

Chromatin Immunoprecipitation ChIP-seq technology

Mikhail Dozmorov

2021-04-14

Outline

Introduction to ChIP-seq experiment

- Biological motivation
- Experimental procedure

Methods and software for ChIP-seq peak calling

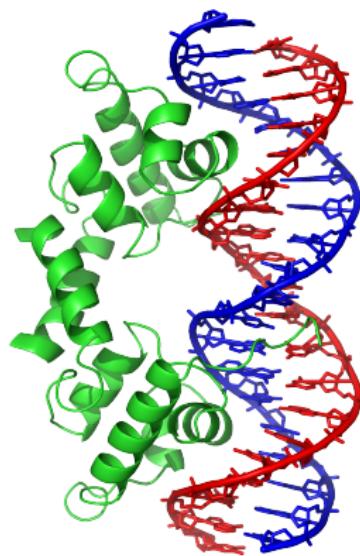
- Protein binding ChIP-seq
- Histone modifications

Higher-order ChIP-seq data analysis

- Peak/motif detection
- Differential binding
- Correlate with other data such as RNA-seq

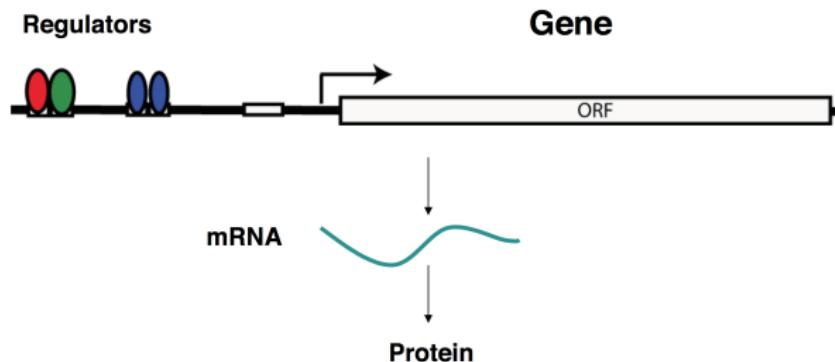
Transcription factors regulate gene expression

- Transcription factors are proteins that bind to specific DNA sequences and act as regulators of gene expression
- Humans have ~2,000 transcription factors



Gene Regulation: DNA -> RNA -> Protein

- What are the transcription factors (TFs) that control gene expression?
- At what genes do these TFs operate?
- Understanding transcriptional regulatory network will
 - Reveal how cellular processes are connected and coordinated
 - Suggest new strategies to manipulate phenotypes and combat disease



ChIP-seq big picture

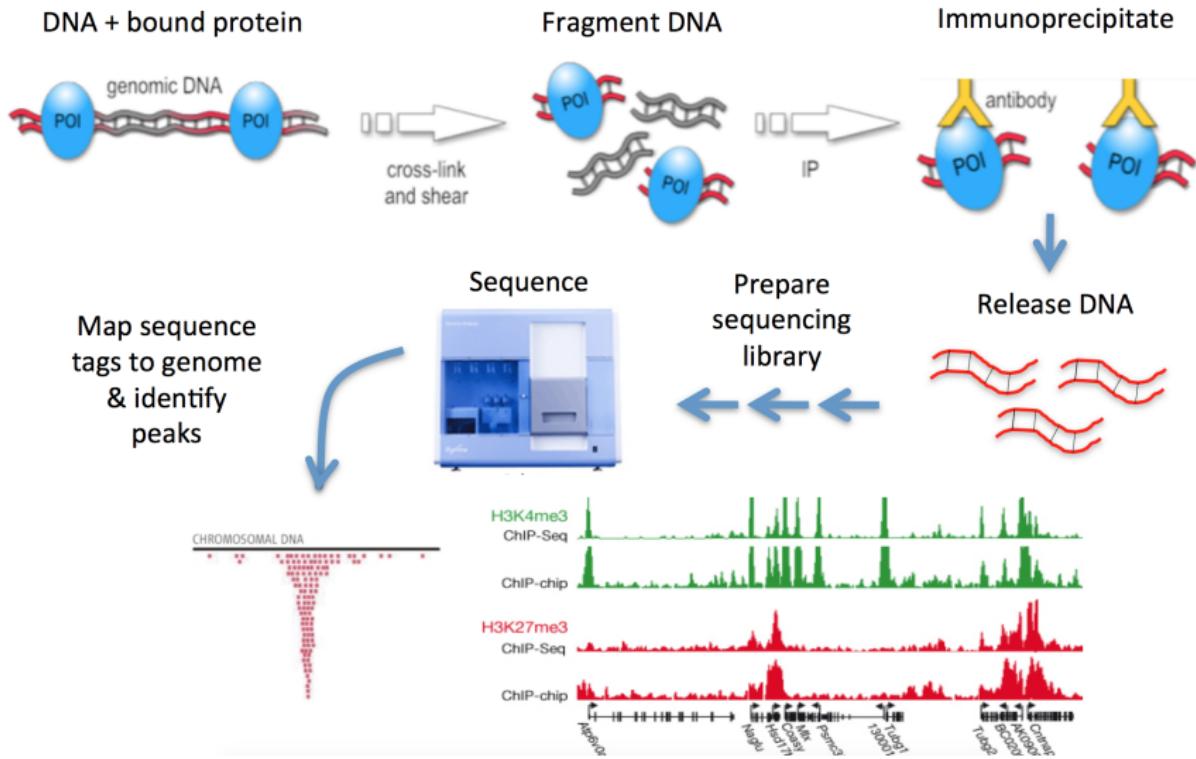
Combine high-throughput sequencing with Chromatin Immunoprecipitation to identify specific protein-DNA interactions genome-wide, including:

- Transcription factors
- Histone marks (various types and modifications)
- RNA Polymerase binding (survey of transcription)
- DNA polymerase binding (investigate DNA replication)
- DNA repair enzyme binding
- Fragments of DNA that are modified (e.g., methylated)

Experimental procedures

- ① ** Crosslink**: fix proteins on Isolate genomic DNA.
- ② **Sonicate**: cut DNA in small pieces of ~200bp.
- ③ **Immunoprecipitate (IP)**: use an antibody to capture DNA segments with specific proteins.
- ④ **Reverse crosslink**: remove protein from DNA.
- ⑤ **Sequence** the DNA segments.

ChIP-seq overview



How many sequences are needed to find ChIP-seq peaks?

- Effective analysis of ChIP-seq data requires sufficient coverage by sequence reads (sequencing depth)
- The required depth depends mainly on the size of the genome and the number and size of the binding sites of the protein
- For mammalian transcription factors (TFs) and chromatin modifications such as enhancer-associated histone marks, which are typically localized at specific, narrow sites and have on the order of thousands of binding sites, 20 million reads may be adequate (4 million reads for worm and fly TFs)
- Proteins with more binding sites (e.g., RNA Pol II) or broader factors, including most histone marks, will require more reads, up to 60 million for mammalian ChIP-seq

Bailey, Timothy, Paweł Krajewski, István Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang. "Practical Guidelines for the Comprehensive Analysis of ChIP-Seq Data." Edited by Fran Lewitter. PLoS Computational Biology 9, no. 11 (November 14, 2013): e1003326. <https://doi.org/10.1371/journal.pcbi.1003326>.

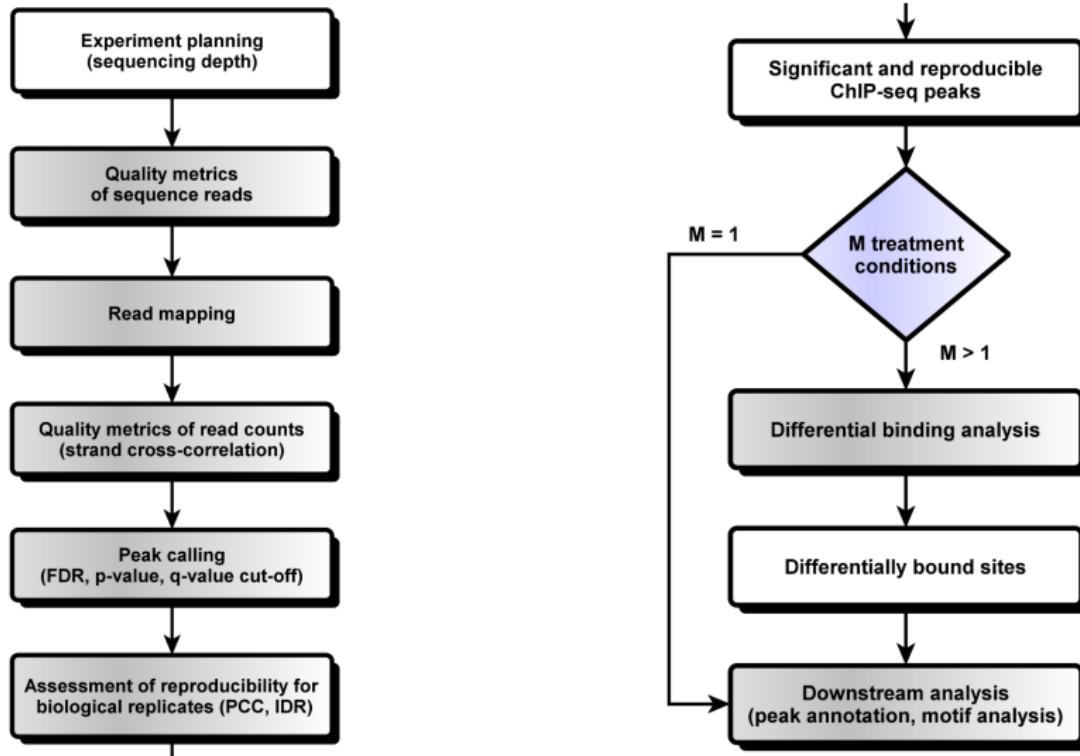
Data from ChIP-seq

- 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 (e.g., 50bp)
 - ② Count the number of reads started at each bin
- Alternatively, consider counts in the regions of interest, e.g., promoters of protein-coding genes (2,000bp upstream and 500bp downstream of transcription start site)

Section 1

Methods and software for ChIP-seq peak/block calling

ChIP-seq analysis workflow



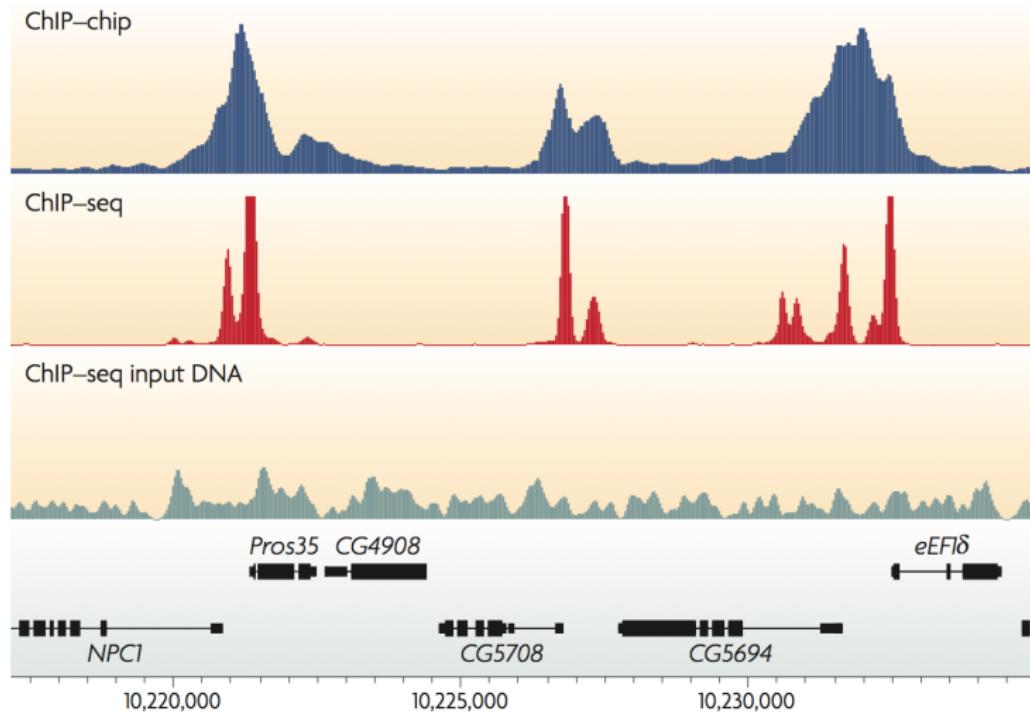
ChIP-seq “peak” detection

- 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

Peak calling: a classic signal versus noise problem



Park, Peter J. "ChIP-seq: Advantages and Challenges of a Maturing Technology." *Nature Reviews Genetics* 10, no. 10 (October 2009): 669–80. <https://doi.org/10.1038/nrg2641>.

Simple ideas for peak detection

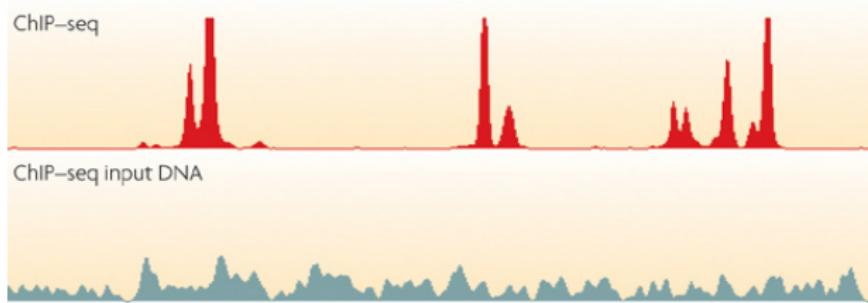
- Peaks are regions with reads clustered, so they can be detected from binned read counts
- Counts from neighboring windows need to be combined to make an inference (so that it's more robust)
- To combine counts:
 - Smoothing based: moving average (MACS, CisGenome), HMM-based (Hpeak)
 - Model clustering 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.
- Importantly, control samples should be sequenced significantly deeper than the ChIP ones in a TF experiment and in experiments involving diffused broad-domain chromatin data. This is to ensure sufficient coverage of a substantial portion of the genome and non-repetitive autosomal DNA regions



Control sample is important

- There are three commonly used types of control sample:
 - ① **Input DNA** (a portion of the DNA sample removed prior to immunoprecipitation (IP))
 - ② **Mock IP DNA** (DNA obtained from IP without antibodies)
 - ③ **DNA from nonspecific IP** (IP performed using an antibody, such as immunoglobulin G, against a protein that is not known to be involved in DNA binding or chromatin modification)

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

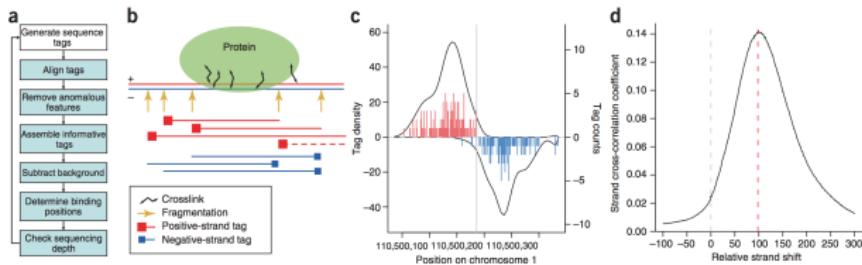
Table 1 Genome mappability fraction

Organism	Genome size (Mb)	Nonrepetitive sequence		Mappable sequence	
		Size (Mb)	Percentage	Size (Mb)	Percentage
<i>Caenorhabditis elegans</i>	100.28	87.01	86.8%	93.26	93.0%
<i>Drosophila melanogaster</i>	168.74	117.45	69.6%	121.40	71.9%
<i>Mus musculus</i>	2,654.91	1,438.61	54.2%	2,150.57	81.0%
<i>Homo sapiens</i>	3,080.44	1,462.69	47.5%	2,451.96	79.6%

Rozowsky, Joel, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. "PeakSeq Enables Systematic Scoring of ChIP-Seq Experiments Relative to Controls." *Nature Biotechnology* 27, no. 1 (January 2009): 66–75. <https://doi.org/10.1038/nbt.1518>.

How do peak-finders map binding sites?

- Fragments contain the TF binding site at a (mostly) random position within them
- Reads are randomly generated from left or right edges (sense or antisense) of fragments
- Binding site position = mid-way between sense tag peak and antisense tag peak
- To get binding site peak, shift sense downstream by 1/2 fragsize & antisense upstream by 1/2 fragsize



Kharchenko, Peter V, Michael Y Tolstorukov, and Peter J Park. "Design and Analysis of ChIP-Seq Experiments for DNA-Binding Proteins." *Nature Biotechnology* 26, no. 12 (December 2008): 1351–59. <https://doi.org/10.1038/nbt.1508>.

Regions to watch out for abnormal peaks

- **Centromeres** - “acen” regions from <http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/cytoBand.txt.gz>
- **ENCODE blacklisted regions** -
<https://sites.google.com/site/anshulkundaje/projects/blacklists>
- **Gaps** - unsequenced parts of human genome, <http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/gap.txt.gz>
- **SuperDups** - large genomic duplications, <http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/genomicSuperDups.txt.gz>

Peak detection

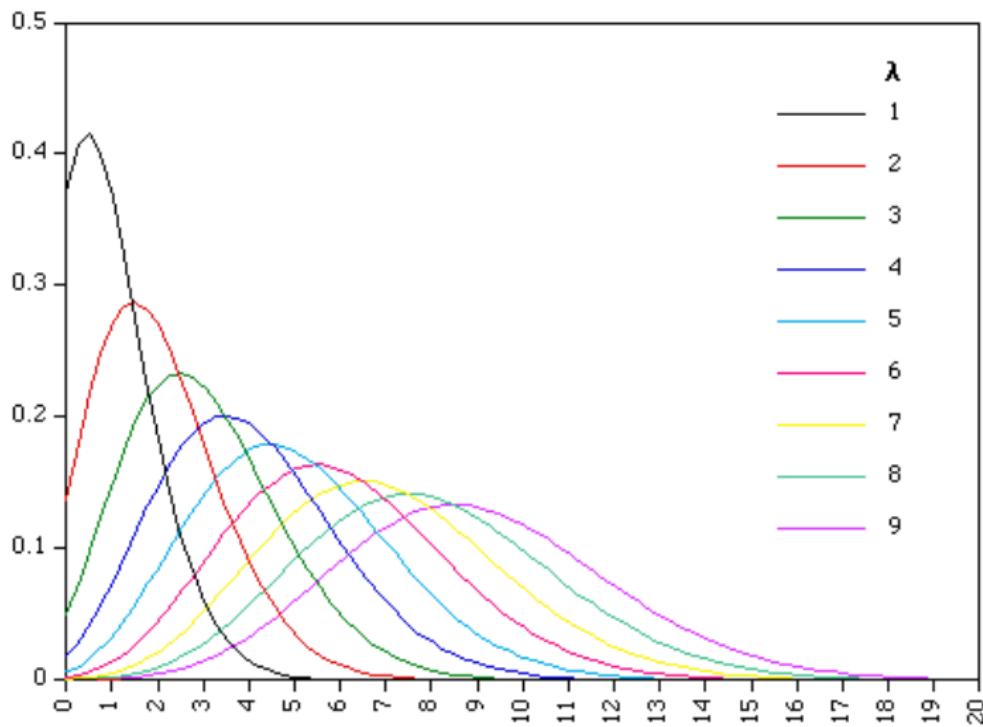
- Calculate read count at each position (bp) in the genome. Don't use a sliding window
- Determine if read count is greater than expected (at each position, bp)
- We need to correct for input DNA reads (control)
 - non-uniformly distributed
 - vastly different numbers of reads between ChIP and Input

The Poisson distribution: discrete distribution to model coverage

- $P(k \text{ discrete events}) = \frac{\lambda^k e^{-\lambda}}{k!}$ Where e is Euler's constant (2.718), and λ is the average number of occurrences of an event
- Example: the “hundred year flood”. Thus $\lambda = 1$ (1 catastrophic flood every 100 years)

https://en.wikipedia.org/wiki/Poisson_distribution

Poisson distribution with different values of λ



Is the observed read count at a given genomic position greater than expected?

Read counts follow a Poisson distribution.

$$P(X \geq x) = 1 - \sum_0^{x-1} \frac{\lambda^x e^{-\lambda}}{x!}$$

- x - observed read count
- λ - expected read count

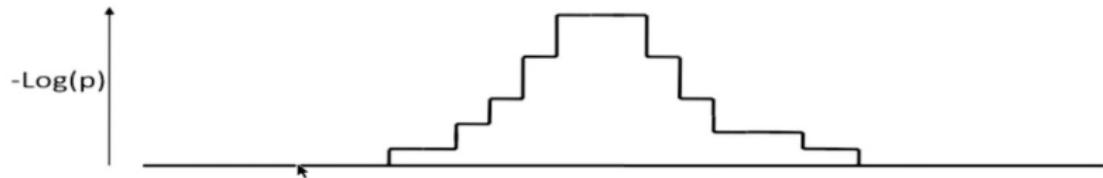
Example:

- $x = 10$ reads, observed
- $\lambda = 0.5$ reads, expected
- $P(X \geq 10) = 1.7 \times 10^{-10}$
- $-\log_{10} P(X \geq 10) = 9.77$
- In R, $P(X \geq 10 | \lambda = 0.5)$ is `1 - sum(dpois(0:9, 0.5))`

Is the observed read count at a given genomic position greater than expected?

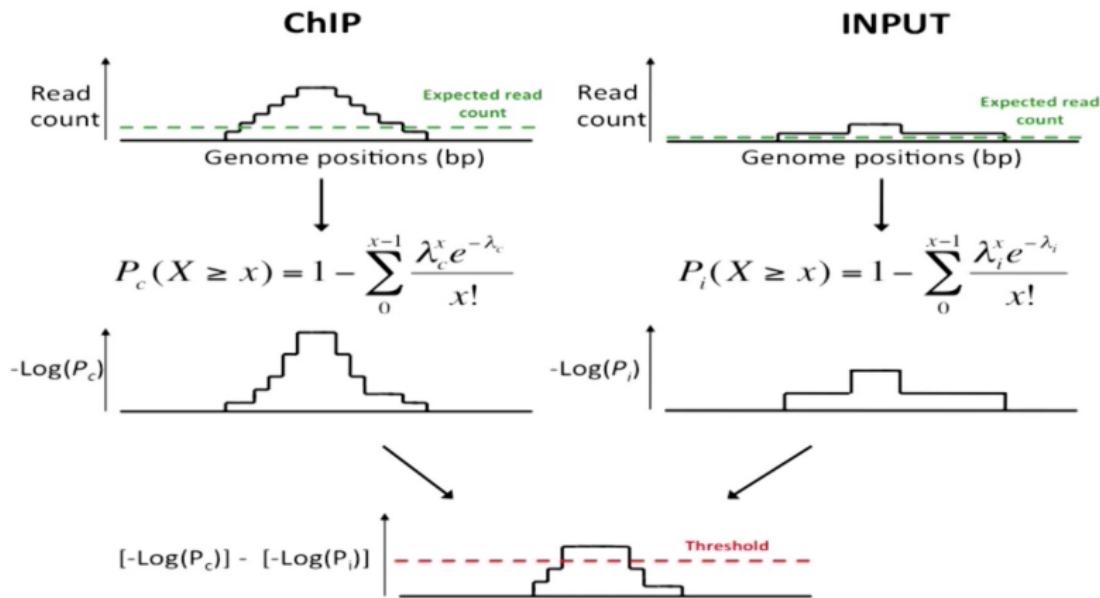


$$P_c(X \geq x) = 1 - \sum_0^{x-1} \frac{\lambda_c^x e^{-\lambda_c}}{x!}$$



Expected read count = total number of reads * extended frag len / chr len

Is the observed read count at a given genomic position greater than expected?



Normalized Peak score at each bp

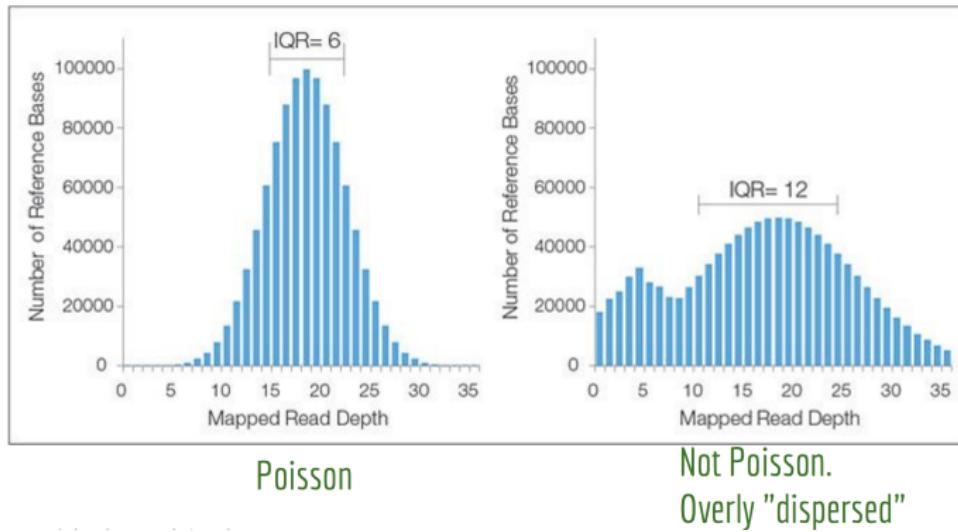
$$R = -\log_{10} \frac{P(X_{ChIP})}{P(X_{Input})}$$

Will detect peaks with high read counts in ChIP, low in Input

Works when no input DNA ($X_{Input} = 0$)

$$P(X \geq x) = 1 - \sum_0^{x-1} \frac{\lambda^x e^{-\lambda}}{x!}$$

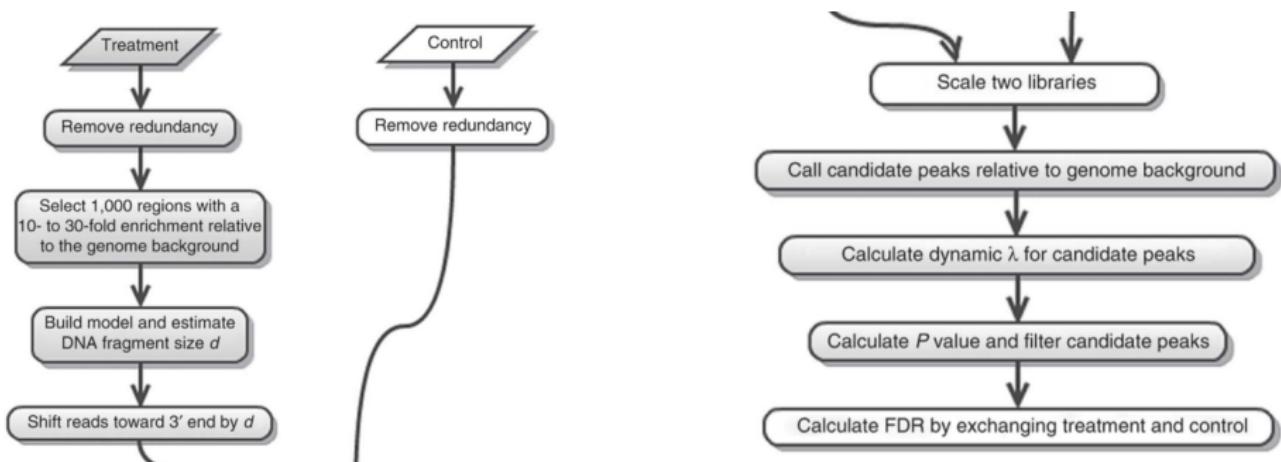
Ideally, sequencing coverage will follow a Poisson distribution. But...



- Negative binomial fits sequencing coverage data much better

<https://www.youtube.com/watch?v=HK7WKsL3c2w>, Pei Fen Kuan, et al., "A Statistical Framework for the Analysis of ChIP-Seq Data," Journal of the American Statistical Association 106, no. 495 (September 2011): 891–903,
<https://doi.org/10.1198/jasa.2011.ap09706>.

MACS (Model-based Analysis of ChIP-Seq)



Written in Python, runs in the command line. `macs14 -t sample.bed -c control.bed -n result`

<http://liulab.dfci.harvard.edu/MACS/index.html>

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, et al. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9, no. 9 (2008): R137.

<https://doi.org/10.1186/gb-2008-9-9-r137>.

MACS (Model-based Analysis of ChIP-Seq)

- 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 Poisson model to scan the genome and score peaks.
 - Counts in a window are assumed to follow Poisson distribution with rate : $\lambda_{local} = \max(\lambda_{BG}, \lambda_{1K}, \lambda_{5K}, \lambda_{10K})$ where λ_{1K} , λ_{5K} and λ_{10K} are λ estimated from the 1 kb, 5 kb or 10 kb window centered at the peak location in the control samples and call peaks.
- FDR estimates from sample swapping: flip the IP and control samples and call peaks. The number of peaks detected under each p-value cutoff will be used as null and used to compute FDR
 - The number of “negative” peaks should be ~10 times less

MACS fine-tuning

MACS can't build a model:

- Adjust the *mfold* values (the fold over background ranges MACS considers for paired peaks)
- Tell MACS to not build a model but instead use the shiftsize you specify.

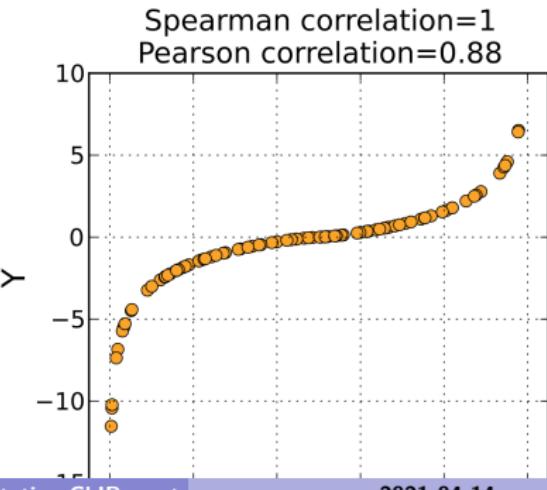
Peaks/Negative Peaks ratio is poor, or too few peaks are detected:

- Adjust model settings to see if you can improve both. Otherwise, you may have to conclude that 1) your library was no good or 2) the factor just doesn't bind to many places in the genome

Reproducibility between samples

- Spearman's rank correlation provides a metric for replicate consistency but does not select consistent events
- Consider two ranked lists of n detected events X and Y , one from each replicate, each ranked by scores from most significant to least significant
- For matched event i ranks are x_i and y_i in X and Y

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$



Irreproducible Discovery Rate (IDR) Analysis

- Consider that the lists X and Y are a mixture of two kinds of events - reproducible and irreproducible
- Model the ranking scores as a two-component mixture and learn the parameters of the reproducible and irreproducible components
- For IDR α , select top l pairs using their scores such that the probability that the rate of pairs from the irreproducible part of the mixture is α

Irreproducible Discovery Rate (IDR) Analysis

- $\Psi_n(t)$ is the fraction of the n events that are paired in the top $n \times t$ events in both X and Y . It is roughly linear from $t = 0$ to the point when events are no longer reproducible (not shared between replicates within the ranking)
- $\Psi'_n(t)$ is first derivative of $\Psi_n(t)$ with respect to t . It allows us to visualize when we transition from reproducible to irreproducible events as t increases

Irreproducible Discovery Rate (IDR) Analysis

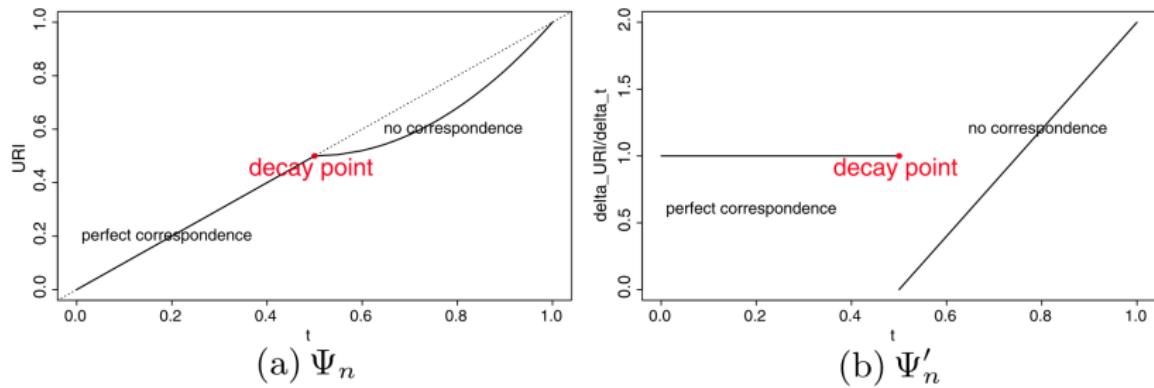


FIG. 1. An illustration of the correspondence profile in an idealized case, where top 50% are genuine signals and bottom 50% are noise. In this case, all signals are ranked higher than noise; two rank lists have perfect correspondence for signals and no correspondence for noise. (a) Correspondence curve. (b) Change of correspondence curve.

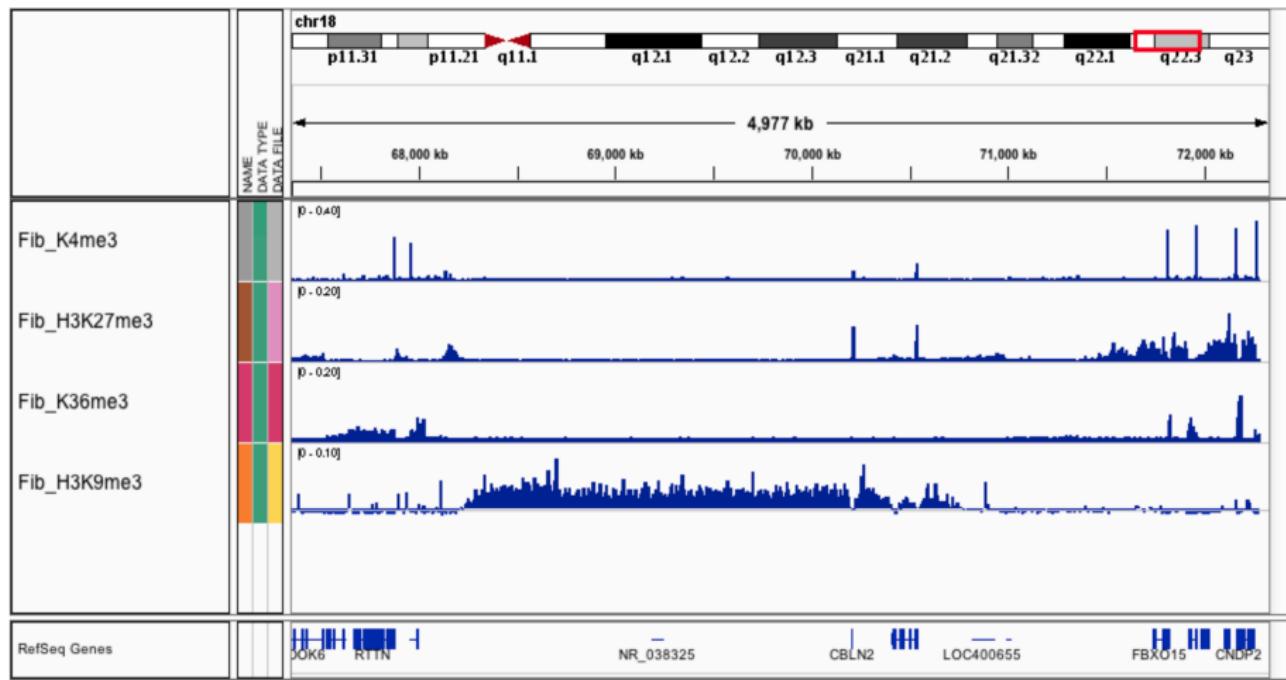
URI - upper ranked identifications

Li, Qunhua, James B. Brown, Haiyan Huang, and Peter J. Bickel. "Measuring Reproducibility of High-Throughput Experiments." *The Annals of Applied Statistics* 5, no. 3 (September 2011): 1752–79. doi:10.1214/11-AOAS466.

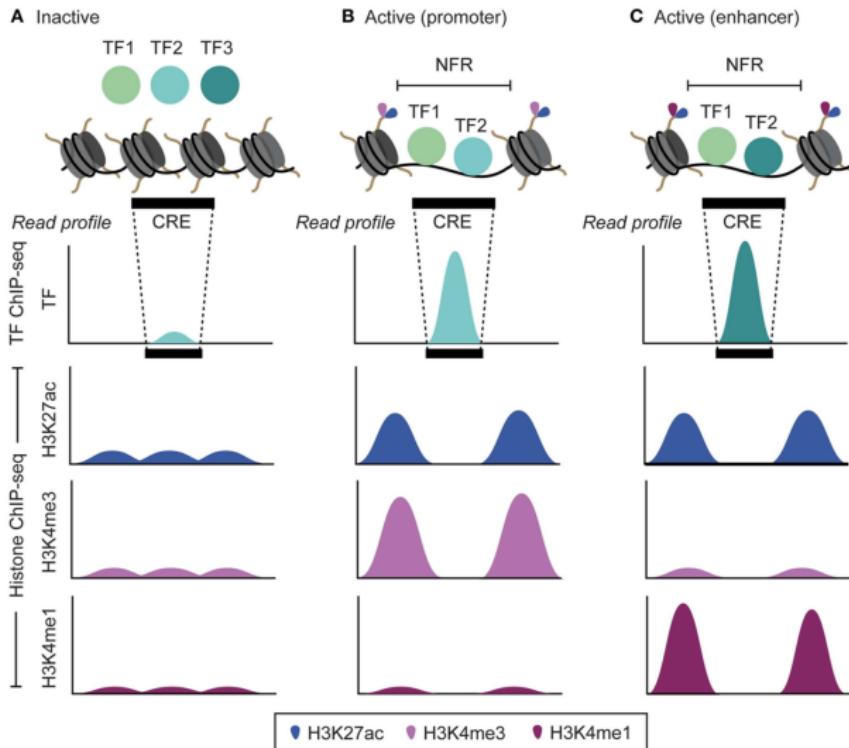
ChIP-seq for histone modification

- 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. Also, RNA Pol II - stalled binding shows as peaks, moving along with transcription shows as broad stretches

Histone modification ChIP-seq data



The transcription factor and histone modification landscape



Peak/block calling from histone ChIP-seq

Use the software developed for TF data:

- Works fine for some data (H3K4me1/2/3, H3K27ac, H3K36me3)
- Not ideal for H3K9me1: 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 a bigger smoothing span, which hurts boundary detection
- Data with mixed peak/block (H3K27me3, H3K36me3) requires varied span: the adaptive fitting is computationally infeasible

HMM-based method:

- Tend to overfit. 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

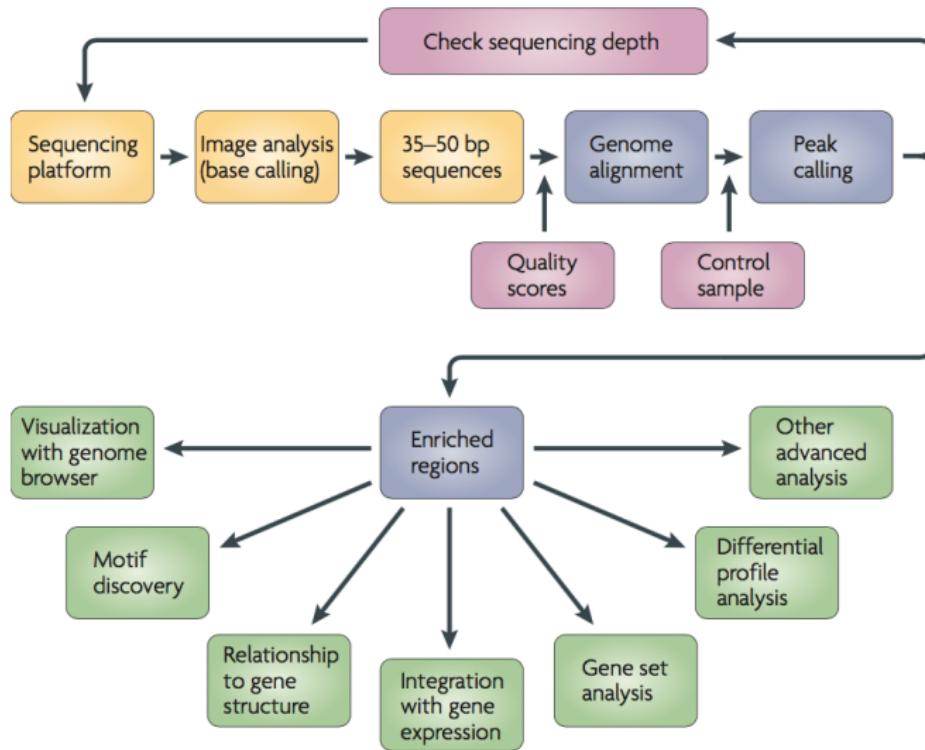
- An updated version of MACS:
<https://github.com/taoliu/MACS/blob/master/README.rst>.
- Has an option for broad peak calling, which uses a post hoc approach to combine nearby peaks.
- Syntax:

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

Summary for ChIP-seq peak/block calling

- Detect regions with reads enriched
- Control sample is important
- Incorporate some special characteristics of the data to improve results
- Calling wide peaks is harder
- Many software available

After peak/block calling



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

Differential binding (DB)

This is different from the overlap analysis because it considers quantitative changes.

Straightforward methods:

- Call peaks from individual datasets
- Union the called peaks to form candidate regions
- Treat the candidate regions as genes, then use the 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 variation 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
- **DIME** (Taslim et al. 2009, 2011, Bioinformatics): finite mixture model on differences of normalized IP counts
- **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
- **ChIPnorm** (Nair et al. 2012, PLoS One): quantile normalization for each data. Ad hoc method for detecting 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 the linear model framework, works for general design

Software packages for the analysis of differential binding in ChIP-seq

Software tool	Availability	Notes
ChIPDiff [36]	http://cmb.gis.a-star.edu.sg/ChIPSeq/paperChIPDiff.htm	Differential histone modification sites using a hidden Markov model
Comparative ChIP-seq [25]	http://www.starklab.org/data/bardet_natprotoc_2011/	Fold change ratio between normalized peak heights
DBChIP [33]	http://pages.cs.wisc.edu/~kliang/DBChIP/	
DESeq§ [31]	http://www.bioconductor.org/packages/release/bioc/html/DESeq.html	Assigns uncertainty measures in a test of non-differential binding (uses edgeR)
DiffBind	http://www.bioconductor.org/packages/release/bioc/html/DiffBind.html	Test based on a model using the negative binomial distribution
DIME [35]	http://cran.r-project.org/web/packages/DIME/	Differential binding affinity analysis (uses edgeR and DESeq)
edgeR§ [32]	http://www.bioconductor.org/packages/release/bioc/html/edgeR.html	Differential identification using mixtures ensemble
MACS [17] (version 2)	https://github.com/taoliu/MACS/	Empirical Bayes estimation and exact tests based on the negative binomial distribution
MAnorm [34]	http://bcb.dfci.harvard.edu/~gcyuan/MAnorm/MAnorm.htm	Differential peak detection based on paired four bedGraph files
MMDiff	http://bioconductor.org/packages/release/bioc/html/MMDiff.html	Robust regression to derive a linear model
NarrowPeaks	http://bioconductor.org/packages/release/bioc/html/NarrowPeaks.html	Differences in shape using Kernel methods
POLYPHEMUS [37]	http://cran.r-project.org/web/packages/polyphemus/	Shape-based analysis of variation using functional PCA
		Non-linear normalization on RNA Pol II profiling

Combine ChIP- and RNA-seq

It is of great interest to study how 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

Downstream analysis

- Find the nearest gene to each peak
- Check distribution relative to gene features (start site, exon, intron, upstream/downstream)
- Find overrepresented motifs in peak region (TFBS binding sites of our factor + possible co-binders) - kmers/logos
- Check if peaks are clustered or co-occur with other binding events
- Sequence conservation (or conservation of binding event, if data is available)
- Gene set functional analysis

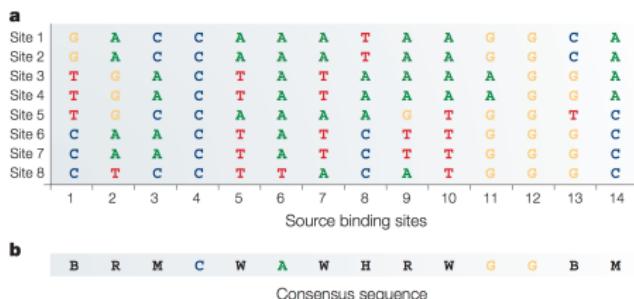
What are motifs?

- Short, recurring patterns in DNA that are presumed to have a biological function
- Often indicate sequence-specific binding sites for proteins such as nucleases and transcription factors (TF)
- In this example, if allowing 1 base mismatch, there are two motifs:
TTGACA and GCATC:

```
GCACGCGGTATCGTTAGCTTGACAATGAAGAATCCCCCCGCTCGACAGT  
GCATACTTTGACACTGACTTCGCTTCTTAATGTTAATGAAACATGCG  
CCCTCTGGAAATTAGTGCGCATCTCACAAACCGAGGAATGACCAAATG  
GTATTTGAAAGTAAGGCAACGGTGATCCCCATGACACCAAAGATGCTAAG  
CAACGCTCAGGCAACGTTGACAGGTGACACCGTTGACCTTTGACTGCGGCCTCTGC  
GTCTCTTGACCCGCTTAATCTAAAGGCCTCCTATTAGTATCCGCAATGT  
GAACAGGAGCGCGAGCCATCAATTGAAGCGAAGTTGACACCTAATAACT
```

Building models for predicting transcription factor binding sites

- Collect sequences, align, create a consensus sequence by selecting a degeneracy nucleotide sequence for each position
 - Note unusual binding sites, like one and eight, that have only 50% of the identical nucleotides



Position frequency matrix

To more accurately reflect the characteristics at each position, a matrix that contains the number of observed nucleotides at each position is created.

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13
A	0	4	4	0	3	7	4	3	5	4	2	0	0
C	3	0	4	8	0	0	0	3	0	0	0	0	2
G	2	3	0	0	0	0	0	0	1	0	6	8	5
T	3	1	0	0	5	1	4	2	2	4	0	0	1

Position weight matrix

A position weight matrix (PWM) is constructed by dividing the nucleotide probabilities by expected background probabilities and converting the values to a log-scale

$$W_{b,i} = \log_2 \frac{p(b, i)}{p(b)}$$

where $p(b)$ - background probability of base b ; $p(b, i)$ - corrected probability of base b in position i ; $W_{b,i}$ - PWM value of base b in position i

Position weight matrix

- The frequency matrix is usually converted to a position weight matrix (PWM) by normalizing frequency values to a log-scale
- PWMs are also known as position-specific scoring matrices (PSSMs)
- A quantitative score for any DNA sequence can be generated by summing the values corresponding to the observed nucleotide at each position. These scores are proportional to binding energies

d Position weight matrix (PWM)

A	-1.93	0.79	0.79	-1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	-1.93	-1.93	0.79
C	0.45	-1.93	0.79	1.68	-1.93	-1.93	-1.93	0.45	-1.93	-1.93	-1.93	-1.93	0.00	0.79
G	0.00	0.45	-1.93	-1.93	-1.93	-1.93	-1.93	-1.93	0.66	-1.93	1.30	1.68	1.07	-1.93
T	0.15	0.66	-1.93	-1.93	1.07	0.66	0.79	0.00	0.00	0.79	-1.93	-1.93	-0.66	-1.93

e Site scoring

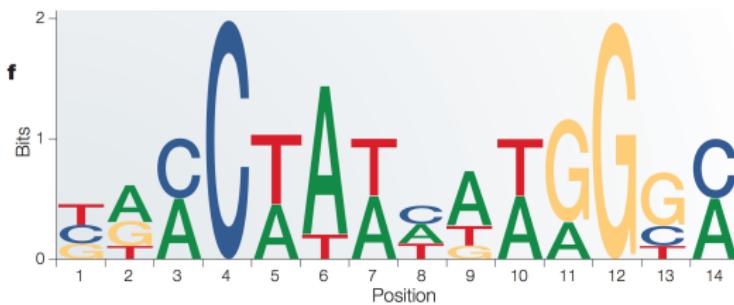
0.45	-0.66	0.79	1.68	0.45	-0.66	0.79	0.45	-0.66	0.79	0.00	1.68	-0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

$$\Sigma = 5.23, 78\% \text{ of maximum}$$

The quantitative PWM score for a putative site is the sum of the PWM values for each nucleotide in the site.

Sequence logo

- The specificity in each column of the alignment can be measured in terms of information content
- A sequence logo scales each nucleotide by the total bits of information multiplied by the relative occurrence of the nucleotide at the position
- Sequence logos enable fast and intuitive visual assessment of pattern characteristics



Probability values can be used to determine the total information content (in bits) in each position.

Motif finding

- Sequence K reads. Goal, to find the locations of motif sites
 - Sequence 1: site starts at a_1
 - Sequence 2: site starts at a_2
 - ...
 - Sequence N: site starts at a_N
- We don't know the alignment variable $A = \{a_1, a_2, \dots, a_N\}$
- We seek within each sequence mutually similar segments of specified width W

Motif finding

- The Gibbs motif sampling (a Monte-Carlo algorithm)
- Every position i in the motif sequence follows probability distribution with $q_{ij} = \{q_{i,1}, \dots, q_{i,4}\}$
 - i ranges from 1 to motif length w
 - j ranges across the nucleotides (paper describes amino acids)
- Background - each non-motif position follows a common probability distribution $p_j = \{p_1, \dots, p_4\}$
- b_j - background frequency of symbol j
- $c_{i,j}$ - count of symbol j at position i

Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment." *Science* (New York, N.Y.) 262, no. 5131 (October 8, 1993): 208–14.

HOMER

- Tries to identify regulatory elements enriched in one set of sequences compared to another
- Motif discovery algorithm designed for regulatory element analysis in genomics applications (DNA sequences only)
 - Known motifs counting
 - De novo motifs identification
 - Attempts to match *de novo* motifs to known ones

<http://homer.ucsd.edu/homer/motif/>

Software tools for motif analysis of ChIP-seq peaks and their uses

See <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003326#s5> for full table

Category	Software tool	Web Server	Obtain peak regions	Motif discovery	Motif comparison	Central motif enrichment analysis	Local motif enrichment analysis	Motif spacing analysis	Motif prediction/mapping
Obtaining sequences	Galaxy [50-52]	X	X						
	RSAT [53]	X	X						
	UCSC Genome Browser [54]	X	X						
Motif discovery + more	ChIPMunk [55]	X			X				
	CisGenome [56]			X	X				
	CompleteMOTIFS [48]			X	X				
	MEME-ChIP [57]	X		X	X		X		
	peak-motifs [58]	X		X	X				X
	Cistrome	X	X	X	X	X	X	X	X

Motif databases

- MotifMap: genome-wide maps of regulatory elements
- Databases for known motifs: TRANSFAC and JASPAR

<http://motifmap.ics.uci.edu/>

<http://gene-regulation.com/pub/databases.html>

<http://jaspar.genereg.net/>

<https://gimmemotifs.readthedocs.io/en/master/>

Summary

- ChIP-seq detects TFBS or measure histone modifications along the genome.
- Peak (short and long) detection is the major goal of data analysis.
- Motif finding helps to understand master regulators of gene expression.
- Many similar technologies and the method are more or less the same.