High-throughput chromatin immunoprecipitation assays

BioConductor 2011

Xuekui Zhang, Eloi Mercier and Arnaud Droit
Fred Hutchinson Cancer Reserach Center, Seattle

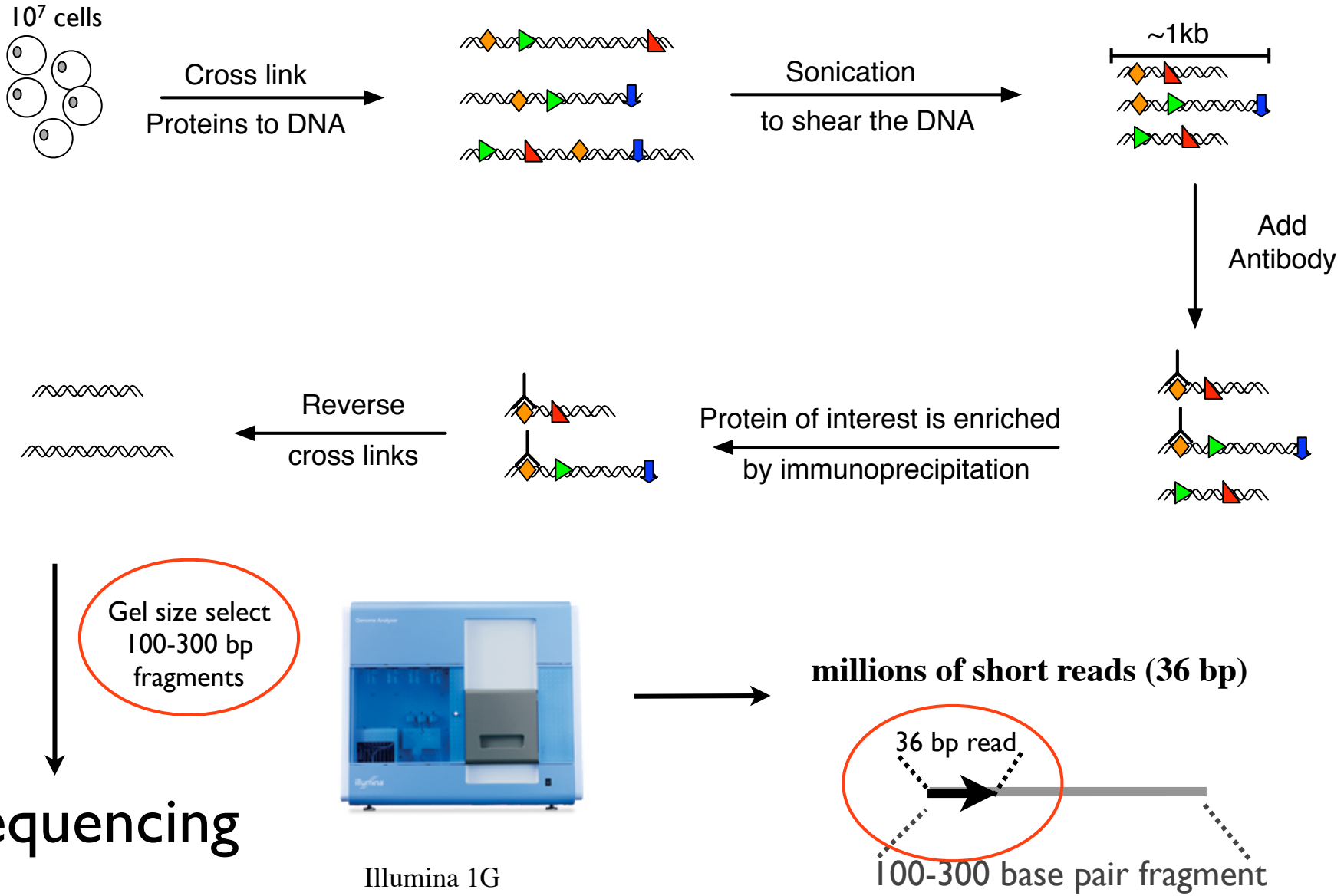
Outline

- High throughput sequencing and ChIP-Seq
- A probabilistic approach to the analysis of ChIP-Seq
- Example on real data and comparison to other approaches

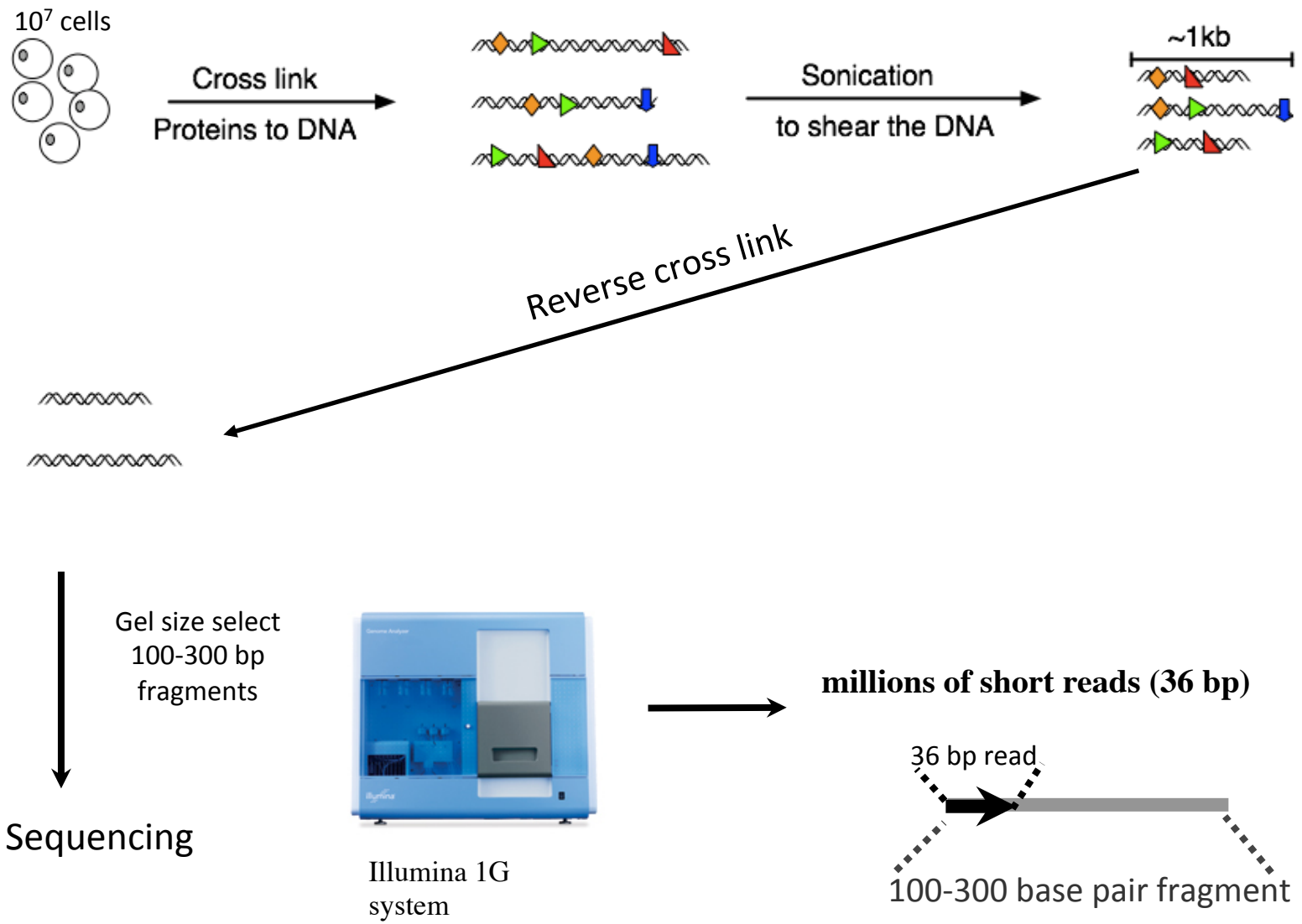
ChIP-Seq

- Couple ChIP with HTS
- A typical ChIP-Seq experiment generates 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
Illumina 1G
millions of short reads (36 bp)
36 bp read
100-300 base pair fragment

ChIP-Seq (Control sample)



Aligners

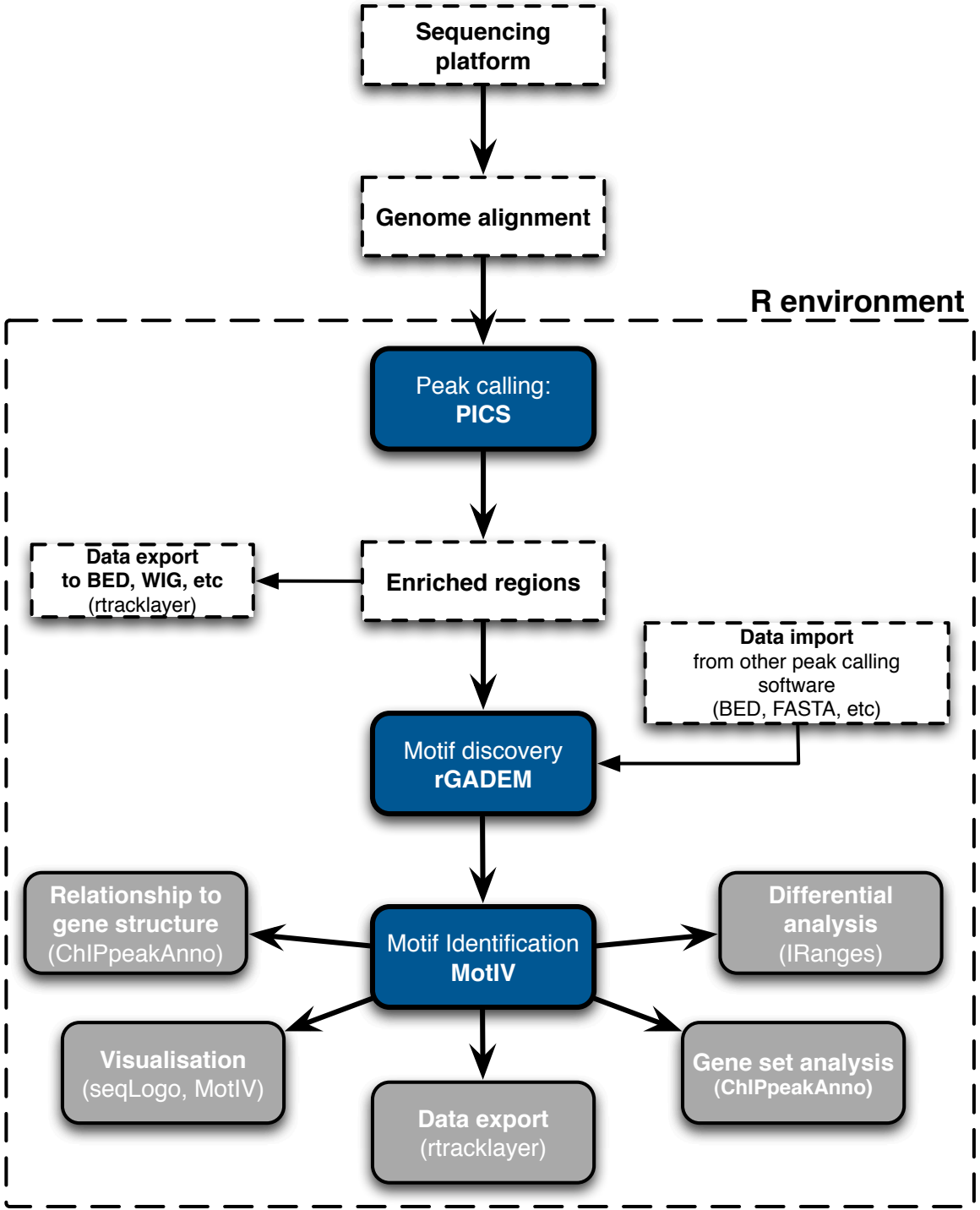
- The first step consists of aligning raw reads to the reference genome.
- There exists numerous “aligners” or “mappers”
- Here are a few popular ones: ELAND, MAQ, Bowtie, SOAP, Gnumap, etc
- Aligning raw reads can take from several hours to several days (depends on data, software and cpu)
- Most aligners 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)
- ShortReads can read most aligner data formats

Peak calling

- Aligned read data are transformed into a form that reflects local densities of immunopre- cipitated DNA fragments → Peaks
- Estimate locations where transcription factors 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


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)

Peak callers

- MACS

Yong Zhang et al. (2008) Model-based Analysis of ChIP-Seq (MACS), *Genome Biology 9:R137*

- cisGenome

Hongkai Ji et al. (2008) An integrated system CisGenome for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol 26(11) 1293-1300*

- …

- PICS

Xuekui Zhang et al. (2010) PICS: Probabilistic Inference for ChIP-seq. *Biometrics* 67(1):151-163

Our approach: PICS (Probabilistic Inference of ChIP-Seq data)

- Probabilistic model:
  - Model binding events
  - Use prior information (Fragment length distribution)
  - Bidirectional reads
- Measures of uncertainty
- Estimation of missing reads
- Resolve adjacent binding sites using mixture models
- Implemented in BioConductor

PICS R package

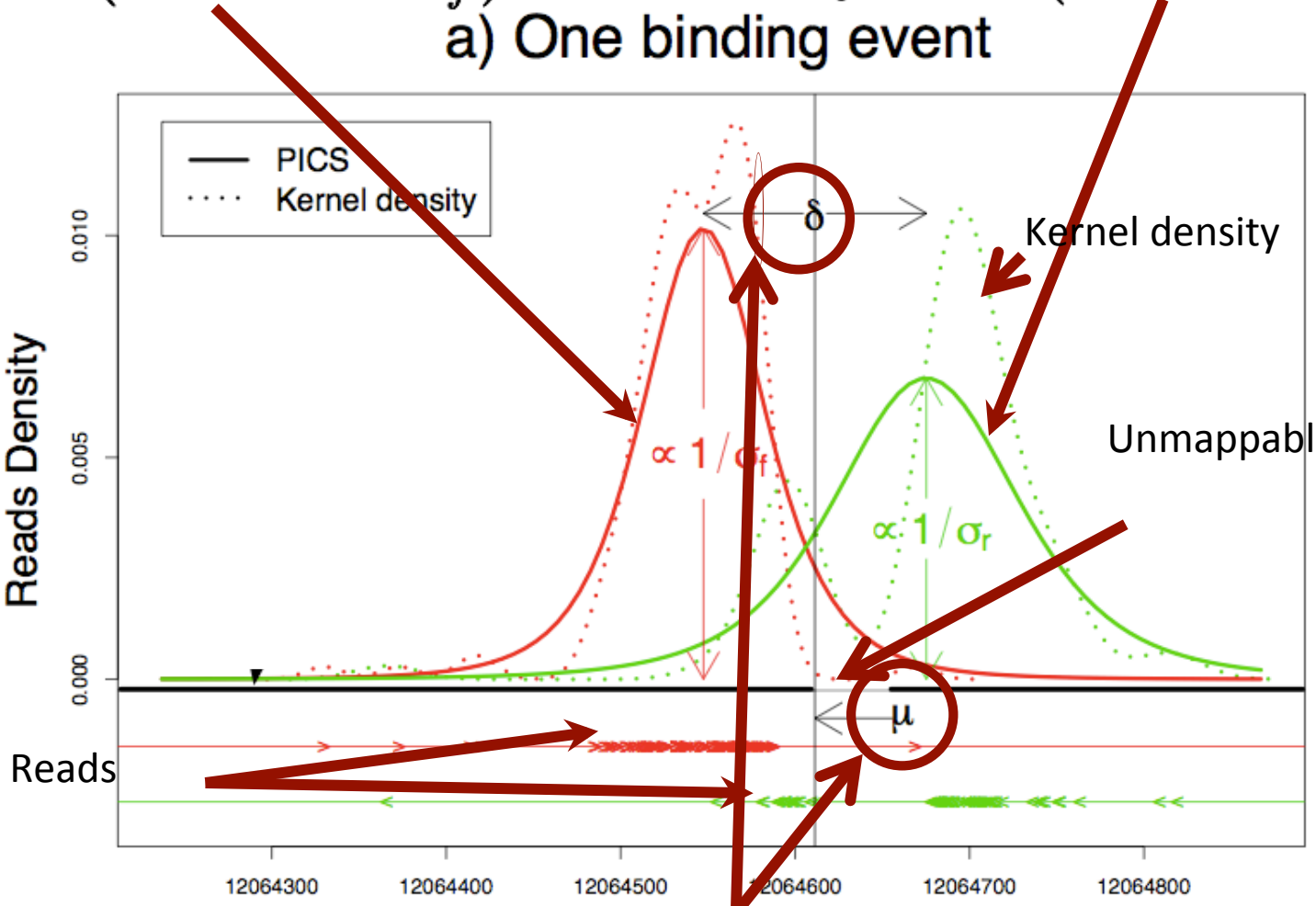
- Perform the segmentation and PICS fitting
- Efficient implementation in C and
- Can be run in parallel using multiple CPUs
- Estimate the FDR and plot the FDR vs. score
- Export to bed/wig
- Can be fine tuned based on your fragment length distribution

Preprocessing

- Divide the genomic reads into regions with “enough” F/R reads
- 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 regions
- Model each region separately

Modeling bi-directional reads

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

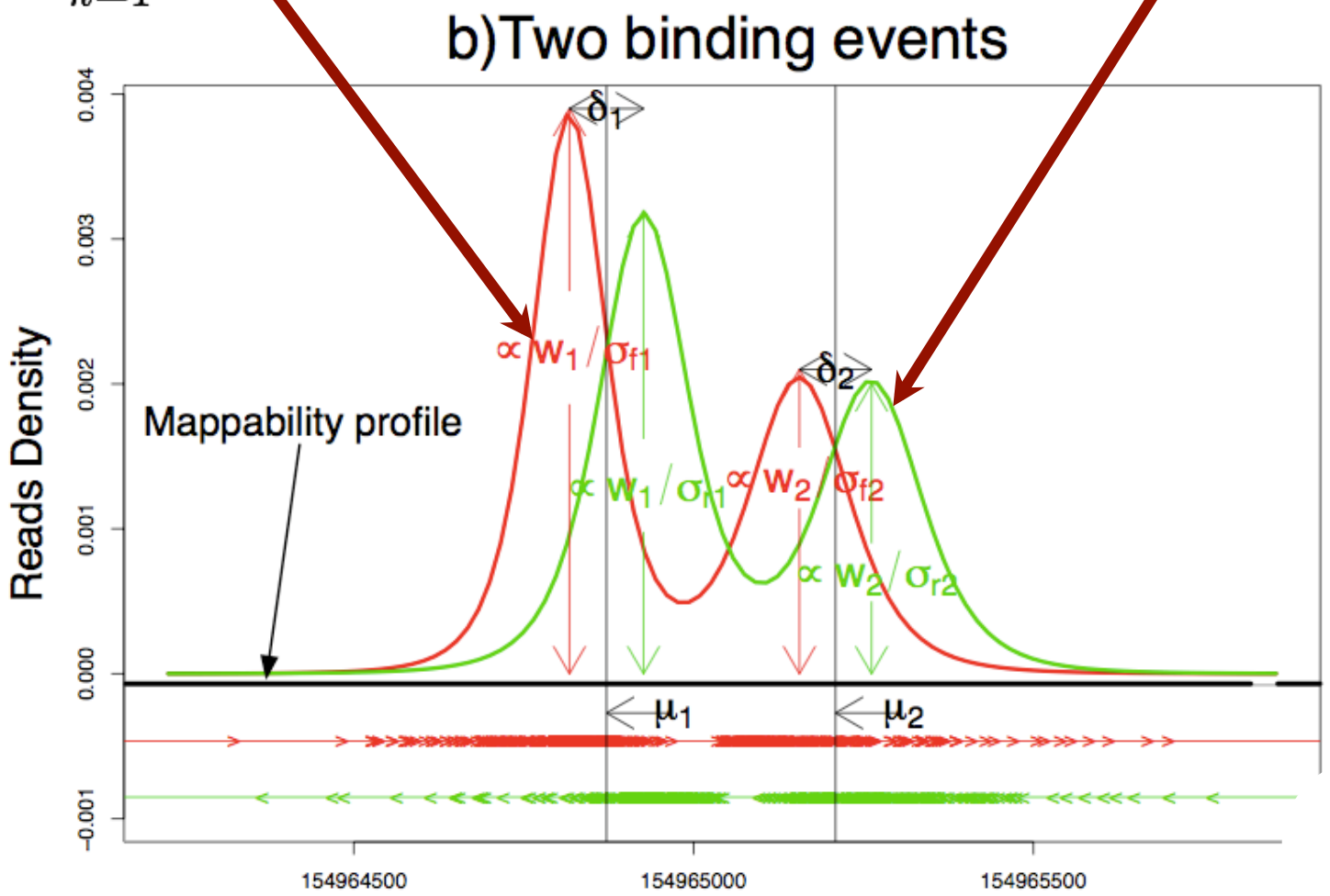


μ : TF binding site position

δ : average fragment length

Modeling bi-directional reads

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



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

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

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

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

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

The missing reads

- 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)
- 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

Application

Application to ER and FOXA1

- FOXA1 data in human MCF7 human cells (Zhang et al., 2008).
- 3,909,507 treatment 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 ChipSeq

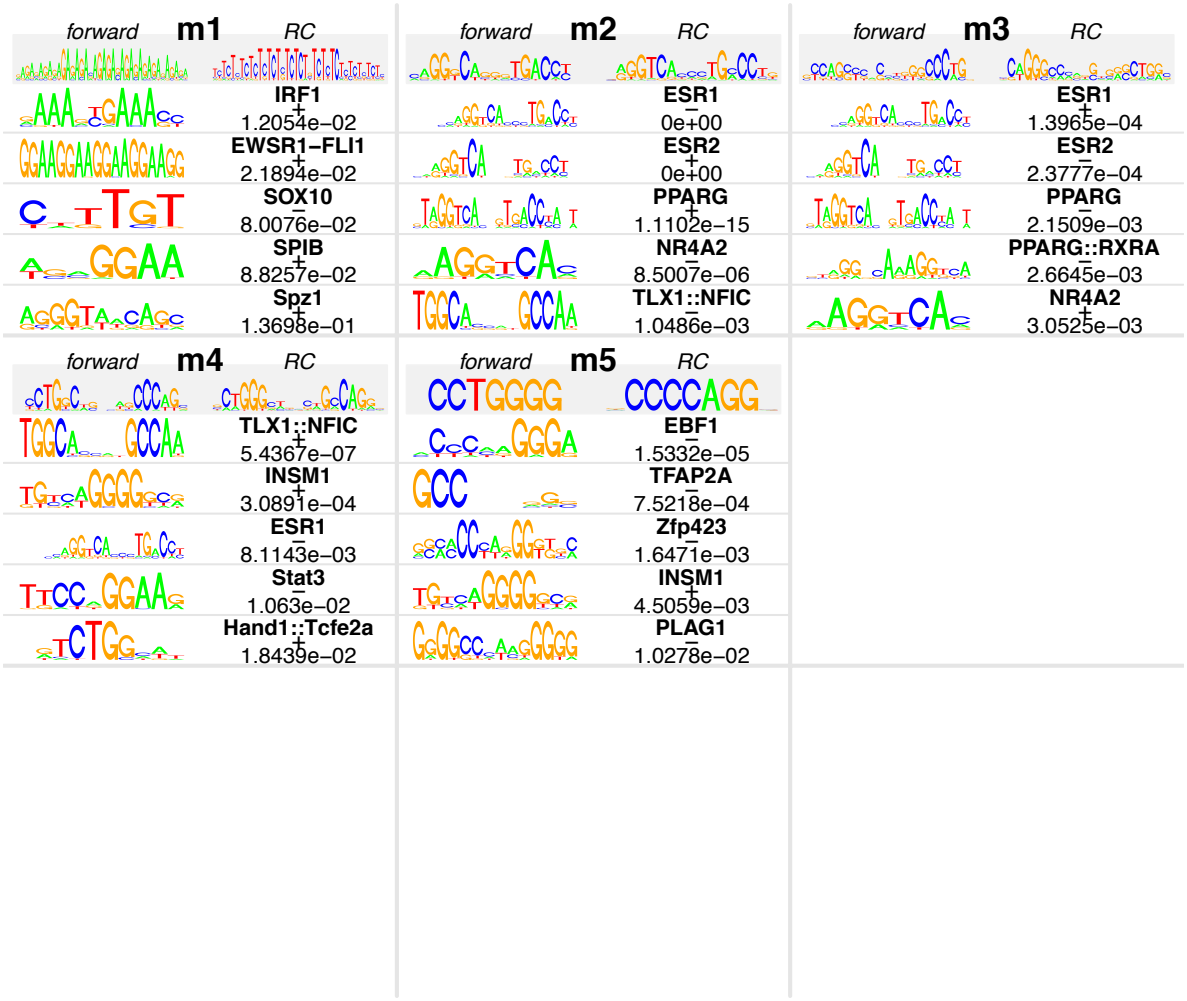
- Packages:
    - ShortRead: to read data
    - BSGenome: to access genomic information
    - PICS: to identify peak list
    - rGADEM: de novo motif discovery
    - MotIV: motifs identifications
    - Rtracklayer: Visualisation: interface to genome browser
    - GenomeGraphs: Visualisation

Validation

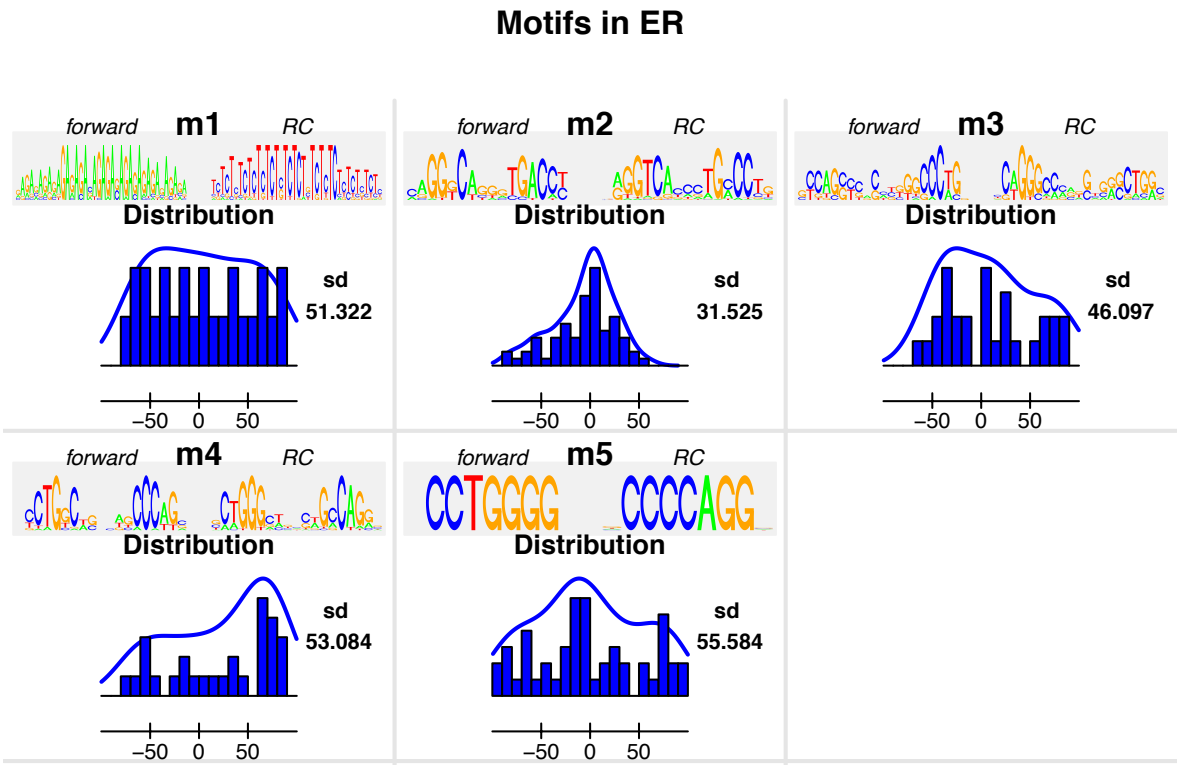
- *de novo* motif search
- Decided to use rGADEM because it is fast and can be used to process 10K+ sequences (binding site estimates +/- 100bps)
- Identified motifs were then fed into MotIV and analyzed with Jaspar

rGADEM + MoTiV results

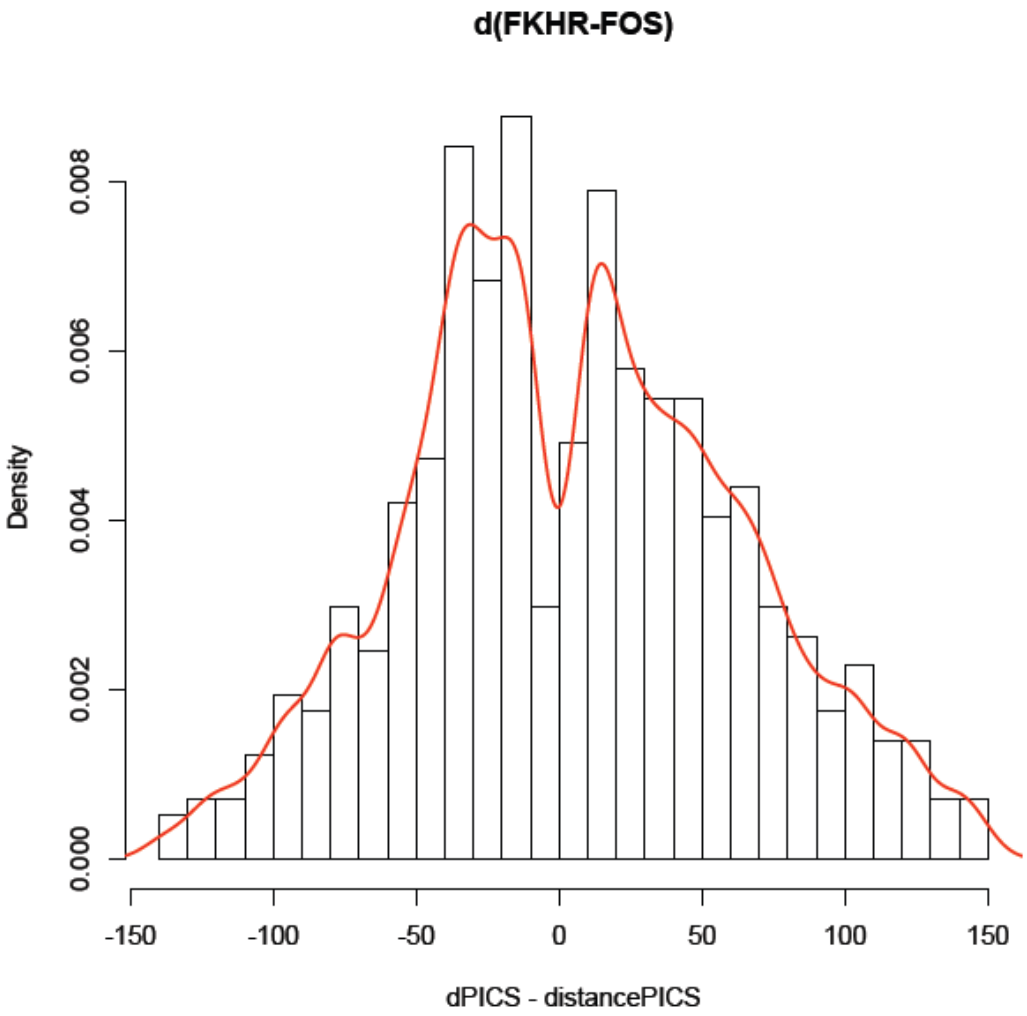
Motifs in ER



rGADEM + MoTiV results



rGADEM + MoTiV results



Installing GSL

- GNU Scientic Library (GSL) : http://www.gnu.org/software/gsl/

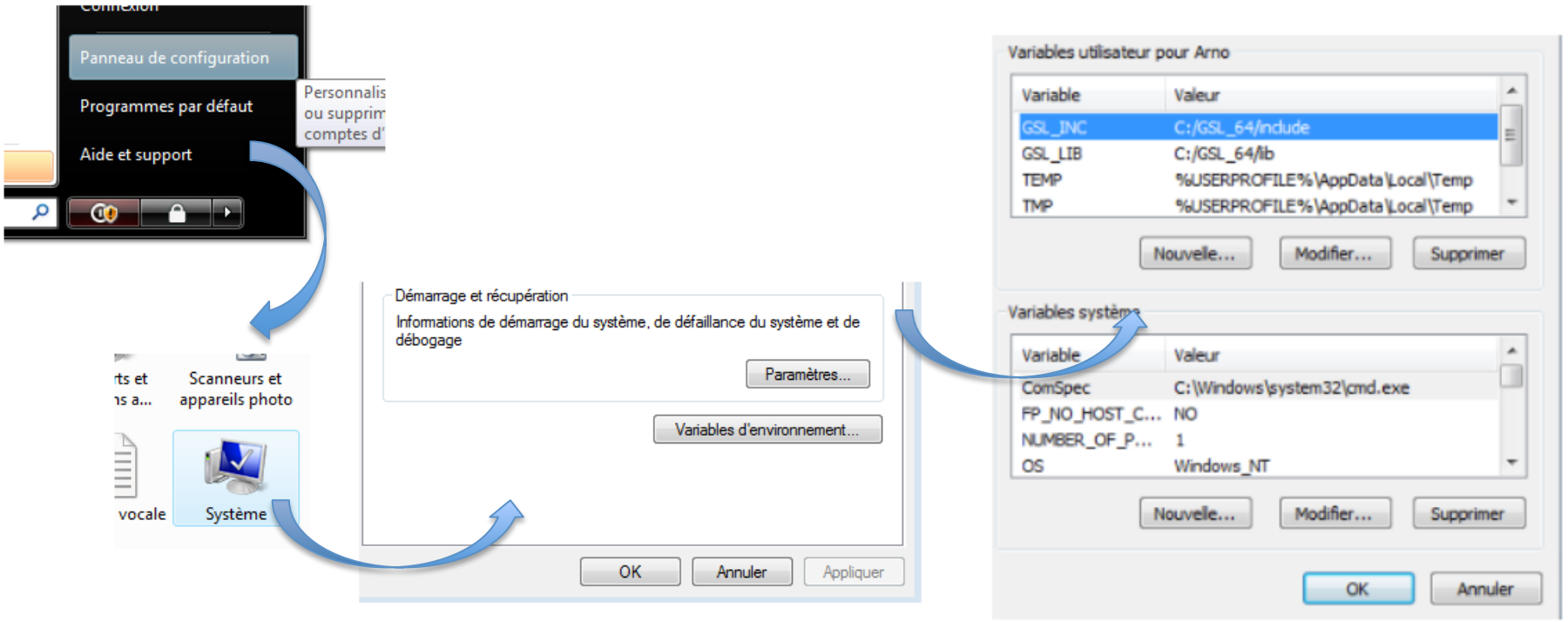
- Basic Linear Algebra Subprograms (BLAS) : http://www.netlib.org/blas/faq.html

Configuring GSL

Linux systems :

export LD_LIBRARY_PATH='/path/to/GSL/:/path/to/BLAS/':$LD_LIBRARY_PATH

Windows plateform :



STEP-BY-STEP ANALYSIS

Reading Data

1. Load the libraries
2. Treatment file : E2
3. Control file : ethl
4. Apply filters
5. Read the data using ShortReads

6. Do the same for FOXA1

PICS processing

1. Preprocess the read data by segmenting the genome into regions
2. Fit PICS to each region
3. Look into the *picsList* object

Detecting enriched regions

- Create filters
- Export as a RangedData Object
- FDR calculation
- Visualization of enriched regions

De Novo motif discovery

1. Load rGADEM and hg18
2. Create a RangedData object
3. Run rGADEM

Visualization

1. Load MotIV
2. Alignments visualization
3. Motifs distribution
4. Filtering the results

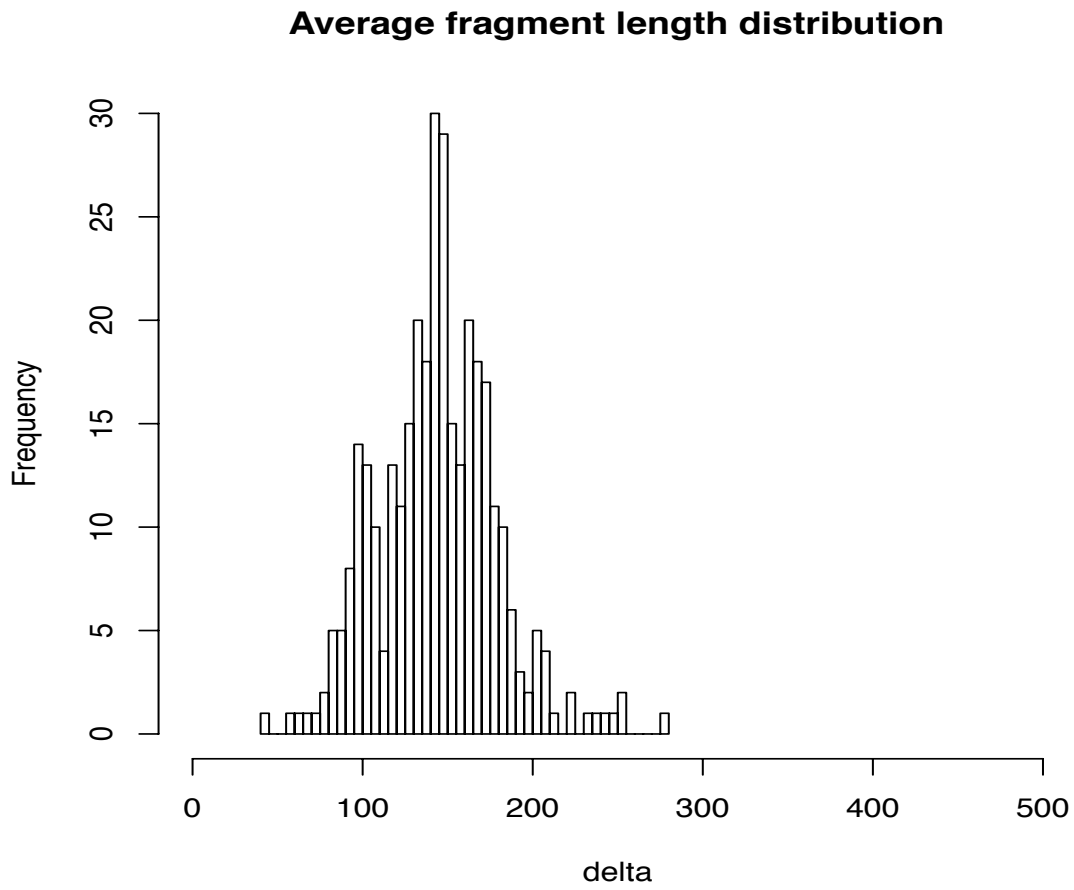
Annotation of enriched regions

1. Load ChIPpeakAnno
2. Compute the distance to the TSS
3. Plot the motif occurrences.

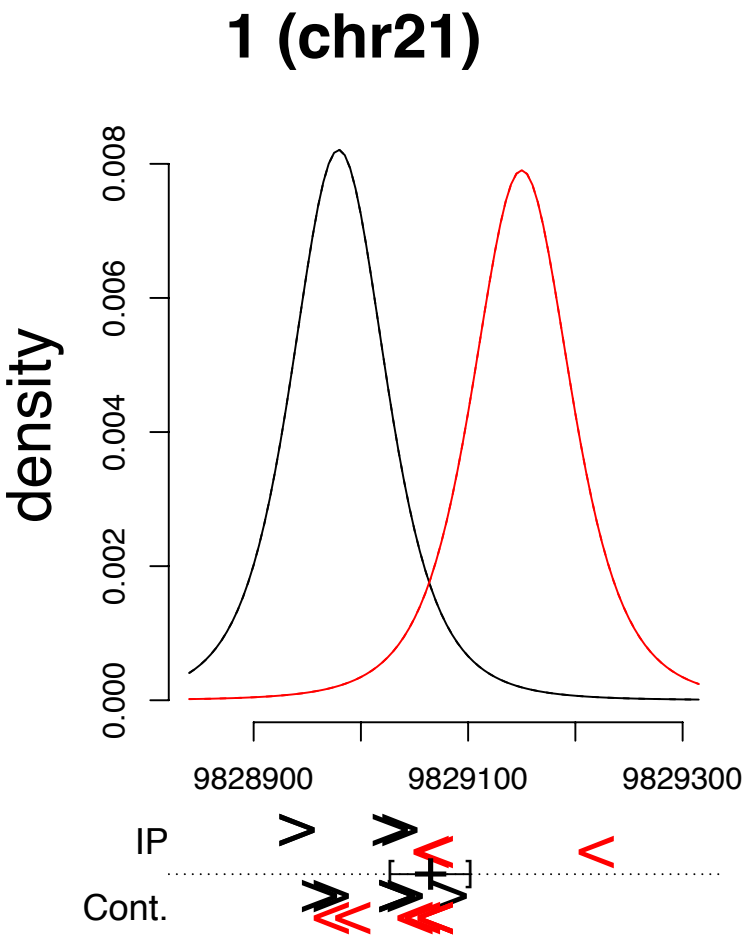
Conclusions

- ChIP is a powerful tool
  - Transcription factors
  - Epigenetics/Epigenomics
- Statistics/Bioinformatics challenges
  - Alignment, detecting binding events, etc
  - Still many challenges with ChIP-Seq

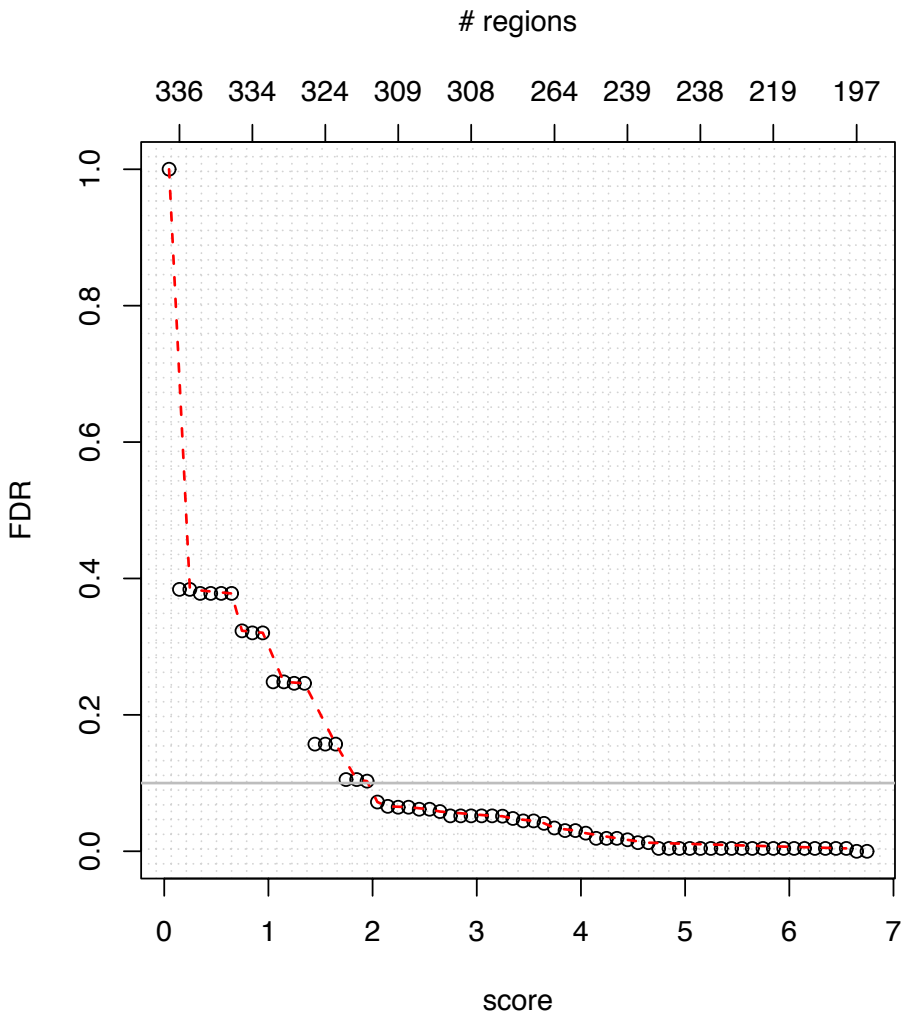
Average fragment length distribution



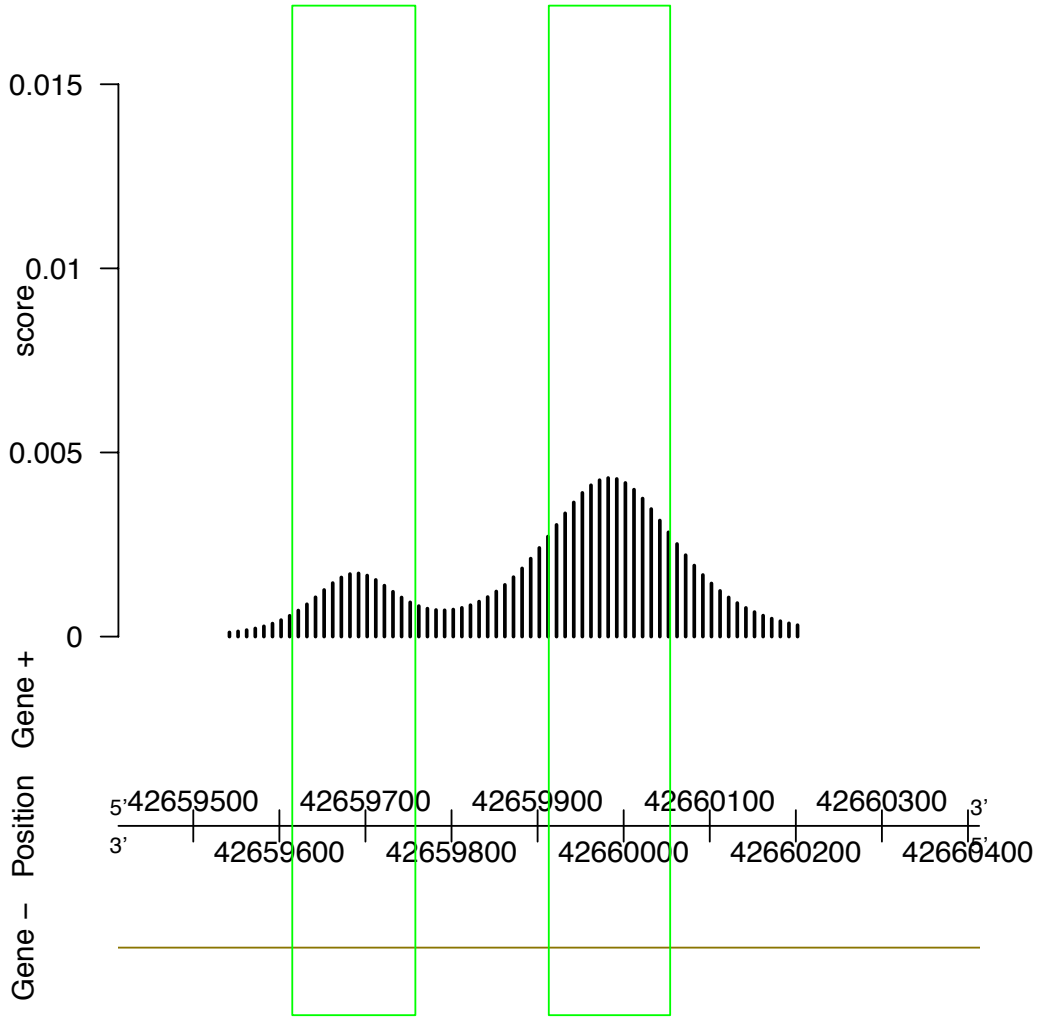
Visualizing candidate region



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: GenomeGraphs



Vizualisation: rtracklayer