ChIP-seq

Mikhail Dozmorov

Spring 2018

Outline

Introduction to ChIP-seq experiment

- Biological motivation
- Experimental procedure

Methods and software for ChIP-seq peak calling

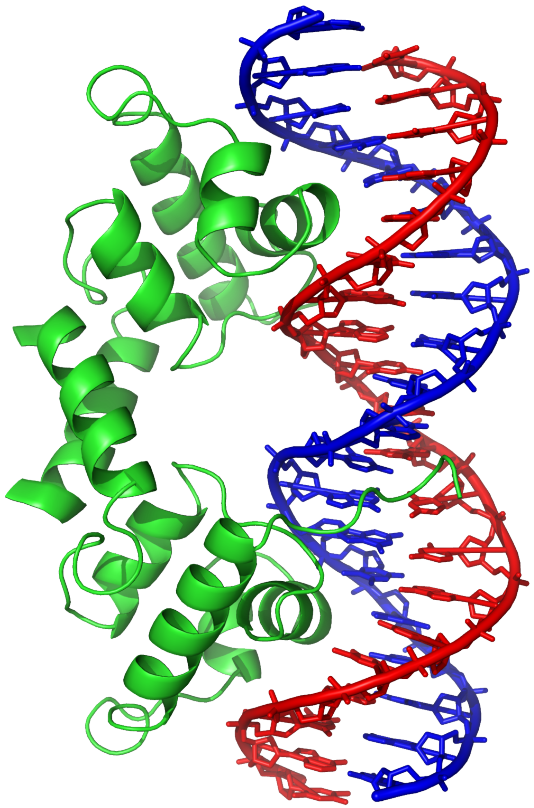
- Protein binding ChIP-seq
- Histone modifications

Higher order ChIP-seq data analysis

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

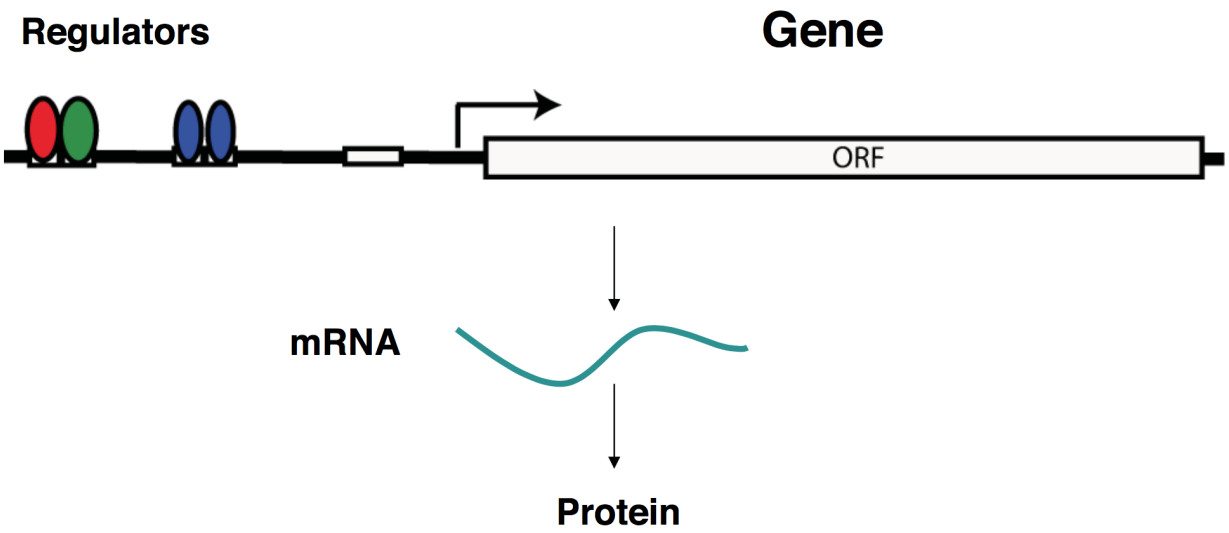
Transcription factors regulate gene expression

- 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

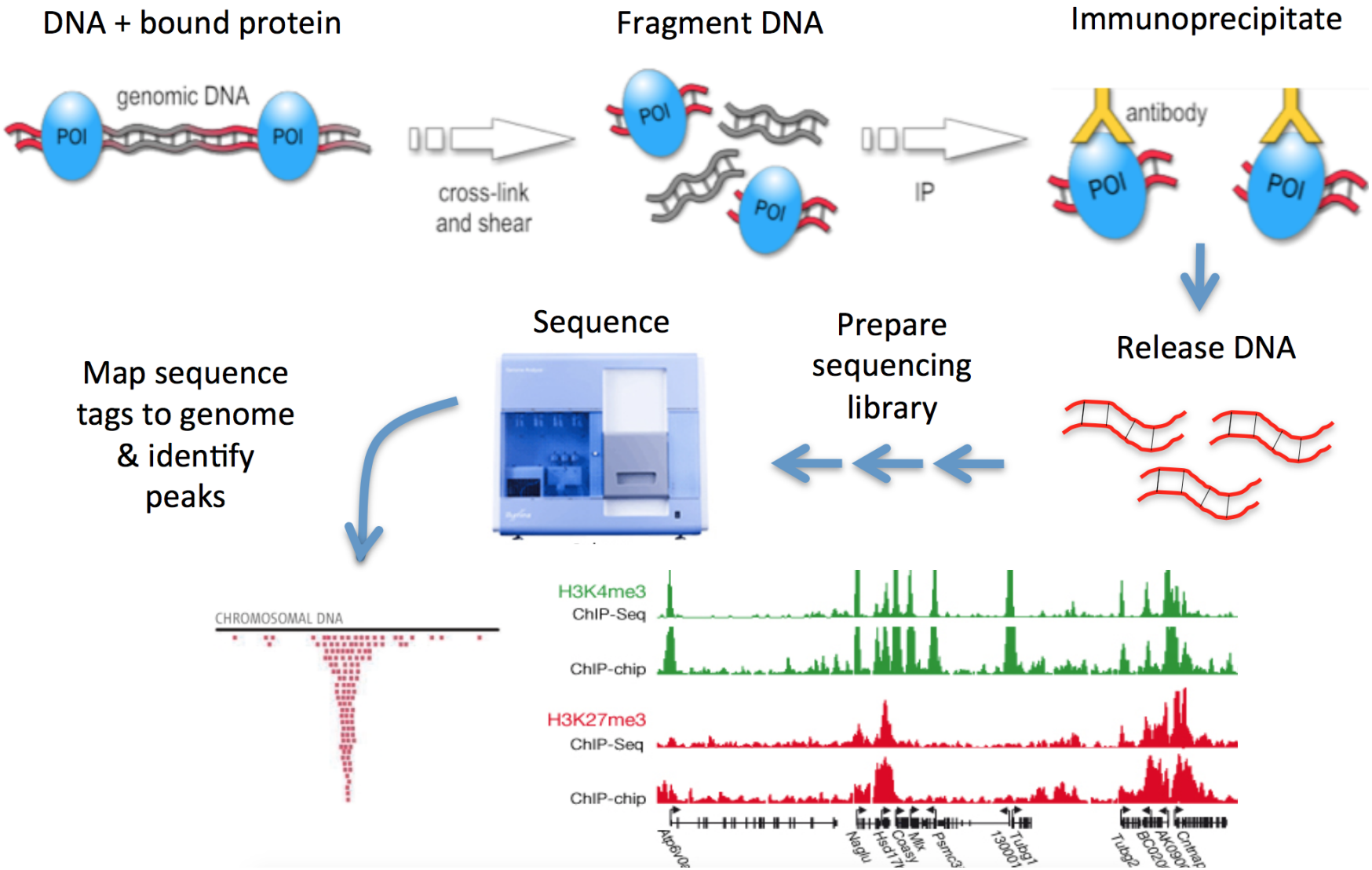
Combine high-throughput sequencing with Chromatin Immunoprecipitation to identify specific protein-DNA interactions genome-wide, including those of:

- Transcription factors
- Histones (various types and modifications)
- RNA Polymerase (survey of transcription)
- DNA polymerase (investigate DNA replication)
- DNA repair enzymes
- 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 antibody to capture DNA segments with specific proteins.
4. **Reverse crosslink**: remove protein from DNA.
5. **Sequence** the DNA segments.

ChIP-seq overview



Advantages of ChIP-seq over ChIP-chip

	ChIP–chip	ChIP–seq
Maximum resolution	Array-specific, generally 30–100 bp	Single nucleotide
Coverage	Limited by sequences on the array; repetitive regions are usually masked out	Limited only by alignability of reads to the genome; in with read length; many repetitive regions can be cove
Cost	US$400–800 per array (1–6 million probes); multiple arrays may be needed for large genomes	Currently US$1,000–2,000 per lane (using the Illumina Genome Analyzer); 6–15 million reads before alignme
Source of platform noise	Cross-hybridization between probes and nonspecific targets	Some GC bias can be present
Experimental design	Single- or double-channel, depending on the platform	Single channel
Cost-effective cases	Profiling of selected regions; when a large fraction of the genome is enriched for the modification or protein of interest (broad binding)	Large genomes; when a small fraction of the genome i enriched for the modification or protein of interest (sharp binding)
Required amount of ChIP DNA	High (a few micrograms)	Low (10–50 ng)
Dynamic range	Lower detection limit; saturation at high signal	Not limited
Amplification	More required	Less required; single-molecule sequencing without amplification is available
Multiplexing	Not possible	Possible

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

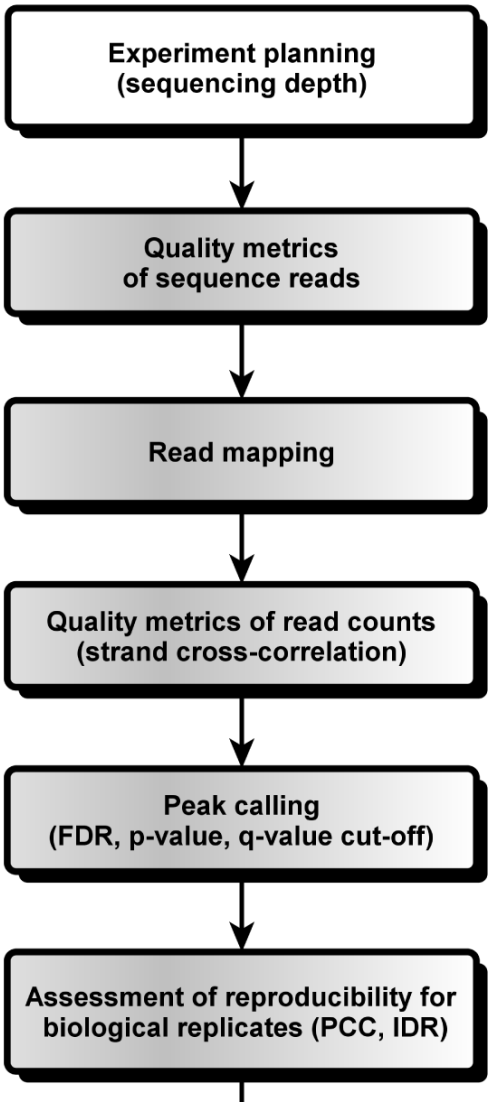
Bailey, Timothy, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang. "Practical Guidelines for the Comprehensive Analysis of ChIP-Seq Data." Edited by Fran Lewitter. PLoS Computational Biology 9, no. 11 (November 14, 2013): e1003326. https://doi.org/10.1371/journal.pcbi.1003326.

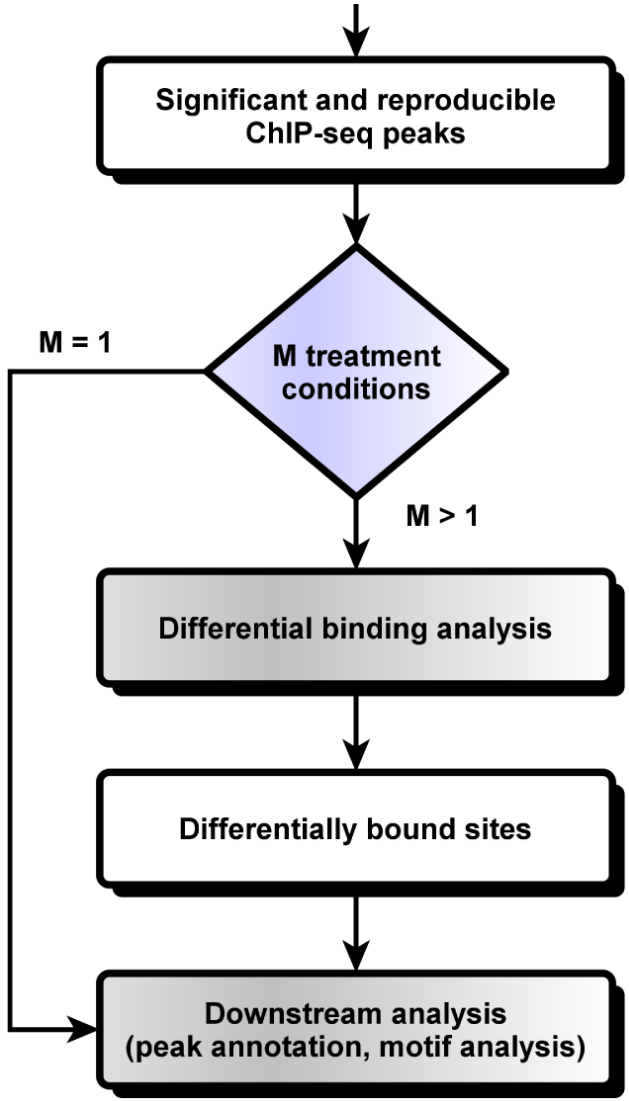
Data from ChIP-seq

- 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 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)

Methods and software for ChIP-seq peak/block calling

ChIP-seq analysis workflow



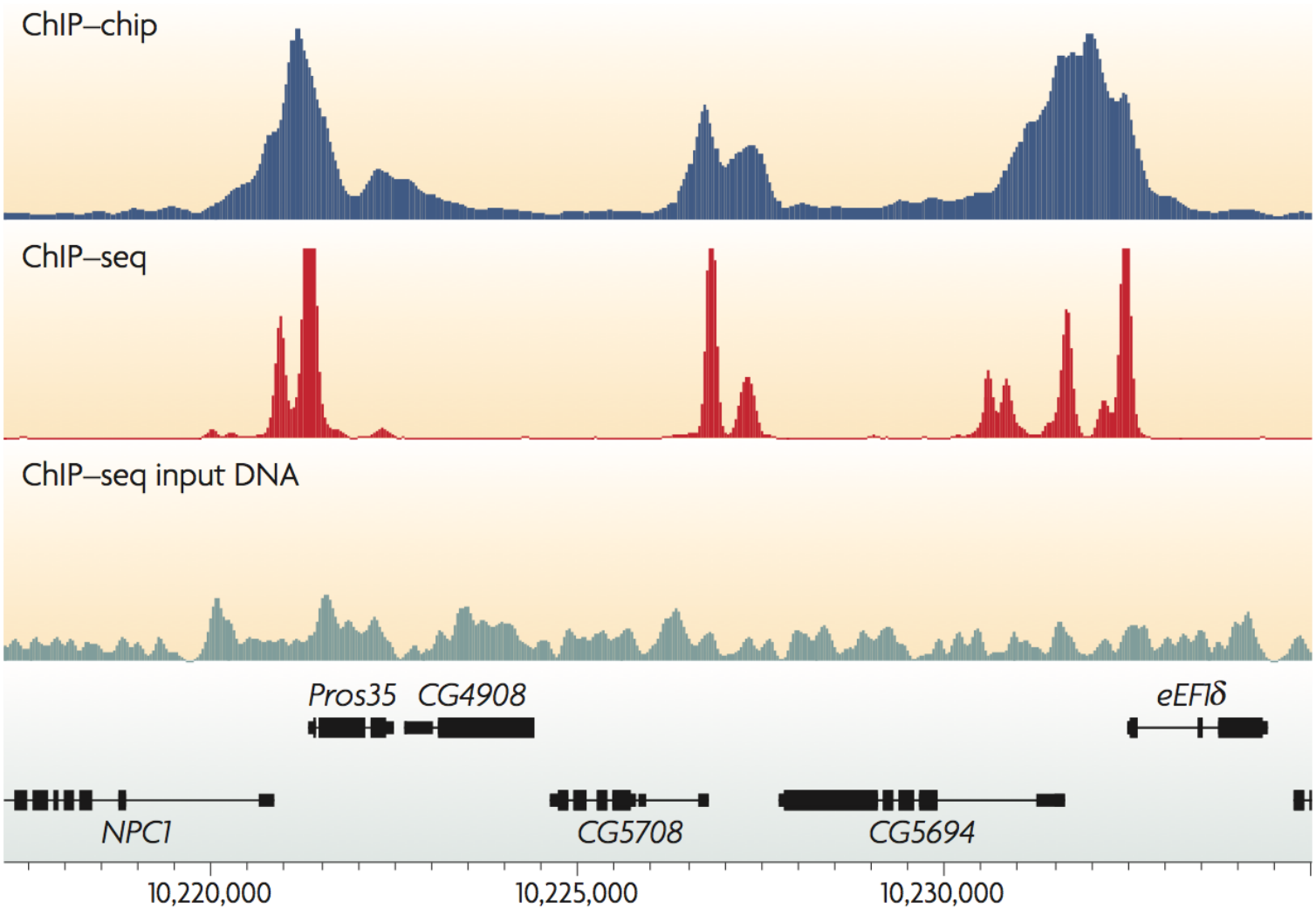


ChIP-seq "peak" detection

- 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

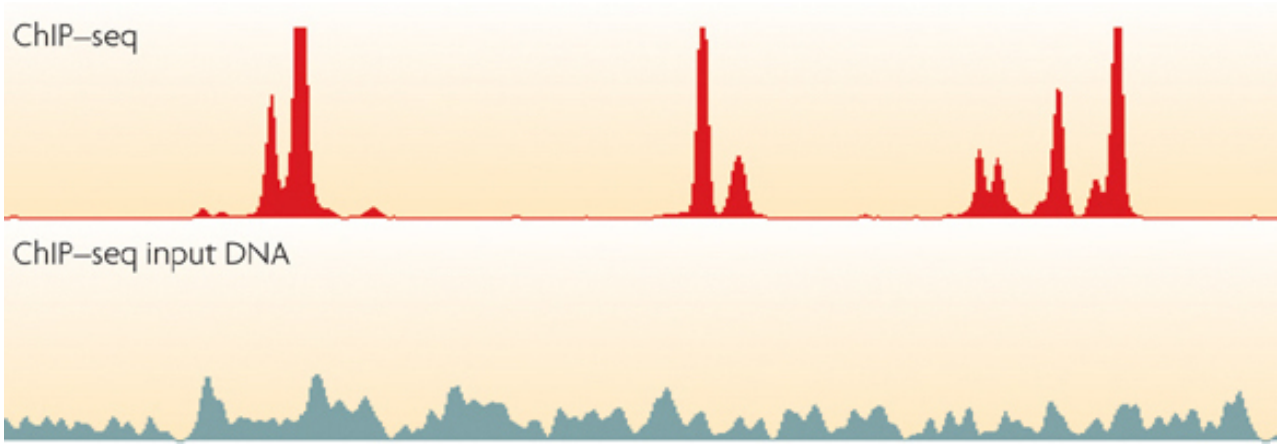
- Peaks are regions with reads clustered, so they can be detected from binned read counts
- Counts from neighboring windows need to be combined to make inference (so that it's more robust)
- To combine counts:
  - Smoothing based: moving average (`MACS`, `CisGenome`), HMM-based (`Hpeak`)
  - Model clustering 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

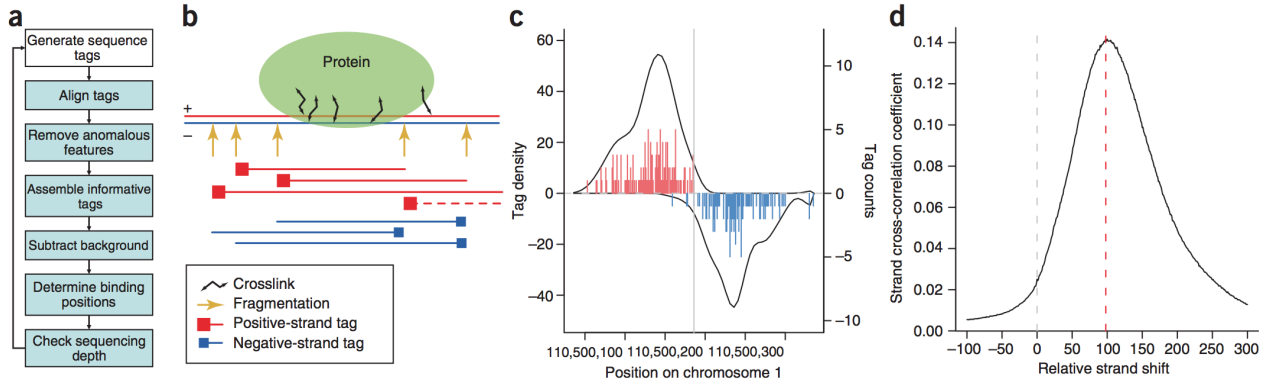
Table 1 Genome mappability fraction

		Nonrepetitive sequence		Mappable sequence	
Organism	Genome size (Mb)	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%

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?

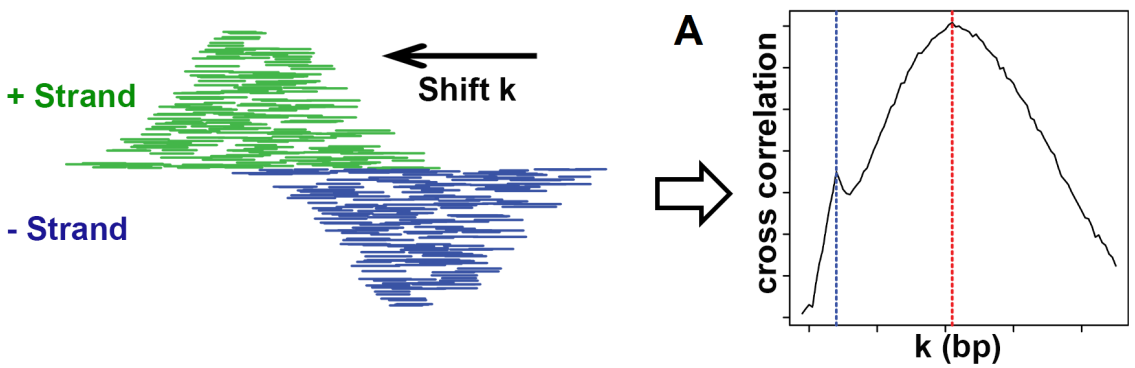
- Fragments contain the TF binding site at a (mostly) random position within them
- Reads are randomly generated from left or right edges (sense or antisense) of fragments
- Binding site position = mid-way between sense tag peak and antisense tag peak
- To get binding site peak, shift sense downstream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

Strand cross-correlation

- A high-quality ChIP-seq experiment often shows a significant clustering of enriched DNA sequence tags at the locations bound by the protein
- The enriched sequence tags on the forward and reversed strands are positioned at a distance from the binding site center that depends on the fragment size distribution
- The cross-correlation between the two strands, i.e., the Pearson correlation between the strand-specific read density profiles as a function of the shift (k) applied to one of the two strands



Bailey, Timothy, Pawel Krajewski, Istvan 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

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 software

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

Software tool	Version	Availability	Point- source (peaks)	Broad regions (domains)
BayesPeak [88]	1.10.0	http://bioconductor.org/packages/release/bioc/html/BayesPeak.html	Yes	
BEADS§ [84]	1.1	http://beads.sourceforge.net/	Yes	Yes
CCAT [91]	3	http://cmb.gis.a-star.edu.sg/ChIPSeq/paperCCAT.htm		Yes
CisGenome [56]	2	http://www.biostat.jhsph.edu/~hji/cisgenome/	Yes	
CSAR [85]	1.10.0	http://bioconductor.org/packages/release/bioc/html/CSAR.html	Yes	
dPeak	0.9.9	http://www.stat.wisc.edu/~chungdon/dpeak/	Yes	
GPS/GEM [67,18]	1.3	http://cgs.csail.mit.edu/gps/	Yes	
HPeak [87]	2.1	http://www.sph.umich.edu/csg/qin/HPeak/	Yes	
MACS [17]	2.0.10	https://github.com/taoliu/MACS/	Yes	Yes
NarrowPeaks*a*	1.4.0	http://bioconductor.org/packages/release/bioc/html/NarrowPeaks.html	Yes	
PeakAnalyzer/ PeakSplitter*a* [89]	1.4	http://www.bioinformatics.org/peakanalyzer	Yes	
PeakRanger [93]	1.16	http://ranger.sourceforge.net/	Yes	Yes
PeakSeq [24]	1.1	http://info.gersteinlab.org/PeakSeq	Yes	
polyaPeak*a*	0.1	http://web1.sph.emory.edu/users/hwu30/polyaPeak.html	Yes	
RSEG [92]	0.6	http://smithlab.usc.edu/histone/rseg/		Yes
SICER [90]	1.1	http://home.gwu.edu/~wpeng/Software.htm		Yes
SIPeS [21]	2	http://gmdd.shgmo.org/Computational-Biology/ChIP-Seq/download/SIPeS	Yes	

Peak detection

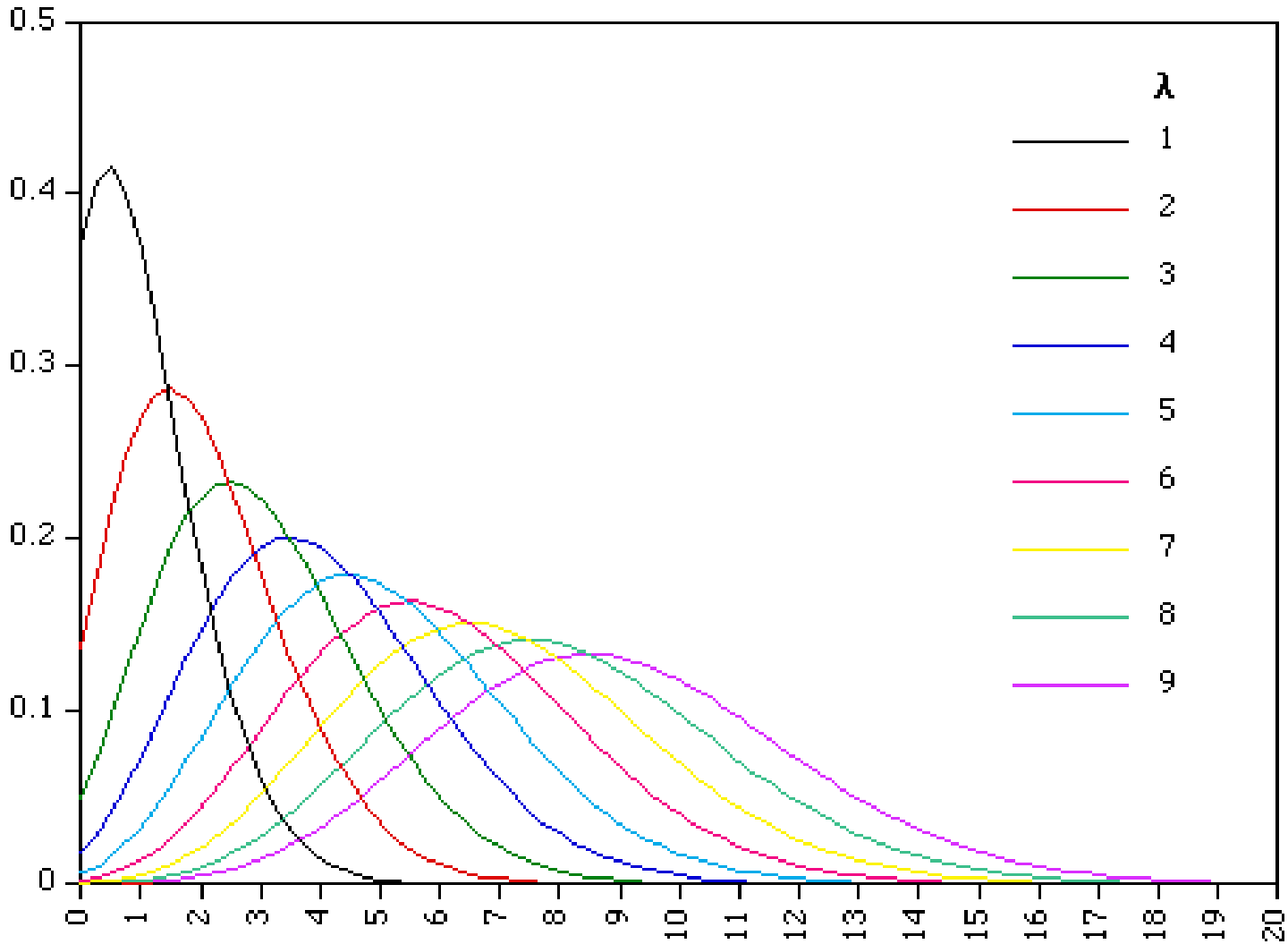
- Calculate read count at each position (bp) in genome. Don't use a sliding window
- Determine if read count is greater than expected (at each position, bp)
- We need to correct for input DNA reads (control)
  - non-uniformly distributed
  - vastly different numbers of reads between ChIP and input

The Poisson distribution: discrete distribution to model coverage

- $P(k \textit{ discrete events}) = \frac{\lambda^k e^{-\lambda}}{k!}$ 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

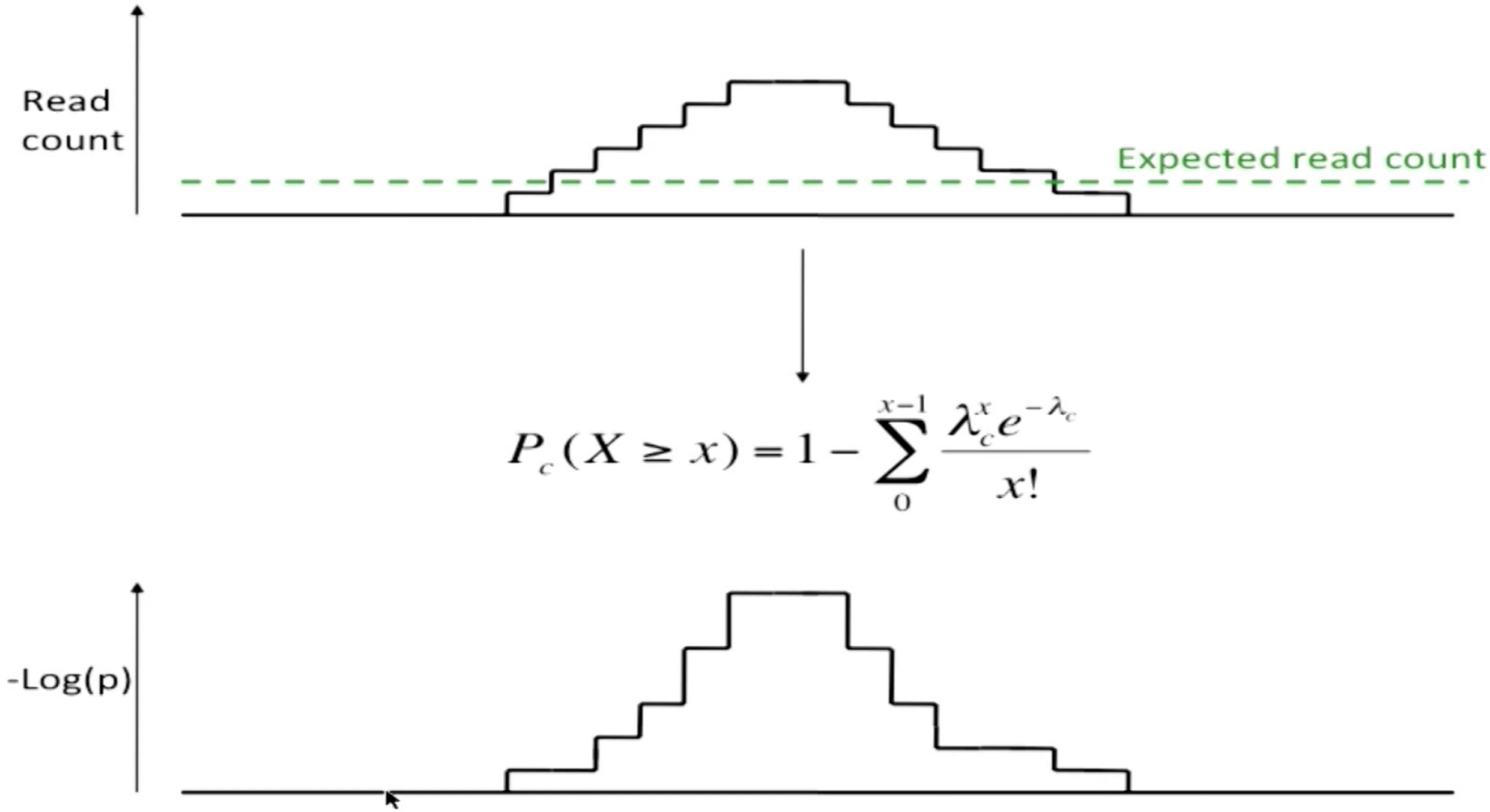
$$P(X \geq x) = 1 - \sum_{0}^{x-1} \frac{\lambda^x e^{-\lambda}}{x!}$$

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

Example:

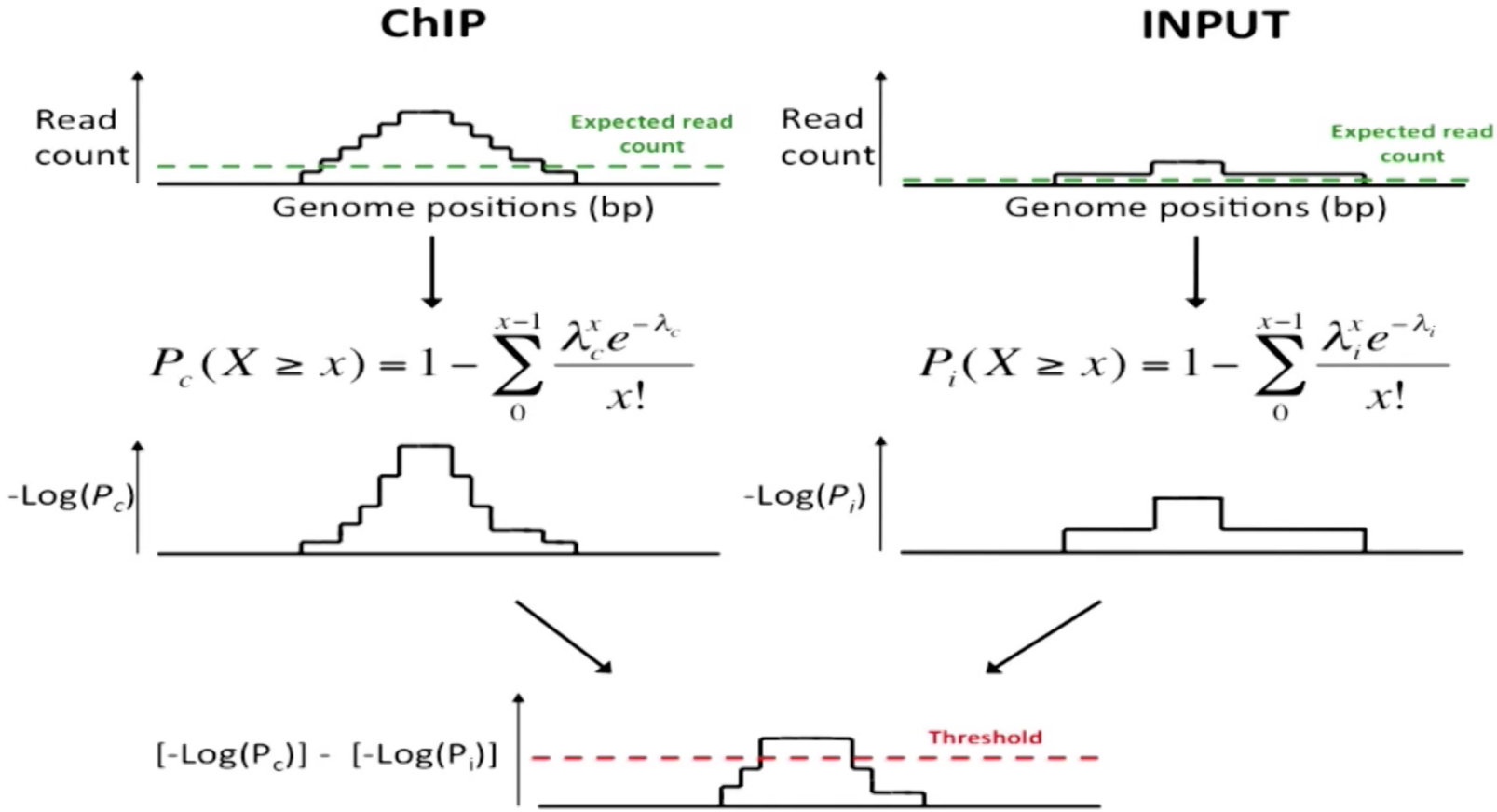
- $x = 10$ reads, observed
- $\lambda = 0.5$ reads, expected
- $P(X \geq 10) = 1.7x10^{-10}$
- $-log_{10}P(X \geq 10) = 9.77$
- In R, $P(X \geq 10|\lambda = 0.5)$ is `1 - sum(dpois(0:9, 0.5))`

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



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

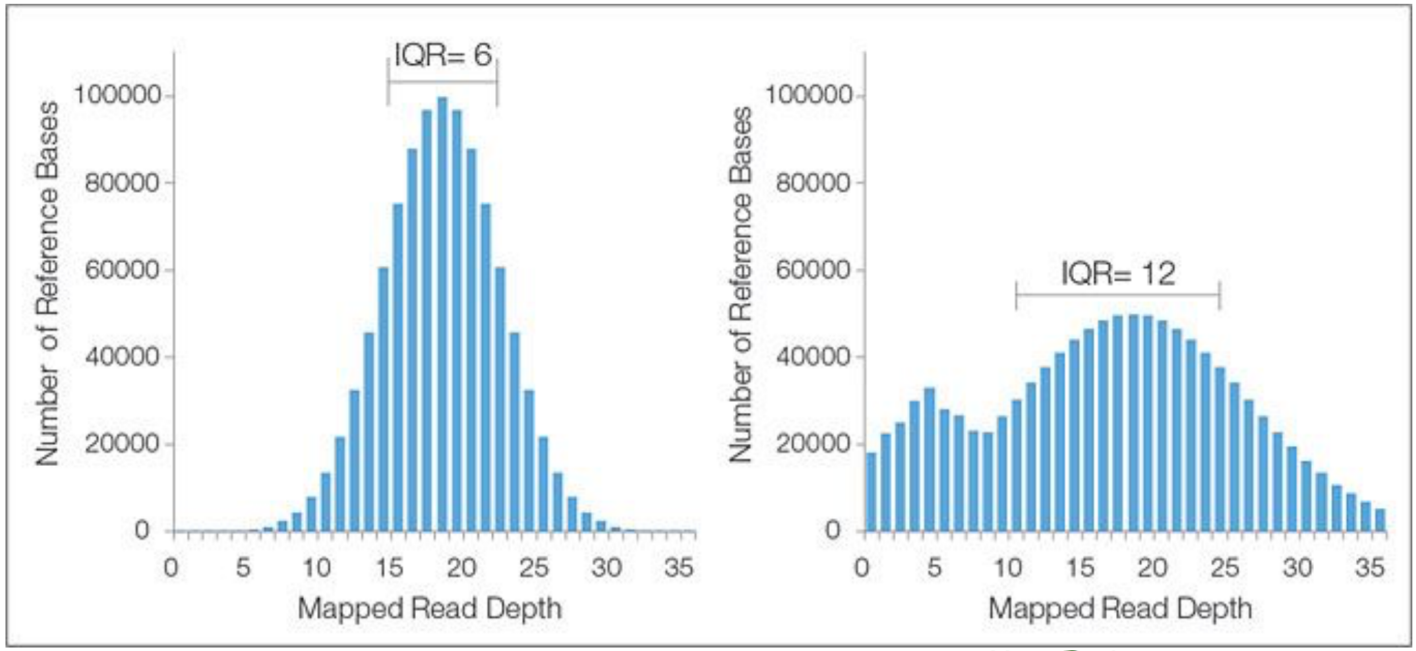
$$R = -log_{10} \frac{P(X_{ChIP})}{P(X_{Input})}$$

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...



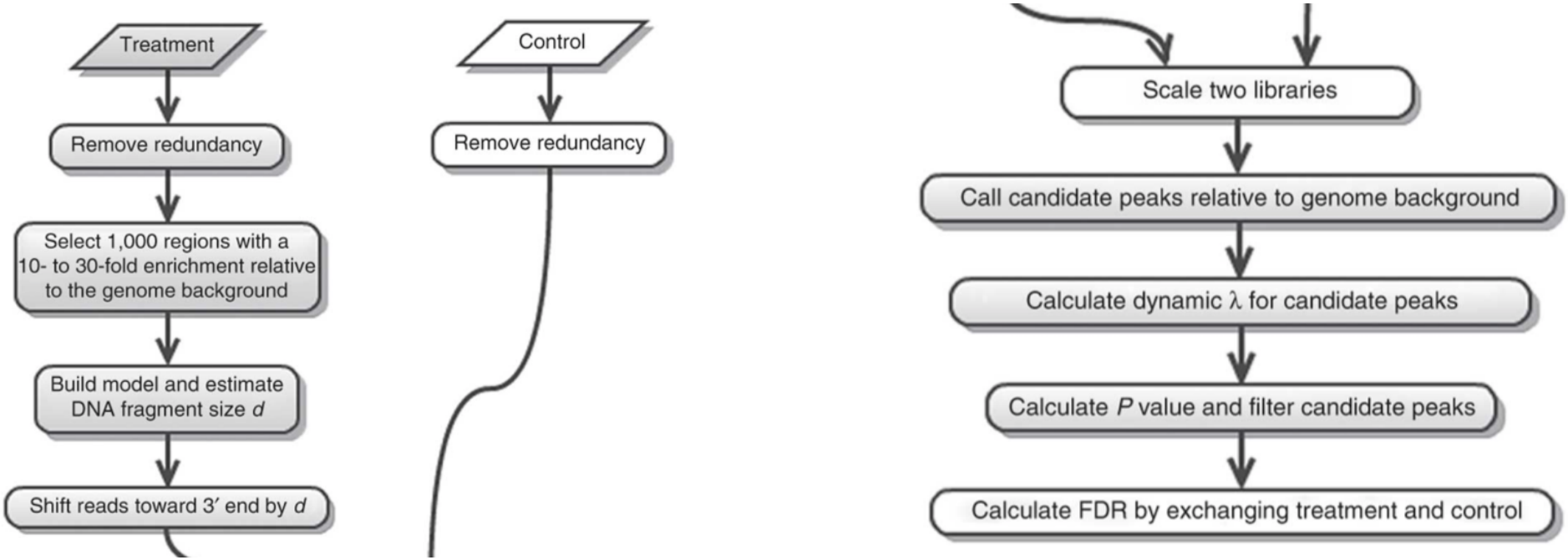
Poisson

Not Poisson.
Overly "dispersed"

- 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 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 Possion model to scan genome and score peaks.
  - Counts in a window are assumed to following Poisson distribution with rate : $\lambda_{local} = max(\lambda_{BG}, \lambda_{1K}, \lambda_{5K}, \lambda_{10K})$ where $lambda_{1K}$, $lambda_{5K}$ and $lambda_{10K}$ 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. 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:

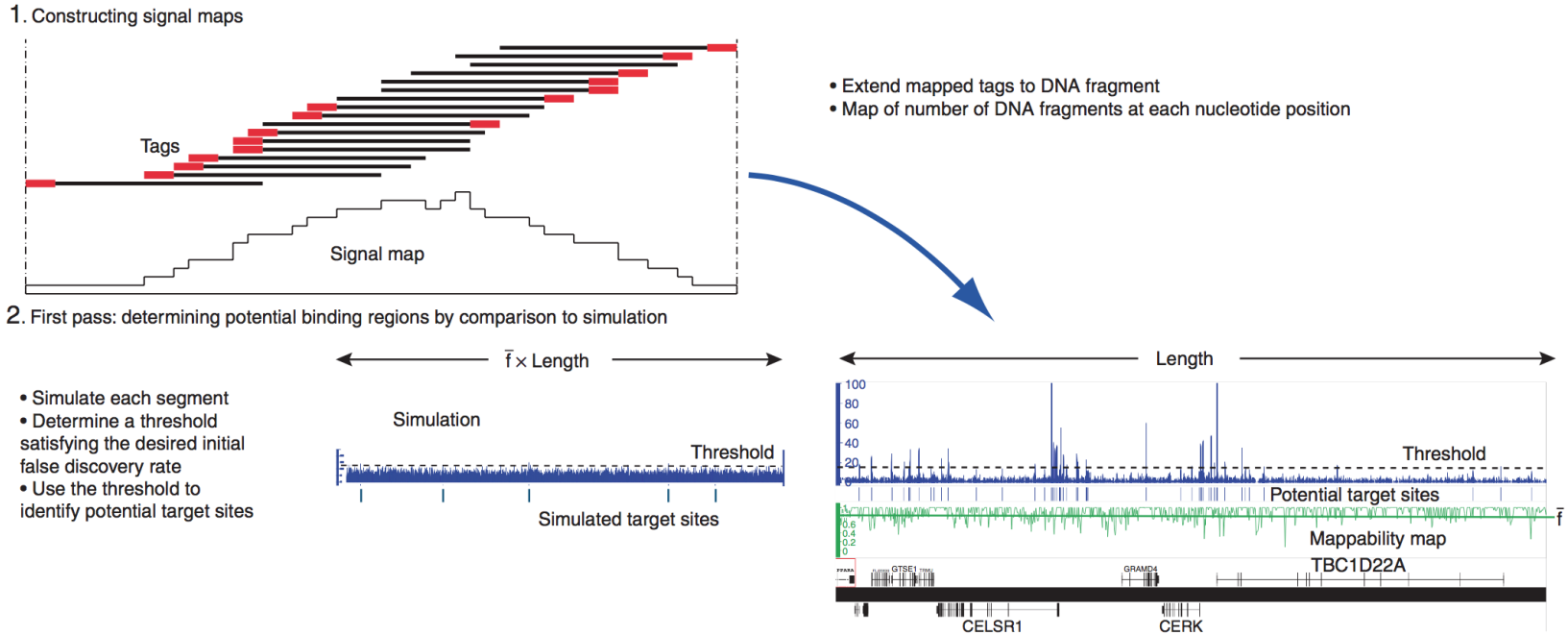
- Adjust the *mfold* values (the fold over background ranges MACs considers for paired peaks)
- Tell MACS to not build a model, but instead use the shiftsize you specify.

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

Consider mappability: PeakSeq

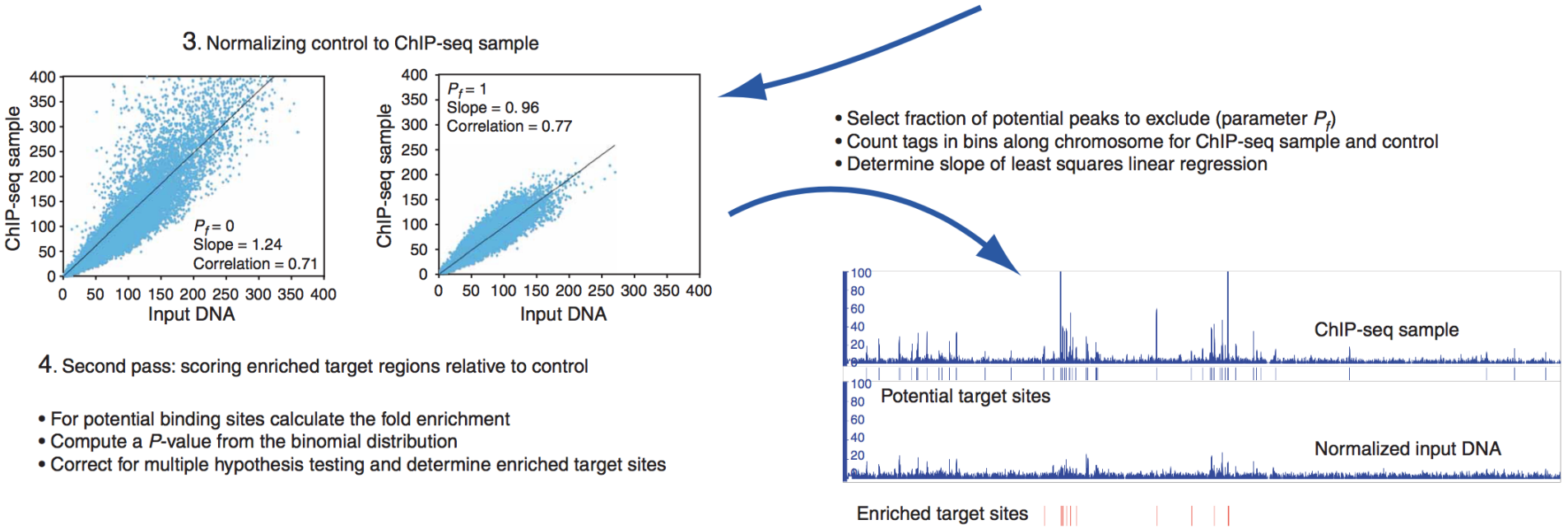
- Two-pass analysis. First round - detect possible peak regions by identifying threshold while considering mappability
  - Cut genome into segments ($L = 1Mb$). Within each segment, the same number of reads are permuted in a region of $f \times L$, there f is the proportion of mappable bases in the segment



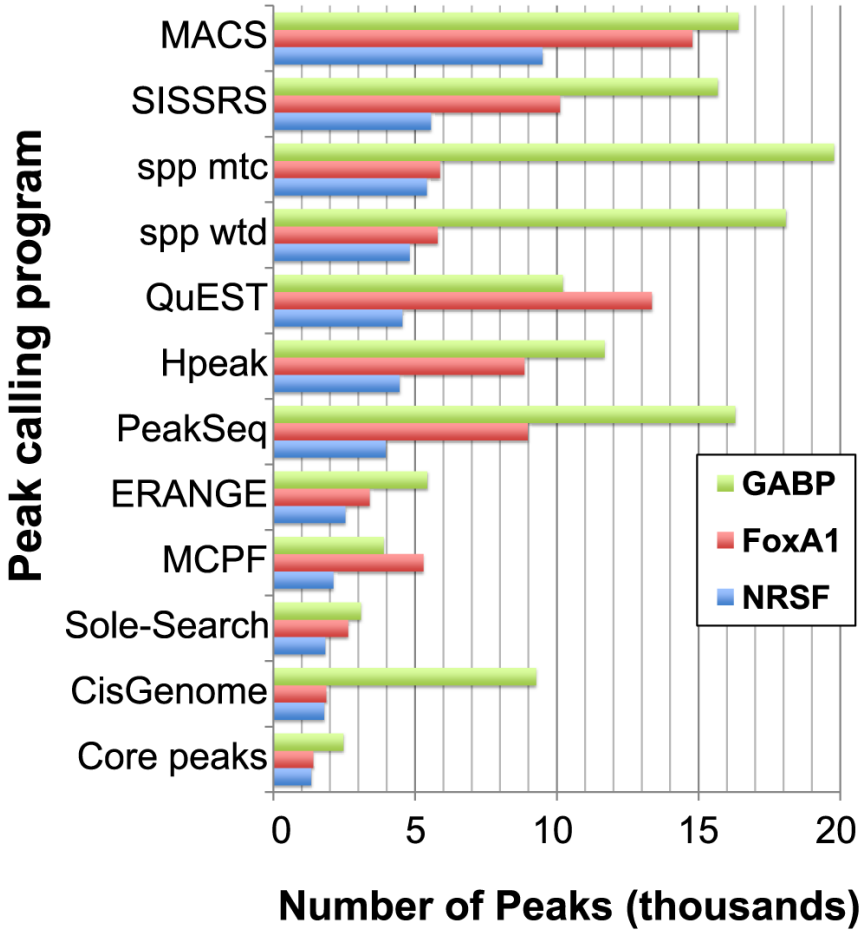
Rozowsky et.al. (2009) NBT https://www.nature.com/nbt/journal/v27/n1/full/nbt.1518.html

Consider mappability: PeakSeq

- 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 regions with control data, using binomial p-value with Benjamini-Hochberg correction



Comparing peak calling algorithms



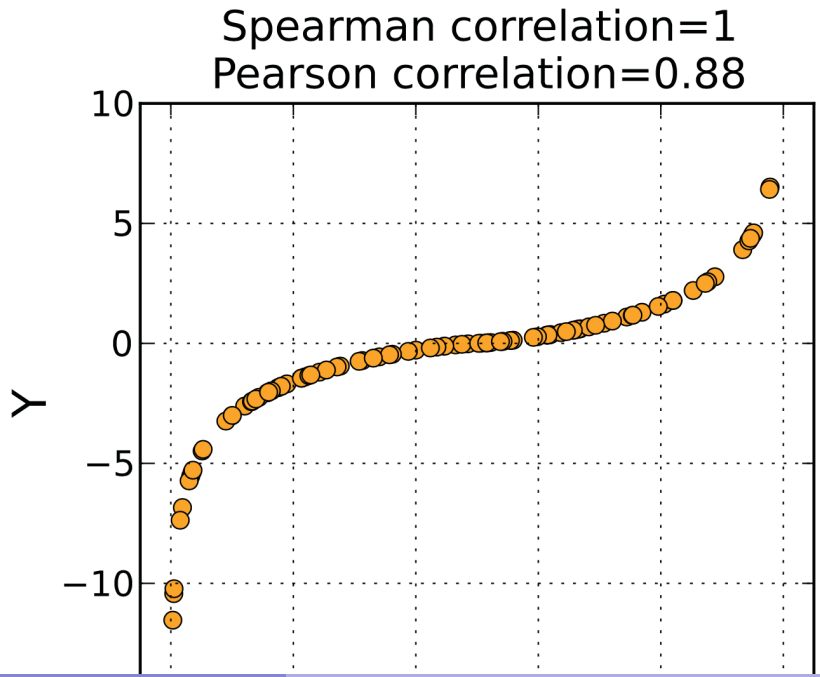
Wilbanks et.al. (2010) PloS One http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0011471

Laajala et.al. (2009) BMC Genomics https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-10-618

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

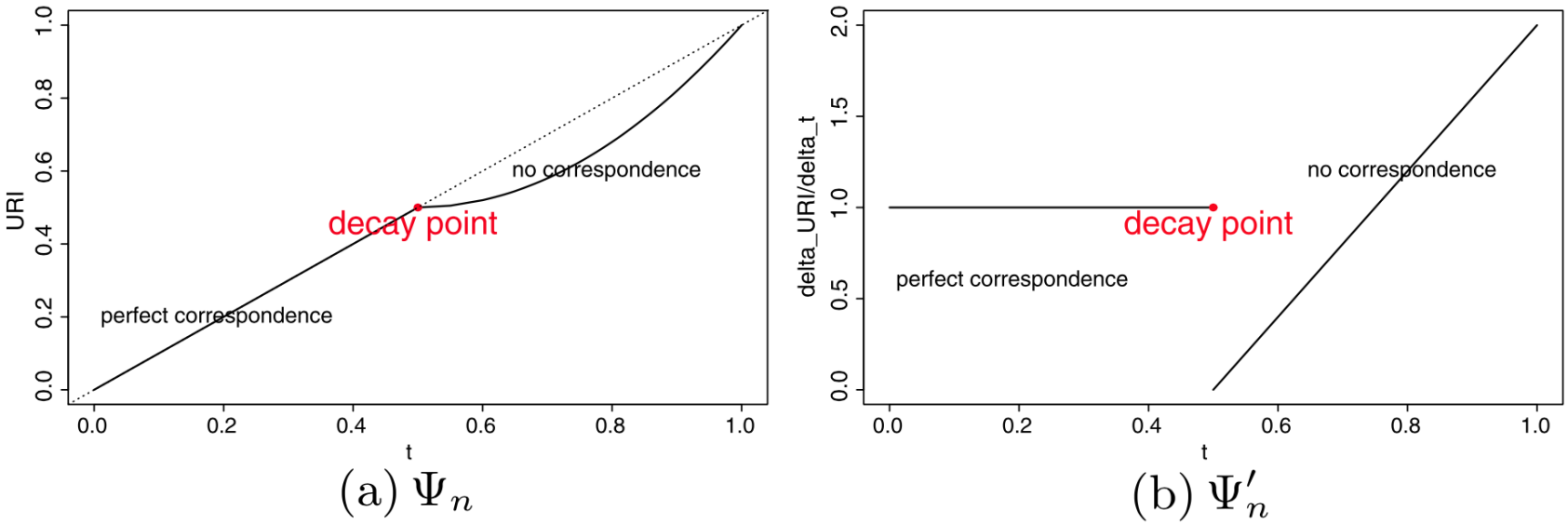
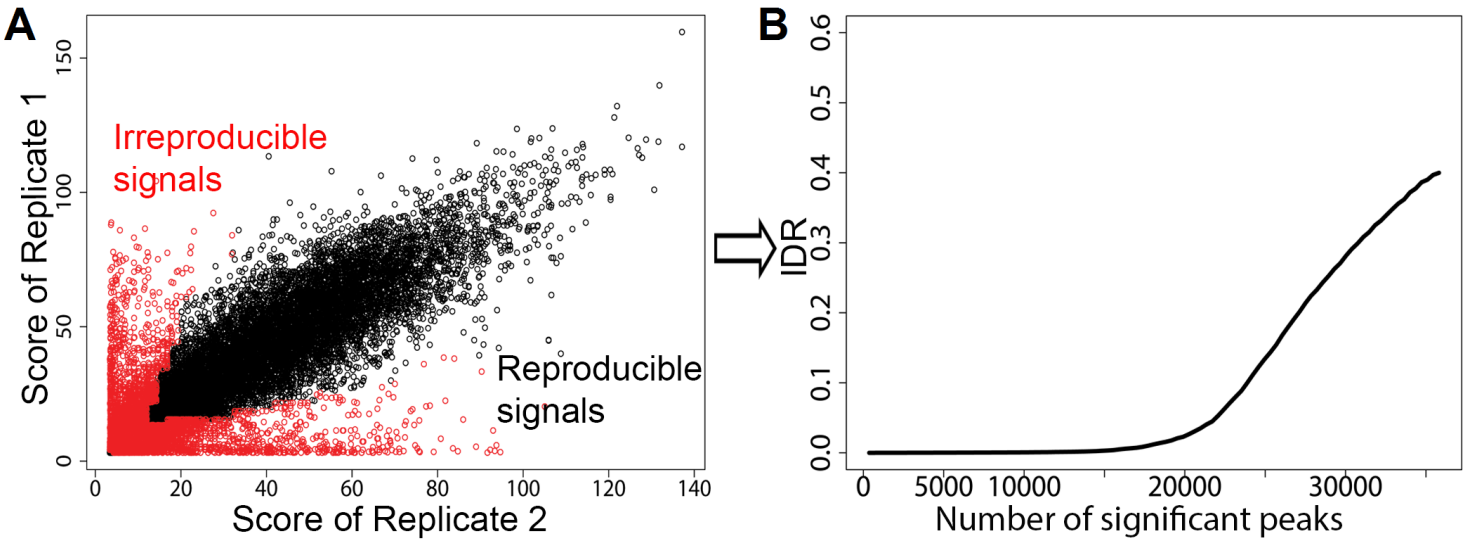


FIG. 1. *An illustration of the correspondence profile in an idealized case, where top* 50% *are genuine signals and bottom* 50% *are noise. In this case, all signals are ranked higher than noise; two rank lists have perfect correspondence for signals and no correspondence for noise.* (a) *Correspondence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

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.

The irreproducible discovery rate (IDR) framework for assessing reproducibility of ChIP-seq data sets



Panel A shows a scatterplot of the significance scores of peaks identified in two replicate ChIP-seq experiments. The IDR method classifies peaks into reproducible (black) and irreproducible (red) groups, and computes for each peak the probability that the peak belongs to the irreproducible group. It ranks and selects peaks according to this probability, and computes IDR, the expected rate of irreproducible discoveries in the selected peaks. Panel B shows the estimated IDR at different rank thresholds when the peaks are sorted by the original significance score.

Bailey, Timothy, Pawel Krajewski, Istvan 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

Irreproducible Discovery Rate Results

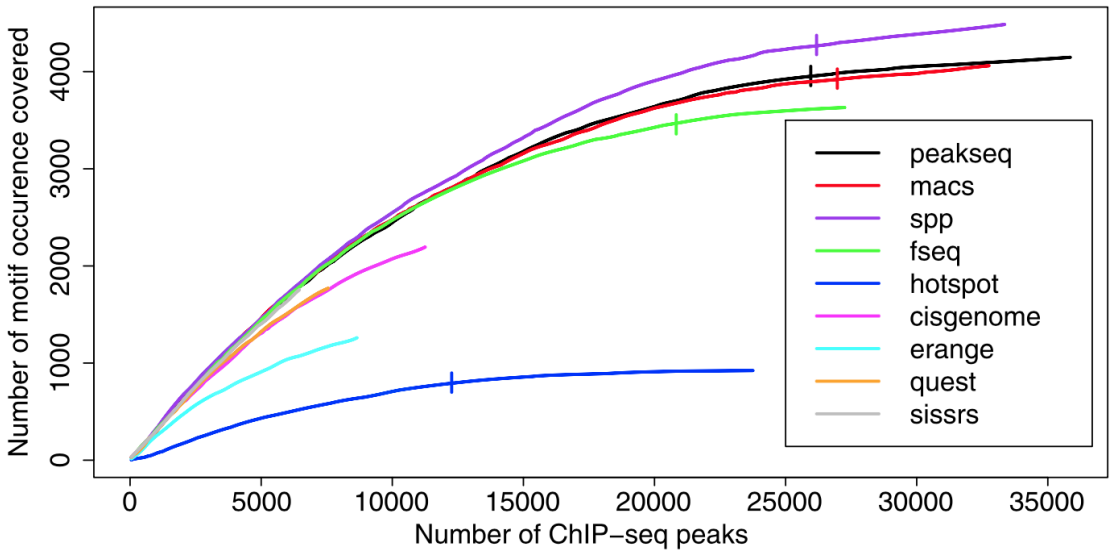


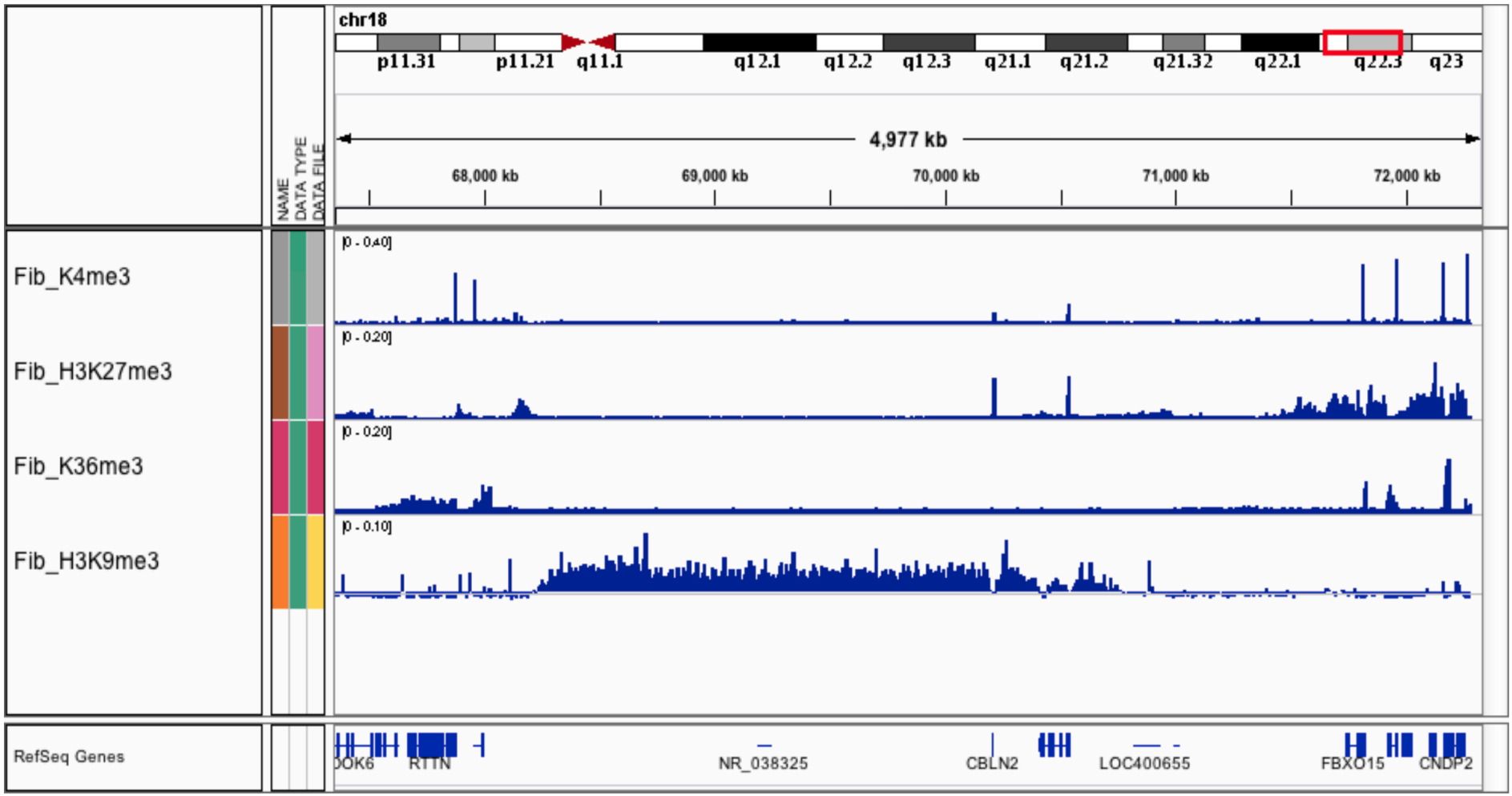
FIG. 7. *The coverage of high-confidence CTCF motif at different numbers of selected ChIP-seq peaks, plotted at various idr cutoffs for nine peak callers on a CTCF Chip-seq experiment from ENCODE. The bars on the curves of Peakseq, MACS, SPP, Fseq and Hotspot show the number of peaks selected at IDR* $= 0.05$. *No selection is made for the rest of the peak callers because model selection favors the one-component model for peaks identified by these callers.*

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.

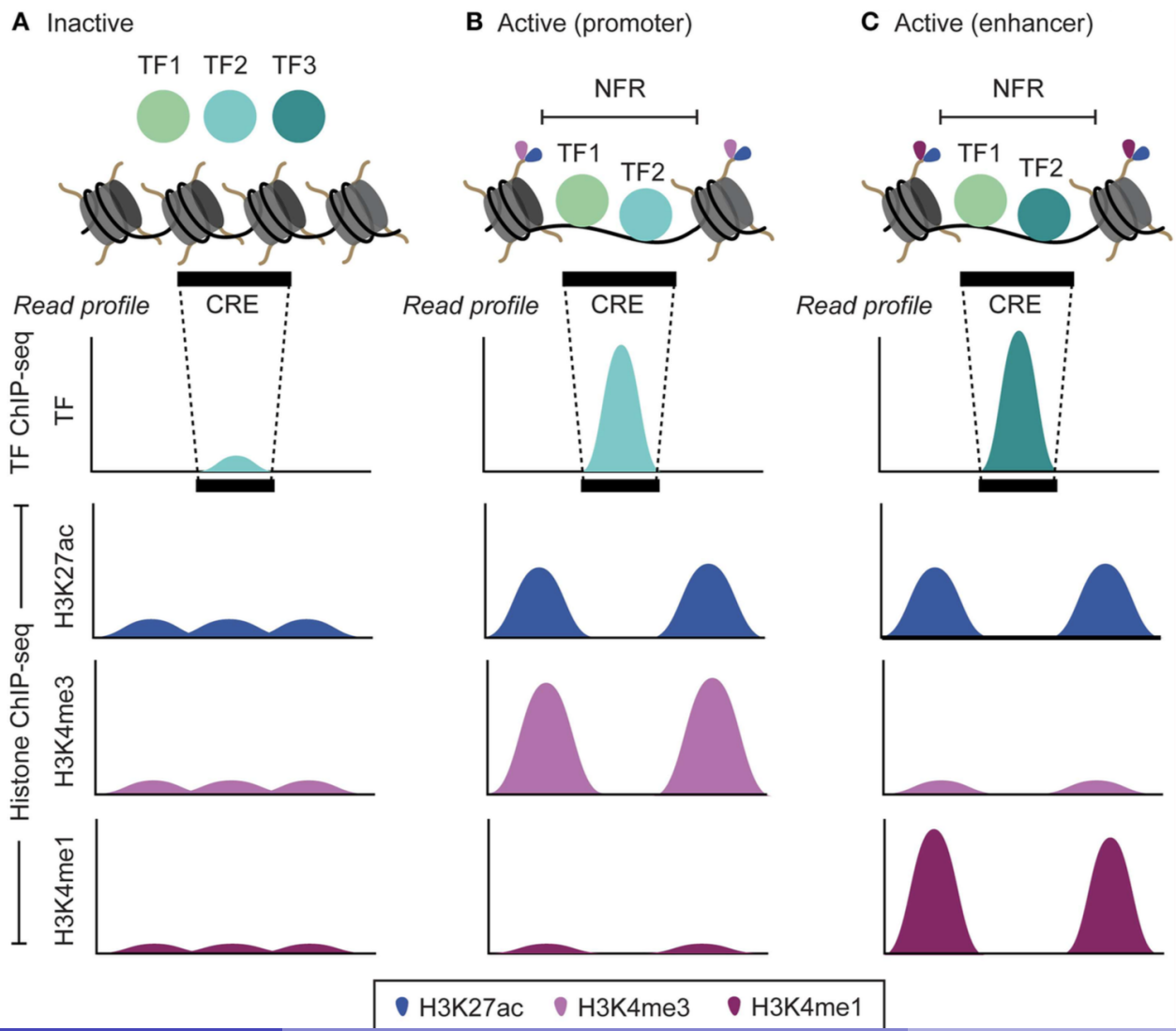
ChIP-seq for histone modification

- 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 (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 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 post hoc approach to combine nearby peaks.
- Syntax:

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

SICER

Algorithm:

- Cut genome of length L into non-overlapping windows w and compute a score s for each window with I reads out of N total based on a Poisson model, $s(I) = -logP(I, \lambda)$, $\lambda = wN/L$
- Identify "islands" vs. "non-islands" by thresholding the scores and clustering windows with significant scores
- For each island, compute the probability of observing the island with a given score. Constructing score distribution is involved

Zang, Chongzhi, Dustin E. Schones, Chen Zeng, Kairong Cui, Keji Zhao, and Weiqun Peng. "A Clustering Approach for Identification of Enriched Domains from Histone Modification ChIP-Seq Data." Bioinformatics 25, no. 15 (August 1, 2009): 1952–58. https://doi.org/10.1093/bioinformatics/btp340.
https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp340

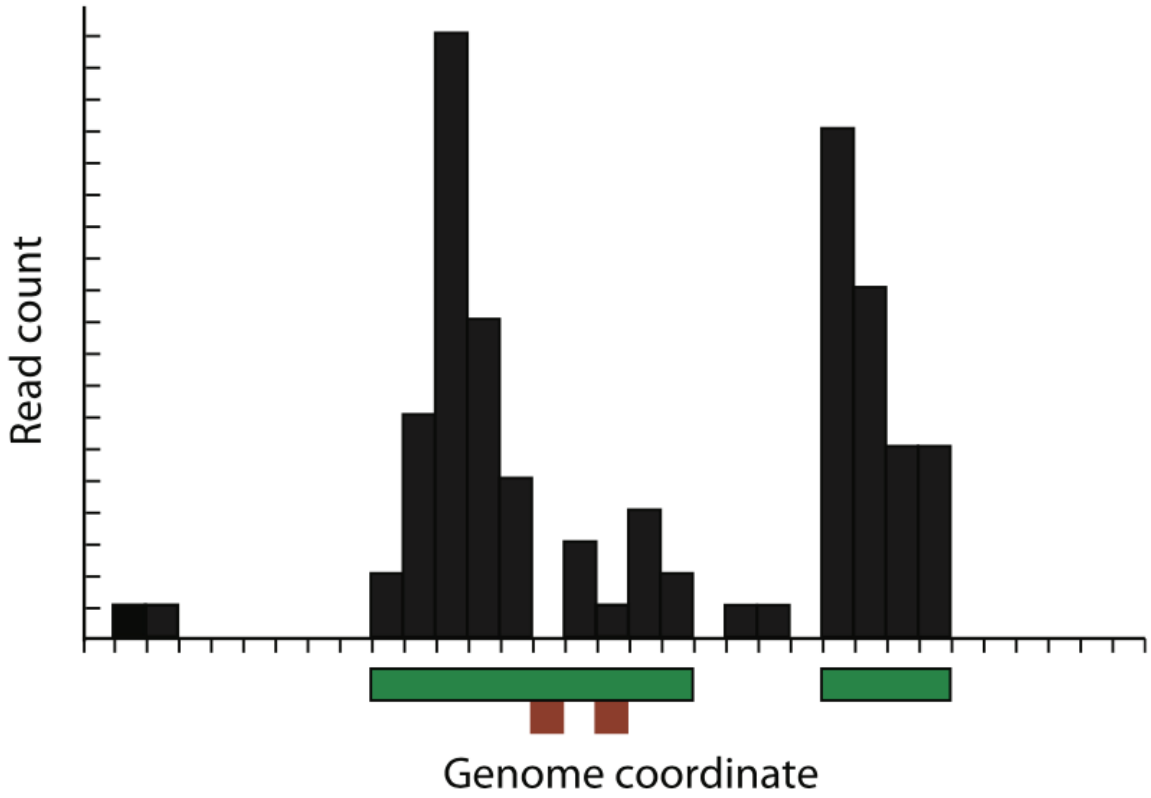
http://home.gwu.edu/~wpeng/Software.htm

SICER: Definition of island

- Eligible and ineligible windows

$$\sum_{l=l_0}^{\infty} P(l, \lambda) \leq p_0$$

- Eligible windows are separated by gaps of ineligible windows
- Island - cluster of eligible windows separated by gaps of size at most g windows



Example islands for read count threshold $l_0 = 2$ and number of gaps $g = 2$

SICER: Scoring islands

- The scoring function is based on the probability of finding the observed tag count in a random background
- For a window with m reads,
  - The probability of finding m reads is Poisson $P(m, \lambda)$
  - $\lambda = wN/L$ is the average number of reads in each window (w - genome window size, L - genome length, N - total number of reads in the ChIP-seq library)
- Scoring function for an eligible window:

$$s = -logP(m, \lambda)$$

- Key quantity: the score of an island
  - Aggregate score of all eligible windows in the island
  - It corresponds to the background probability of finding the observed pattern

SICER: Island score statistics

- Probability distribution of scores for a single window in a random background model ($\delta(*)$ - Dirac delta function):

$$p(s) = \sum_{l \geq l_0} \delta(s - s(l)) P(l, \lambda)$$

- Probability of a window being 'ineligible':

$$t = P(0, \lambda) + P(1, \lambda) + ... + P(l_0 - 1, \lambda)$$

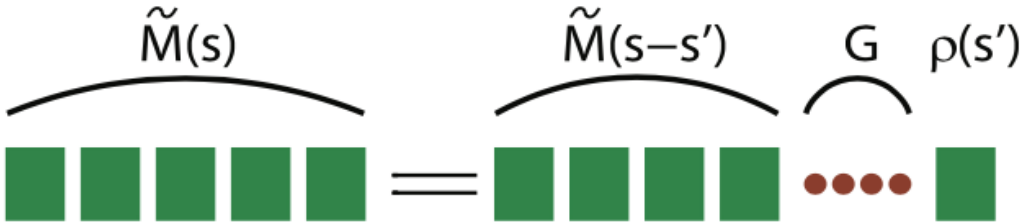
- The number of 'ineligible' windows in a gap ranges from zero to g. Gap factor:

$$G = 1 + t + t^2 + ... + t^g$$

SICER: Island score statistics

- Probability of finding an island of score s:

$$M(s) = t^{g+1} \tilde{M}(s) t^{g+1}$$



- The island score can be partitioned between the last 'eligible' window and the rest in a combinatorial manner, therefore a recursion relation can be constructed for the kernel $\tilde{M}(s)$:

$$\tilde{M}(s) = G(\lambda, l_0, g) \int_{s_0}^{s} \tilde{M}(s - s') p(s') ds'$$

SICER: Island score statistics

- Asymptotics of island score distribution in the random background

$$\tilde{M}(s) = \alpha exp^{(-\beta s)}$$

- Estimate β from

$$G(\lambda, I_0, g) \sum_{I \geq I_0} P(I, \lambda)^{1-\beta} = 1$$

- then, estimate α by fitting

SICER: Island score statistics

- Significance determination with random background model:
  - E-value determines an island score threshold
- Threshold score value s_T can be determined by requiring the expected number of islands with scores above the threshold s_T to be less than a E-value threshold e:

$$\sum_{s \geq s_T} LM(s) \leq e$$

Zang, Chongzhi, Dustin E. Schones, Chen Zeng, Kairong Cui, Keji Zhao, and Weiqun Peng. "A Clustering Approach for Identification of Enriched Domains from Histone Modification ChIP-Seq Data." Bioinformatics 25, no. 15 (August 1, 2009): 1952–58. https://doi.org/10.1093/bioinformatics/btp340.

SICER: Choosing parameters

- Fragment size
- Window size: data resolution
- Gap size
- Compared with other methods, SICER focuses on the clustered enrichment rather than local enrichment

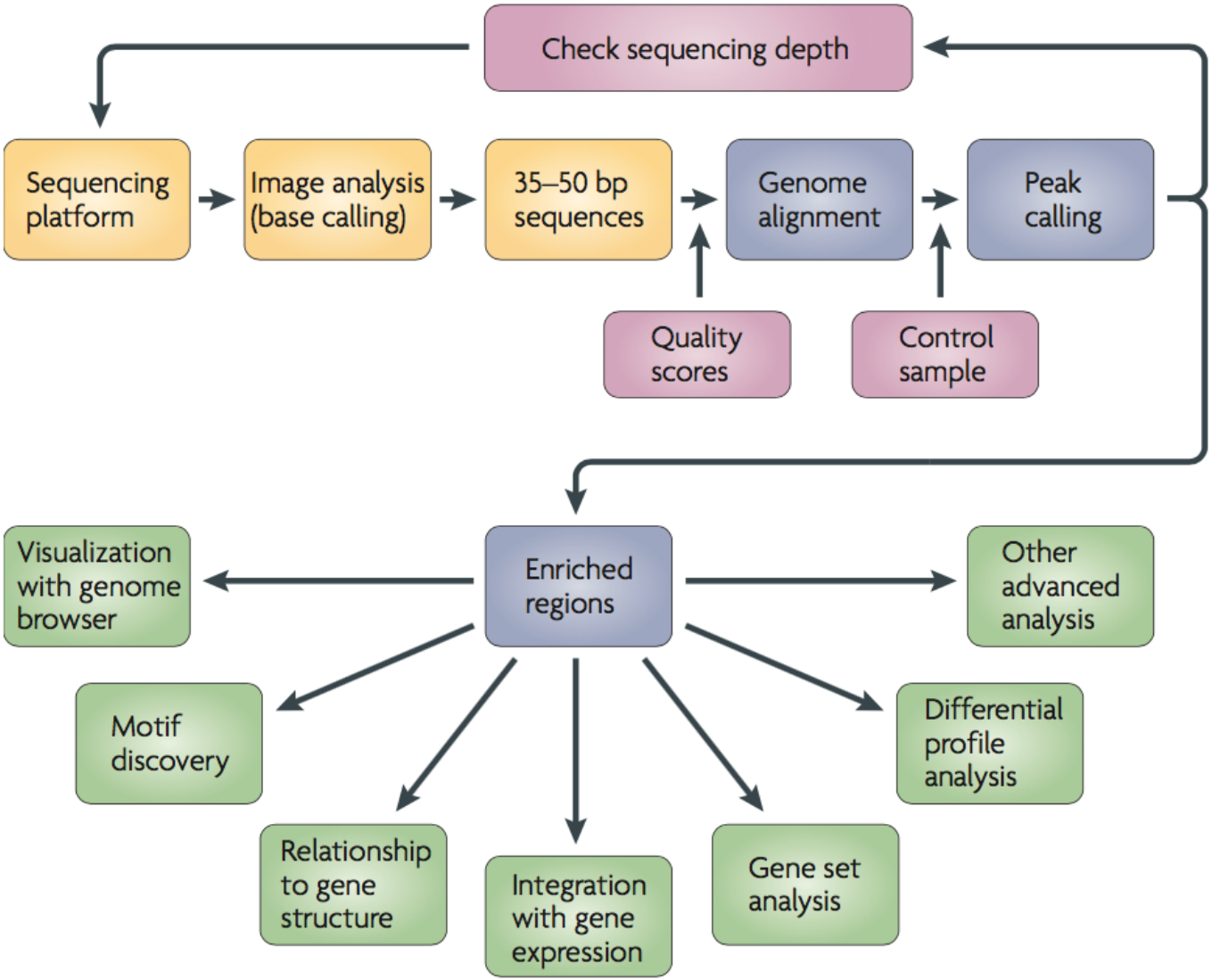
Use SICER

- 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`

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 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 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 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 peak
- **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

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/	Assigns uncertainty measures in a test of non-differential binding (uses edgeR)
DESeq§ [31]	http://www.bioconductor.org/packages/release/bioc/html/DESeq.html	Test based on a model using the negative binomial distribution
DiffBind	http://www.bioconductor.org/packages/release/bioc/html/DiffBind.html	Differential binding affinity analysis (uses edgeR and DESeq)
DIME [35]	http://cran.r-project.org/web/packages/DIME/	Differential identification using mixtures ensemble
edgeR§ [32]	http://www.bioconductor.org/packages/release/bioc/html/edgeR.html	Empirical Bayes estimation and exact tests based on the negative binomial distribution
MACS [17] (version 2)	https://github.com/taoliu/MACS/	Differential peak detection based on paired four bedGraph files
MAnorm [34]	http://bcb.dfci.harvard.edu/~gcyuan/MAnorm/MAnorm.htm	Robust regression to derive a linear model
MMDiff	http://bioconductor.org/packages/release/bioc/html/MMDiff.html	Differences in shape using Kernel methods
NarrowPeaks	http://bioconductor.org/packages/release/bioc/html/NarrowPeaks.html	Shape-based analysis of variation using functional PCA
POLYPHEMUS [37]	http://cran.r-project.org/web/packages/polyphemus/	Non-linear normalization on RNA Pol II profiling

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

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 paderns 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:

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 $A = \{a_1, a_2, ..., a_N\}$
- We seek within each sequence mutually similar segments of specified width W

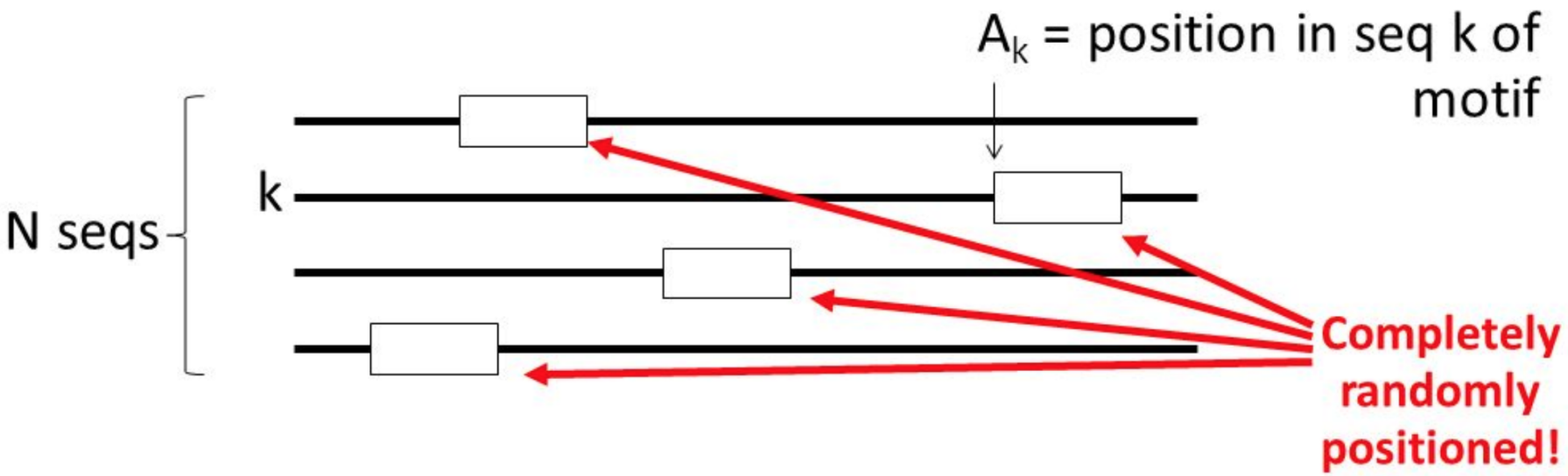
Motif finding

- The Gibbs motif sampling (a Monte-Carlo algorithm)
- Every position i in the motif sequence follows probability distribution with $q_{ij} = \{q_{i,1}, ..., q_{i,4}\}$
  - 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 $p_j = \{p_1, ..., p_4\}$
- 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.

Motif finding

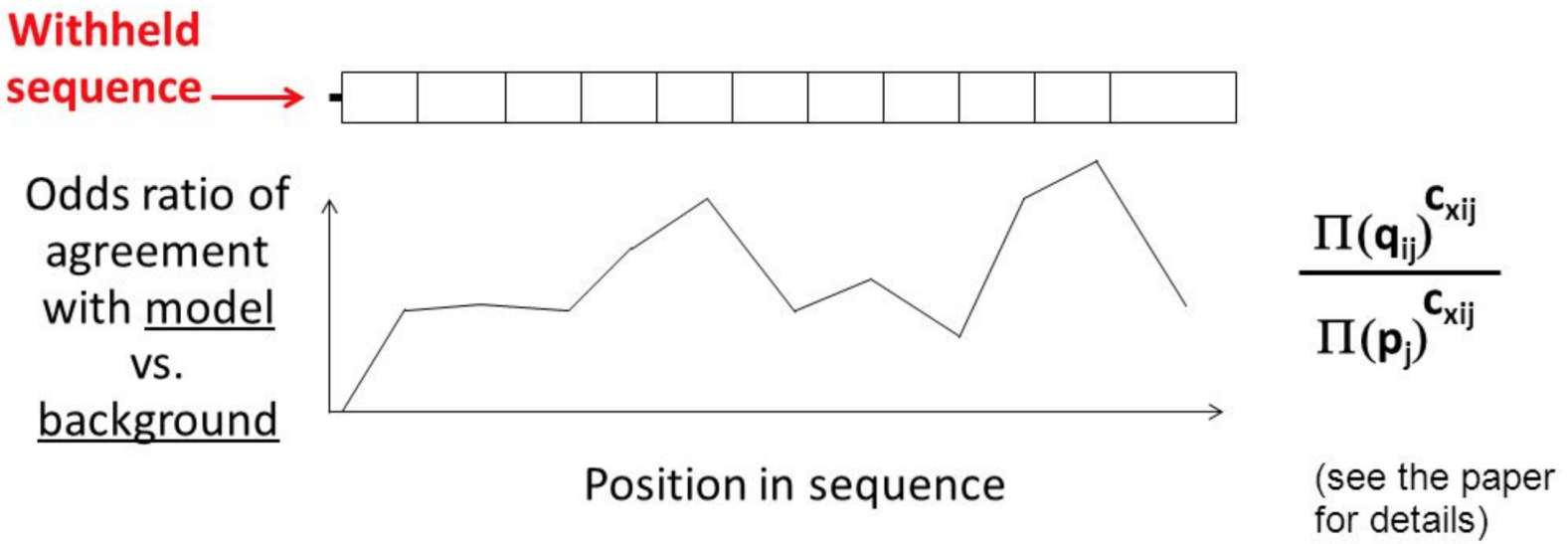
- Start by choosing w and randomly positioning each motif
- Predictive update step: Randomly choose one sequence, calculate $q_{i,j}$ and p_j from $N-1$ remaining sequences



$$q_{i,j} = \frac{c_{i,j} + b_j}{N - 1 + \sum b_j}$$

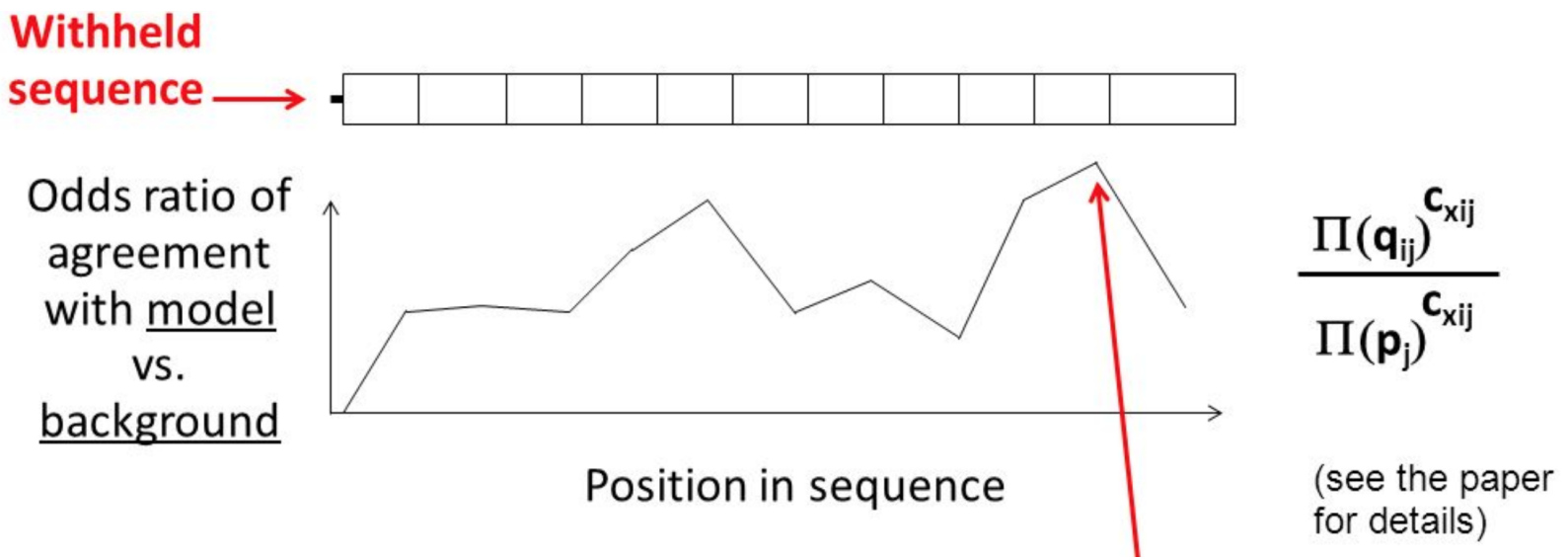
Motif finding

- Stochastic sampling step: For withheld sequence, slide motif down sequence and calculate agreement with model



Motif finding

- Don't just choose the maximum
- Instead, select a new A_k position proportional to this odds ratio
- Then, choose a new sequence to withhold, and repeat everything
- Stop the iteration if there is no change in maximizing the likelihood function $F = \sum_{i=1}^{W} \sum_{j=1}^{4} c_{i,j} log \frac{q_{i,j}}{p_j}$



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.

Summary

- 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

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 predicti
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	X				
	MEME-ChIP [57]	X		X	X	X			
	peak-motifs [58]	X		X	X				X
	Cistrome [49]	X	X	X		X	X		X

More tools

- 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/