A *Bioconductor* pipeline for the analysis of ChIP-Seq experiments.

BioConductor 2013

Sangsoon Woo, Renan Sauteraud, Arnaud Droit, Xuekui Zhang,

Fred Hutchinson Cancer Reserach Center,
Seattle

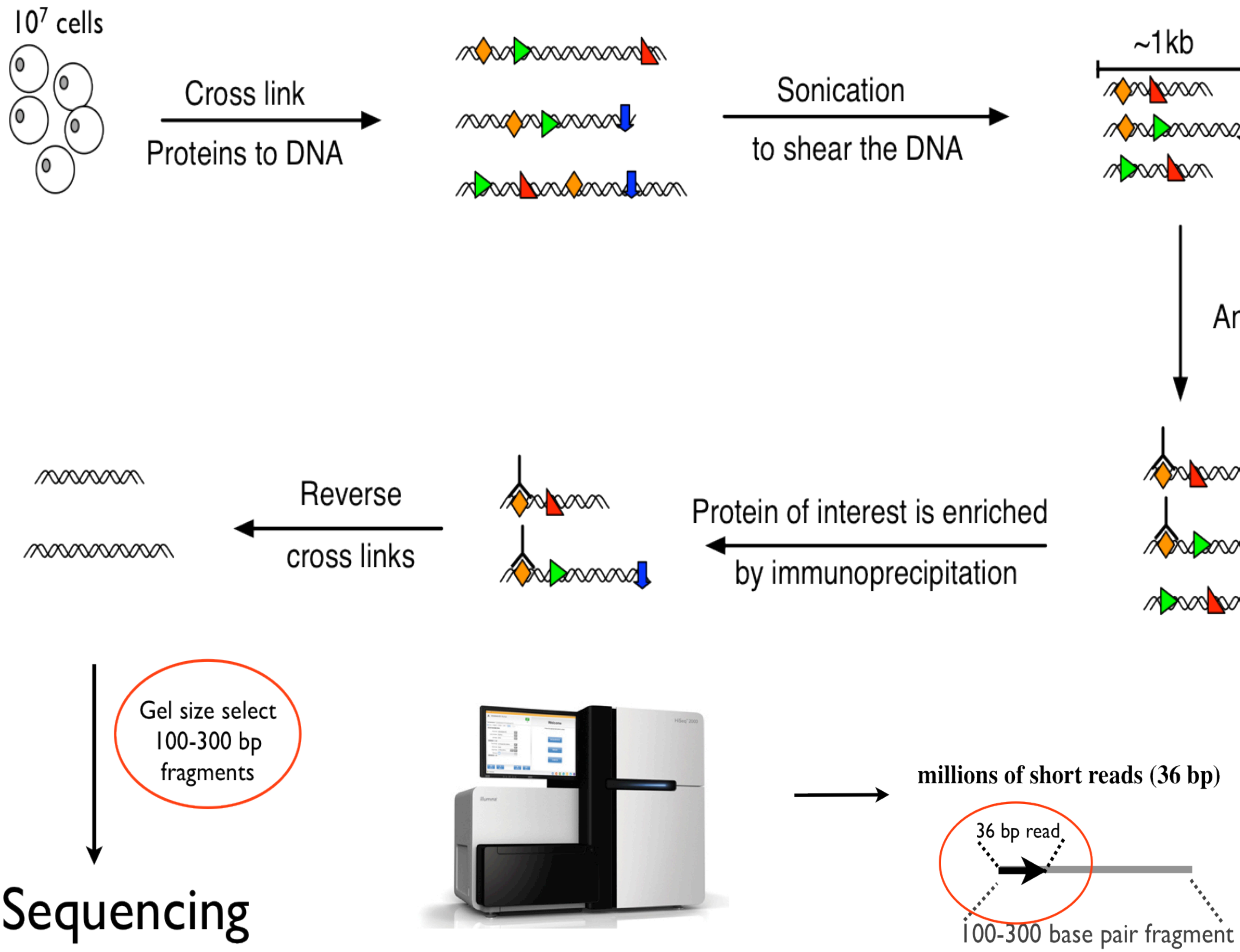
Outline

- Introduction of ChIP-Seq
- Transcription factor binding sites
- Real data example
- Nucleosome positioning

ChIP-Seq

- Couple ChIP with HTS
- A typical ChIP-Seq experiment generates tens of millions of short reads
- Read lengths are in the order of 50-150bps
- Because of chromatin, antibodies and alignment biases, a control sample is still recommended

ChIP-Seq



10^7 cells

Cross link

Proteins to DNA

Sonication

to shear the DNA

~1kb

Add Antibody

Protein of interest is enriched

by immunoprecipitation

Reverse

cross links

Gel size select 100-300 bp fragments

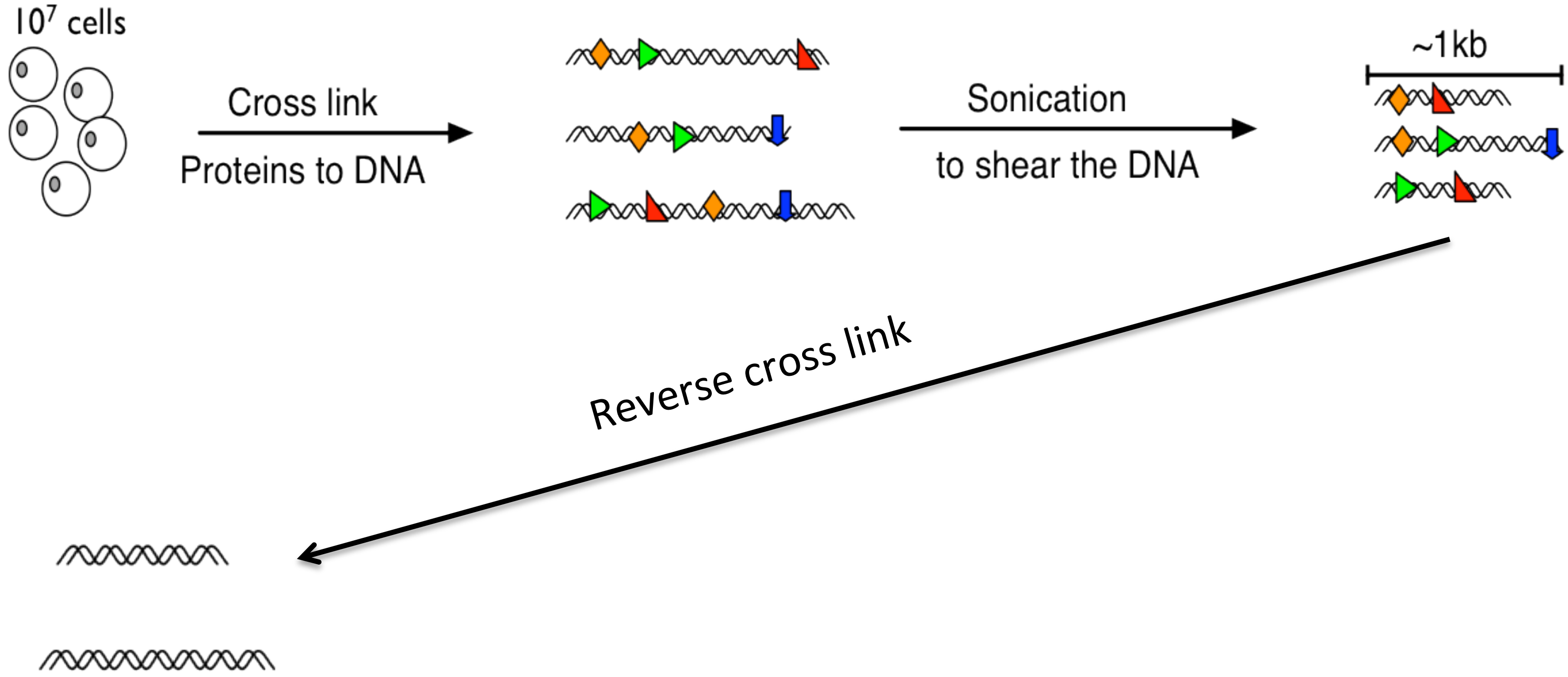
Sequencing

millions of short reads (36 bp)

36 bp read

100-300 base pair fragment

ChIP-Seq: control



10^7 cells

Cross link

Proteins to DNA

Sonication

to shear the DNA

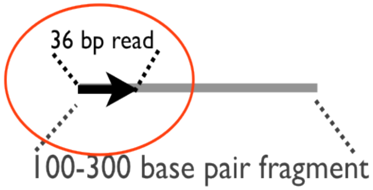
~1kb

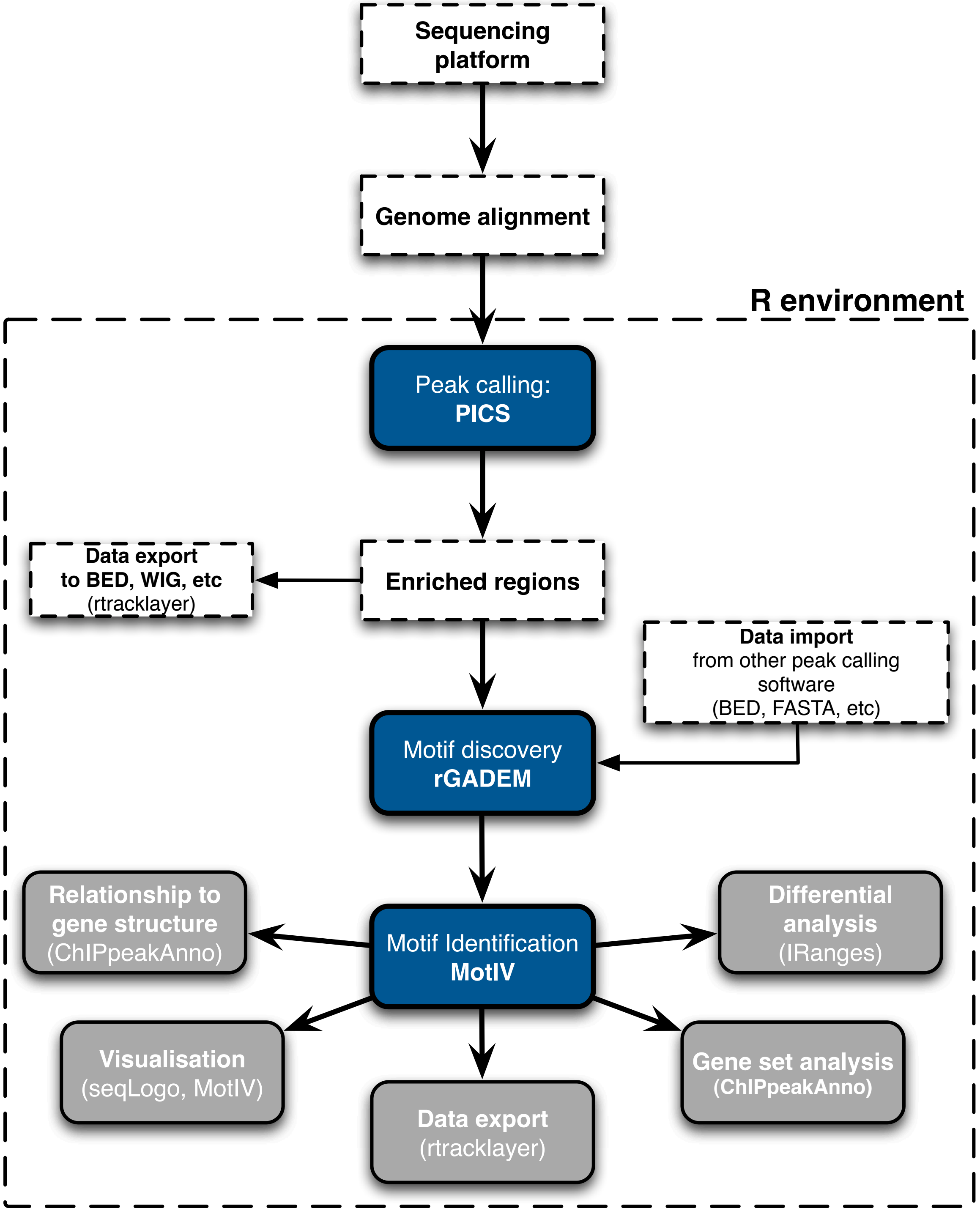
Reverse cross link

Gel size select 100-300 bp fragments

Sequencing

millions of short reads (36 bp)




Sequencing platform
Genome alignment
R environment
Peak calling: PICS
Data export to BED, WIG, etc (rtracklayer)
Enriched regions
Data import from other peak calling software (BED, FASTA, etc)
Motif discovery rGADEM
Relationship to gene structure (ChIPpeakAnno)
Motif Identification MotIV
Differential analysis (IRanges)
Visualisation (seqLogo, MotIV)
Data export (rtracklayer)
Gene set analysis (ChIPpeakAnno)

Aligners

- The first step consists of aligning raw reads to the reference genome.
- There exists numerous “aligners” or “mappers”
- Here are a few popular ones: Bowtie, BWA, ELAND, MAQ, etc
- Aligning raw reads of a sample can take from several minutes to several days (depends on data, software and cpu)
- Most aligners will perform “just fine” for ChIP- Seq

Aligned Reads

- Once reads have been aligned, we obtained a bed like file with *chromosome, start, end* and *strand* information for each sequence

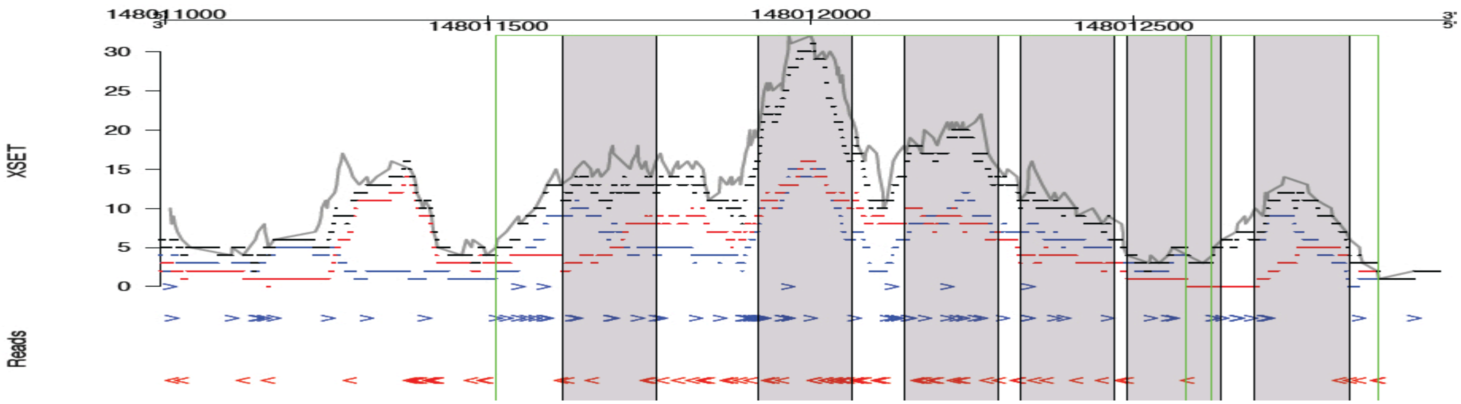
- Some reads cannot be uniquely aligned, and are typically discarded

- R and Bioconductor provide basic sequence alignment capabilities and great input support (Biostrings, ShortReads, Rsamtools)

- ShortReads can read most aligner data formats

Peak calling

- Aligned read data are transformed into a form that reflects local densities of immunoprecipitated DNA fragments → Peaks
- Estimate locations where transcription factors(TF) were associated with DNA → Peak summit
- Assign a score to each of these locations → Enrichment score
- Estimate a score threshold that leads to a desired false positive rate (or FDR) → thresholding



Peak callers for TF

- MACS → Yong Zhang et al
- cisGenome → Hongkai Ji et al
- USEQ → David Nix et al
- **PICS** (our approach)
- …

Why PICS?

- Measures of uncertainty
- Bidirectional reads
  - (Automatically pair forward peaks with reverse peaks, and estimate the DNA fragment length for **each** binding site)
- Correction for bias due to missing reads
- Resolve adjacent binding sites using mixture models
- Parallel running with multiple CPUs
- Implemented in BioConductor

PICS R package

- Perform the segmentation and PICS fitting
- Efficient implementation in C
- Parallel running with multiple CPUs
- Estimate the FDR and plot the FDR vs. score
- Export to bed/wig
- Can be fine tuned based on your fragment length distribution

Preprocessing

- Divide the genomic into regions by removing low reads regions

- Scan the genome every 10 pbs with a sliding window of size 150 bps

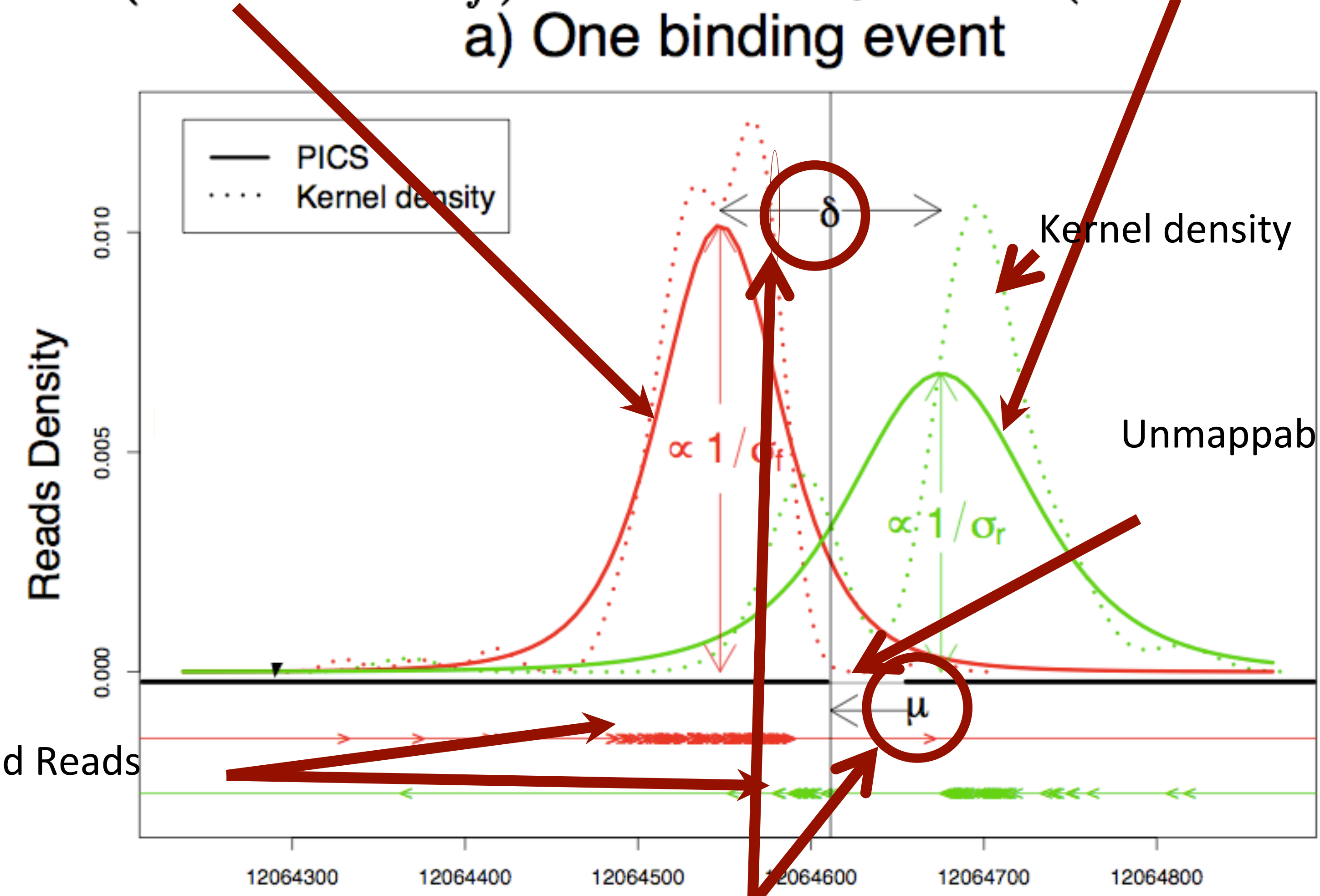
  - Minimum number of F reads on the left and R reads on the right
  - Merge overlapping regions

- N disjoint candidate regions

- Model each region separately and process them in parrallel

Modeling bi-directional reads

$$f_i \sim t_4\left(\mu - \delta/2, \sigma_f^2\right) \qquad r_j \sim t_4\left(\mu + \delta/2, \sigma_r^2\right)$$



Kernel density

Unmappable region

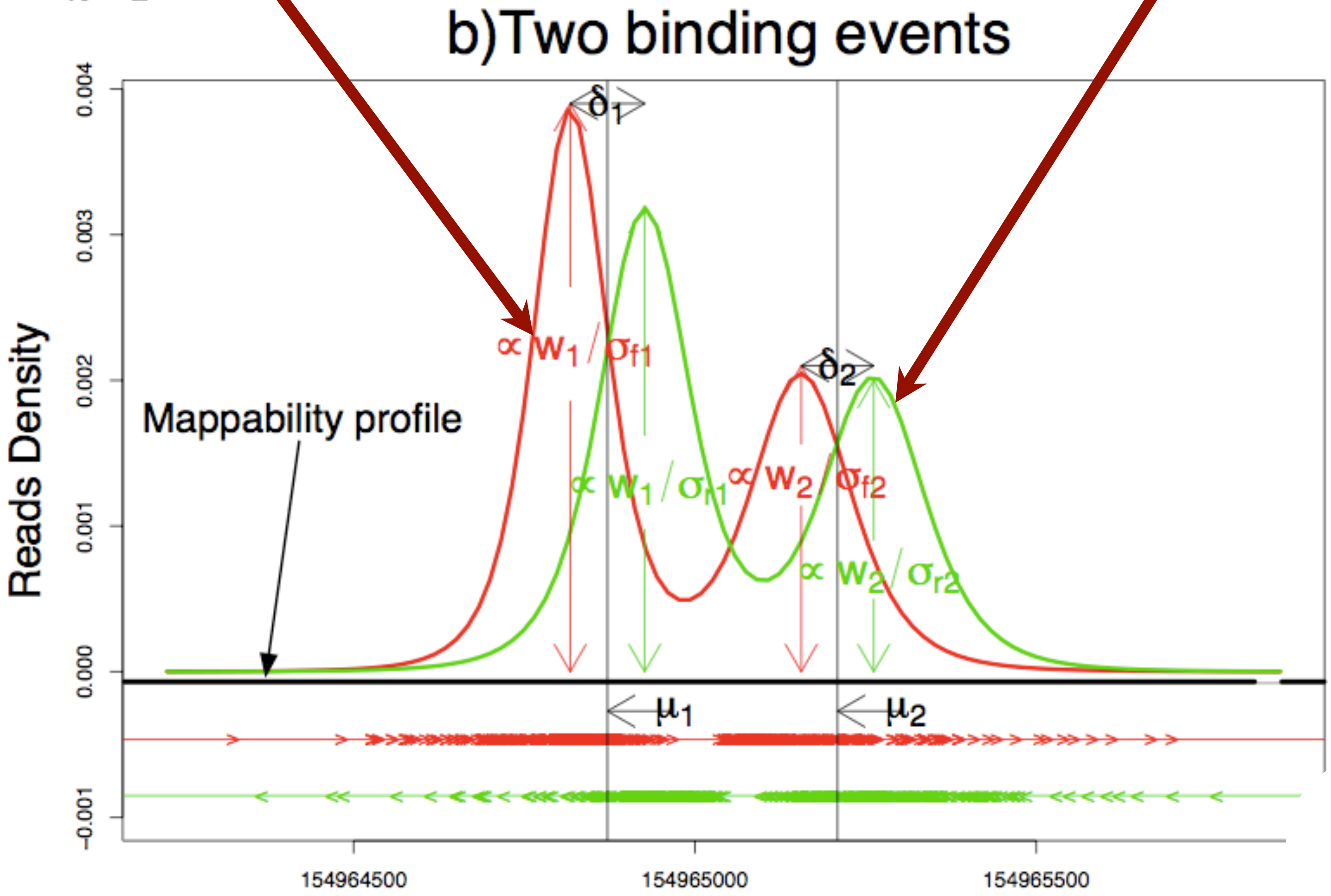
Aligned Reads

μ : TF binding site position

δ : average fragment length

Modeling bi-directional reads

$$f_i \sim \sum_{k=1}^{K} w_k t_4\left(\mu_{fk}, \sigma_{fk}^2\right) \qquad r_j \sim \sum_{k=1}^{K} w_k t_4\left(\mu_{rk}, \sigma_{rk}^2\right)$$



$$\mu_{fk} = \mu_k - \delta_k/2 \qquad \mu_{rk} = \mu_k + \delta_k/2$$

Parameter estimation

- Use an ECM type algorithm
- E-step: Missing data are the cluster memberships and the weights of the normal distribution. Explicite formulation for the E-step
- Mstep: No closed form estimates, so split into two M steps

Prior distributions

- Use Normal Inverse Gamma conjugate prior for computational convenience

$$\sigma_{fk}^{-2}, \sigma_{rk}^{-2} \quad \sim \quad \mathcal{G}a(\alpha, \beta)$$

$$(\delta_k | \sigma_{fk}^2,\ \sigma_{rk}^2) \quad \sim \quad \mathrm{N}(\xi, \rho^{-1}/(\sigma_{fk}^{-2} + \sigma_{rk}^{-2}))$$

- Hyper-parameters are chosen to match our prior knowledge (eg. DNA fragment length 80-300 bps)

The missing reads – the problem

- Genome is made of a short alphabet (A,G,C,T), hence sequence repeats can occur! So many short reads are discarded due to no uniquely aligned positions.

- The amount of missing reads is unknown in each unmappable region.

- Boundaries of unmappable regions are known -- (the 0/1 mappability profile obtained by exhaustive enumeration)

The missing reads – our solution

- Use an idea of McLachlan and Jones (1998) for grouped and truncated data -- introducing latent variables:
  - amount of missing reads (negative multinomial)
  - positions of missing reads (same dist'n as observed reads)

- We use EM algorithm for fitting hierarchical mixture models incorporating these latent variables

Scoring binding events

- Compute an enrichment score to rank and identify an interesting list of binding events.

- The enrichment score is defined as the ratio (IP/ Control) of the observed F/R reads falling in the 90% contours of the F/R distributions.

- By swapping the IP/Control samples, we can get an estimate of the number of false positives for a given threshold, and thus compute an estimate of the FDR

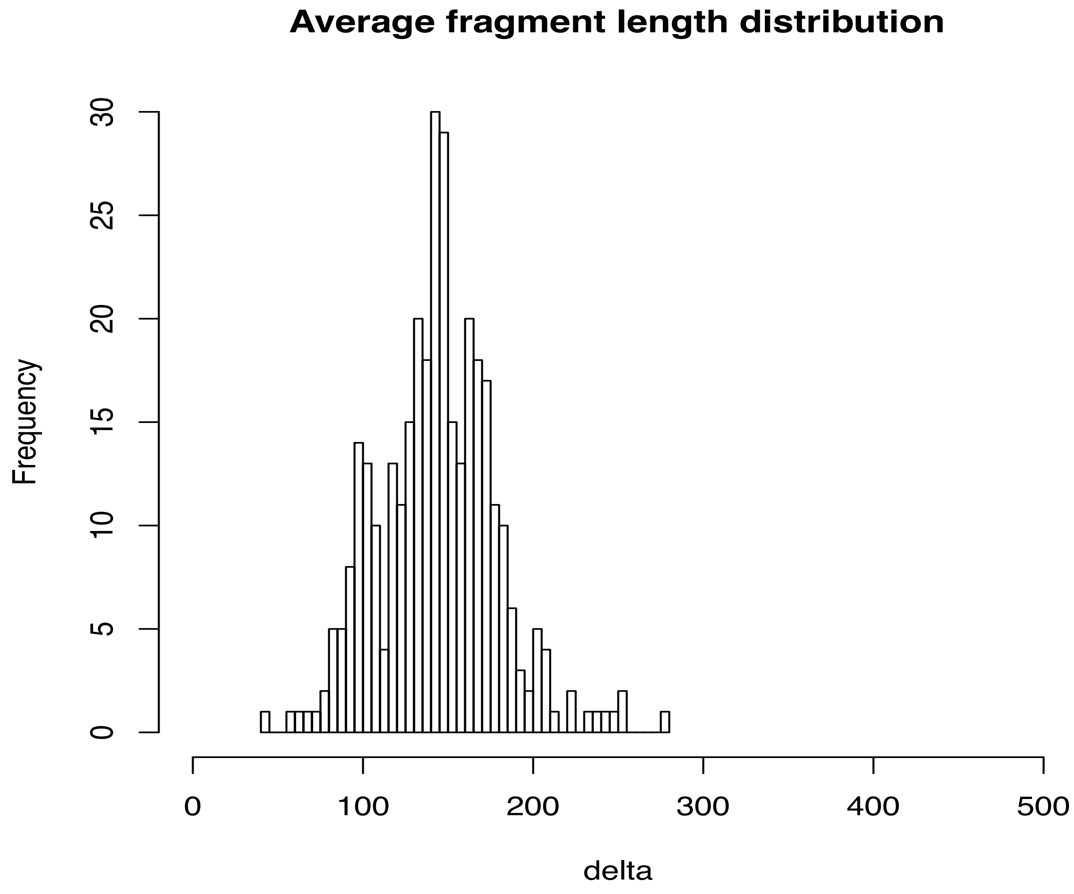
Application to ER and FOXA1

- FOXA1 data in human MCF7 human cells (Zhang et al., 2008).
- 3,909,507 ChIP-seq reads and 5,233,322 input DNA control reads
- ER data data in human MCF7 human cells (Hu et al., 2010)
- Use: PICS, rGADEM and MoTiV

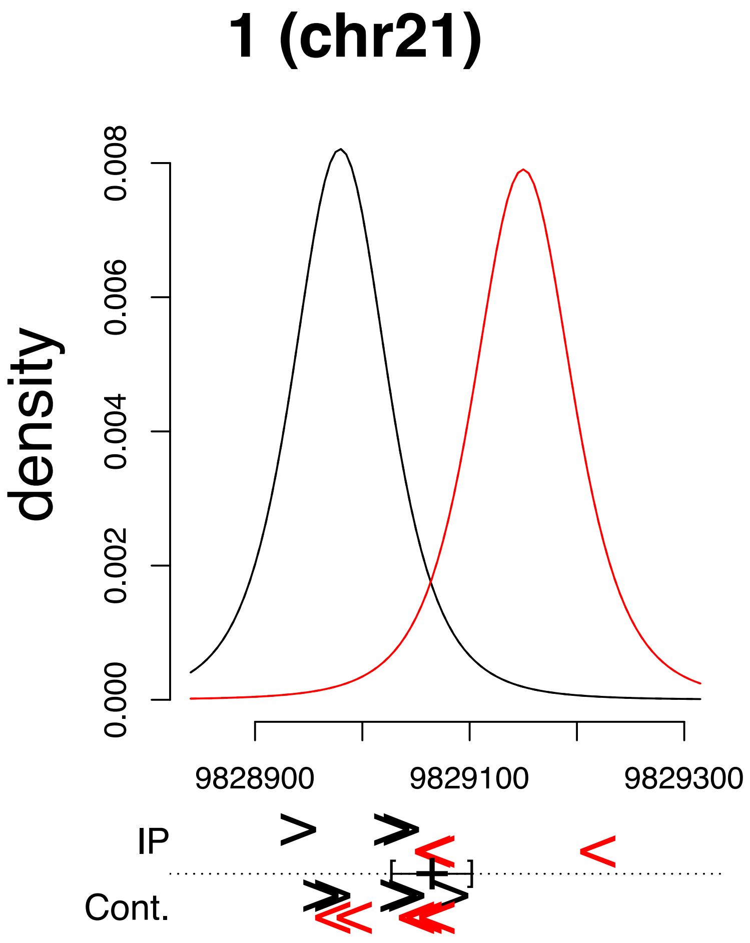
Package ChipSeqBioC

- Packages:
  - ShortRead: to read data
  - BSGenome: to access genomic information
  - PICS: to identify peak list
  - rGADEM: de novo motif discovery
  - MotIV: motifs identifications
  - Rtracklayer: visualisation: interface to genome browser
  - GenomeGraphs: visualisation
  - Gviz: visualisation
  - PING: to identify nucleosome positioning

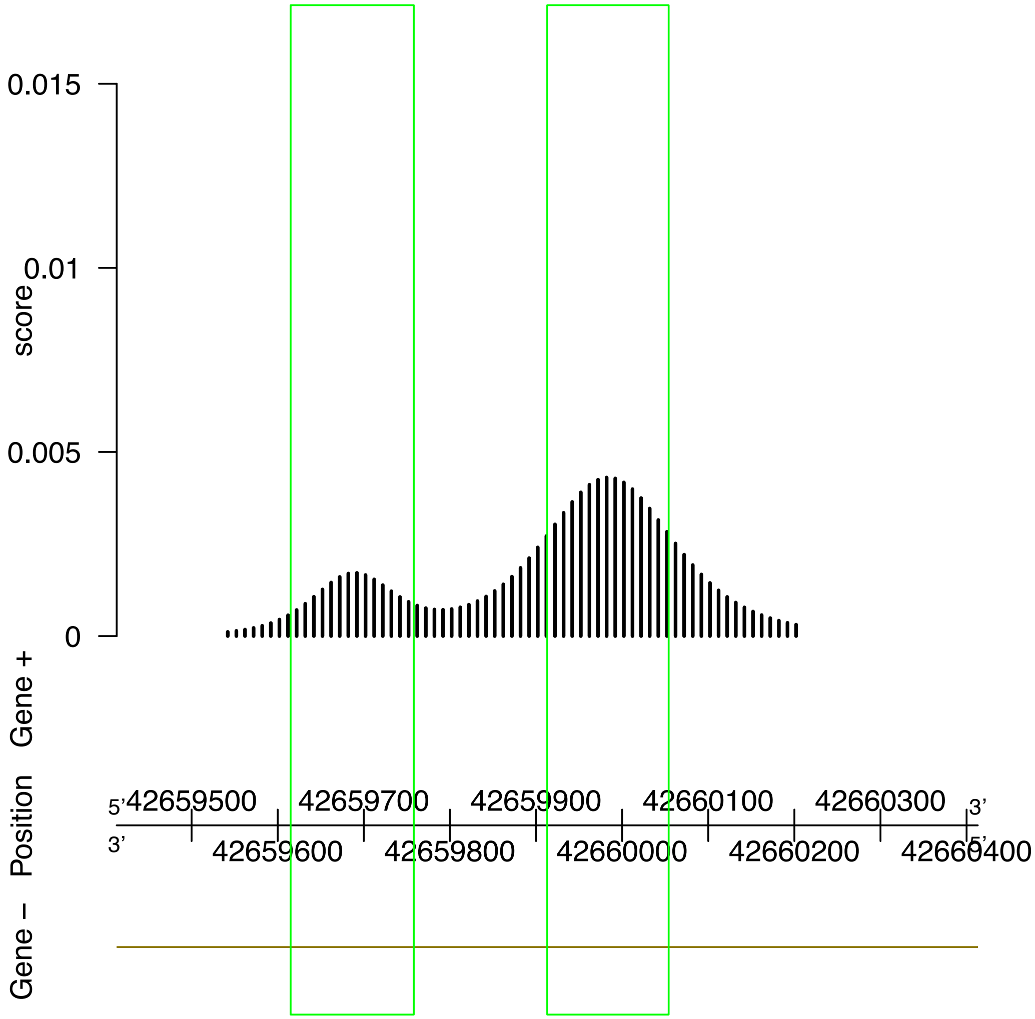
Average fragment length distribution



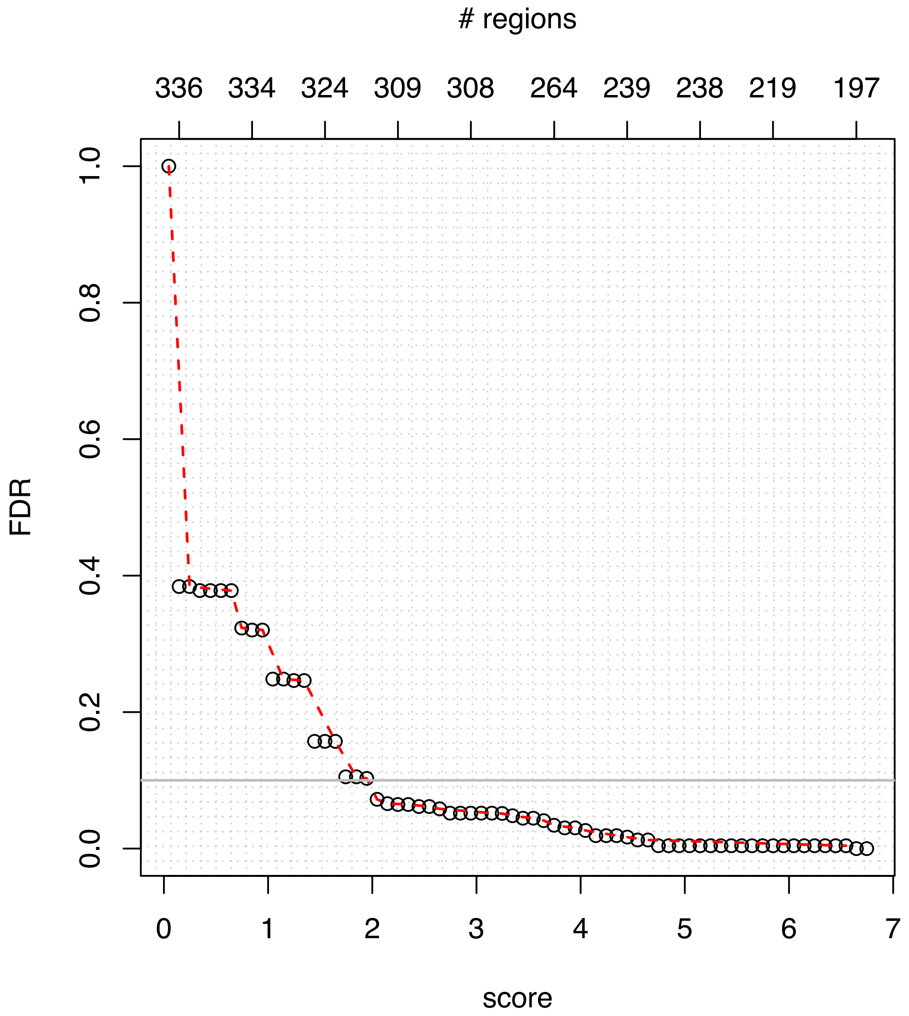
Visualizing candidate region



Vizualisation: GenomeGraphs

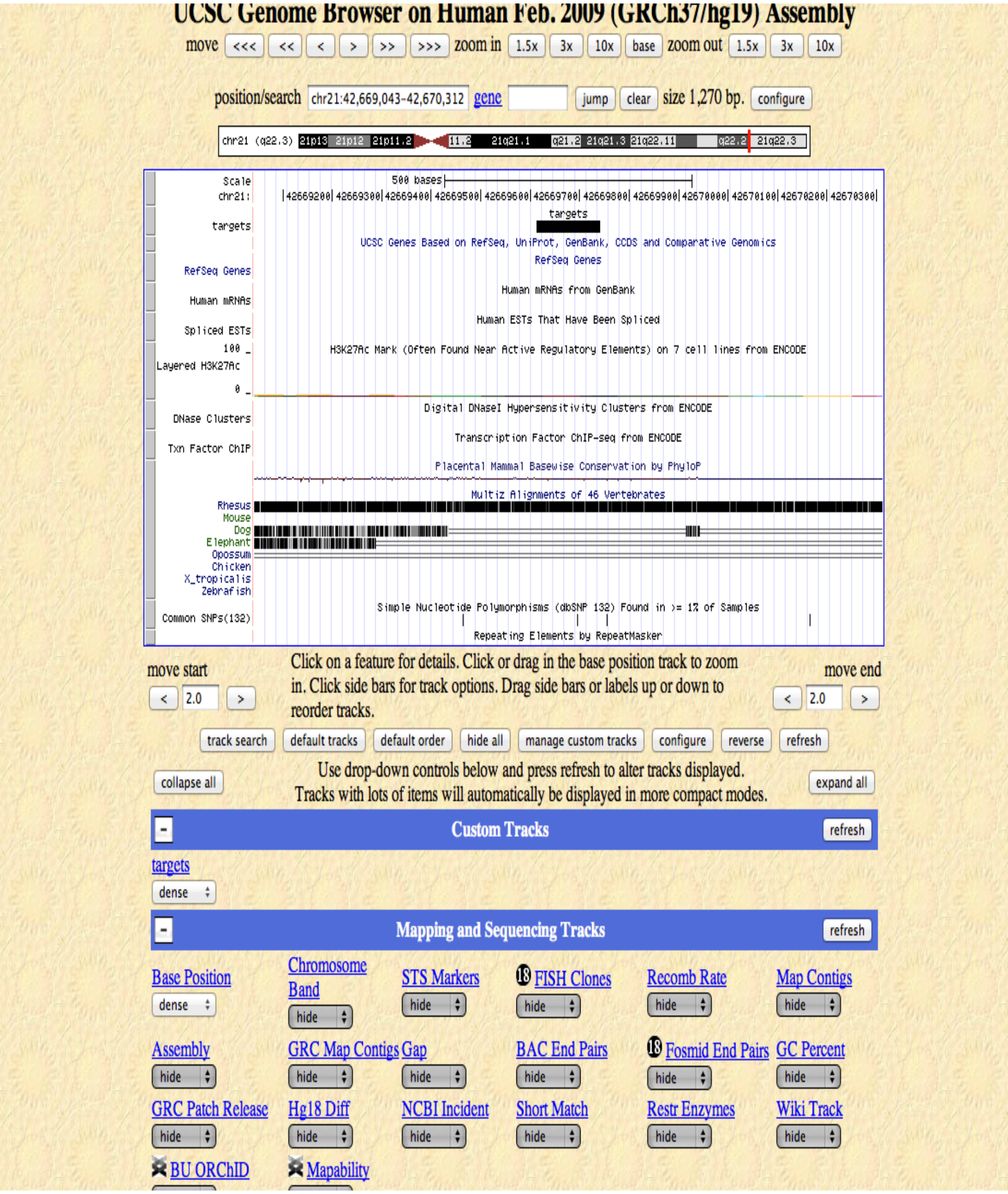


FDR



# regions
336
334
324
309
308
264
239
238
219
197
FDR
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
score

Vizualisation: rtracklayer



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

position/search chr21:42,669,043-42,670,312 size 1,270 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks.

Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

Validation

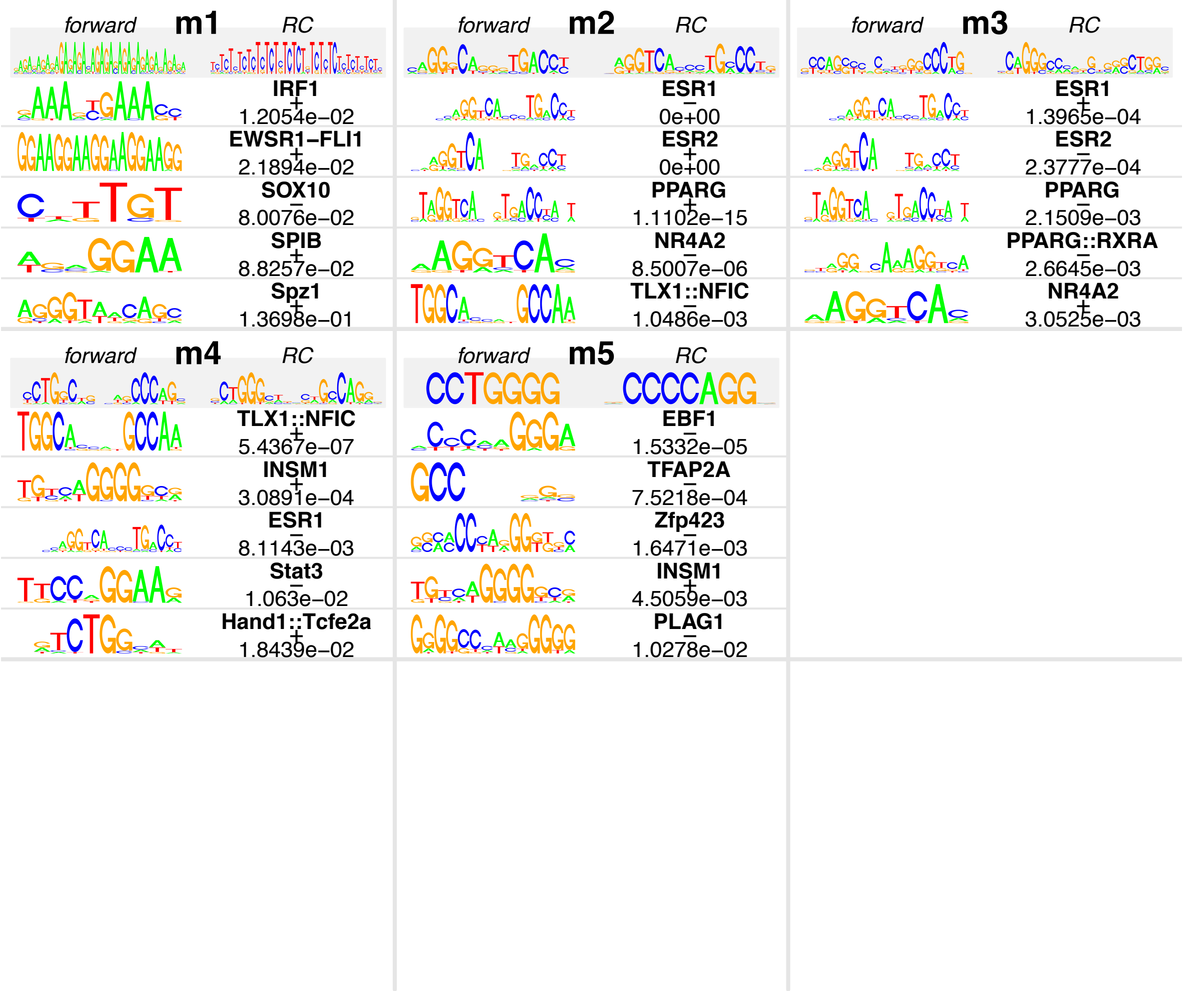
- *de novo* motif search

- rGADEM is fast and can be used to process 10K+ sequences (binding site estimates +/- 100bps)

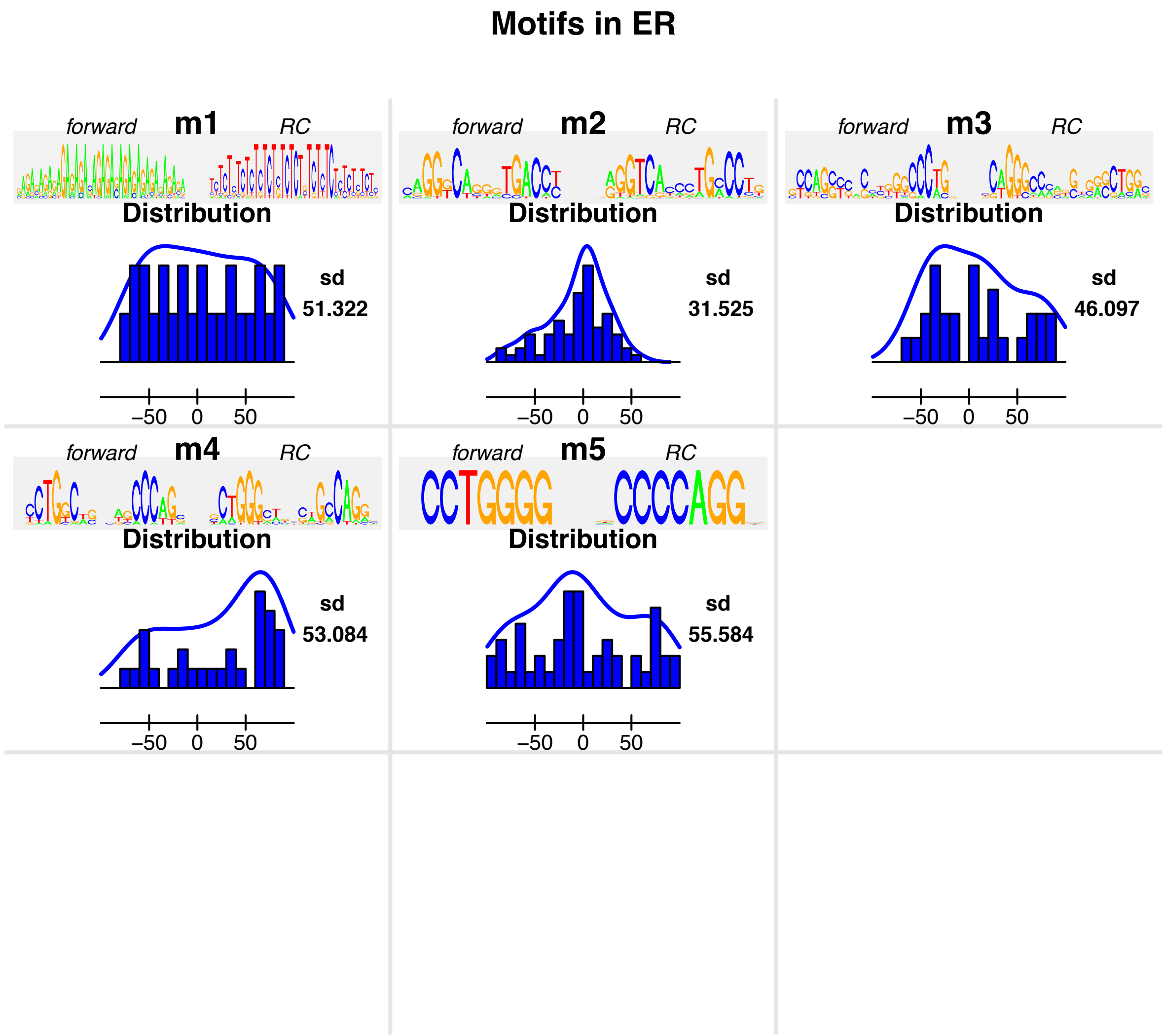
- Identified motifs were then fed into MotIV and analyzed with Jaspar

rGADEM + MoTiV results

Motifs in ER



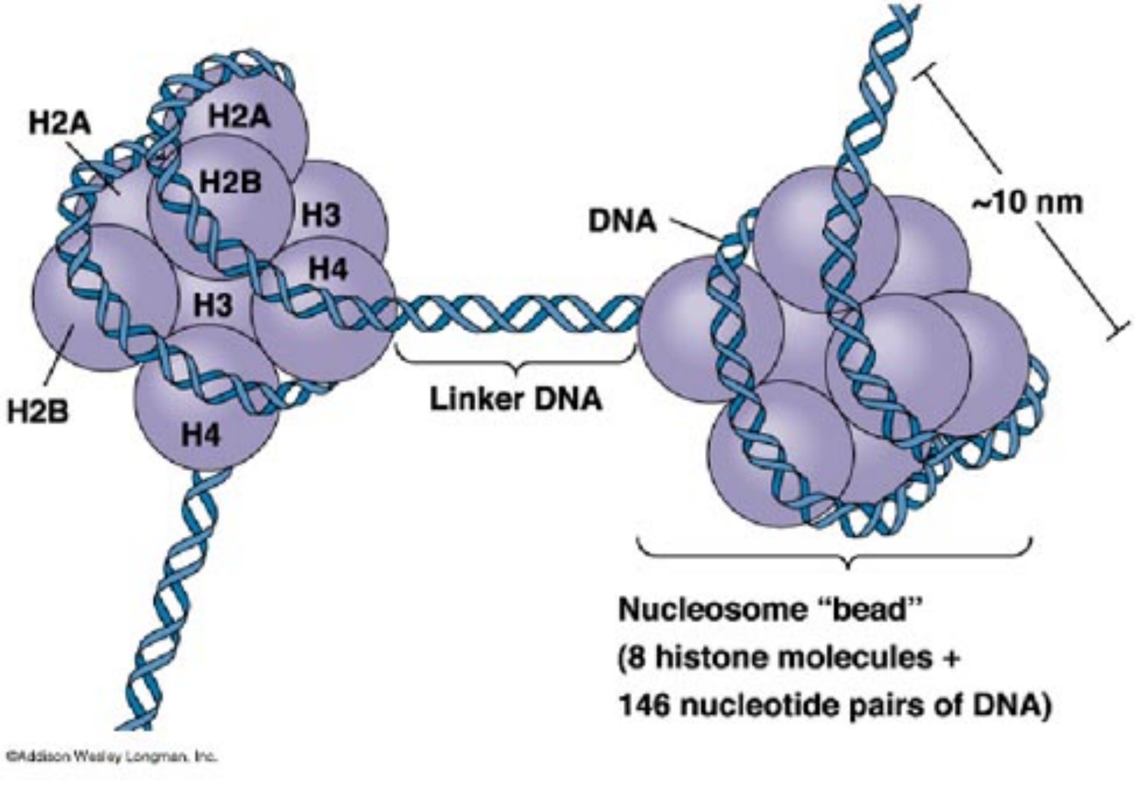
rGADEM + MoTiV results



The biology – nucleosomes (1)

- The nucleosome core particle (shown in the figure) consists of about 147 bps of DNA wrapped around the histone octamer. (H2A, H2B, H3, and H4)

- Adjacent nucleosomes are joined by 10-80 bp of 'linker' DNA.



The biology – nucleosomes (2)

- DNA wrapped around nucleosomes is less accessible to DNA binding proteins. Hence nucleosomes can regulate processes that require access to DNA.

  e.g. DNA replication or transcription

- Many gene regulatory proteins interact with nucleosomes, such as modifying amino acids on N-terminal histone tails.

- So genome-wide profiling nucleosome positions is important in understanding how transcriptional machinery functions in vivo.

PING

- We developed a new method, PING, for identifying nucleosome positioning from sequencing data.

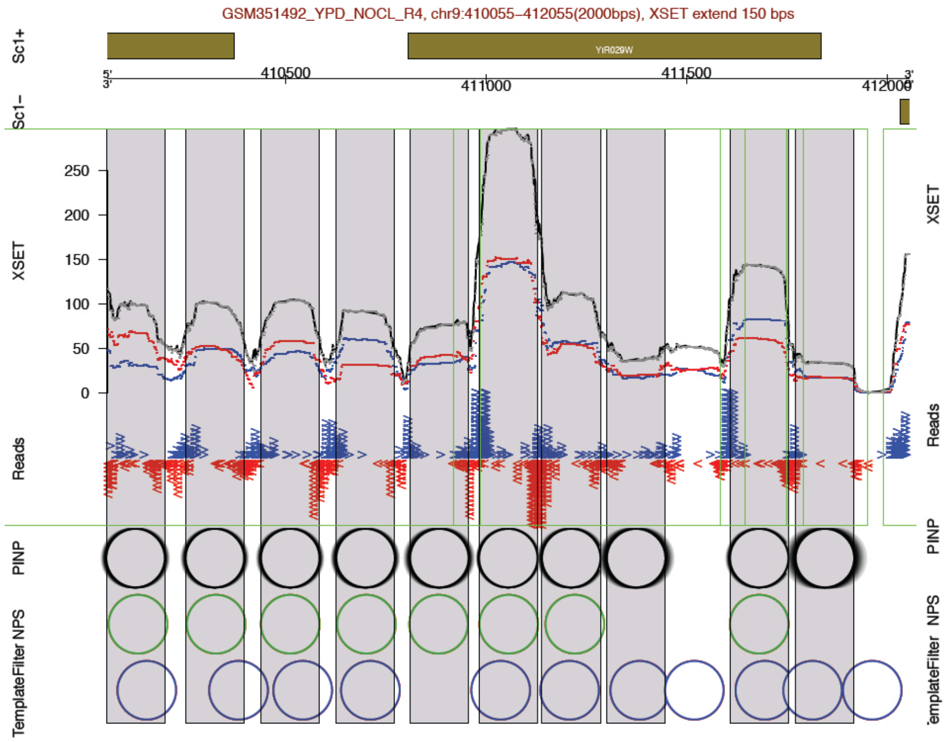
- PING is developed based on PICS framework, hence inherits all PICS features discussed above.

- PING is different from PICS in:
  - Address spatial relations of nucleosomes (Gaussian Markov Random Field (GMRF) prior on nucleosome locations)
  - Other details. (New segmentation, new model selection criteria, new tuning parameters, and additional post-process step)

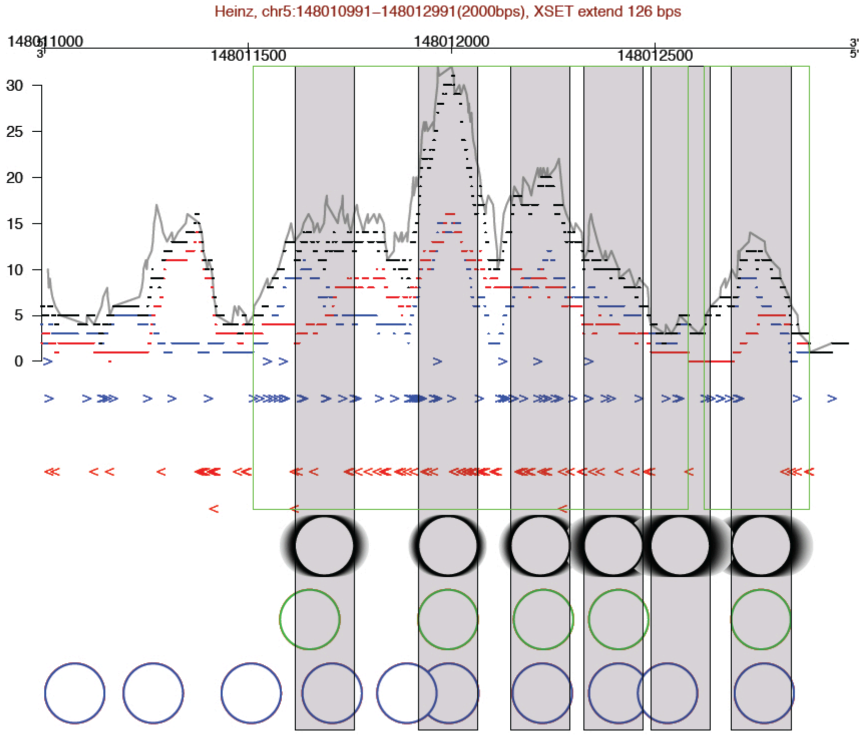
PING features

- PING handle data from large genome (e.g. mammal) in ~ 1hr.
- PING is robust to low read densities (simulation comparisons shown later)
- PING handle both Sonication data and MNase data

MNase-seq



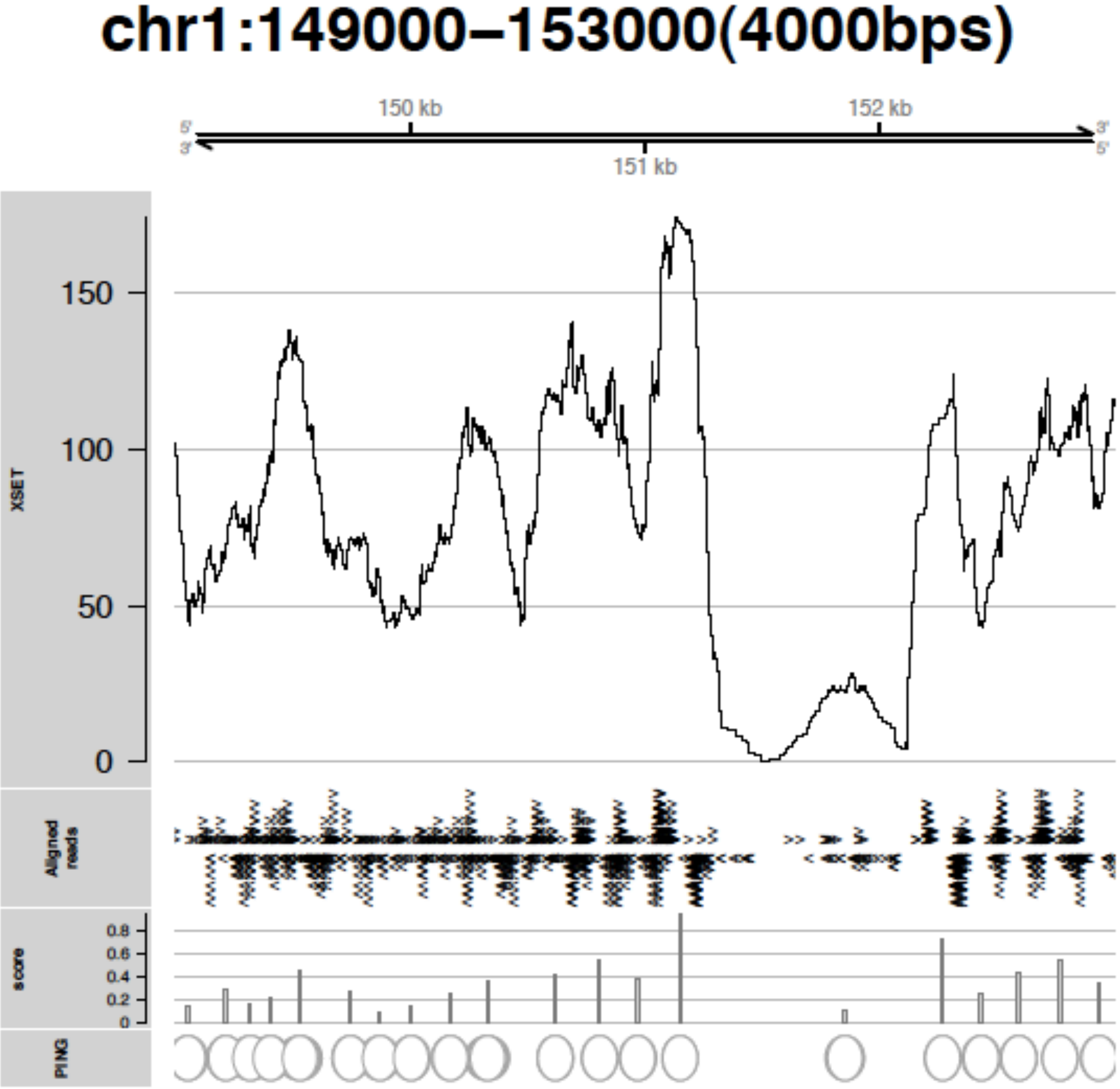
Sonicated ChIP-seq



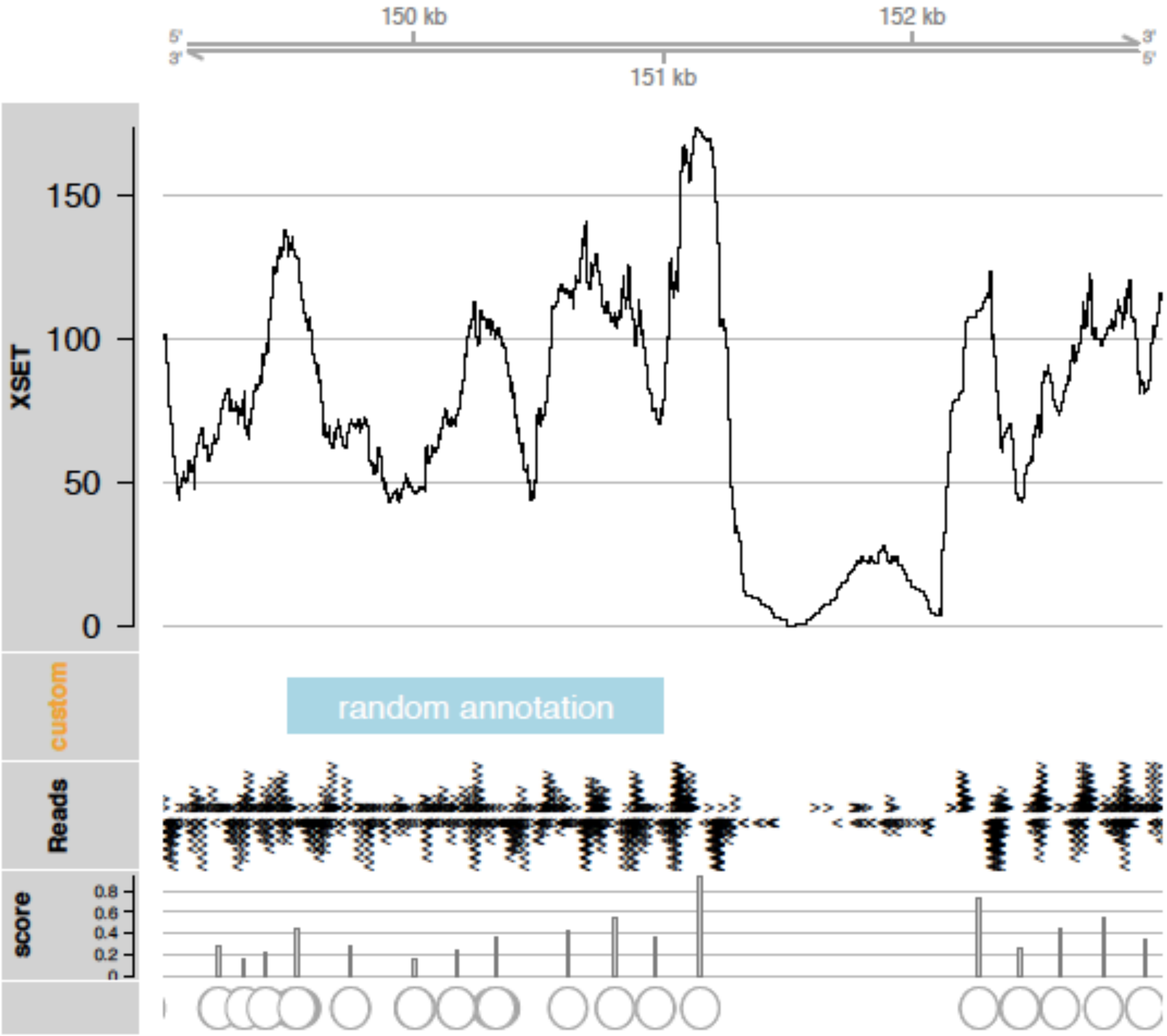
PING R package

- Work for MNase and Sonicated with Single-End and Paired-End sequencing data
- Perform the segmentation and PING fitting
- Efficient implementation in C
- Parallel running with multiple CPUs
- Export PING and postPING results to bed/wig
- Built-in plotting function for Visualization

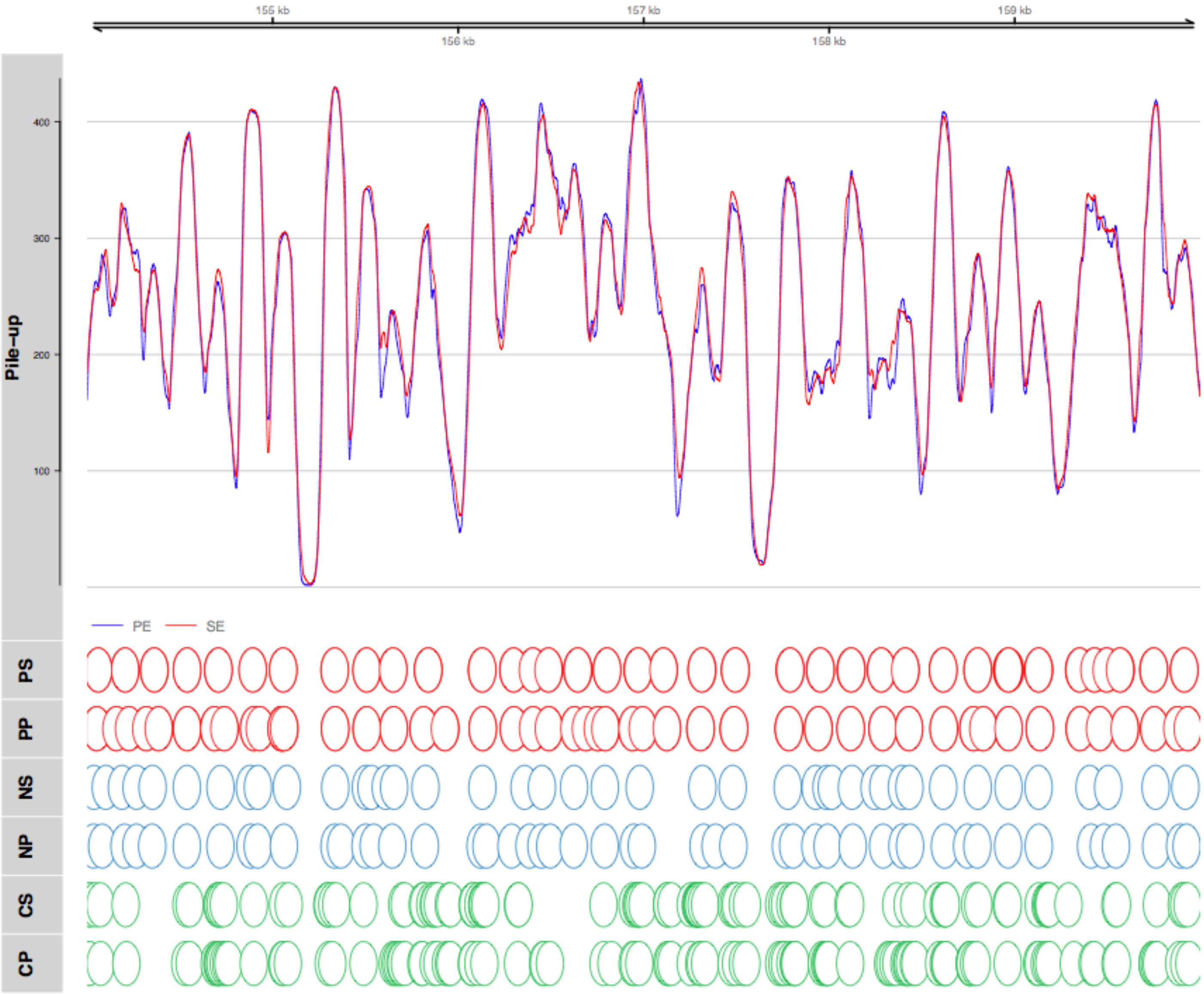
plotSummary()



Custom plot with Gviz



Custom plot with Gviz



Conclusions

- ChIP is a powerful tool
  - Transcription factors
  - Epigenetics/Epigenomics

- Statistics/Bioinformatics challenges
  - Alignment, detecting binding events, etc
  - Still many challenges with ChIP-Seq