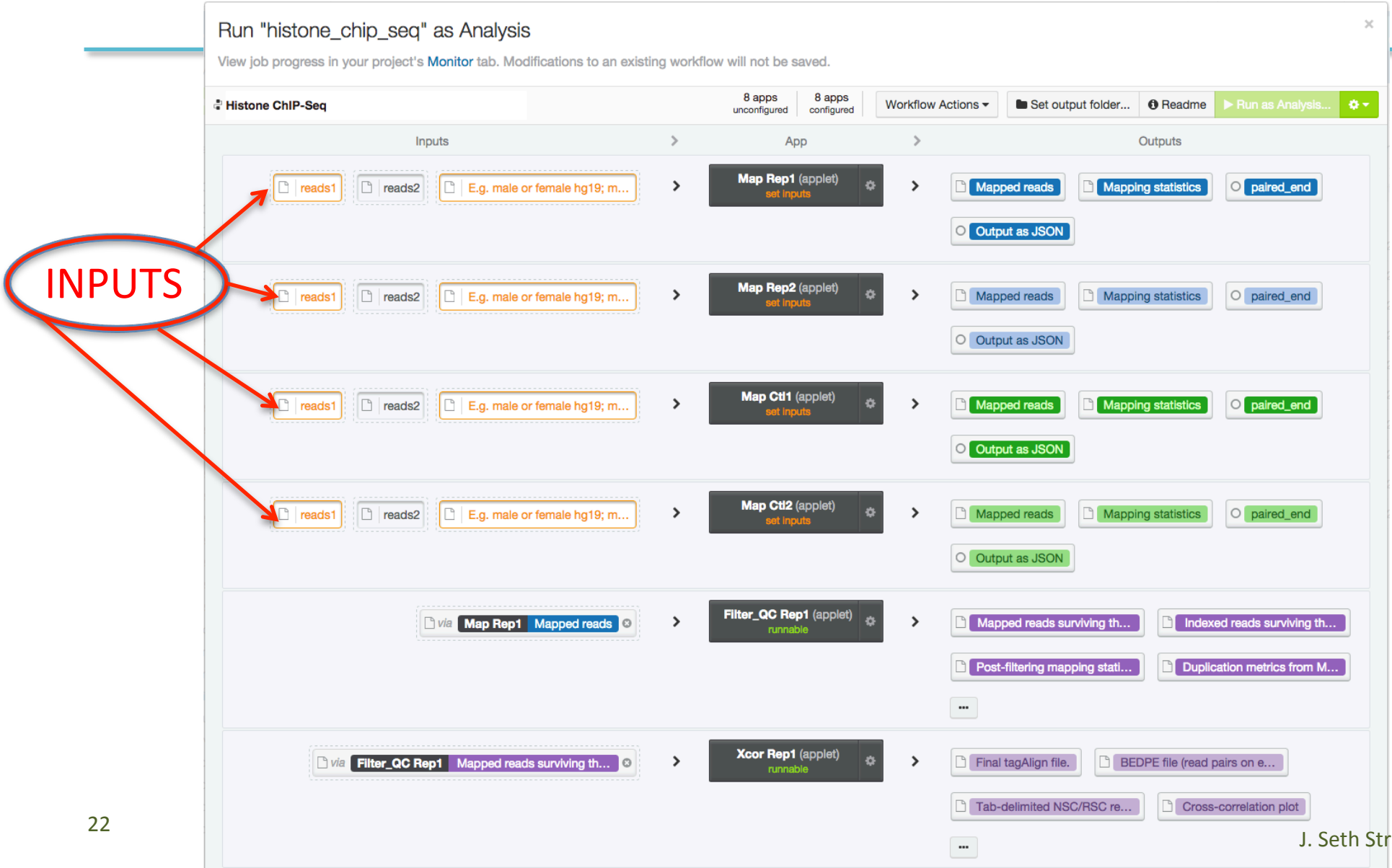


# Select a Workflow to Run



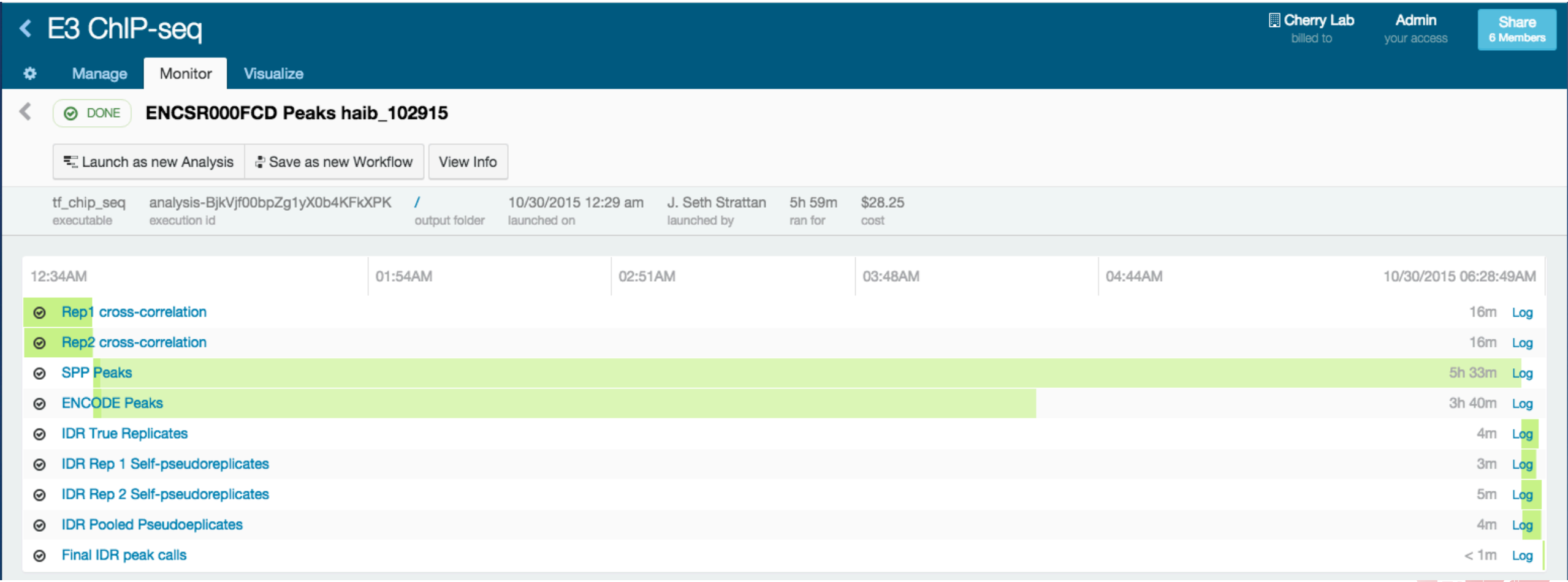

Select the workflow "Run Analysis …"

# A Workflow

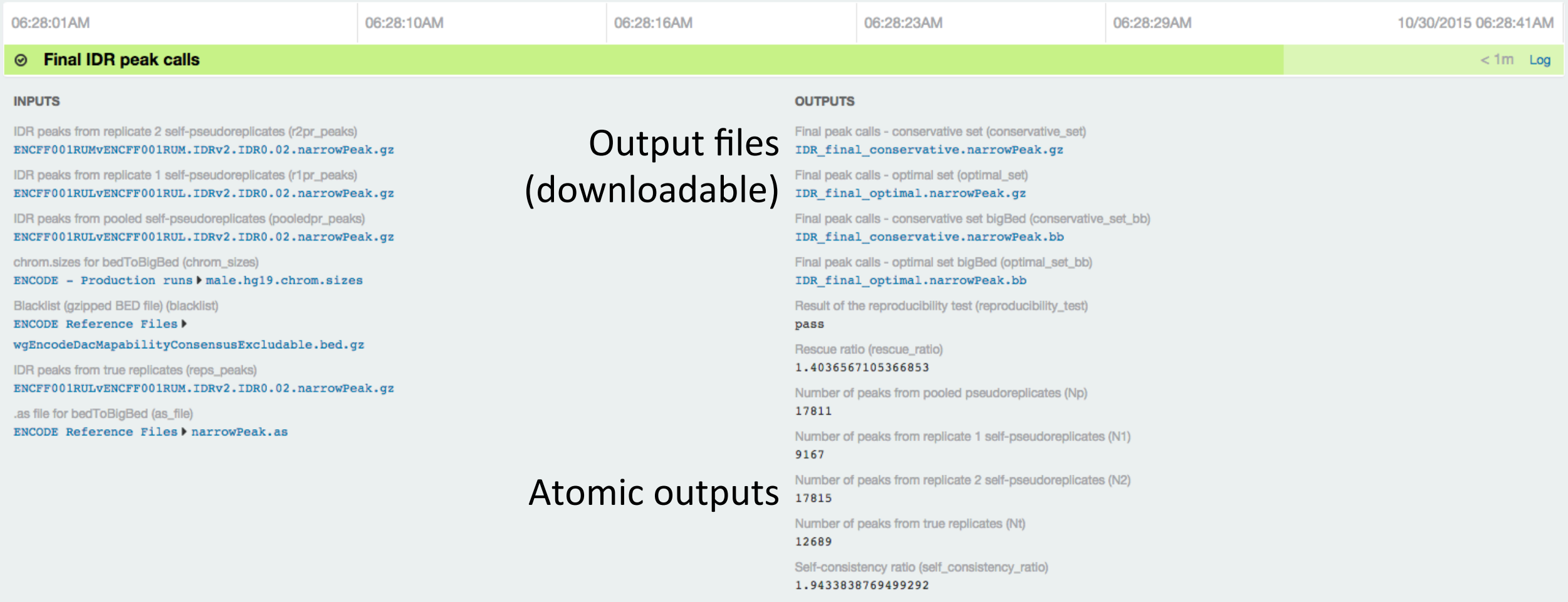



Once input requirements are met … "Run as Analysis"

# The “Monitor” Tab



# Gather Results




Output files (downloadable)

Atomic outputs

# Summary and Development Plans

- DCC have implemented ChIP analysis pipelines as specified by the DAC.
- Signal tracks and peak calls are produced, along with QC metrics.
- First deployment is to the DNAnexus cloud platform.
- All ChIP experiments from the October 2015 freeze have been run through the mapping, and TF or histone pipelines.
- Anyone can run these pipelines on DNAnexus.
- Currently under development (early 2016 release):
    - Add GEM and PeakSeq peak callers to TF pipeline.
    - Add motif analysis.
    - Refine IDR metrics for histone pipeline.
    - Improved documentation.
    - GRCh38 processing of all experiments.

# 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

**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 Irizarryt, 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

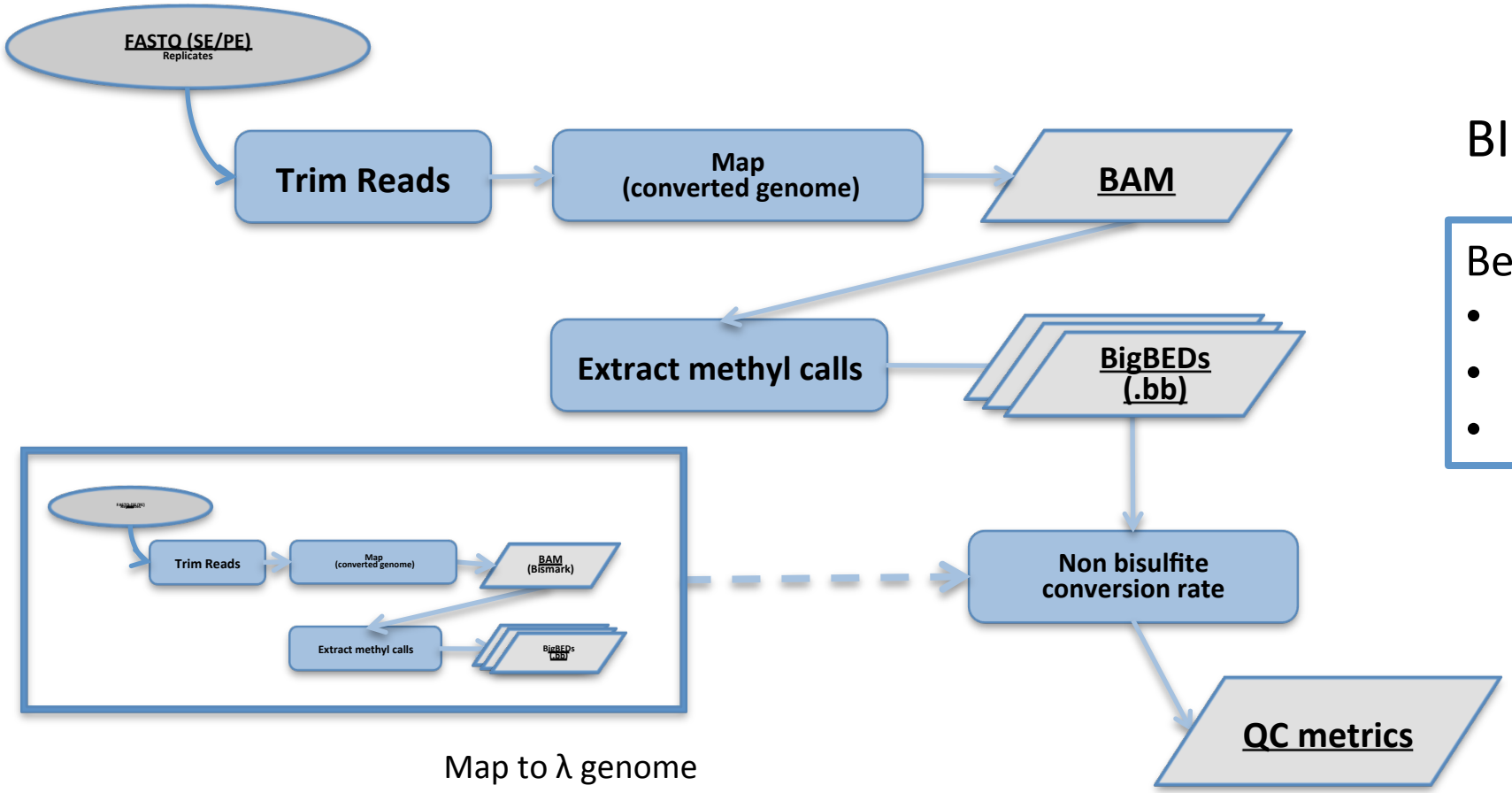

@encodedcc


encode-help@lists.stanford.edu

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

# Schema: ENCODE WGBS Pipeline

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




Map to λ genome

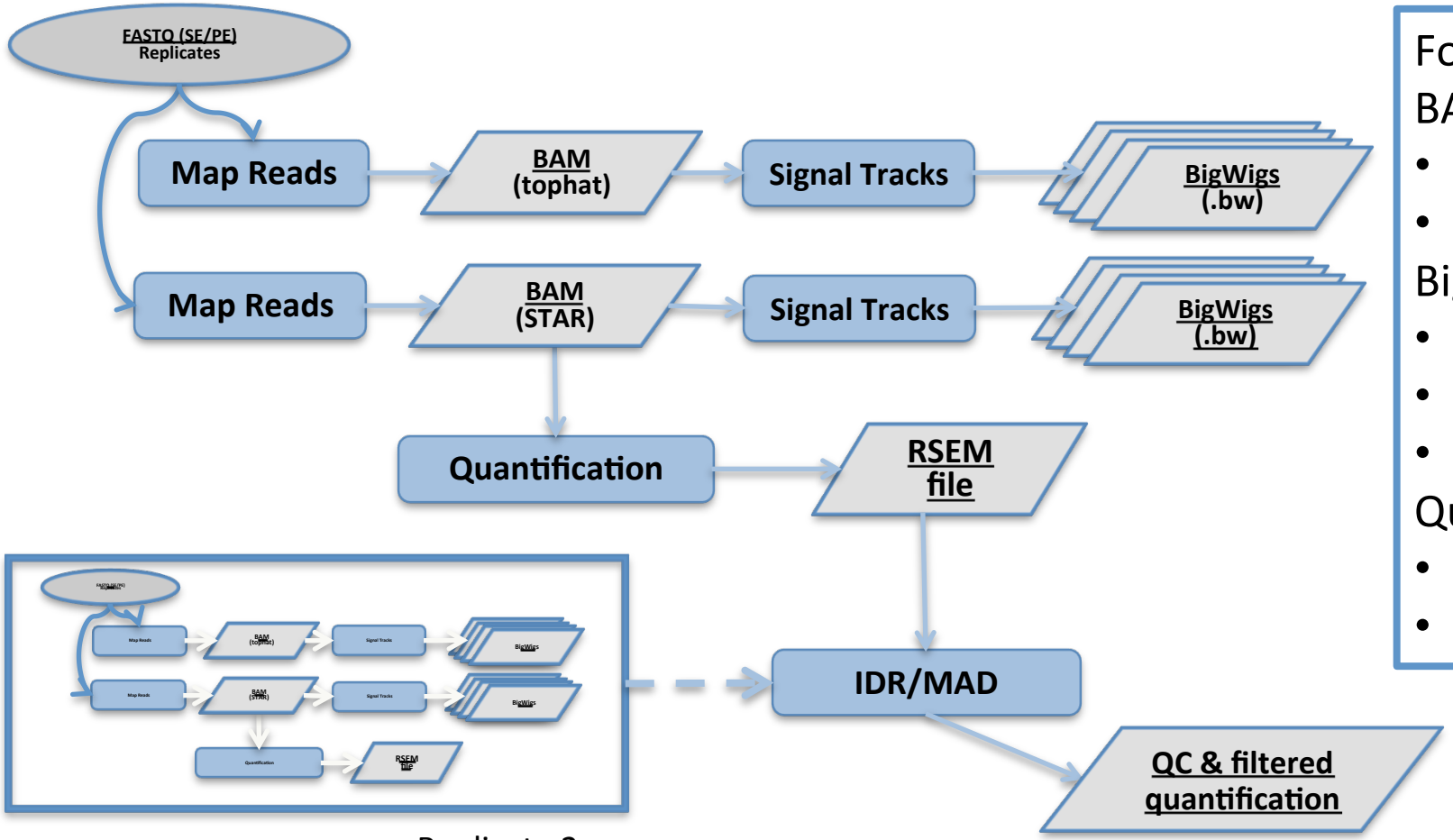
BISMARK (v 0.10)

Bed/BigBed files for:

- CG context
- CHG context
- CHH context

# Schema: ENCODE RNA-seq Pipeline

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



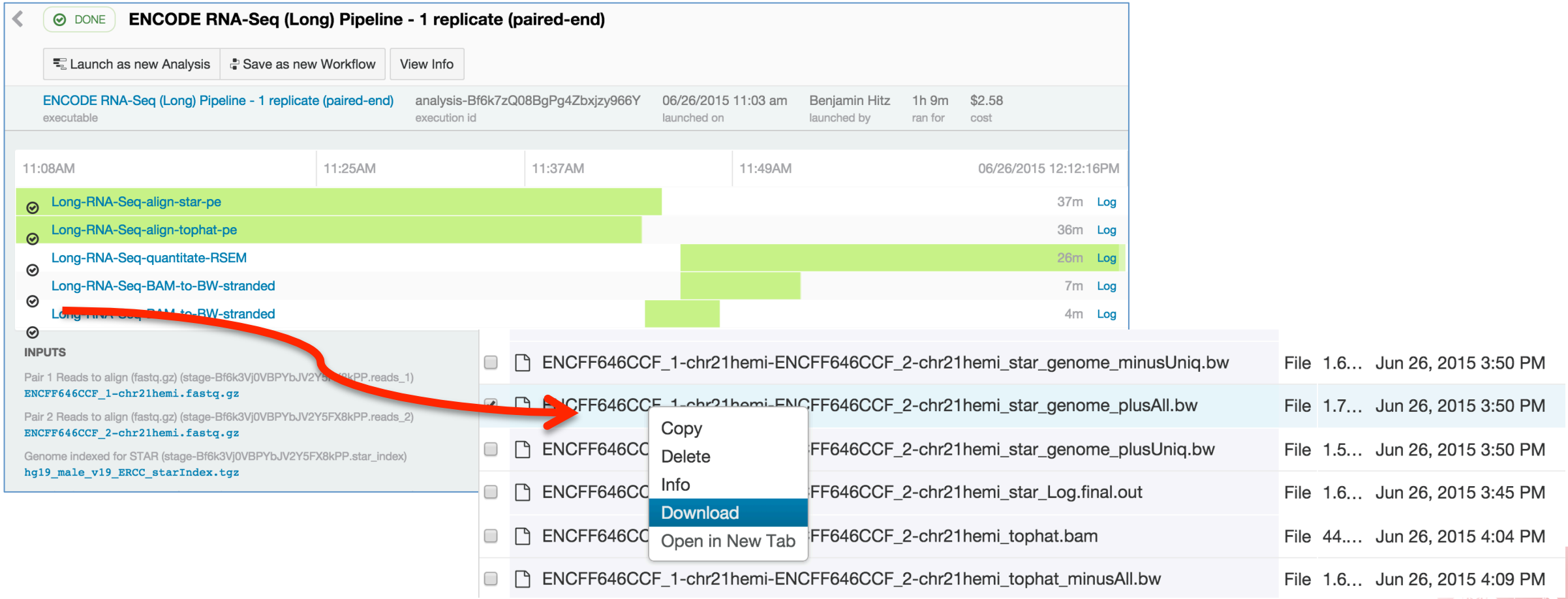

Replicate 2

For *each* Mapper (STAR, tophat)
BAM files:
- mapped to genome
- mapped to transcriptome

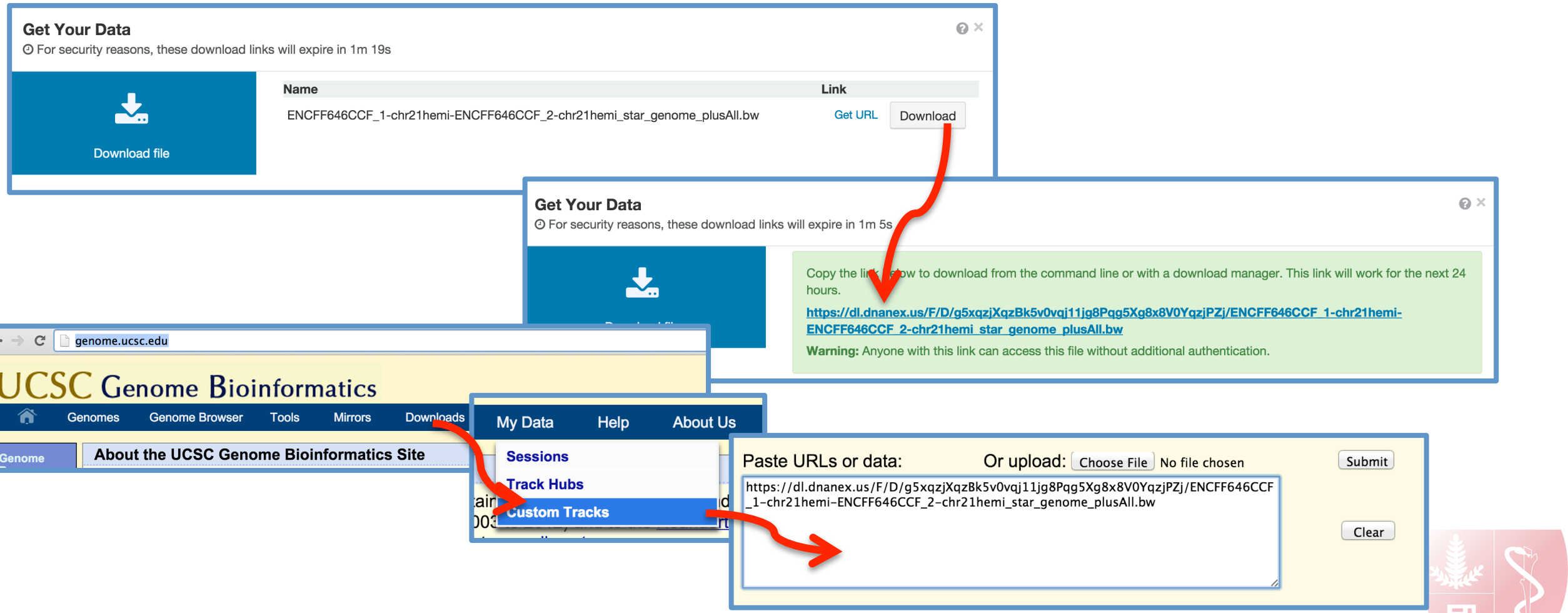
BigWig files:
- plus/minus strand (paired)
- uniquely mapped
- multi+uniquely mapped

Quantifications (RSEM):
- genome
- transcriptome

# Pick up the Results



# Visualize!


Get Your Data
For security reasons, these download links will expire in 1m 19s
Download file
Name
Link
ENCFF646CCF_1-chr21hemi-ENCFF646CCF_2-chr21hemi_star_genome_plusAll.bw
Get URL
Download
Get Your Data
For security reasons, these download links will expire in 1m 5s
Copy the link below to download from the command line or with a download manager. This link will work for the next 24 hours.
https://dl.dnanex.us/F/D/g5xqzjXqzBk5v0vqj11jg8Pqg5Xg8x8V0YqzjPZj/ENCFF646CCF_1-chr21hemi-ENCFF646CCF_2-chr21hemi_star_genome_plusAll.bw
Warning: Anyone with this link can access this file without additional authentication.
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:
Choose File
No file chosen
Submit
https://dl.dnanex.us/F/D/g5xqzjXqzBk5v0vqj11jg8Pqg5Xg8x8V0YqzjPZj/ENCFF646CCF_1-chr21hemi-ENCFF646CCF_2-chr21hemi_star_genome_plusAll.bw
Clear

# Visualize!

**Manage Custom Tracks**

**genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19) [hg19]

| Name | Description | Type | Doc | delete |
|---|---|---|---|---|
| ENCFF646CCF_1-chr21hemi-ENCFF646CCF_2-chr21hemi_star_genome_plusAll | ENCFF646CCF_1-chr21hemi-ENCFF646CCF_2-chr21hemi_star_genome_plusAll | bigWig | | |

add custom tracks
go to genome browser
go to table browser
go to variant annotation integrator

**Custom Tracks** refresh

ENCFF646CCF_1-chr21hemi-ENCFF646CCF_2-chr21hemi_star_genome_plusAll

dense