Introduction to ChIP-Seq data analyses
Outline

• Introduction to ChIP-seq experiment.
  – Biological motivation.
  – Experimental procedure.

• Method and software for ChIP-seq peak calling.
  – Protein binding ChIP-seq.
  – Histone modifications.

• Higher order ChIP-seq data analysis.
  – Overlaps of peaks.
  – Differential binding.
  – Correlate with other data such as RNA-seq.
Introduction to ChIP-seq experiment
ChIP-seq: Chromatin ImmunoPrecipitation + sequencing

• Scientific motivation: measure specific biological modifications along the genome:
  – Detect binding sites of DNA-binding proteins (transcription factors, pol2, etc.) .
  – quantify strengths of chromatin modifications (e.g., histone modifications).
Experimental procedures

1. Crosslink: fix proteins on Isolate genomic DNA.
2. Sonication: cut DNA in small pieces of ~200bp.
3. IP: use antibody to capture DNA segments with specific proteins.
4. Reverse crosslink: remove protein from DNA.
5. Sequence the DNA segments.
Protein/DNA Crosslinking *in vivo*

By Richard Bourgon at UC Berkley
Sonication (cut DNA into pieces)

By Richard Bourgon at UC Berkley
Capture using TF-specific Antibody

By Richard Bourgon at UC Berkley
Immunoprecipitation (IP)
Reverse Crosslink and DNA Purification

By Richard Bourgon at UC Berkley
Amplification (PCR)

By Richard Bourgon at UC Berkley
Other similar sequencing technologies

• “Captured sequencing” – enrich and then sequence selected genomic regions.
• Similar technologies:
  – MeDIP-seq: measure methylated DNA.
  – DNase-seq: detect DNase I hypersensitive sites.
  – FAIRE-seq: detect open chromatin sites.
  – Hi-C: study 3D structure of chromatin conformation.
  – GRO-seq: map the position, amount and orientation of transcriptionally engaged RNA polymerases.
  – Ribo-seq: detect ribosome occupancy on mRNA. Captured mRNA-seq.
  – MeRIP-seq: measure RNA methylation. Captured mRNA-seq.
• Analysis techniques are more or less similar.
Methods and software for ChIP-seq peak/block calling
Data from ChIP-seq

• Raw data: sequence reads.
• After alignments: genome coordinates (chromosome/position) of all reads.
• Often, aligned reads are summarized into “counts” in equal sized bins genome-wide:
  1. segment genome into small bins of equal sizes (50bps).
  2. Count number of reads started at each bin.
ChIP-seq “peak” detection

• When plot the read counts against genome coordinates, the binding sites show a tall and pointy peak. So “peaks” are used to refer to protein binding or histone modification sites.

• Peak detection is the most fundamental problem in ChIP-seq data analysis.
Simple ideas for peak detection

• Regions with reads clustered are likely to be peaks.
• Counts from neighboring windows need to be combined to make inference (so that it’s more robust).
• To combine counts:
  – Smoothing based: moving average (MACS, CisGenome), HMM-based (Hpeak).
  – Model clustering of reads starting position (PICS, GPS).
• Moreover, some special characteristics of the data can be incorporated to improve the peak calling performance.
Before peak detection: what do we know about ChIP-seq?

• Artifacts need to be considered.
  – DNA sequence: can affect amplification process or sequencing process
  – Chromatin structure (e.g., open chromatin region or not): may affect the DNA sonication process.
  – A control sample is necessary to correct artifacts.

• Reads clustered around binding sites to form two distinct peaks on different strands.

• Alignment issue: mappability.
Control sample is important

- A control sample is necessary for correcting many artifacts: DNA sequence dependent artifacts, chromatin structure, repetitive regions, etc.
Reads aligned to different strands

- Number of Reads aligned to different strands form two distinct peaks around the true binding sites.
- This information can be used to help peak detection.

Valouev et al. (2008) *Nature Method*
Mappability

• For each basepair position in the genome, whether a 35 bp sequence tag starting from this position can be uniquely mapped to a genome location.

• Regions with low mappability (highly repetitive) cannot have high counts, thus affect the ability to detect peaks.

| Table 1 Genomic mappability fraction |
|--------------------------------------|-------------------|------------------|
| **Organism** | **Genome size (Mb)** | **Nonrepetitive sequence** | **Mappable sequence** |
| | **Size (Mb)** | **Percentage** | **Size (Mb)** | **Percentage** |
| Caenorhabditis elegans | 100.28 | 87.01 | 86.8% | 93.26 | 93.0% |
| Drosophila melanogaster | 168.74 | 117.45 | 69.6% | 121.40 | 71.9% |
| Mus musculus | 2,654.91 | 1,438.61 | 54.2% | 2,150.57 | 81.0% |
| Homo sapiens | 3,080.44 | 1,462.69 | 47.5% | 2,451.96 | 79.6% |
Peak detection software

- MACS
- Cisgenome
- QuEST
- Hpeak
- PICS
- GPS
- PeakSeq
- MOSAiCS
- ...

MACS (Model-based Analysis of ChIP-Seq)  
Zhang et al. 2008, GB

- Estimate shift size of reads $d$ from the distance of two modes from + and − strands.
- Shift all reads toward 3’ end by $d/2$.
- Use a dynamic Possion model to scan genome and score peaks. Counts in a window are assumed to following Possion distribution with rate:  
  \[ \lambda_{\text{local}} = \max(\lambda_{\text{BG}}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k}) \]
  - The dynamic rate capture the local fluctuation of counts.
- FDR estimates from sample swapping: flip the IP and control samples and call peaks. Number of peaks detected under each p-value cutoff will be used as null and used to compute FDR.
Using MACS

- [http://liulab.dfci.harvard.edu/MACS/index.html](http://liulab.dfci.harvard.edu/MACS/index.html)
- Written in Python, runs in command line.
- Command:
  ```
  macs14 -t sample.bed -c control.bed -n result
  ```
Cisgenome (Ji et al. 2008, *NBT*)

- Implemented with Windows GUI.
- Use a Binomial model to score peaks.

\[ n_i = k_{1i} + k_{2i} \]

\[ k_{1i} \mid n_i \sim \text{Binom}(n_i, p_0) \]
Consider mappability: PeakSeq
Rozowsky et al. (2009) NBT

- First round analysis: detect possible peak regions by identifying threshold considering mappability:
  - Cut genome into segment (L=1Mb). Within each segment, the same number of reads are permuted in a region of \( f \times \text{Length} \), where \( f \) is the proportion of mappable bases in the segment.
• Second round analysis:
  – Normalize data by counts in background regions.
  – Test significance of the peaks identified in first round by comparing the total count in peak region with control data, using binomial p-value, with Benjamini-Hochberg correction.

3. Normalizing control to ChIP-seq sample

4. Second pass: scoring enriched target regions relative to control
   • For potential binding sites calculate the fold enrichment
   • Compute a P-value from the binomial distribution
   • Correct for multiple hypothesis testing and determine enriched target sites
Comparing peak calling algorithms

- Wilbanks et al. (2010) *PloS One*
- Laajala et al. (2009) *BMC Genomics*
rather than the wider regions reported by other methods. For both the FoxA1 and NRSF datasets, the median peak width was between 250 and 400 bp for all methods reporting peak width ranges, with the exception of CisGenome which had smaller median peak width (72 bp for NRSF and 90 bp for GABP; Figure S8 and S9). In contrast, peaks called from the GABP dataset tended to be wider, with median peak widths ranging from 300 to 800 bp, excepting CisGenome which was only 90 bp (Figure S10). This observed variance between datasets emerges either from actual differences in transcription factor binding (GABP binding in a more distributed manner), from variation in the preparation of samples (such as differences in antibody specificity or size selection during the preparation of the sequencing library) or a combination of such factors.

In general, programs also provide an estimate of the exact binding position, given as a single coordinate calculated either as the highest point of tag coverage in the peak or by some other scoring metric. This coordinate is meant to aid the researcher in honing in on section of DNA originally cross-linked by the target protein during the ChIP-enrichment step. Though there is no single nucleotide at which cross-linking occurs, this estimate is meant to facilitate the precise discovery of \( \text{cis} \)-regulatory elements. To assess the positional accuracy of these estimates made by different programs, the distance was calculated between each predicted binding coordinate and the centers of high confidence binding motifs within 250 bp (Figure 7, Table S3). Binding positions were estimated as the center of the reported peak region, if the program did not provide a predicted binding coordinate (HPeak, PeakSeq and Sole-Search; starred in Figure 7). Unsurprisingly, all three datasets revealed that these centered estimates provided much less positional resolution than the precise predictions of binding positions by other programs.

For all programs, the positional accuracy was lower for the GABP dataset (Figure 7C) than for either FoxA1 or NRSF.

Figure 5. Sensitivity assessment.

A

NRSF qPCR verified true positives

B

High confidence NRSE2 motifs

C

GABP qPCR verified true positive

D

High confidence GABP motifs
Another type of approach: modeling the read locations

- Regions with more reads clustered tend to be binding sites.
- This is similar to using binned read counts.
- Reads mapped to forward/reverse strands are considered separately.
- Peak shapes can be incorporated.
PICS: Probabilistic Inference for ChIP-seq
(Zhang et al. 2010 Biometrics)

• Use shifted t-distributions to model peak shape.
• Can deal with the clustering of multiple peaks in a small region.
• A two step approach:
  – Roughly locate the candidate regions.
  – Fit the model at each candidate region and assign a score.
• EM algorithm for estimating parameters.
• Computationally very intensive.
In this section, we use the binding event in question, and having segmented the read data into candidate regions, as described in single candidate region were generated by multiple closely-spaced binding events, along with the corresponding PICS parameter for which we have more prior information. Figure 1a displays that we do not model the sequence counts, but rather the distribution of reverse reads has been biased by a random offset.

Figure 1b displays a candidate region that has two binding events, along with the corresponding PICS parameter.

\[
\begin{align*}
  f_i & \sim \sum_{k=1}^{K} w_k t_4 \left( \mu_{f_k}, \sigma_{f_k}^2 \right) \overset{d}{=} g_f(f_i | \mathbf{w}, \mathbf{\mu}, \mathbf{\delta}, \mathbf{\sigma}_f) \\
  r_j & \sim \sum_{k=1}^{K} w_k t_4 \left( \mu_{r_k}, \sigma_{r_k}^2 \right) \overset{d}{=} g_r(r_j | \mathbf{w}, \mathbf{\mu}, \mathbf{\delta}, \mathbf{\sigma}_r)
\end{align*}
\]
GPS

Guo et al. 2010, Bioinformatics

- The general idea is very similar to PICS.
- Use non-parametric distribution to model the peak shape.
- Estimation of peak shape and peak detection are iterated until convergence.
Use GPS

• Run following command:
  
  ```
  java -Xmx1G -jar gps.jar --g mm8.info --d Read_Distribution_default_default.txt --expt IP.bed --ctrl control.bed --f BED --out result
  ```

• It’s much slower than MACS or CisGenome. So we won’t do it in the lab.
A little more comparison

- I found that using peak shapes helps. GPS tend to perform better. PICS seems not stable.
Bioconductor packages for protein binding ChIP-seq

- There are several packages: chipseq, ChIPseqR, BayesPeak, PICS, etc., but not very popular.
- Most people use command line driven software like MACS or CisGenome GUI.
ChIP-seq for histone modification

- Histone modifications have various patterns.
  - Some are similar to protein binding data, e.g., with tall, sharp peaks: H3K4.
  - Some have wide (mega-bp) “blocks”: H3k9.
  - Some are variable, with both peaks and blocks: H3k27me3, H3k36me3.
Histone modification ChIP-seq data
Peak/block calling from histone ChIP-seq

- Use the software developed for TF data:
  - Works fine for some data (K4, K27, K36).
  - Not ideal for K9: it tends to separate a long block into smaller pieces.
- Many existing methods, mostly based on smoothing, HMM or wavelet.
Complications in histone peak/block calling

• Smoothing-based method:
  – Long block requires bigger smoothing span, which hurts boundary detection.
  – Data with mixed peak/block (K27me3, K36me3) requires varied span: adaptive fitting is computationally infeasible.

• HMM based method:
  – Tend to over fit. Sometimes need to manually specify transition matrix.
Available methods/software for histone data peak calling

- MACS2
- BCP (Bayesian change point caller)
- SICER
- RSEG
- UW Hotspot
- BroadPeak
- mosaicsHMM
- WaveSeq
- ZINBA
- ARHMM
- ...

MACS2

- Has an option for broad peak calling, which uses post hoc approach to combine nearby peaks.
- Syntax:

```
macs2 callpeak -t ChIP.bam -c Control.bam --broad -g hs --broad-cutoff 0.1
```
RSEG

• By Andrew Smith at USC: http://smithlabresearch.org/software/rseg/

• Use negative binomial distribution to model the bin counts, NBDiff distribution for differences between IP and control.

• HMM (3-state for TF data, 2-state for epigenomic domains) for genome segmentation. Use permutation to calculate p-values and determine boundaries.
Use RSEG

• Inputs are bed files.
• First determine “deadzone” (low or unmappable regions). Deadzones for different species can be obtained from their website.

  deadzone -s fa -k 32 -o deadzones-mm9-k32.bed mm9

• Then call blocks:

  rseg-diff -c mouse-mm9-size.bed -o output.bed -i 20 -v -mode 2 -d deadzone-mm9-k32.bed IP.bed control.bed
SICER
Zang et al. 2009, Bioinformatics

• Algorithm:
  – Cut genome into non-overlapping windows and compute a score for each window based on a Poisson model.
  – Identify “islands” by thresholding the scores.
  – Compute a score for each island. This is the tricky part.
Use SICER

- The software is written in python.
- Inputs are bed files for IP and control.
- Good computational performance.
- Results are sometimes sensitive to the parameters.
- A typical command is like:

```
SICER.sh  . h3k27me3.bed control.bed . hg19
2 200 150 0.74 600 0.01
```
• Use ARHMM (auto-regressive HMM) to model the binned read counts.
  – The AR part has smoothing effects which overcomes the problem of HMM that it tends to generate smaller blocks.
• Has capability to include more covariates, and do model selection.
  – Consider IP counts are response, covariates can be control counts, GC content, mappability, TF bindings, etc.
• According to my limited experience, the results seem to be desirable.
• An R package is available at https://code.google.com/p/hmmcov/, but not in very good shape.
Summary for ChIP-seq peak/block calling

• Detect regions with reads enriched.
• Control sample is important.
• Incorporate some special characteristics of the data improves results.
• Calling blocks (long peaks) is harder.
• Many software available.
After peak/block calling

• Compare results among different samples:
  – Presence/absence of peaks.
  – Differential binding.
  – Combinatory patterns.

• Compare results with other type of data:
  – Correlate TF binding with gene expressions from RNA-seq.
Comparison of multiple ChIP-seq

• It’s important to understand the co-occurrence patterns of different TF bindings and/or histone modifications.
• Post hoc methods: look at overlaps of peaks and represent by Venn Diagram.
  – This can be done using different tools. We’ll practice using Bioconductor packages in the lab.
Differential binding (DB)

• This is different from the overlapping analysis, because it considers quantitative changes.

• Straightforward methods:
  – Call peaks from individual dataset.
  – Union the called peaks to form candidate regions.
  – Treat the candidate regions as genes, then use RNA-seq method to test. Or model the differences of normalized counts from two conditions.
Issues to consider in DB analysis

• How to use control data:
  – Need to model the IP-control relationship.
  – Simply subtracting control might not be ideal.

• Normalization between experiments:
  – Signal to noise ratios (SNRs) are different due to technical and biological artifacts.

• Biological variations and experimental design (same as in RNA-seq).
Existing method/software for DB analysis

- ChIPDiff (Xu et al. 2008, Bioinformatics): HMM on differences of normalized IP counts between two groups.
- MAnorm (Shao et al. 2012, Genome Biology): normalization based on MA plot of counts from two groups, then use normalized “M” values to rank differential peaks.
- DBChIP (Liang et al. 2012 Bioinformatics) and DiffBind: Bioconductor packages, based on RNA-seq method.
- ChIPComp (Chen et al. 2015 Bioinformatics): Based on linear model framework, works for general design.
Combine ChIP- and RNA-seq

• It is of great interest to study how the gene expressions are controlled by protein bindings and epigenetic modifications.

• Easy approach:
  – Look at the correlation of promoter TF binding (from ChIP-seq), and gene expression (from RNA-seq).

• More advanced approaches:
  – Build a model to predict gene expression (from RNA-seq) from protein binding and epigenetic data (from ChIP-seq).
  – Build a network for all ChIP- and RNA-seq data.
Predict expression from TF binding
Ouyang et al. (2009) *PNAS*

- **Goal:** to build a model to predict gene expressions using 12 TF binding datasets.
- **Data:** mouse ESC TF data from a cell paper by a Singapore group.
- **Method:** regression based.
- **A similar paper using histone modification to predict gene expression is Karlic et al. (2010) *PNAS.*
Procedures in Ouyang et al.

- Read counts are first summarized into gene level.
- Association strength between TF j and gene is:

\[ a_{ij} = \sum_k g_k e^{-d_k/d_0}, \]

where \( g_k \) is the intensity (number of reads aligned to the coordinate) of the \( k \)th binding peak of the TF \( j \), \( d_k \) is the distance (number of nucleotides) between the TSS of gene \( i \) and the \( k \)th binding peak in the reference genome, and \( d_0 \) is a constant. In theory, the summation is over all binding peaks of a given TF.

- Result \( a_{ij} \) is a matrix of ngenes by nTF.
- PCA on \( a_{ij} \) to avoid having one TF dominating.
- log-linear model:

\[ \log Y_i = \mu + \sum_{j=1}^{M} \beta_j X_{ij} + \epsilon_i, \]
Prediction results from TF binding

A

B

C

D

Log10(Predicted ESC expr.) vs. Log10(RNA-Seq ESC expr.)

Log10(Predicted ESC expr.) vs. Log10(RNA-Seq ESC expr.)

Log10(Predicted ESC expr.) vs. Log10(Array ESC expr.)

Log10(Predicted ESC expr.) vs. Log10(Array ESC expr.)

$r = 0.652$

$r = 0.806$

$r = 0.512$

$r = 0.727$
Prediction results from histone modification

**A** Scatterplot with the measured CD4+ log(expression) vs. predicted CD4+ log(expression) in CD4+ cells. The correlation coefficient is $r = 0.77$

**B** Comparison between the models ranked according to prediction accuracy. The models are ranked from the lowest to the highest prediction accuracy.

**C** Bar plot showing the frequency of appearance of different histone modifications in best scoring three-modifications models (142 models) for CD4+ cells.

**D** Scatterplot with the measured CD36+ log(expression) vs. predicted CD36+ log(expression) in CD36+ cells. The correlation coefficient is $r = 0.75$

**E** Scatterplot with the measured CD133+ log(expression) vs. predicted CD133+ log(expression) in CD133+ cells. The correlation coefficient is $r = 0.63$
Network based analysis of multiple ChIP-seq

• Yu et al. (2008) *Genome Research*.

• Data used: human CD4+ T-cell chip-seq for 23 histones and TF binding (from Keji Zhao’s Cell paper). Read counts are summarized into TSS +/- 1kb region.

• Method:
  – Bayesian network on discretized counts using WinMine. A randomization procedure is implemented to select the robust edges.
Result from BN

```
Ac h a i no ft h r e en o d e sn e g a t i v e l ya s s o c i a t dw i t hg e n e expression, H3K9me3 \rightarrow H3K9me2 \rightarrow H3K79me2, is down-stream of H3K27me3 and H4K20me3 (Fig. 3C). Three other nodes, H3K27me2, H4R3me2, and H3K36me3, are also negatively correlated with gene expression, with H4R3me2 dependent on H3K27me2 and H3K4me3, and H3K36me3 on H3K4me3 only (Fig. 3C). The rest of the nodes, including all the monomethylations, H3K4me3, Pol II, H2A.Z, and CTCF, are all directly or indirectly associated and positively correlated with gene expression. Among them, a chain of causal relationships formed among four monomethylations, H3K4me1 \rightarrow H2BK5me1 \rightarrow H3K36me1 \rightarrow H3K79me1, seems to be significantly longer than expected (P = 0.044 assuming normal distribution) (see Methods) (Fig. 3C). The biological meaning of this observation is currently unknown.

The modifications or binding events, such as H2A.Z and CTCF binding at the bottom of the cascade, are not predicted to be causal to gene expression (Fig. 3C).

Existing experimental support for the inferred relationships since epigenetic modifications reflect the gene expression status in a particular tissue and state, the same gene is likely to be modified differently in distinct tissues or conditions. Even the methyltransferase or demethylase complexes catalyzing the modifications might be different. However, the relationship of each modification to gene expression status, and the relationships among various modifications present in T cells might reflect their interaction patterns in general toward forming the
```

Figure 3. Causal relationships among histone modifications and gene expression. (A) Flowchart of a Bayesian network construction using sequence counts within TSS \pm 1kb. See text for details. (B) The coverage and accuracy of models derived from sequence counts within TSS \pm 600bp, TSS/H11506 1kb, and TSS/H11506 2kb. For each N (an integer from one to 10) nine out of 10 group combinations, the models' accuracy and coverage are calculated generating a curve for each sequence range used to construct the models. We performed random grouping 100 times, and hence, the coverage and accuracy at each N is the average of 100 trials. The vertical and horizontal bars on the curve denote the standard deviations of accuracy and coverage at each point. (C) The common Bayesian network (see text for details) consisted of only compelled edges agreed by all 100 trials. The model is based on the sequence counts in TSS/H11506 1kb. The edge colors indicate the correlations (measured by Pearson correlation coefficient [PCC]) among the various modification/binding factors; nodes are colored by their correlation to gene expression. Colors are scaled as shown in the color legend. The edge directions have the same meanings as in Fig. 1B. (D) The causal relationships in the Bayesian network model are not expected by shuffled sequence counts among genes for each ChIP. Comparing to that of the real data, when the sequence counts are shuffled among genes, each Bayesian network contains zero compelled edge. Each point on a curve represents the average results of 100 tests or 100 simulations, with the vertical bars on the curve denoting the standard deviations at each point.
Review

• ChIP-seq detects TFBS or measure histone modifications along the genome.
• Peak (short and long) detection is the major goal of data analysis.
• Number of aligned reads are input data. Data in neighboring regions need to be combined to call peaks.
• Many similar technologies, and the method are more or less the same.