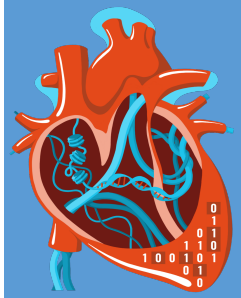
Epigenomics tutorial

Part2 - Approaches for identifying target genes of regulatory elements

Nina Baumgarten, Dennis Hecker, Sivarajan Karunanithi, Marcel H. Schulz



Computational Epigenomics and Systems Biology Group
Institute for Cardiovascular Regeneration
Goethe University Frankfurt

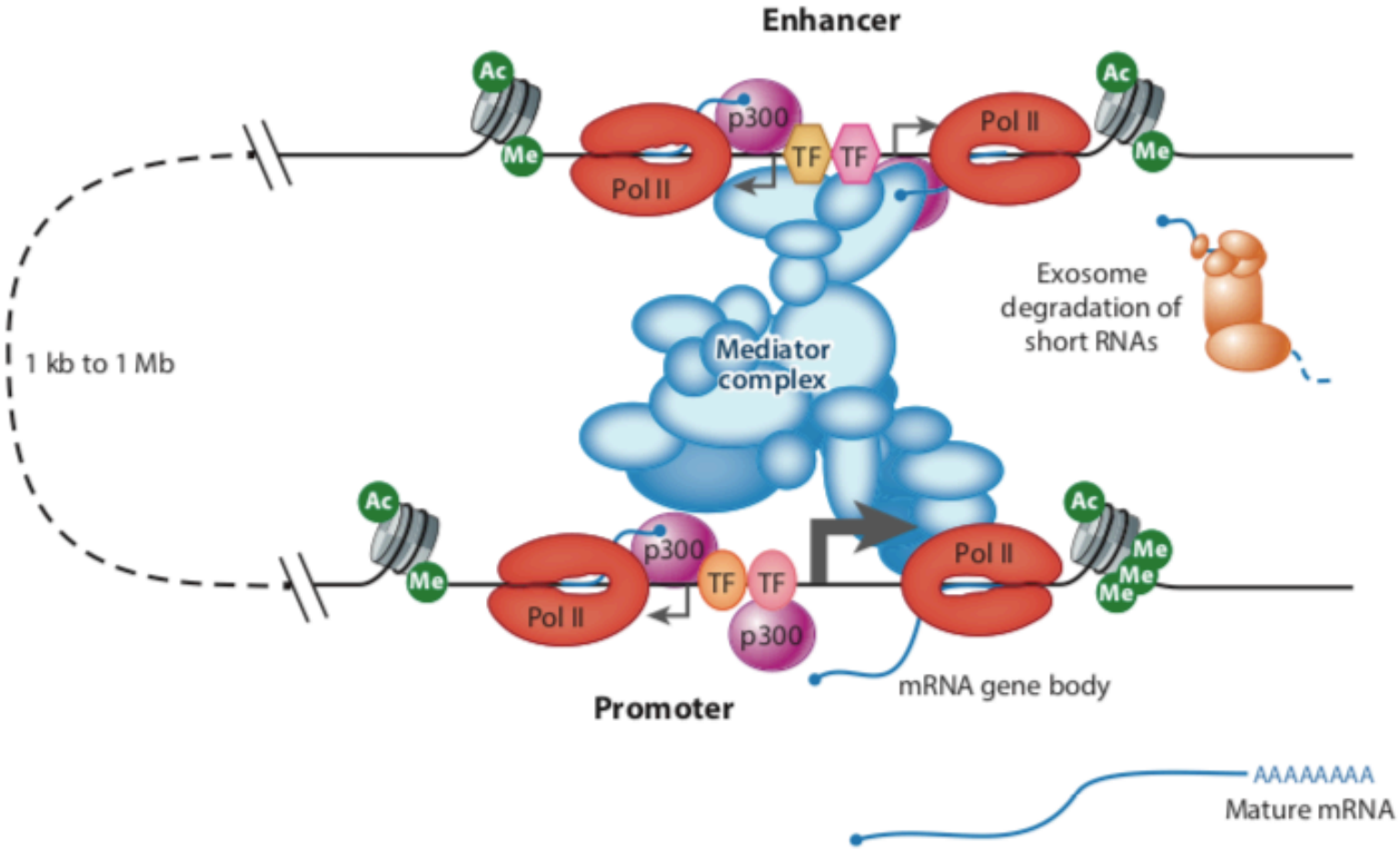
GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

Questions addressed in this part

- How to identify transcription factors that may be involved in regulating genes of interest
- How to identify possible target genes of regulatory elements?
- How to predict biological functions of a transcription factor?

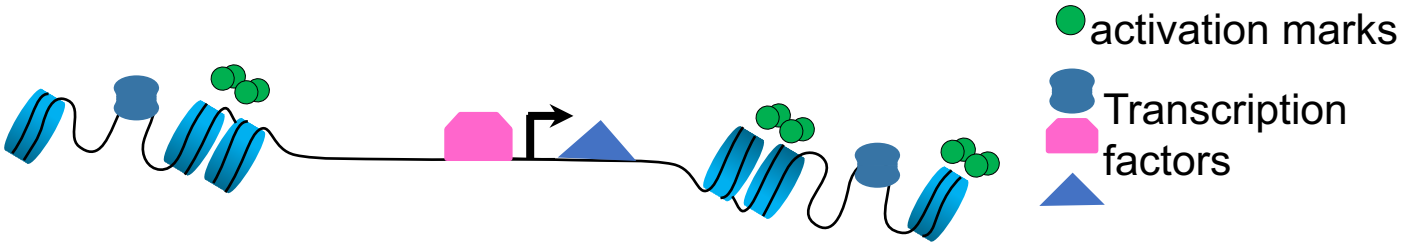
Part 1: Prediction of target genes of regulatory elements

Enhancers – promoter complex

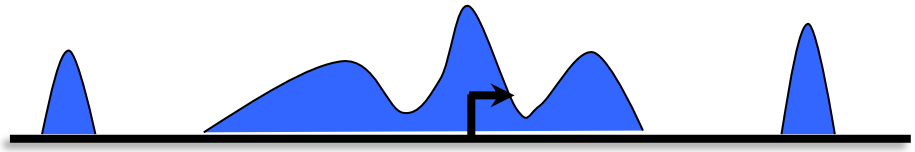


Field and Adelman 2020
https://doi.org/10.1146/annurev-biochem-011420-095916

Integrating epigenetic changes along the DNA



open-chromatin
Regions / footprints
(DNase1, ATAC, NOMe)



Integrative prediction of TF binding
Ernst et al. Gen Res 2010
Natarajan et al. Gen Res 2012
ENCODE consortium Nature 2012
Sherwood et al. Nat Biotech. 2014

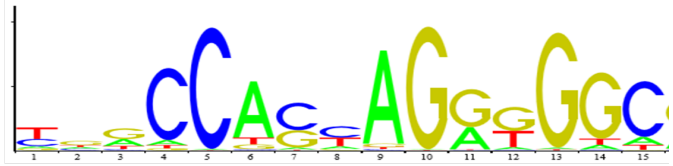
Gene expression prediction
Budden et al. Brief Bioinfor 2015
Mcleay et al. Bioinformatics 2012
TEPIC - Schmidt et al. NAR 2017

Approaches for linking elements to genes

1. Window-based
2. Nearest-gene
3. Using chromatin-conformation measurements
4. Association-based

Hit-based vs. Affinity based TF occupancy prediction

PWM scores



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

"Hits" are determined using a significance threshold relying on a statistical null model

PWM scores

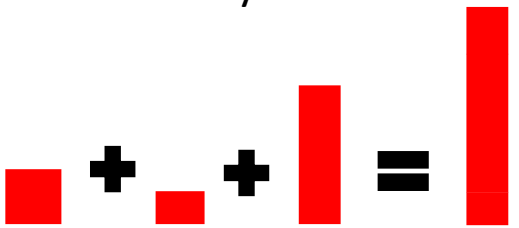
significance threshold

e.g. Grant et al. 2011 Bioinformatics (FIMO)

Estimate average number of bound TF molecules (affinity) in a region

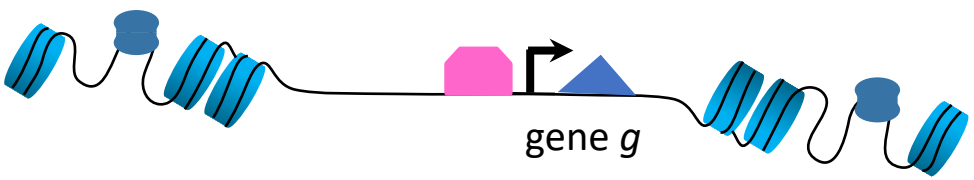
$$p(S) = \frac{[TF \cdot S]}{[S] + [TF \cdot S]} = \frac{K \cdot [TF]}{1 + K \cdot [TF]}.$$

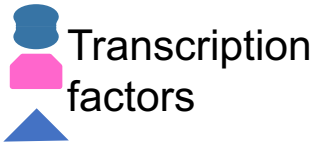
Affinity score



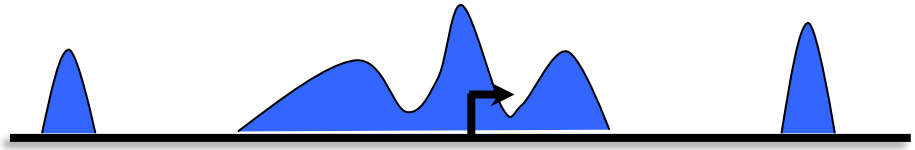
Roider et al. 2006 Bioinformatics (TRAP)

Use epigenomic data to predict TF binding (TEPIC)





open-chromatin regions



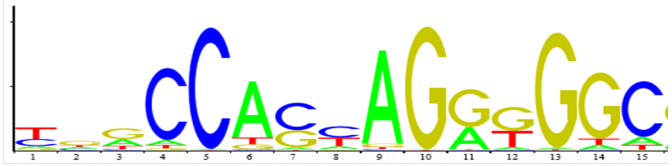
peak calling

annotation for TF *i* (TRAP^)

3.4 10.8 0.2

726 TFs
Jaspar
Hocomoco
ENCODE motifs

*Mathelier et al. 2015,
Kulakovskiy et al. 2016,
Kheradpour & Kellis 2014



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

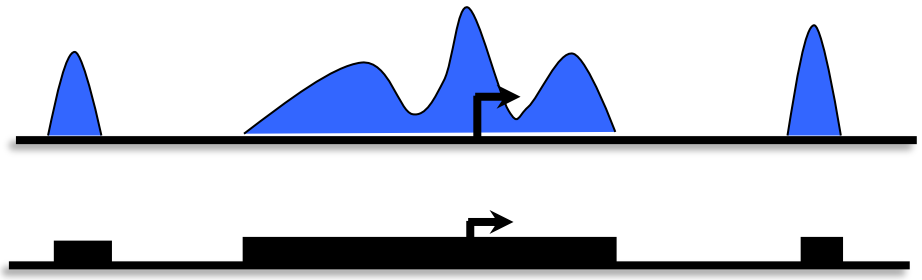
^ Roider et al. 2008, Bioinformatics

Use epigenomic data to predict TF binding (TEPIC[&])

open-chromatin signal

peak calling

annotation for TF *i* (TRAP)



$$a_{g,i} = \sum_{p \in P_{g,y}} a_{p,i} s_p e^{-\frac{d_{p,g}}{d_0}}$$

$a_{g,i}$ Affinity score for TF i at gene g

$P_{g,y}$ Set of all peaks p within a window of size y around the TSS of gene g

$d_{p,g}$ Distance of peak p to TSS of gene g

d_0 Constant used in weighting the distance $d_{p,g}$

$a_{p,i}$ Affinity score for TF i in peak p

s_p Magnitude of the signal within peak p

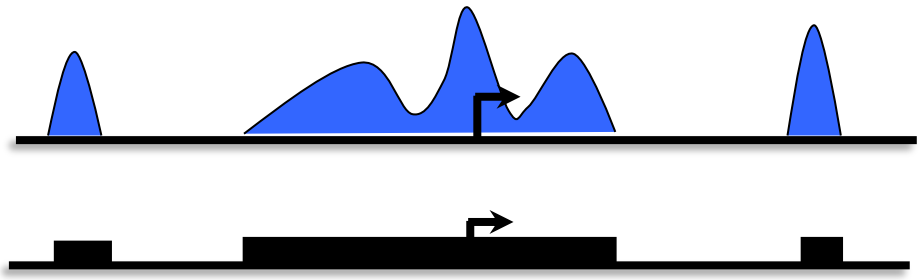
& Schmidt et al. *NAR* 2017

Use epigenomic data to predict TF binding (TEPIC[&])

open-chromatin signal

peak calling

annotation for TF *i* (TRAP)



$$a_{g,i} = \sum_{p \in P_{g,y}} a_{p,i} s_p e^{-\frac{d_{p,g}}{d_0}}$$

$a_{g,i}$ Affinity score for TF i at gene g

$P_{g,y}$ Set of all peaks p within a window of size y around the TSS of gene g

$d_{p,g}$ Distance of peak p to TSS of gene g

d_0 Constant used in weighting the distance $d_{p,g}$

$a_{p,i}$ Affinity score for TF i in peak p

s_p Magnitude of the signal within peak p

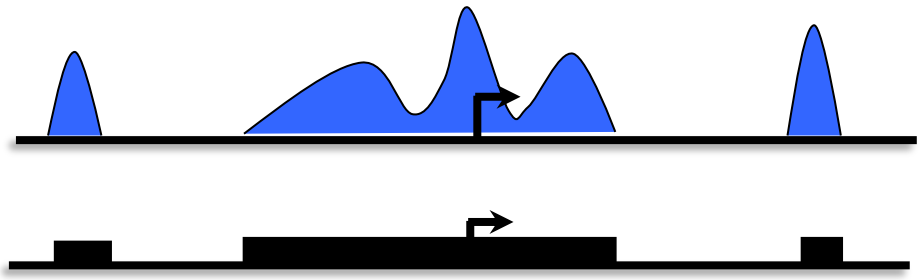
& Schmidt et al. *NAR* 2017

Use epigenomic data to predict TF binding (TEPIC[&])

open-chromatin signal

peak calling

annotation for TF *i* (TRAP)



$$a_{g,i} = \sum_{p \in P_{g,y}} a_{p,i}\, s_p\, e^{-\frac{d_{p,g}}{d_0}}$$

distance weight*

$a_{g,i}$ Affinity score for TF i at gene g

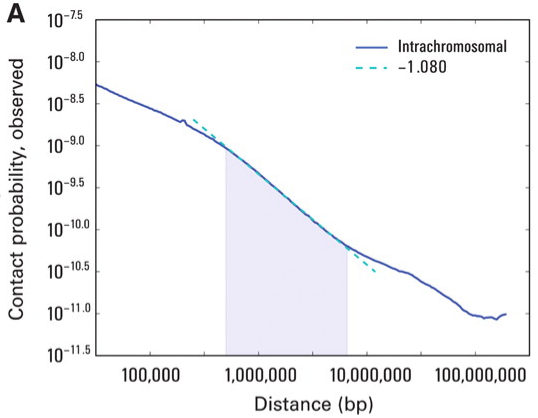
$P_{g,y}$ Set of all peaks p within a window of size y around the TSS

$d_{p,g}$ Distance of peak p to TSS of gene g

d_0 Constant used in weighting the distance $d_{p,g}$

$a_{p,i}$ Affinity score for TF i in peak p

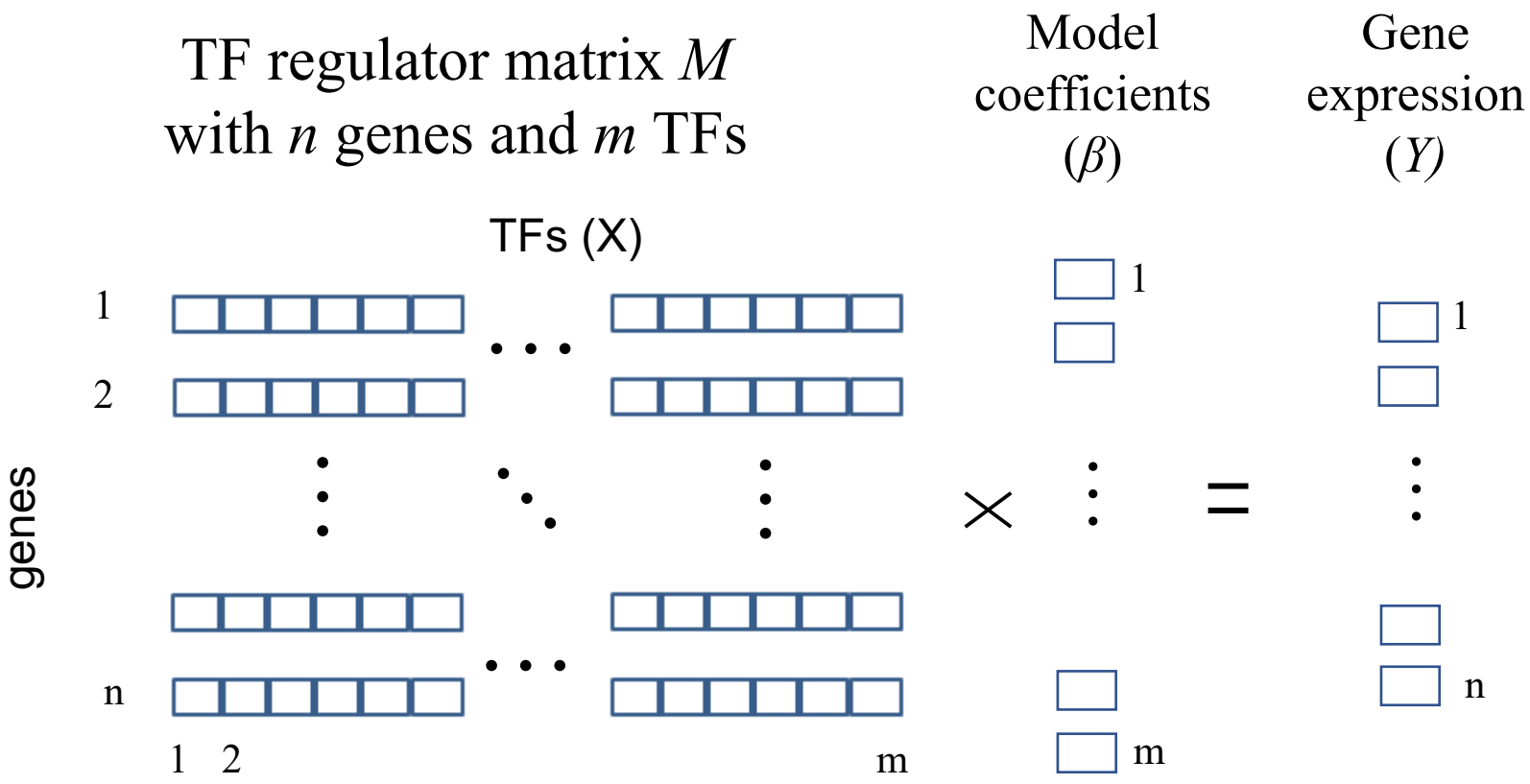
s_p Magnitude of the signal within peak p



*Ouyang et al. PNAS 2009
Lieberman-Aiden Science 2009

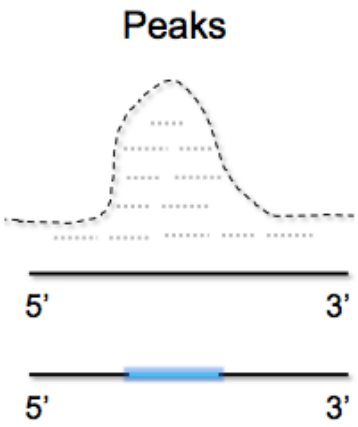
& Schmidt et al. *NAR* 2017

Predicting gene expression with linear regression (INVOKE)



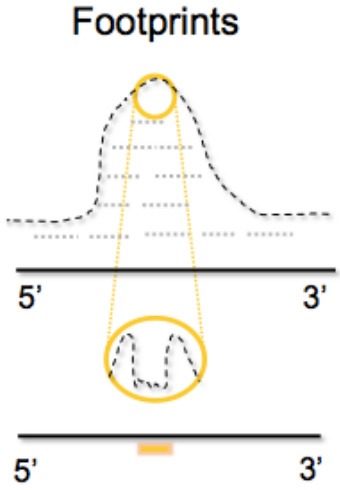
Solved using Elastic-net regularized linear regression, parameter optimization using CV
glmnet R package (Friedman et al. 2010)

Gene expression performance comparison: peaks vs footprints

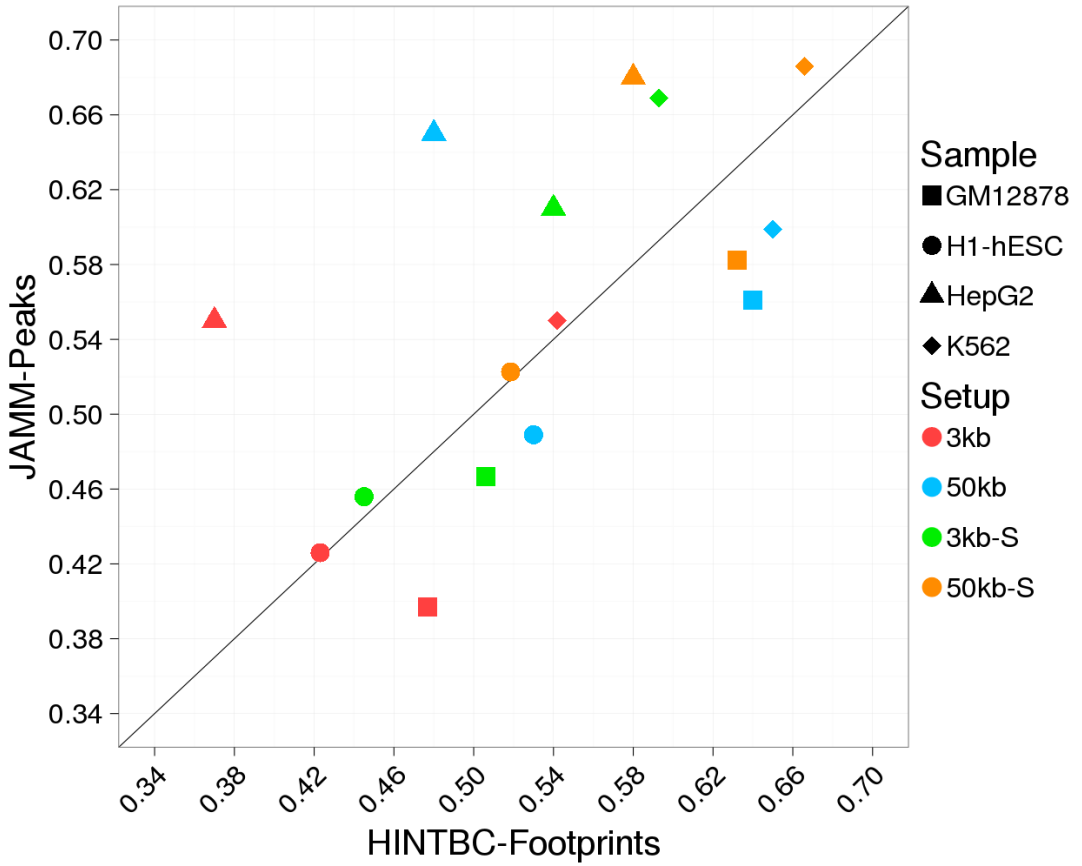


MACS2
(Zhang et al. Gen Biol 2008),

JAMM
(Ibrahim et al. Bioinf 2015)

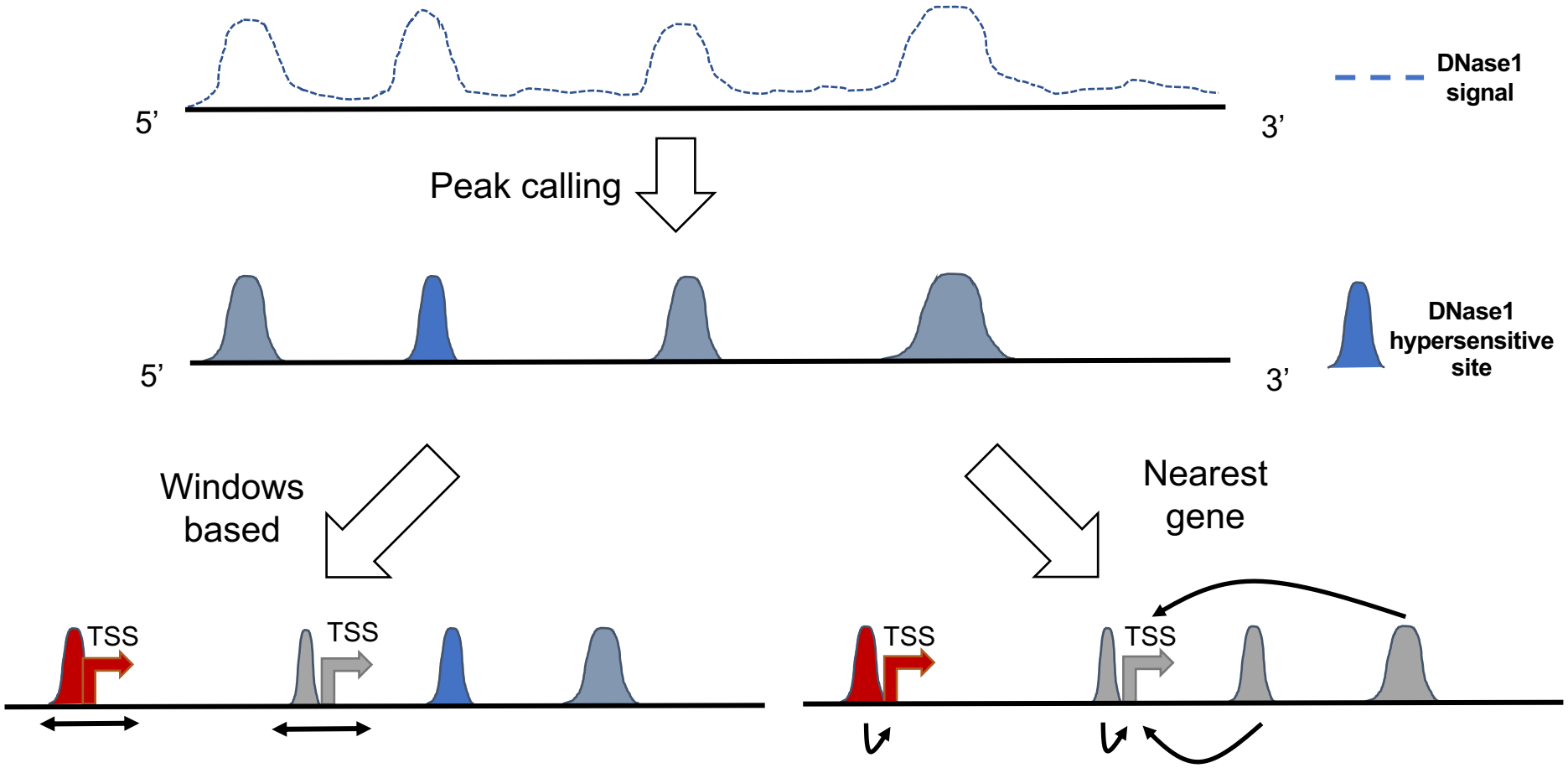


HINT-BC
(Gusmao et al. Nat Meth 2016)



Schmidt et al. NAR 2017

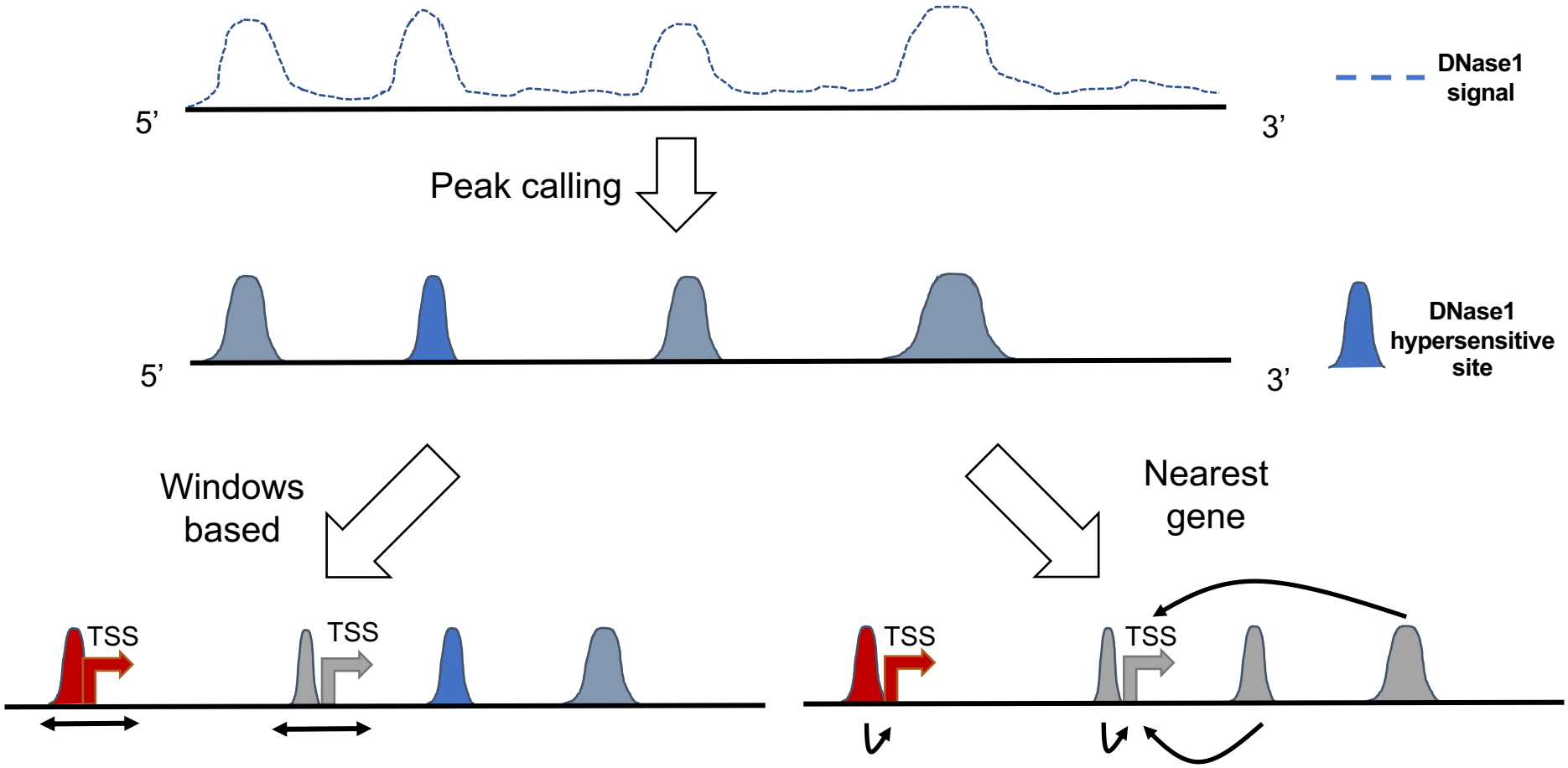
Window-based versus *Nearest gene based* peak association



Schmidt et al. 2020

Window-based versus *Nearest gene based* peak association

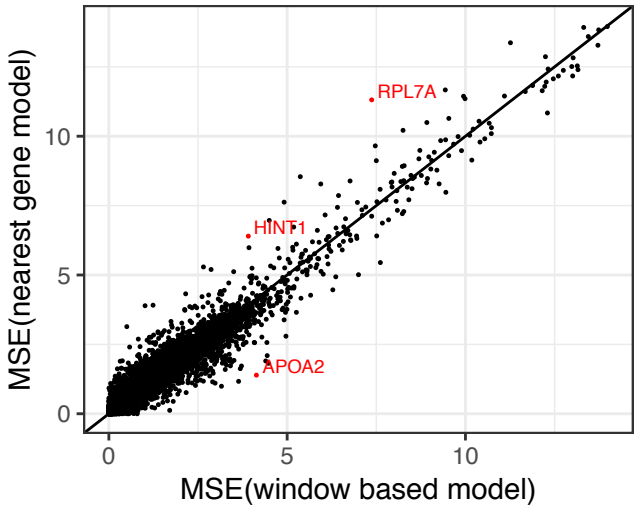
Do interactive Pol!



Schmidt et al. 2020

Both models are imperfect

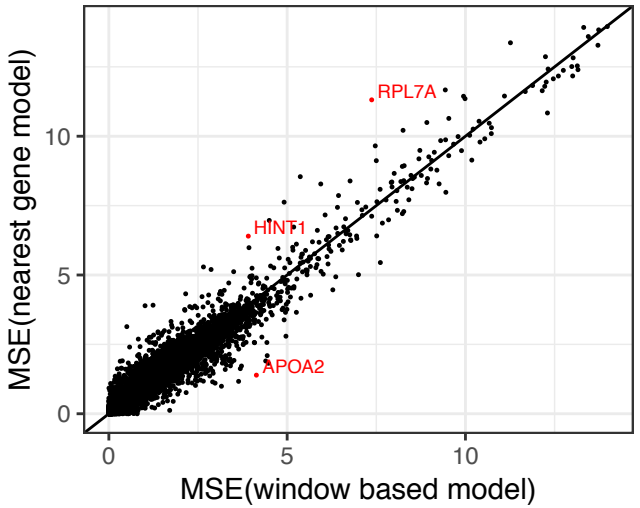
DNase1-seq experiments in HeLa cells

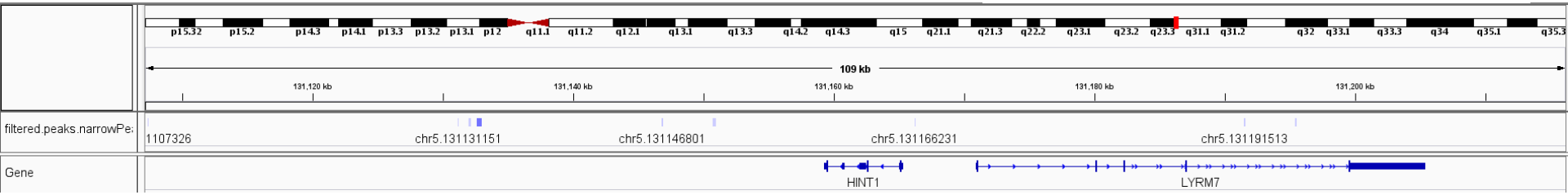


Schmidt et al. 2020

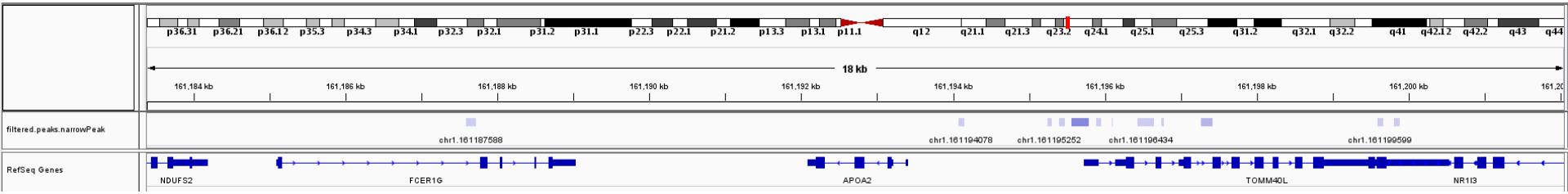
Both models are imperfect

DNase1-seq experiments in HeLa cells





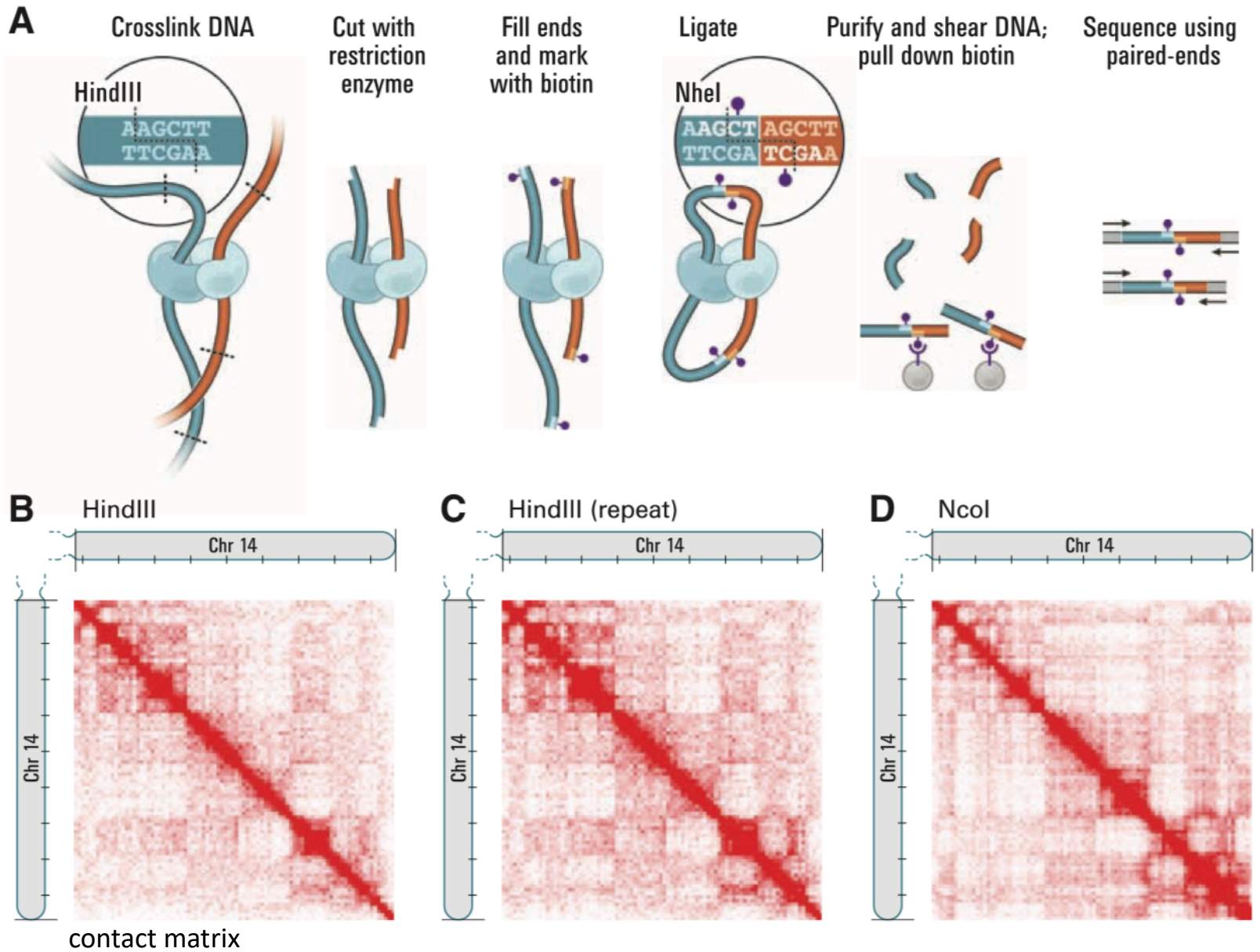
HINT1



APOA2

Schmidt et al. 2020

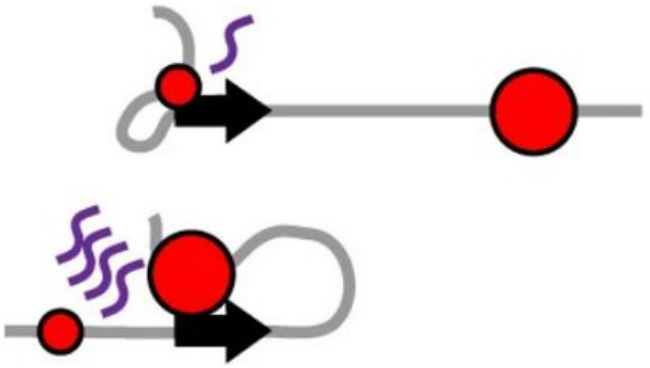
Hi-C measurement of DNA-DNA interactions in cells



contact matrix

Lieberman-Aiden et al. 2009

ABC-score for integration of Hi-C contact counts



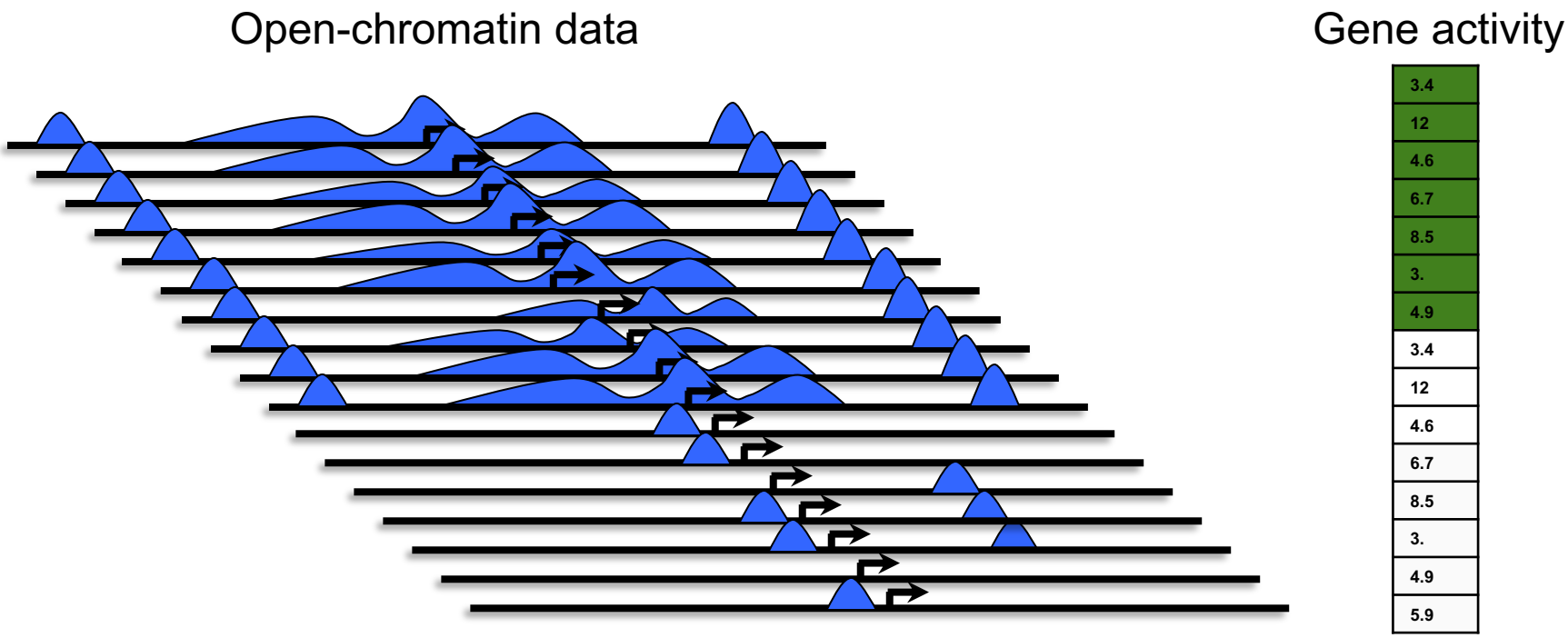
	Activity × contact = A × C					ABC score
e1:	1	×	12	=	12	0.75
e2:	4	×	1	=	4	0.25

Activity = H3K27ac signal in peaks residing in Hi-C window
Contact = Number of contact counts HiC-window to gene promoter

Fulco et al. 2019
https://www.nature.com/articles/s41588-019-0538-0

Association-based learning of enhancer-gene regulatory maps from BIG data

- Paired DNAse1-seq and gene expression data from the same cell type / tissue from consortia such as ENCODE, BLUEPRINT, IHEC, DEEP

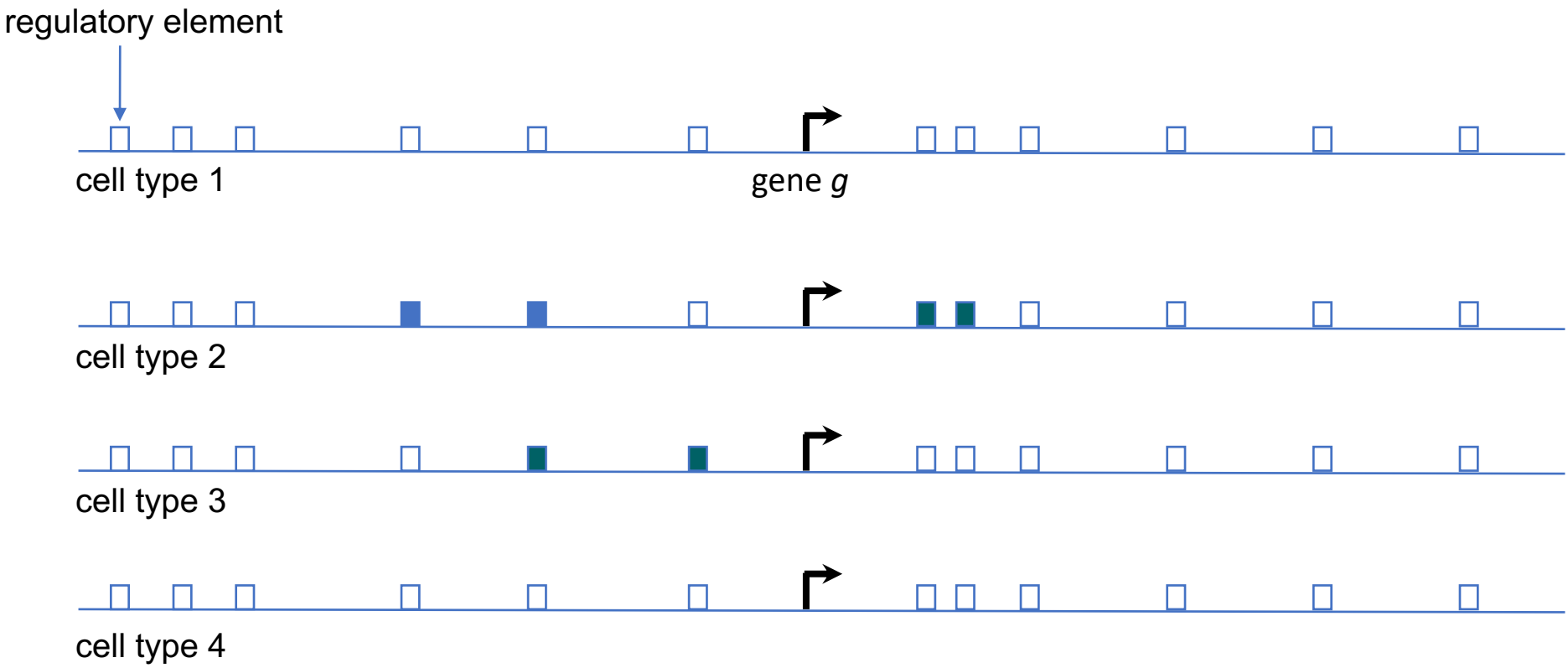


Software STITCHIT can do this: https://github.com/SchulzLab/STITCHIT

IHEC International Human Epigenome Consortium

EpiRegio DB knowledgebase

- STITCHIT predictions using 100kb windows for BLUEPRINT and ROADMAP data
- Contains 2.4 million regulatory elements linked to putative target genes in human



EpiRegioDB

EpiRegio DB knowledgebase

Questions that can be easily addressed:

1. What are putative regulatory elements for a set of genes ?
2. Which regions overlap putative regulatory elements ?
3. What are the potential target genes of regulatory elements ?

Data can be retrieved using:

- Webserver (https://epiregio.de)
- REST API (e.g. Python‘s requests module)

EpiRegioDB

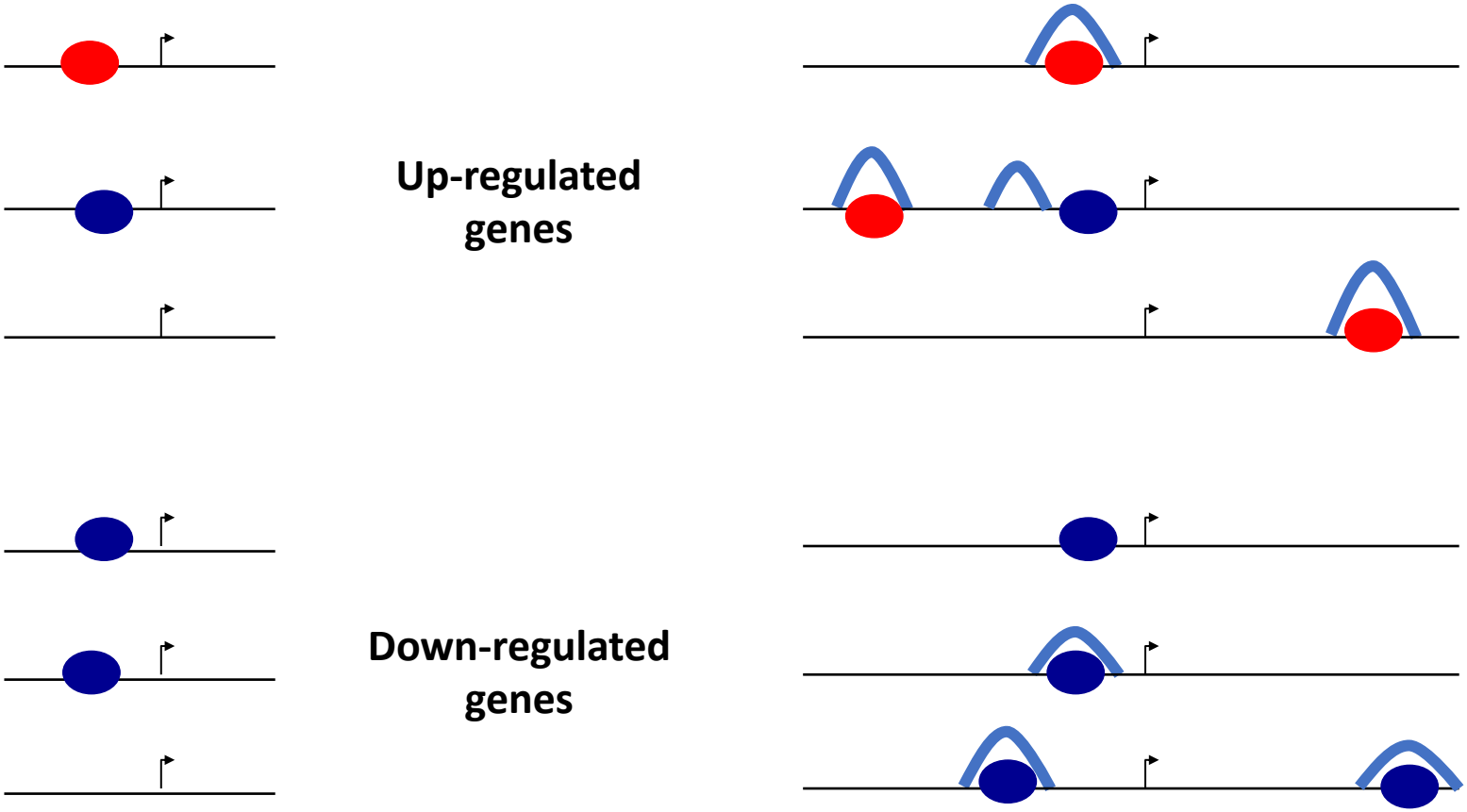
Part 2 – Integrating gene expression data for prioritization of important transcription factors

Interpretation of differential gene expression results

Which TFs regulate diffferences in gene expression?

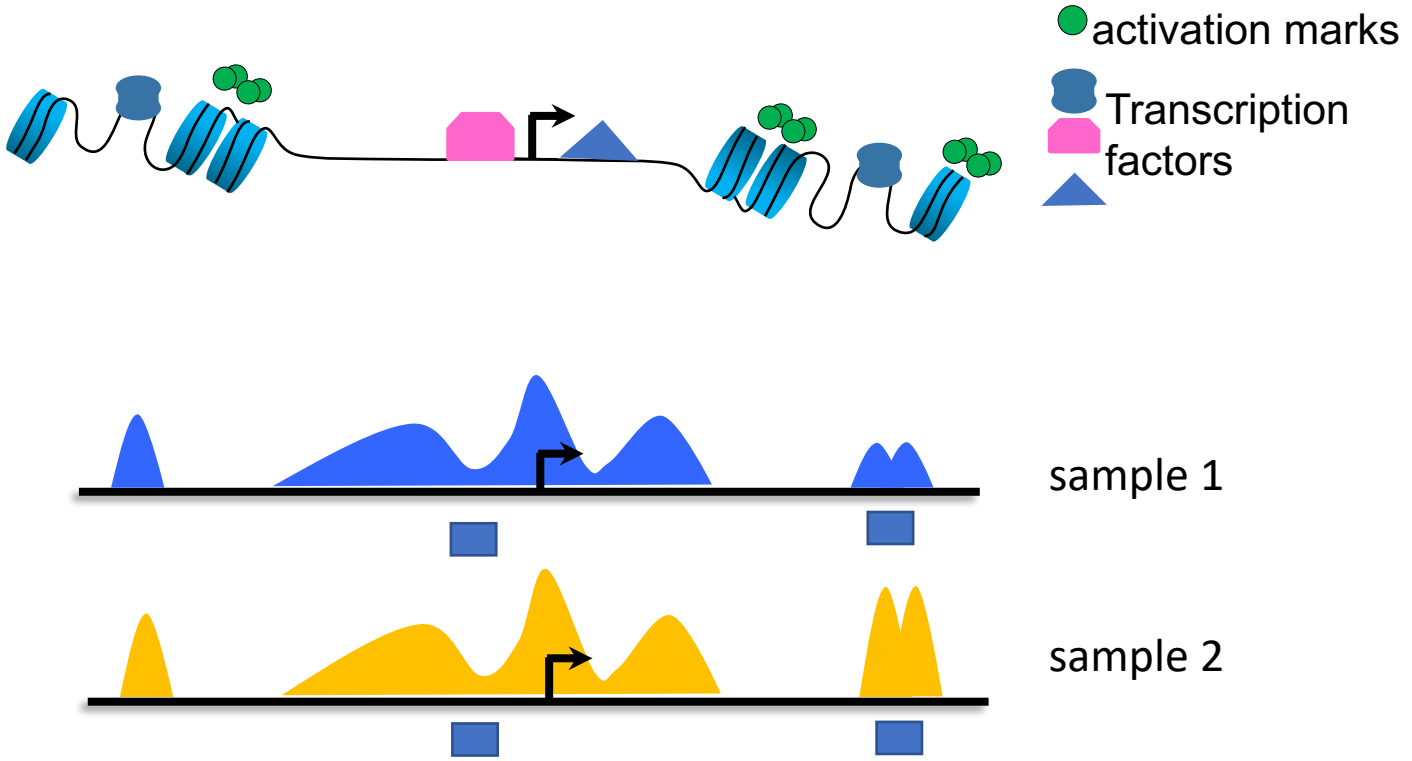
TF Enrichment Analysis

Dynamite analysis



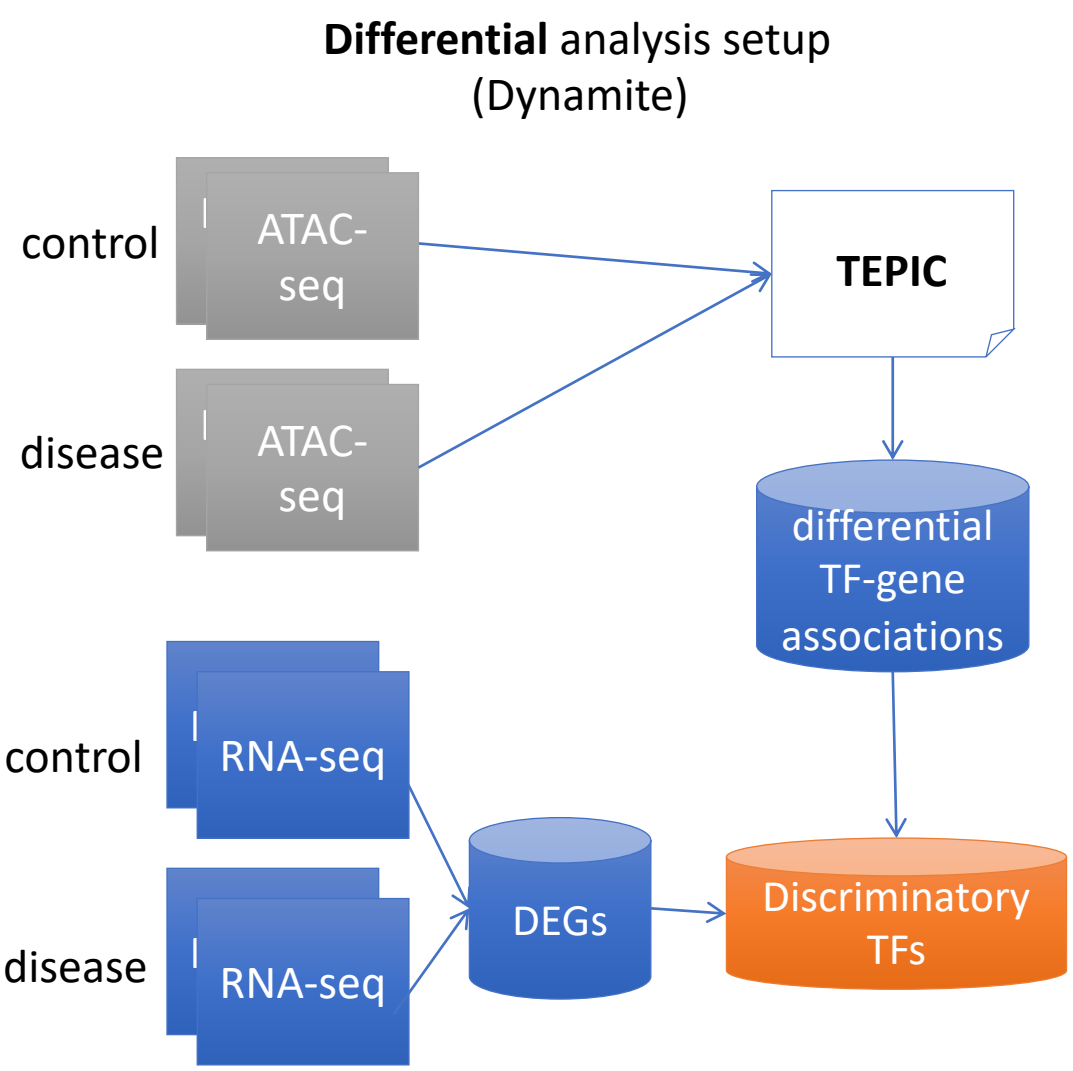
Schmidt et al 2018

Comparison to differential footprinting analysis



Analysis of differential footprinting

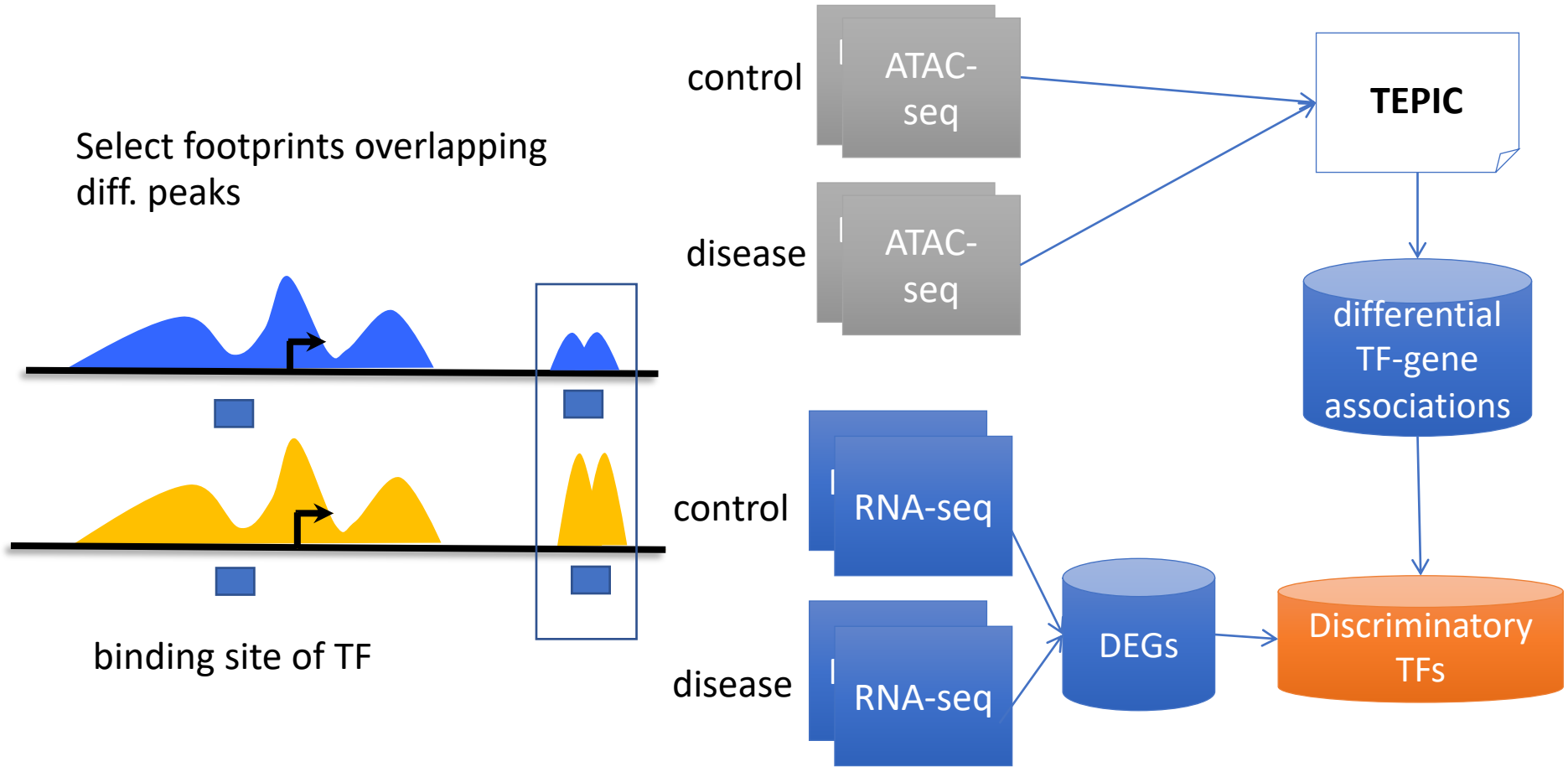
Dynamite workflow for differential analysis



Schmidt et al 2018

Dynamite workflow for differential analysis

Differential analysis setup (Dynamite)



Schmidt et al 2018

Dynamite uses a classifier setup

Data Matrix: Combine expression class and TF affinity ratios

Fold changes Diff. exp. genes

Gene	Expression Changes	TF1	TF2	…
A	Up	1.2	3.9	
B	Down	4.2	0.7	
C	Down	0.8	1.7	
D	Up	0.4	1.6	
E	Up	1.0	1.2	
…				

Schmidt et al 2018

Classification Setup – TF Affinity Data

TF affinities replicate 1, subtype 1

Gene	TF1	TF2	…
A	14	1	
B	0.5	12	
…			

TF affinities replicate 2, subtype 1

Gene	TF1	TF2	…
A	12	3	
B	0.5	8	
…			

.
.
.

Mean TF affinities subtype 1

Gene	TF1	TF2	…
A	13	2	
B	0.5	10	
…			

Mean TF affinities subtype 2

Gene	TF1	TF2	…
A	1	10	
B	10	20	
…			

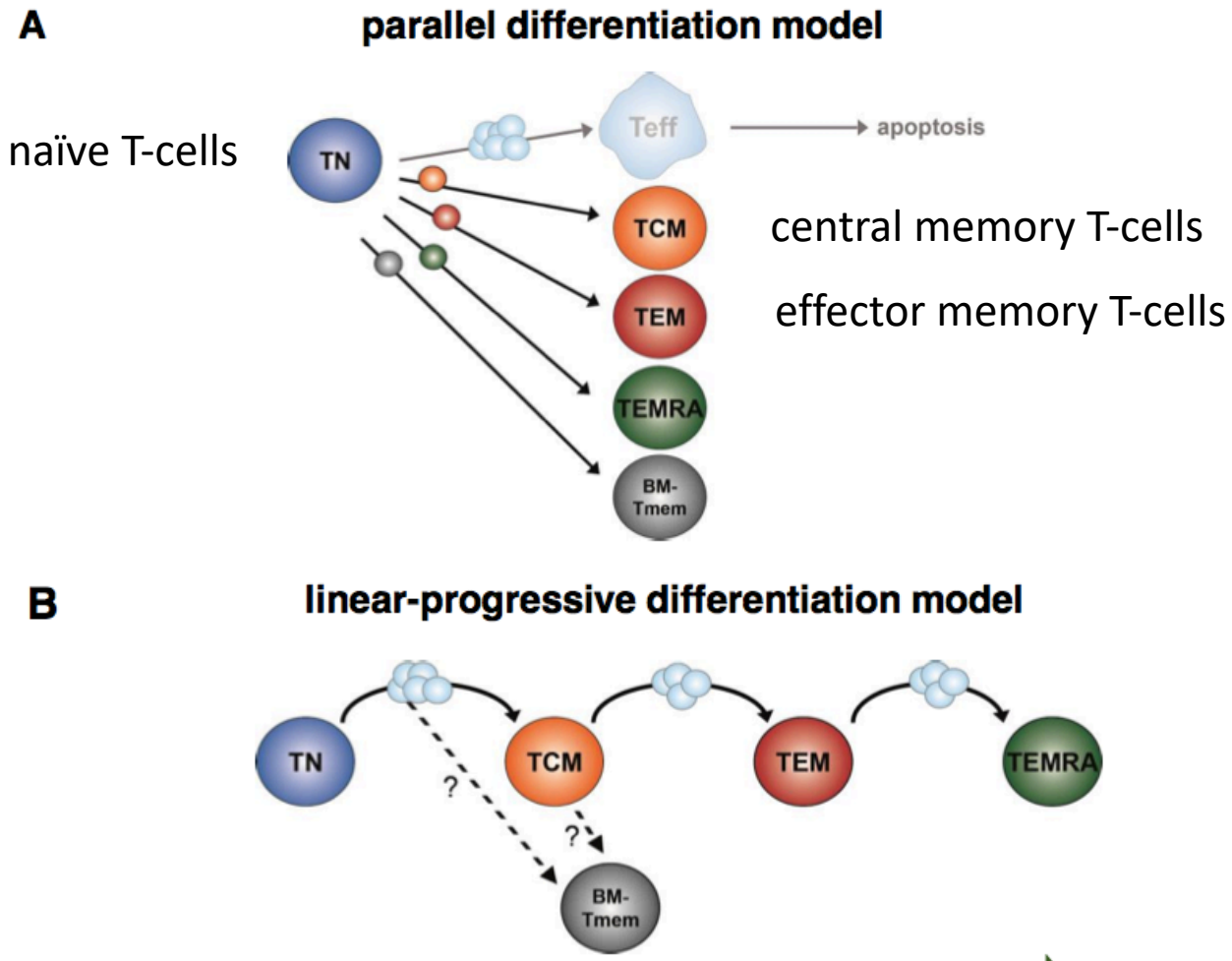
TF affinity ratios between subtype 1 and 2

Gene	TF1	TF2	…
A	13	0.2	
B	0.05	0.5	
…			

.
.
.

Schmidt et al 2018

Complete Epigenomes for different CD4+ memory T-cells

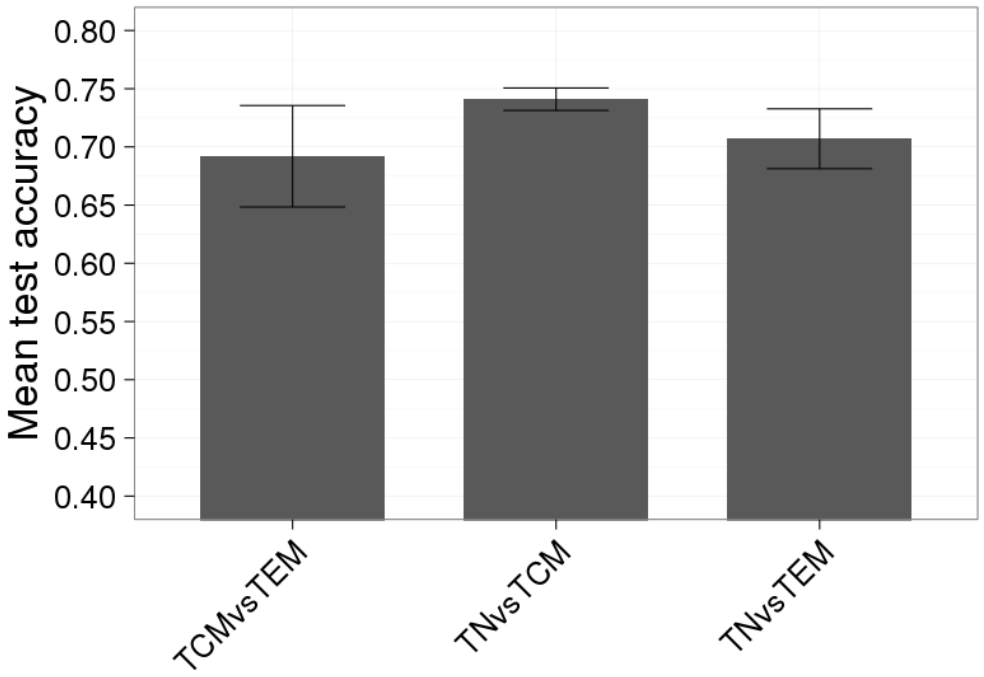


complete epigenomes (RNA, histone ChIP-seq, NOMe-seq, methylation), 3-10 female donors pooled

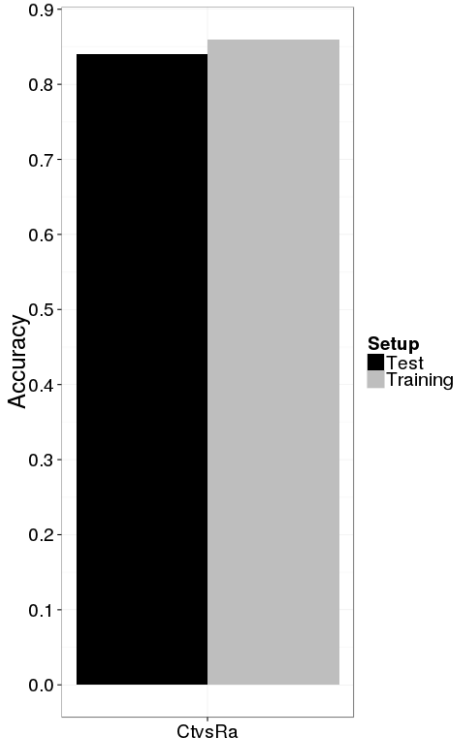
Durek et al. Cell Immunity 2016

Classification performance using logistic regression

Comparing CD4+ T-cells



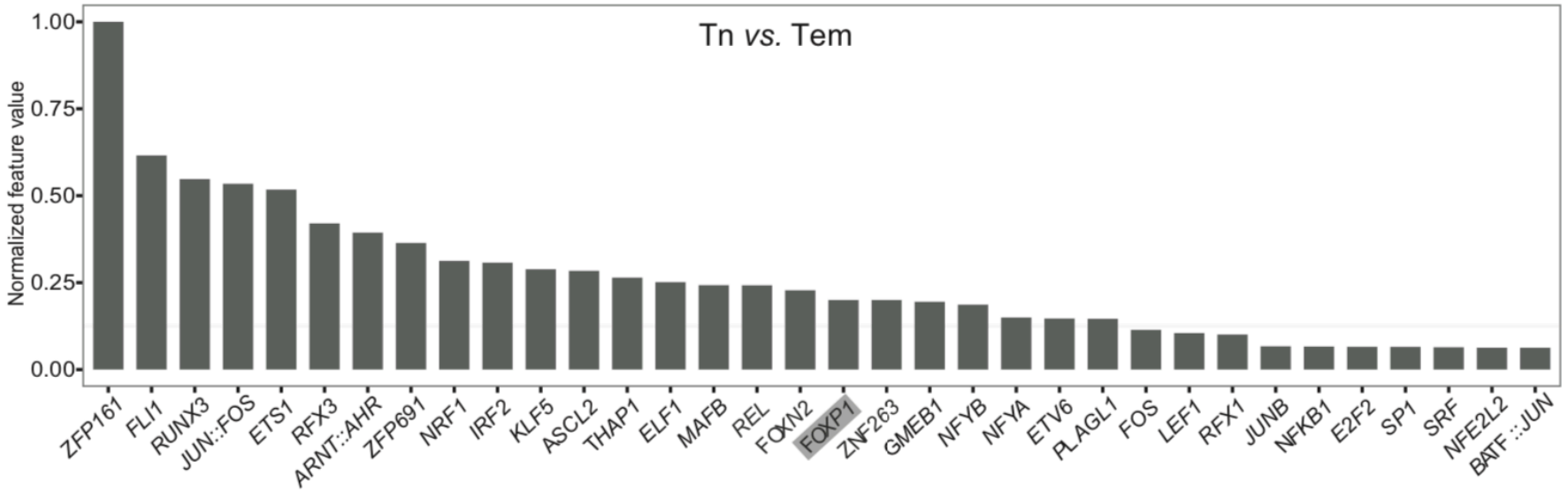
Mouse Fibroblasts
Control / Diseased



Classifier: Elastic-net regularized logistic regression

Durek et al. Cell Immunity 2016

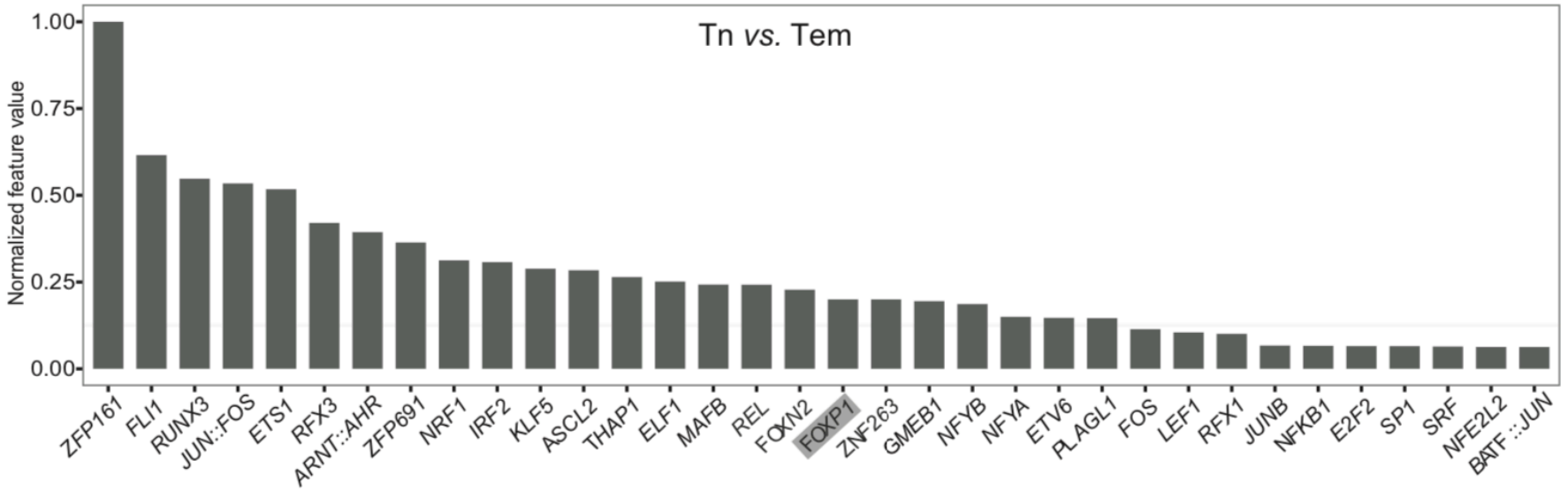
Interpretation of coefficients



Coefficients of logistic regression were normalized to the interval [0,1]

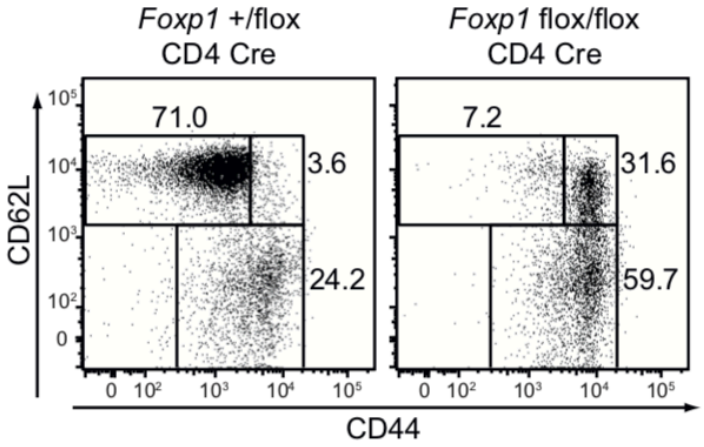
Durek et al. Cell Immunity 2016

Interpretation of coefficients



Coefficients of logistic regression were normalized to the interval [0,1]

FOXP1 was shown
to be a „gate-keeper“
of naive CD4+ T-cells



Durek et al. Cell Immunity 2016

Filtering of non-expressed TFs

- Dynamite analysis and diff. Footprinting use only motif information
- remove TFs that are not expressed in your samples, if you want to move on to experimental validation
- TF expression < 0.5 TPM are removed

Acknowledgements

Nina Baumgarten

Sivarajan Karunanithi

Dennis Hecker

Funding:

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

CARDIO PULMONARY INSTITUTE

DFG Deutsche Forschungsgemeinschaft

DZHK DEUTSCHES ZENTRUM FÜR HERZ-KREISLAUF-FORSCHUNG E.V.

Bundesministerium für Bildung und Forschung

What you will do now

Practical 2 : Use Dynamite to prioritize important TFs using the HINT footprints made before

Practical 3: Compare different methods of linking TF-footprint occurrences to genes

List of references for further reading

- **TEPIC 2-an extended framework for transcription factor binding prediction and integrative epigenomic analysis** Florian Schmidt, Fabian Kern, Peter Ebert, Nina Baumgarten, Marcel H Schulz, DOI:10.1093/bioinformatics/bty856

- **EpiRegio: analysis and retrieval of regulatory elements linked to genes,** Nina Baumgarten, Dennis Hecker, Sivarajan Karunanithi, Florian Schmidt, Markus List , Marcel H, DOI: 10.1093/nar/gkaa382

- **Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction** , Florian Schmidt et al. DOI: 10.1093/nar/gkw1061

- **Epigenomic Profiling of Human CD4 + T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development** Durek et al. DOI:10.1016/j.immuni.2016.10.022