monaLisa

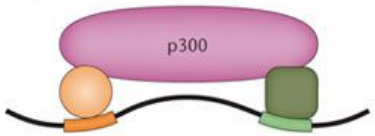
MOtif aNAlysis with Lisa

European Bioconductor Meeting 2019

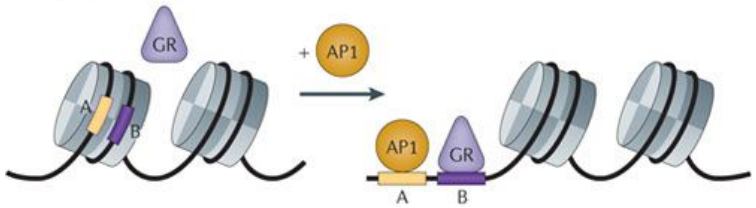
Dania Machlab
Lukas Burger
Michael Stadler

Friedrich Miescher Institute for Biomedical Research

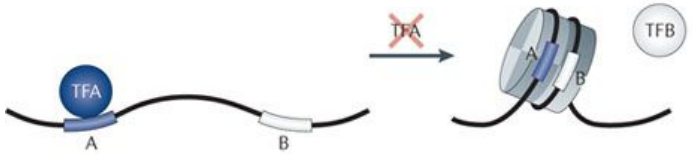
Background and Motivation



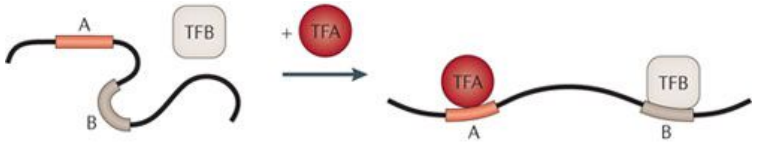
Co-binding



Chromatin remodeling



Blocking repositioning



Architectural role

HOMER

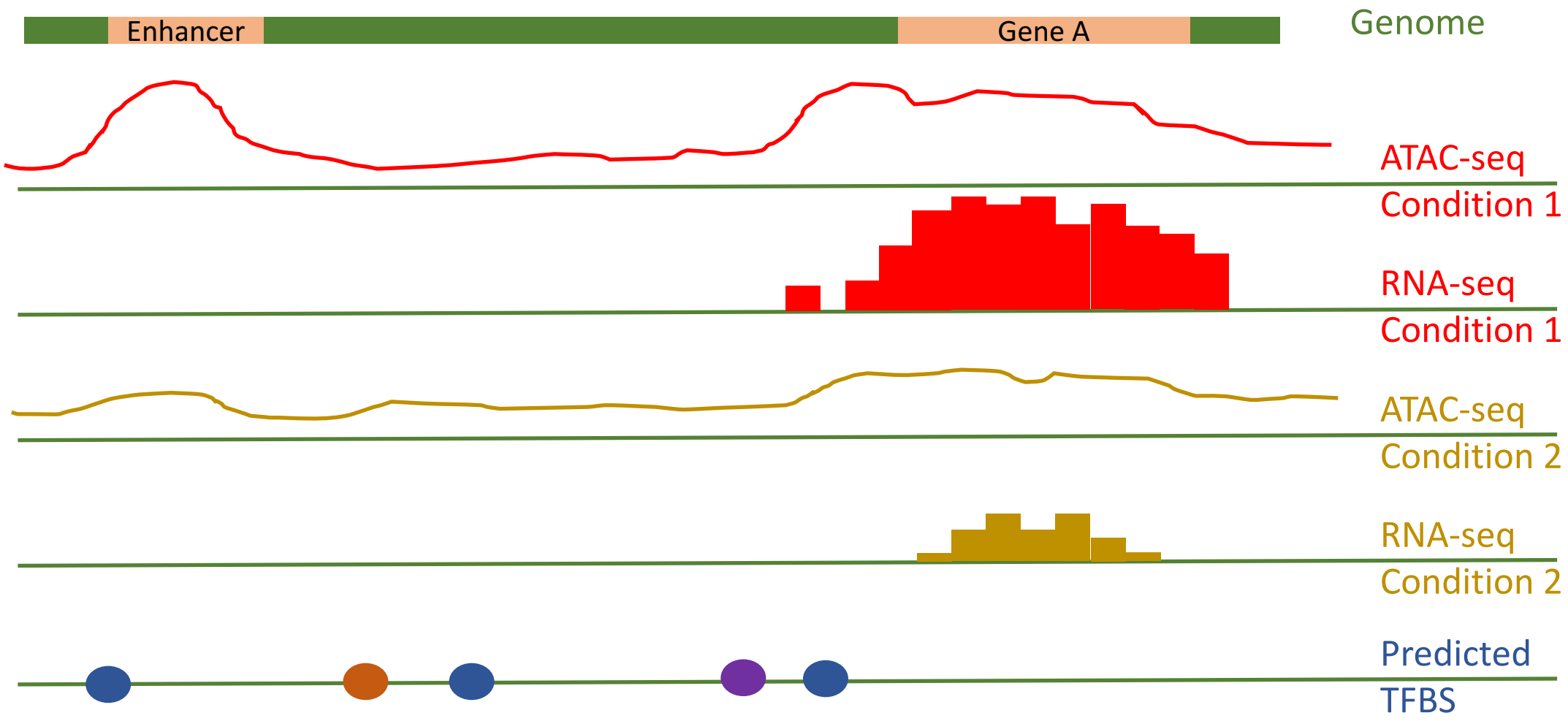
Software for motif discovery and next-gen sequencing analysis

Use monaLisa to:

- Identify Enriched motifs
- Select motifs explaining observed changes

Francois Spitz & Eileen E. M. Furlong (2012) *Nature Reviews Genetics*

Background and Motivation



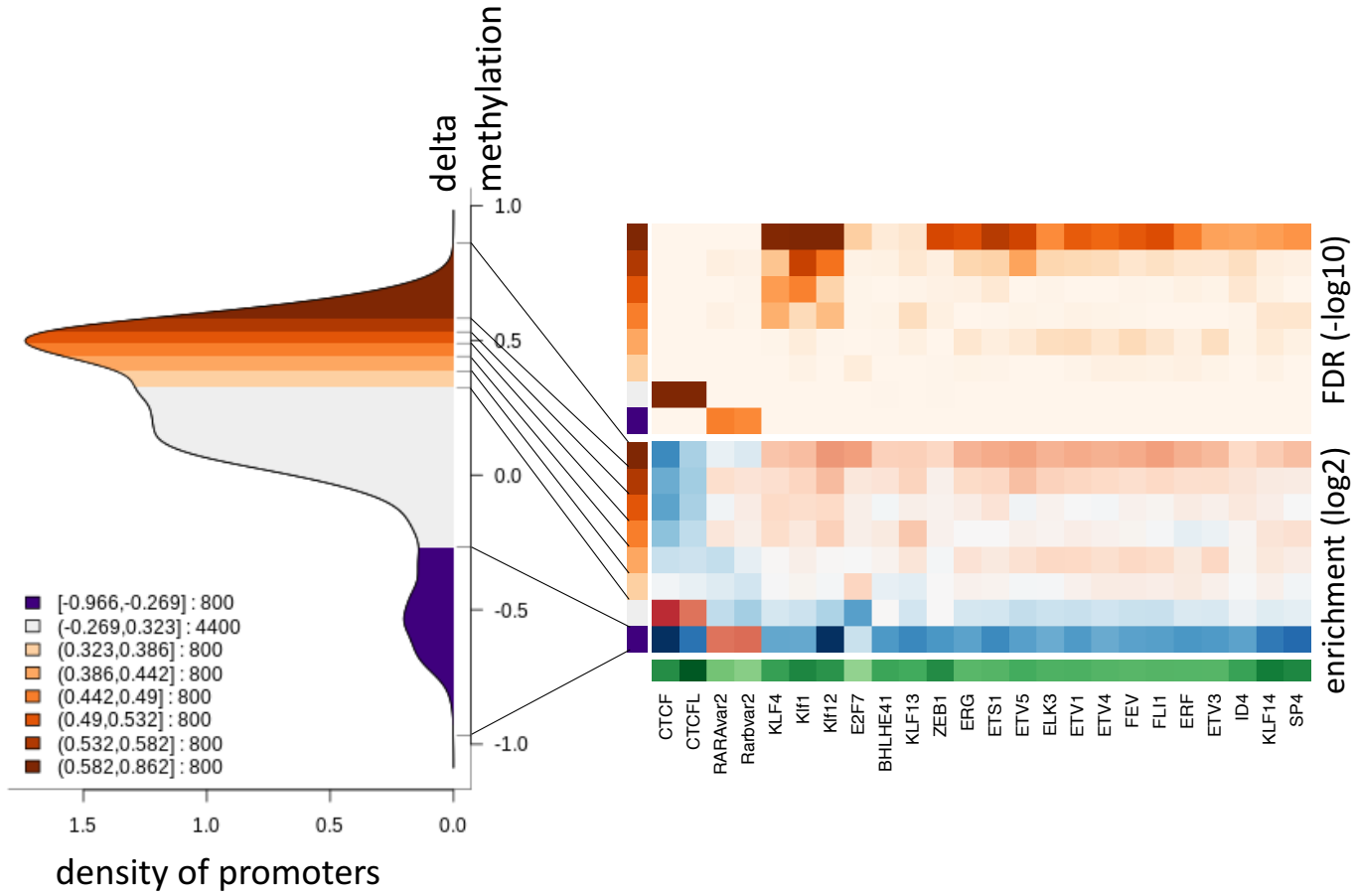
Identify Enriched Motifs

```
# bin regions by delta methylation
bins <- bin(x = lmrsel$deltaMeth, binmode = "equalN",
            nElement = 800, minAbsX = 0.3)

# dump motifs into file for use by Homer
motiffile <- tempfile(fileext = ".motif")
dumpJaspar(motiffile, pkg = "JASPAR2018")

# find Homer (findMotifsGenome.pl)
homerfile <- findHomer(dirs = "/work/gbioinfo/Appz
                       /Homer/Homer-4.10.4/bin/")

# run analysis
outdir <- tempfile(fileext = ".output")
se <- runHomer(gr = lmrsel, b = bins,
               genomedir = "/work/gbioinfo/DB/genomes/mm9",
               outdir = outdir, motifFile = motiffile,
               homerfile = homerfile,
               regionsize = "given", Ncpu = 20L)
```



Select Motifs using Stability Selection

$$\hat{\beta}^{\lambda,W} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} \frac{|\beta_k|}{W_k} \right)$$

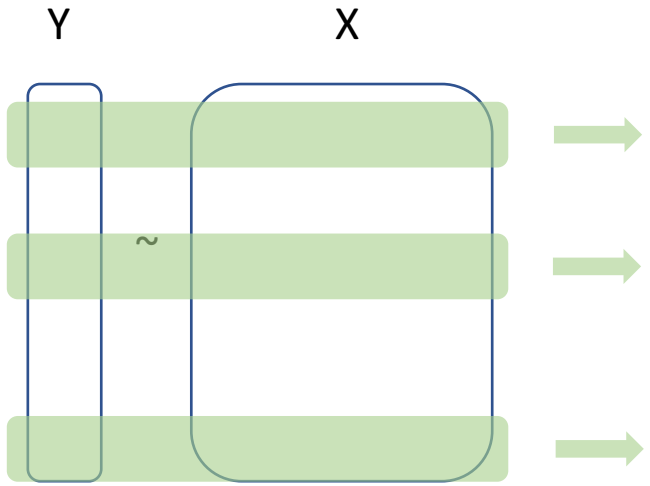
← Randomized lasso stability selection

observed logFC

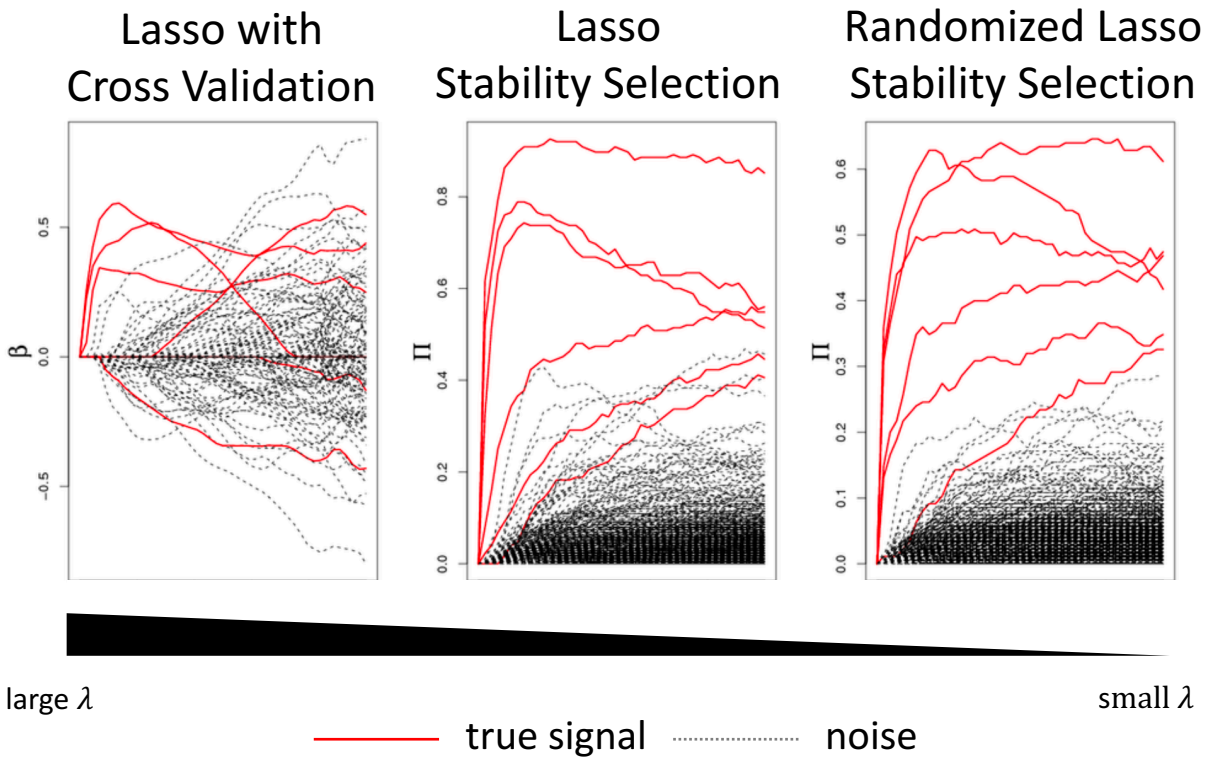
predicted TFBS

regularization parameter

weakness parameter

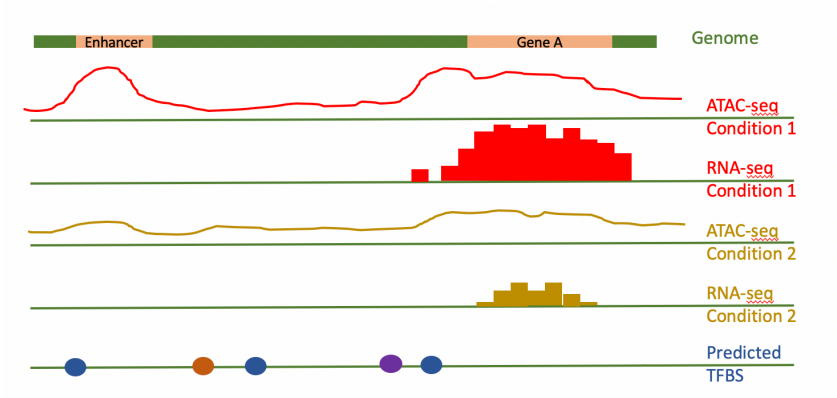


perform regularized regression



Meinshausen & Bühlmann (2010) *Journal of the Royal Statistical Society*

Select Motifs Explaining Observed Changes in Accessibility



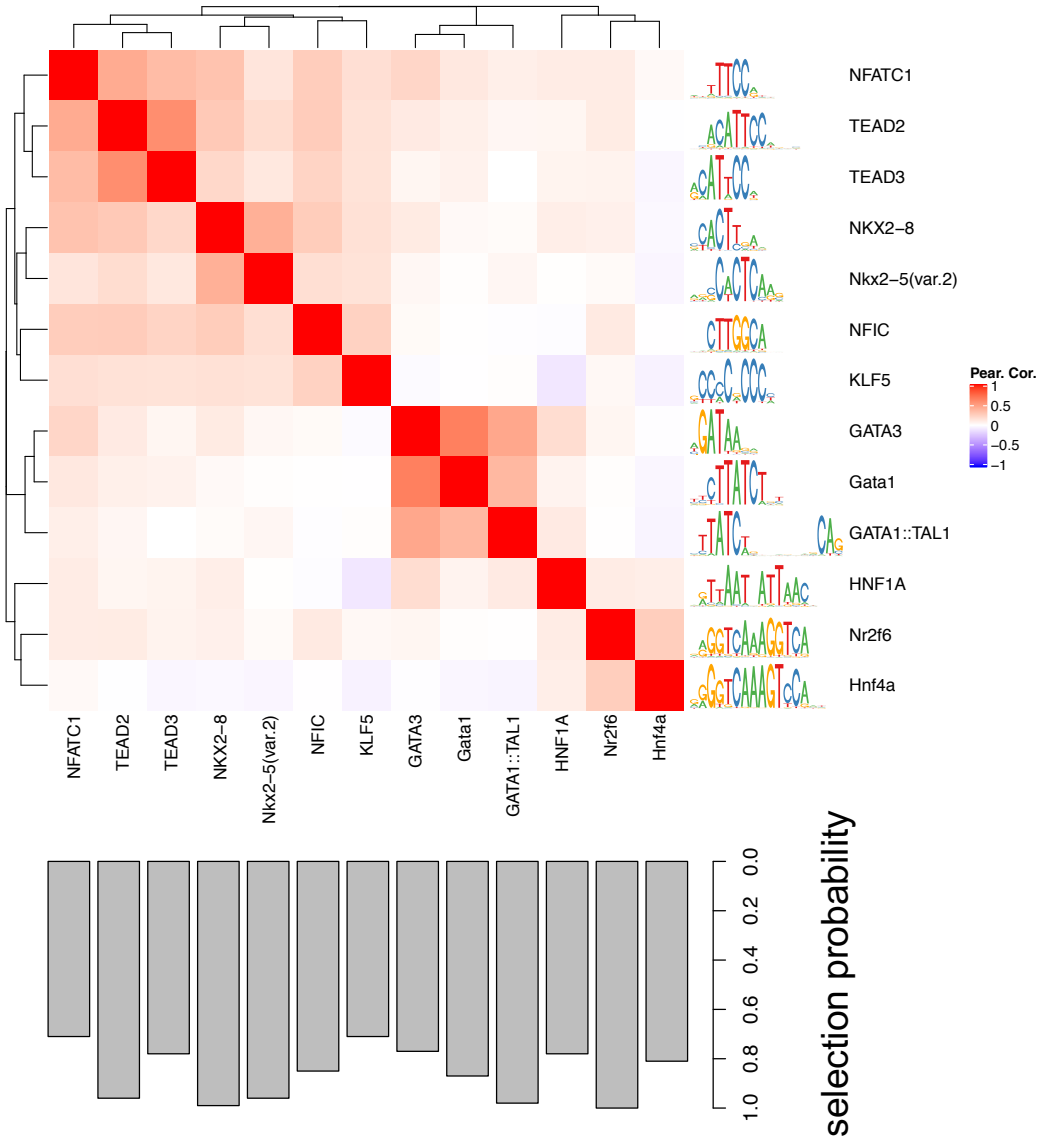

```r
# Genome
genome <- BSgenome.Mmusculus.UCSC.mm10

# Get PWMs
pfms <- getMatrixSet(JASPAR2018, list(matrixtype = "PFM", tax_group = "vertebrates"))
pwms <- toPWM(pfms)

# Get TFBS on given GRanges
homerfile <- findHomer(homerfile = "homer2", dirs = "/work/gbioinfo/Appz/Homer/Homer-4.10.4/bin/")
hits <- findMotifHits(query = pwms, subject = peaks, min.score = 6.0, method = "homer2",
                      homerfile = homerfile, genome = genome, Ncpu = 2)

# Get predictor matrix
predictor_matrix <- as.matrix(as.data.frame.matrix(table(seqnames(hits), as.character(hits$pwmname))))

# Perform randomized lasso stability selection
stabs <- randomized_stabsel(x = predictor_matrix, y = response,
                            weakness = 0.8, cutoff = 0.7, PFER = 2, mc.cores = 2)
```



glmnet::glmnet and stabs::stabsel used

Summary and Outlook

- We can identify TFs enriched in regions of interest that display certain log-fold changes
- We can select TFs that are likely to explain the observed log-fold changes using stability selection
- We can be use any fold-change defined on regions of interest (ATAC-seq, methylation, expression, ChIP-seq …) to select motifs explaining the observed logFC

- We want to look at motif enrichment without using existing databases (unbiased view)
- Enriched k-mers, grouping them, aligning them to predict the motif
- Submit to Bioconductor

- https://github.com/fmicompbio/monaLisa