Basics of ChIP-Seq
Lauren Mills Ph.D
RISS @ MSI

Slides: https://www.msi.umn.edu/tutorial-materials
Outline

• What can you learn from ChIP?
• Planning an experiment
• Sequencing considerations
• Identifying Peaks
• Looking at your data
• Interacting with data
• Data Sources
ChIP-Seq

Chromatin immunoprecipitation to isolate fragments of DNA bound by protein of interest. Can be used to identify regions of the genome bound by specific Transcription Factors, Histone Modifications and RNA Polymerase.
ChIP-Seq

- Chomatin immunoprecipitation to isolate fragments of DNA bound by protein of interest.
- Can be used to identify regions of the genome bound by specific Transcription Factors, Histone Modifications or RNA Polymerase.
- Biological function of non-coding parts of the genome i.e., enhancers, silencers, insulators.
ChIP-Seq

• Transcription Factors
  – Where are they?
  – Sequence preference
  – Correlation with gene expression

• Chromatin Marks
  – Where are they?
  – What combinations do they come in
  – How do they relate to biological status
Public Data Repositories

- **Roadmap Epigenomics**
  - Human focused
  - Large number of histone marks, DNA methylation, DNase hypersensitivity across many many cell types

- **ENCODE**
  - Human and Mouse (some)
  - Started with ChIP-chip data has moved into ChIP-seq and other enrichment based sequencing
  - Came before Roadmap

- **modENODE**
  - ENCODE but for Fruit Fly (7 species) and Worm (4 species)
  - Same goals as ENCODE different organisms
Basic ChIP-Seq Workflow

- Isolation
- Sequencing
- Mapping
- Peak Finding

Target List
Motif Analysis
Overlap w/ Other Data
Isolation

• Results are very dependent on “wet-lab” protocols
  • Use a validated antibody
  • Validate enrichment using PCR before prepping for sequencing
• Input samples are needed for each cell type, antibody and sonication.
• Antibody list from ENCODE: https://www.encodeproject.org/search/?type=antibody_lot
Basic ChIP-Seq Workflow

- Isolation
- Sequencing
- Mapping
- Peak Finding

Target List
Motif Analysis
Overlap w/ Other Data
Sequencing Recommendations

- Large Genomes (human, plant)
  - TF: 10 million uniquely mapped reads
  - Chromatin: 20 million uniquely mapped reads

- Small Genomes (worm, fly, yeast)
  - TF: 2 million uniquely mapped reads
  - Chromatin: 5 million uniquely mapped reads

- At least Two biological replicates are recommended
Sequencing Recommendations

• Enrichment regions will be at least as large as the fragments of DNA sequenced. (Sonication)

• Longer reads do not give you better information. 100bp max is recommended.

• Paired end data is not required and some algorithms cannot use paired end data.
Experimental Design

Cell State 1

- DNA Input
- Chromatin IP

Cell State 2

- DNA Input
- Chromatin IP
Basic ChIP-Seq Workflow

Isolation → Sequencing → Mapping → Peak Finding

→ Target List → Motif Analysis → Overlap w/ Other Data
Mapping Your Reads

• Map biological replicates independently using standard mapping tools i.e., BWA Bowtie.
• Remove PCR duplicates, these are 100% identical reads. SAMtools rmdup
• Depending on the peak finding software used you may need to convert from SAM/BAM to BED format (BedTools)
• Check your mapping stats, if < 80% of your reads map then there may be an issue.
Basic ChIP-Seq Workflow

- Isolation
- Sequencing
- Mapping
- Peak Finding

Target List
Motif Analysis
Overlap w/ Other Data

© 2014 Regents of the University of Minnesota. All rights reserved.
Peak Calling

- Peak callers are algorithms that identify regions that have more reads than background.
- Three types: point-source, broad peak callers and those that can do both.
- MACS, MACS2, ZINBA, SICER, CCAT

<table>
<thead>
<tr>
<th>chr1</th>
<th>3504555</th>
<th>3504771</th>
<th>MACS_peak_1</th>
<th>51.97</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>3505222</td>
<td>3505607</td>
<td>MACS_peak_2</td>
<td>92.84</td>
</tr>
<tr>
<td>chr1</td>
<td>3506002</td>
<td>3506130</td>
<td>MACS_peak_3</td>
<td>53.32</td>
</tr>
<tr>
<td>chr1</td>
<td>4485926</td>
<td>4486384</td>
<td>MACS_peak_4</td>
<td>91.62</td>
</tr>
<tr>
<td>chr1</td>
<td>4613288</td>
<td>4613545</td>
<td>MACS_peak_5</td>
<td>73.09</td>
</tr>
<tr>
<td>chr1</td>
<td>4758830</td>
<td>4759138</td>
<td>MACS_peak_6</td>
<td>74.85</td>
</tr>
<tr>
<td>chr1</td>
<td>4759374</td>
<td>4760160</td>
<td>MACS_peak_7</td>
<td>163.21</td>
</tr>
<tr>
<td>chr1</td>
<td>4773712</td>
<td>4774711</td>
<td>MACS_peak_8</td>
<td>283.47</td>
</tr>
</tbody>
</table>
Peak Data

• WIG family- continuous data
  – WIG, BigWIG, bedGraph

• BED family – discrete locations
  – Bed, Bed6, Bed12

• Your data can be visualized in genome browsers such as IGV, UCSC and Ensembl
Which Peaks?

- Depending on the software peaks can be evaluated by FDR, reads in the peak, enrichment ratio.
- It is common to get 10s of thousands of peaks in a sample.
- The more reads you have the more peaks you will get.
Which Peaks?

• What do positive control regions look like? P-value, enrichment, score, rank

• Which peaks are found in all of the biological replicates?

• Top 10, 20, 30% of peaks?
Visualize the Data

- What genomic context do your peaks reside in?
- Do they co-localize with other peaks? Differentially expressed genes? Know TFBS?
- What changes between samples, states, time?

Ren and colleagues

© 2014 Regents of the University of Minnesota. All rights reserved.
Visualize the Data

- Most visualizations are custom (R, MATLAB)
- CEAS, venn Diagram and heatmap tool in Galaxy
Next Steps

• Motif Identification
  – Use only highest confidence peaks
  – Multiple tools and methods, MEME is very popular

• Differential peak calling
  – Still a new problem but tools do exist
  – Samples need to be treated uniformly

• More Advanced Integrative analysis
  – Machine learning techniques can identify patterns that exist in large data sets
  – ChromHMM
Links for more Info

- MACS: https://github.com/taoliu/MACS/blob/macs_v1/README.rst
- MACS2: https://github.com/taoliu/MACS/
- SCIER: http://home.gwu.edu/~wpeng/Software.htm
- ZINBA: https://code.google.com/p/zinba/
- chromHMM: http://compbio.mit.edu/ChromHMM/
- MEME: http://meme.nbcr.net/meme/doc/meme-chip.html
- ENCODE guidelines paper: http://genome.cshlp.org/content/22/9/1813.full?sid=bf1bf8f4-9ecf-4dd6-9761-78af1e542311
- Roadmap Epigenomics: http://www.roadmapepigenomics.org/
- ENCODE: https://www.encodeproject.org/
- modENCODE: http://www.modencode.org/
- UCSC epigenome browser: http://www.epigenomebrowser.org/