What is ChIP-Seq?

- ChIP – Seq = Chromatin Immunoprecipitation Sequencing
- The sequencing of genomic DNA fragments that co-precipitate with a DNA-binding protein
- ‘Unbiased’ – doesn’t rely prior knowledge of precise DNA binding sites (like ChIP-ChIP)
- Results
  - The regulatory sites for any transcription factor
  - Direct downstream targets of any transcription factor
  - The DNA sequence motif recognized by the binding protein
Transcription Factors

- DNA binding proteins that attach themselves to the genome with an affinity for a specific DNA sequence
- Function: Bind to specific sites in the genome, recruit cofactors, and regulate transcription
- ChIP-seq – identify binding transcription factor binding sites across entire genomes

Summary of ChIP Seq Assay

1. Collect and fractionate DNA

2-3. Enrich binding sites using IP

3b. PCR (Not Shown)

4. Sequence short reads

5. Align reads to reference genome

Photos: U.S. Department of Energy Genome Programs
ChIP-seq considerations

- Peaks are detected on a per sample basis
- Control samples are not required, but encouraged
- # reads for each sample don’t have to match
- Antibody selection
  - Must have specificity for the protein
  - Must be able to immunoprecipitate with target protein; even if they do, they may not do well with ChIP-seq
- Sequencing – platform dependent bias, error rates
- Algorithm – short reads ambiguous in repeat regions, account for sequence errors
Sample Data Set

Study mapped the genomic binding sites of the NRSF transcription factor across the entire genome.

Two samples: NRSF-enriched ChIP sample (chip.txt) and control sample (mock.txt) DNA immunoprecipitated by a non-specific control antibody.


Imported ChIP-Seq Data

<table>
<thead>
<tr>
<th>Current Selection Chip.txt</th>
<th>Sample</th>
<th>Number of Reads</th>
<th>Number of Alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>chip.txt</td>
<td>1</td>
<td>168709</td>
<td>168709</td>
</tr>
<tr>
<td>mock.txt</td>
<td>2</td>
<td>2319153</td>
<td>2319153</td>
</tr>
</tbody>
</table>

Sequence Input Wizard
Please describe the data you wish to import:

- Apply Bedtools script ([EDGAR](http://edgar.ibisc.usq.fr)):
  - Input common data in a SOUC Bed file
  - Input common data in a DRIP Bed file
  - Input common data in a DRIP Bed file

- Apply BED file:
  - Input BED file

- Apply BAM file:
  - Input BAM file

- Input BED file:
  - Input BED file

- Input BED file:
  - Input BED file

- Input BED file:
  - Input BED file

Current Selection Chip.txt

<table>
<thead>
<tr>
<th>Current Selection Chip.txt</th>
<th>Sample</th>
<th>Number of Reads</th>
<th>Number of Alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>chip.txt</td>
<td>1</td>
<td>168709</td>
<td>168709</td>
</tr>
<tr>
<td>mock.txt</td>
<td>2</td>
<td>2319153</td>
<td>2319153</td>
</tr>
</tbody>
</table>
QA/QC--Fragment Length Analysis

- Single end reads – phase shift between the forward and reverse reads
  - Maximum
  - Only on IP samples
- Paired-end reads – distribution of fragment lengths between paired end fragments

Peak Detection

1) Extend Reads by Estimated frag length(small)
2) Calculate Midpoints
3) Divide Genome into windows of estimated fragment length or 100bp
4) Count number of midpoints in each window
5) Sliding window to merge overlapping windows
6) Fit to ZTNB; peak cutoff determined based on FDR
Detect Peaks Results

Peaks are detected in each sample separately reported one peak at a time.
Create a List of Enriched Regions

- Regions of DNA which have many reads mapped to them
- They will occur only in our protein bound sample
Detecting motifs—Discover de novo motifs

Height = binding importance; how well a base is preserved

Gibbs Motif Sampler

Search for instances of Motif in Sequences
Create new Motif out of discovered instances
### De Novo Motif Instances

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Start</th>
<th>End</th>
<th>Strand</th>
<th>Motif ID</th>
<th>Instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1675836</td>
<td>1675846</td>
<td>-</td>
<td>motif1</td>
<td>CCGACTCTCCGGTCCGTCCTGAC</td>
</tr>
<tr>
<td>10</td>
<td>2264959</td>
<td>2264969</td>
<td>-</td>
<td>motif1</td>
<td>GTATCGGCTCCAGTGTACACATAC</td>
</tr>
<tr>
<td>10</td>
<td>4127863</td>
<td>4127873</td>
<td>+</td>
<td>motif1</td>
<td>GGAGCTGCTCCATGTCTGCCTGCT</td>
</tr>
<tr>
<td>10</td>
<td>6795273</td>
<td>6795283</td>
<td>+</td>
<td>motif1</td>
<td>AGGCTCTTCGCCGTTGTCCTGCC</td>
</tr>
<tr>
<td>10</td>
<td>6795281</td>
<td>6795291</td>
<td>-</td>
<td>motif1</td>
<td>GGAGCTGCTCCAGTGTACACATAC</td>
</tr>
<tr>
<td>10</td>
<td>13306684</td>
<td>13306694</td>
<td>+</td>
<td>motif1</td>
<td>GGCCTCCTCCCTGCCTGTCCTGAAA</td>
</tr>
<tr>
<td>10</td>
<td>14642422</td>
<td>14642432</td>
<td>+</td>
<td>motif1</td>
<td>CTAGCTTTCAAGCTATGGGCCTG</td>
</tr>
<tr>
<td>10</td>
<td>16694446</td>
<td>16694456</td>
<td>+</td>
<td>motif1</td>
<td>TGCTCTTCCTCGTGCTGTCCTGAAA</td>
</tr>
</tbody>
</table>

### Detect Known Motifs

#### IUPAC Nucleotide Code

<table>
<thead>
<tr>
<th>Base</th>
<th>IUPAC Nucleotide Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A or G</td>
</tr>
<tr>
<td>C</td>
<td>C or T</td>
</tr>
<tr>
<td>G</td>
<td>G or C</td>
</tr>
<tr>
<td>T (or U)</td>
<td>Thymine (or Uracil)</td>
</tr>
<tr>
<td>R</td>
<td>A or G</td>
</tr>
<tr>
<td>Y</td>
<td>C or T</td>
</tr>
<tr>
<td>S</td>
<td>G or C</td>
</tr>
<tr>
<td>W</td>
<td>A or T</td>
</tr>
<tr>
<td>K</td>
<td>G or T</td>
</tr>
</tbody>
</table>

#### Search for Motif(s) in Sequences

- **Choose Motifs to Search:**
  - Advanced: Remove custom motifs

- **Search for motif(s):**
  - Choose motifs to search
  - Search for motif(s)

- **Sequence Quality Threshold:**
  - Sequence Quality (default: 0.7)
Detecting Motifs -- Find Known Motif

Overlap with Databases

- Databases of ChIP binding available from UCSC such as Oreganno
- Databases of known genes such as RefFlat
  - Genes which overlap with peaks, nearby to motif instance
Find Overlapping Genes

(PAZAR-hg19) soon - public database of regulatory sequences and transcription factors

Overlapping Genes
DNA Methylation

- An important epigenetic modification involved in control of gene expression (usually repression)
- The addition of a methyl group to cytosine (C)
- 1.5 out of 100 bases are methylated
- Linked to many diseases including cancer
Me-DIP Seq

SONICATION

DENATURE

5mC ANTIBODIES

IMMUNOPRECIPITATE

- Bisulfite Sequencing
- Methylminer(AB)

HT SEQUENCING

M -Methylated

Me-DIP Seq workflow:

- Import & Quality Control
- Detect Methylated Regions – peak finding, Differential Methylation
- Create List of methylated regions
- Find overlapping genes
- Classify regions by gene section
- Measure distance of methylation to gene transcription start site

NGS Methylation Workflow
Methylated Overview – Hilbert Curve

Densely Methylated

Centromere – no reads

End of data – padded with zeros

Peak Finding

• Similar to ChIP-Seq (peaks not as sharp)
• Peaks reported on a per sample basis
• Filter by Peaks in Case but not in Control (Input)

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Start</th>
<th>End</th>
<th>Sample 1 ID</th>
<th>Sample 2 ID</th>
<th>Start in Sample 1</th>
<th>End in Sample 1</th>
<th>Start in Sample 2</th>
<th>End in Sample 2</th>
<th>Mean M in Sample 1</th>
<th>Mean M in Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>200</td>
<td>Sample1</td>
<td>Sample2</td>
<td>100</td>
<td>200</td>
<td>100</td>
<td>200</td>
<td>1.234</td>
<td>1.567</td>
</tr>
</tbody>
</table>

Higher M in Sample1 vs. Sample2

Higher M in Sample2 vs. Sample1
Find Overlapping Genes

(PAZAR-hg19) soon -public database of regulatory sequences and transcription factors

Classify Regions by Gene Section
Methylated Region with Multiple Peaks

Detected Peaks in Chromosome Browser
Distance of Methylation to TSS

A few examples

Integrated Genomics
RNA-seq data and Exon array data

Integration of ChIP-seq & RNA-Seq data
What Else?

Histone Modification
SOM Fingerprinting – Look for Clustering Genes

Gene Patterns by Category – Profile Trellis
Biological Interpretation:
Pathway Analysis (coming soon)

Partek® Genomics Suite™

Your Start-to-Finish Solution for Analysis of Next Generation Sequencing Data

Sequencing  Alignment  Statistical Analysis  Visualization  Biological Interpretation  Publication

✓ RNA-Seq
✓ sRNA-Seq
✓ ChIP-Seq
✓ DNA-Seq
✓ MeDIP-Seq

Get your FREE trial today
www.partek.com