ENCODE Uniform Data Processing Pipelines Workshop

Ben Hitz and J. Seth Strattan
ENCODE DCC
ENCODE Research Applications and Users Meeting
July, 2015

To set up your environment
See the link below “Workshop Session 3: ENCODE Uniform Processing”
https://www.encodeproject.org/tutorials/encode-users-meeting-2015/

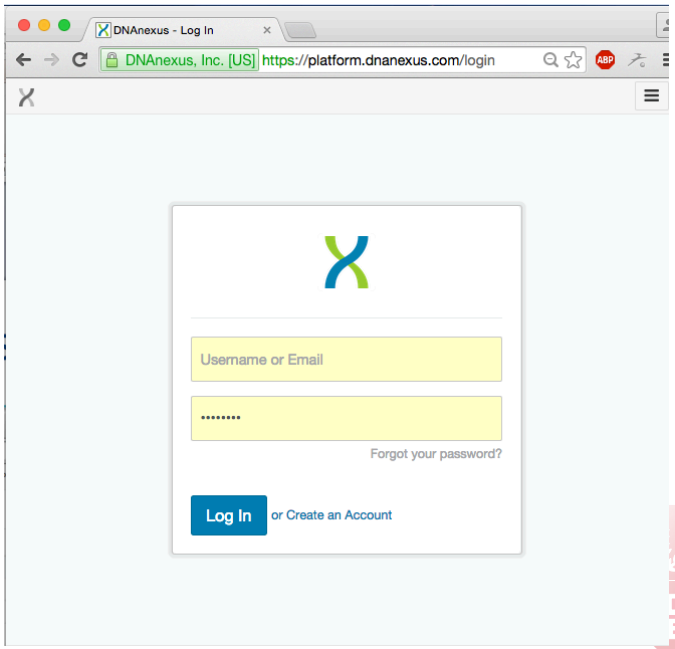
Pipelines Demonstration and Exercise

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 (Franklin Ballroom)
Prepare to run web-based pipelines | Presentation | Hands-on Tutorial
J. Seth Strattan, Stanford University
Ben Hitz, Stanford University

Log in ->



What would you like to learn?

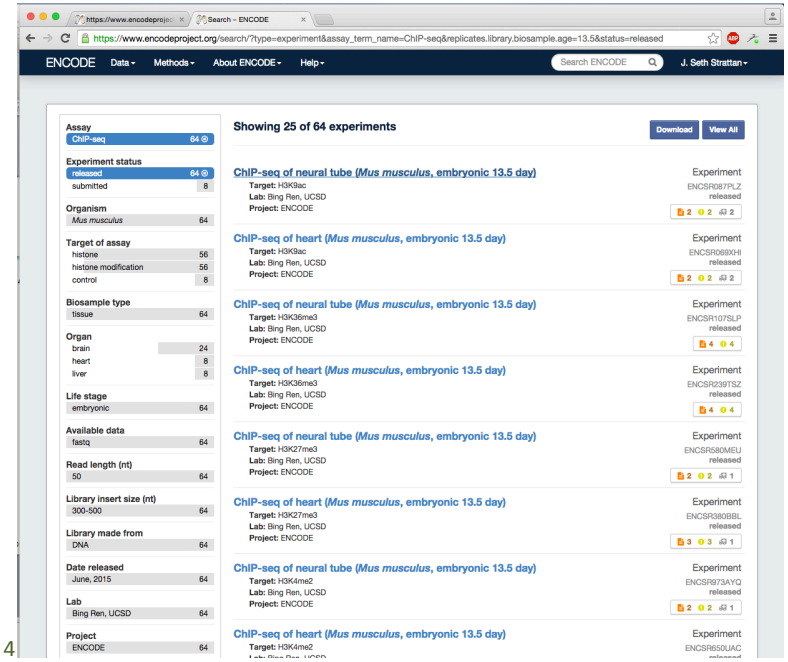
How many of you:

1. … have downloaded ENCODE data and intersected it with other data?
2. … have already implemented an analysis pipeline based on 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?

Pipelines Workshop in Context

Data Access → Visualization → Interpretation → Processing → Advanced Analysis

Eurie: ENCODE Portal

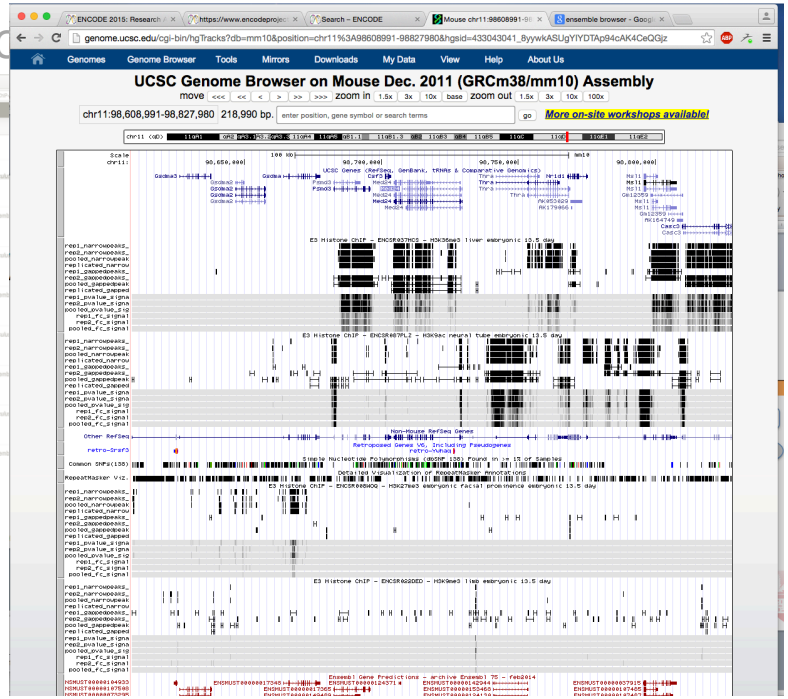


Pipelines Workshop in Context

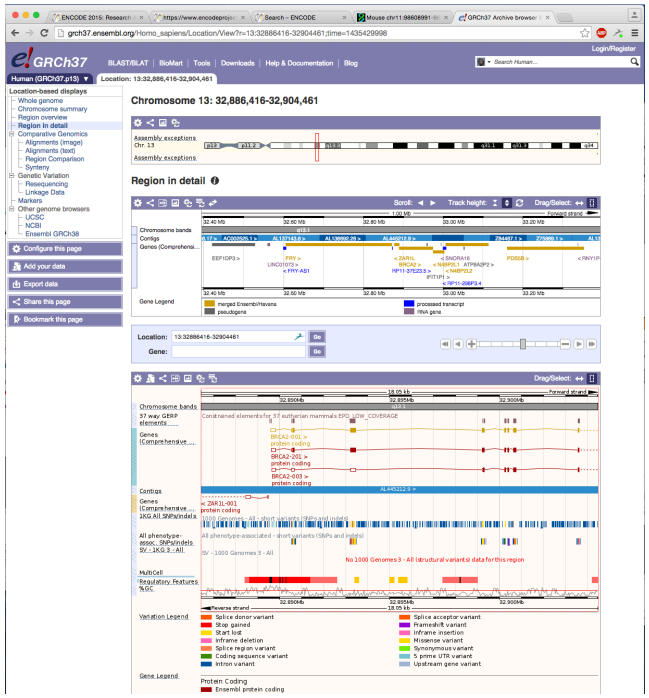
Data Access → Visualization → Interpretation → Processing → Advanced Analysis

Eurie: ENC

Pauline: UCSC Genome Browser



Emily: ENSEMBL Browser

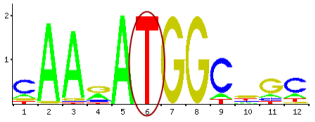


Pipelines Workshop in Context

Data Access → Visualization → Interpretation → Processing → Advanced Analysis

Emily: VEP

Does it affect TF binding?



GAACATGGCGGC

Score = 10.414

T/C

GAACACG

Score = 10

Jill: HaploReg and RegulomeDB

rs17293632 is Associated with IBD and Crohn's Disease

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

Query SNP: rs4637157 and variants with r^2 >= 0.8

chr	pos (hg19)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL tissues	Motifs changed	GENCODE genes	dbSNP func annot
2	29422	0.82	1	rs4263140	A	G	0.48	0.13	0.20	0.09			NHEK, HMEC	4 cell types	CEBPB		7 altered motifs	9.4kb 3' of FAM110C	
2	29443	1	1	rs4637157	T	C	0.39	0.12	0.17	0.08			NHEK, HMEC	4 cell types	CEBPB		8 altered motifs	9.4kb 3' of FAM110C	
2	30091	0.8	0.98	rs28446791	C	G	0.47	0.13	0.20	0.09								8.7kb 3' of FAM110C	
2	31318	0.96	0.98	rs6732811	G	C	0.40	0.12	0.16	0.08							6 altered motifs	7.5kb 3' of FAM110C	
2	31324	0.96	0.98	rs6706828	C	T	0.40	0.12	0.16	0.08							Ets,ZNF263	7.5kb 3' of FAM110C	
2	31791	0.98	1	rs28433318	C	T	0.52	0.13	0.20	0.08			NHEK				BAF155,CHD2	7kb 3' of FAM110C	
2	38733	0.8	0.98	rs112074103	GA	G	0.47	0.13	0.20	0.09			NHEK, HMEC	Fibrobl			TATA	80bp 3' of FAM110C	
2	39340	0.8	0.98	rs4530399	A	G	0.47	0.13	0.20	0.09			HMEC, NHEK				GCNF,Nr2f2,Zbtb3	FAM110C	3'-UTR
2	40569	0.8	0.98	rs6731388	T	C	0.52	0.14	0.20	0.09			HMEC, NHEK	Chorion,HeLa-S3	4 bound proteins		Pou2f2,Pou6f1,Rhox11	FAM110C	3'-UTR
2	41404	0.8	0.98	rs10173732	G	A	0.36	0.13	0.20	0.09		NHEK		H9ES			Spz1	FAM110C	3'-UTR
2	50092	0.96	0.98	rs6749595	T	C	0.54	0.13	0.20	0.08							4 altered motifs	3.2kb 5' of FAM110C	
2	53652	0.96	0.98	rs4438516	G	A	0.47	0.13	0.20	0.08							7 altered motifs	6.8kb 5' of FAM110C	
2	55007	0.96	0.98	rs112988427	CAG	C	0.47	0.13	0.20	0.08							GR,NF-I,TLX1::NFIC	8.1kb 5' of FAM110C	
2	55237	0.95	0.98	rs10188860	T	C	0.47	0.14	0.20	0.08							4 altered motifs	8.4kb 5' of FAM110C	
2	61687	0.98	1	rs10197241	A	T	0.44	0.13	0.20	0.08							4 altered motifs	15kb 5' of FAM110C	
2	66839	0.96	0.98	rs10200966	C	T	0.56	0.13	0.20	0.08			NHEK				GR	20kb 5' of FAM110C	
2	67321	0.96	0.98	rs11680031	G	A	0.56	0.13	0.20	0.08		K562	HMEC, NHEK				Ets,GR	20kb 5' of FAM110C	
2	70074	0.95	0.98	rs300761	A	G	0.56	0.14	0.20	0.08			NHEK, HMEC	Jurkat,PrEC	STAT1		Myc,Sox	23kb 5' of FAM110C	

Pipelines Workshop in Context

Data Access → Visualization → Interpretation → Processing → Advanced Analysis

Ben & Seth: ENCODE Processing Pipelines

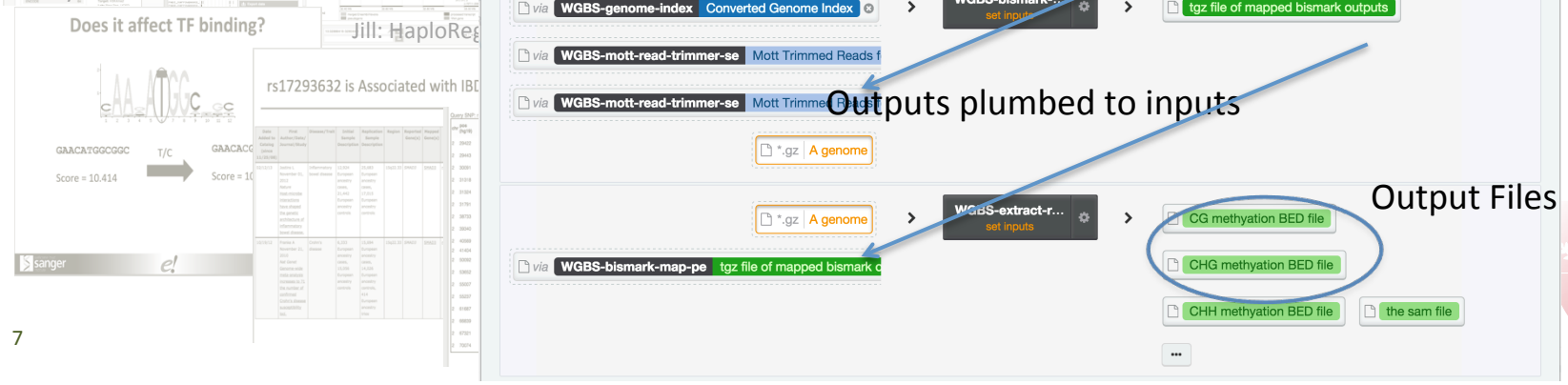
Eurie: ENCODE Portal

Pauline: UCSC Genome Browser

Emily: ENSEMBL Browse

Emily: VEP

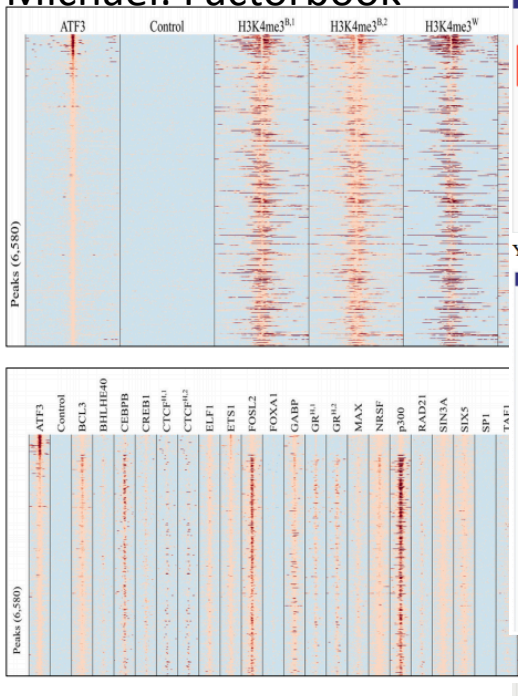
Jill: HaploReg



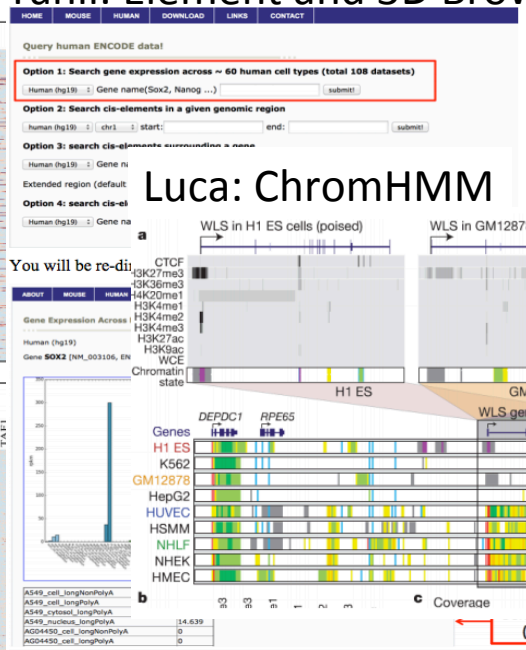
Pipelines Workshop in Context

Data Access → Visualization → Interpretation → Processing → Advanced Analysis

Michael: Factorbook

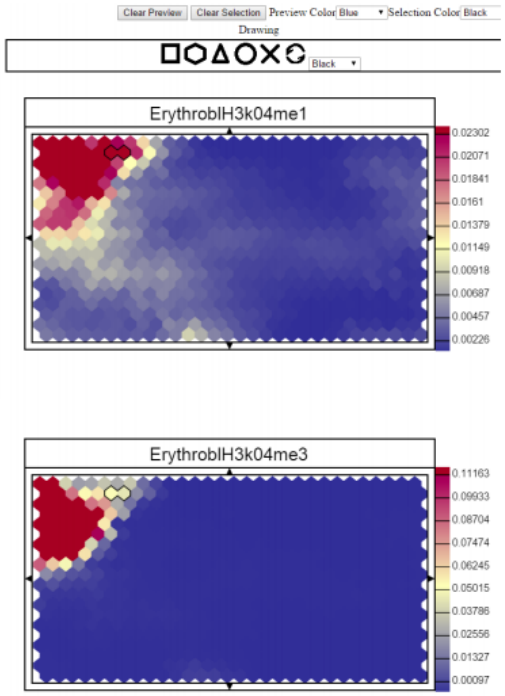


Yanli: Element and 3D Browser



Luca: ChromHMM

Camden: SOM's

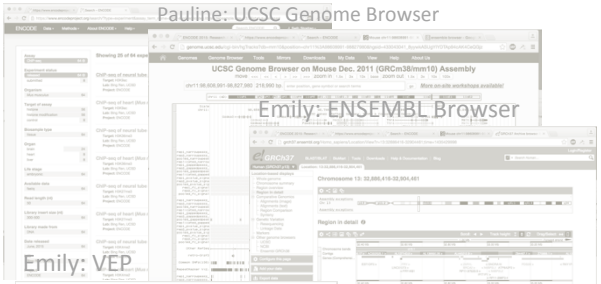


Pipelines Workshop in Context

Data Access → Visualization → Interpretation → <u>Processing</u> → Advanced Analysis

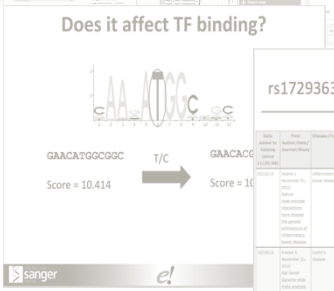
Eurie: ENCODE Portal

Pauline: UCSC Genome Browser

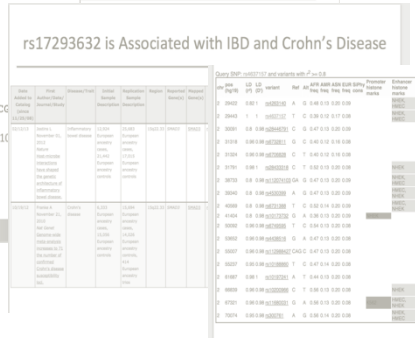


Emily: ENSEMBL Browser

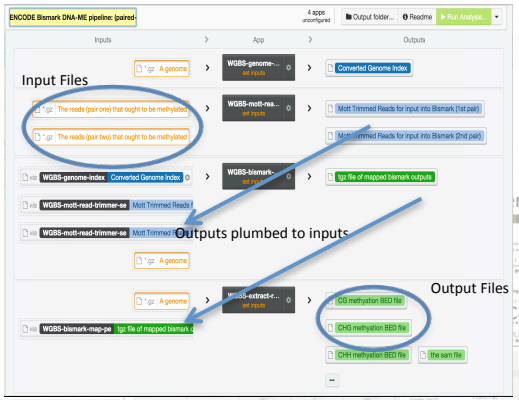
Emily: VEP

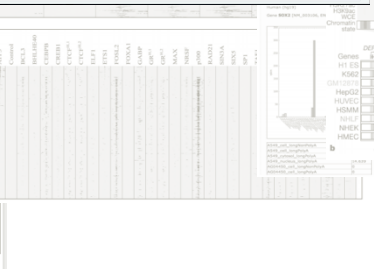


Jill: HaploReg and RegulomeDB



Ben & Seth: ENCODE Processing Pipelines

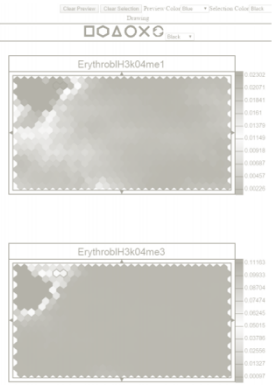




nt and 3D Browser

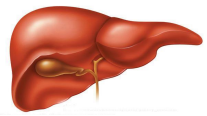
ca: ChromHMM

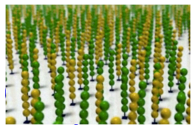
Camden: SOM's



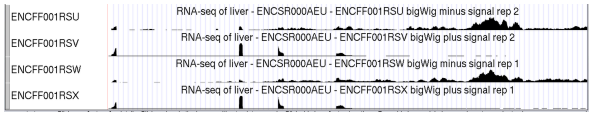
J. Seth Strattan, PhD ENCODE DCC

DCC Delivers ENCODE Data

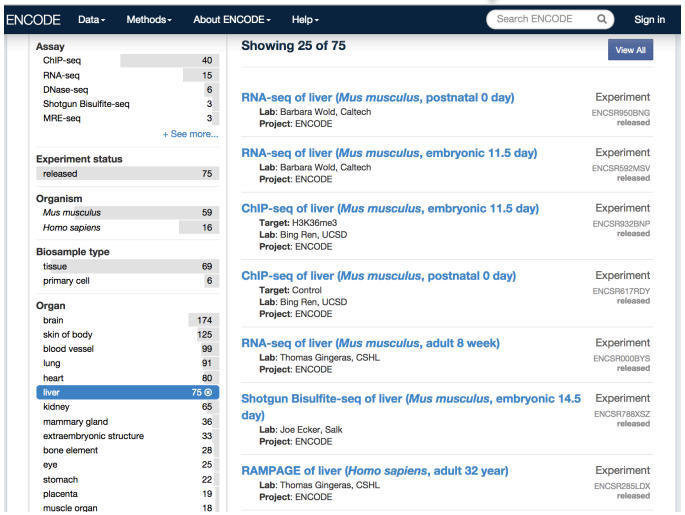




```
+
CCCFFFFFHHHHHGIJJIGGHEIIEGGEGGIJJBHIG
@BI:SL-HAB:D0RRAACXX:8:2309:21201:7829 1:X:0:GCCGTCGA
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
+
CCCFFFFFHHHHHHJJJJJJGJJJJIIJJJJGGIGJ
@BI:SL-HAB:D0RRAACXX:8:2113:4623:40045 1:X:0:GCCGTCGA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
+
??@ADDBDH:CDHHI+AEFHI?GGHII:EFIII?F=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



Sample → Library → Primary Data → Processed Data

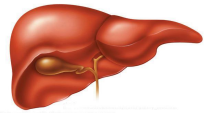


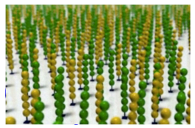
ENCFF001RTZ	fastq	reads	2	(1, 1)	Esther Chan	2013-08-15	Download	pending
ENCFF001RSV	bigWig	plus signal		(2, 1)	Esther Chan	2013-07-26	Download	pending
ENCFF001RSU	bigWig	minus signal		(2, 1)	Esther Chan	2013-07-26	Download	pending

AWS S3 Bucket
ENCODE Files

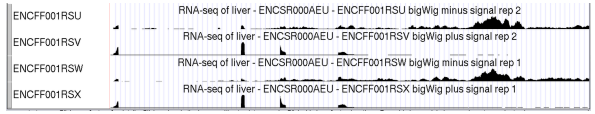
J. Seth Strattan, PhD ENCODE DCC

ENCODE DCC Delivers ENCODE Metadata

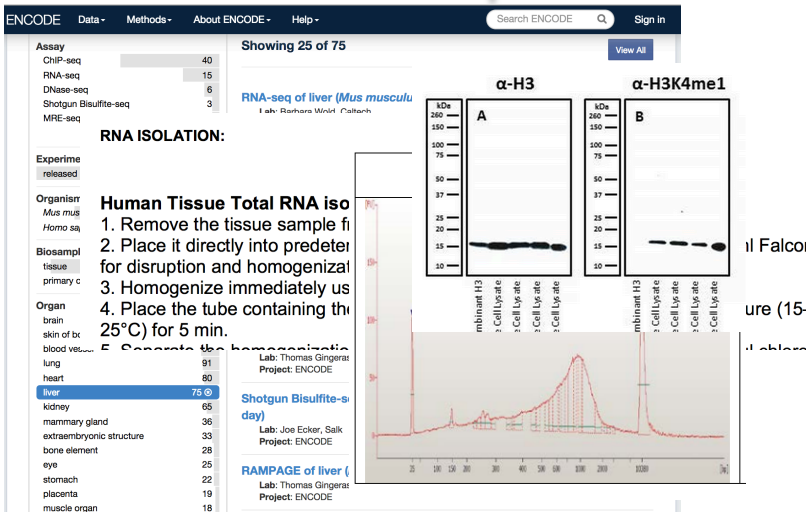




+
CCCFFFFFHHHHHGIJJIGGHEIIEGGEGGIJJBHIG
@BI:SL-HAB:D0RRAACXX:8:2309:21201:7829 1:X:0:GCCGTCGA
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
+
CCCFFFFFHHHHHHJJJJJJGJJJJIIJJJJGGIGJ
@BI:SL-HAB:D0RRAACXX:8:2113:4623:40045 1:X:0:GCCGTCGA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
+
??@ADDBDH:CDHHI+AEFHI?GGHII:EFIII?F=
@BI:SL-HAB:D0RRAACXX:8:2206:11680:21762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

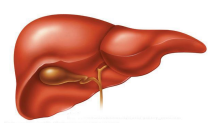


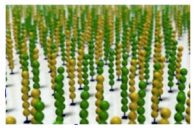
Sample → Library → Primary Data → Processed Data



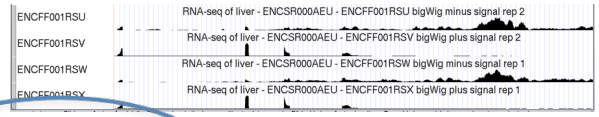
J. Seth Strattan, PhD ENCODE DCC

ENCODE Analysis Pipelines as Deliverables





```
+
CCCFFFFFHHHHHGIJJIGGHEIIEGGEGGIJJBHIG
@BI:SL-HAB:D0RRAACXX:8:2309:21201:7829 1:X:0:GCCGTCGA
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
+
CCCFFFFFHHHHHHJJJJJJGJJJJIIJJJJGGIGJ
@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
+
```



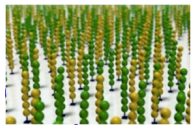
Sample → Library → Primary Data → Processed Data

Goals:

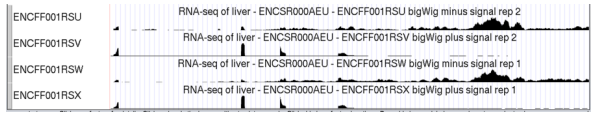
1. Deploy ENCODE-defined pipelines for ChIP-seq, RNA-seq, DNase-seq, methylation.
2. Use those pipelines to generate the standard ENCODE peaks, quantitations, CpG.
3. Capture metadata to make clear what software, versions, parameters, inputs were used.
4. Capture, accession, and distribute the output.
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

Deployment Platform Considerations



```
+
CCCFFFFFHHHHHGIJJIGGHEIIEGGEGGIJJBHIG
@BI:SL-HAB:D0RRAACXX:8:2309:21201:7829 1:X:0:GCCGTCGA
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
+
CCCFFFFFHHHHHHJJJJJJGJJJJIIJJJJGGIGJ
@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 and adaptable for deployment to your HPC environment

https://github.com/ENCODE-DCC

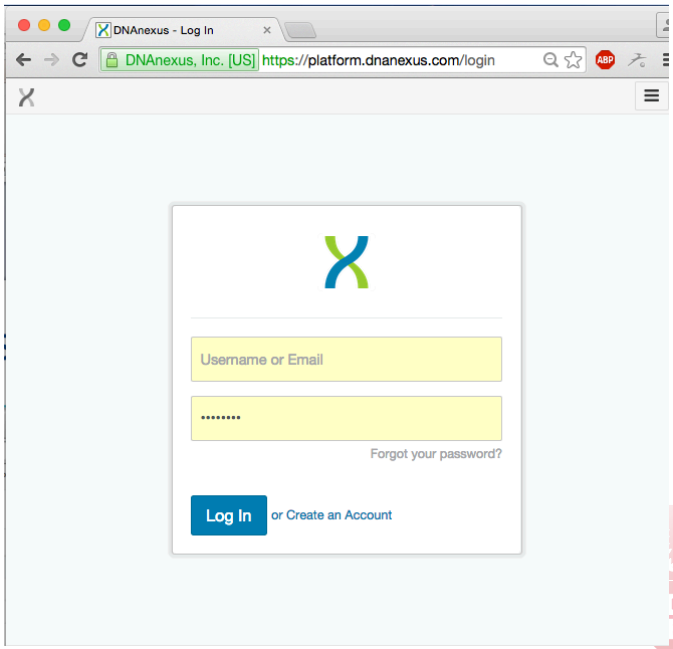
Pipelines Demonstration and Exercise

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 (Franklin Ballroom)
Prepare to run web-based pipelines | Presentation | Hands-on Tutorial
J. Seth Strattan, Stanford University
Ben Hitz, Stanford University

Log in ->

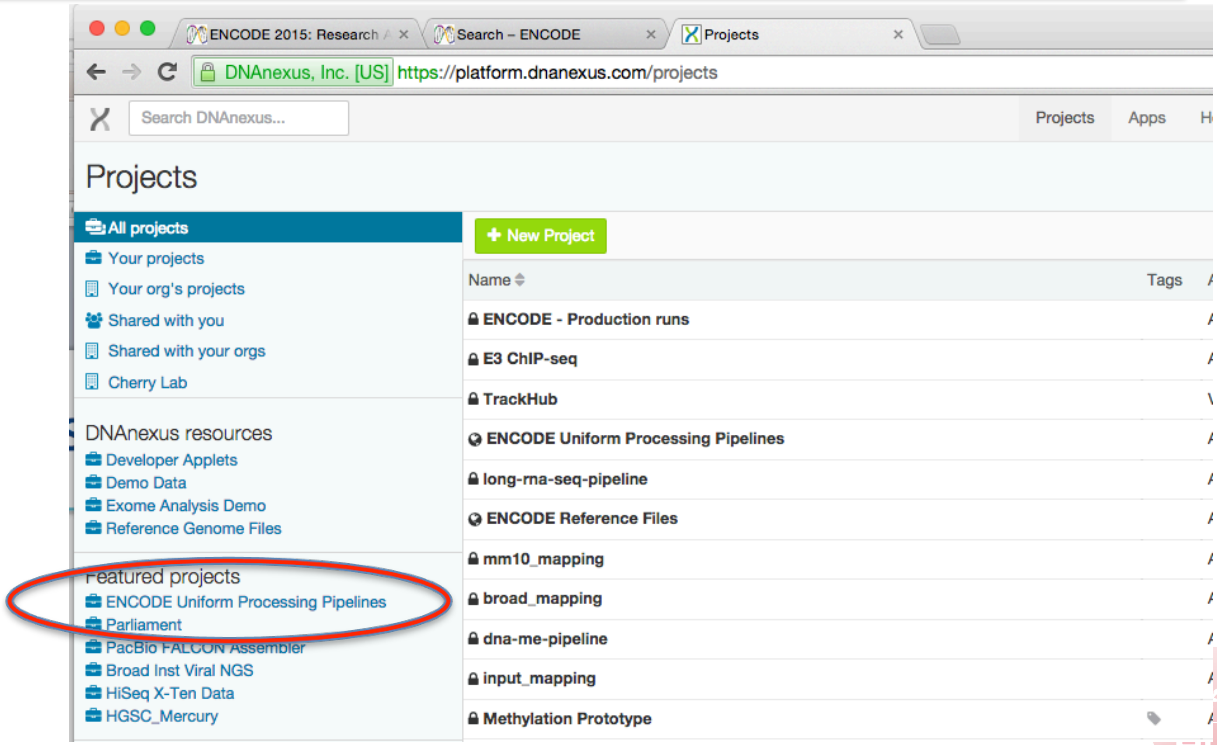


Pipelines Demonstration and Exercise

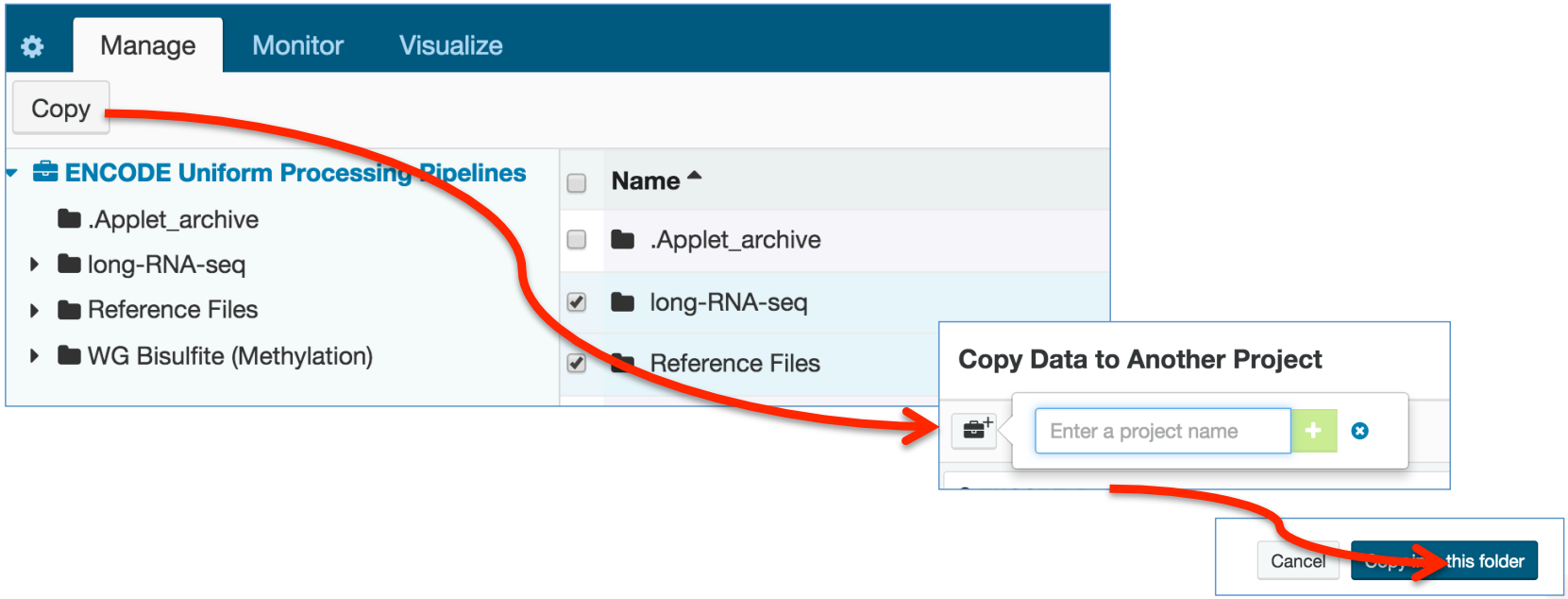
Select:

Featured Projects:

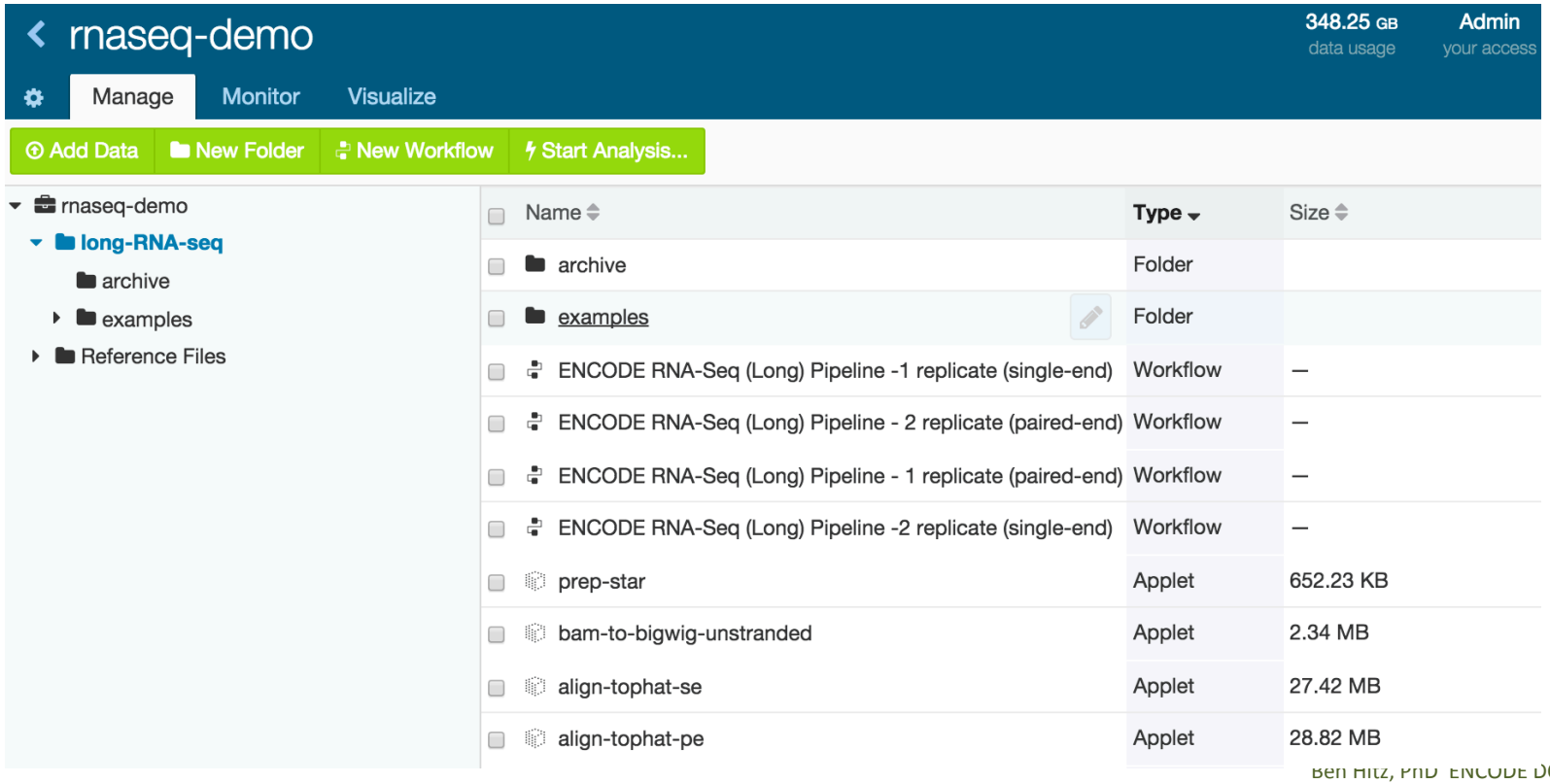
.. ENCODE Uniform Processing Pipelines



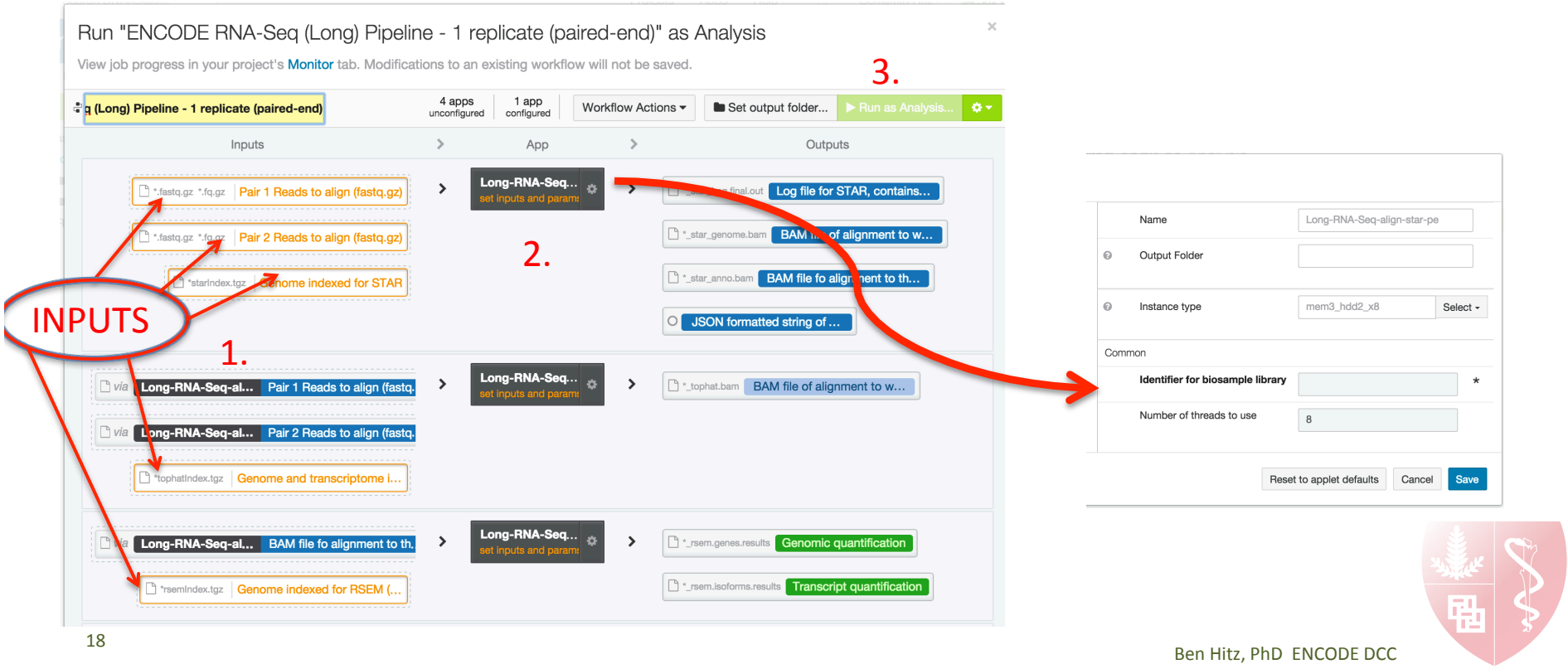
Setup the Demo Project



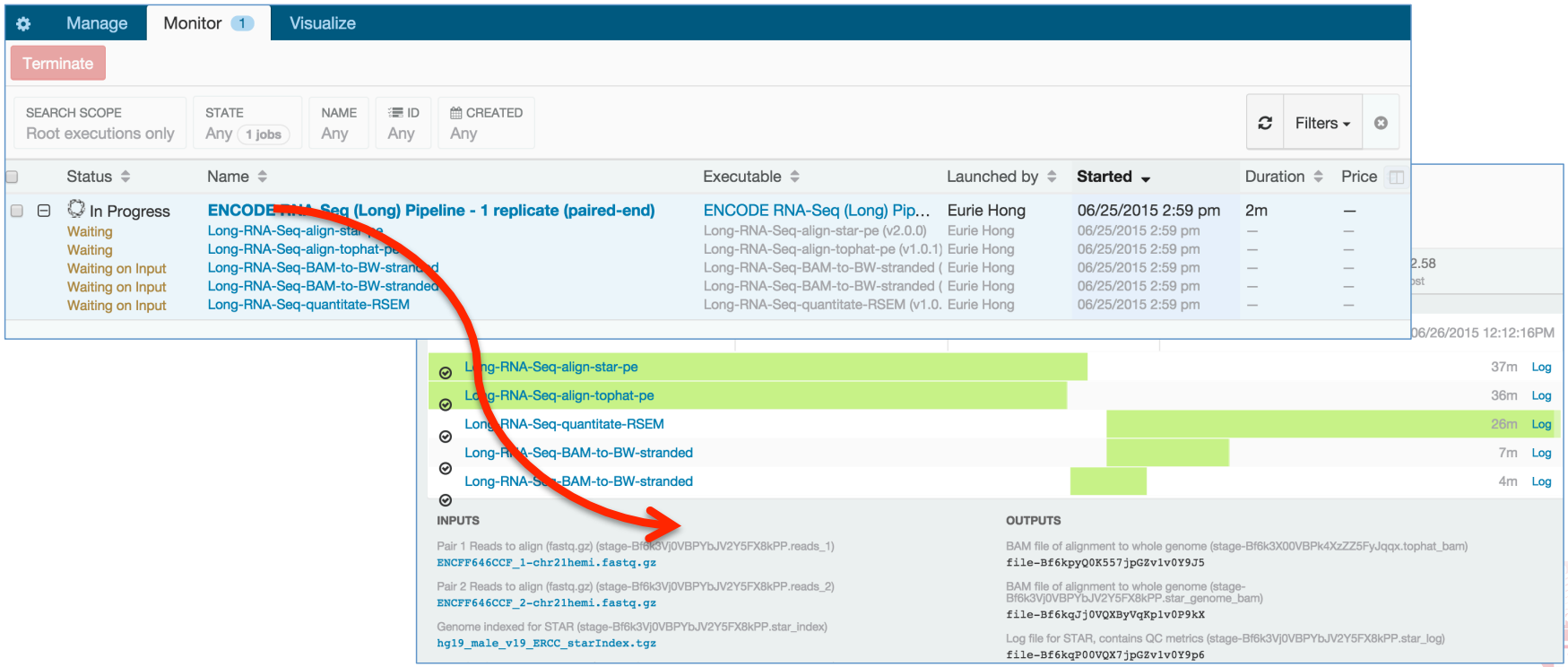
Project Overview



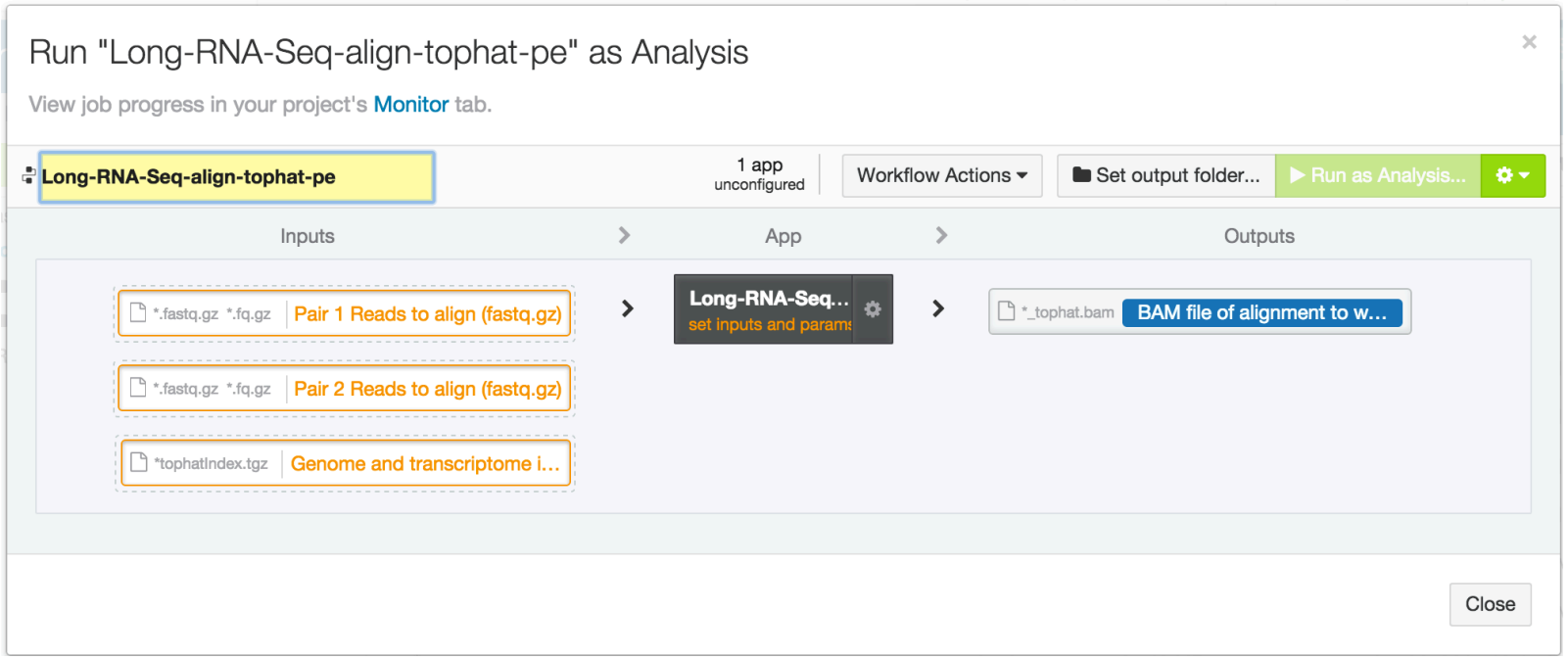
A Workflow


Run "ENCODE RNA-Seq (Long) Pipeline - 1 replicate (paired-end)" as Analysis
View job progress in your project's Monitor tab. Modifications to an existing workflow will not be saved.
3.
(Long) Pipeline - 1 replicate (paired-end)
4 apps unconfigured
1 app configured
Workflow Actions
Set output folder...
Run as Analysis...
Inputs
App
Outputs
INPUTS
1.
2.
*.fastq.gz *.fq.gz
Pair 1 Reads to align (fastq.gz)
Pair 2 Reads to align (fastq.gz)
*starIndex.tgz
Genome indexed for STAR
Long-RNA-Seq...
set inputs and params
Log file for STAR, contains...
*_star_genome.bam
*_star_anno.bam
JSON formatted string of ...
via
Long-RNA-Seq-al...
Pair 1 Reads to align (fastq.
Pair 2 Reads to align (fastq.
*tophatIndex.tgz
Genome and transcriptome i...
*_tophat.bam
BAM file of alignment to w...
BAM file fo alignment to th.
*rsemIndex.tgz
Genome indexed for RSEM (...
*_rsem.genes.results
Genomic quantification
*_rsem.isoforms.results
Transcript quantification
Name
Long-RNA-Seq-align-star-pe
Output Folder
Instance type
mem3_hdd2_x8
Select
Common
Identifier for biosample library
Number of threads to use
8
Reset to applet defaults
Cancel
Save

The Monitor Tab



An Applet


Run "Long-RNA-Seq-align-tophat-pe" as Analysis
View job progress in your project's Monitor tab.
Long-RNA-Seq-align-tophat-pe
1 app
unconfigured
Workflow Actions
Set output folder...
Run as Analysis...
Inputs
App
Outputs
*.fastq.gz *.fq.gz
Pair 1 Reads to align (fastq.gz)
Long-RNA-Seq...
set inputs and params
*_tophat.bam
BAM file of alignment to w...
*.fastq.gz *.fq.gz
Pair 2 Reads to align (fastq.gz)
*tophatIndex.tgz
Genome and transcriptome i...
Close

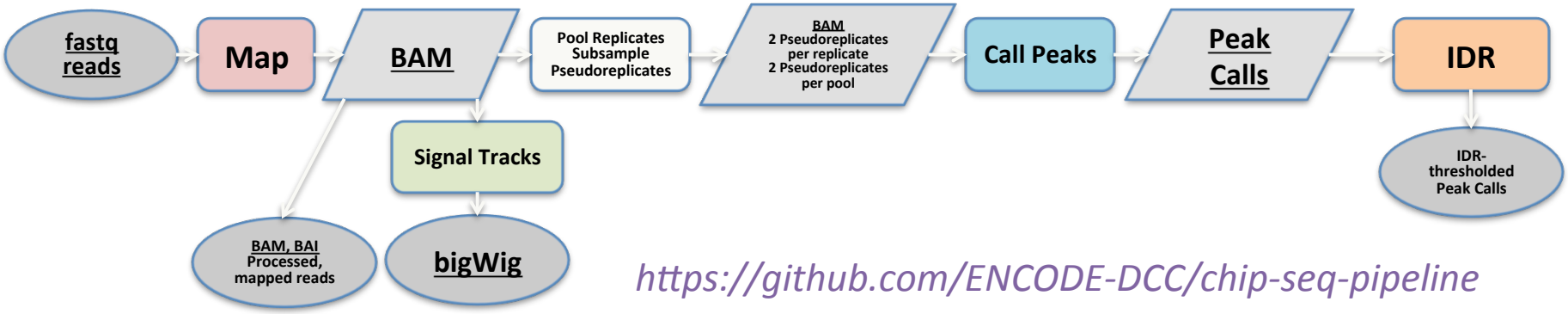
An RNA-seq Workflow

Interregnum
What/Where are the ENCODE Results?

<u>While the example is running:</u>

- What steps do the pipelines run?
- What inputs do they take?
- What outputs do they produce?
- Where are the Uniform ENCODE Results?

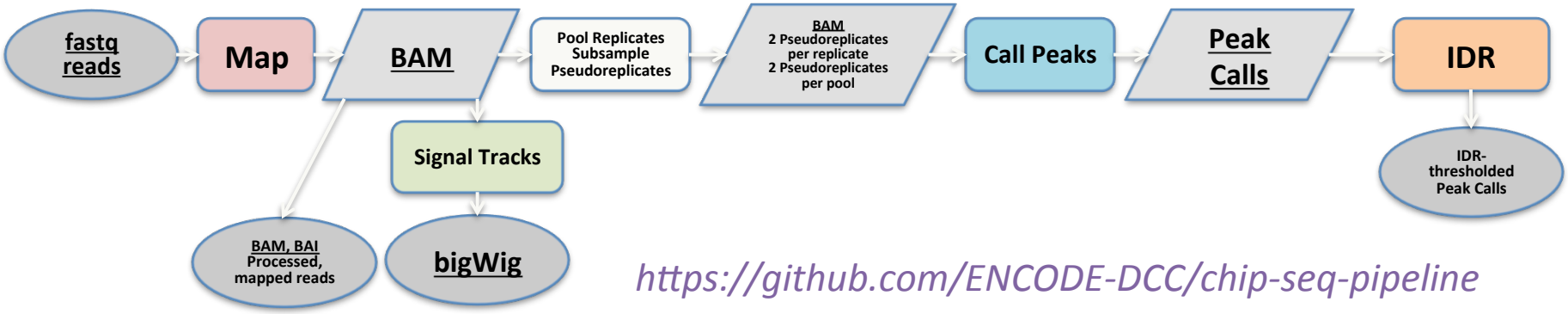
Schema: ENCODE ChIP-seq IDR Pipeline



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

Target	Key Software	Input Files	Output Files	QA Metrics
TF's	bwa Picard markDuplicates samtools MACS2 (Signal tracks) SPP (PeakSeq, GEM future) IDR2	fastq's (SE or PE) Two biological replicates Matched controls	One bam per replicate bigWig fold signal over control bigWig p-value signal over control bed/bigBed true replicates peaks bed/bigBed pooled replicates peaks bed/bigBed IDR thresholded peaks	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
Histone Mods	MACS2 for peaks Overlap thresholding IDR2 (future)		bed/bigBed Replicated peaks	

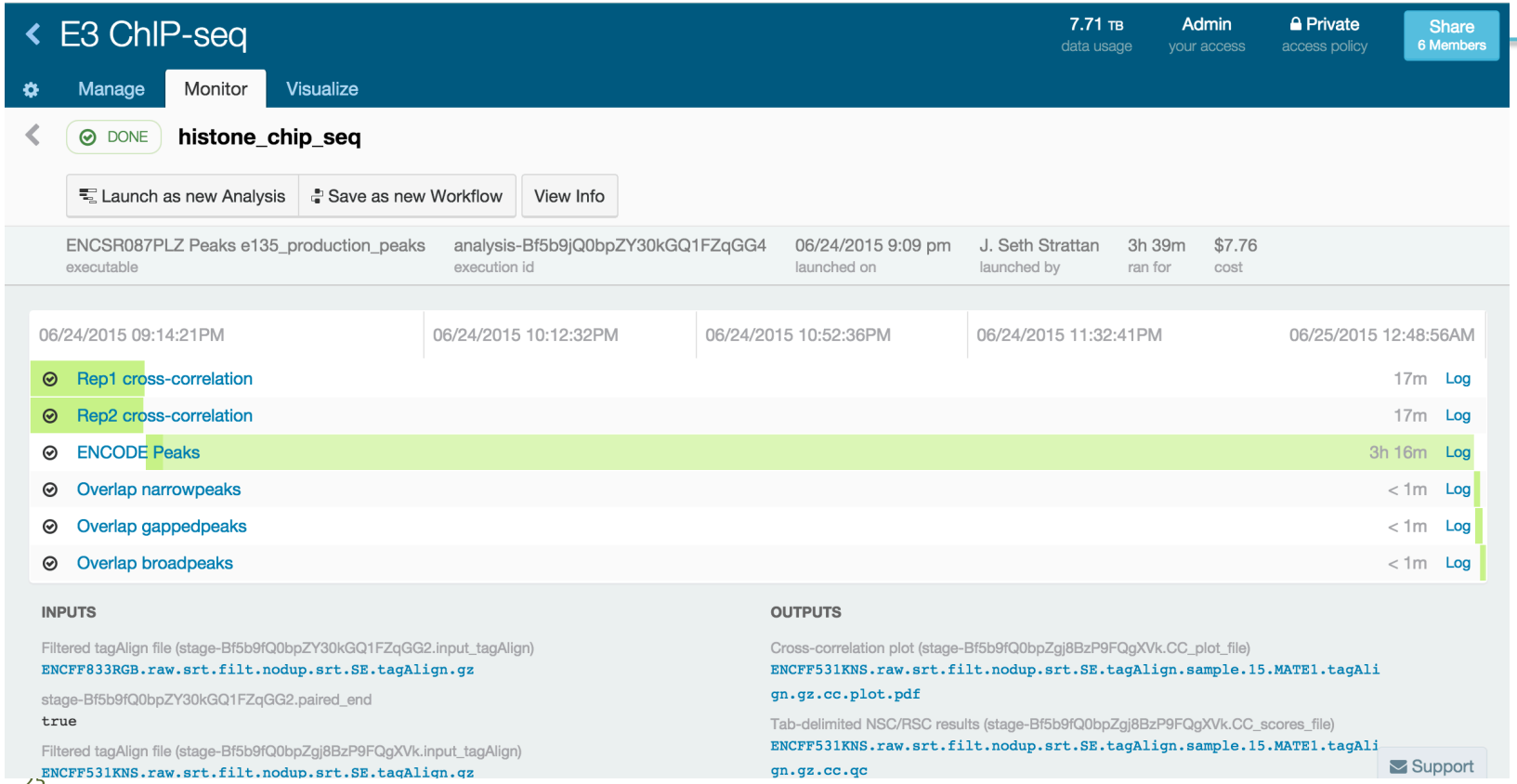
ENCODE ChIP-seq Quality Metrics: Resources



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

Estimates	Description	References
Depth	Number of uniquely mapping reads Number of distinct uniquly mapping reads	Jung YL, et al. Nucleic Acids Research. 2014;42(9):e74
Library Complexity **ChIP Quality**	Non-Redundant Fraction PCR Bottleneck Coefficient Normalized Strand Cross-Correlation Relative Strand Cross-Correlation	Landt S, et al. Genome Res. 2012. 22: 1813-1831
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

ENCODE ChIP-seq on the Cloud



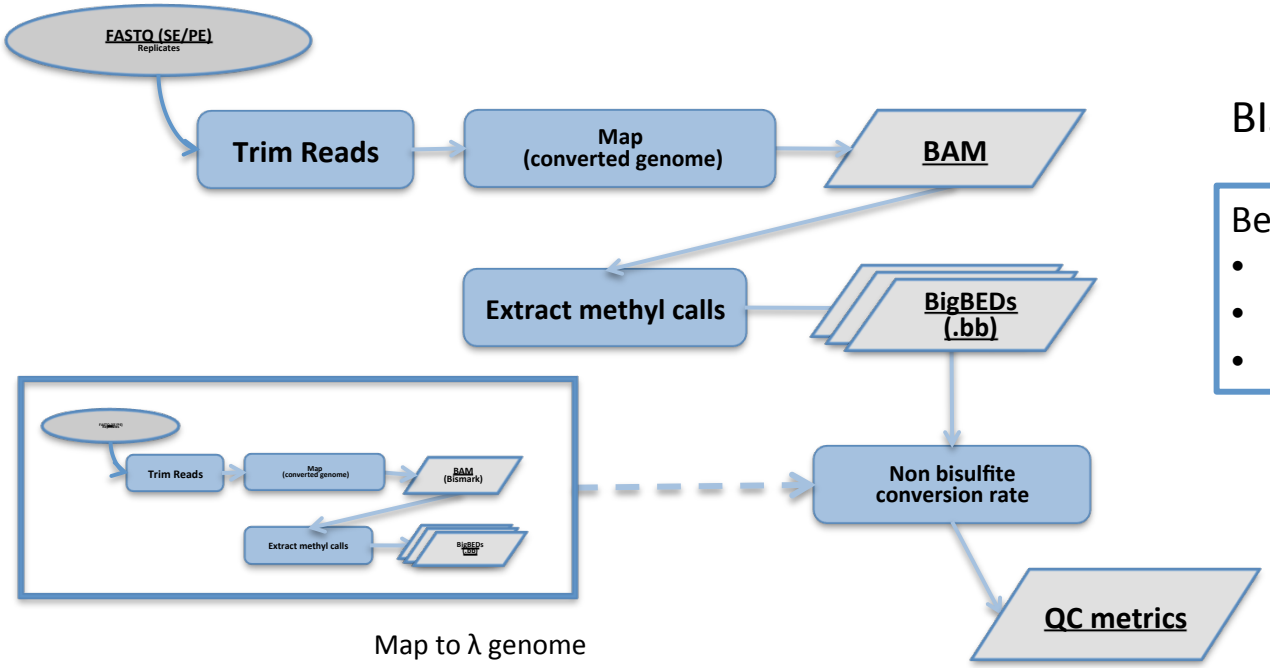
Uniformly Processed Data On the ENCODE Portal

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

Schema: ENCODE WGBS Pipeline

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



Map to λ genome

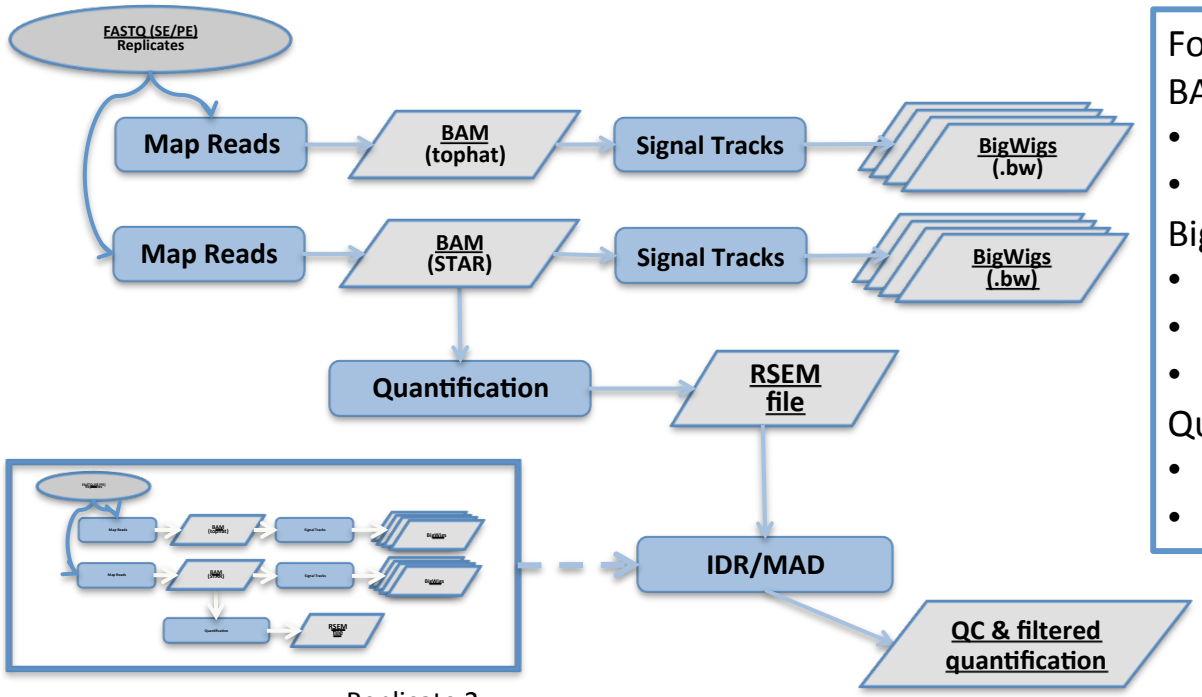
BISMARK (v 0.10)

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

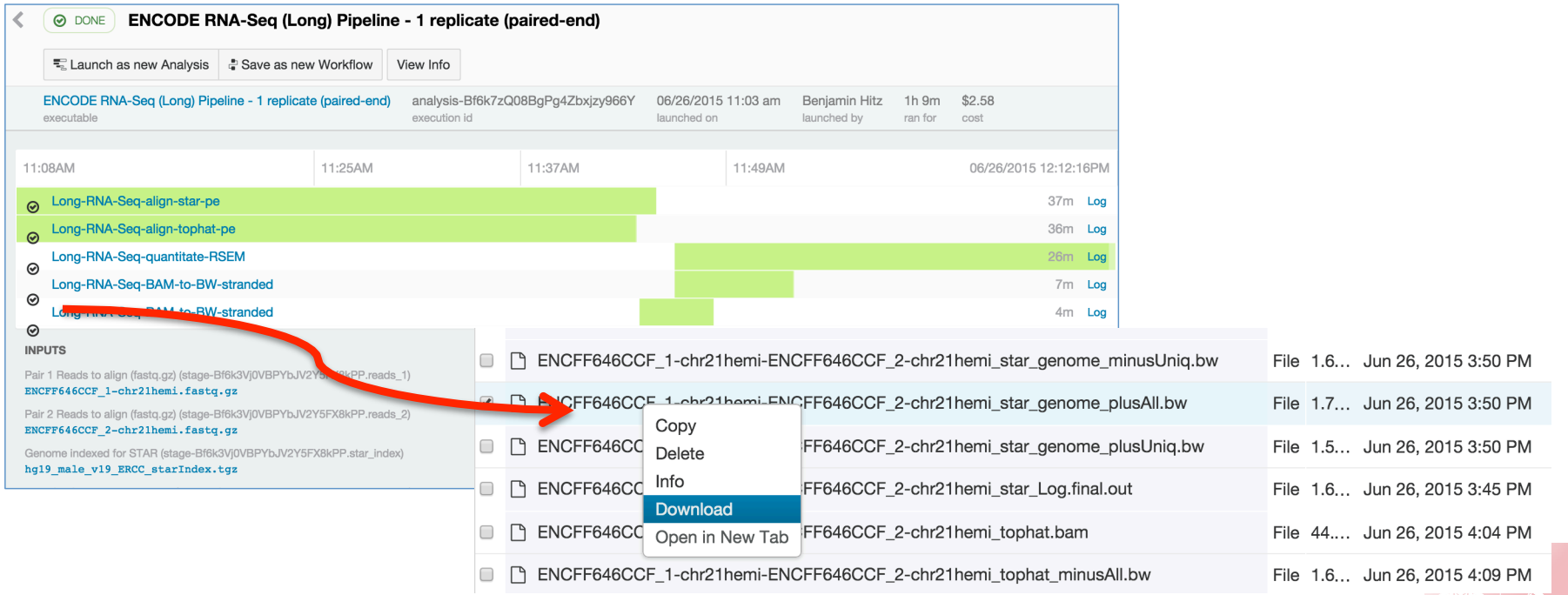
Quantifications (RSEM):
- genome
- transcriptome

Uniformly Processed Data On the ENCODE Portal

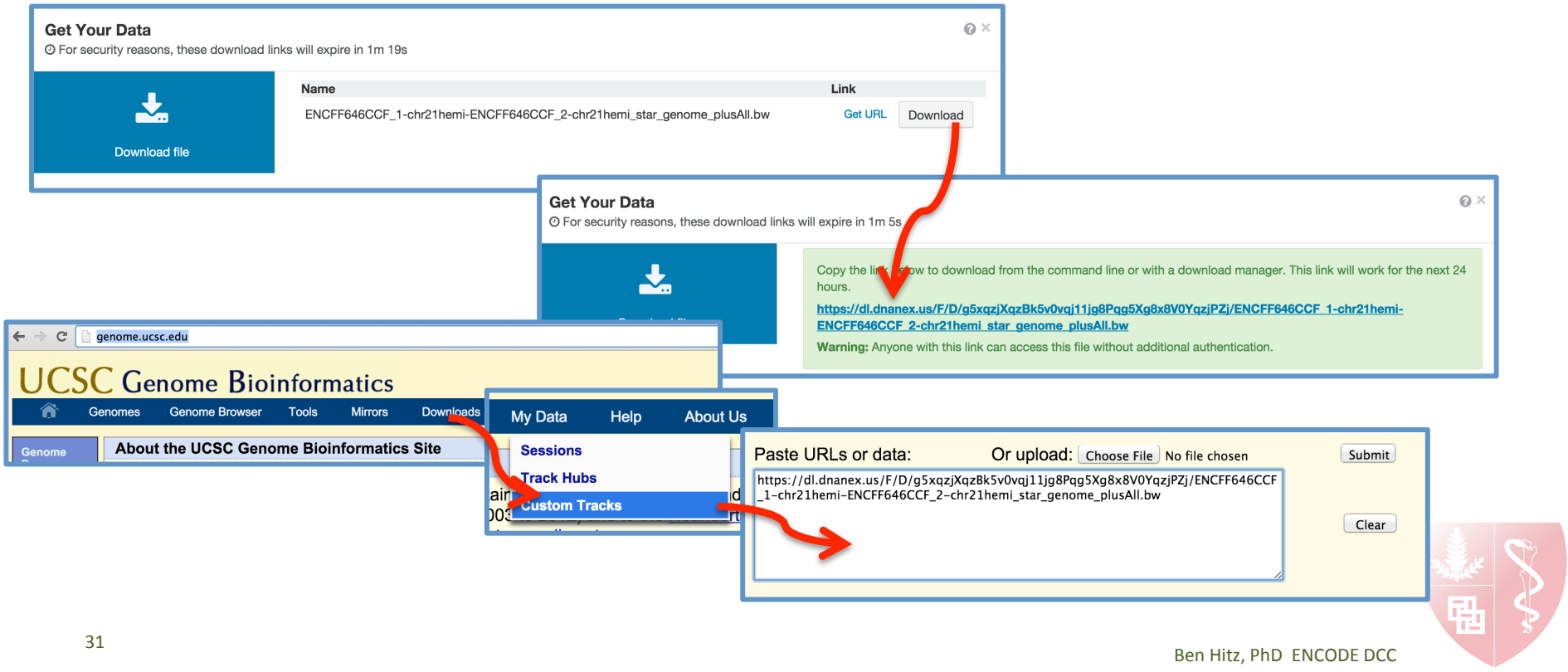
RNA-seq Example
https://www.encodeproject.org/experiments/ENCSR368QPC/

- 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

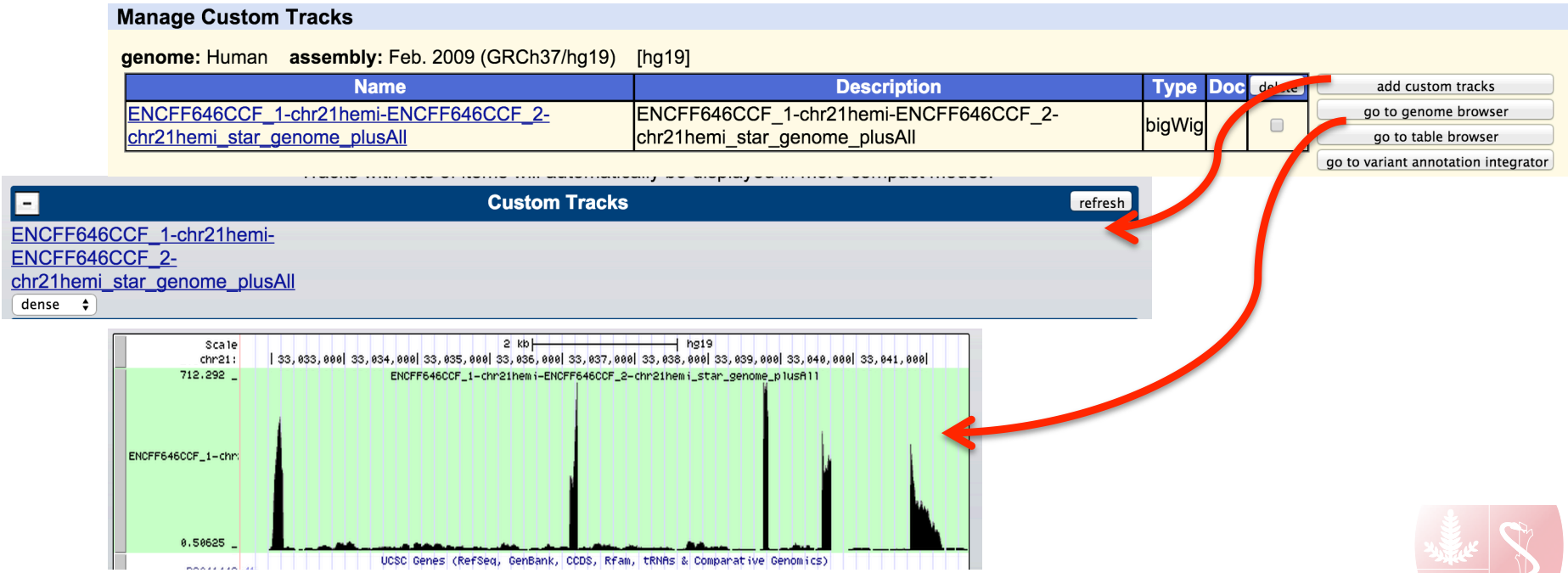
Pick up the Results



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
Sessions
Track Hubs
Custom Tracks
About the UCSC Genome Bioinformatics Site
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!

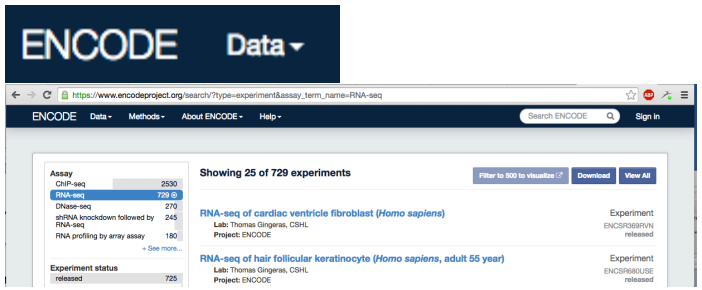

Manage Custom Tracks
genome: Human assembly: Feb. 2009 (GRCh37/hg19) [hg19]
Name
Description
Type
Doc
ENCFF646CCF_1-chr21hemi-ENCFF646CCF_2-chr21hemi_star_genome_plusAll
ENCFF646CCF_1-chr21hemi-ENCFF646CCF_2-chr21hemi_star_genome_plusAll
bigWig
add custom tracks
go to genome browser
go to table browser
go to variant annotation integrator
Custom Tracks
refresh
ENCFF646CCF_1-chr21hemi-ENCFF646CCF_2-chr21hemi_star_genome_plusAll
dense
Scale
chr21:
2 kb
hg19
33,033,000 33,034,000 33,035,000 33,036,000 33,037,000 33,038,000 33,039,000 33,040,000 33,041,000
712.292
ENCFF646CCF_1-chr21hemi-ENCFF646CCF_2-chr21hemi_star_genome_plusAll
ENCFF646CCF_1-chr
0.50625
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)

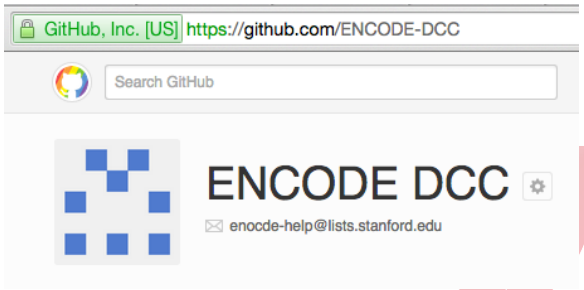
Pipeline Workshop Summary

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





J. Seth Strattan, PhD ENCODE DCC

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/