ChIP-Seq Analysis

Simon Andrews
simon.andrews@babraham.ac.uk
@simon_andrews
v2023-05
What this course covers

- The theory of ChIP-Seq
- ChIP-Seq library properties
- Sequencing, Data processing and QC
- Data visualisation and exploration
- Types of analysis
  - Peak Calling
  - Differential Binding
What is ChIP-Seq?

ChIP-Seq is a technology which uses high-throughput sequencing to infer the positions of any mark associated with DNA which can be captured by an antibody.
Types of antibody

• Transcription factors / repressors
  – nanog, CTCF

• Histones and histone modifications
  – H3, H3K4me3

• DNA modifications
  – Methyl-Cytosine, Formyl cytosine

• Chromatin remodelling proteins
  – BMI1, EZH2

• Transcription machinery
  – Pol2
How Does ChIP-Seq work

1. Cross-link proteins to DNA
2. Fragment DNA (sonication, MNase etc)
3. Capture
4. Elute
5. Remove Crosslinks
6. Size Select
7. Add Adapters
8. Sequence
Related Techniques

- **ATAC-Seq**
  - Uses transposases to digest exposed DNA to enrich for accessible DNA.

- **Cut and Tag**
  - Uses transposases fused to antibodies to find marked, accessible chromatin
What can you sequence?

ChIP Material

Mock (IgG)ChIP Material

Input Material (sonicated / Mnase / H3 etc)
Sequencing for ChIP

Barcode Adapter ChIP Fragment Adapter Barcode

Primer Read 1

Barcode Adapter ChIP Fragment

Primer Read 2

Barcode Adapter ChIP Fragment

Barcode Read
Single End vs Paired End

Paired
(= >$£€¥)

Single
(= Cheaper!)

Insert

Read 1

Inferred

Read 2

Inferred

Read 1
What you end up with

Original
40bp
Reads

Extended
by
250bp
Types of Enrichment

• Single points (typical TF, some histone marks)
Types of Enrichment

- Broad Regions (some histone marks, PolII)
Types of Enrichment

• Virtually everywhere (h3)
Types of Enrichment

- Artefactual (GC in this case)
What are you actually measuring?

• ChIP Seq measures **RELATIVE** enrichment
  – Region A has twice as much signal as Region B

• Without some external calibration, **NOTHING** in ChIP-Seq gives an **ABSOLUTE** measure.
What can affect enrichment?

Starting

More Sites

Poorer Signal
What sort of questions can you answer?

• Where is this mark present?
  – General - it's in promoters, gene bodies etc.
  – Specific - it's at these loci

• How does this mark change when I do XXX?
  – Categorical: A peak disappears
  – Quantitative: The enrichment of a locus changes
ChIP-Seq Data Processing and QC
A typical ChIP Library

- Potential technical problems
  - Adapter contamination
  - PCR Duplication

- Potential biological problems
  - Lack of enrichment
  - Other selection biases
QC of raw sequence
Base Call Quality
QC of raw sequence
Sequence Composition
QC of raw sequence
Sequence Composition

⚠️ Overrepresented sequences

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Count</th>
<th>Percentage</th>
<th>Possible Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGTAGGGTTAGGTTAGGTTAGGTTAGGTTAGGTTAGGTTAGG</td>
<td>11933</td>
<td>0.1717855635921134</td>
<td>No Hit</td>
</tr>
<tr>
<td>CCTAACTACCTAACCTAACCTAACCTAACCTAACCTAACCTAC</td>
<td>8917</td>
<td>0.1283677089207136</td>
<td>No Hit</td>
</tr>
<tr>
<td>CCTAACTACCTAACCTAACCTAACCTAACCTAACCTAACCTAC</td>
<td>7298</td>
<td>0.1050608432996936</td>
<td>No Hit</td>
</tr>
<tr>
<td>CTAACCTACCTAACCTAACCTAACCTAACCTAACCTAACCTAC</td>
<td>7282</td>
<td>0.104630509653161</td>
<td>No Hit</td>
</tr>
<tr>
<td>GTTAGGGTTAGGTTAGGTTAGGTTAGGTTAGGTTAGGTTAGG</td>
<td>7040</td>
<td>0.10134671647135503</td>
<td>No Hit</td>
</tr>
</tbody>
</table>

A telomere is a region of repetitive nucleotide sequences at each end of a chromosome, which protects the end of the chromosome from deterioration or from fusion with neighboring chromosomes. ... This sequence of TTAGGG is repeated approximately 2,500 times in humans.

Telomere - Wikipedia
https://en.wikipedia.org/wiki/Telomere
QC of raw sequence
Adapter Contamination

Trim Galore!
Quality and Adapter Trimming
Mapping ChIP Data

• All regions should be linear genomic stretches

• Standard genomic aligners are fine
  – Bowtie2  http://bowtie-bio.sourceforge.net/bowtie2/
  – BWA       http://bio-bwa.sourceforge.net/
Example Bowtie2 Mapping

• Create Genome Index (once - slow!)

```
bowtie2-build yeast_genome.fa yeast_index
```

• Map a single FastQ file

```
bowtie2 \
-x yeast_index  \
-U data.fastq.gz  \
| samtools view \
-bS \
-o data.bam
```
Post Alignment QC
Mapping Statistics

41523294 reads; of these:
41523294 (100.00%) were unpaired; of these:
1851792 (4.46%) aligned 0 times
32175322 (77.49%) aligned exactly 1 time
7496180 (18.05%) aligned >1 times
95.54% overall alignment rate
Post Alignment Processing
MAPQ Filtering

- ChIP-Seq relates sequences to positions in a reference genome
- You need to be confident that the reported position is correct
- Filtering on MAPQ value (likelihood of reported position being incorrect) is an easy way to do this
- MAPQ filtering should be performed in most cases

`samtools view -q 20 -b -o filtered.bam data.bam`
Standard Processing Workflow

- FastQ File
  - Trimmed FQ File
    - BAM File
      - Filtered BAM
        - Visualisation and Assessment
      - MultiQC Report
    - SAM Tools
      - Trimmed FQ File
    - Bowtie
      - Trimmed FQ File
      - Trimmed FQ File
  - FastQC
    - FastQC Report
      - MultiQC Report
    - Trim Galore

- MultiQC Report
  - Mapping Stats
Data Processing Exercise
Running programs in Linux

• Open a shell (text based OS interface)

• Type the name of the program you want to run
  – Add on any options the program needs
  – Press return - the program will run
  – When the program ends control will return to the shell

• Run the next program!
Running programs

babraham@babraham-VirtualBox:~$ ls
Desktop  Documents  Downloads  examples.desktop
Music  Pictures  Public  Templates  Videos

babraham@babraham-VirtualBox:~$

- Command prompt - you can't enter a command unless you can see this
- The command we're going to run (ls in this case, to list files)
- The output of the command - just text in this case
The structure of a unix command

```
lst -ltd --reverse Downloads/ Desktop/ Documents/
```

- **Program name**
- **Switches**
- **Data** (normally files)

Each option or section is separated by spaces. Options or files with spaces in must be put in quotes.
Command line switches

• Change the behaviour of the program
• Come in two flavours (each option often has both types available)
  – Minus plus single letter (eg \(-x\) \(-c\) \(-z\))

  – Two minuses plus a word (eg \(--extract\) \(--gzip\))

• Some take an additional value
  \(-f\) somfile.txt (specify a filename)
  \(--width=30\) (specify a value)
Specifying file paths

• Specify names from whichever directory you are currently in
  – If I'm in /home/simon
  – Data/big_data.fq.gz
  • is the same as /home/simon/Data/big_data.fq.gz

• Move to the directory with the data and just use file names
  – cd Data
  – big_data.fq.gz
Command line completion

• Most errors in commands are typing errors in either program names or file paths

• Shells (ie BASH) can help with this by offering to complete path names for you

• Command line completion is achieved by typing a partial path and then pressing the TAB key (to the left of Q)
Command line completion

List of files / folders:

- Desktop
- Documents
- Downloads
- Music
- Public
- Published
- Templates
- Videos

T [TAB] ➔ Templates
P [TAB] ➔ Public
Do [TAB] ➔ [beep]
Do [TAB] [TAB] ➔ Documents Downloads
Doc [TAB] ➔ Documents

You should ALWAYS use TAB completion to fill in paths for locations which exist so you can't make typing mistakes
(it obviously won't work for output files though)
Debugging Tips

• Be wary of anything which finishes suspiciously quickly!

• Look for errors before asking for help. They will either be
  – The last piece of text before the program exited
  – The first piece of text produced after it started (followed by the help file)

• Programs which are stuck can be cancelled with Control+C
**Some useful commands**

`cd mydir`  
Change directory to `mydir`

`ls -ltrxh`  
List files in the current directory, show details and put the newest files at the bottom

`less x.txt`  
View the `x.txt` text file
  
  Return = down one line  
  Space = down one page  
  q = quit
Data Processing Exercise
Exploring and Understanding ChIP-Seq data
Some Basic Questions

• Is there any enrichment?
  – What is the size / patterning of enrichment?

• How well are my controls behaving?

• What is the best way to quantitate this data?

• Are there any technical artefacts?
Start with a visual inspection

- Is there any enrichment?
- What is the size / patterning of enrichment?
- How well are my controls behaving?
Start with a visual inspection

- Is there any enrichment?
- What is the size / patterning of enrichment?
- How well are my controls behaving?
Extending reads if necessary

For point enrichment, insert size is roughly peak width/2
Examine Controls

• IgG or other Mock IP
  – Good result is no material at all
  – Not worth sequencing. Reads are only informative if the ChIP hasn't worked.
  – May be justified for Cut and Run where there is no real input

• Input material (sonicated / Mnase etc)
  – Genomic library - everywhere equally
  – Technical issues can cause variation
Examine Controls

- Does the coverage look even
- If there are multiple inputs to do they look similar
Why do controls misbehave?

• Low coverage
  – Repetitive unmappable regions
  – Holes in the assembly

• High coverage
  – Mismapped reads from outside the assembly

• Biases
  – GC content
  – Segmental Duplication

  Blacklist these regions and remove them from the analysis (ignore hits within these regions)

  Input normalisation might help, but requires further examination
Making Blacklists

- Look at distribution of Input counts
  - Set limits on unusually high/low values
  - Remove regions outside those limits
Comparison of samples
Initial Quantitation

• Always start with a simple unbiased quantitation (not focussed on features/peaks)

• Tiled measures over the whole genome
  – Use approximate insert size as window size
  – Something around 500bp is normally sensible

• Linear read count quantitation corrected for total library size
Compare samples
Visual comparison against raw data

- Similar apparent overall enrichment
- Any obvious differences?
Compare samples
Scatterplot input vs ChIP

Raw

Filtered
Compare samples
Scatterplot input vs input

- Any suggestion of differential biases in inputs
- Can we merge them to use as a common input

Extensive genomic copy number variation in embryonic stem cells
Qi Liang, Nathalie Conte, William C. Skarnes, and Allan Bradley
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

Of 26 clones that could not contribute to the mouse germ line, trisomies were detected in 7 which involved chromosomes 1, 6, 8, and 11. In 5 cases, loss of the Y chromosome was detected.
Compare samples
Scatterplot ChIP vs ChIP

Look at examples for different parts of the plot

• Look for outgroups (differentially enriched)
• Compare level of enrichment (compare to diagonal)
Compare samples
Summarise distributions

• QQPlot
  – Percentile though measures(x) vs Percentile through total quantitation (y)
  – Perfect input is on the diagonal
  – More enrichment moves the curve down and right
  – How flat is your input? How consistent are the ChIPs?
Compare samples
Higher level clustering

Correlation Matrix

Correlation Tree

PCA Plot

tSNE Plot
Associate enrichment with features
Trend Plots

• Graphical way to look at overall enrichment relative to positions in features
  – Gene bodies
  – Promoters
  – CpG islands

• May influence how we later quantitate and analyse the data
  – Analyse per feature
  – Look for exceptions to the general rule
• Overall average
• Says nothing about the number / proportion of features affected
Check trend plot results against data
Aligned Probes Plots give more detail

• Information per feature instance
• Comparison of equivalent features in different marks/samples
After exploration you should...

• Know whether your ChIP is really enriched
• Know the nature / shape of the enrichment
• Know whether your controls behave well
• Know whether you're likely to have differential enrichment
• Know if you will need additional normalisation
• Know the best strategy to measure your data
Data Exploration Exercise
Analysing ChIP-Seq Data
Steps in Analysis

• Define enriched regions
  – Based around features
  – De-novo peak prediction

• Quantitate
  – Corrections and Normalisation

• Compare
  – Categorical
  – Quantitative
Defining Regions - Should I peak call?

• Choices
  – Make measurements around features (promoters / genes / CpG islands etc)
  – Make measurements around enriched regions (peaks)

• Can I use features?
  – Do you see a strong and complete linkage between enrichment and the type of feature you want to use?
  – If not, then you should peak call
How MACS Works

1. Optimise the starting data
2. Build a background model
3. Test sliding windows
4. Apply per-site adjustment
5. Report

macs2 callpeak --broad -t chip.bam -c input.bam
Optimise the starting data

- Correct the for/rev offset
- Deduplicate
Build a background model

Lambda value

Observed
Build a background model

Lambda value

Critical p-value (n=18)

Model
Build a background model

Lambda value

Critical p-value (n=18)

Observed + Model
Test Sliding Windows

• Generally use half of the library fragment size

• Windows whose count exceeds the critical value are kept

• Merge adjacent windows over the critical value to form peaks

• Generates candidate (not final) peak set
Correct for local variation

Generate localised model if input density is higher than the global value
Most pessimistic p-value is kept
Broad Peaks

- Added in MACS2 – suitable where larger regions with variable enrichment exist
- Uses two thresholds for enrichment
How should you apply peak callers

• Multiple ChIPs (over multiple conditions)
• Multiple Inputs
Input variability is generally consistent

– Mapability
– Genome Assembly
– Fragmentation biases

Unless you see substantial variability between inputs it’s better to combine them into a single reference input sample
Multiple ChIPs

BAM Files

- WT ChIP 1
- WT ChIP 2
- KO ChIP 1
- KO ChIP 2

Peak Sets

- Peaks
  - WT ChIP 1
  - WT ChIP 2
  - KO ChIP 1
  - KO ChIP 2

WT ChIP 1 + WT ChIP 2 + KO ChIP 1 + KO ChIP 2
Multiple ChIPs

BAM Files

- WT ChIP 1
- WT ChIP 2
- KO ChIP 1
- KO ChIP 2

Peak Sets

- WT Peaks 1
- WT Peaks 2
- KO Peaks 1
- KO Peaks 2

WT Peaks 1 And WT Peaks 2
KO Peaks 1 And KO Peaks 2

WT Peaks 1 And WT Peaks 2
Or
KO Peaks 1 And KO Peaks 2
Why isn't a peak called

Fewer peaks are called by just sub-sampling the same data
Why isn't a peak called

With no input the region around the peak is used to model the background. Broader peaks can be missed.

For ATAC data (no input) you should skip the rescoring step (--nomodel)
Reporting on Peak sets

• Don’t make claims based solely on the number of peaks (“there were more WT peaks than KO peaks” for example)

• Don’t make claims based on regions being peaks in 1 set but not another (there were 465 peaks which were specific to KO)

• It is OK to make statements about overlap (there were 794 peaks which were common to WT and KO)

• You have to address differential enrichment problems quantitatively
Quantitating ChIP data for analysis

• Quantitation of ChIP is not a simple problem

• Can start with something simple but in many cases you will need to refine this

• Globally corrected log counts are a good place to start
Should I normalise to input?

• Only consider input normalisation if:
  1. You have substantial variation in the coverage of your input (excluding outliers)
  2. Your ChIP signal is correlated with the input level
Why not just always do "fold over input"?

• Inputs are generally poorly measured
  – Poor coverage compared to ChIP

• Fold change values are more influenced by input than ChIP

• Biases in the input are smaller than enrichment power of the antibody
Hits with increased enrichment

Hits with decreased enrichment
Evaluating and Normalising Enrichment

- **Good Enrichment**
- **Worse Enrichment**
- **Similar Enrichment**
- **Small Difference**
- **Large Difference**
Evaluating and Normalising Enrichment

![Diagram showing normalised read count against percentile through data]

- Normalised Read Count
- Percentile through data

Legend:
- run1_Control
- run2_Hordei_Ctrl
- run3_Choice1
- run4_Hordei_i01
- run5_Hordei_i02
- run6_Hordei_i03
- run7_Hordei_i04
- run8_Hordei_Ctrl
Normalising Enrichment

Size Factor
Single point of comparison
Works well for small differences
Insufficient for large differences
Allows the use of count based stats

Enrichment
Two points of comparison
Corrects for larger differences
Not directly compatible with count based stats

Quantile
Forces distributions to be identical
Corrects any differences, easy to apply
Normalising Enrichment
Checking Normalisation

Before Normalisation

After Normalisation
Look for systematic enrichment changes (real biology!!)

Use replicates to build a case for a biological rather than technical difference
Differential enrichment analysis

• Needs to be quantitative

• Needs to operate on non-deduplicated data

• Two statistical options
  – Count based stats on raw uncorrected counts
    • DESeq
    • EdgeR
  – Continuous quantitation stats on normalised enrichment values
    • LIMMA
Which statistic to pick?

• If enrichment is roughly similar
  – Raw counts, then DESeq/EdgeR

• If there are large differences in enrichment
  – Enrichment normalisation
  – LIMMA statistics
Visualisation of hits

• Map onto scatterplot for simple verification

• Normally makes sense to use log transformed counts

• Look at the data underneath candidates you make specific claims about
Hit validation

- Look whether hits make sense
- Look at points which change but were not selected
- Log scale should be used
- Keep the context of non-hits
Hit validation
Directionality

• Most ChIP enrichments are not strand-specific
• Should expect to see enrichment on both strands
Hit validation

Heatmap

• You should be able to see consistency between replicates
Data Analysis Exercise
Experimental Design
Experimental Design Considerations

• All normal rules apply
  – Think about sources of variation
  – Don't confound variables
  – Think about what batch effects might exist

• Test your antibody well before starting
  – By far the biggest factor in success
  – Good performance on Western / in-situ is not a guarantee, but it's a good start
Experimental Design Considerations

• Number of replicates
  – Lots of studies use 2 replicates
  – Fine for just finding binding sites (motif analysis)
  – Not really enough for differential binding
    • Huge reliance on 'information sharing'
    • No accurate measurement of variance per peak
    • Potentially over-predicts differential binding
  – Should think about likely levels of variability and make replicates to match
Experimental Design Considerations

• Amount of sequencing
  – Can be difficult to predict
  – Depends on
    • Genome size
    • Proportion of genome which is enriched
    • Efficiency of enrichment
  – ENCODE standard is ~20M reads per sample
    • Can get away with fewer (K4me3 for example)
    • Will need more for some marks (H3 for example)
    • Sequencing depth will affect ability to detect changes
Experimental Design Considerations

• Type of sequencing
  – Single end is fine for most applications
    • ATAC-Seq can require paired end for some analyses
  – Moderate read length is required
    • Can map anywhere in the genome
    • 50bp is probably OK. 100bp would be preferable
Downstream Analyses
Composition / Motif Analysis

• Composition
  – Good place to start, can provide either biological or technical insight
  – See if hits (up vs down) cluster based on the underlying sequence composition

• Motifs
  – Great for defining putative binding sites
  – Interesting to do sensitivity check
  – Can do differential motif calling (for hit/non-hit)
Compter - composition analysis

www.bioinformatics.babraham.ac.uk/projects/compter
MEME - Motif Analysis
Gene Ontology / Pathway

• Be careful how you relate hits to genes
  – Really need to have a global link between peak positions and genes
  – Random positions will give significant GO hits if you just use closest/overlapping genes