Prediction of Regulatory Networks from Expression and Chromatin Data

Ivan G. Costa, RWTH Aachen University, Germany

Marcel Schulz, Saarland University & Max Planck Institute for Informatics, Germany

Matthias Heinig, Helmholtz Center Munich, Germany

RWTH AACHEN UNIVERSITY

max planck institut informatik

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Overview

Time	Topic	Who
2:30 - 2:45	Introduction / gene regulation / transcription / chromatin	IC
2:45 - 3:00	Introduction ChIP-seq peak calling	MH
3:00 - 3:50	Practical peak calling	MH & JH
4:15 - 4:30	Introduction Footprints	IC
4:30 - 4:45	Introduction Regulatory networks	MS
4:45 - 5:50	Practical Regulatory Networks	IG, MS & FS
5:50 - 6:00	Q & A session	all

Material - https://github.com/SchulzLab/EpigenomicsTutorial-ISMB2017

Team

Ivan Costa (IC)

Matthias Heinig (MH)

Johann Hawe (JH)

Marcel Schulz(MH)

Florian Schmidt (FS)

Introduction - Regulatory networks

Marcel H. Schulz

Saarland University & Max Planck Institute for Informatics, Germany

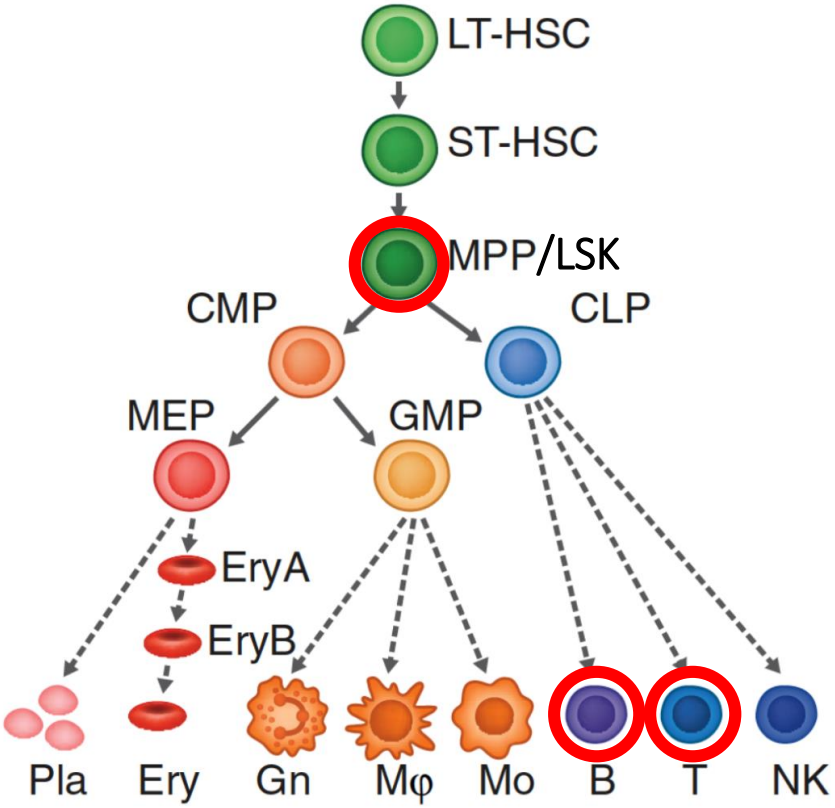
http://hgsb.mpi-inf.mpg.de/

CLUSTER OF EXCELLENCE

max planck institut informatik

UNIVERSITÄT DES SAARLANDES

Identification of key transcriptional regulators



Which transcription factors are related to the gene expression changes in the marked cells?

Identifying TF binding sites

Experimentally

- ChIP-seq
- Protein binding microarrays (PBMs)
- SELEX

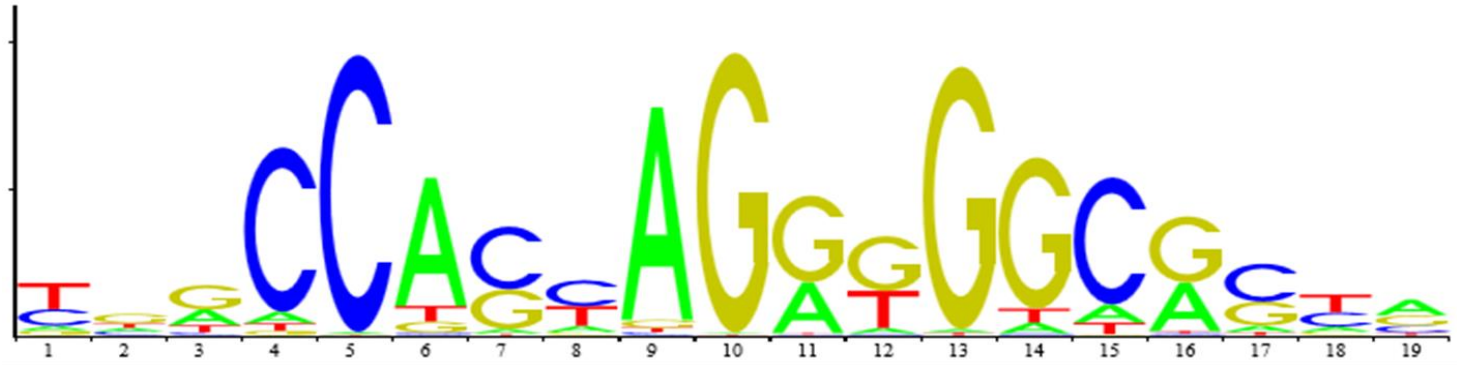
Computationally

- Site-centric
- Segmentation based

Position Weight Matrices (PWMs)

Experimentally identified binding site for CTCF

3’ 5’



http://jaspar.genereg.net/cgi-bin/jaspar_db.pl?ID=MA0139.1

Databases containing PWMs

TRANSFAC*	http://www.biobase-international.com/product/transcription-factor-binding-sites
JASPAR	http://jaspar.genereg.net/
UniPROBE	http://the_brain.bwh.harvard.edu/uniprobe/
Hocomoco	http://hocomoco.autosome.ru/

*commercial database

Site-centric TF annotation

Sliding window

For all positions in the genome:

Calculate a score for each PWM, if it is significant, report a putative TF binding site.

FIMO, Grant et al., Bioinformatics, 2011

Site-centric TF annotation

Sliding window

Reported TF binding site

For all positions in the genome:

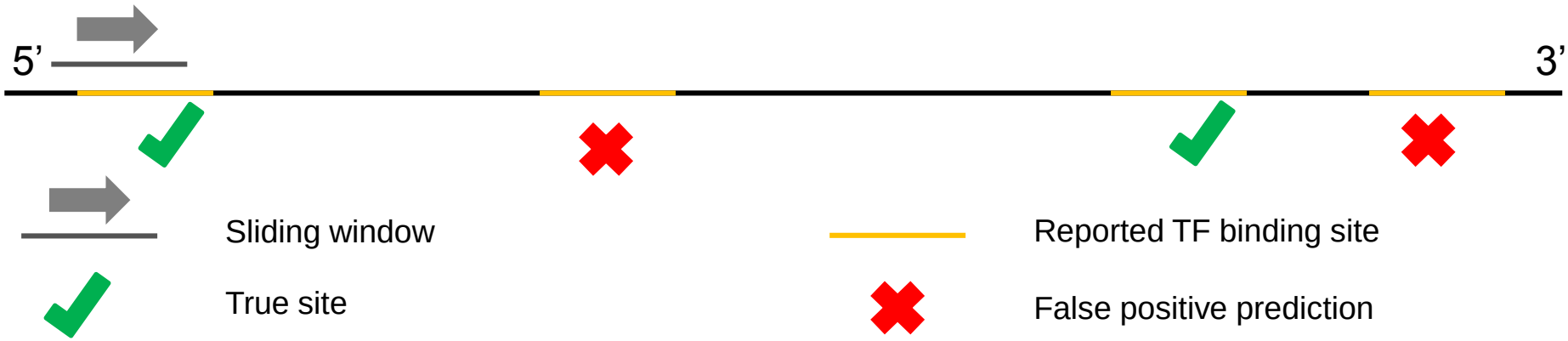
Calculate a score for each PWM, if it is significant, report a putative TF binding site.

Delivers a binary view of TF binding.

Therefore this type of annotation is also known as hit-based.

FIMO, Grant et al., Bioinformatics, 2011

Site-centric TF annotation



For all positions in the genome:

Calculate a score for each PWM, if it is significant, report a putative TF binding site.

Delivers a binary view of TF binding.

Therefore this type of annotation is also known as hit-based.

What is causing the errors?

FIMO, Grant et al., Bioinformatics, 2011

Site-centric TF annotation

(1) Hits by chance:

Genome is about 3,000,000,000bp long
TF binding motifs consist of about ~10bp

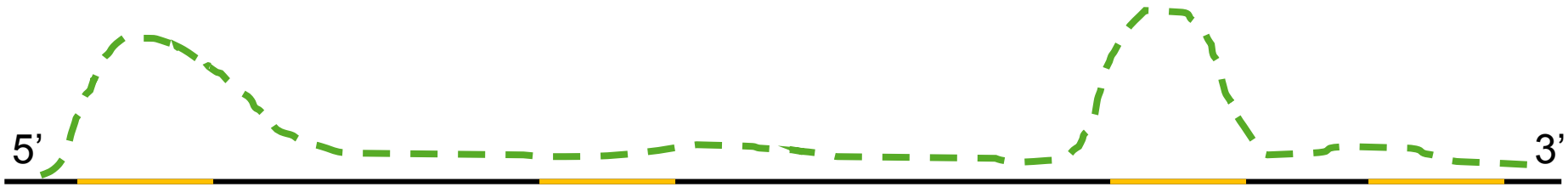
Many false positive predictions

(2) Chromatin state:

A gene should not be expressed

The binding site of activating TFs is blocked by nucleosomes

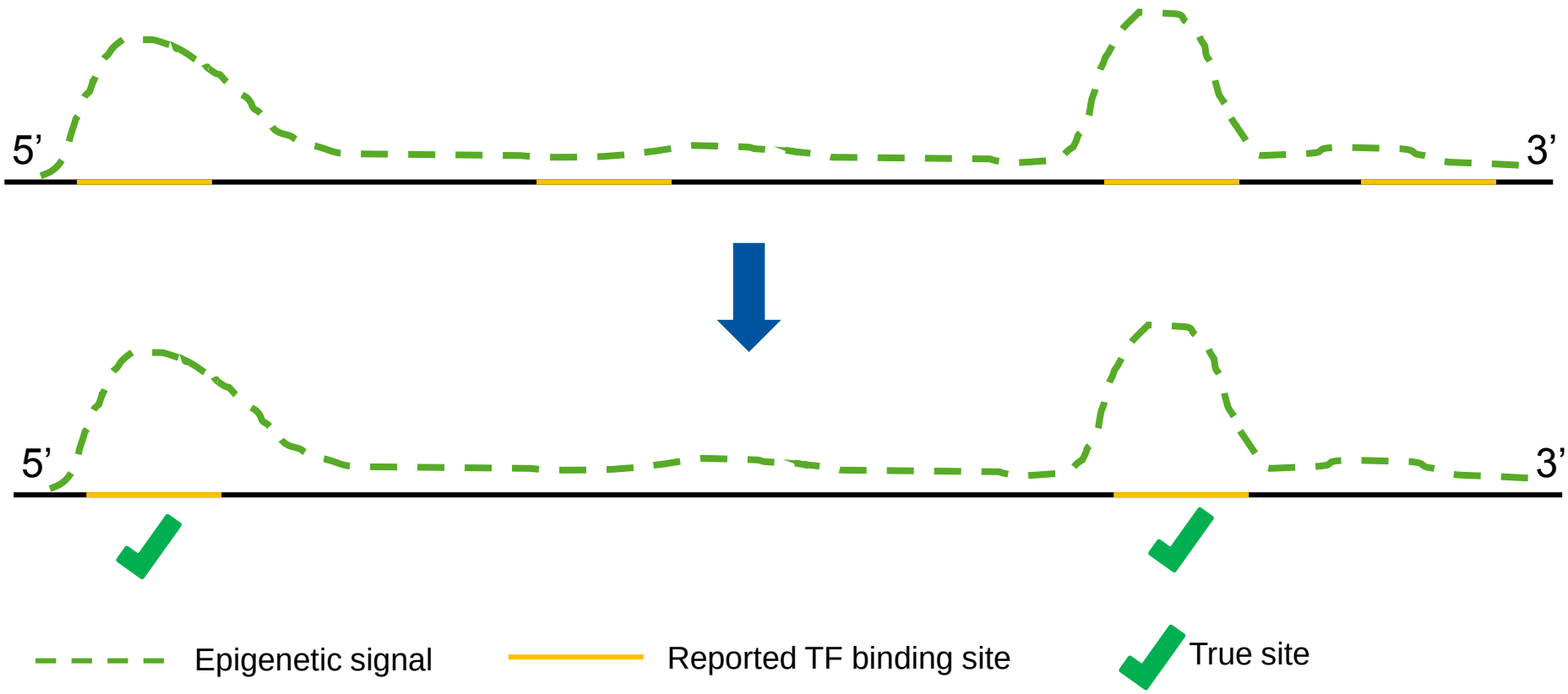
Improving predictions with epigenetics data



Epigenetic signal Reported TF binding site

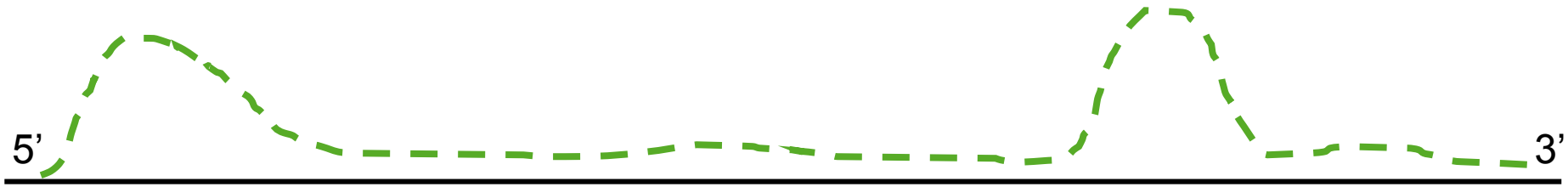
CENTIPEDE, Pique-Regi et al., Genome Research, 2011.
FIMO, Cuellar Partida et al., Bioinformatics, 2011.
MILLIPEDE, Luo and Hartemink, Pac Symp Biocomput, 2013.
PIQ, Sherwood et al., Nature Biotechnology, 2013.
BinDNase,, Kähärä J, Lähdesmäki H, Bioinformatics, 2015.

Improving predictions with epigenetics data



CENTIPEDE, Pique-Regi et al., Genome Research, 2011.
FIMO, Cuellar Partida et al., Bioinformatics, 2011.
MILLIPEDE, Luo and Hartemink, Pac Symp Biocomput, 2013.
PIQ, Sherwood et al., Nature Biotechnology, 2013.
BinDNase,, Kähärä J, Lähdesmäki H, Bioinformatics, 2015.

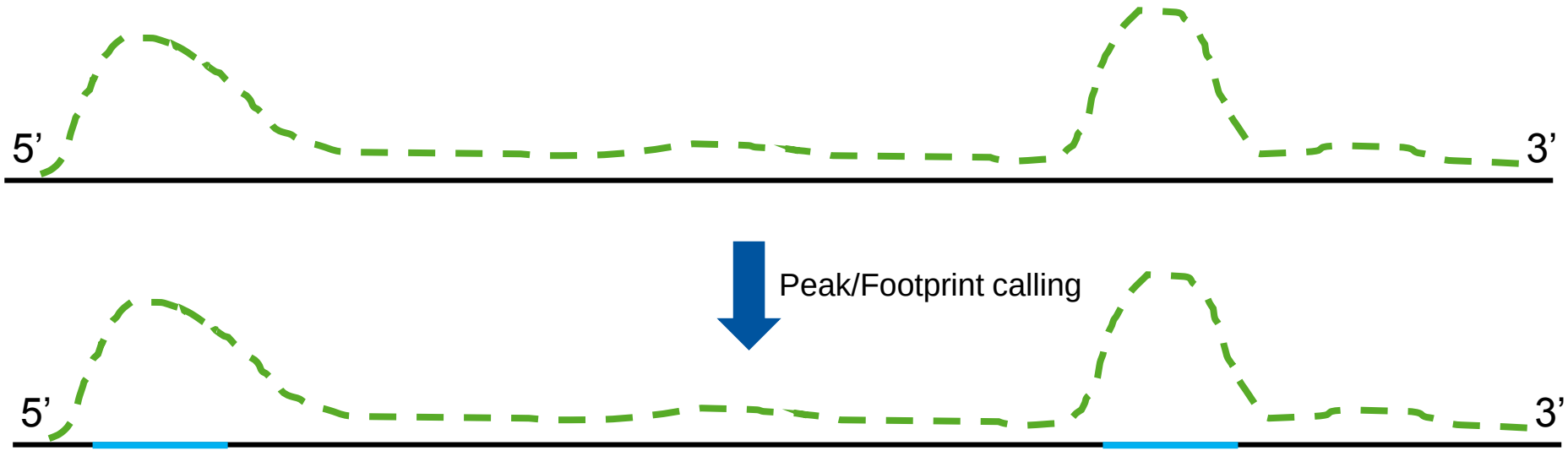
Segmentation based TF annotation



Epigenetic signal Candidate region

Wellington, Piper et al., Nucleic Acids Res., 2013
HINT, Gusmao et al., Bioinformatics, 2014
TEPIC, Schmidt et al., NAR, 2016

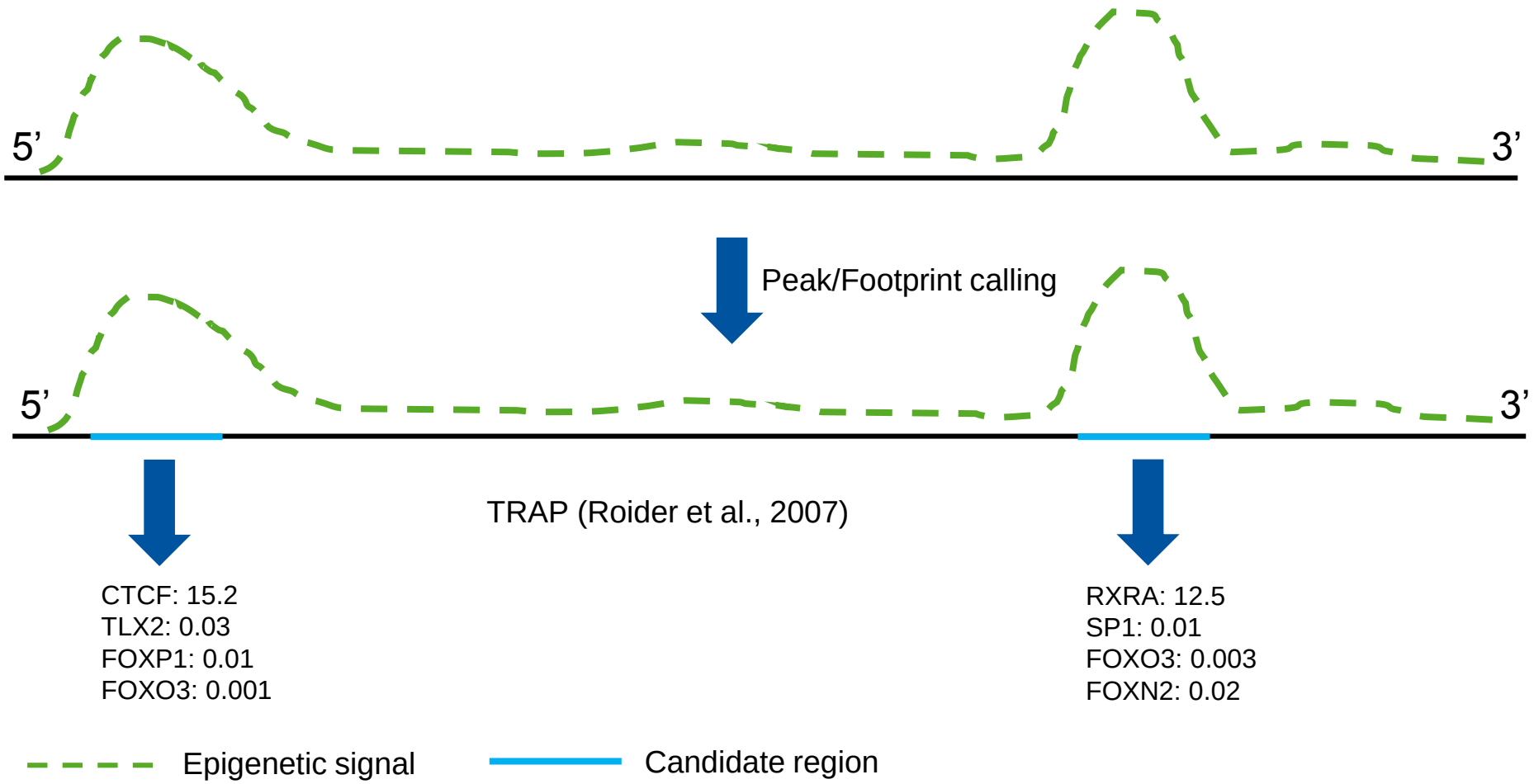
Segmentation based TF annotation



Epigenetic signal Candidate region

Wellington, Piper et al., Nucleic Acids Res., 2013
HINT, Gusmao et al., Bioinformatics, 2014
TEPIC, Schmidt et al., NAR, 2016

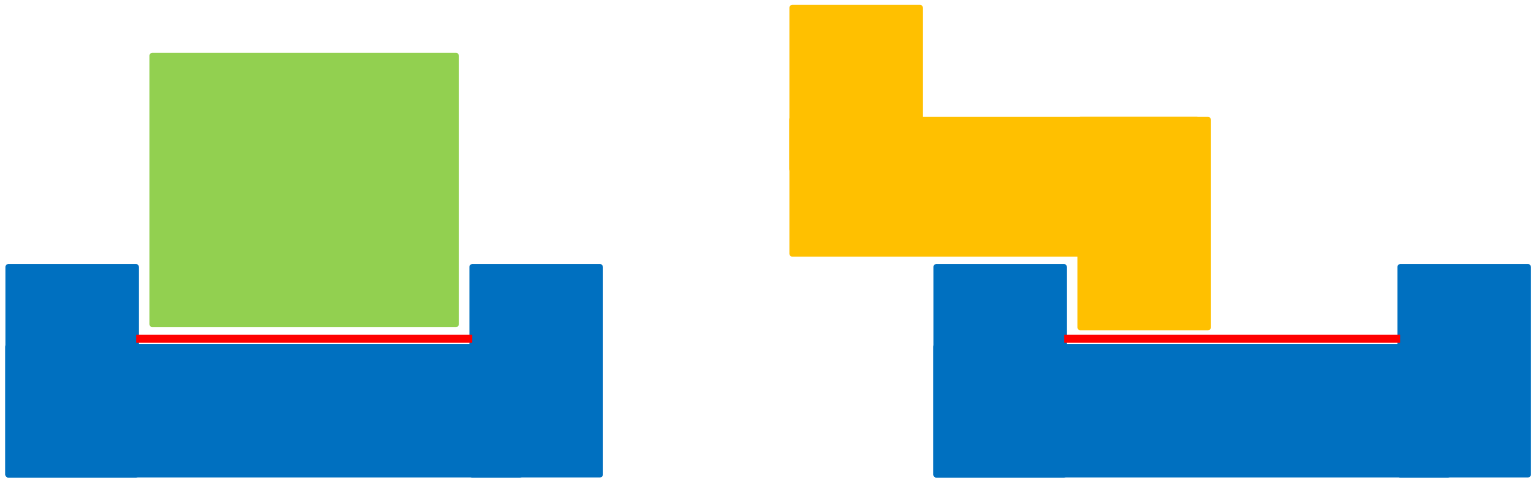
Segmentation based TF annotation



Wellington, Piper et al., Nucleic Acids Res., 2013
HINT, Gusmao et al., Bioinformatics, 2014
TEPIC, Schmidt et al., NAR, 2016

Why are we using TRAP?

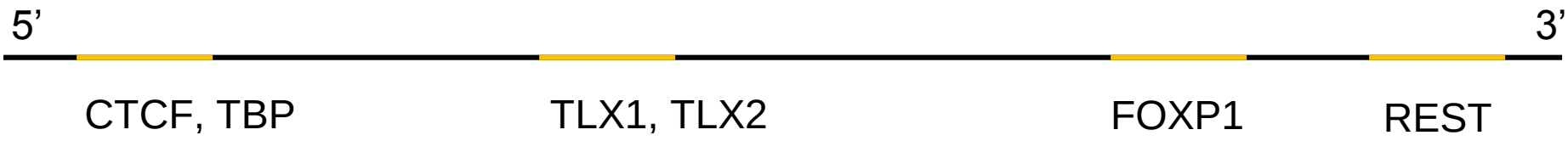
TF binding is not binary!



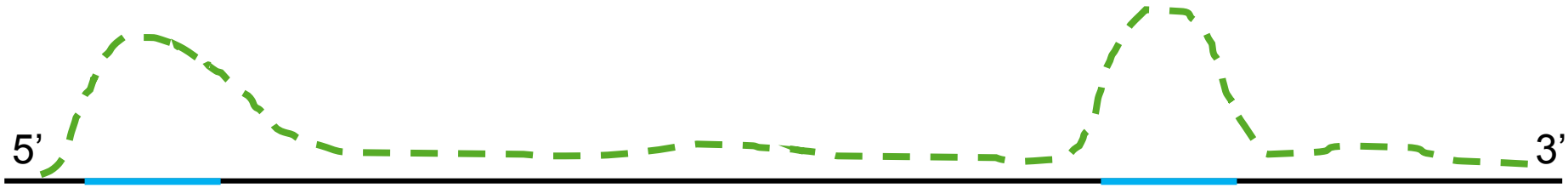
Von Hippel, P.H., and Berg, O.G, Proc. Natl. Acad. Sci. , 1985
Crocker, J. et al., Cell, 2015
Amos Tanay, Genome Research, 2006

What do we get from any of the methods?

Site-centric



Segmentation based

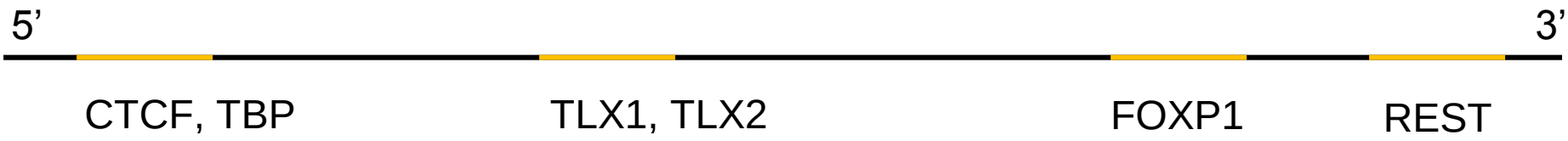


CTCF: 15.2
TLX2: 0.03
FOXP1: 0.01
FOXO3: 0.001

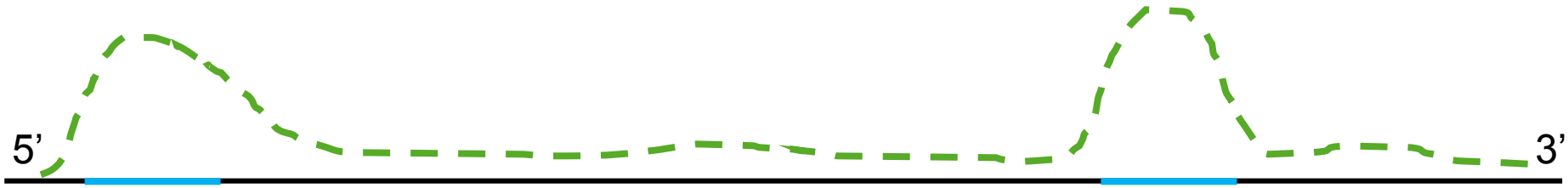
FOXP1: 12.3
TLX2: 0.03
CTCF: 0.01
FOXO3: 0.001

What do we get from any of the methods?

Site-centric



Segmentation based



CTCF: 15.2
TLX2: 0.03
FOXP1: 0.01
FOXO3: 0.001

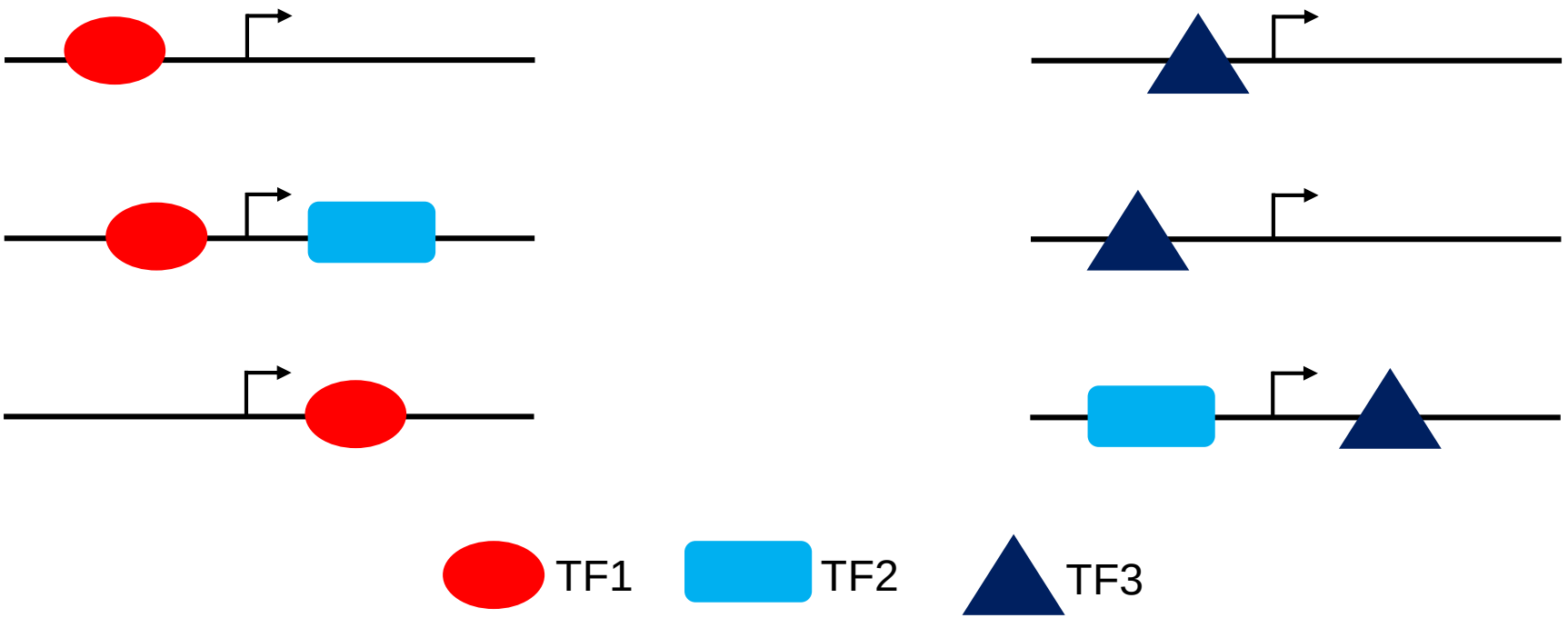
How do we know which TFs are important for our biological question?

FOXP1: 12.3
TLX2: 0.03
CTCF: 0.01
FOXO3: 0.001

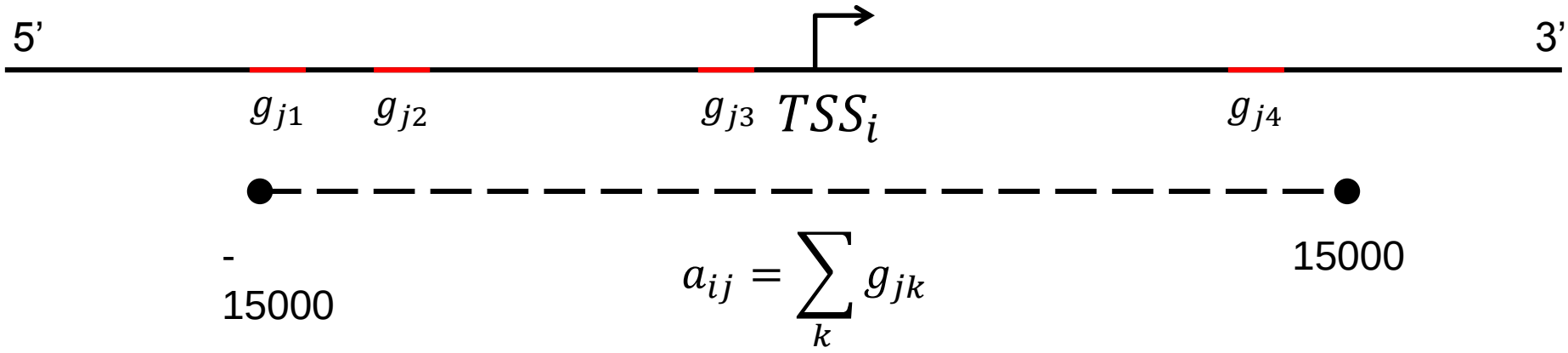
Identify TFs that bind specifically to up/down regulated genes

Up-regulated genes

Down-regulated genes



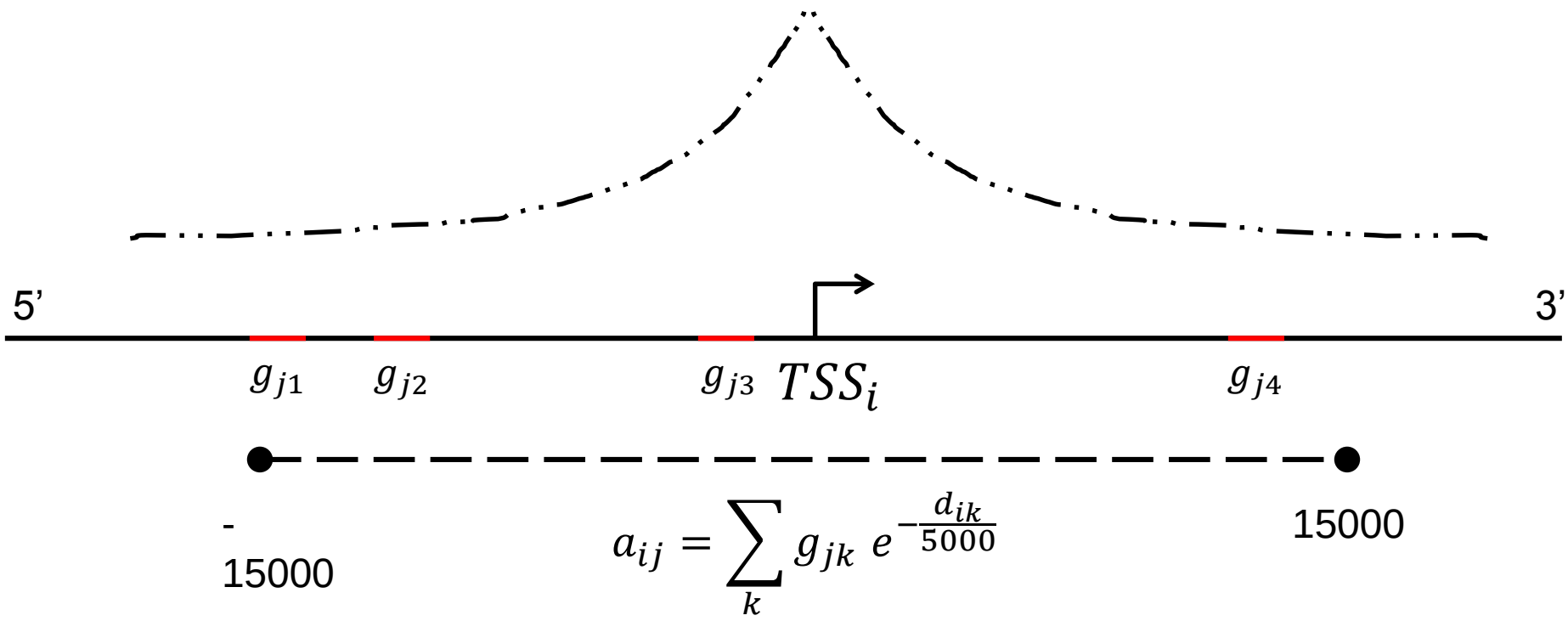
Computing TF-gene scores



a_{ij} = Affinity of TF j in gene I

g_{jk} = Affinity of TF j in peak k

Computing TF-gene scores



a_{ij} = Affinity of TF j in gene I

g_{jk} = Affinity of TJ j in peak k

d_{ik} = Distance between the TSS of gene i to the middle of peak k

Constructing TF features

TF affinities, cell type1

Gene	TF1	TF2	…
A	13	2	
B	0.5	10	
…			

TF affinity ratios between cell types 1 and 2

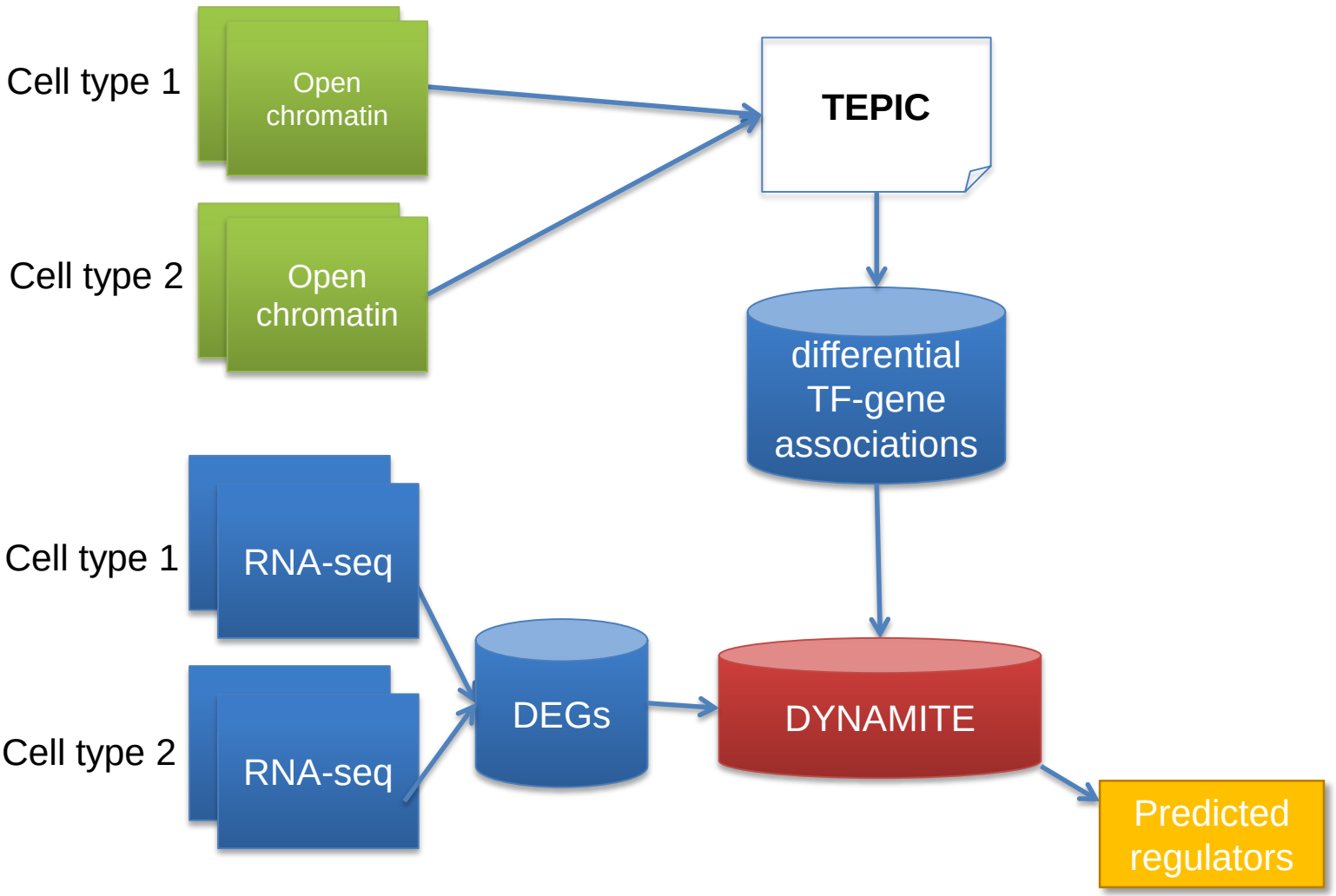
Gene	TF1	TF2	…
A	13	0.2	
B	0.05	0.5	
…			

.
.
.

TF affinities, cell type2

Gene	TF1	TF2	…
A	1	10	
B	10	20	
…			

Input to DYNAMITE



Data matrix used as input for the classifier

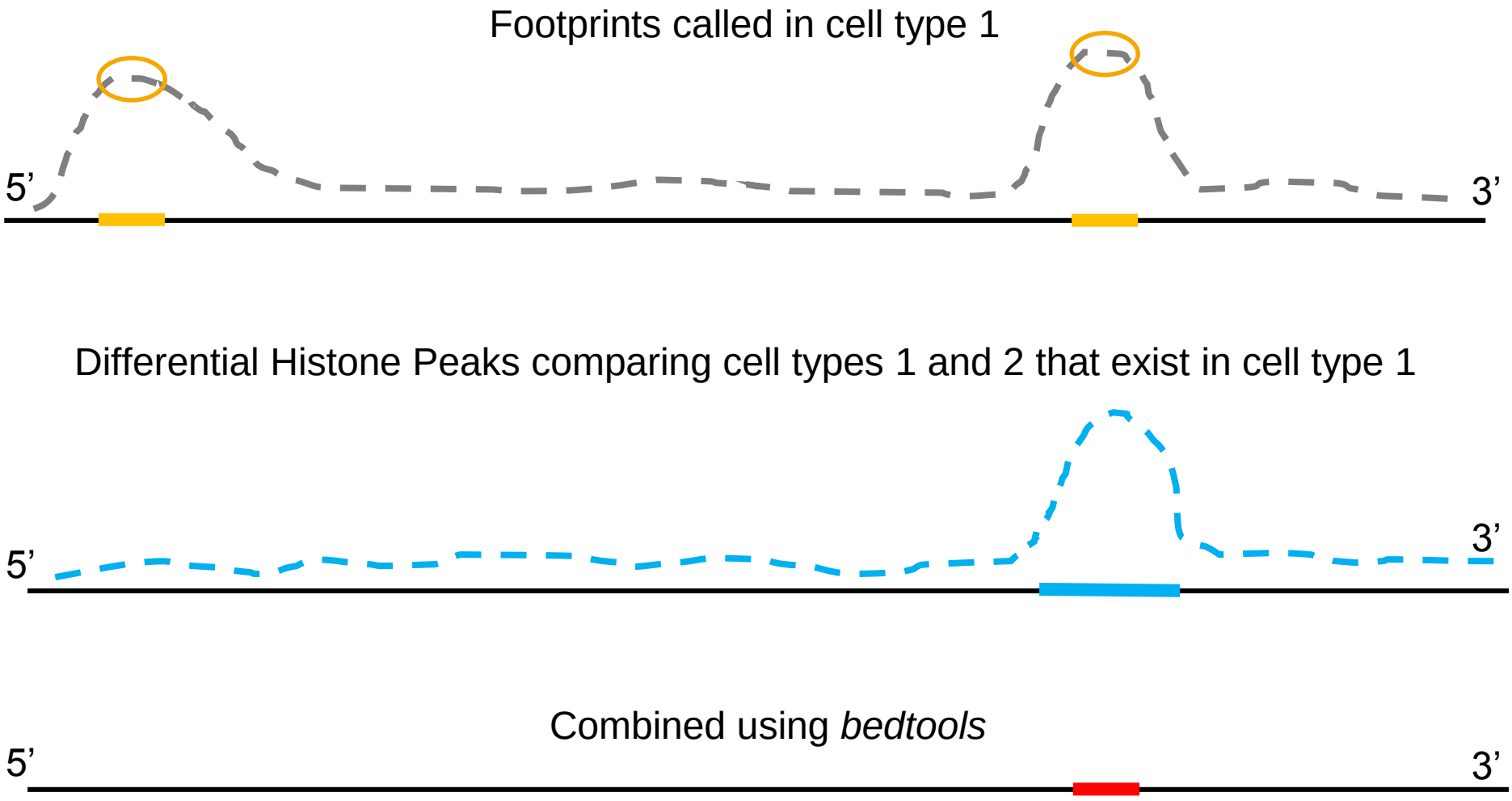
Gene	TF1	…	TF*n*
A	1.2		3.9
B	4.2		0.7
C	0.8		1.7
D	0.4		1.6
E	1.0		1.2
…			

~

Expression Changes*
Up
Down
Down
Up
Up
…

*discretized log2 fold ratios, methods to compute dif. Exp. Genes are e.g. *Cuffdiff* or *DESeq2*.

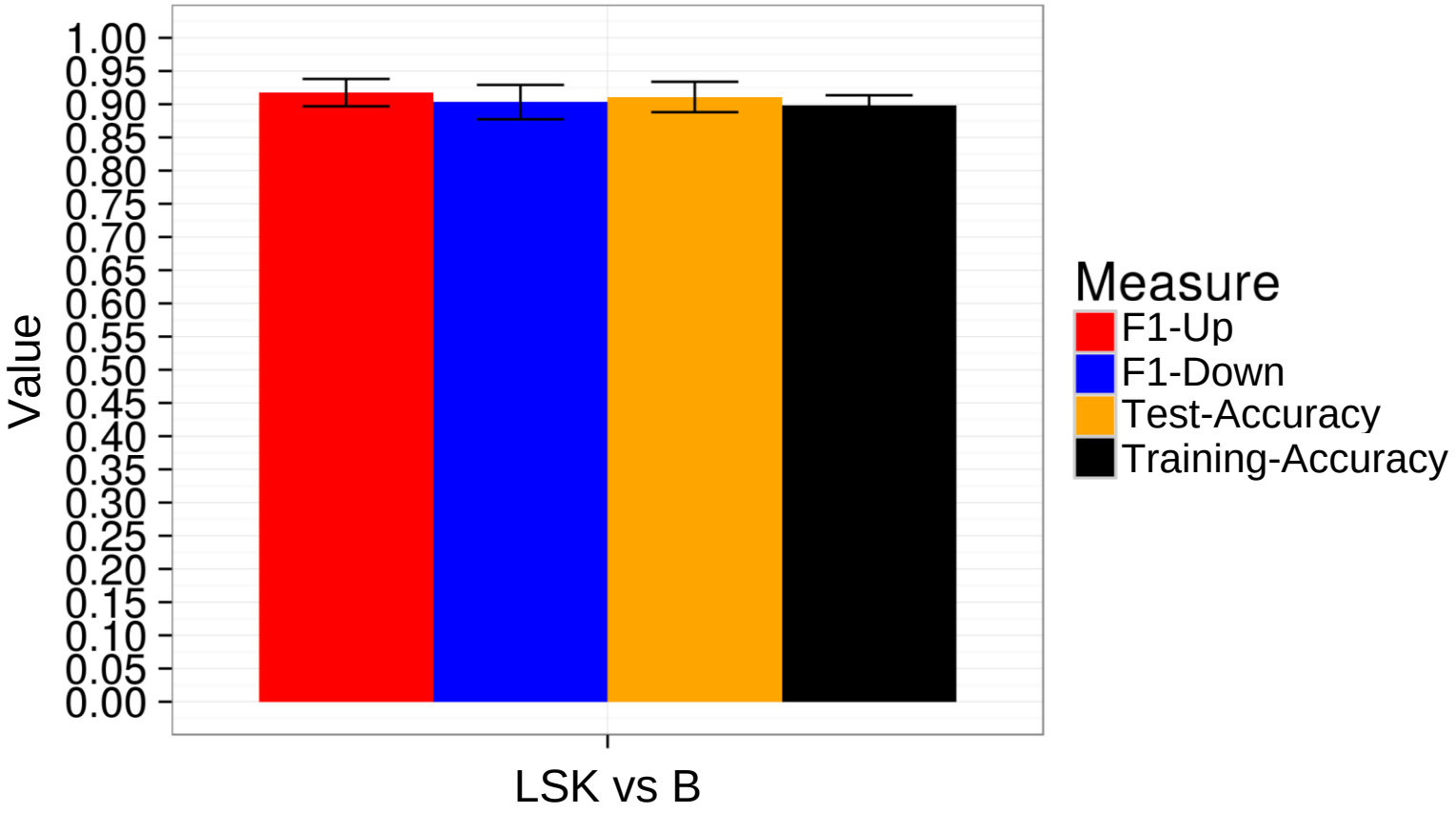
How to prepare the input data from the footprints and the differential Histone peaks?



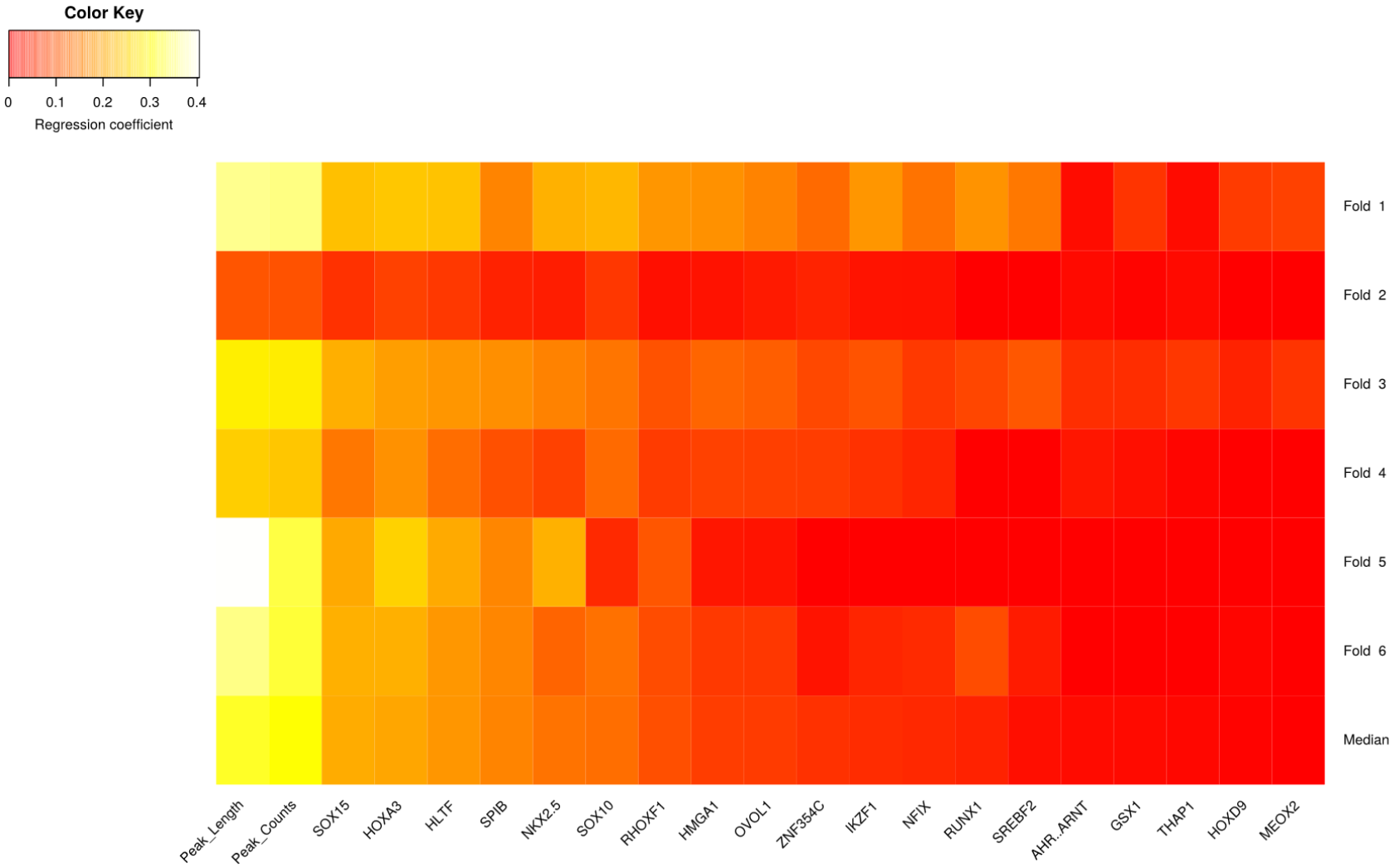
Example run of DYNAMITE: LSK vs B

	LSK	B
Candidate regions for TF binding	4,254	12,404
Differentially expressed genes	2239	
Number of genes used as input	466	

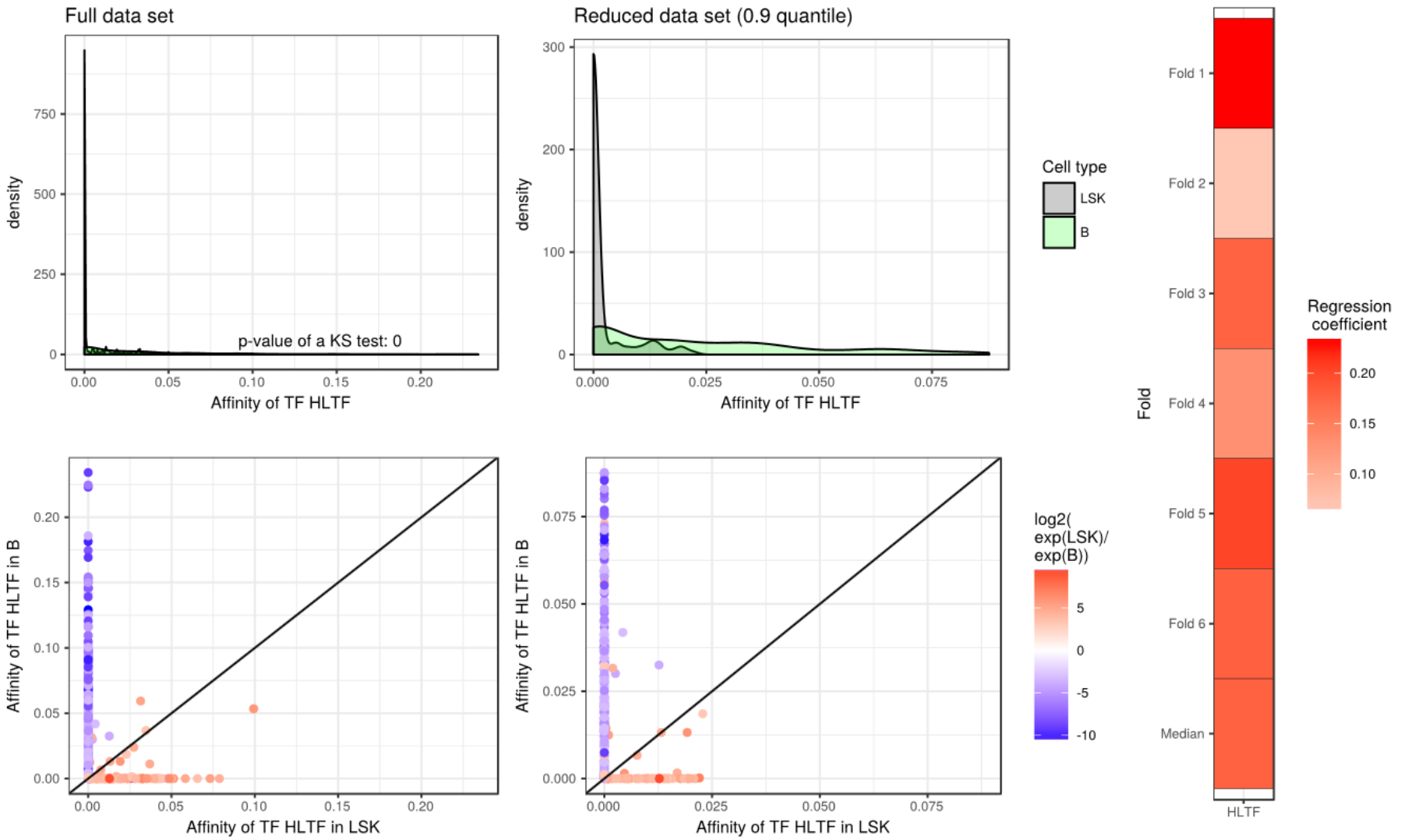
Model performance



Resulting Feature Heatmap



Closer feature investigation



Next

Time	Topic	Who
2:30 - 2:45	Introduction / gene regulation / transcription / chromatin	IC
2:45 - 3:00	Introduction ChIP-seq peak calling	MH
3:00 - 3:50	Practical peak calling	MH & JH
4:15 - 4:30	Introduction Footprints	IC
4:30 - 4:45	Introduction Regulatory networks	MS
4:45 - 5:50	Practical Regulatory Networks	IG, MS & FS
5:50 - 6:00	Q & A session	all

Material - https://github.com/SchulzLab/EpigenomicsTutorial-ISMB2017

Team

Ivan Costa (IC)

Matthias Heinig (MH)

Johann Hawe (JH)

Marcel Schulz(MH)

Florian Schmidt (FS)

Learning setup of DYNAMITE