

High-throughput chromatin immunoprecipitation assays

BioConductor 2011

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

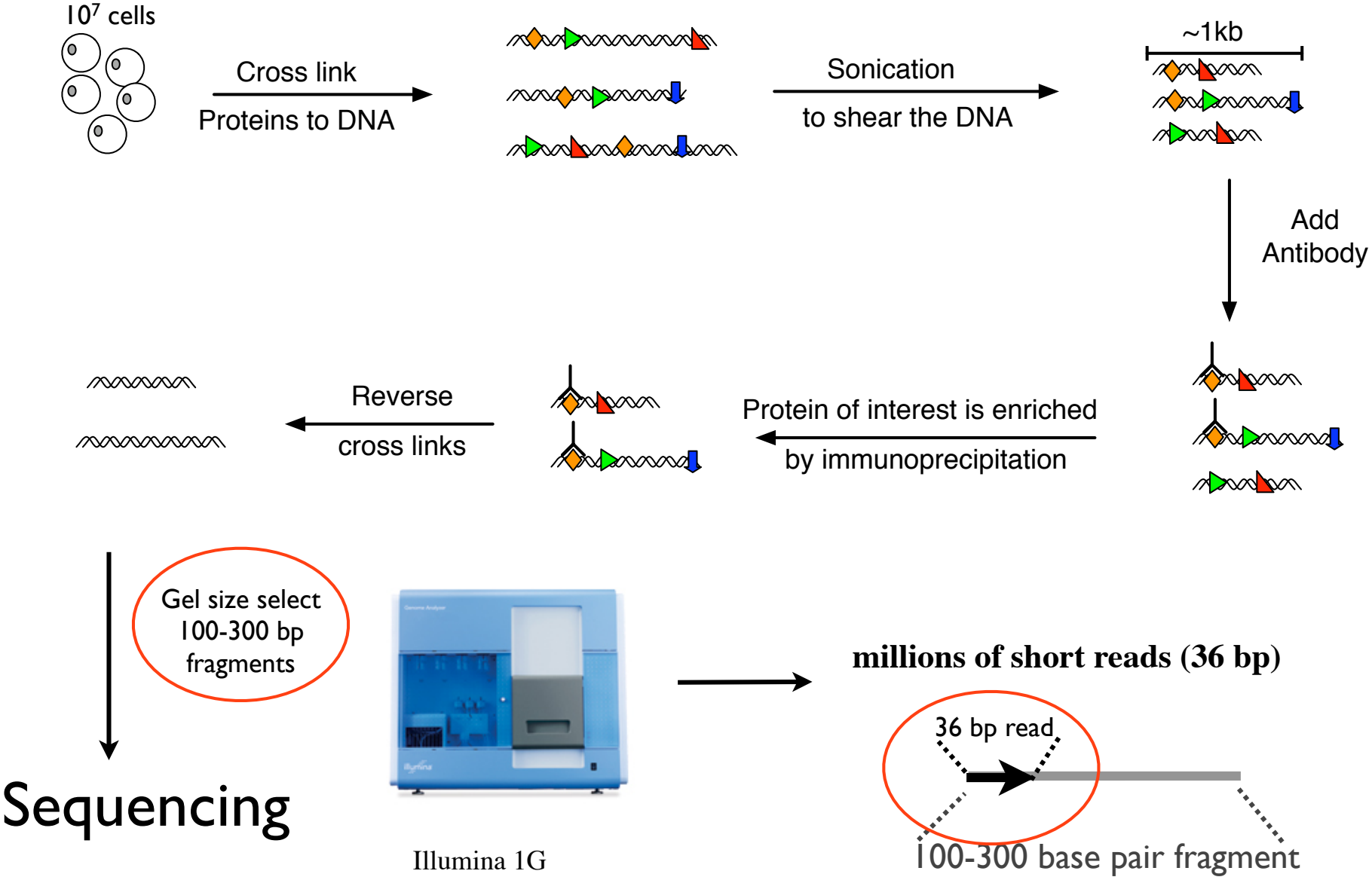
Outline

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

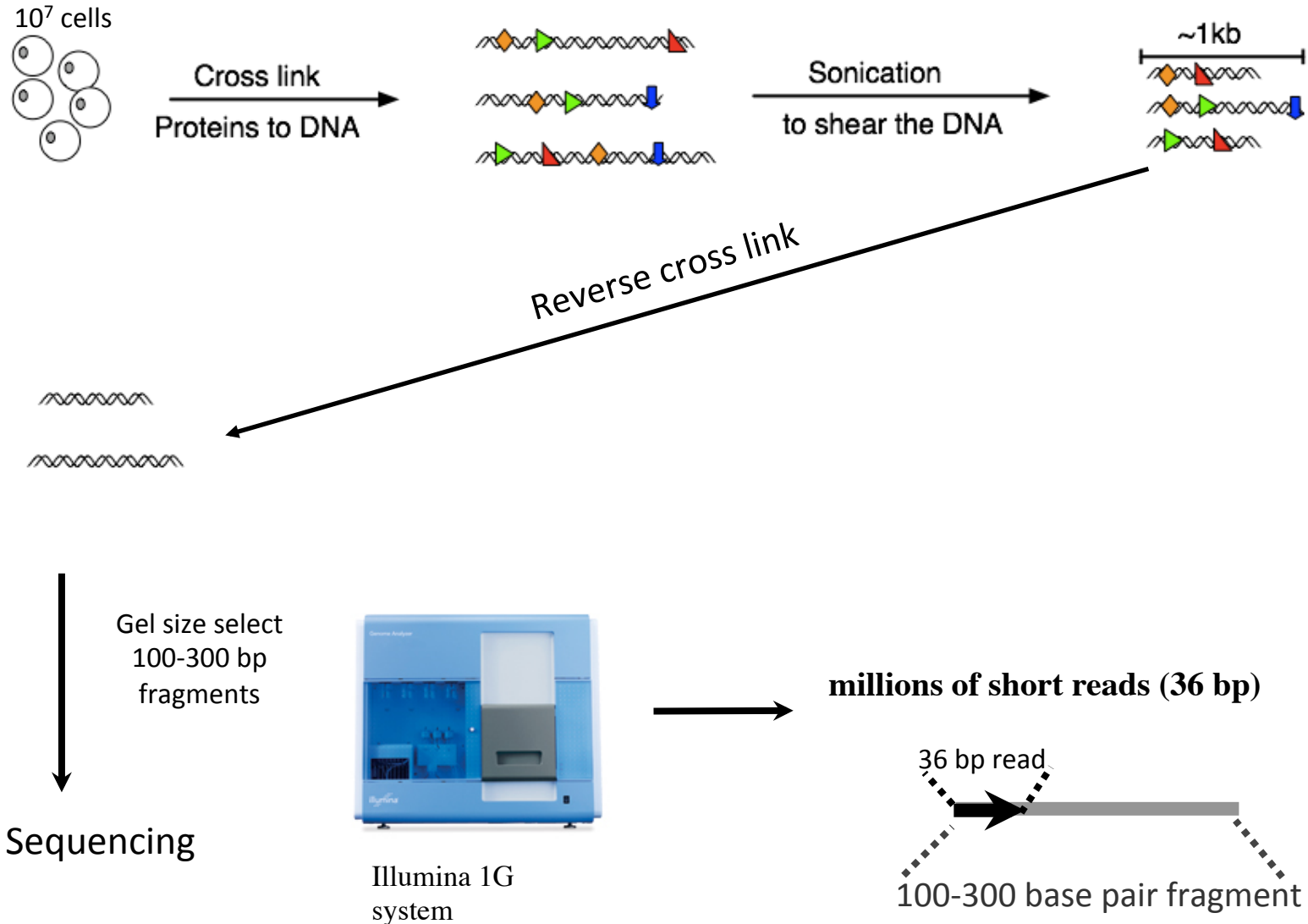
ChIP-Seq

- 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



ChIP-Seq (Control sample)



Aligners

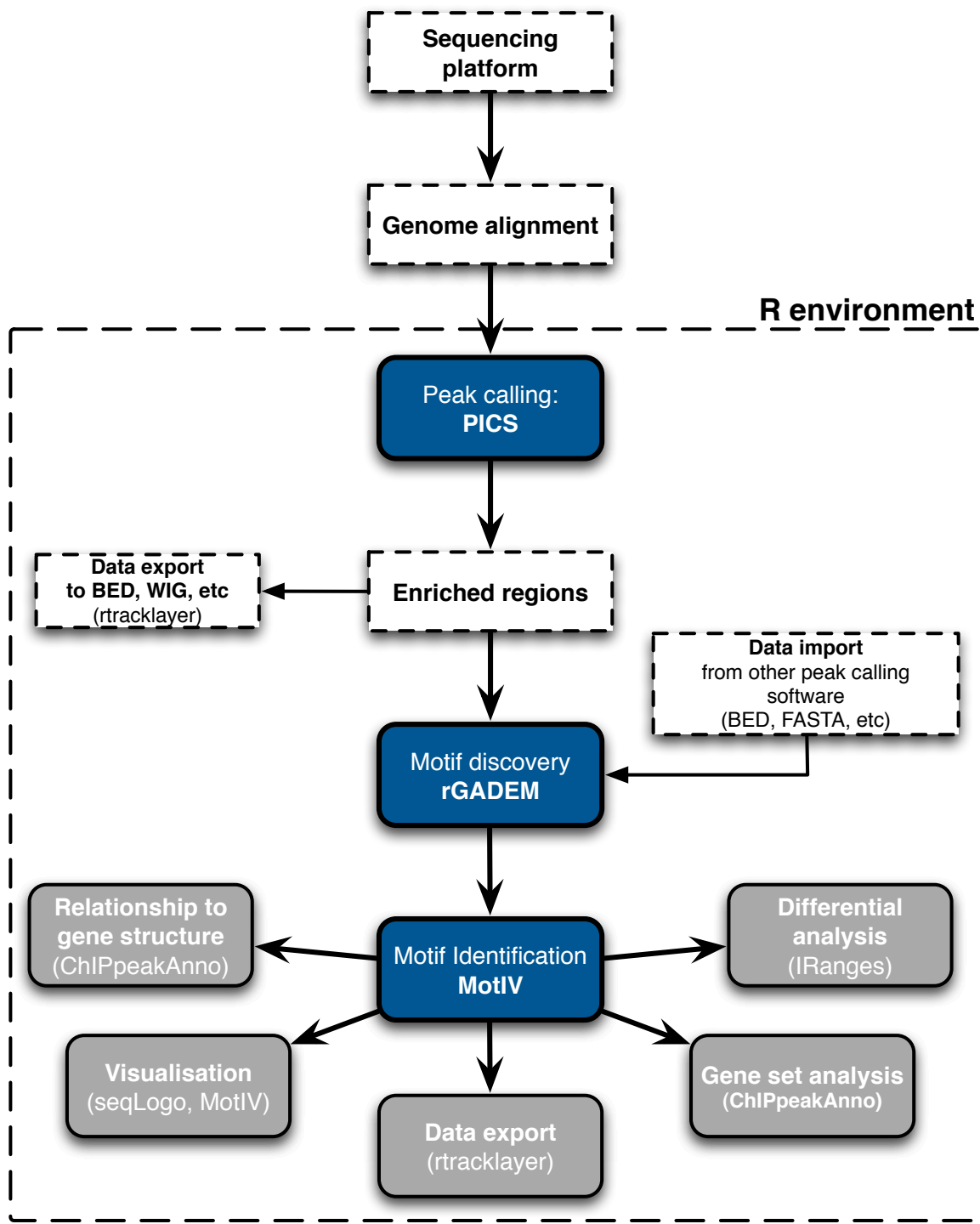
- The first step consists of aligning raw reads to the reference genome.
- There exists numerous “aligners” or “mappers”
- Here are a few popular ones: ELAND, MAQ, Bowtie, SOAP, Gnumap, etc
- Aligning raw reads can take from several hours to several days (depends on data, software and cpu)
- Most aligners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



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

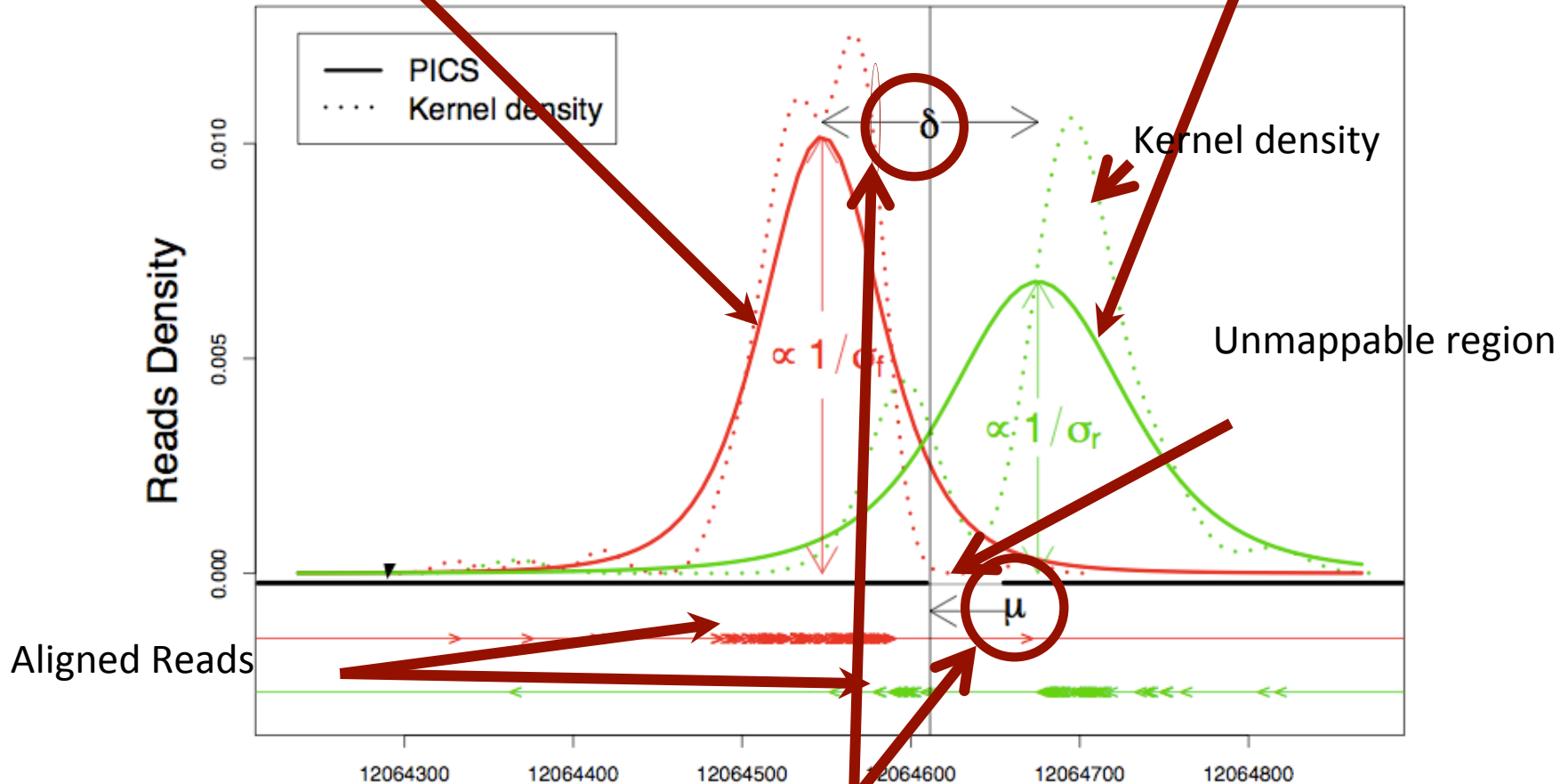
Preprocessing

- 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

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

a) One binding event



μ
 δ

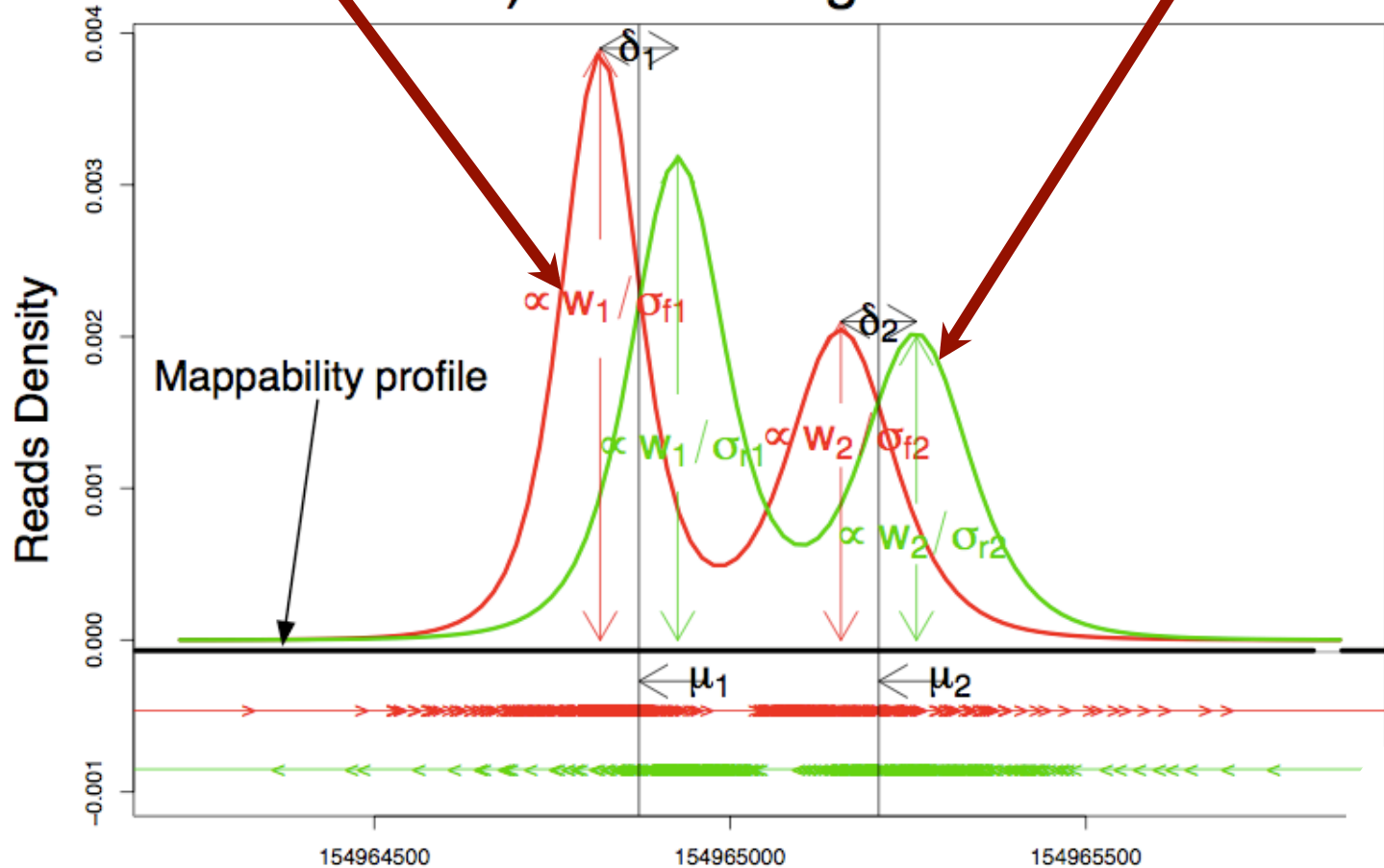
: TF binding site position

: average fragment length

Modeling bi-directional reads

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

b) Two binding events



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

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

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

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

- 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

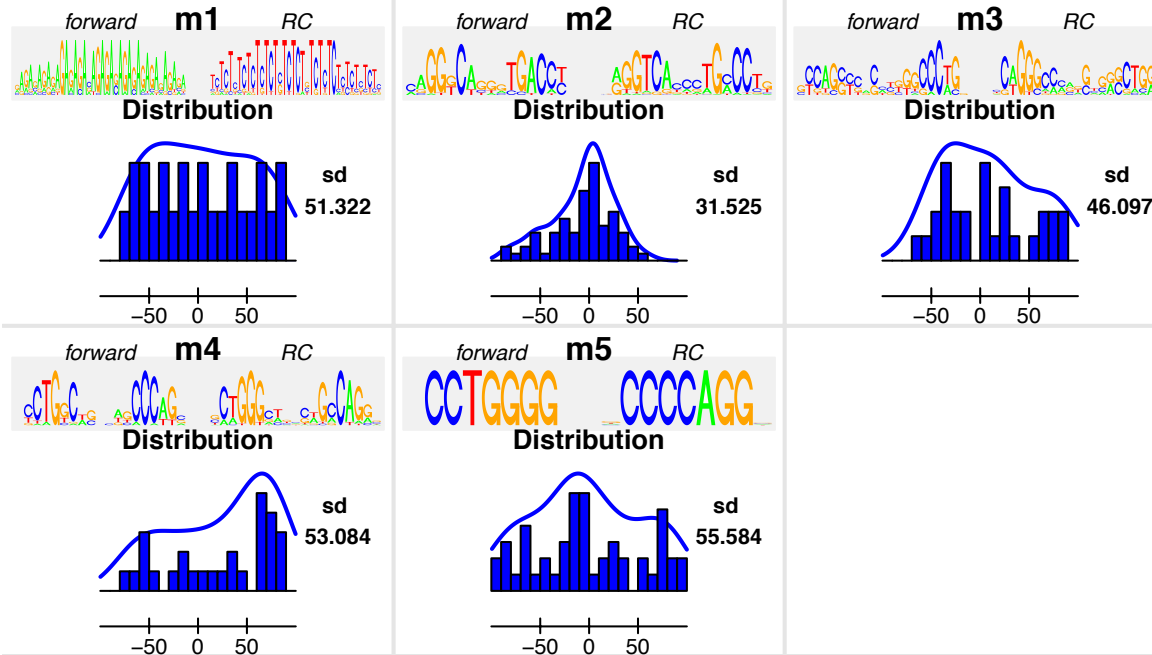
rGADEM + MoTiV results

Motifs in ER

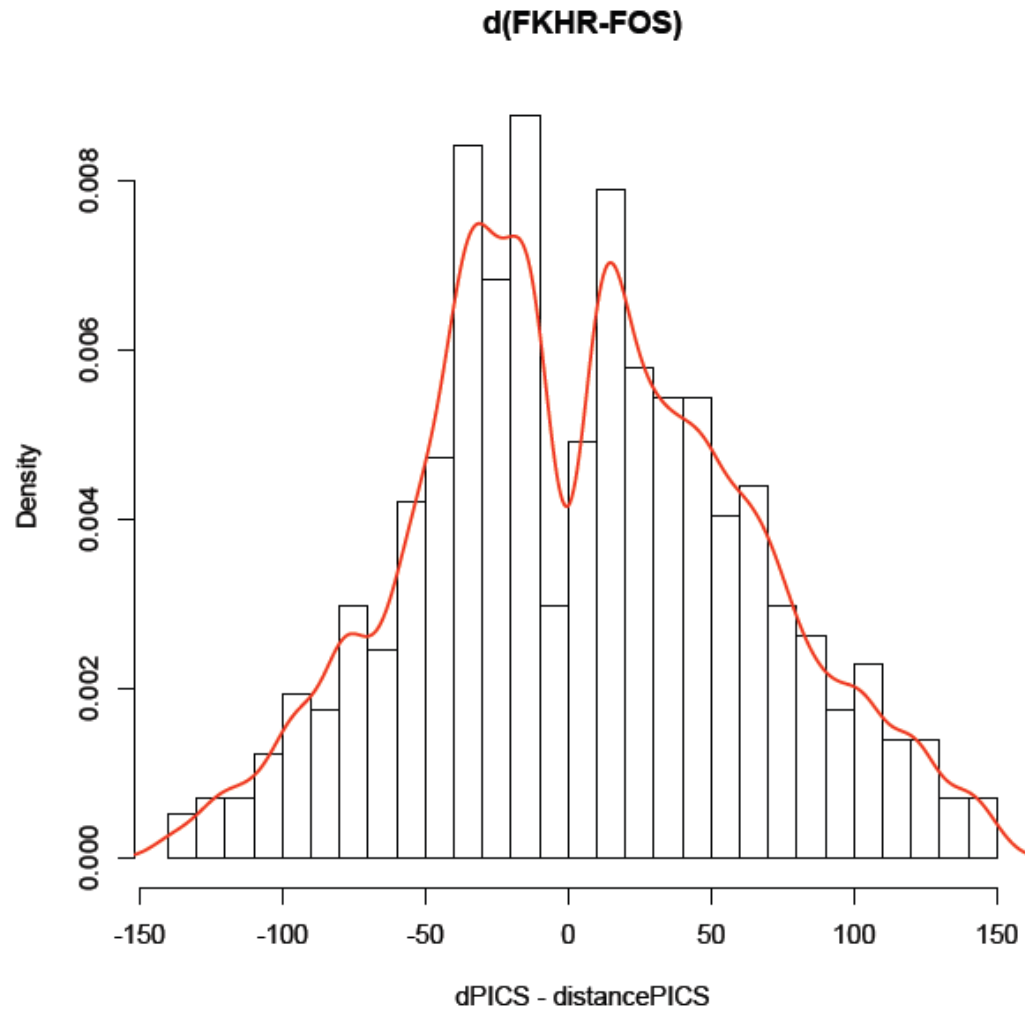
forward	m1	RC	forward	m2	RC	forward	m3	RC
AAA TGAAA	IRF1	1.2054e-02	AGGCA TGACCT	ESR1	0e+00	AGGCA TGACCT	ESR1	1.3965e-04
GGAAGGAAGGAAGGAAGG	EWSR1-FLI1	2.1894e-02	AGGCA TGACCT	ESR2	0e+00	AGGCA TGACCT	ESR2	2.3777e-04
C ITGT	SOX10	8.0076e-02	TAGTCA TCACTA I	PPARG	1.1102e-15	TAGTCA TCACTA I	PPARG	2.1509e-03
A GGAA	SPIB	8.8257e-02	AGGCA	NR4A2	8.5007e-06	AGGCA	PPARG::RXRA	2.6645e-03
ACGGTA CAGC	Spz1	1.3698e-01	TGCCA GCCAA	TLX1::NFIC	1.0486e-03	AGGCA	NR4A2	3.0525e-03
forward	m4	RC	forward	m5	RC			
TGCCA GCCAA	TLX1::NFIC	5.4367e-07	CCTGGGG CCCCAGG	EBF1	1.5332e-05			
TGAGGGG	INSM1	3.0891e-04	C C GGG	TFAP2A	7.5218e-04			
AGGCA TGACCT	ESR1	8.1143e-03	GCC	Zfp423	1.6471e-03			
TICC GGAA	Stat3	1.063e-02	AGGCA TGACCT	INSM1	4.5059e-03			
ICTGG	Hand1::Tcf2a	1.8439e-02	TGAGGGG	PLAG1	1.0278e-02			

rGADEM + MoTiV results

Motifs in ER



rGADEM + MoTiV results



Installing GSL

- GNU Scientific 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 platform :

The image shows a sequence of steps to configure environment variables in Windows. It starts with the Windows Start menu, where the 'System' icon is highlighted. An arrow points to the 'System' control panel window, which is open to the 'Advanced system settings' tab. Another arrow points to the 'Environment variables' button. A final arrow points to the 'User variables for Arno' dialog box, where the 'GSL_INC' variable is highlighted.

Variables utilisateur pour Arno

Variable	Valeur
GSL_INC	C:/GSL_64/include
GSL_LIB	C:/GSL_64/lib
TEMP	%USERPROFILE%\AppData\Local\Temp
TMP	%USERPROFILE%\AppData\Local\Temp

Variables système

Variable	Valeur
ComSpec	C:\Windows\system32\cmd.exe
FP_NO_HOST_C...	NO
NUMBER_OF_P...	1
OS	Windows_NT

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

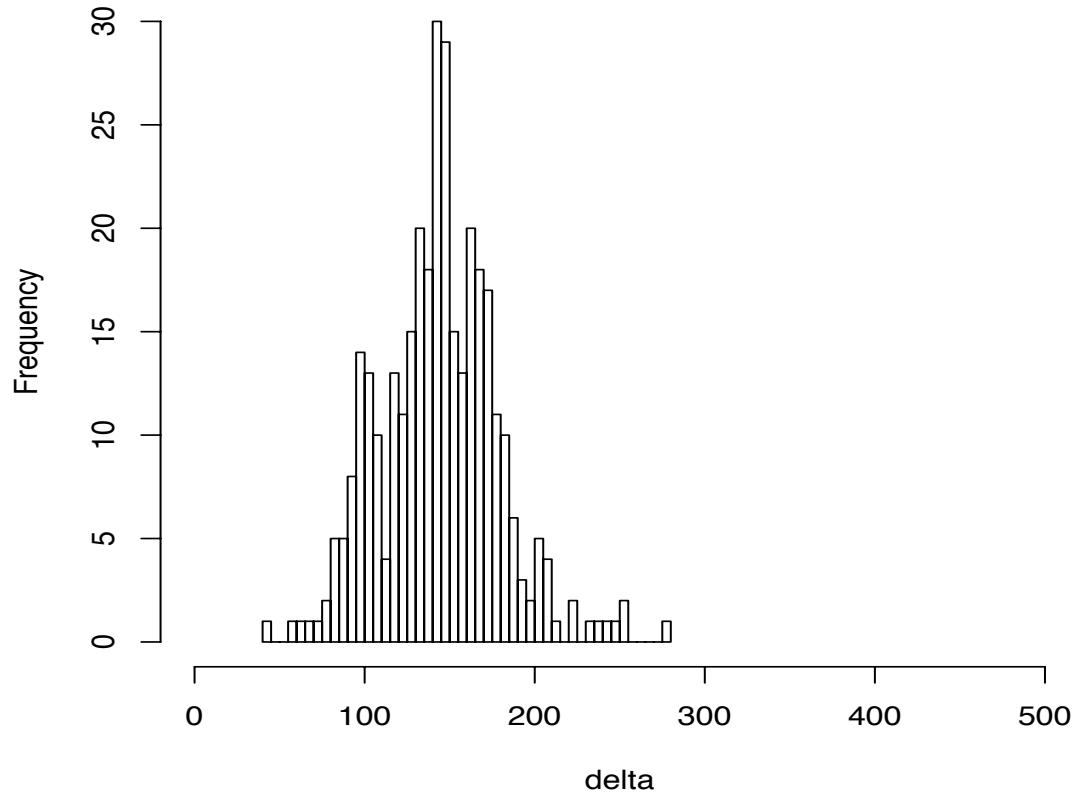
1. Load ChIPpeakAnno
2. Compute the distance to the TSS
3. Plot the motif occurrences.

Conclusions

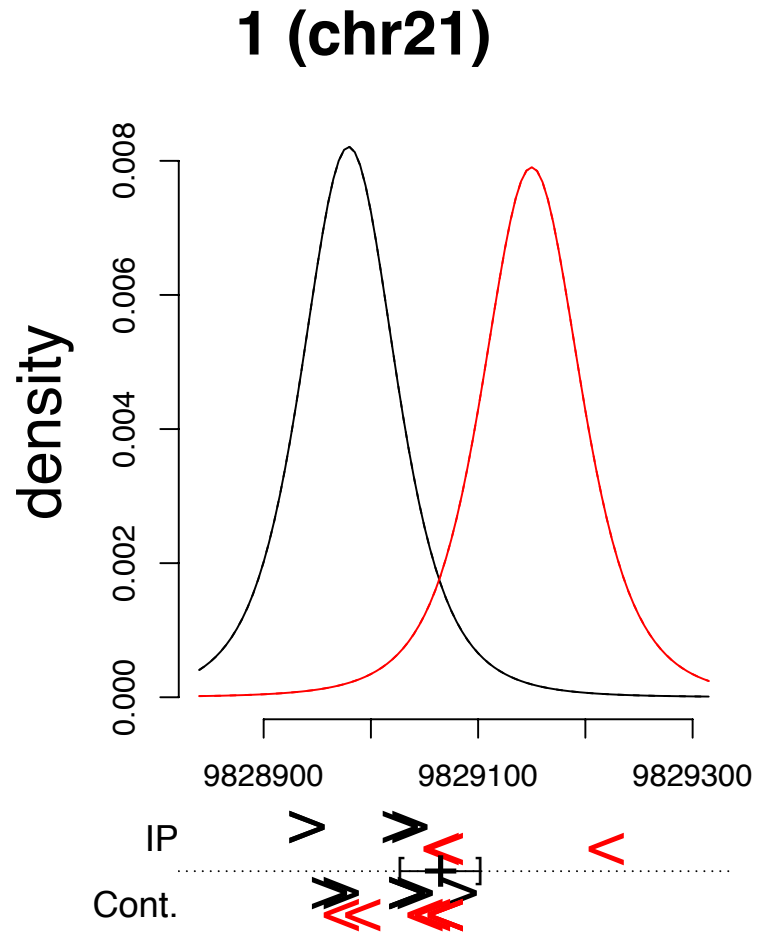
- ChIP is a powerful tool
 - Transcription factors
 - Epigenetics/Epigenomics
- Statistics/Bioinformatics challenges
 - Alignment, detecting binding events, etc
 - Still many challenges with ChIP-Seq

Average fragment length distribution

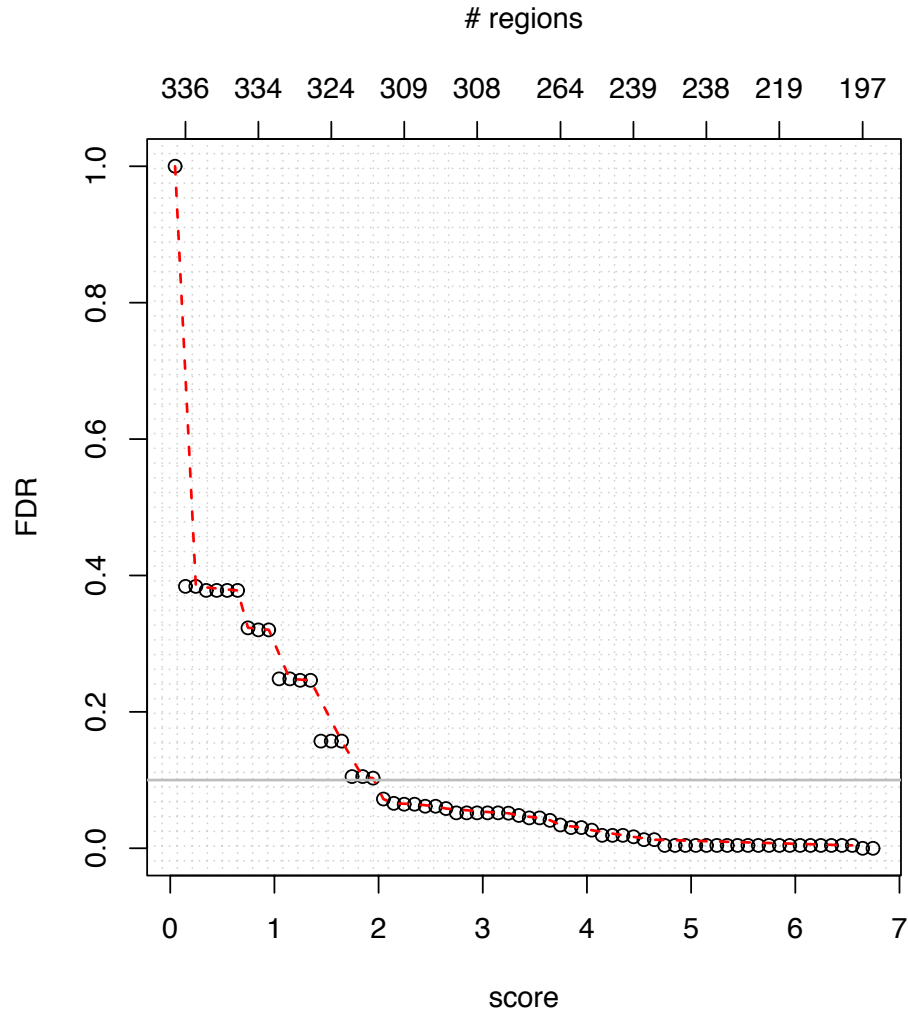
Average fragment length distribution



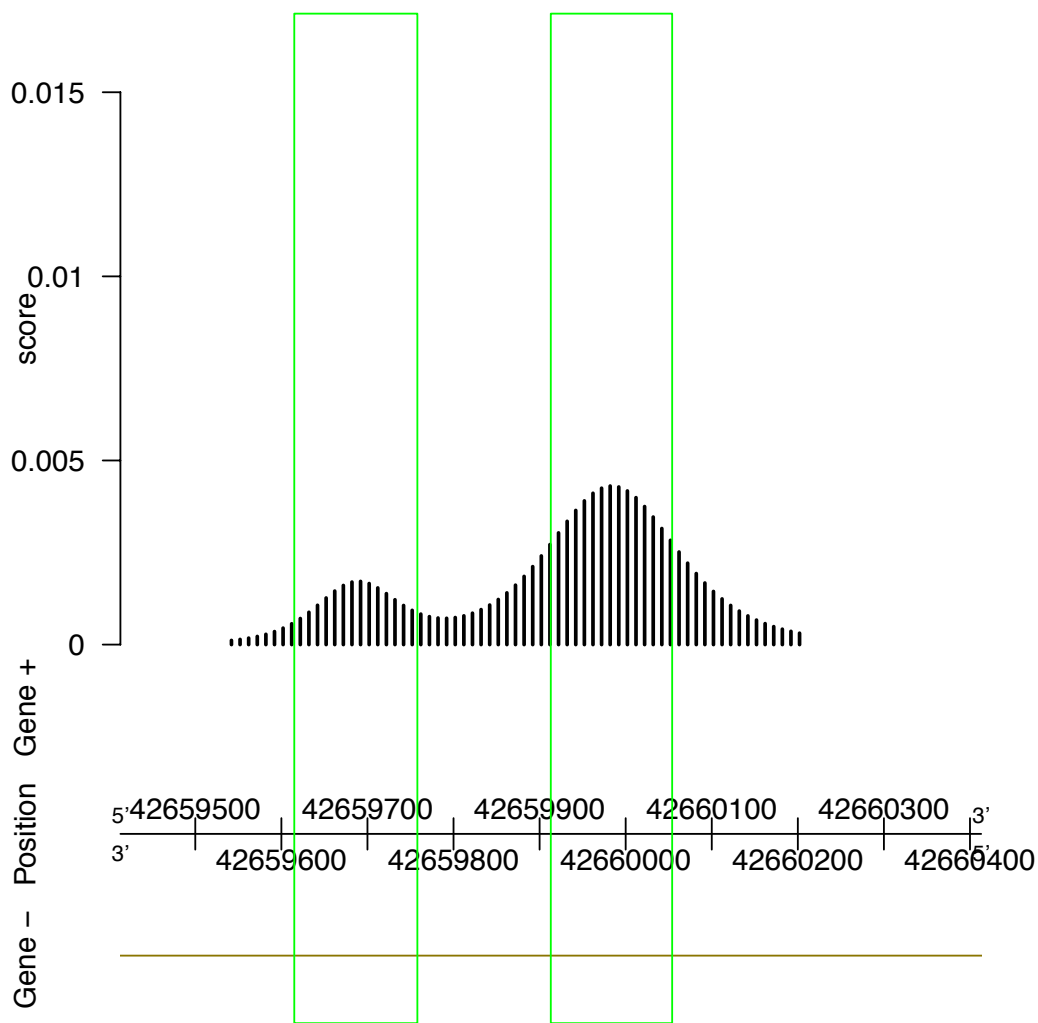
Visualizing candidate region



FDR



Vizualisation: GenomeGraphs



Vizualisation: rtracklayer

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

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr21:42,669,043-42,670,312 [gene](#) jump clear size 1,270 bp. configure

chr21 (q22.3) 21p13 21p12 21p11.2 11q2 21q21.1 21q21.2 21q21.3 21q22.11 q22.2 21q22.3

Scale 500 bases

chr21: |42669200|42669300|42669400|42669500|42669600|42669700|42669800|42669900|42670000|42670100|42670200|42670300|

targets

RefSeq Genes

UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics

RefSeq Genes

Human mRNAs

Human mRNAs from GenBank

Spliced ESTs

Human ESTs That Have Been Spliced

Layered H3K27Ac

100 _

0 _

DNase Clusters

Digital DNaseI Hypersensitivity Clusters from ENCODE

Tn Factor ChIP

Transcription Factor ChIP-seq from ENCODE

Placental Mammal Basepair Conservation by PhyloP

Multiz Alignments of 46 Vertebrates

Rhesus

Mouse

Dog

Elephant

Opossum

Chicken

X_tropicalis

Zebrafish

Common SNPs (132)

Single Nucleotide Polymorphisms (dbSNP 132) Found in >= 12 of Samples

Repeating Elements by RepeatMasker

move start 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. move end

< 2.0 >

track search default tracks default order hide all manage custom tracks configure reverse refresh

collapse all Use drop-down controls below and press refresh to alter tracks displayed. expand all

Tracks with lots of items will automatically be displayed in more compact modes.

Custom Tracks refresh

targets dense

Mapping and Sequencing Tracks refresh

Base Position Chromosome Band STS Markers FISH Clones Recomb Rate Map Contigs

dense hide hide hide hide hide

Assembly GRC Map Contigs Gap BAC End Pairs Fosmid End Pairs GC Percent

hide hide hide hide hide

GRC Patch Release Hg18 Diff NCBI Incident Short Match Restr Enzymes Wiki Track

hide hide hide hide hide

BU ORCHID Mapability