Introduction to ChIP-seq data analysis

ENAR 2014

Hao Wu, Emory University

Outline

- Introduction to ChIP-seq experiment: biological motivation and experimental procedure.
- Method and software for ChIP-seq peak calling:
  - Protein binding ChIP-seq.
  - Histone modifications.
- After peak calling:
  - Overlaps of peaks.
  - Differential analysis.

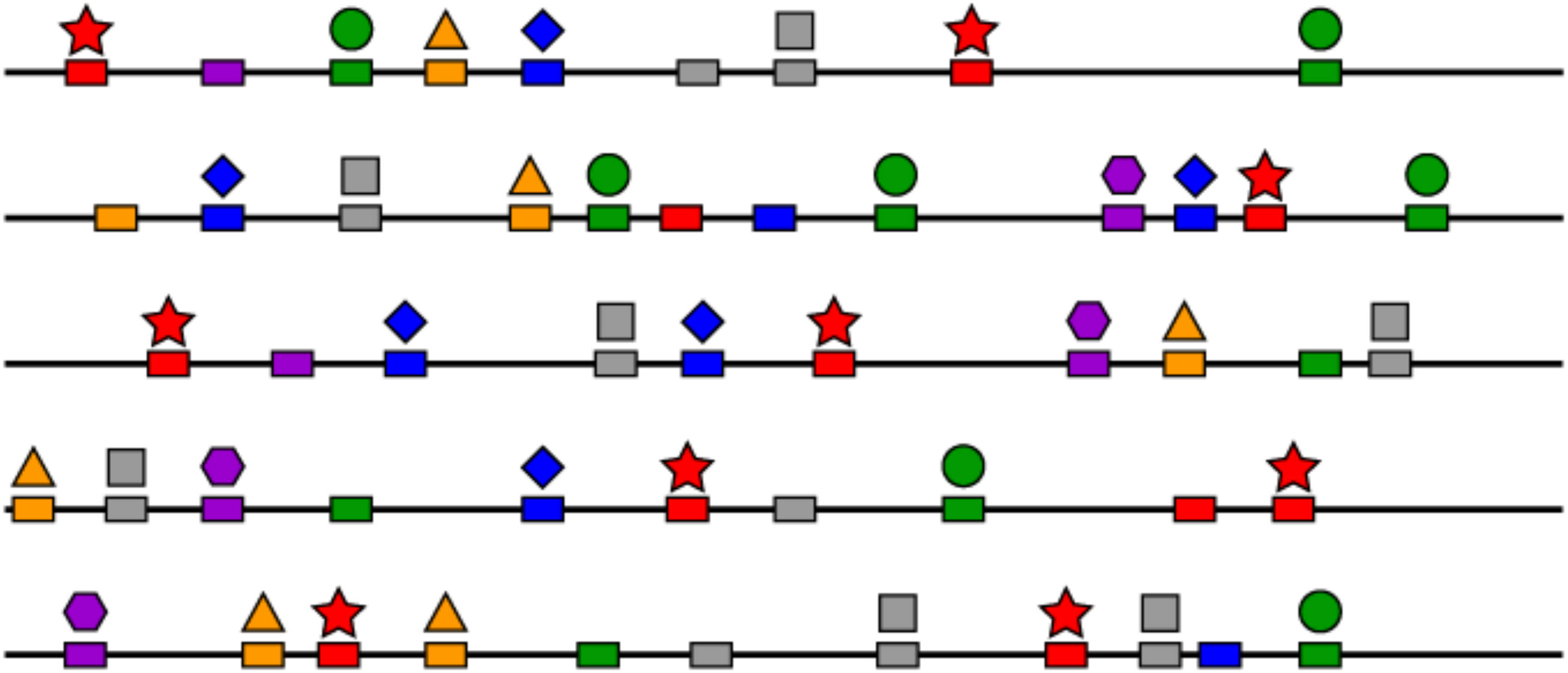
ChIP-seq: Chromatin ImmunoPrecipitation + sequencing

- Biological motivation: detect or measure some type of 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

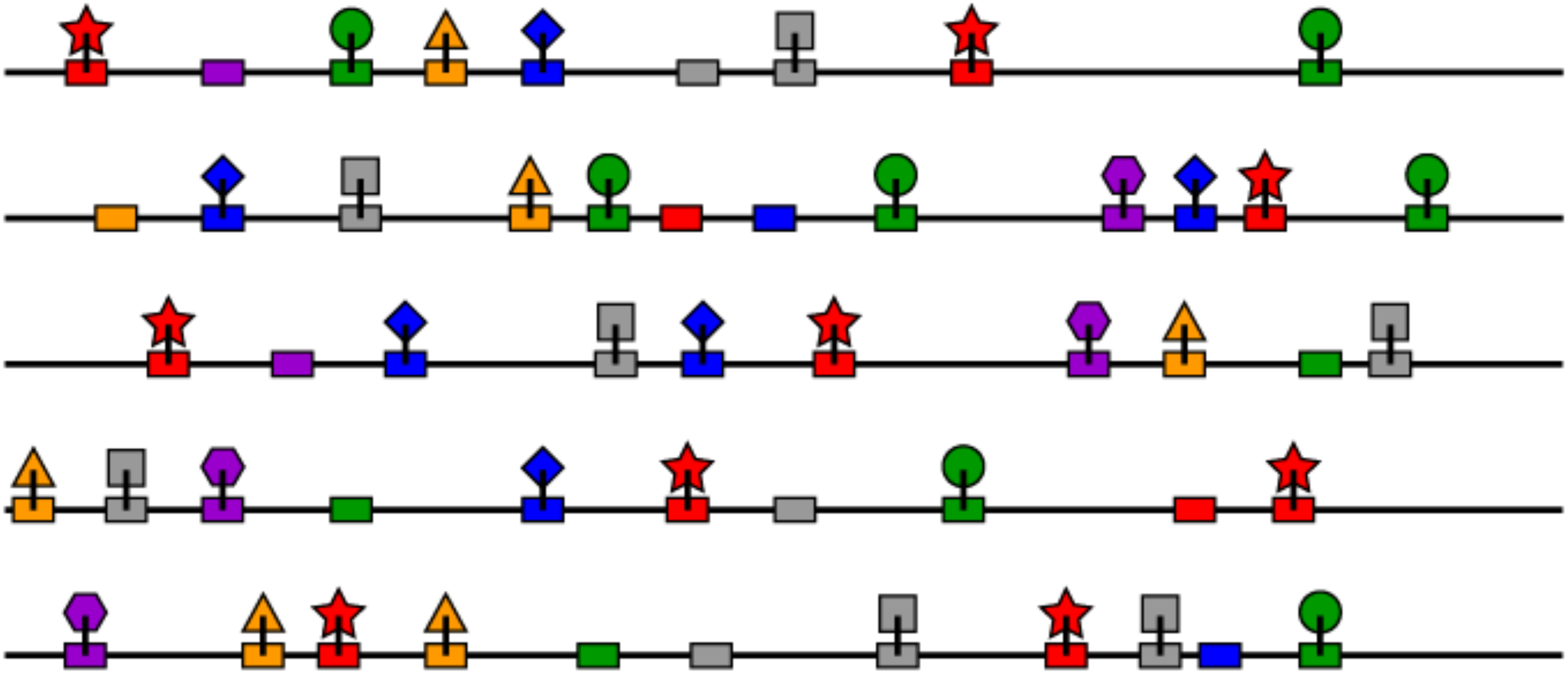
- **Crosslink**: fix proteins on isolate genomic DNA.
- **Sonication**: cut DNA in small pieces of ~200bp.
- **IP**: use antibody to capture DNA segments with specific proteins.
- **Reverse crosslink**: remove protein from DNA.
- Sequence the DNA segments.

Genomic DNA with TF



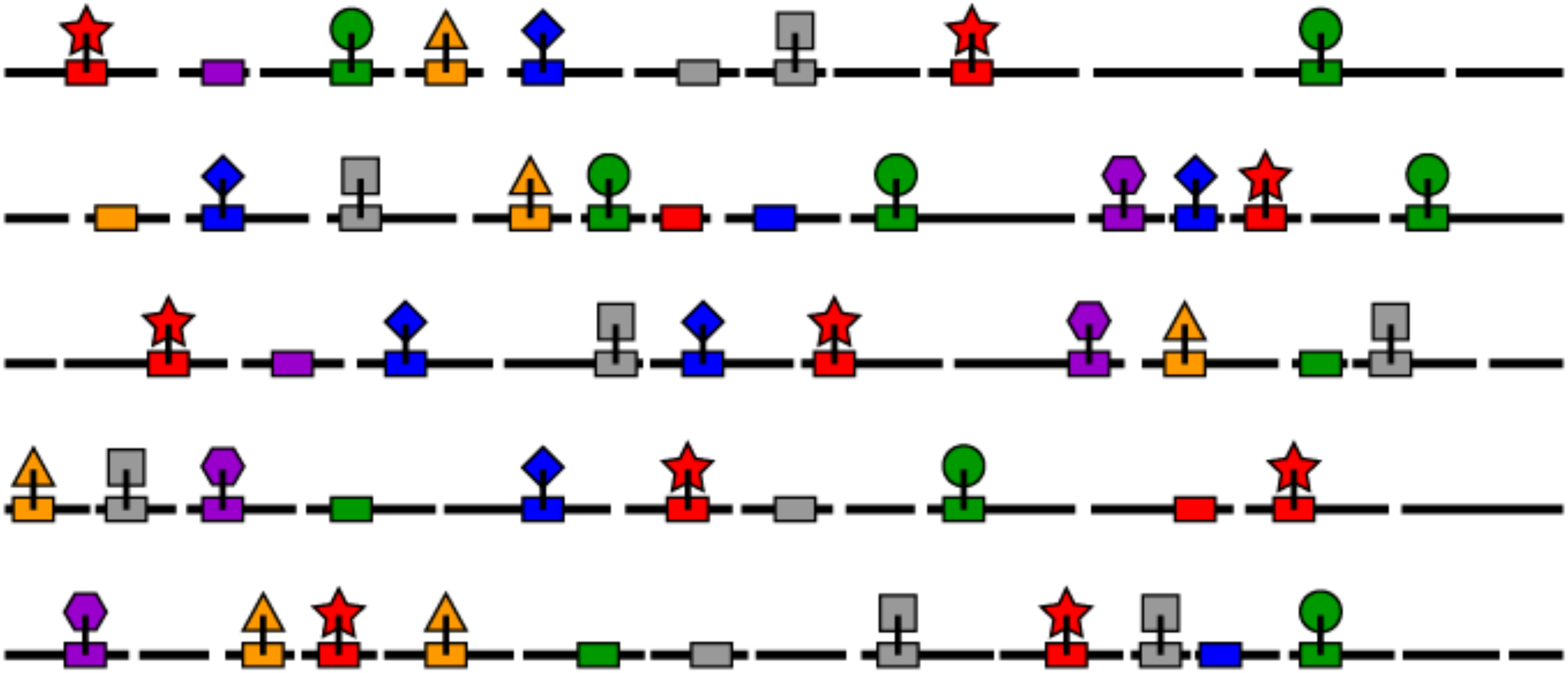
By Richard Bourgon at UC Berkley

TF/DNA Crosslinking *in vivo*



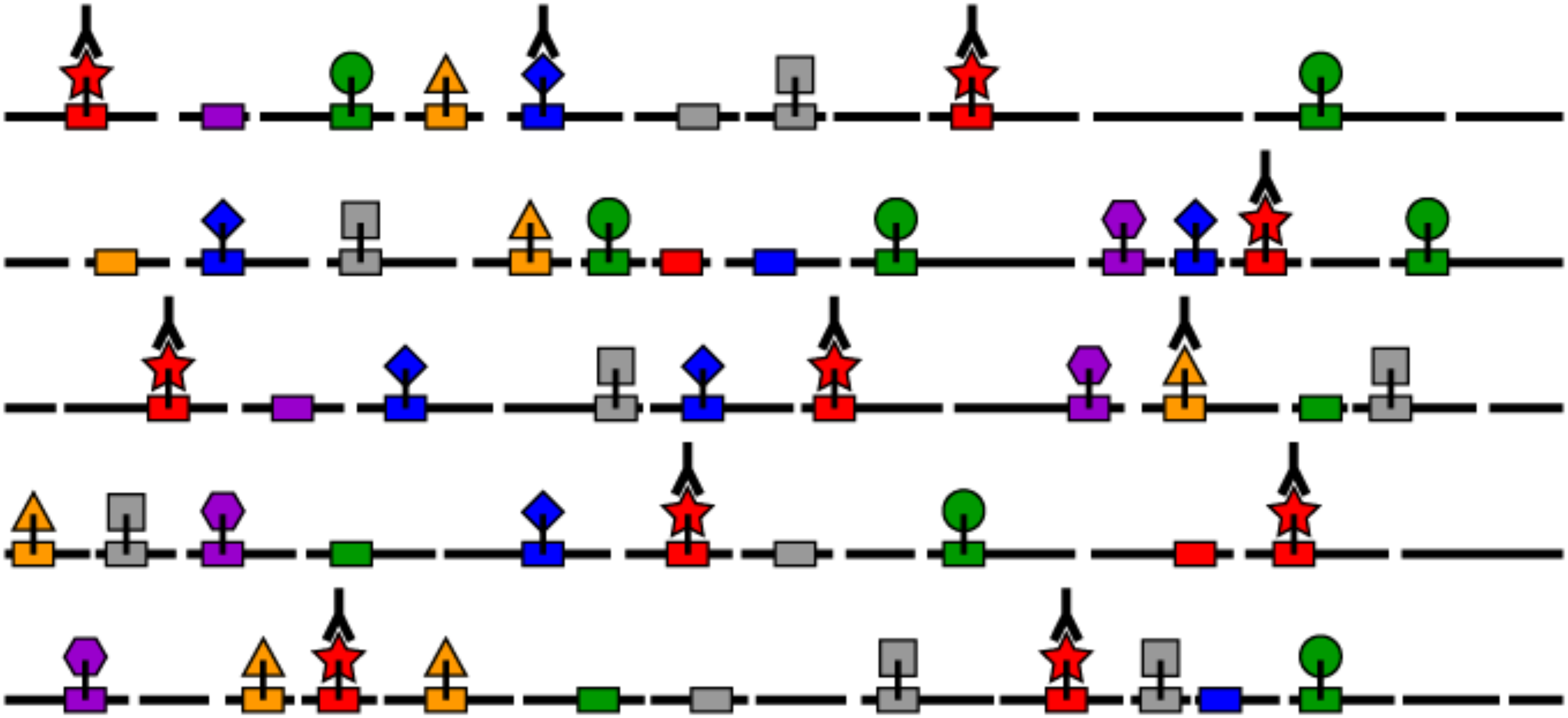
By Richard Bourgon at UC Berkley

Sonication



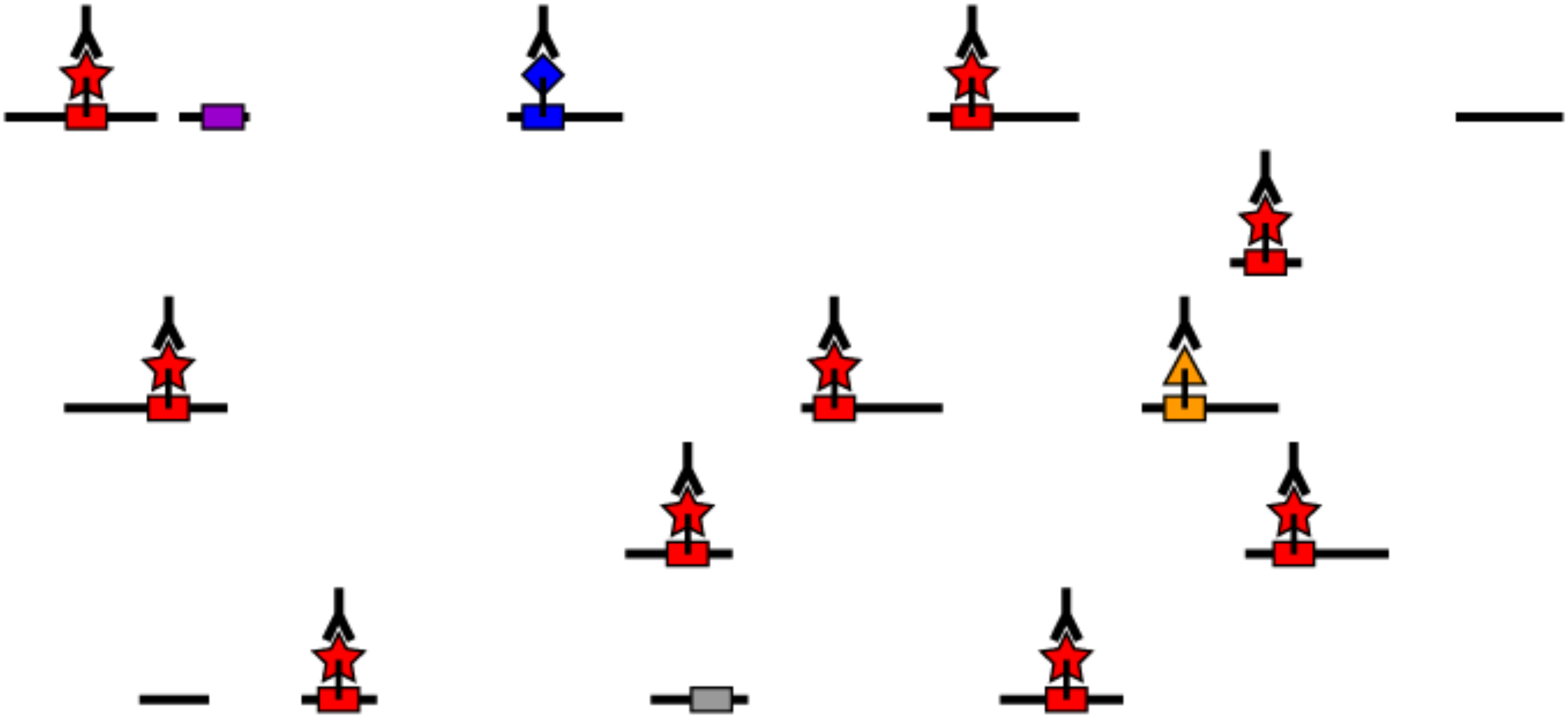
By Richard Bourgon at UC Berkley

TF-specific Antibody



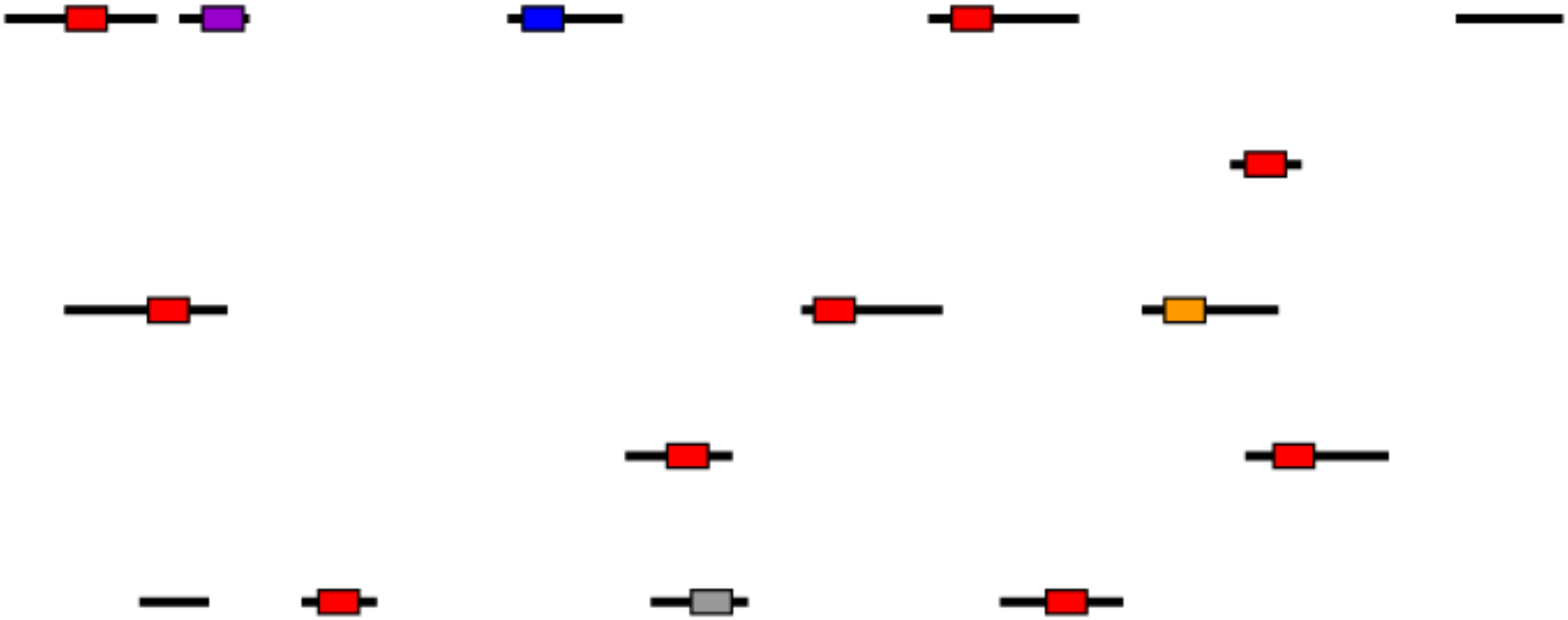
By Richard Bourgon at UC Berkley

Immunoprecipitation (IP)



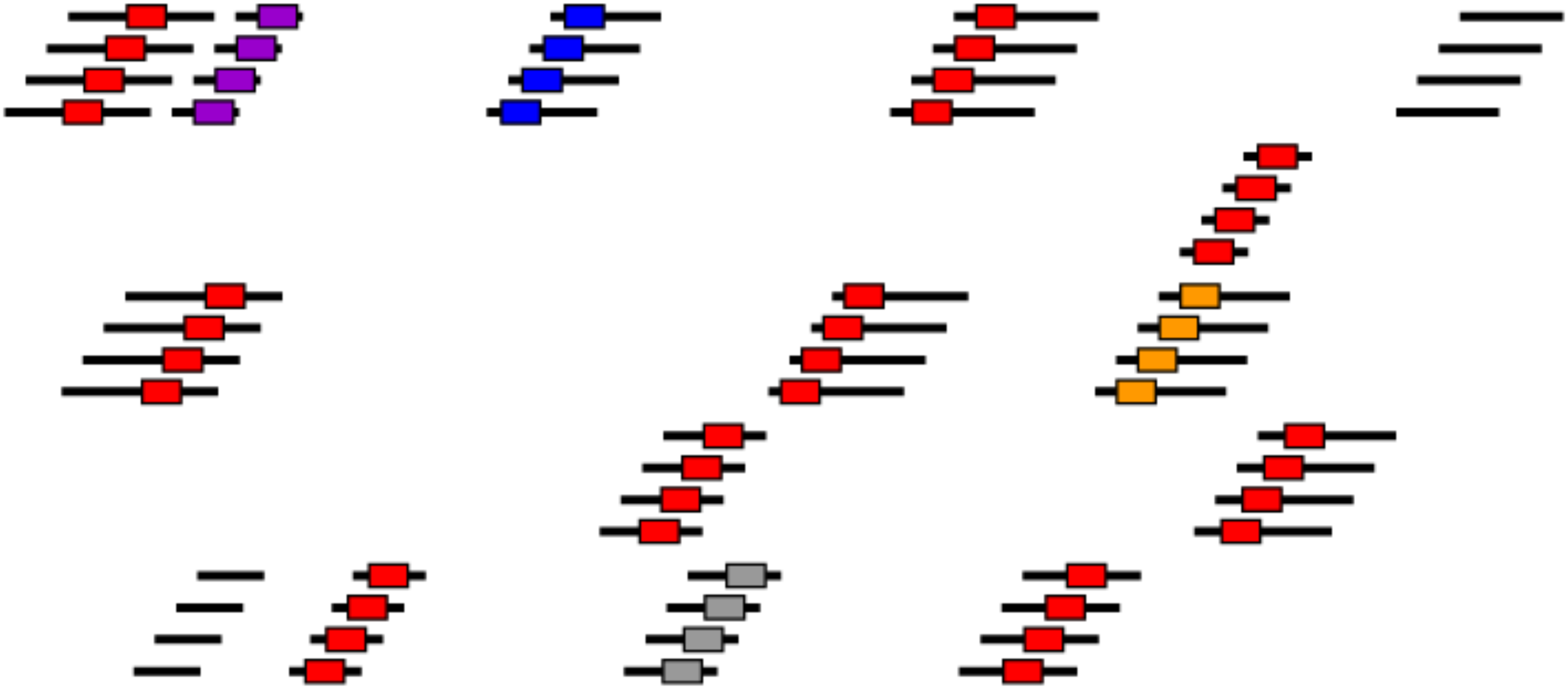
By Richard Bourgon at UC Berkley

Reverse Crosslink and DNA Purification



By Richard Bourgon at UC Berkley

Amplification then sequencing



By Richard Bourgon at UC Berkley

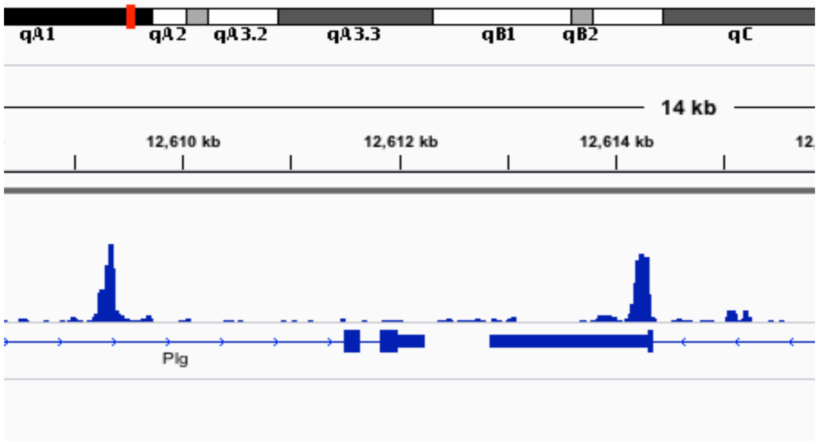
Data from ChIP-seq

- Raw data: sequence reads.
- After alignments: genome coordinates (chromosome/position) of all reads.
- For downstream analysis, aligned reads are often 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.

Methods and software for ChIP-seq peak/block calling

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.



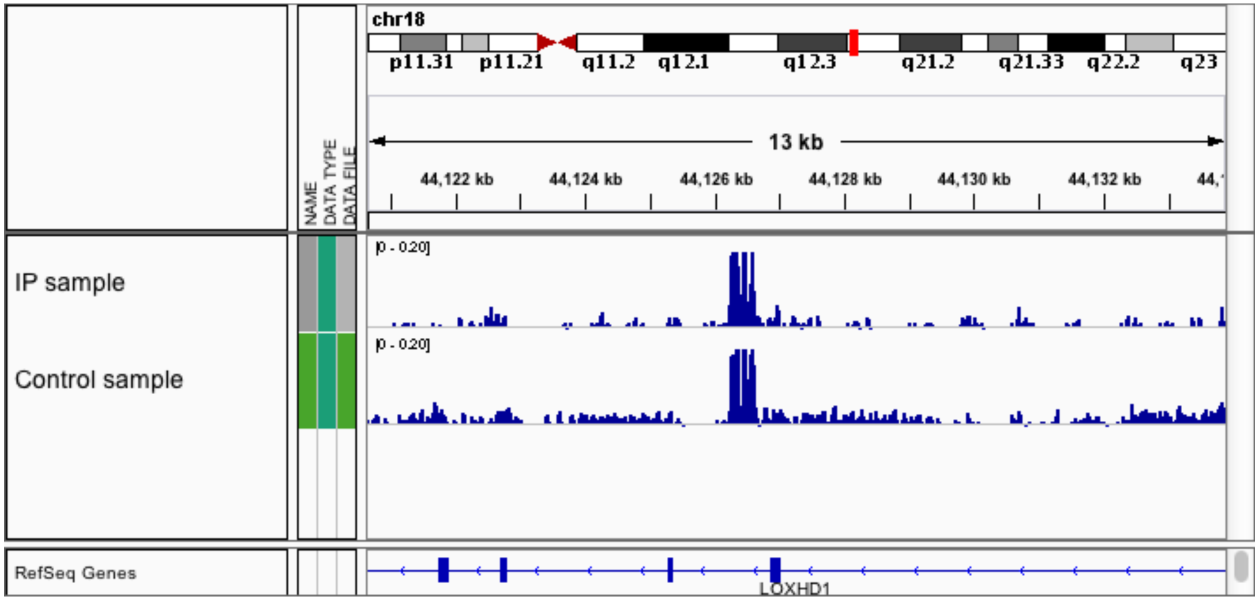
- Peak detection is the most fundamental problem in ChIP-seq data analysis.

Simple ideas for peak detection

- 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 considered to improve the peak calling performance.

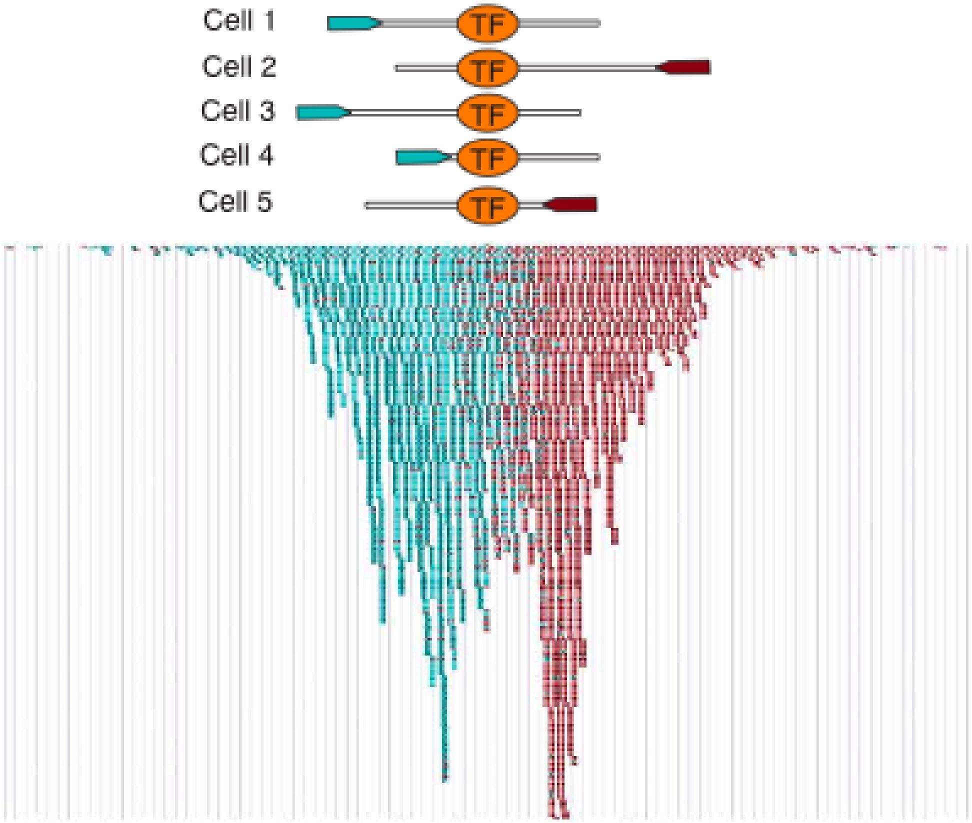
Control sample is important

- A control sample is necessary for correcting many artifacts:
  - DNA sequence contents affect amplification or sequencing process.
  - Repetitive regions affect alignments.
  - Chromatin structures (e.g., open chromatin region or not) affect the DNA sonication process.



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 (because multi-aligned reads are discarded), thus affect the ability to detect peaks.

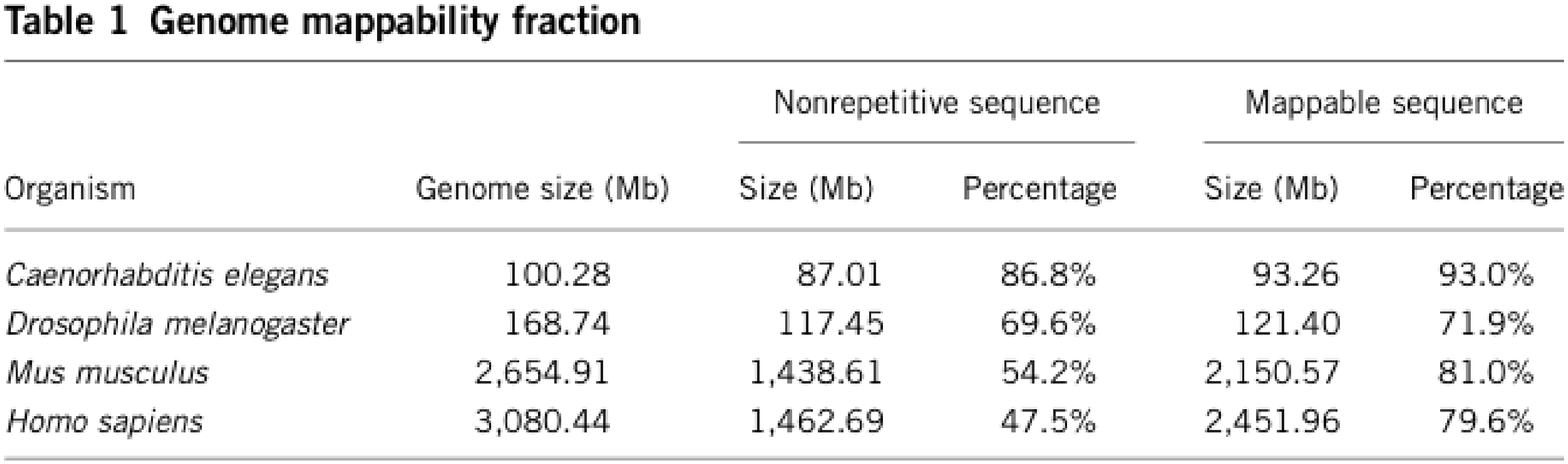


Table 1 Genome mappability fraction

		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%

Normalization issues

- The most common normalization needed is to adjust for total counts.
- Normalize by total counts is conservative, because ChIP sample contains reads mapped to background and peaks, but control sample have reads mapped to background only.
- It's better to normalize using the number of total reads in backgrounds. Two pass algorithm:
  - Roughly find peaks, and exclude those regions.
  - Compute total reads in the leftover regions and normalize based on that.
- Other normalizations (GC contents, MA plot based) available, but don't seems to help much.

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 Poisson distribution with rate: $\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k},] \lambda_{5k}, \lambda_{10k})$
  - The dynamic rate capture the local fluctuation of counts.
- FDR is estimated 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 is easy

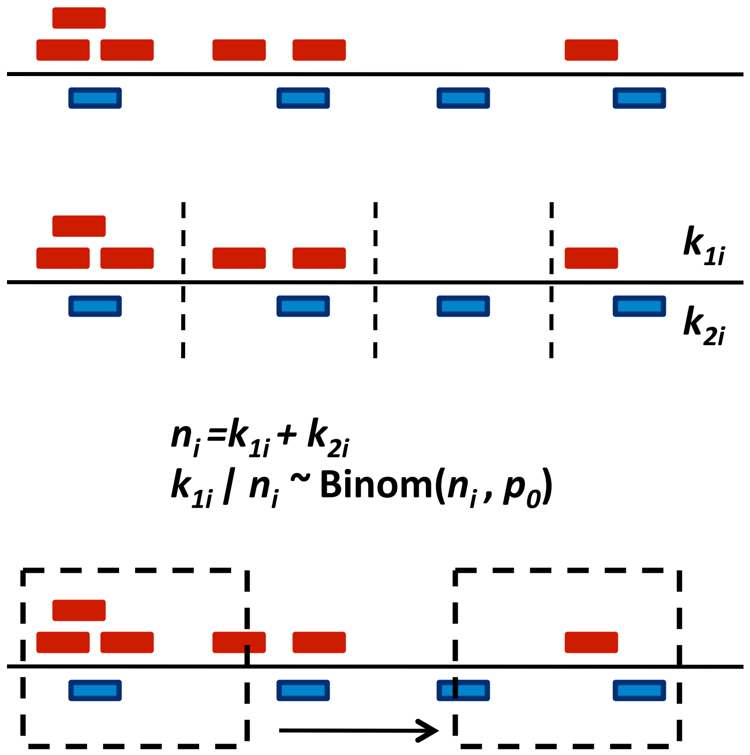
- http://liulab.dfci.harvard.edu/MACS/index.html
- Written in Python, runs in command line.
- Command:

```
macs14 -t sample.bed -c control.bed -n result
```

- A problem: doesn’t consider replicates. Data from replicated samples need to be merged.

Cisgenome (Ji et al. 2008, *NBT*)

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



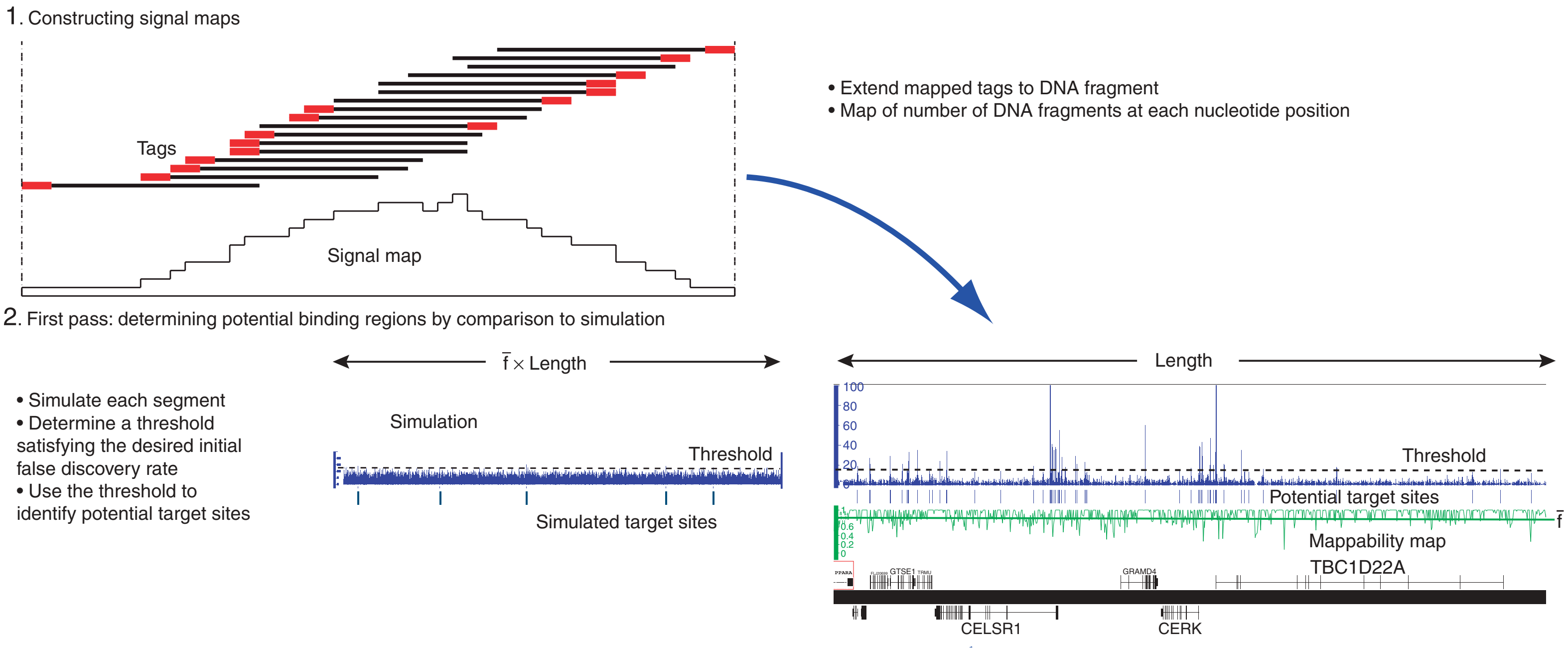
$n_i = k_{1i} + k_{2i}$

$k_{1i} \mid n_i \sim \text{Binom}(n_i, p_0)$

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* × *Length*, where *f* is the proportion of mappable bases in the segment.



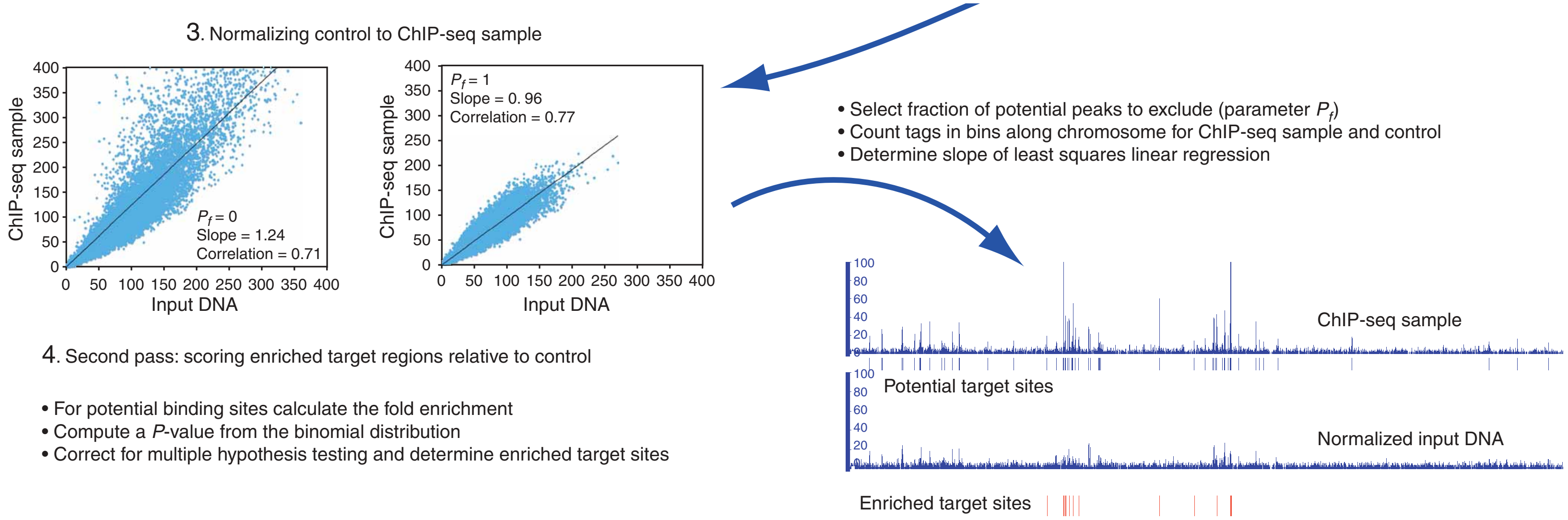
1. Constructing signal maps

- Extend mapped tags to DNA fragment
- Map of number of DNA fragments at each nucleotide position

2. First pass: determining potential binding regions by comparison to simulation

- Simulate each segment
- Determine a threshold satisfying the desired initial false discovery rate
- Use the threshold to identify potential target sites

- 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

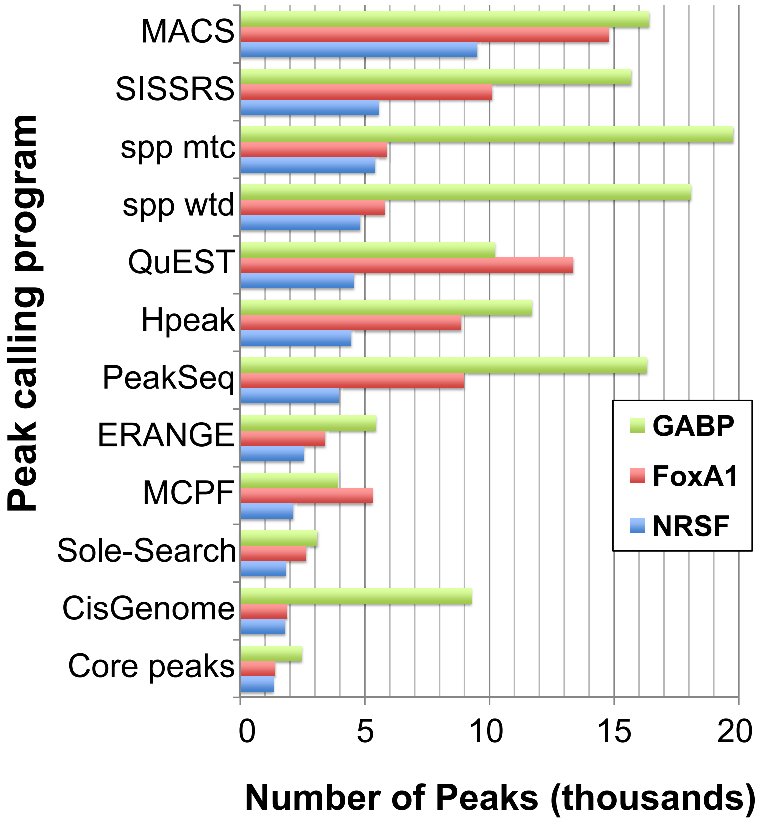
- Select fraction of potential peaks to exclude (parameter P_f)
- Count tags in bins along chromosome for ChIP-seq sample and control
- Determine slope of least squares linear regression

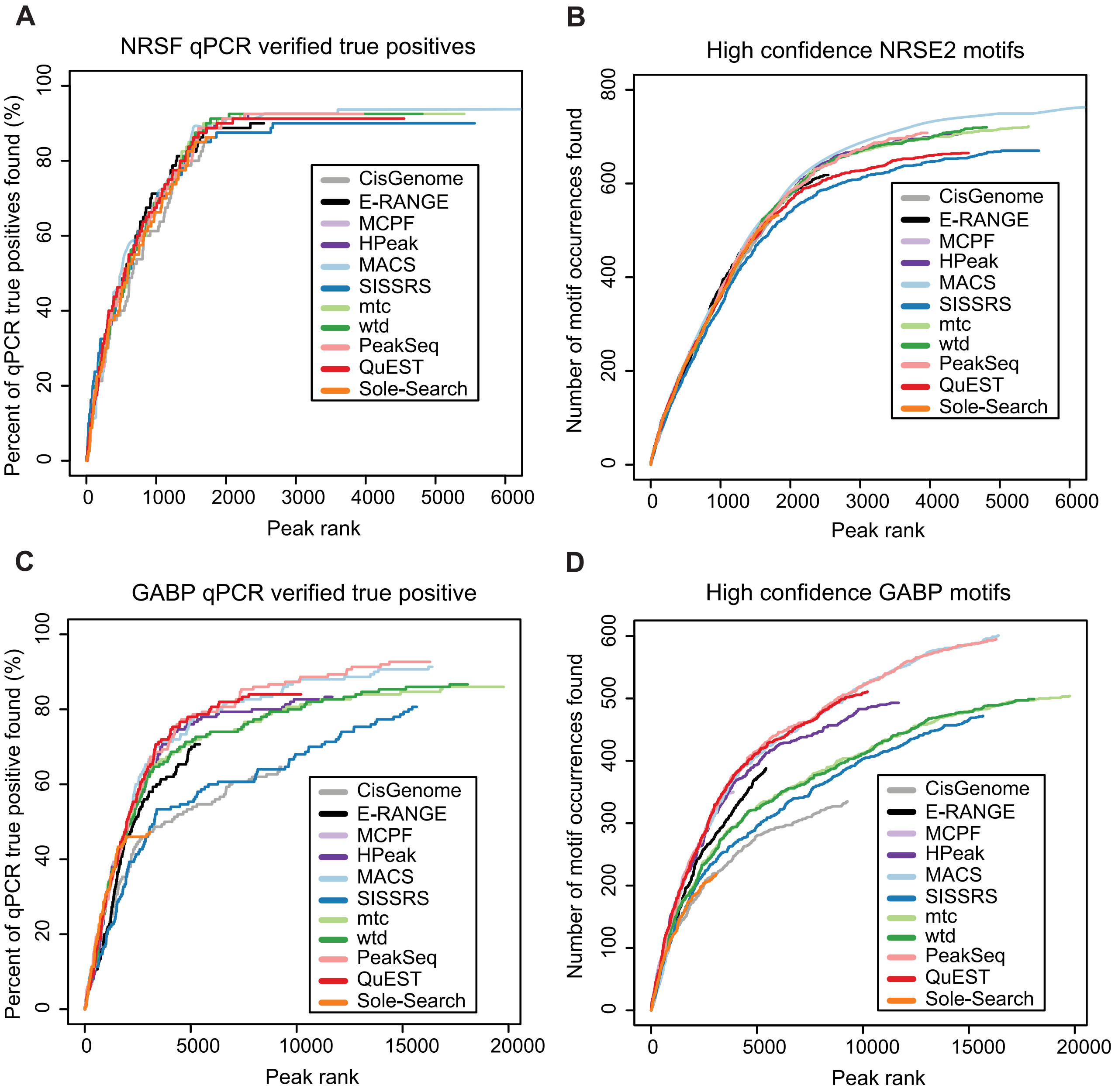
4. Second pass: scoring enriched target regions relative to control

- 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*



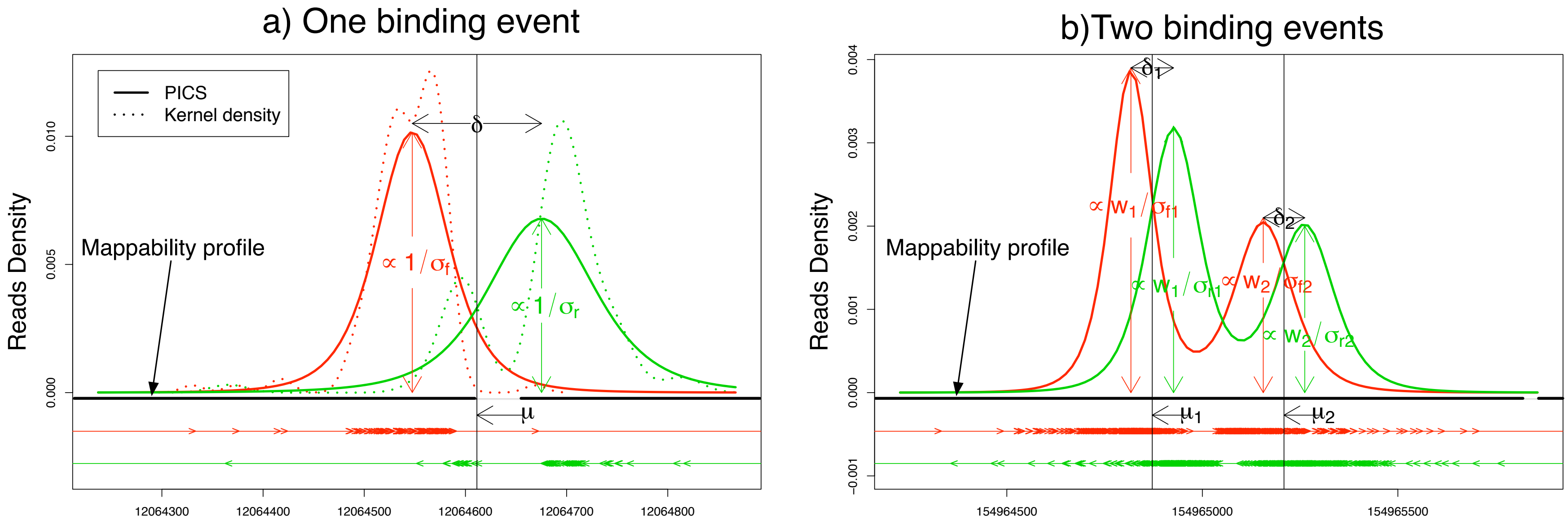

A
NRSF qPCR verified true positives
Percent of qPCR true positives found (%)
0
20
40
60
80
100
0
1000
2000
3000
4000
5000
6000
Peak rank
CisGenome
E-RANGE
MCPF
HPeak
MACS
SISSRS
mtc
wtd
PeakSeq
QuEST
Sole-Search
B
High confidence NRSE2 motifs
Number of motif occurrences found
0
200
400
600
800
0
1000
2000
3000
4000
5000
6000
Peak rank
CisGenome
E-RANGE
MCPF
HPeak
MACS
SISSRS
mtc
wtd
PeakSeq
QuEST
Sole-Search
C
GABP qPCR verified true positive
Percent of qPCR true positive found (%)
0
20
40
60
80
100
0
5000
10000
15000
20000
Peak rank
CisGenome
E-RANGE
MCPF
HPeak
MACS
SISSRS
mtc
wtd
PeakSeq
QuEST
Sole-Search
D
High confidence GABP motifs
Number of motif occurrences found
0
100
200
300
400
500
600
0
5000
10000
15000
20000
Peak rank
CisGenome
E-RANGE
MCPF
HPeak
MACS
SISSRS
mtc
wtd
PeakSeq
QuEST
Sole-Search

Another class 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 shape 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.
- R/Bioconductor package available.



a) One binding event

b)Two binding events

$$f_i \sim \sum_{k=1}^{K} w_k t_4\left(\mu_{fk}, \sigma^2_{fk}\right) \overset{d}{=} g_f(f_i|\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\sigma}_f)$$

$$r_j \sim \sum_{k=1}^{K} w_k t_4\left(\mu_{rk}, \sigma^2_{rk}\right) \overset{d}{=} g_r(r_j|\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\sigma}_r)$$

GPS (Genome Positioning System)

Guo *et al.* 2010, *Bioinformatics*

- Part of GEM (Genome wide Event finding and Motif discovery) software suite.
- 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.
- Written in Java, runs in command line.

Use GPS

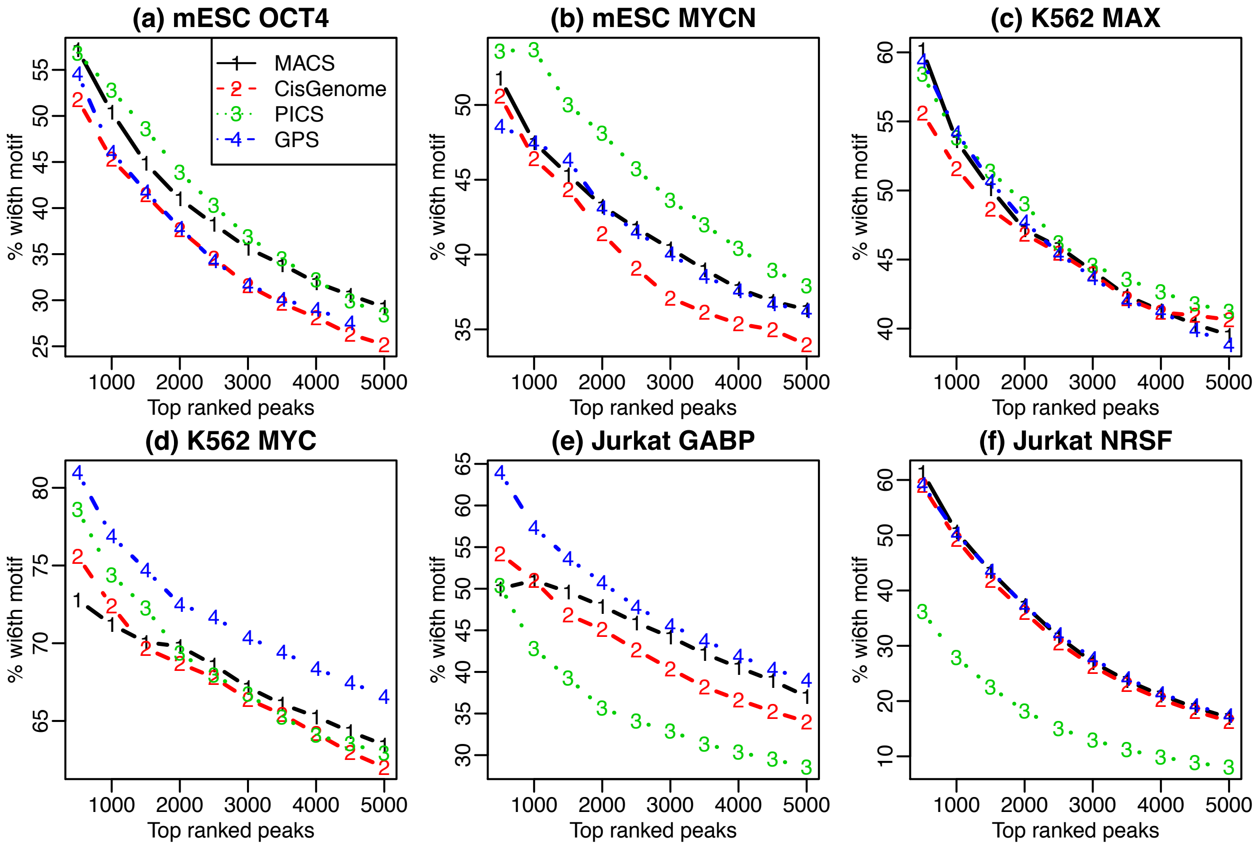
- Run following command:

```
java -Xmx1G -jar gps.jar --g mm8.info --d
Read_Distribution_default.txt --expt IP.bed
--ctrl control.bed --f BED --out result
```

- It’s much slower than MACS or CisGenome.

A little more comparison

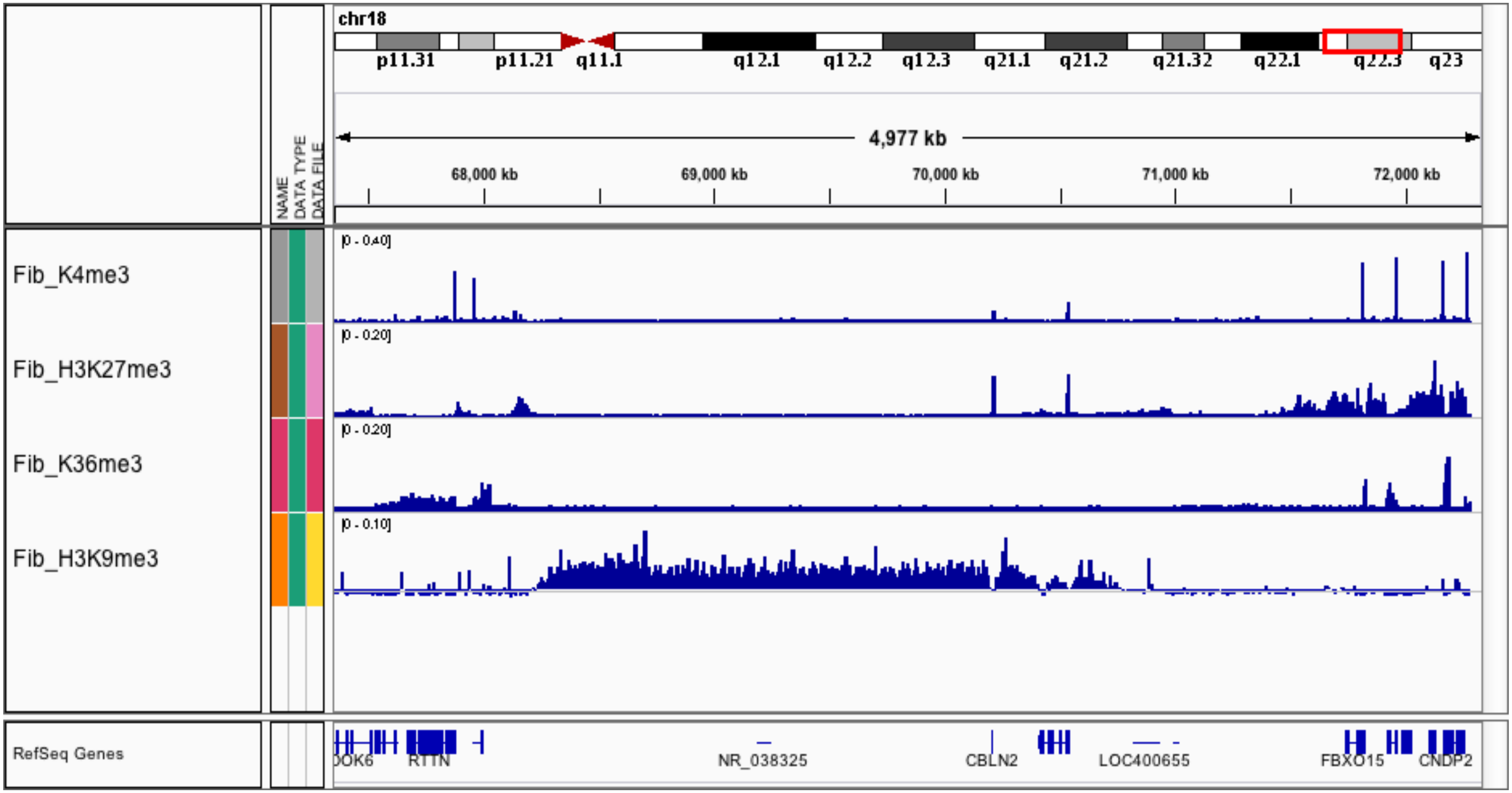
- I found that using peak shapes helps. GPS tend to perform better. PICS seems not stable.



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.
- Existing methods based on: smoothing, HMM, wavelet, etc.
- Method for detecting blocks is relatively under-developed and under-tested:
  - ENCODE is evaluating existing methods.

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
- …

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.

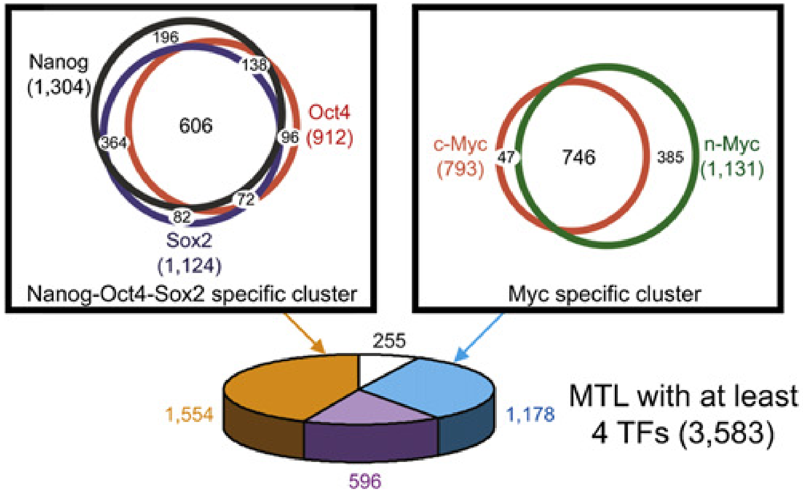
Downstream analysis after peak/block calling

After peak/block calling

- Compare results among different samples:
  - Presence/absence of peaks.
  - Differential binding.
  - Look for Combinatory patterns.
- Compare results with other type of data:
  - Correlate TF binding with gene expressions from RNA-seq or DNA methylation from BS-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: ***BEDtools***, ***Bioconductor***, etc.
  - We will practice in the lab.



Differential binding (DB) analysis

- Problems for the overlapping analysis are:
  - Completely ignores the quantitative differences of peaks.
  - Highly dependent on the thresholds for defining peaks.
- More desirable: quantitative comparison to detect differential protein binding or histone modification (referred to as “DB analysis”).
- Typical DB analysis procedure:
  - Call peaks from individual dataset.
  - Union the called peaks to form candidate regions.
  - Hypothesis testing for each candidate region.

Complications in DB analysis

- Different backgrounds: for example, chromatin structures affect the sequencing efficiency.
- Signal to noise ratios (SNR) from different experiments:
  - Biological: sample with less peak will have taller peaks.
  - Technical: qualities of the experiments are different.
- To summarize:
  - DB is more complicated than RNA-seq DE problem.
  - Methods are relatively under-developed.

Existing methods for DB analysis

- Normalize data first, then compare:
  - **MAnorm** (Shao *et al.* 2012, *Genome Biology*): normalization based on MA plot of counts from two conditions, then use normalized "M" values to rank differential peaks.
  - **ChIPnorm** (Nair *et al.* 2012, *PLoS One*): quantile normalization for each dataset, then define differential peak based on normalized IP differences.
- Based on RNA-seq DE methods:
  - **DBChIP**: Liang *et al.* (2012) Bioinformatics.
  - **DiffBind**: A Bioconductor package.
- Model the differences of data from two IP sample:
  - **DIME** (Taslim *et al.* 2009, 2011, *Bioinformatics*): finite mixture model on differences of normalized IP counts.
  - **ChIPDiff** (Xu et al. 2008, Bioinformatics): HMM on differences of normalized IP counts between two groups.

Review

- NGS provides cost-effective ways for various aspects of genomic research.
- ChIP-seq is a type of NGS for genome-wide regional analysis: detect protein binding or histone modification regions.
- Main goal of ChIP-seq data analysis is “peak/block” calling.
  - Many software available, based on smoothing or HMM.
  - Block calling is harder.
- Comparison of ChIP-seq signals (differential binding analysis) is still an open problem.