Processing, integrating and analysing chromatin immunoprecipitation followed by sequencing (ChIP-seq) data
The regulatory system

Dynamic, sequential gene expression changes required for normal development

- Chromatin accessibility
- Histone modifications
- Chromatin state
- Transcription factor binding
- Sequence features
- Cis regulatory modules
- Distal or proximal interactions
- Gene expression
High throughput sequencing (HTS) data

- Chromatin immunoprecipitation (ChIP-seq)
  - To determine protein binding sites *in vivo*
  - Transcription factor binding & histone modifications
- DNase I hypersensitivity (DHS) (also ATAC-seq and FAIRE-seq)
  - Chromatin accessibility
  - Protein binding footprints
High throughput sequencing (HTS) data

- **RNA-seq** and cap analysis gene expression (CAGE)
- Gene expression
- Alternate transcription start site usage
- Changes in expression (temporal or perturbation)
High throughput sequencing (HTS) data

- Chromosome conformation capture
- DNA looping
- Long distance enhancer/promoter interactions
- 3C, 4C, 5C, Hi-C, CHi-C, ChIA-PET…
ChIP-seq

GENERATING AND INTERPRETING CHIP-SEQ DATA
ChIP-seq experiment

Wet lab

- Extract DNA fragments bound by protein of interest
ChIP-seq sequence analysis

Sequencing & quality control

- **Sequence depth**
  - Depends on size of genome and type of protein
    - **Mammalian transcription factor**
      → 20 million reads

- **Sequence quality**
  - **Read quality** summary
  - Unusual base pair patterns
  - Adapter sequences
  - E.g. FastQC tool
ChIP-seq alignment

Alignment & quality control

- Alignment summary
  - Generated by alignment tools
  - Uniquely aligned reads
ChIP-seq alignment

Alignment & quality control

• Alignment summary
  – Generated by alignment tools
  – Uniquely aligned reads

• Read distribution quality
  – ChIP-seq peaks create **bimodal pattern**
  – Strand cross correlation analysis (SCCA)
Peak calling

GENERATING AND EXPLORING CHIP-SEQ PEAKS
Basic principles of peak calling

Sample
Exposed to antibody

Input
No antibody exposure

Peak
With statistical significance

Compared to
To generate
Peak calling tools

- Everyone seems to have built their own!
- Omic Tools reports **86 ChIP-seq tools**
- 51 in 2016
- **In-house** tools
- Find or adapt a tool that fits your needs
- Potentially, develop your own

Be aware!
Not all tools are well documented or tested
So, how do I choose a tool?

- Choice of tool depends on problem
- Based on experiment itself
- Based on sequence data and read distribution
Combining tools and replicates

- Use **multiple** peak callers
- Combine outcomes (e.g. common peaks)
- Take advantage of strengths of different peak callers
- Bolster weaknesses

- Biological replicates can vary **significantly**
  - Call peaks for replicates individually
  - Compare/overlap to achieve ‘golden standard’
- Comparisons are dominated by poor replicate

<table>
<thead>
<tr>
<th>Peak caller</th>
<th>Total</th>
<th>Unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>MACS2</td>
<td>42,536</td>
<td>12%</td>
</tr>
<tr>
<td>HOMER</td>
<td>45,044</td>
<td>19%</td>
</tr>
<tr>
<td>SPP</td>
<td>19,474</td>
<td>0.7%</td>
</tr>
</tbody>
</table>
Peak quality control

- Number of peaks
- Motif enrichment
- Read coverage
  - Fraction of reads in peaks (FRiP) > 1%
    - Generally observe > 10%
    - E.g. below 6/50 reads in peak -> 12%

CTCF and DNA
5T0U (PDB)
Hashimoto et al. 2017
ChIP-seq complications

- ChIP-seq generates peaks for all of these events
Integrating data

COMBINING DATA SETS TO IMPROVE OUTCOMES
Data integration

- Experiments capture **dependent** regulatory events
  - ChIP-seq – regulatory elements
  - DHS – chromatin accessibility
  - RNA-seq – expression patterns

- Consider multiple datasets to:
  - Improve confidence
  - Improve understanding
  - Support hypotheses
Supporting histone modifications

- Explore **chromatin environment**
- Layer/overlap histone modifications
- **DHS** – chromatin accessibility
Supporting transcription factors

- Transcription factors preferentially bind **open/active chromatin and regulatory regions**
- Alter expression of genes
- RNA-seq on knock-out of transcription factor
  - Identify genes with significant change in expression
System complexity

- Small number of **differentially expressed genes** are bound by target transcription factor
- **System redundancy**
- **Indirect changes in expression**

Ma et al. 2014
More complicated relationships

- Overlapping is simple
- Next steps: pattern identification, prediction, classification
- **Machine learning approaches**
  - Hidden Markov model to classify epigenetic state
  - Bayesian network to predict transcription factor binding events
  - Random forest of decision trees to predict long distance enhancer/promoter interactions
Take home messages

- Understand your data and how best to use it
- Quality control!
- Peak calling
  - Use multiple tools where possible
- Keep up to date with advances
- Data integration
  - Combine resources and data to gain a more complete picture
Resources

- **Data/figures**

- **Useful papers**
ChIP-seq complications

- Possible to observe **multiple states** at one genomic location
- **False negatives**
  - Can’t detect small sub-populations
- **False positives**
  - General non-specific chromatin being pulled down
  - Bias not removed by input
- Replicates can resolve variation
Transcription Factors

- Confirm *in vitro* and *in silico* results
  - Overlapping peaks with motifs

- **Identify consensus motif**
  - For transcription factors which do not have an existing/known motif
  - To identify variations in motif

- **Differential peak binding**
  - To identify differences in binding patterns
  - Compare cell types or time points
Histone Modifications

- **Epigenetic analysis**
  - Generate *epigenetic profiles*
  - Identify *chromatin states* genome wide
    - E.g. ChromHMM
  - Identify *regulatory modules*
    - E.g. promoters or enhancers

- **Differential peak binding**
  - Identify differences in epigenetic patterns
Long distance regulation

- Chromosome conformation capture reports different types of interactions
- Histone modifications can identify enhancer-promoter interactions
- Filter out structural interactions