Introduction to ChIP-Seq data analyses

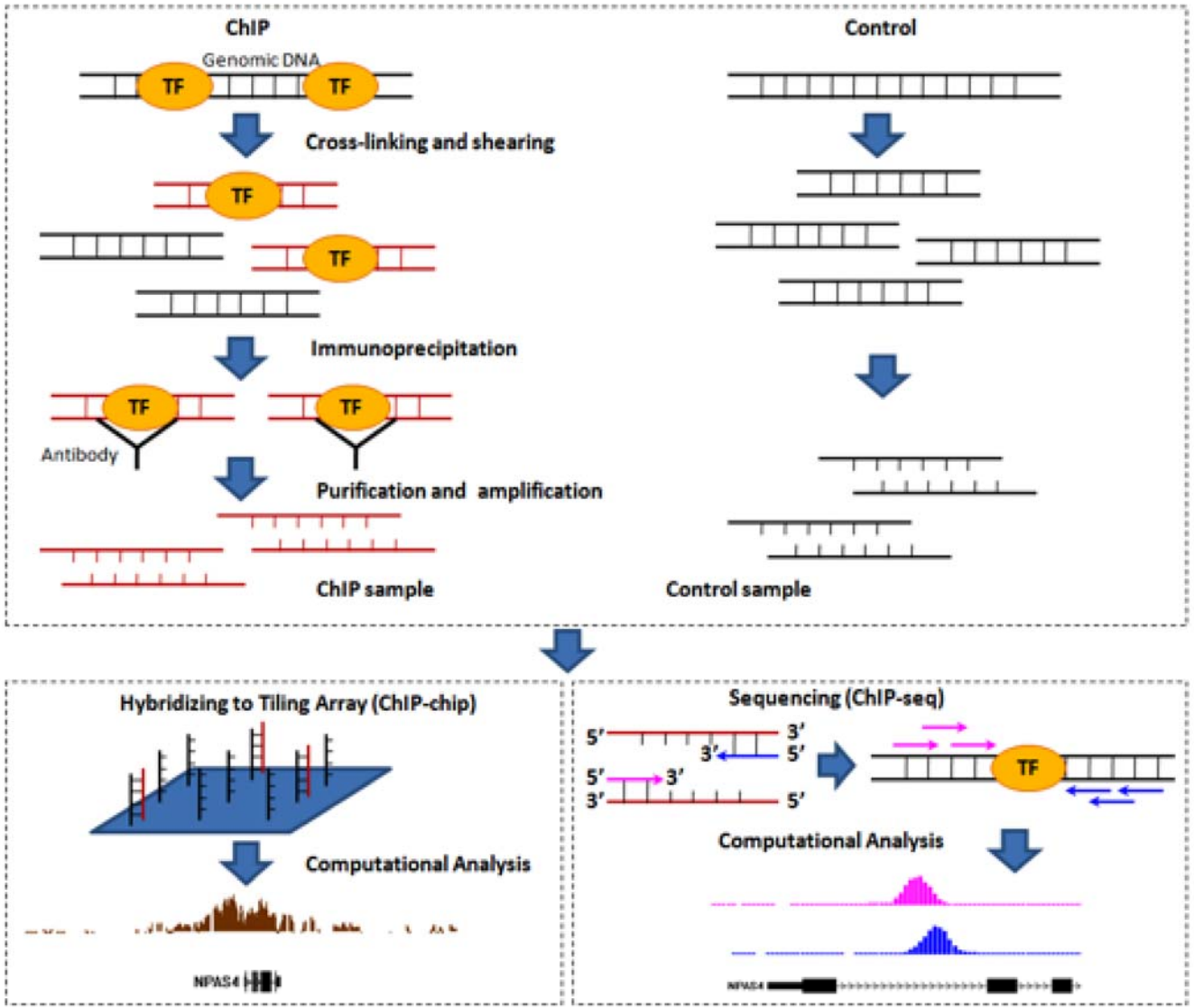
Acknowledgement: slides taken from Dr. H Wu @Emory

ChIP-seq: Chromatin ImmunoPrecipitation + sequencing

- Same biological motivation as ChIP-chip: 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

- Same as ChIP-chip except the last step: sequencing is used to replace microarray.
  - 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.

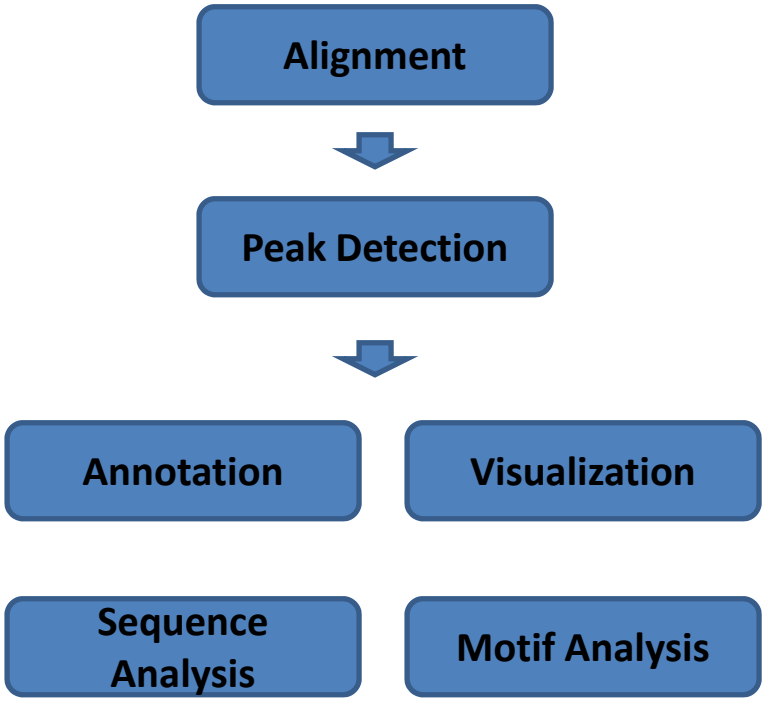


Hongkai Ji, Hopkins Biostat

Advantages of ChIP-seq over ChIP-chip

- Not limited by array design, especially useful for species without commercially available arrays.
- Higher spatial resolution.
- Better signal to noise ratio and dynamic ranges.
- Less starting materials.

ChIP-seq data analyses



Hongkai Ji, Hopkins Biostat

ChIP-seq “peak” detection

- Peaks: protein binding or histone modification sites.
- Data from ChIP-seq:
  - raw data: sequence reads.
  - After alignments: genome coordinates (chromosome/position) of the start of each read.
  - Usually, aligned reads are summarized into “counts” of equal sized bins genome-wide:
    1. segment genome into small bins of equal sizes (50bps).
    2. Count number of reads started at each bin.
- The bin counts are inputs for many peak calling algorithms.
  - Similar to probe signals in ChIP-chip.
  - Difference: discrete (seq) vs. continuous (microarray) data.

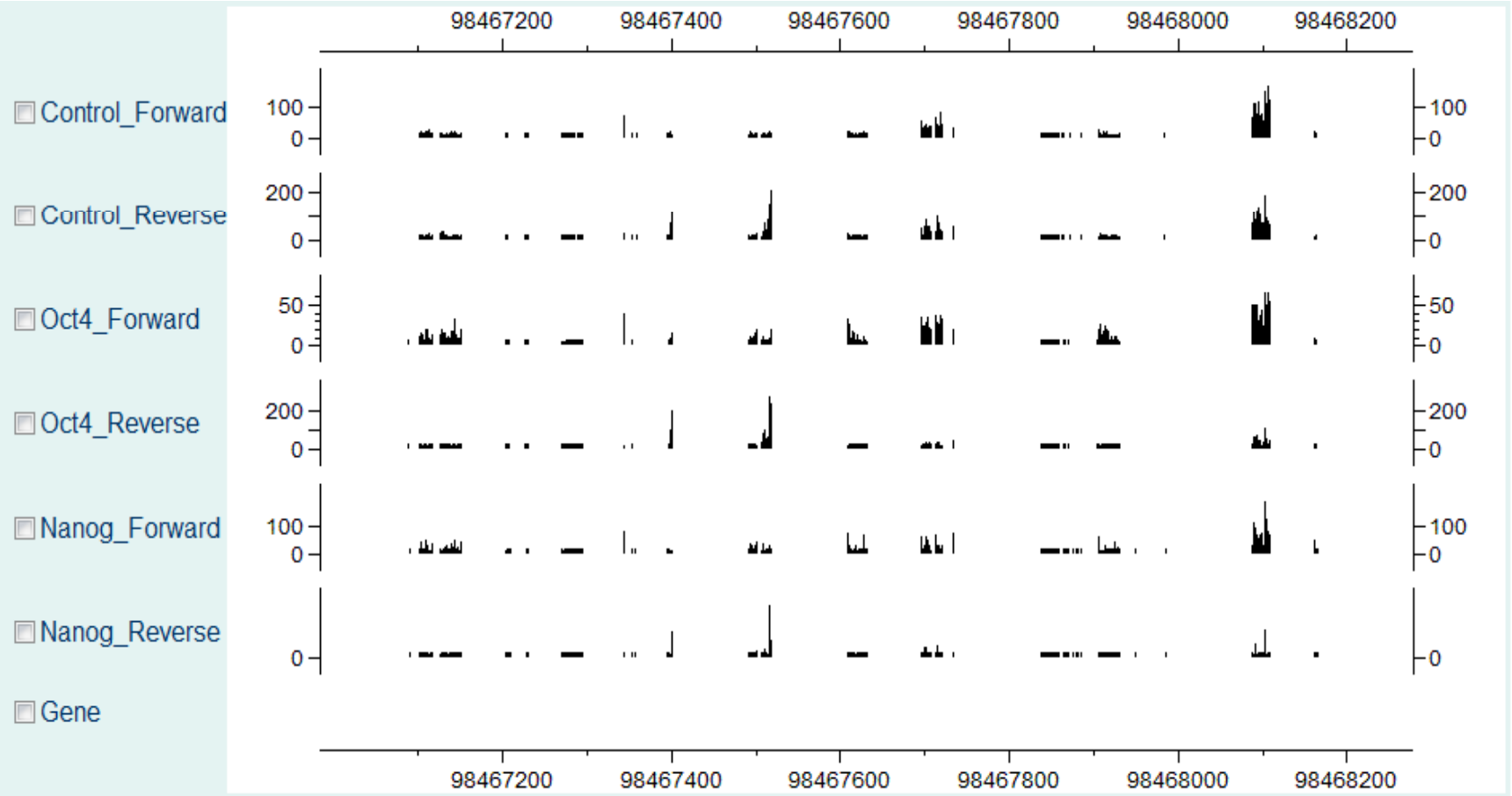
Normalization issues

- The most common normalization needed is to adjust for total counts.
- Divided by total counts is conservative, because ChIP sample contains reads mapped to background and peaks, but control sample have reads mapped to background only.
- Should 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.

Before peak detection: what do we know about ChIP-seq?

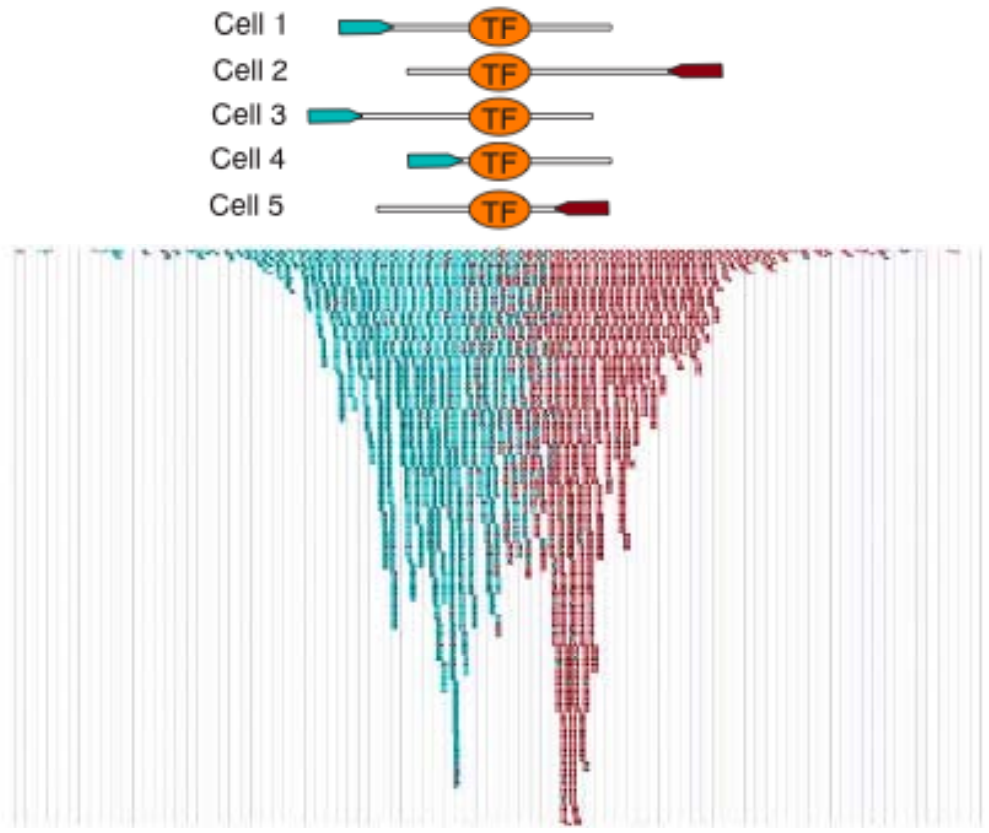
- Sonicated DNA sequences are selected so that the resulting sequence fragments are approximately 200bp.
- In the immunoprecipitated (IP) sample, more sequence reads should be presented around the TF binding sites.
- However, other artifacts need to be considered.
  - DNA sequence: can affect amplification 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.

Control sample is important



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 (or other length now) 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%

Peak detection software

- MACS
- Cisgenome
- QuEST
- Hpeak
- PICS
- PeakSeq
- MOSAiCS
- …

Peak detection methods

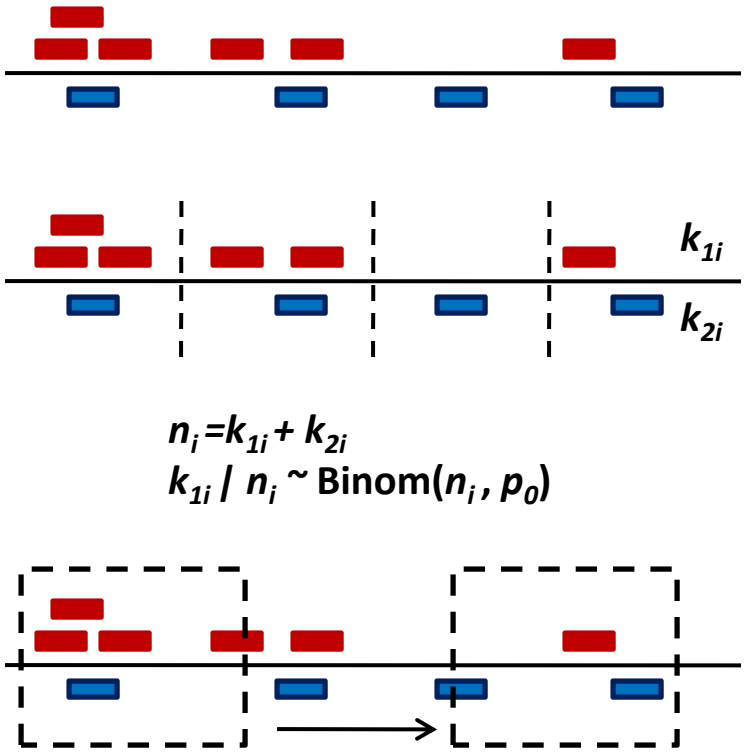
- Counts from neighboring windows need to be combined to make inference.
- To combine counts:
  - Smoothing based: moving average (MACS, Cisgeenome), HMM (Hpeak).
  - Model clustering of reads starting position (PICS).
- Other consideration: mappability of the genome (PeakSeq, MOSAiCS).

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 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.

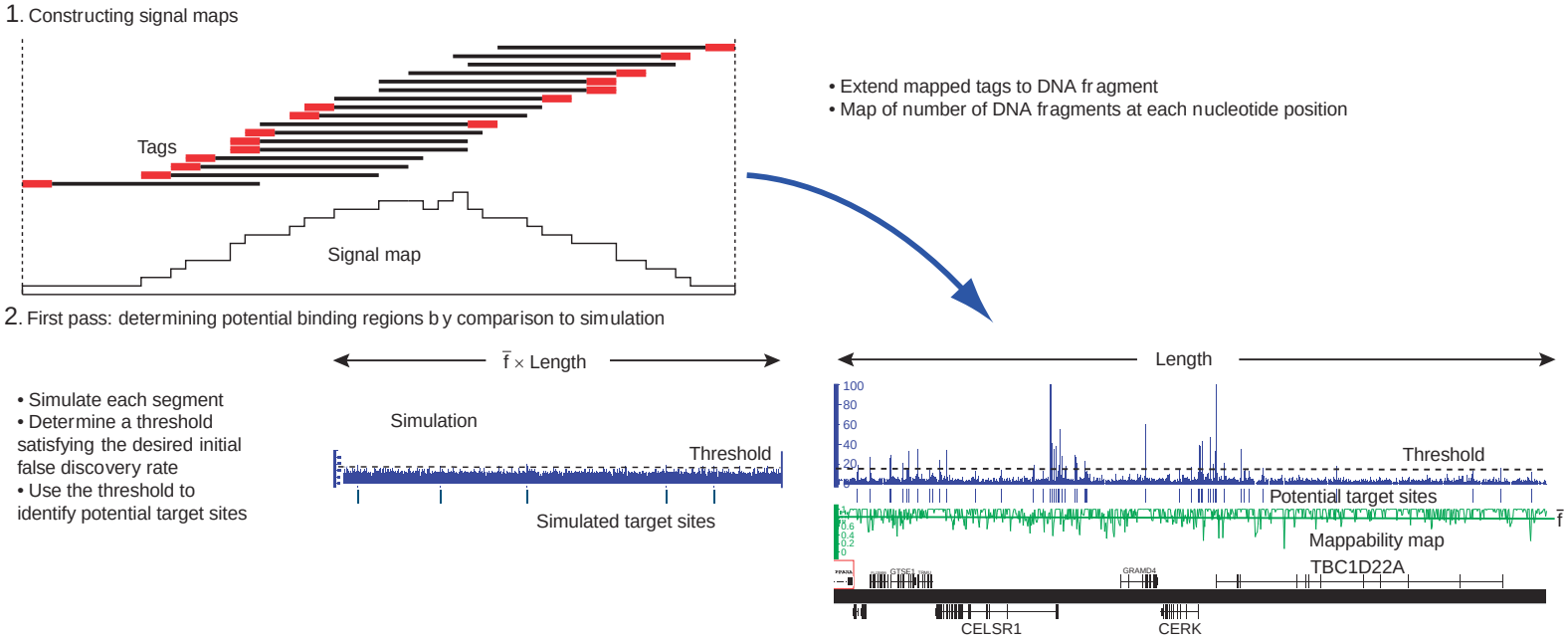
Cisgenome (Ji et al. 2008, *NBT*)

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

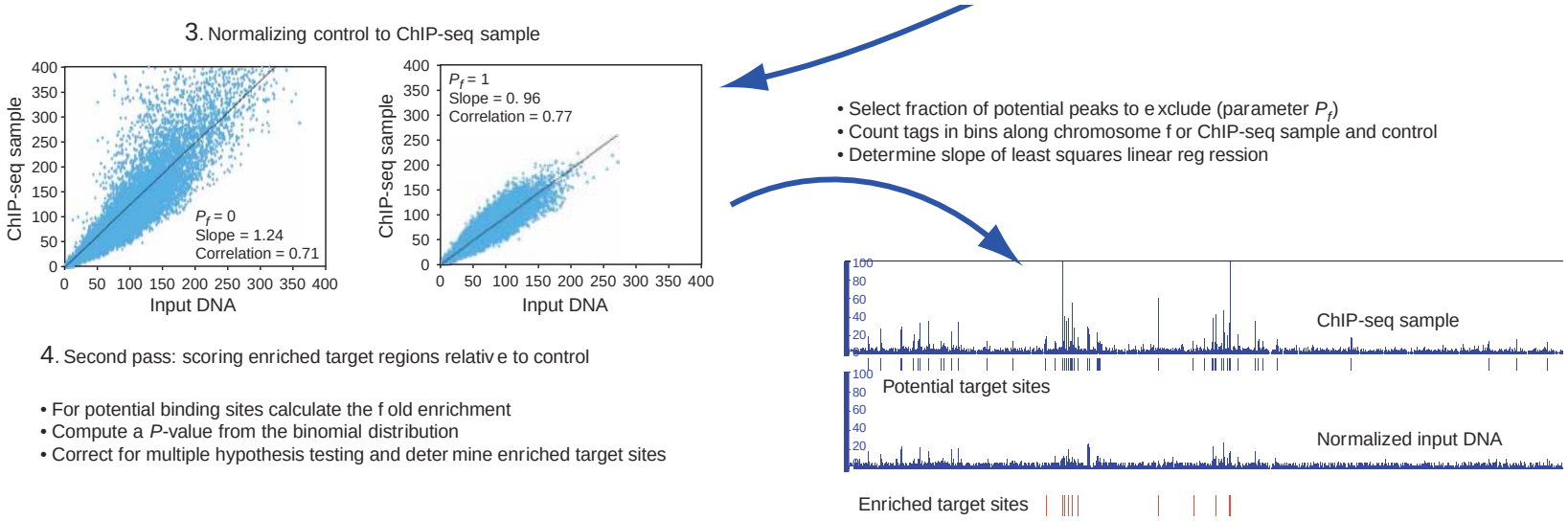


Consider mappability: PeakSeq Rozowsky et al. (2009) *NBT*

- 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.



- 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.



Bioconductor packages for ChIP-seq

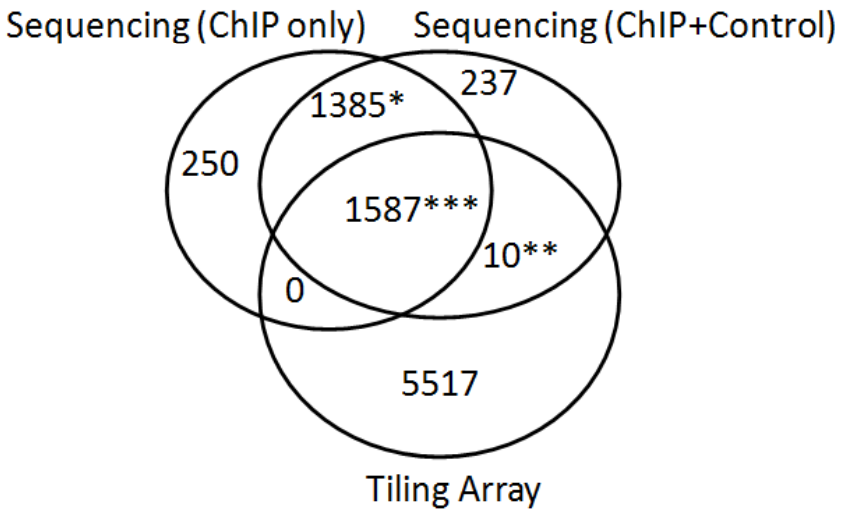
- There are some packages: chipseq, ChIPseqR, BayesPeak, PICS, etc., but not very popular.
- Most people use command line driven software like MACS or CisGenome GUI.

Comparing ChIP-seq and ChIP-chip: Ji et al. (2008) *Nature Biotechnology*

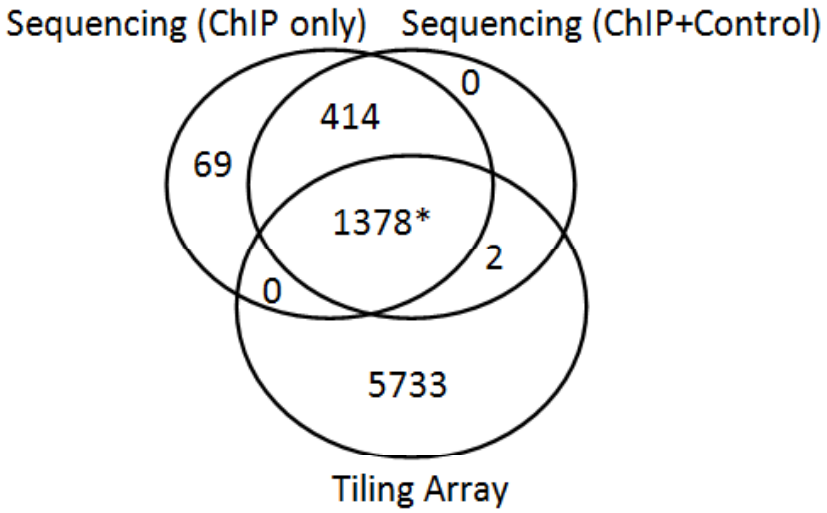
- NRSF ChIP-chip: 2 ChIP + 2 Mock IP in Jurkat cells, profiled using Affymetrix Human Tiling 2.0R arrays.
- NRSF ChIP-seq: ChIP + Negative Control in Jurkat cells, sequenced with the next generation sequencer made by Illumina/Solexa.

Overlaps in peaks

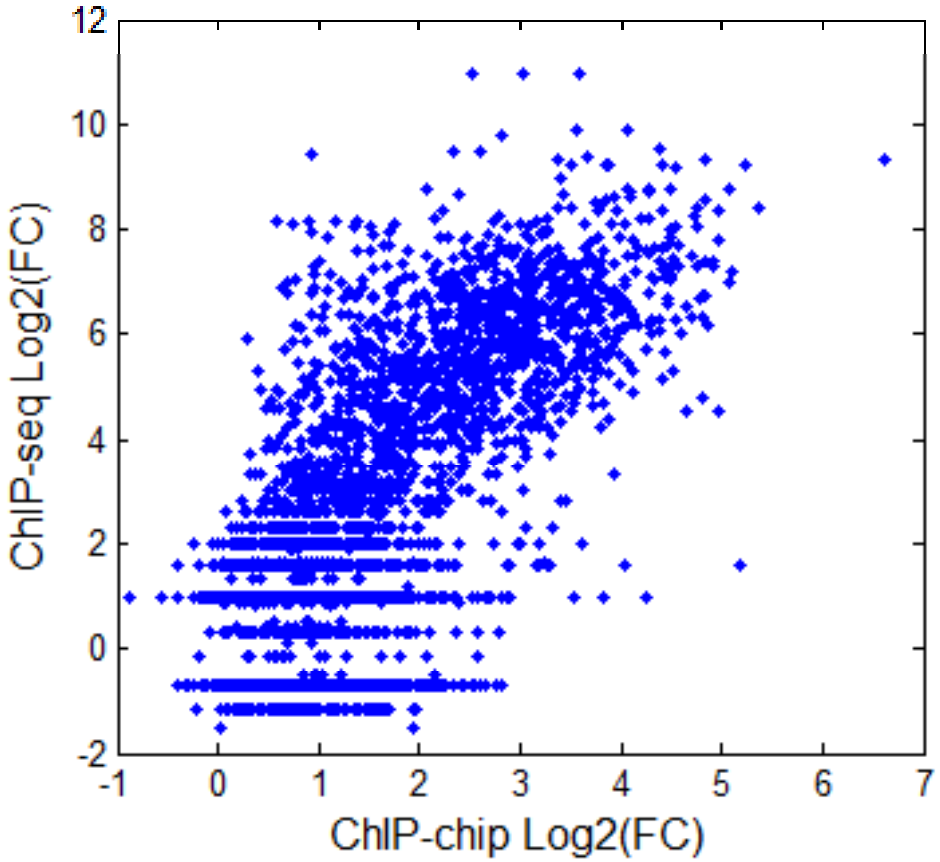
Before post-processing



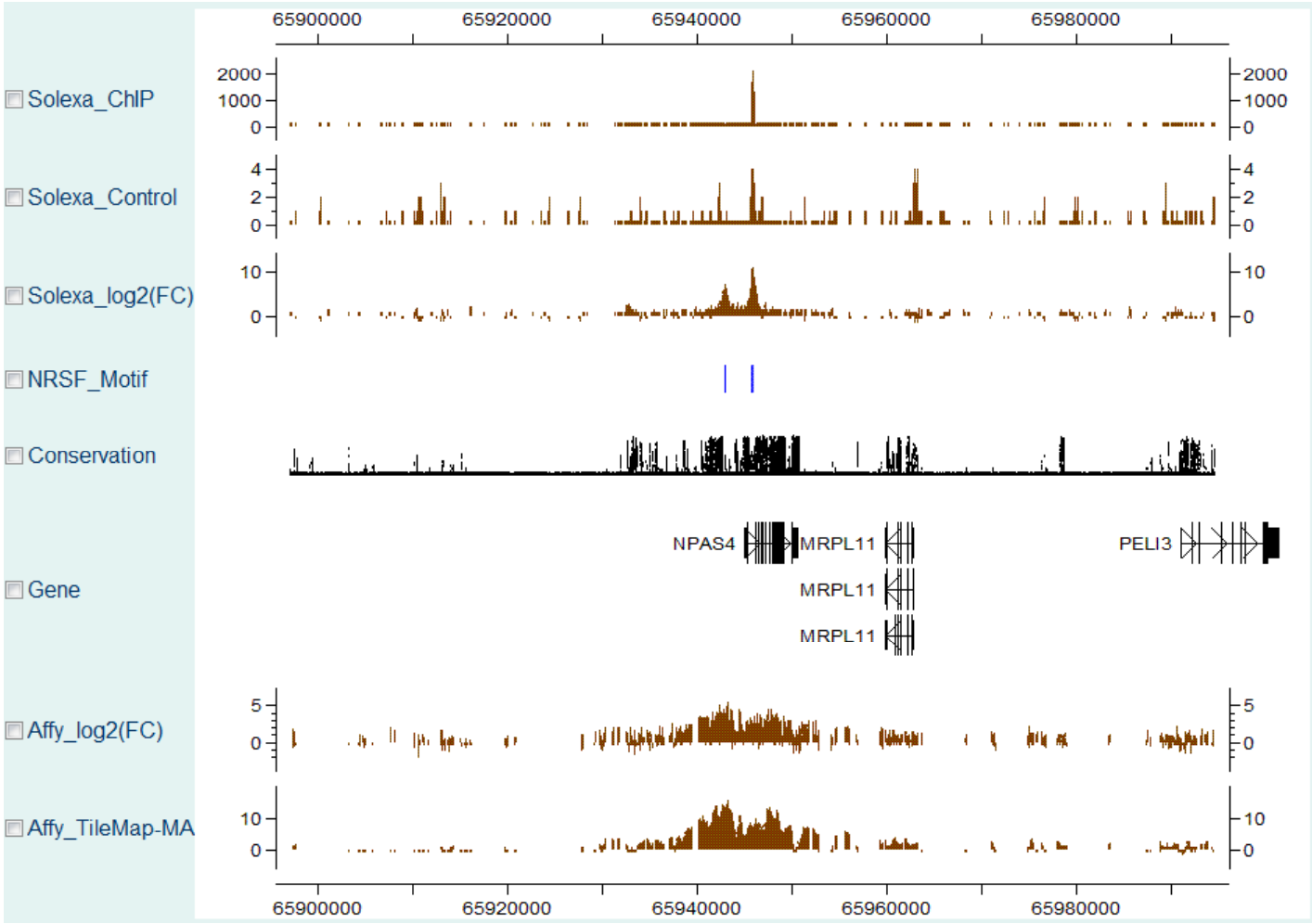
After post-processing



Correlation in signals



ChIP-seq provide better signals



Review

- ChIP-seq is used to detect interesting regions in the genome, same as tiling arrays.
- The enriched DNA segments is sequenced directly, instead of relying on hybridization in tiling arrays.
- Number of aligned reads are input data. More reads in a region indicate stronger signals (for TF binding, etc.).
- Single dataset peak detection is similar to that in ChIP-chip. Data in neighboring regions need to be combined.
- Joint analysis from multiple datasets if of great interests.