

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



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



Marcel Schulz(MH)



Florian Schmidt (FS)

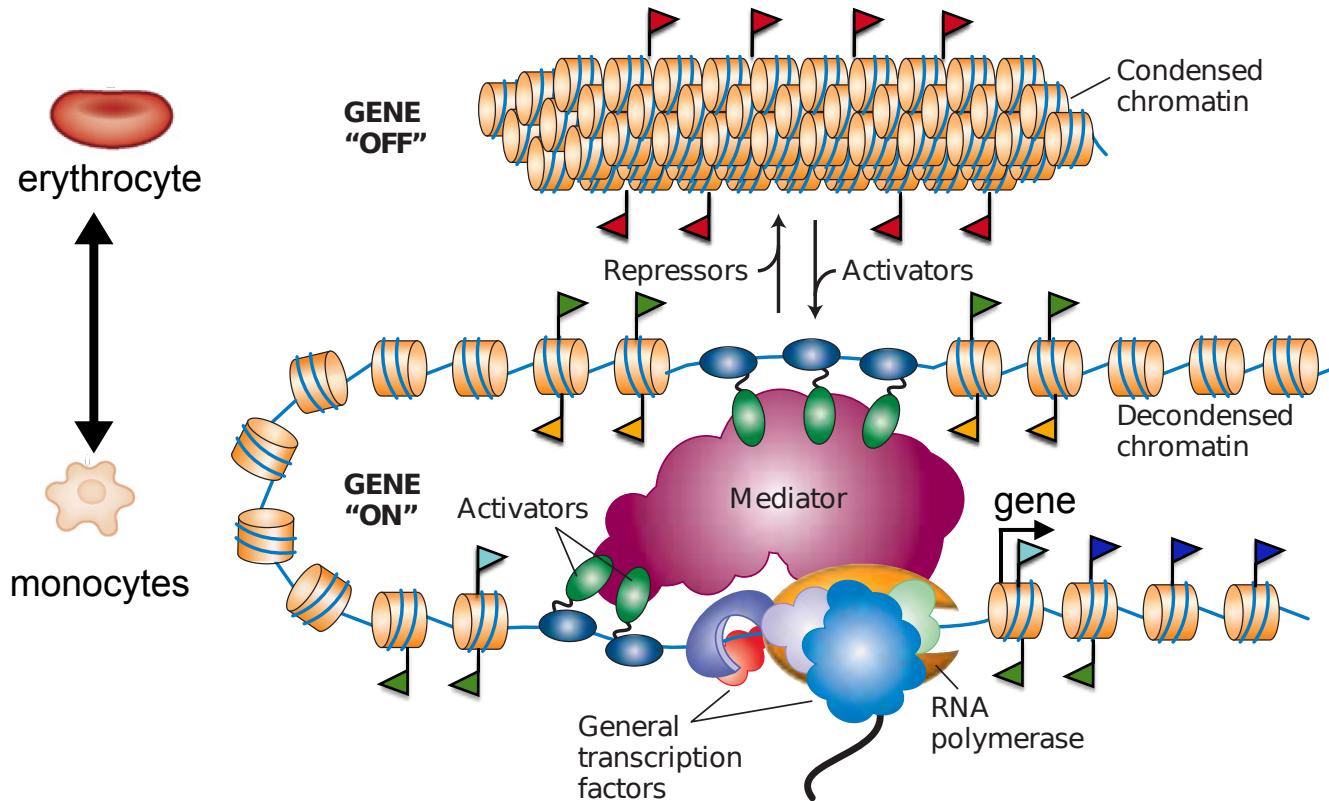
Introduction to Footprints

Ivan G. Costa
RWTH Aachen University, Germany

www.costalab.org



Chromatin and Cell Memory/Plasticity



Histone Code

► Transcription

H3K79me2, H3k36me3

► Active Regions

H3K27ac, H3K9ac

► Active Promoters

H3K4me3

► Active Enhancers

H3K4me1

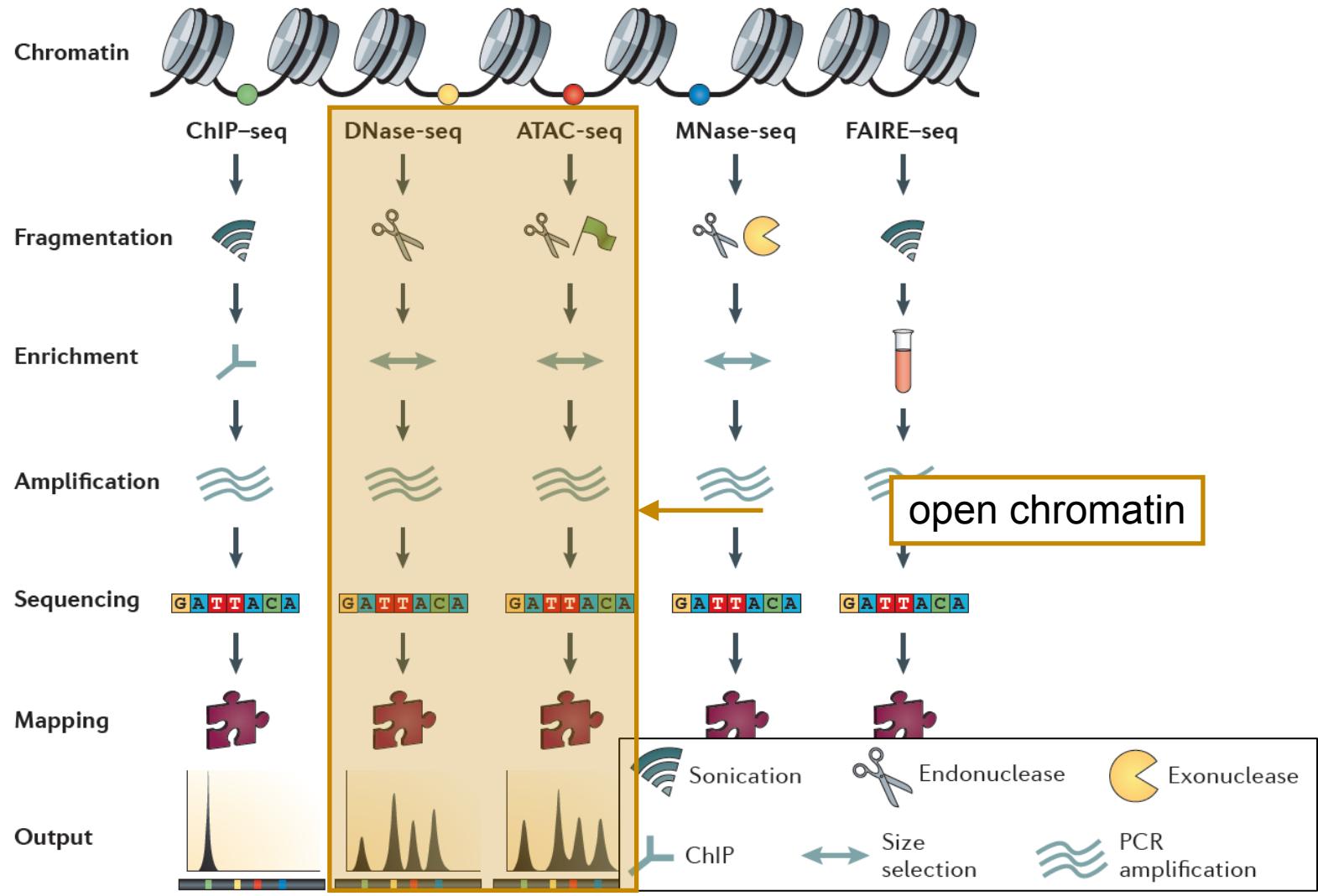
► Repressed Prom.

H3K27me3

► Repressed Regions

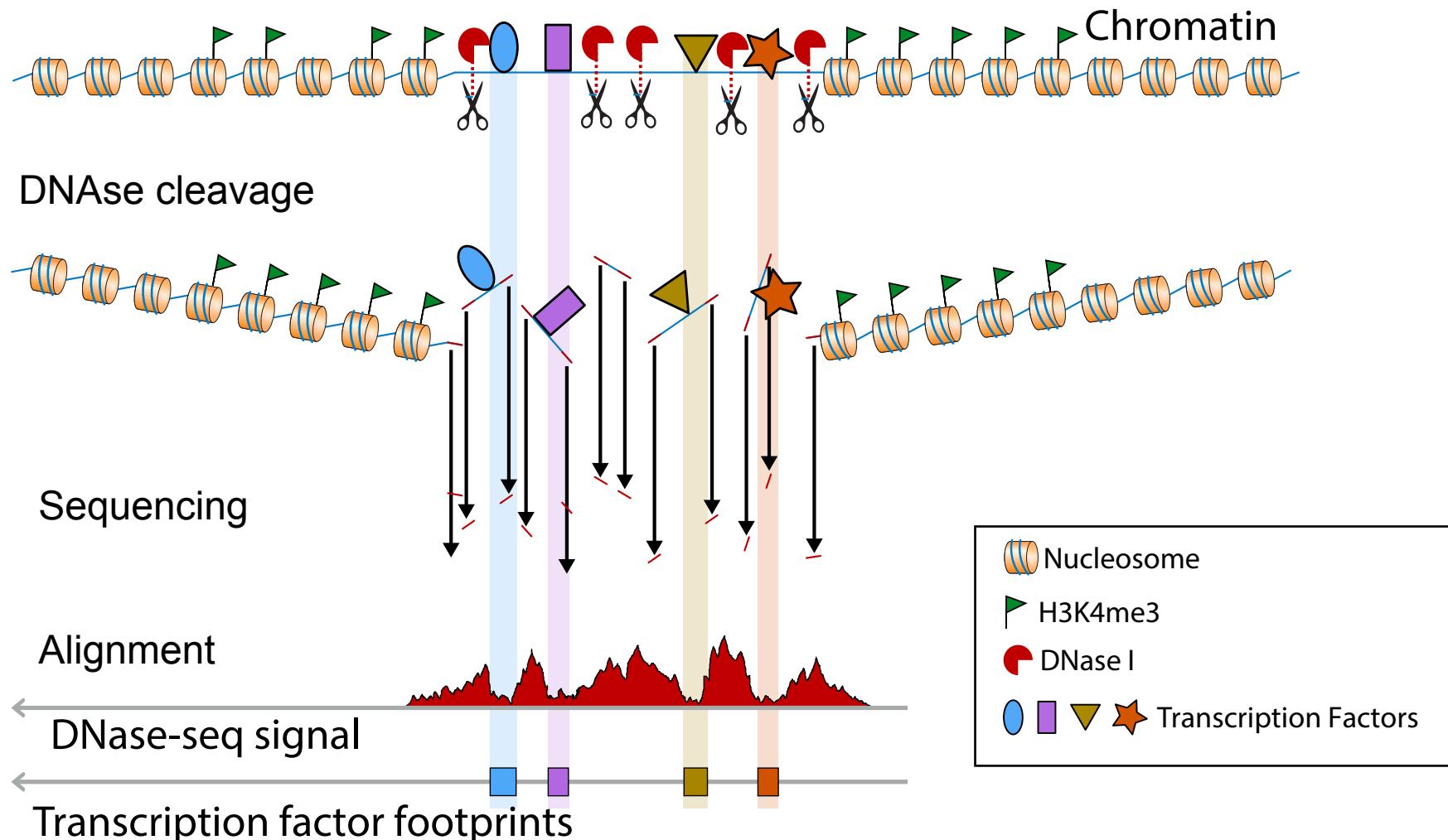
H3K9me3

NGS and Chromatin

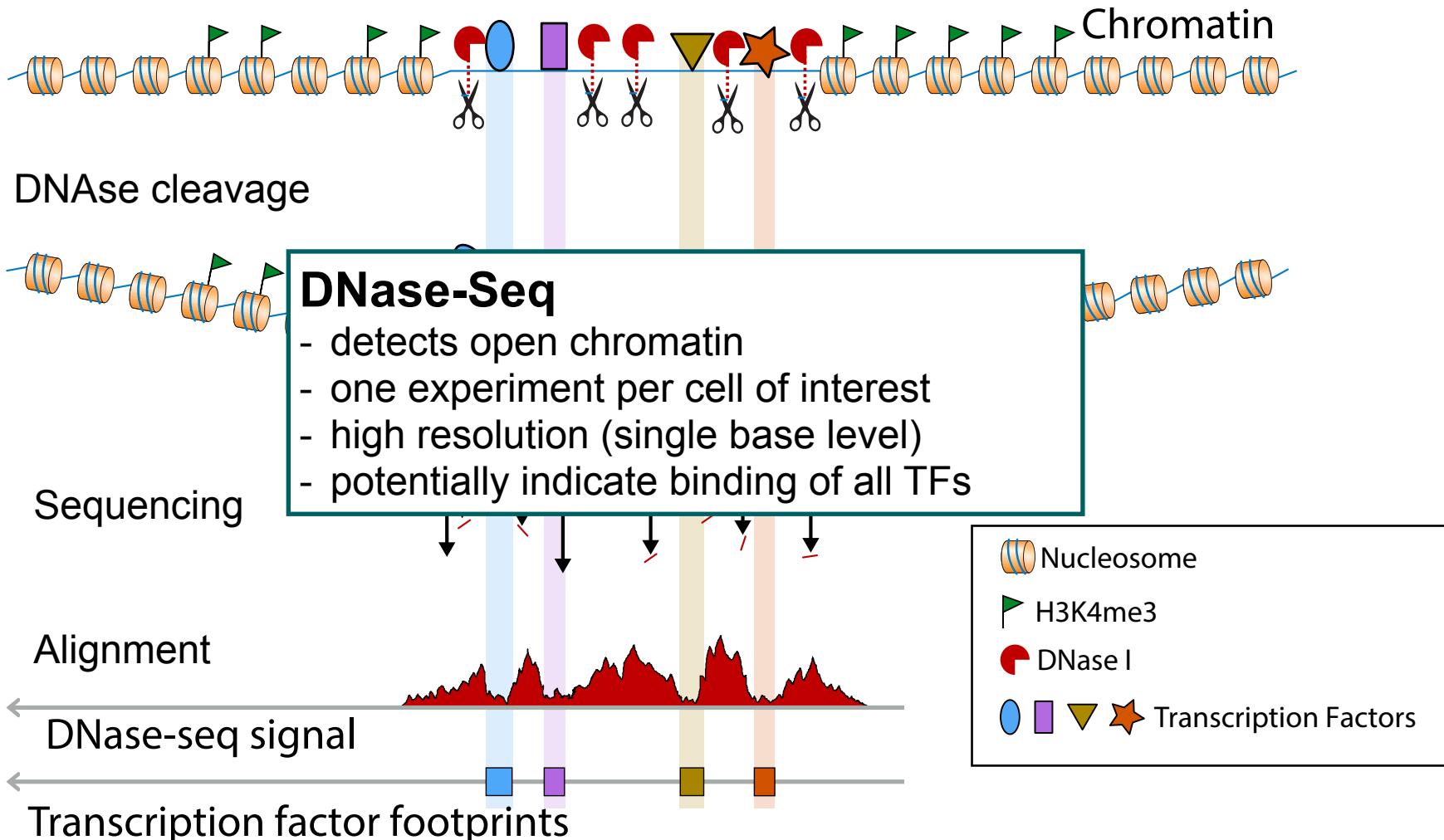


Source: Meyer, C.A. and Liu X.S. (2014). *Nature Reviews Genetics*.

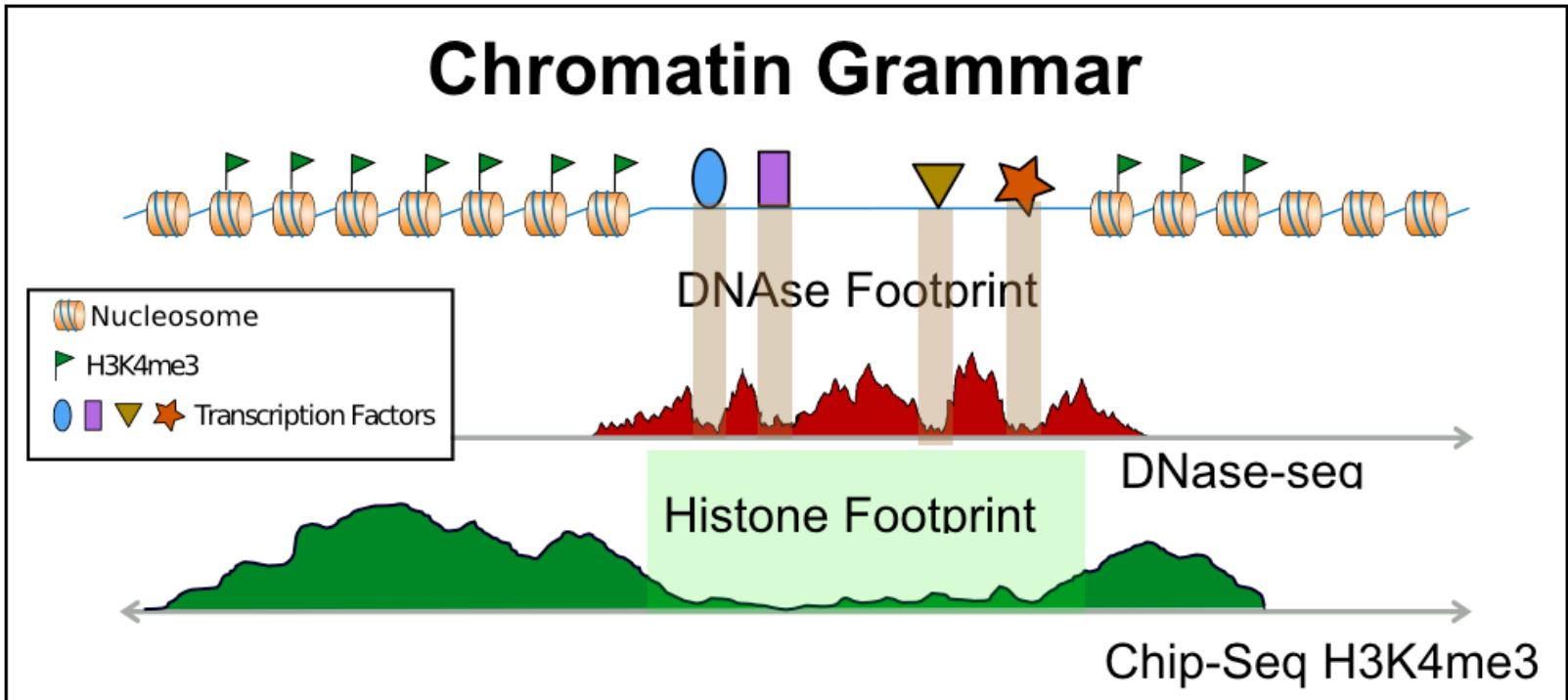
DNA - Protein interactions with DNase-seq



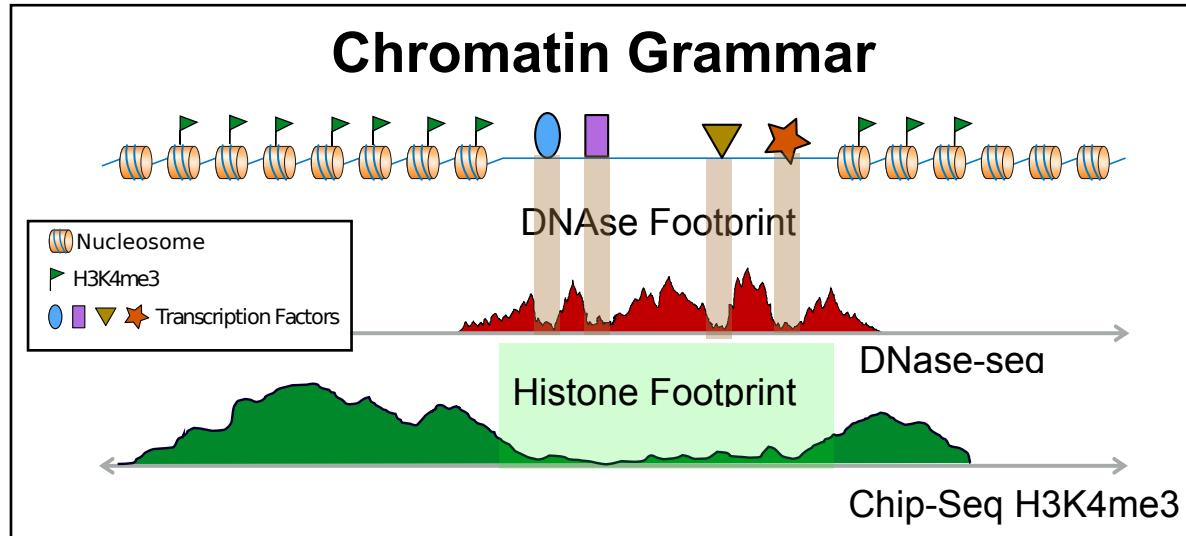
DNA - Protein interactions with DNase-seq



Detection of Active Binding Sites



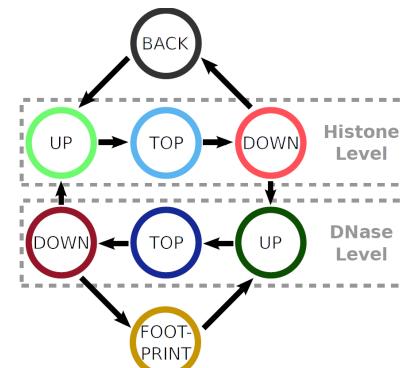
Detection of Active Binding Sites



HINT (Hmm-based IdeNtification of Transcription factor footprints)

- scan DNase and/or ChIP-Seq (activating marks) to predict footprints
- normalization for cleavage bias and global artifacts
- obtain cell independent models

8 State HMM

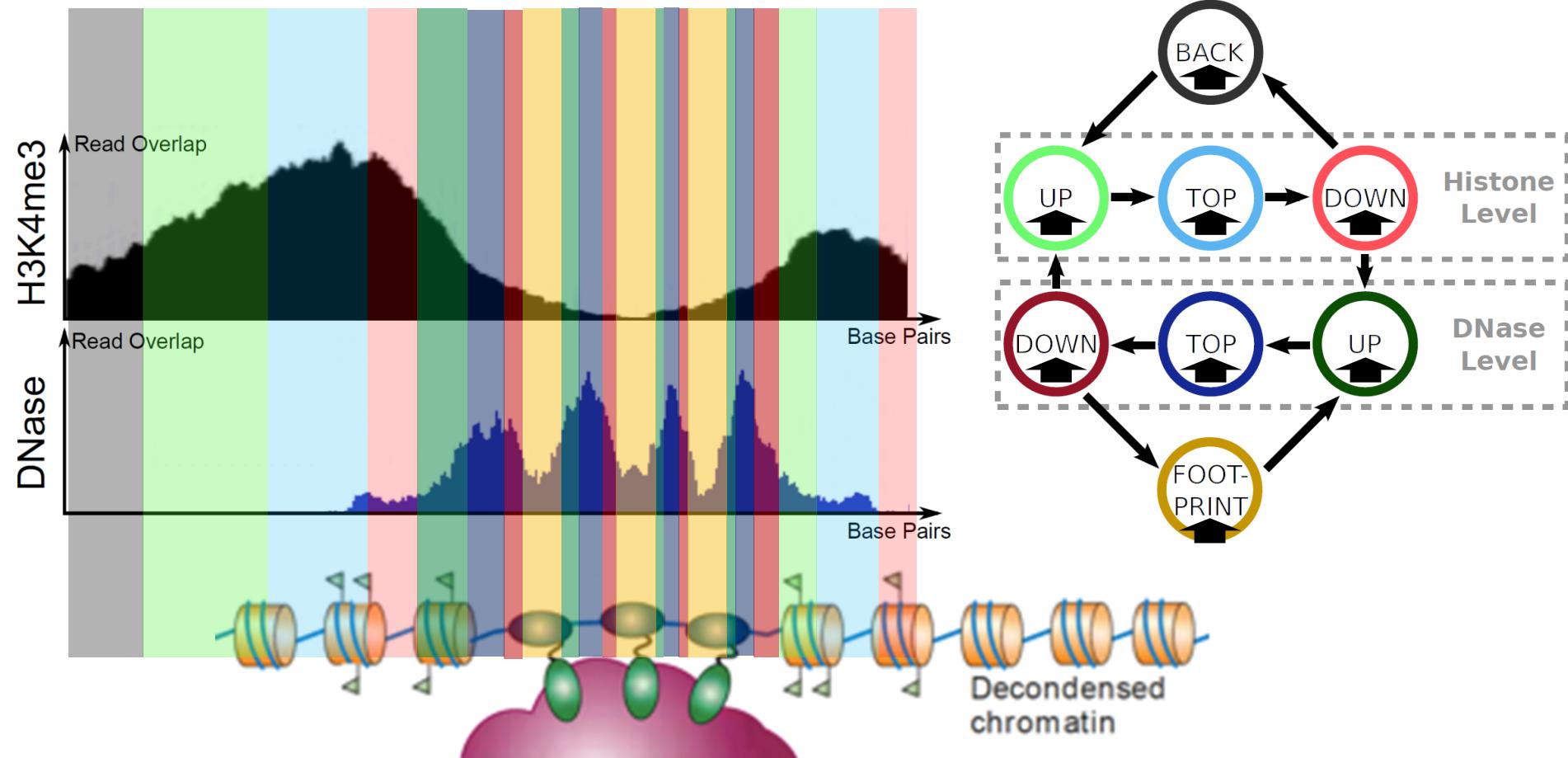


9

Gusmao EG et. al, (2014), Bioinformatics, 30(22):3143-51.

Gusmao EG et. al, (2016) Nature Methods, 13, 303–309.

Method - HINT



- Emissions - multivariate Gaussian (signal and slope of histone and DNase)
- HMM trained on manually annotated region

10

Example - K562 cells

Chr1: 10,487,866-10,491,652



APITD1-CORT

[0 - 100]

H3K4me1

[0 - 150]

H3K4me3

[0 - 200]

DNase

Footprint

Example - K562 cells

Chr1: 10,487,866-10,491,652



[0 - 100]

H3K4me1

[0 - 150]

H3K4me3

[0 - 200]

DNase

CTCF

REST

CTCF

JUN

REST

CTCF

REST

CTCF

Footprint

GCN4

MPBS



Example - K562 cells

Chr1: 10,487,866-10,491,652

APITD1-CORT

[0 - 100]

H3K4me1

[0 - 150]

H3K4me3

[0 - 200]

DNase

Footprint

CTCF

MPBS

REST

CTCF

JUN

REST

GCN4

REST

CTCF

footprint
(active)

no footprint
(inactive)

footprint
(active)

no footprint
(inactive)

Example - K562 cells

Chr1: 10,487,866-10,491,652

APITD1-CORT

[0 - 100]

H3K4me1

[0 - 150]

H3K4me3

[0 - 200]

DNase

Footprint

CTCF

MPBS

REST

CTCF

JUN

REST

GCN4

REST

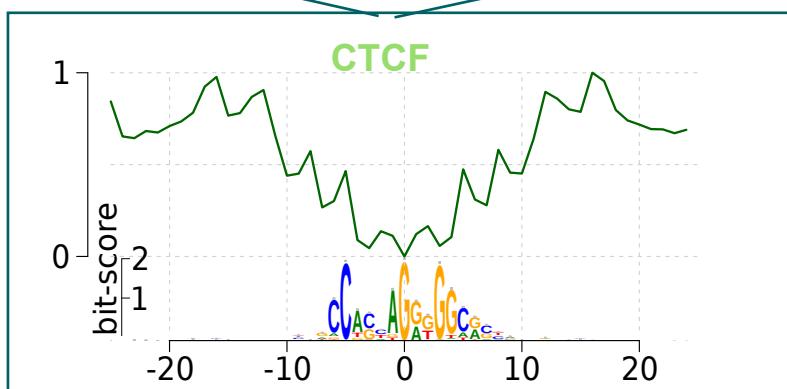
CTCF

footprint
(active)

no footprint
(inactive)

footprint
(active)

no footprint
(inactive)



DNase-seq Artifacts - Cleavage Bias

- DNase I prefers to bind (and cleave) some DNA regions ...

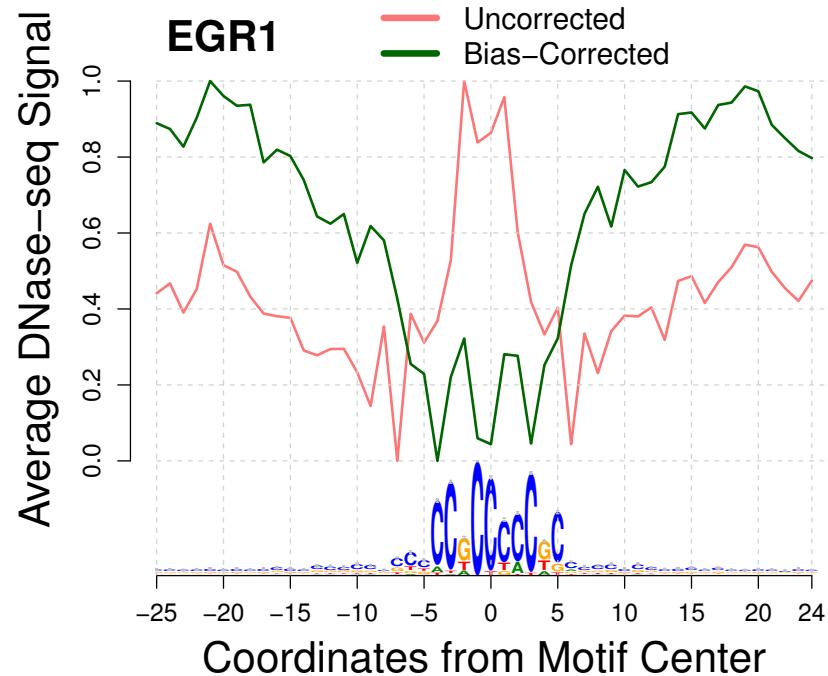


- ... than other DNA regions.



- Example*: We observe a ~3.5 higher frequency of reads starting in **ACCGGG** than the frequency of **ACCGGG** in the genome .
- For a given position i around a k -mer with x_i reads

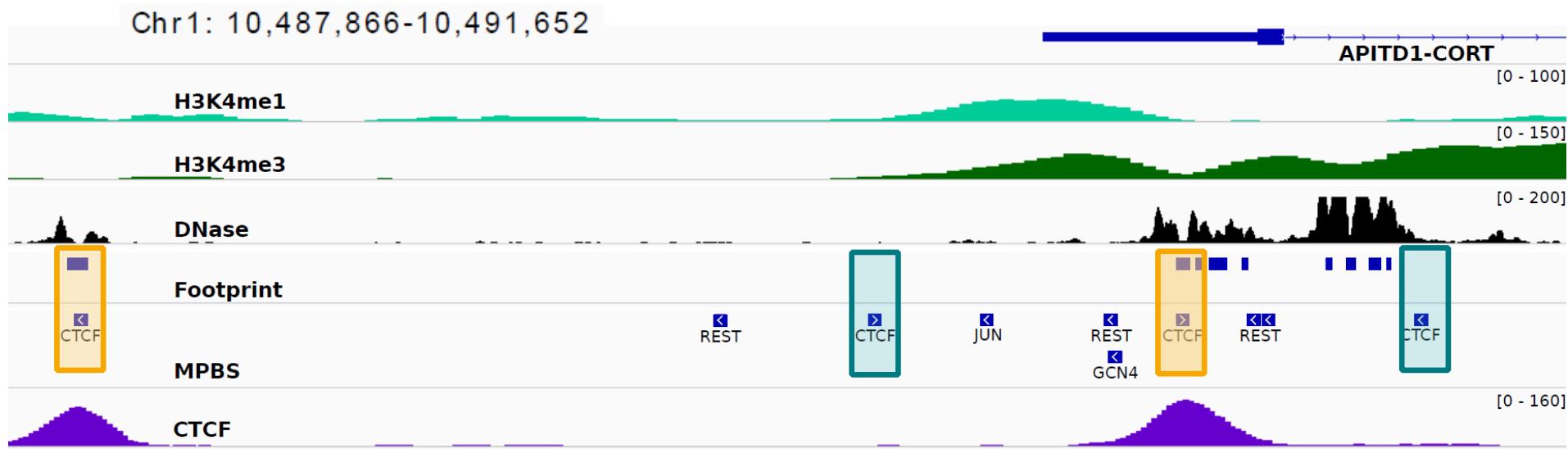
$$\text{corrected } x_i = x_i / \text{bias}_{k\text{-mer}}$$



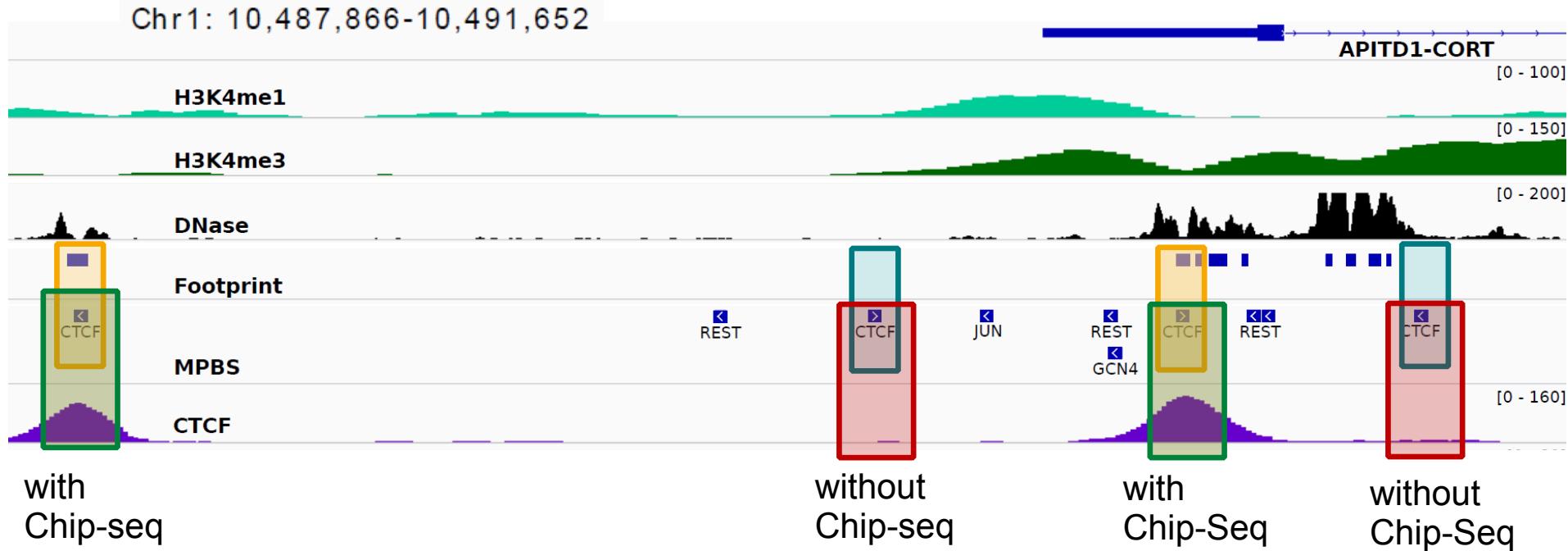
*All examples are for cell type K562
(Single-hit prot. – Crawford's Lab).

He, H.H. et al. (2014). *Nature Methods* 11(1):73-8.

Gold Standard - TF ChIP-Seq and motif search



Gold Standard - TF ChIP-Seq and motif search



ChIP-Seq	Footprint	Outcome	
with	Active	True Positive	Sensitivity
with	Inactive	False Negative	ROC AUC
without	Active	False Positive	AUPR
without	Inactive	True Negative	Specificity

Evaluation on 88 Transcription factors H1-ESC and K562



- Baseline methods

- PWM - sequence based motifs
- TC - number of DNase reads

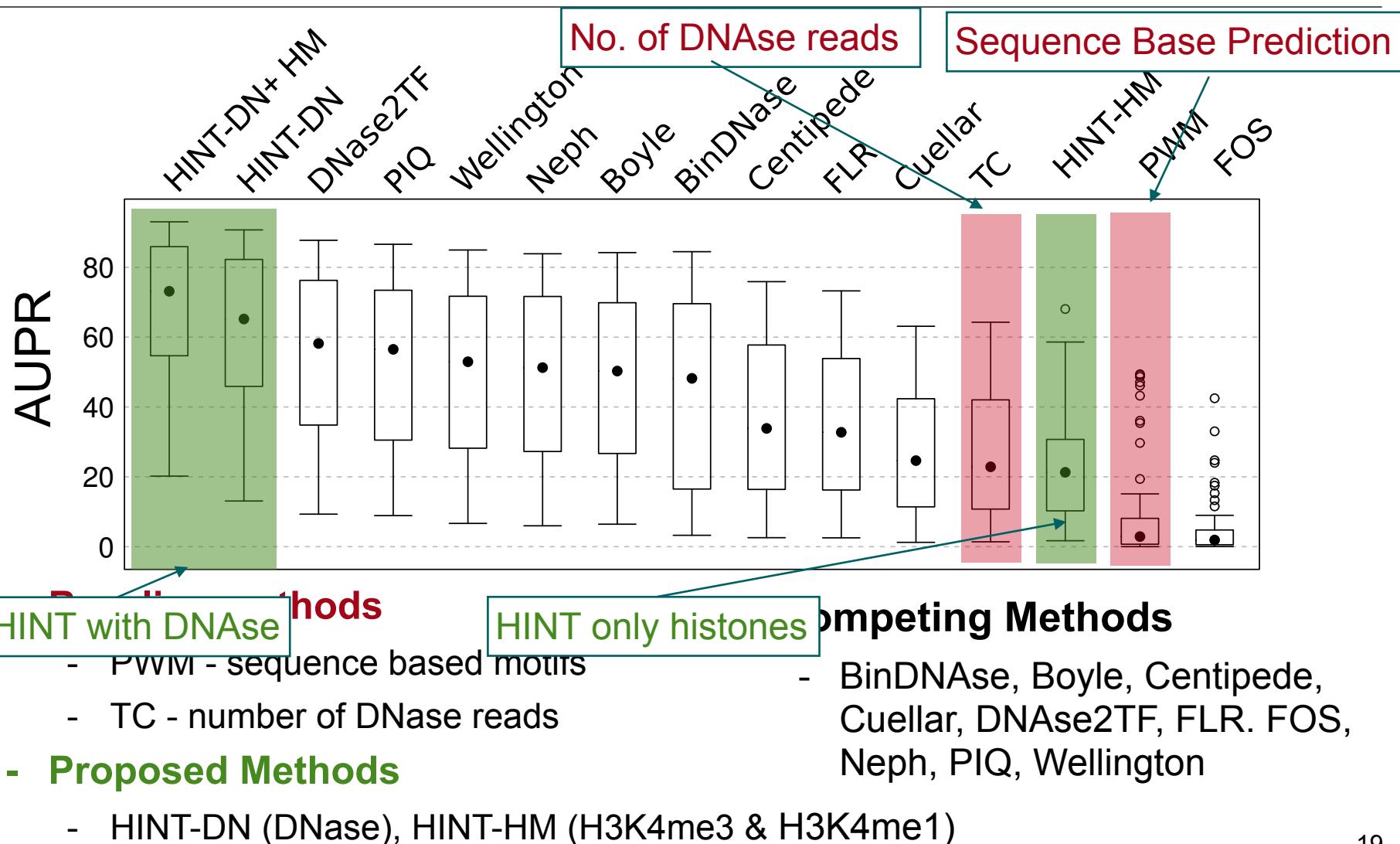
- Proposed Methods

- HINT-DN (DNase), HINT-HM (H3K4me3 & H3K4me1)

- Competing Methods

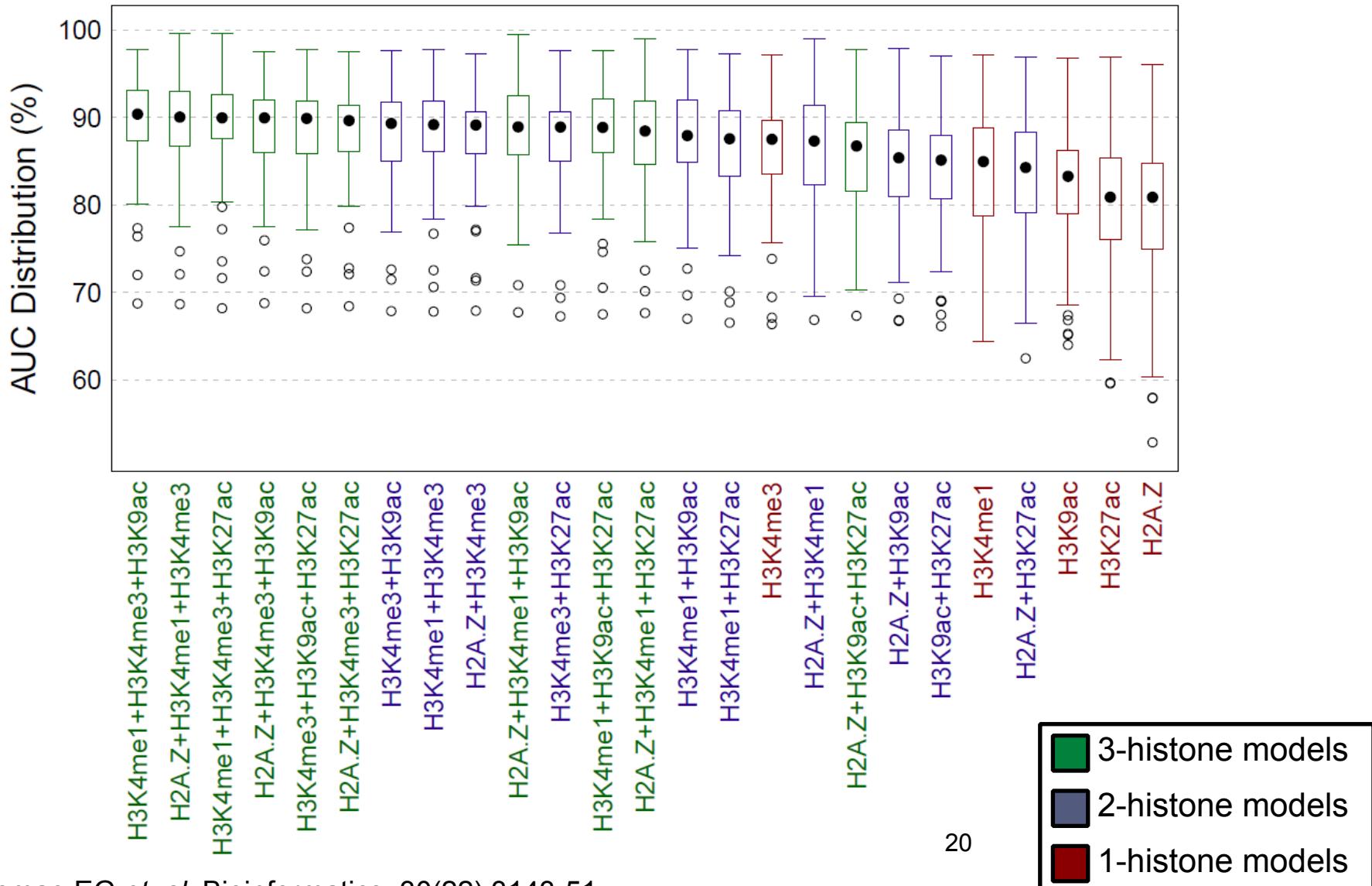
- BinDNase, Boyle, Centipede, Cuellar, DNase2TF, FLR, FOS, Neph, PIQ, Wellington

Evaluation on 88 Transcription factors H1-ESC and K562



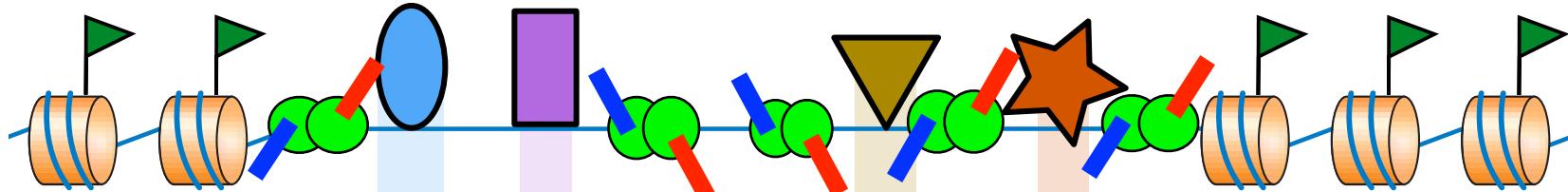
Evaluations - Which histone modifications?

Analysis on 2 cells over 83 TFs (Chr1 only)

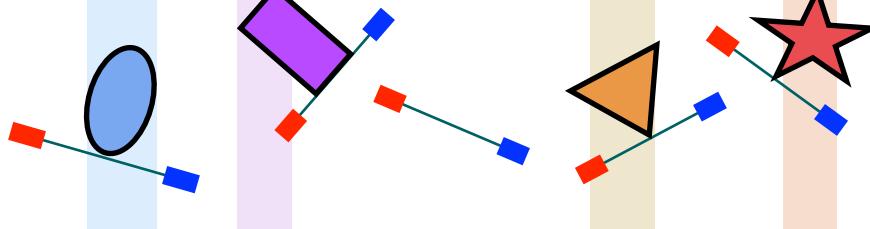


DNA - Protein interactions with ATAC-seq

Tn5 cleavage



Tn5 insertion

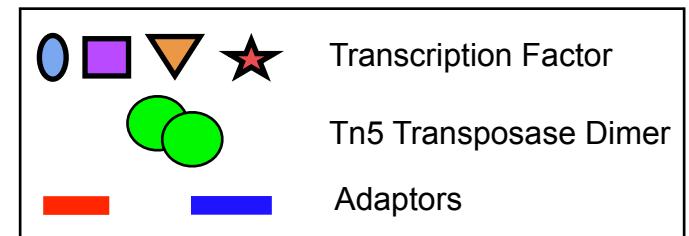


Sequencing



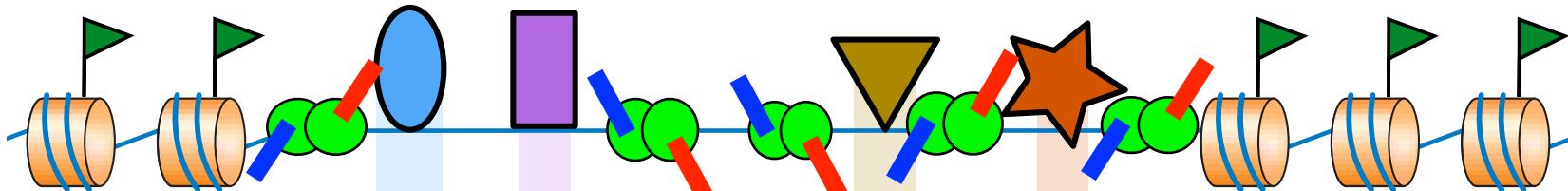
21

Alignment & Downstream Analysis

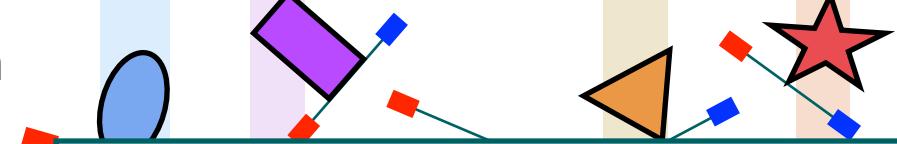


DNA - Protein interactions with ATAC-seq

Tn5 cleavage



Tn5 insertion



Sequencing

ATAC-Seq

- requires less cells / simpler protocols than DNase
- Tn5 are large size and sequence bias

22

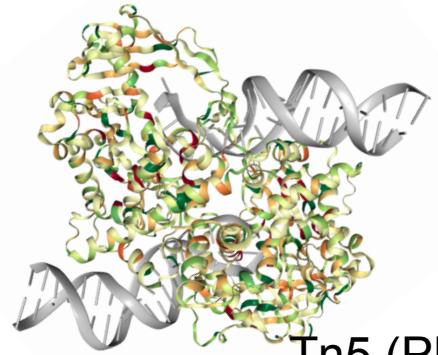
Alignment & Downstream Analysis



Open Chromatin Protocols - Comparison

Structural Features

ATAC-seq

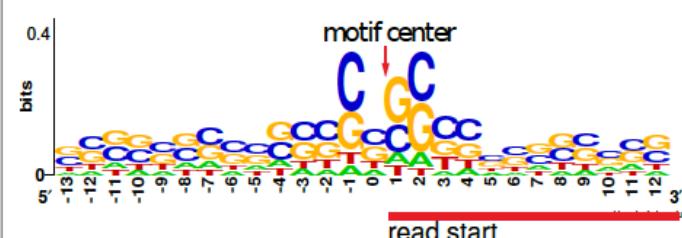
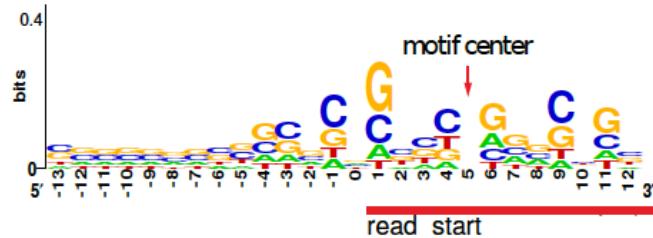


DNAse-seq



DNase-1 (PDB)

Cleavage Bias

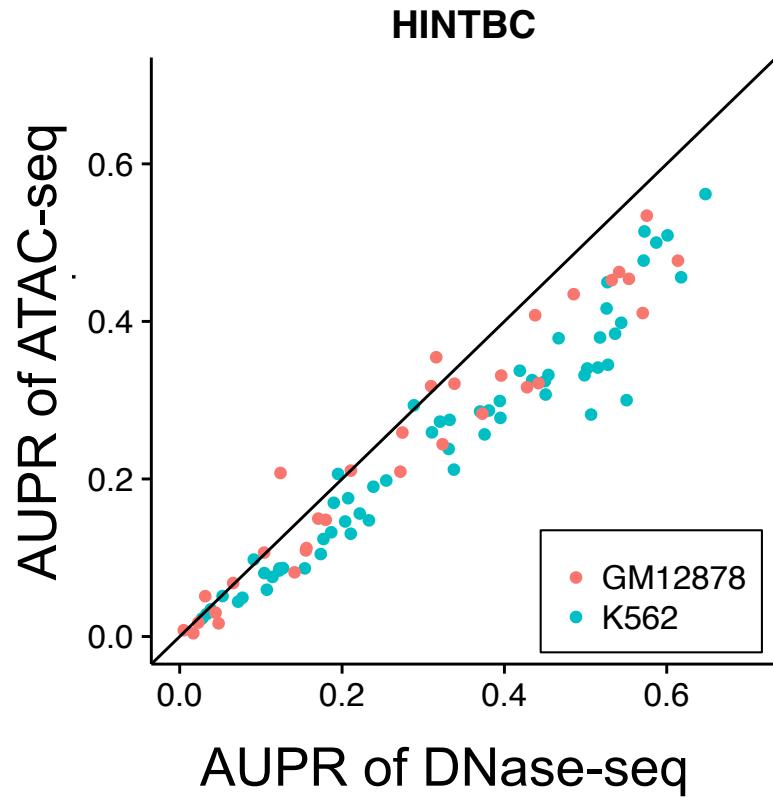


Experimental Issues

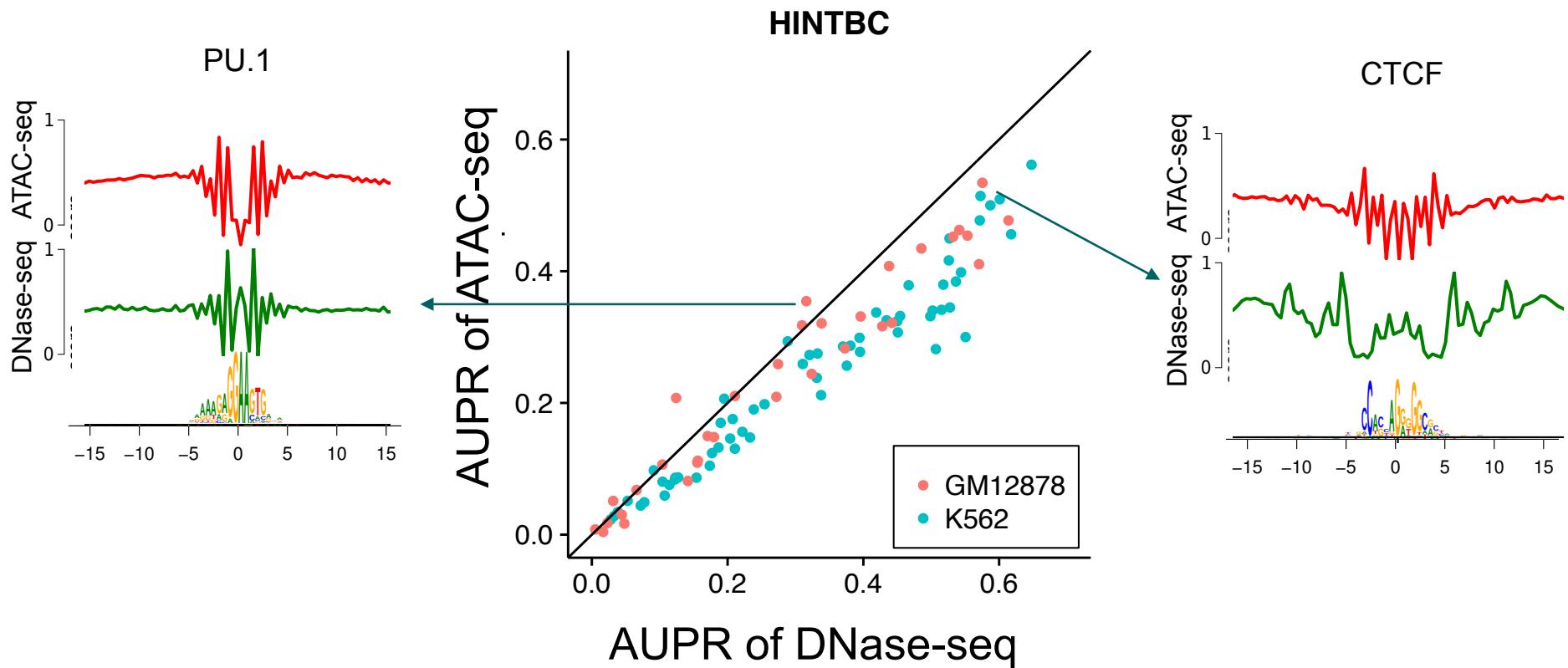
- simple/fast protocol
- requires thousands cells

- complex protocol
- requires millions cells

Open Chromatin Protocols - Comparison



Open Chromatin Protocols - Comparison



Footprint analysis

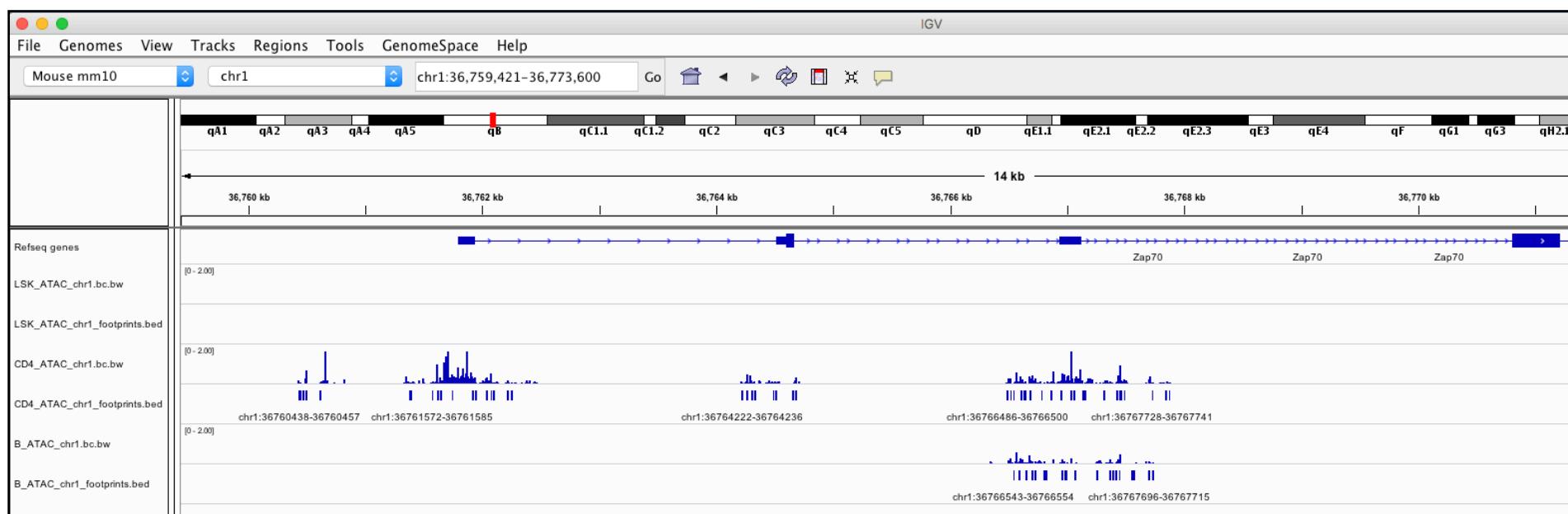
- allow detection of cell specific binding sites
- cleavage bias correction is crucial in DNase-seq

Novel chromatin protocols (ATAC-seq)

- alternative for experiments with low cell counts
- footprinting is comparable to DNase-seq
 - also requires bias correction

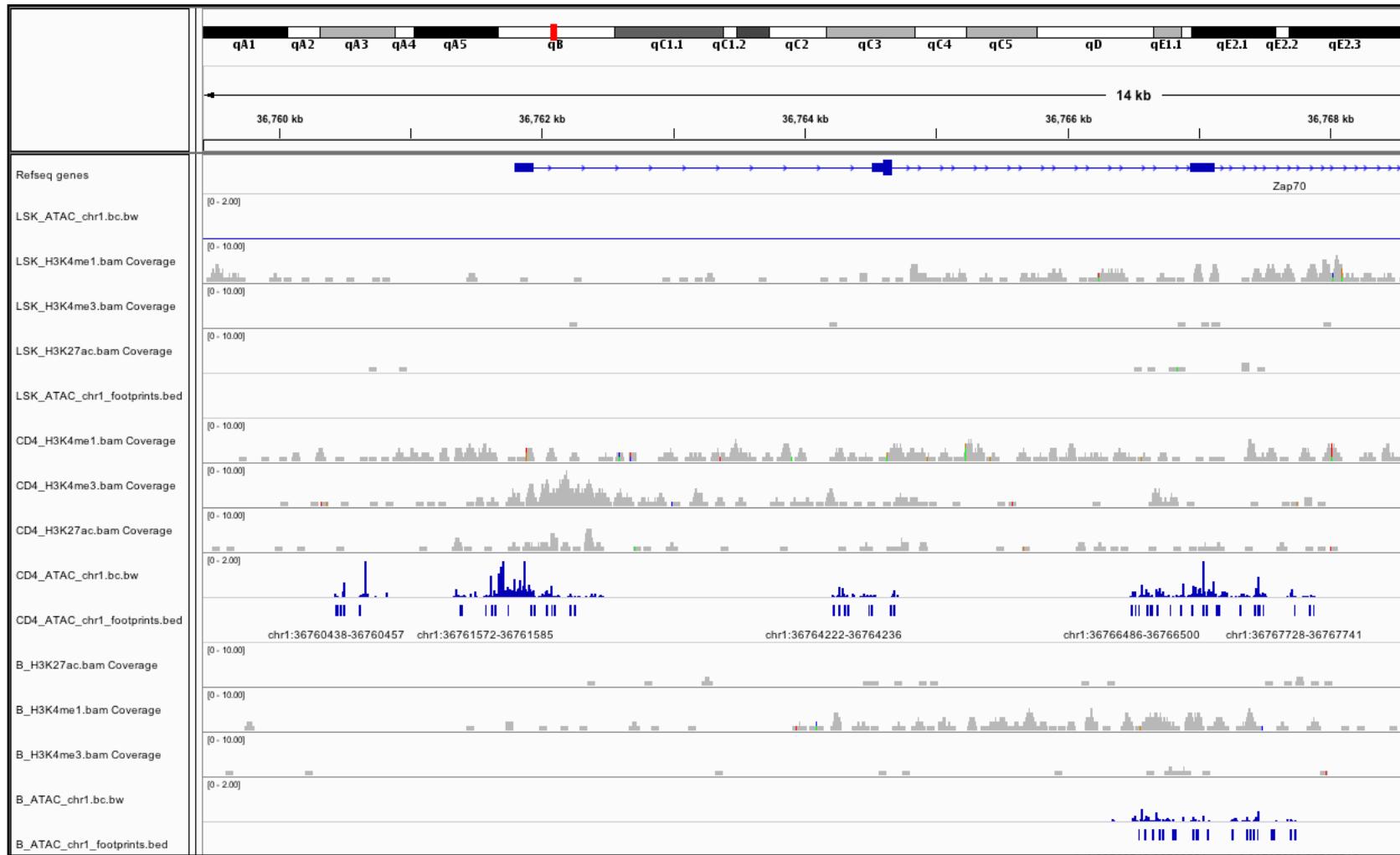
Tutorial / IGV Screenshots

ATAC-seq profiles (bigWig) and Footprints around Zap70 locus



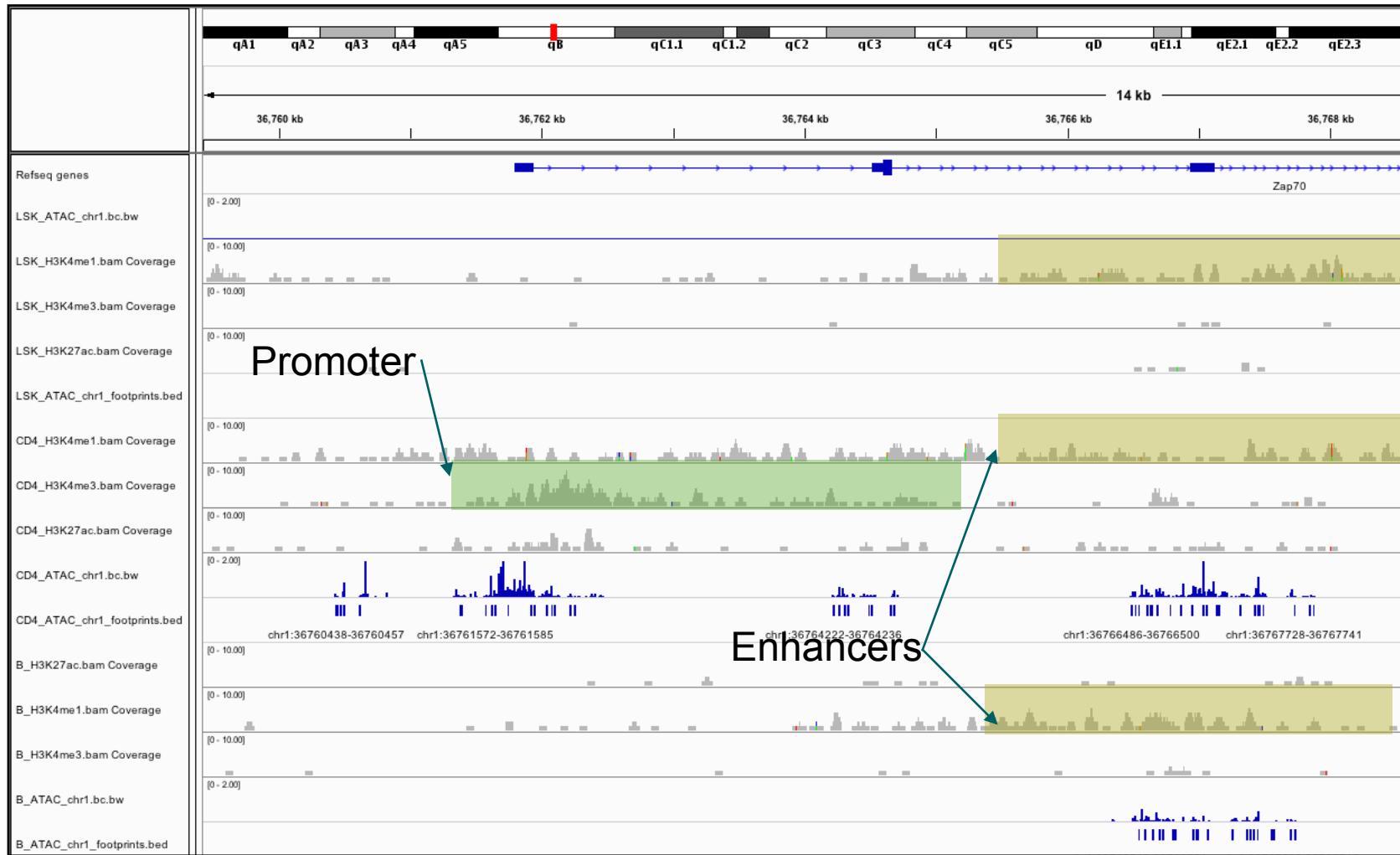
Tutorial / IGV Screenshots

ATAC-seq, Histones and Footprints around Zap70 locus



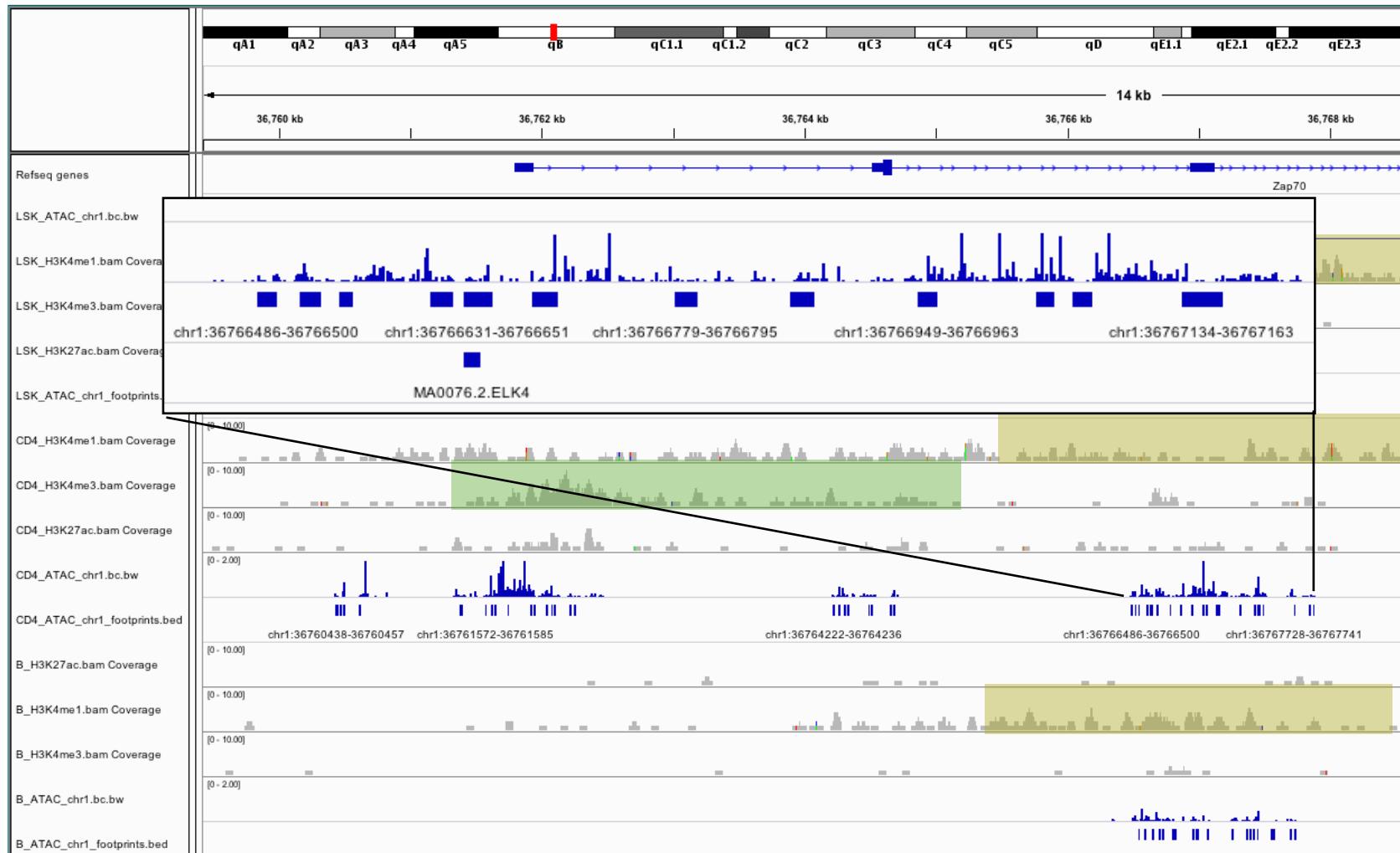
Tutorial / IGV Screenshots

ATAC-seq, Histones and Footprints around Zap70 locus



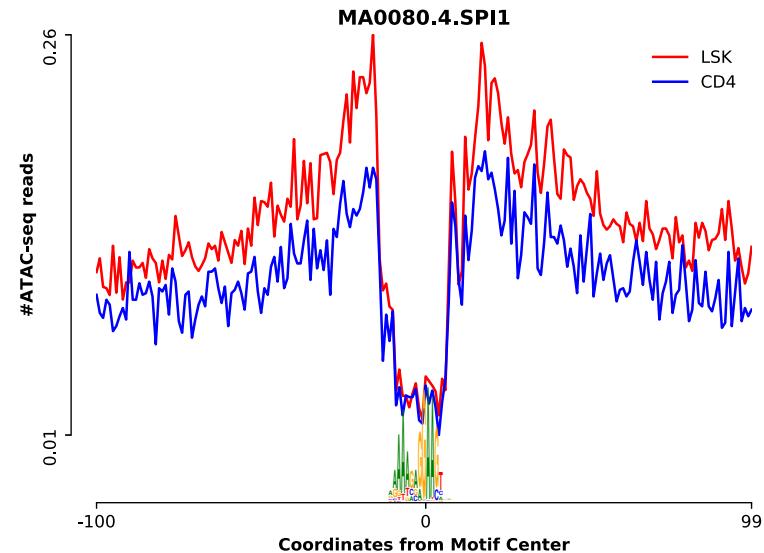
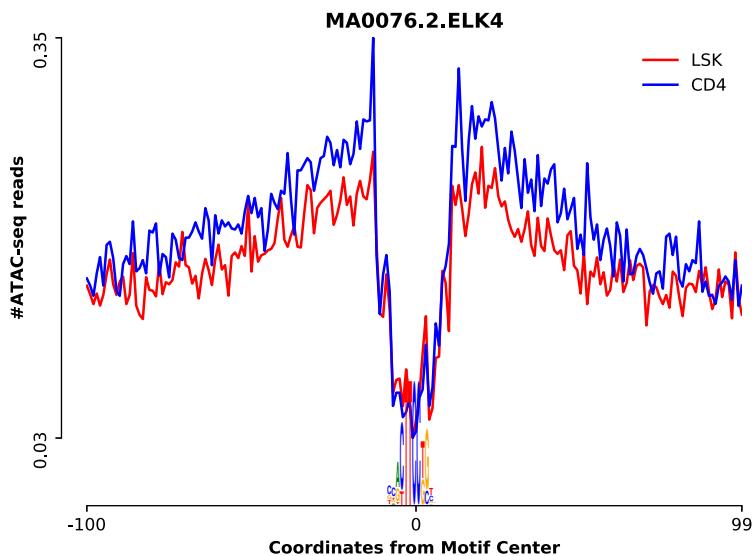
Tutorial / IGV Screenshots

Motif Matching of Pu.1 and Elk4



Tutorial / IGV Screenshots

ATAC-seq profiles around Pu.1(Spi1) and Elk4





www.costalab.org

IZKF Interdisziplinäres
Zentrum für
Klinische Forschung

RWTHAACHEN
UNIVERSITY