ENCODE ANALYSIS PIPELINES

J. Seth Strattan, PhD
ENCODE Data Coordinating Center (DCC)
ENCODE User’s Meeting Workshop
June, 2016
To set up an account:

https://www.encodeproject.org/tutorials/encode-meeting-2016/

Scroll down to "Preparing to Run ENCODE Pipelines"
Workshop Goals

- Introduce the ENCODE Analysis Pipelines.
- Run the transcription factor ChIP-seq pipeline on a ZBED1 ChIP experiment in K562.
- Run the long RNA-seq pipeline on a total-RNA experiment from a human tissue sample.
- Understand the pipeline inputs, outputs, and QC metrics and how to navigate them.
- Visualize the results of your analyses.
- Take home the ability to replicate ENCODE analyses on your own data.
ENCODE DCC Delivers ENCODE Metadata

Sample → Library → Primary Data → Processed Data
DCC Delivers ENCODE Data

Sample → Library → Primary Data → Processed Data

Sample

Library

Primary Data

Processed Data

AWS S3 Bucket

ENCODE Files

J. Seth Strattan, PhD ENCODE DCC
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
Deployment Platform Considerations

<table>
<thead>
<tr>
<th>Sample</th>
<th>Library</th>
<th>Primary Data</th>
<th>Processed Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replicability</td>
<td>Provenance</td>
<td>Ease of Use</td>
<td>Scalability</td>
</tr>
<tr>
<td>Hard</td>
<td>Moderate</td>
<td>Easy</td>
<td>Highly</td>
</tr>
</tbody>
</table>

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
**Schema: ENCODE ChIP-seq IDR 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) | | | |

https://github.com/ENCODE-DCC/chip-seq-pipeline
Analysis Pipeline Demonstration and Workshop

To set up an account:

https://www.encodeproject.org/tutorials/encode-meeting-2016/

Scroll down to “Preparing to Run ENCODE Pipelines”

Exercise 1  
TF Chip-seq  

Exercise 2  
RNA-seq  

Exercise 3  
Histone ChIP-seq
ChIP-seq Results On the ENCODE Portal

Histone ChIP-seq Example
https://www.encodeproject.org/experiments/ENCSR087PLZ/

• 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/ENCSR286PCG/

• 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 of the pooled replicates.
### What are the Pipeline Outputs?

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Content</th>
<th>Stage : Output</th>
<th>Basic Files</th>
<th>Stage : Output</th>
<th>Special-Purpose Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF ChIP</td>
<td>Optimal IDR Peaks</td>
<td>Final IDR Peak Calls : optimal_set_bb</td>
<td>Conservative IDR Peaks (comparing true replicates only, not pool)</td>
<td>Final IDR Peak Calls : conservative_set_bb</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pooled control-normalized signal</td>
<td>Final IDR Peak Calls : pooled_signal</td>
<td>Final IDR Peak Calls : rep[1,2]_signal</td>
<td>Final IDR Peak Calls : rep[1,2]_pvalue_signal</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Per-replicate signals (visual check of replication)</td>
<td></td>
<td>p-value signals (vs the control)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ENCODE Peaks : rep[1,2]_pvalue_signal</td>
</tr>
<tr>
<td>Histone ChIP</td>
<td>Replicated narrowPeaks</td>
<td>Final narrowpeaks : overlapping_peaks</td>
<td>Per-replicate signals (visual check of replication)</td>
<td>Final narrowpeaks : rep[1,2]_signal</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>long RNA-seq</td>
<td>plus/minus strand signal from uniquely</td>
<td>bam to stranded signal : [plus,minus]_uniq_bw</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mapping reads</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Per-gene quantitation</td>
<td>RSEM quantify genes : rsem_gene_results</td>
<td></td>
<td></td>
<td>Isoform quantitation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RSEM quantify genes : rsem_iso_results</td>
</tr>
</tbody>
</table>

**TF ChIP**
- **Optimal IDR Peaks**: Final IDR Peak Calls : optimal_set_bb
- **Pooled control-normalized signal**: Final IDR Peak Calls : pooled_signal

**Histone ChIP**
- **Replicated narrowPeaks**: Final narrowpeaks : overlapping_peaks
- **Pooled control-normalized signal**: Final narrowpeaks : pooled_signal
- **gappedPeaks (a set of connected narrowPeaks)**: Final gappedpeaks : overlapping_peaks
- **broadPeaks (a region of enrichment)**: Final broadpeaks : overlapping_peaks

**long RNA-seq**
- **plus/minus strand signal from uniquely mapping reads**: bam to stranded signal : [plus,minus]_uniq_bw
- **Per-gene quantitation**: RSEM quantify genes : rsem_gene_results
ENCODE ChIP-seq Quality Metrics: Resources

**Estimates** | **Description** | **References**
---|---|---
**Depth** | Number of mapped reads<br>Number of useable fragments after filtration | Jung YL, et al. Nucleic Acids Research. 2014;42(9):e74
**Library Complexity** | Non-Redundant Fraction (NRF)<br>PCR Bottleneck Coefficient (PBC1 and PBC2)<br>Strand Cross-Correlation (NSC and RSC)<br>Cross-Correlation Plot | Landt S, et al. Genome Res. 2012. 22: 1813-1831
**Replicate Concordance** | | |
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
RNA-seq Results 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
Results from the ChIP-seq exercise
Results from the ChIP-seq exercise

OUTPUTS

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

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

Number of peaks in the optimal set (No)
169

Number of peaks in the conservative set (Nc)
117

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

Self-consistency ratio (self_consistency_ratio)
1.2127659574468086

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

Self-consistency ratio (self_consistency_ratio)
1.2127659574468086

Pooled replicates signal (pooled_signal)
pool.fc_signal.bw
Results from the ChIP-seq exercise

Final peak calls - optimal set bigwig:
IDR_final_optimal.narrowPeak

Final peak calls - conservative set bigwig:
IDR_final_conservative.narrowPeak

Number of peaks in the optimal set:
169

Number of peaks in the conservative set:
117

Potential output files:
- IDR_final_optimal.narrowPeak
- IDR_final_optimal.narrowPeak.gz
- IDR_final_optimal.narrowPeak.bw
Results from the ChIP-seq exercise

“Download” to generate temporary URL’s to the selected files
Visualize on the UCSC Genome Browser
Visualize on the UCSC Genome Browser

Manage Custom Tracks

- Name: IDR_final_optimal.narrowPeak
- Description: IDR_final_optimal.narrowPeak
- Type: bigBed

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

- Move: less than, less than, less than, greater than, greater than, greater than
- Zoom in: 1.5x, 3x, 10x, base
- Zoom out: 1.5x, 3x, 10x, 100x

chr21: 33,001,785-33,064,574, 62,790 bp.

Microsatellite
Segmental Dups
COSMIC SNP2(146)
RepeatMaster

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.
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
Contributors

**ENCODE Data Coordinating Center**
Mike Cherry, PI, Stanford
Ben Hitz, Project Manager

**Pipeline Developers**
Ben Hitz, WGBS, Software Lead
Tim Dreszer, RNA-seq, DNAseq-seq
J. Seth Strattan, ChIP-seq

**Portal Developers**
Ulugbek Baymuradov
Laurence Rowe
Forrest Tanaka
Chris Thomas

**Data Wranglers**
Esther Chan
Jean Davidson
Idan Gabdank
Jason Hilton
Cricket Sloan

**QA & Biocuration Assistance**
Marcus Ho
Aditi Narayanan
Kathrina Onate

**Support Staff**
Stuart Miyasato
Matt Simison

**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 Guigo, CRG
Sarah Djebali, CRG

**ChIP-seq**
Anshul Kundaje, Stanford
Nathan Boley, Stanford
Jin Lee, Stanford

**DNA nexus**
Mike Lin
Andey Kislyuk
Singer Ma
Brett Hannigan
Ohad Rodeh
Joe Dale
George Asimenos
Alpha Diallo
Chai Fungtammasan
Naina Thangaraj

@encodedcc  encode-help@lists.stanford.edu  https://github.com/ENCODE-DCC/
Schema: ENCODE WGBS Pipeline

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

BISMARK (v 0.10)

Bed/BigBed files for:
- CG context
- CHG context
- CHH context