Methods and algorithms for epigenome analysis

Ivan G. Costa & Zhijian Li
Institute for Computational Genomics
Joint Research Centre for Computational Biomedicine
RWTH Aachen University, Germany

Institute for Computational Genomics
01011011010101
10100100101

RWTH AACHEN UNIVERSITY

Overview

1. Background and ATAC-seq protocol
2. Standard analysis of ATAC-seq data
   1. Quality check, reads alignment, peak calling
3. Footprinting Analysis
4. Description of practical

Time	Topic
13:45 - 14:15	Methods and algorithms for epigenome analysis: peak/footprint-calling, and transcription factor motif analysis (ZL)
14:15 - 15:00	Hands-on 1: Analysis of ATAC-seq data (ZL)

Team

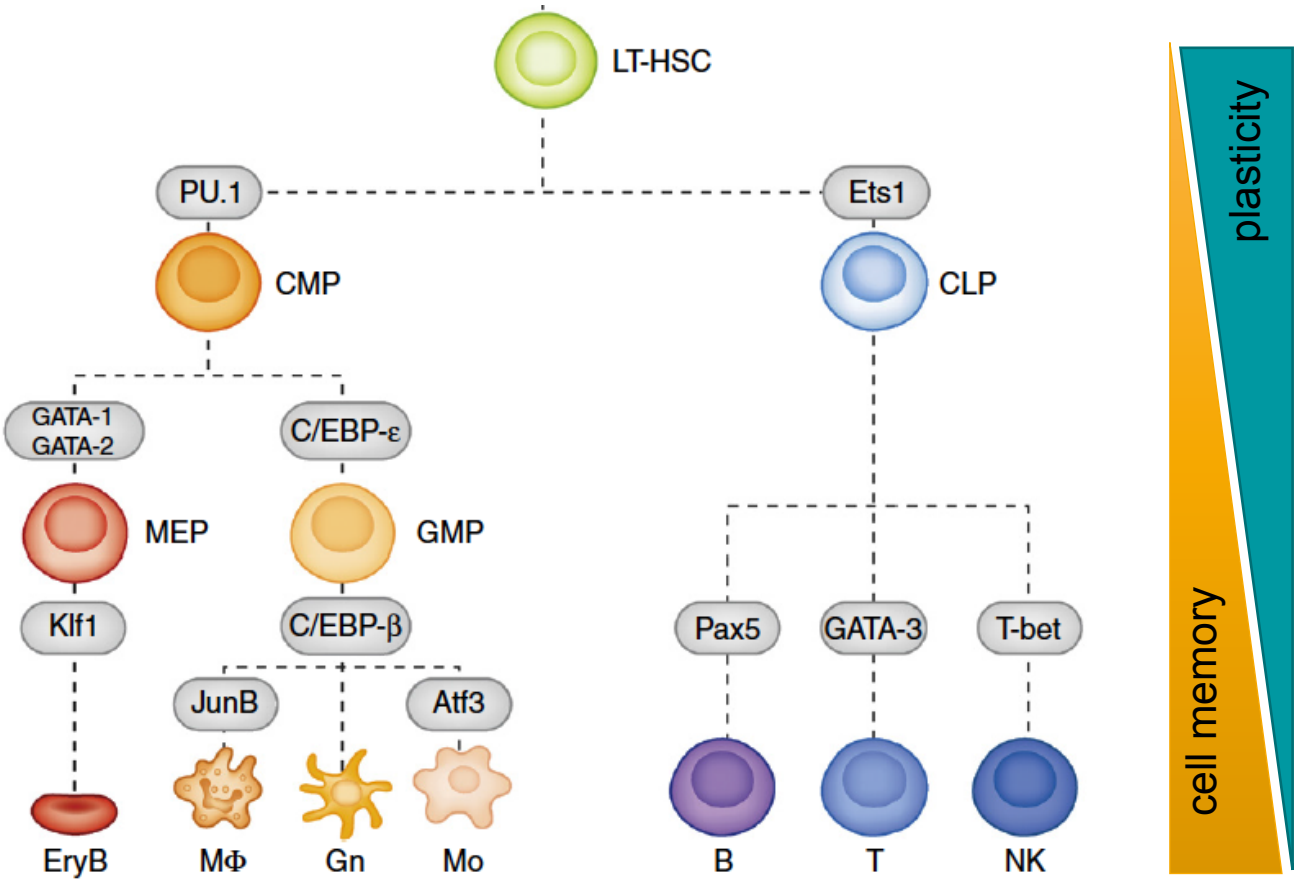
Ivan Costa (IC)

Zhijian Li (ZL)

https://github.com/SchulzLab/EpigenomeAnalysisTutorial-2020

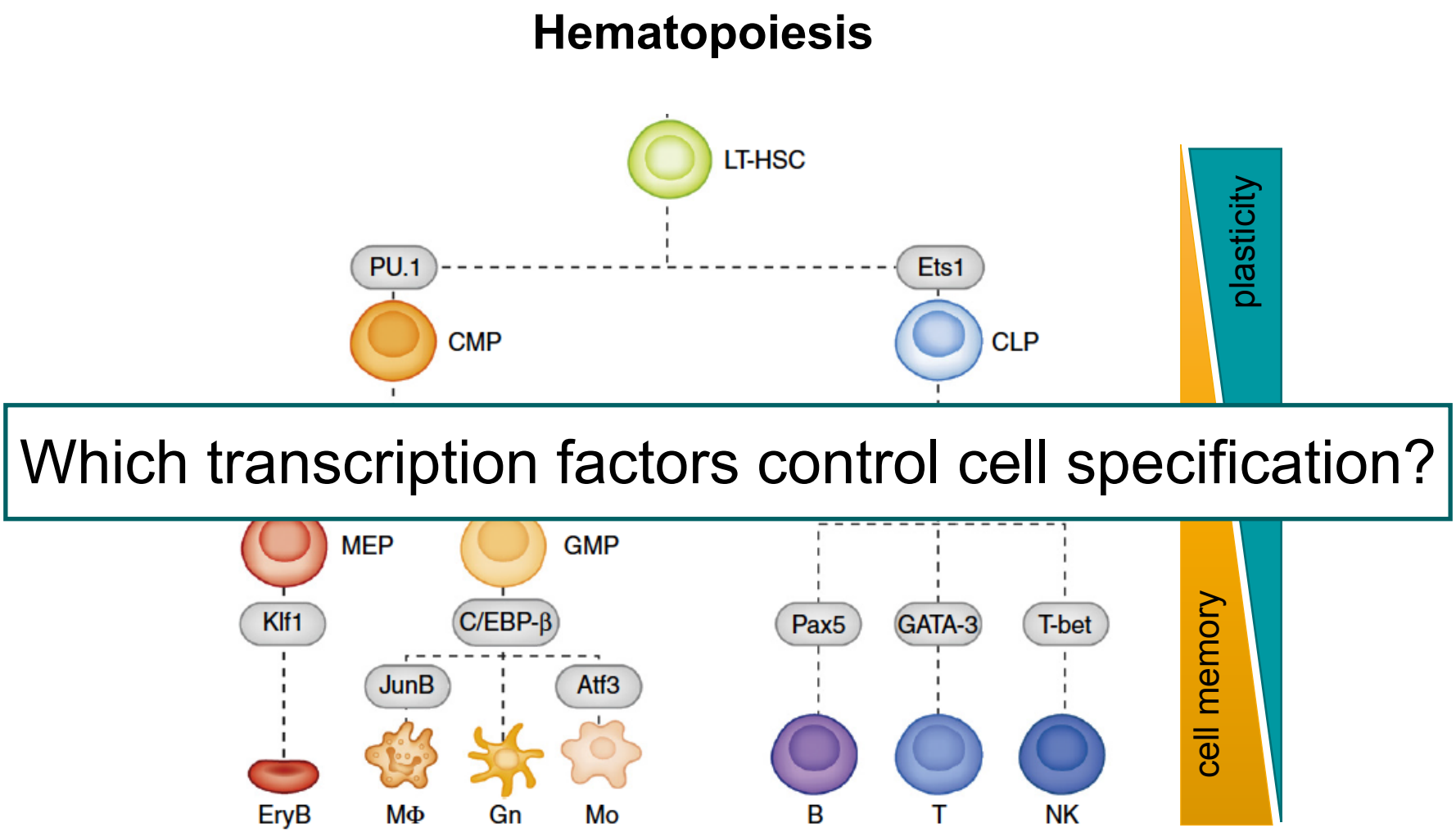
Cell differentiation

Hematopoiesis



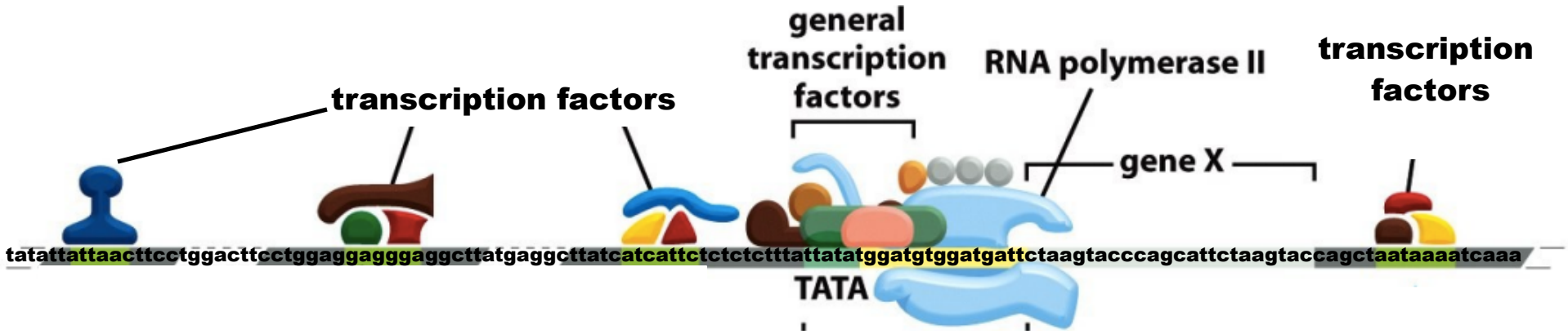
Source: Amit (2016), *Nature Immunoloy.*

Cell differentiation



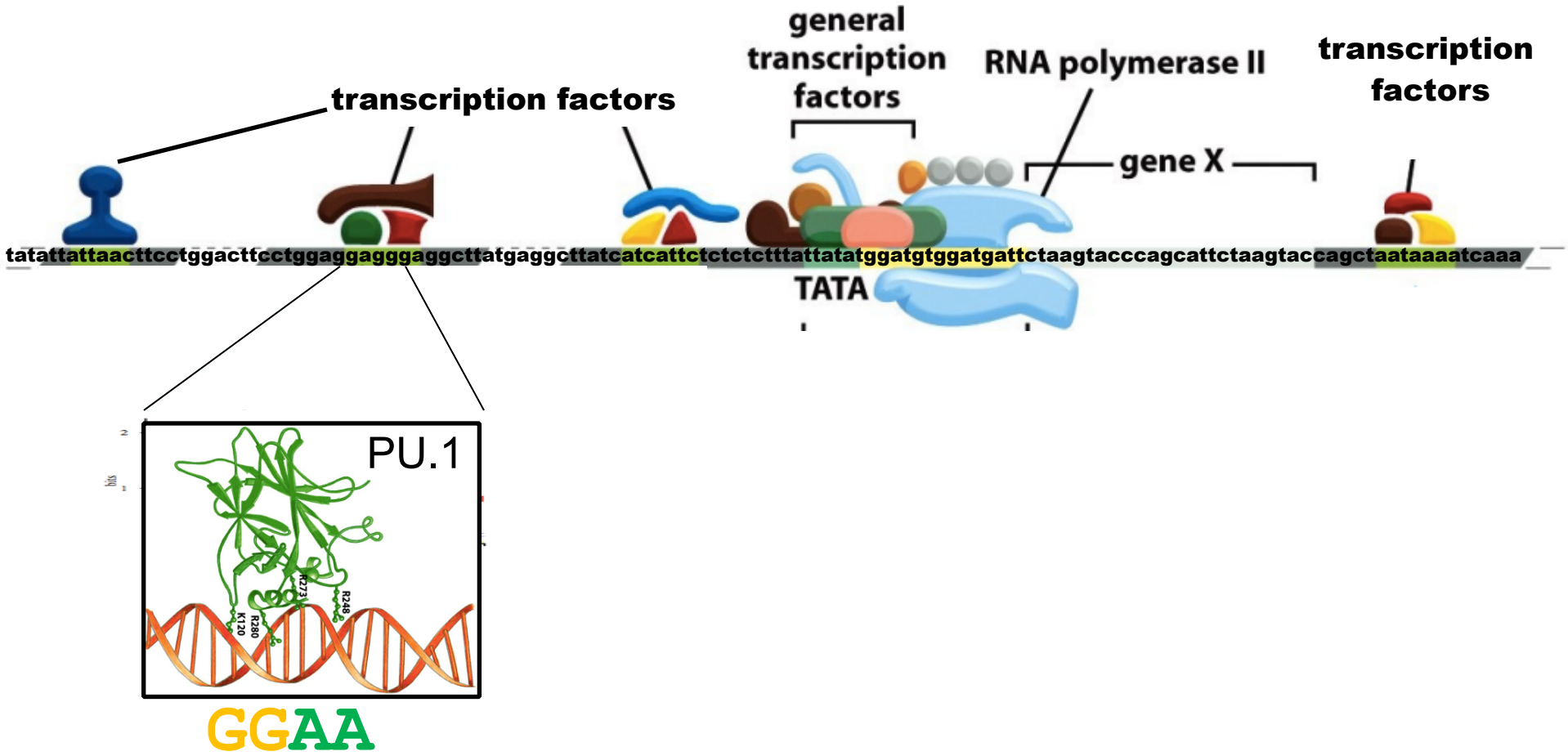
Source: Amit (2016), *Nature Immunoloy.*

Computational approach: motif searching



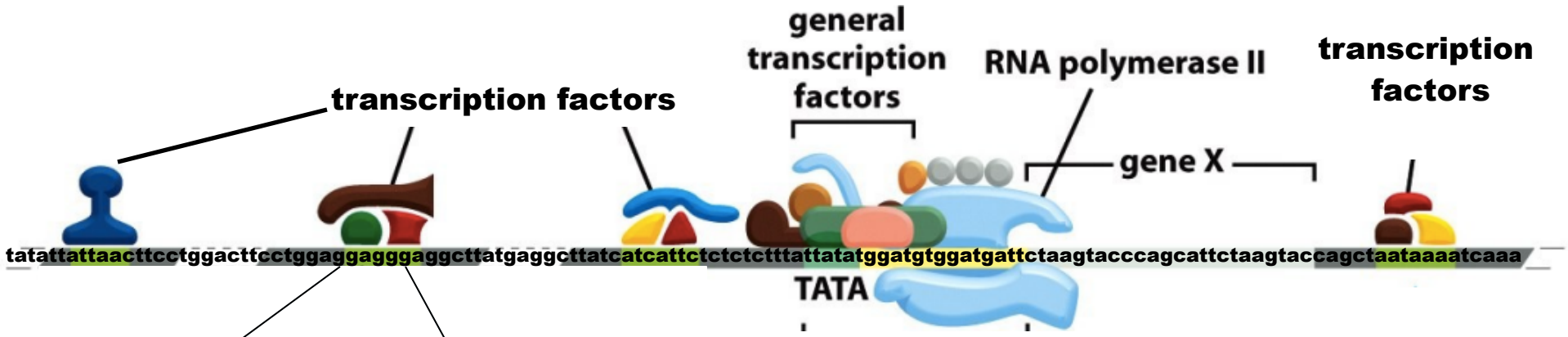
Source: Alberts, B. et al. (2008) Garland Science, 5th ed.

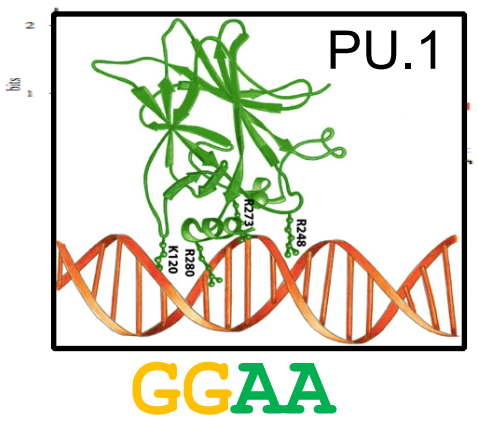
Computational approach: motif searching



Source: Alberts, B. et al. (2008) Garland Science, 5th ed.

Computational approach: motif searching





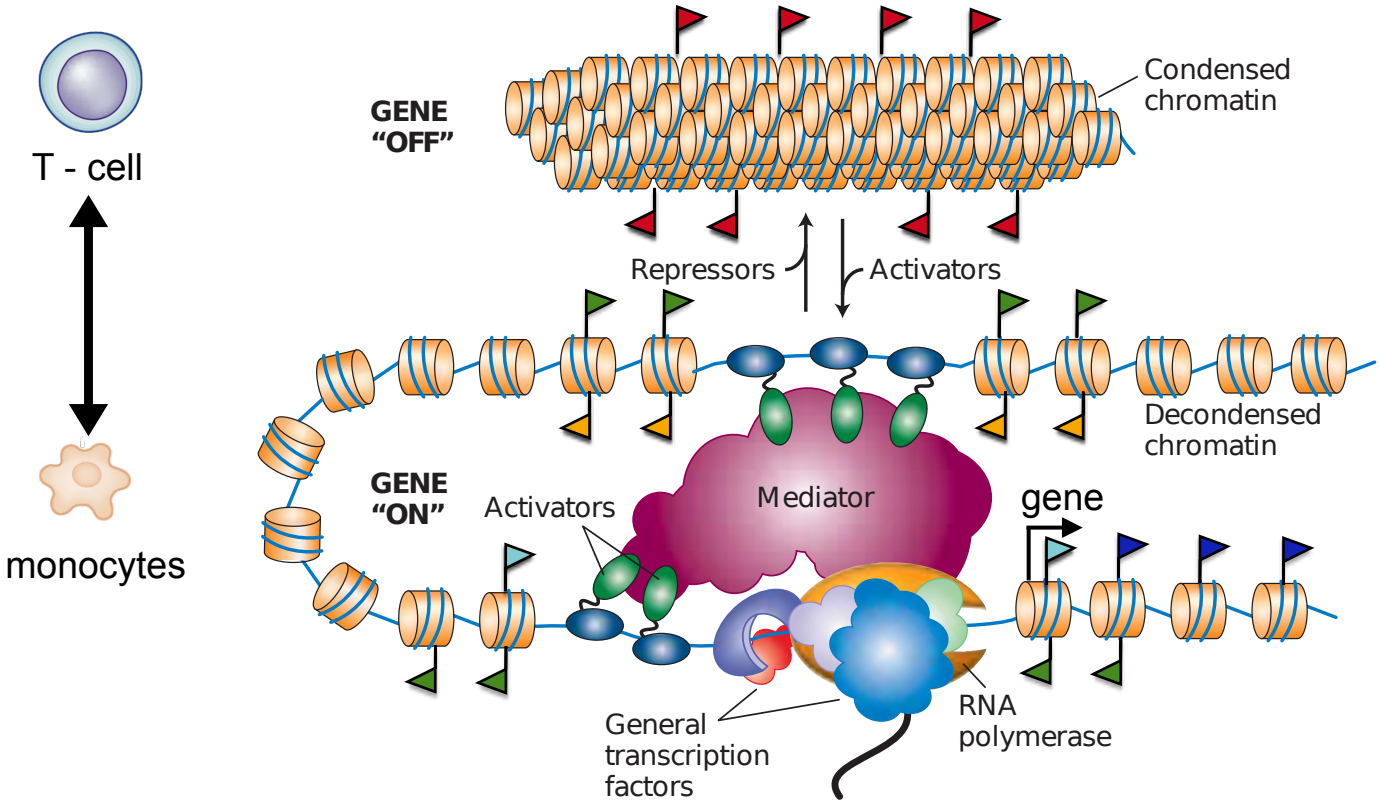
Problems

- motifs are small and degenerate (6-8 conserved positions)
- distal binding sites (>10^6 from genes)
- only part of motifs are known

Too many false positive predictions!

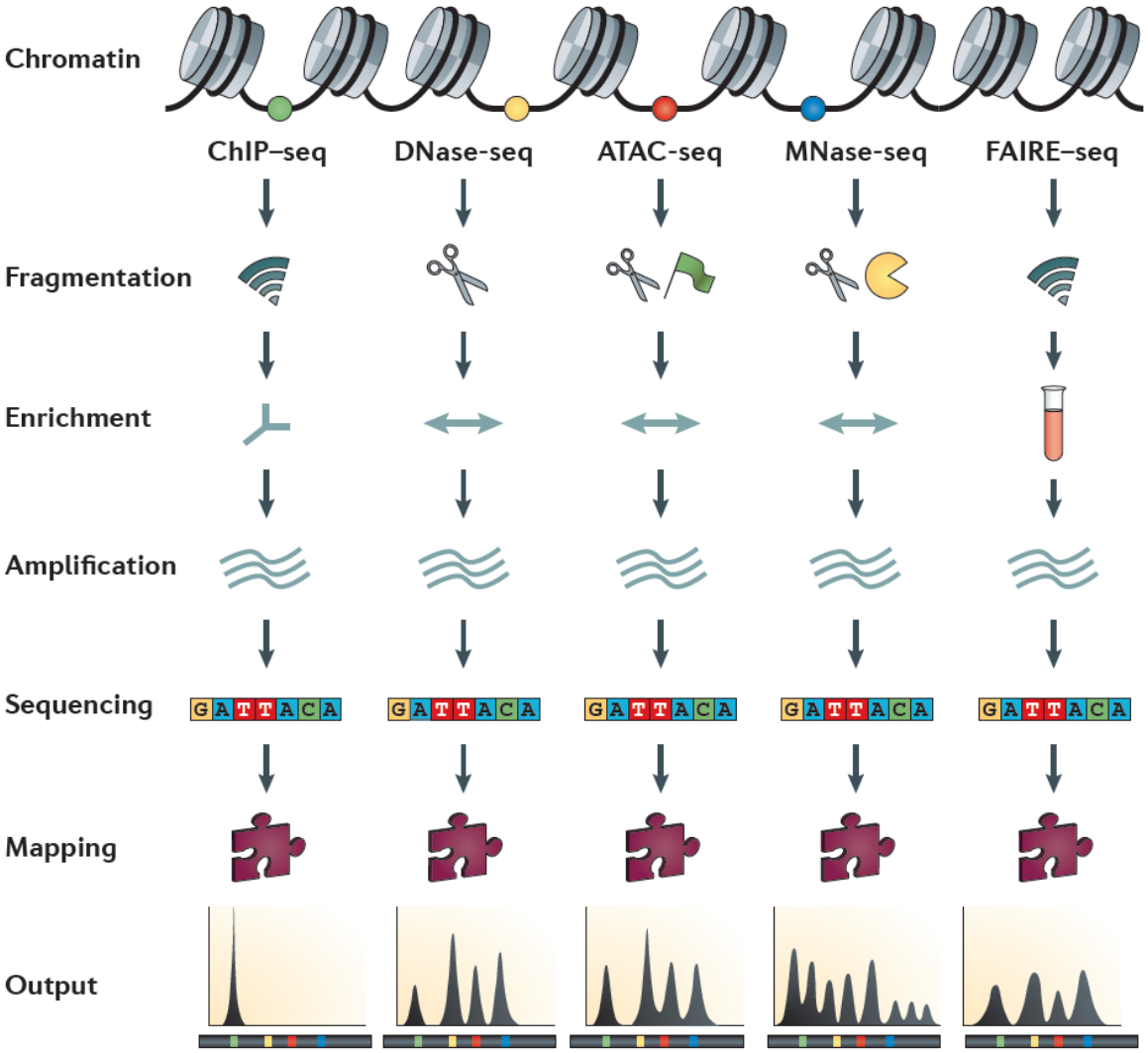
Source: Alberts, B. et al. (2008) Garland Science, 5th ed.

Open chromatin and the regulatory epigenome



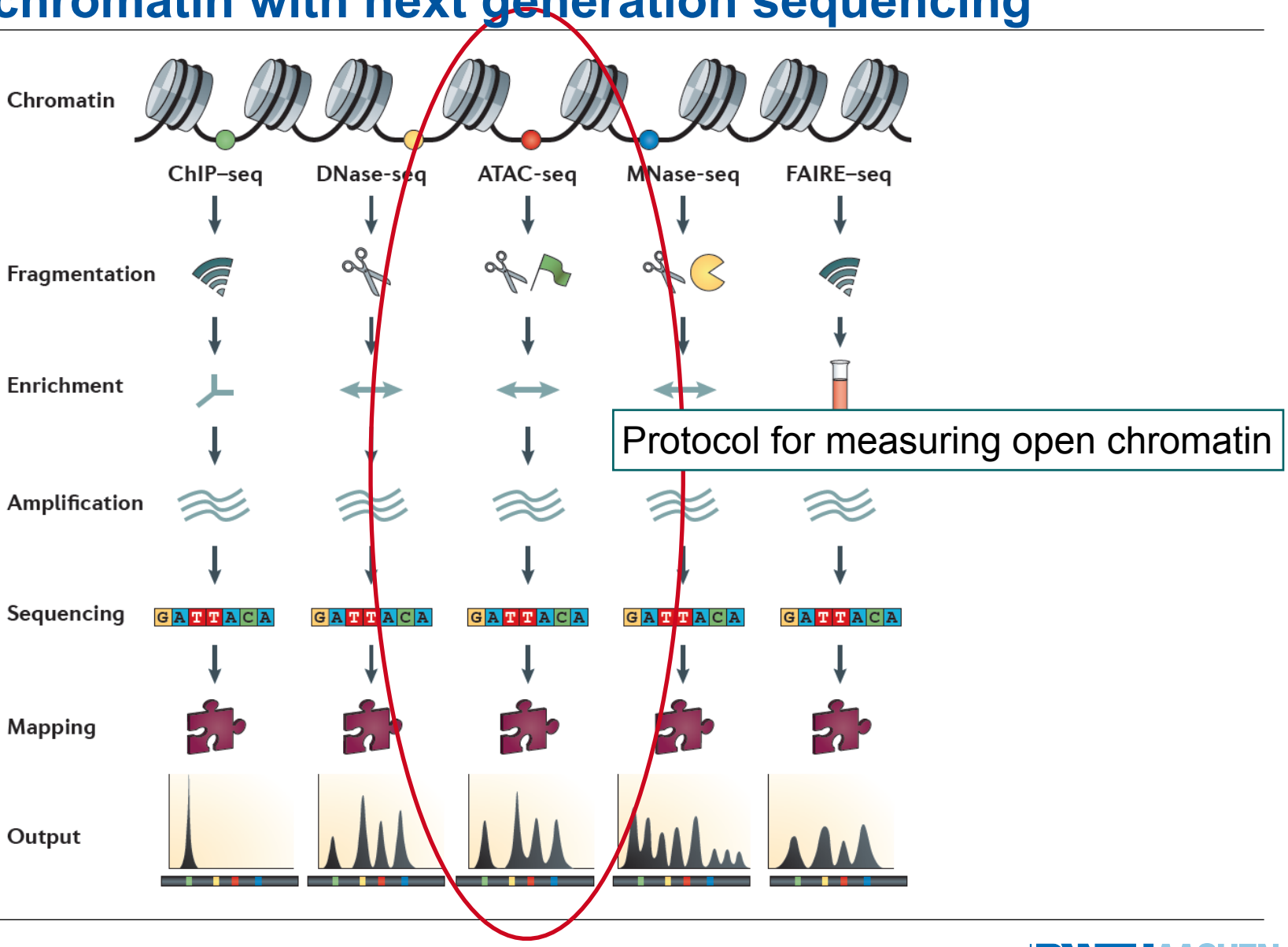
Adapted from Lodish, B. et al. (2004) 5th ed.

Open chromatin with next generation sequencing



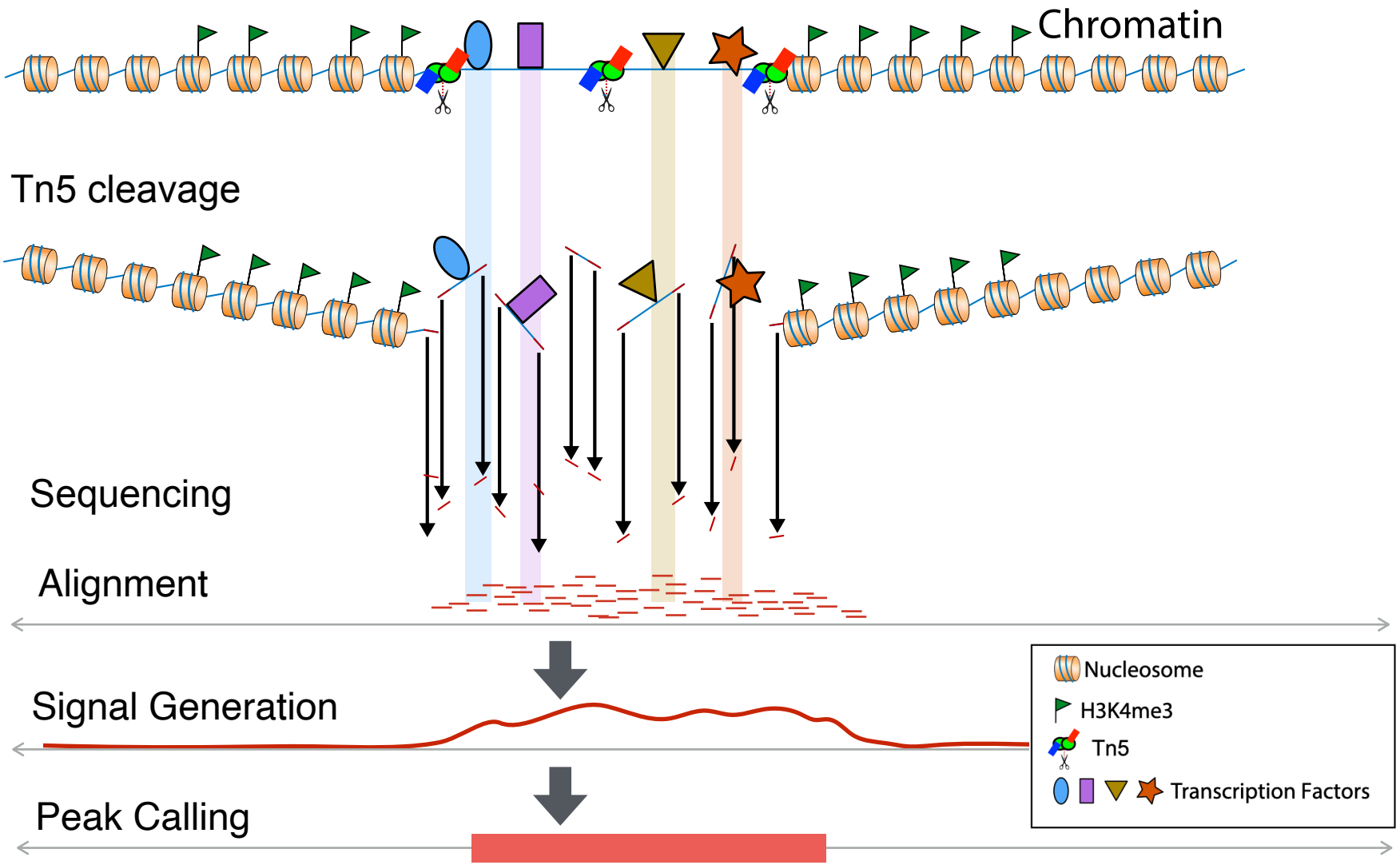
Source: Meyer, C.A. *and Liu X.S.* (2014). *Nature Reviews Genetics*.

Open chromatin with next generation sequencing



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

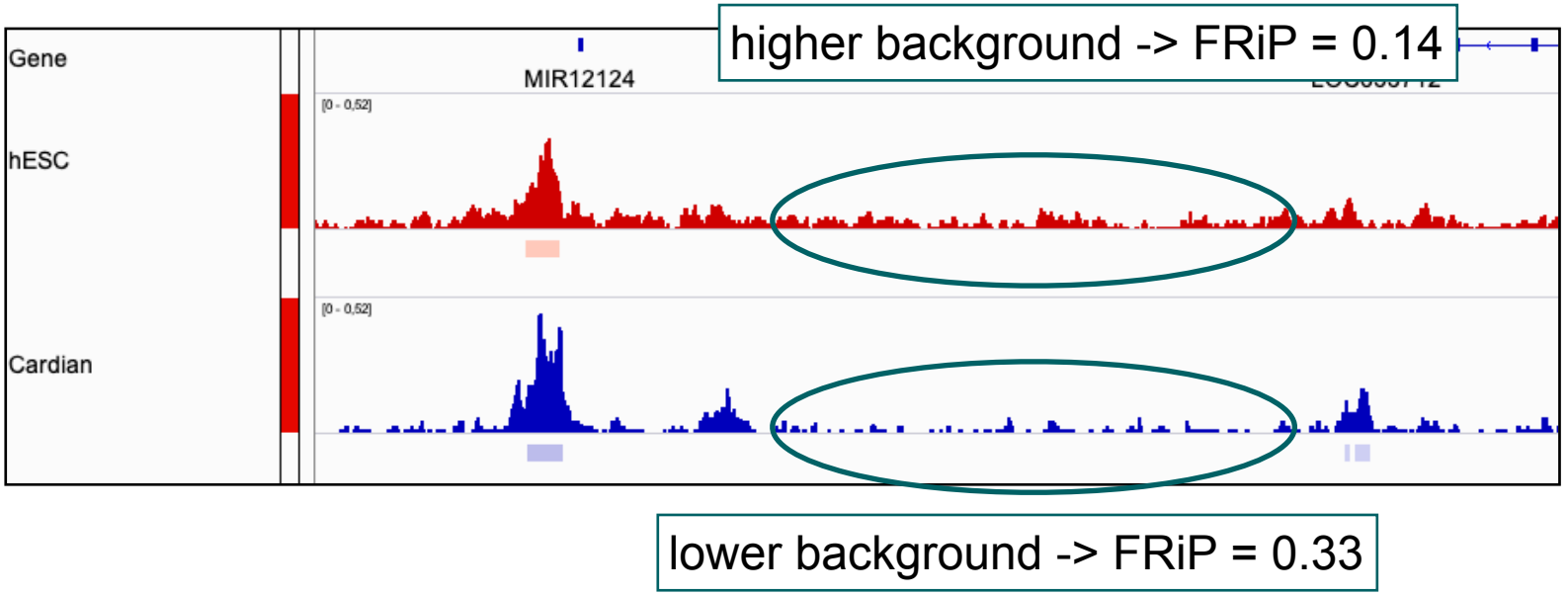
Open chromatin with ATAC-seq



Quality check of ATAC-seq by FRiP

Fraction of reads in peaks (FRiP)

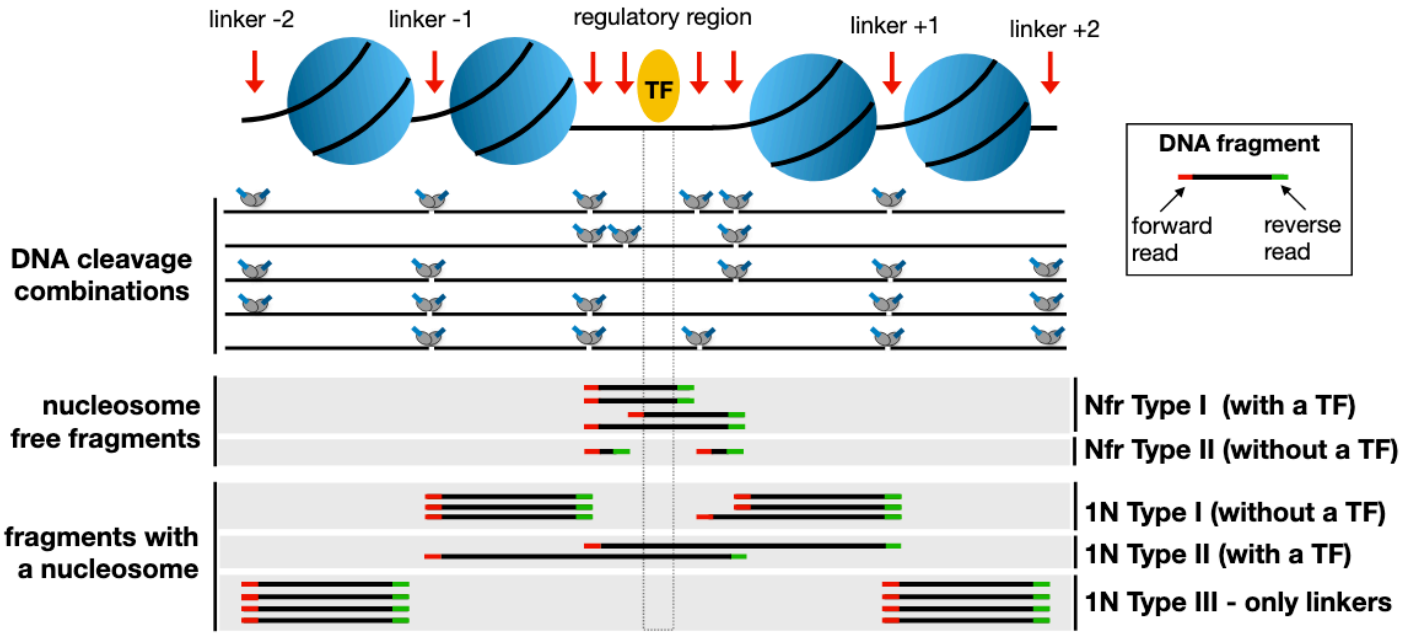
- provides an estimate of the signal (reads in peaks) vs. noise (reads in background) in a library
- FRiP > 0.3 is recommended, e.g. ENCODE



Quality check of ATAC-seq by fragments size distribution

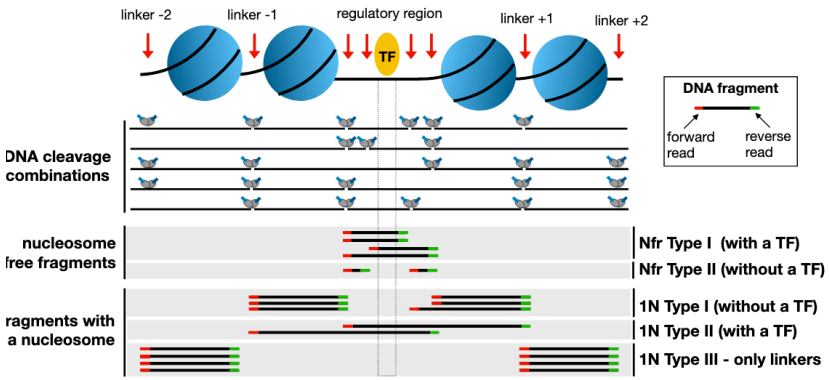
ATAC-seq libraries are mostly based on paired-end reads

- fragments are either from **nucleosome free** or **nucleosome containing regions**

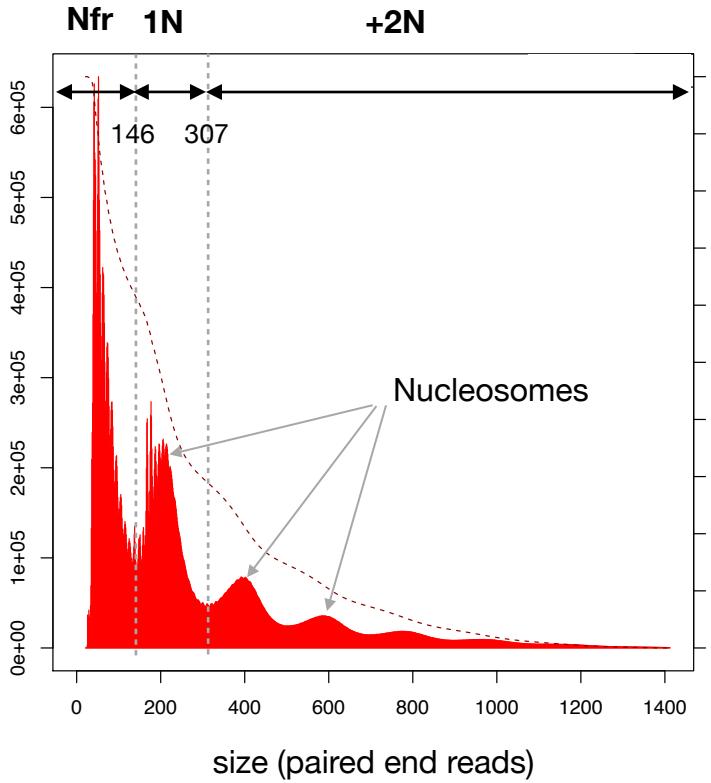


Quality check of ATAC-seq by fragments size distribution

- Fragment size distribution indicates presence of fragments with 1 or 2 nucleosomes
- Good libraries have clear nucleosome patterns



Fragment size distribution



Nfr - nucleosome free reads

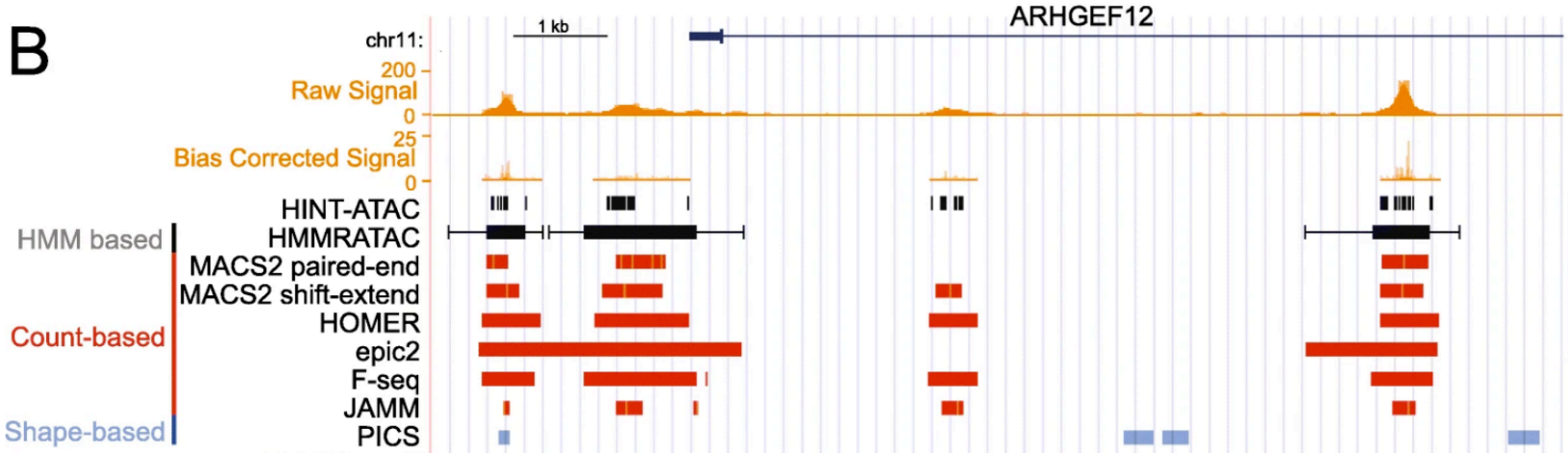
1N - reads with 1 nucleosome **reads**

+2N - reads with 2 or more nucleosomes

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

Peaks calling in ATAC-seq

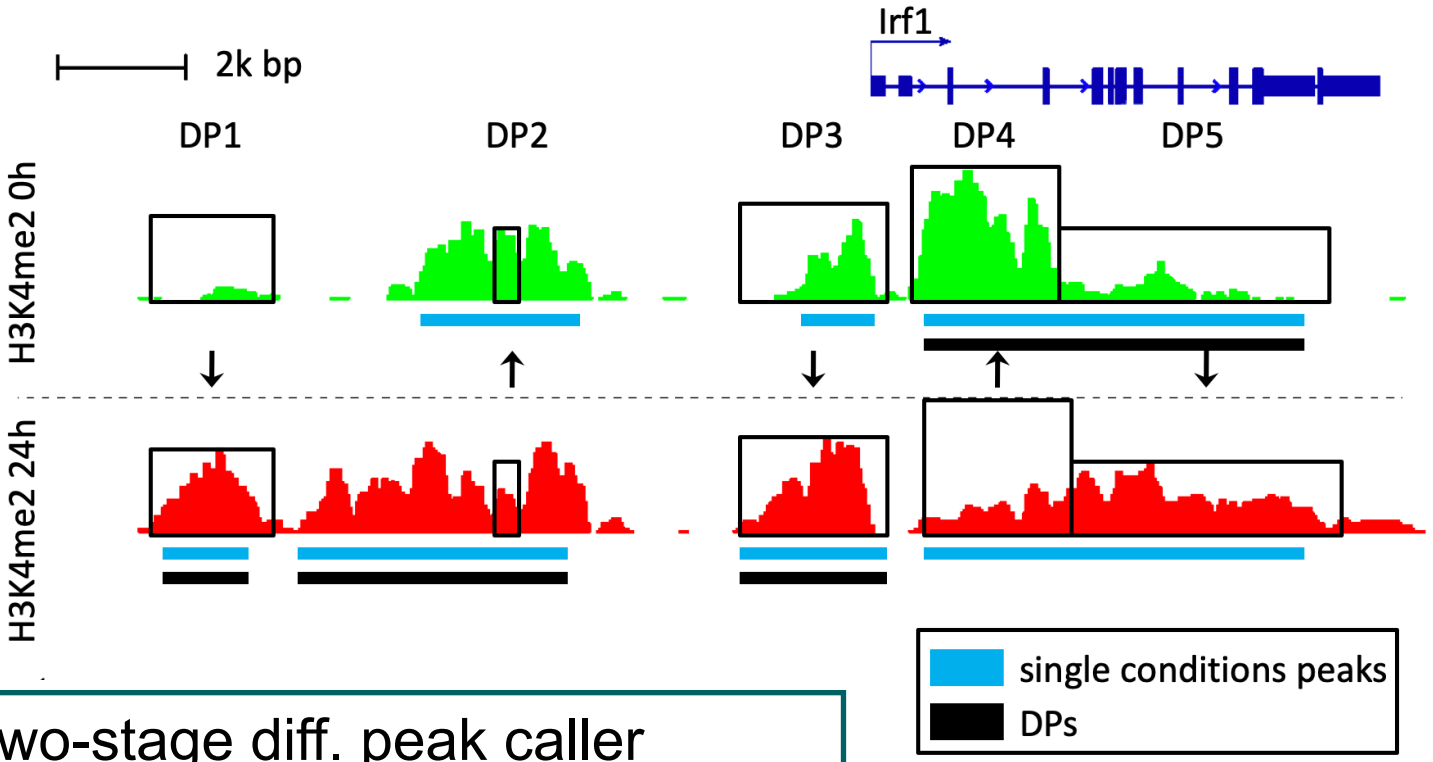


- MACS2
  - most frequently used
- HMMRATAC
  - ATAC-seq specific peak caller
  - ignores reads from large fragments / linker cleavage sites

Source: Yan, Genome Biology, 2020.

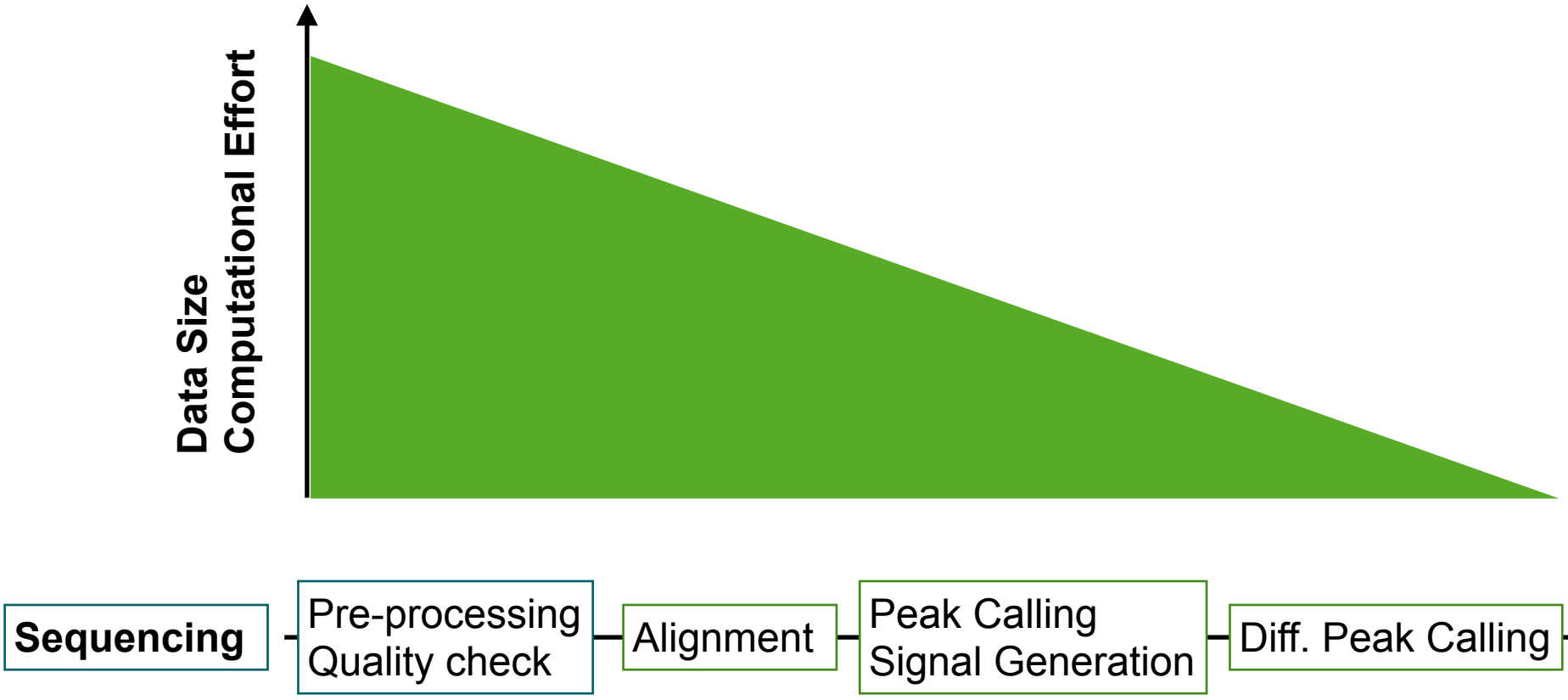
Differential peaks calling in ATAC-seq

Problem definition: Find genomic regions (of arbitrary size) with changes in read density between two conditions.



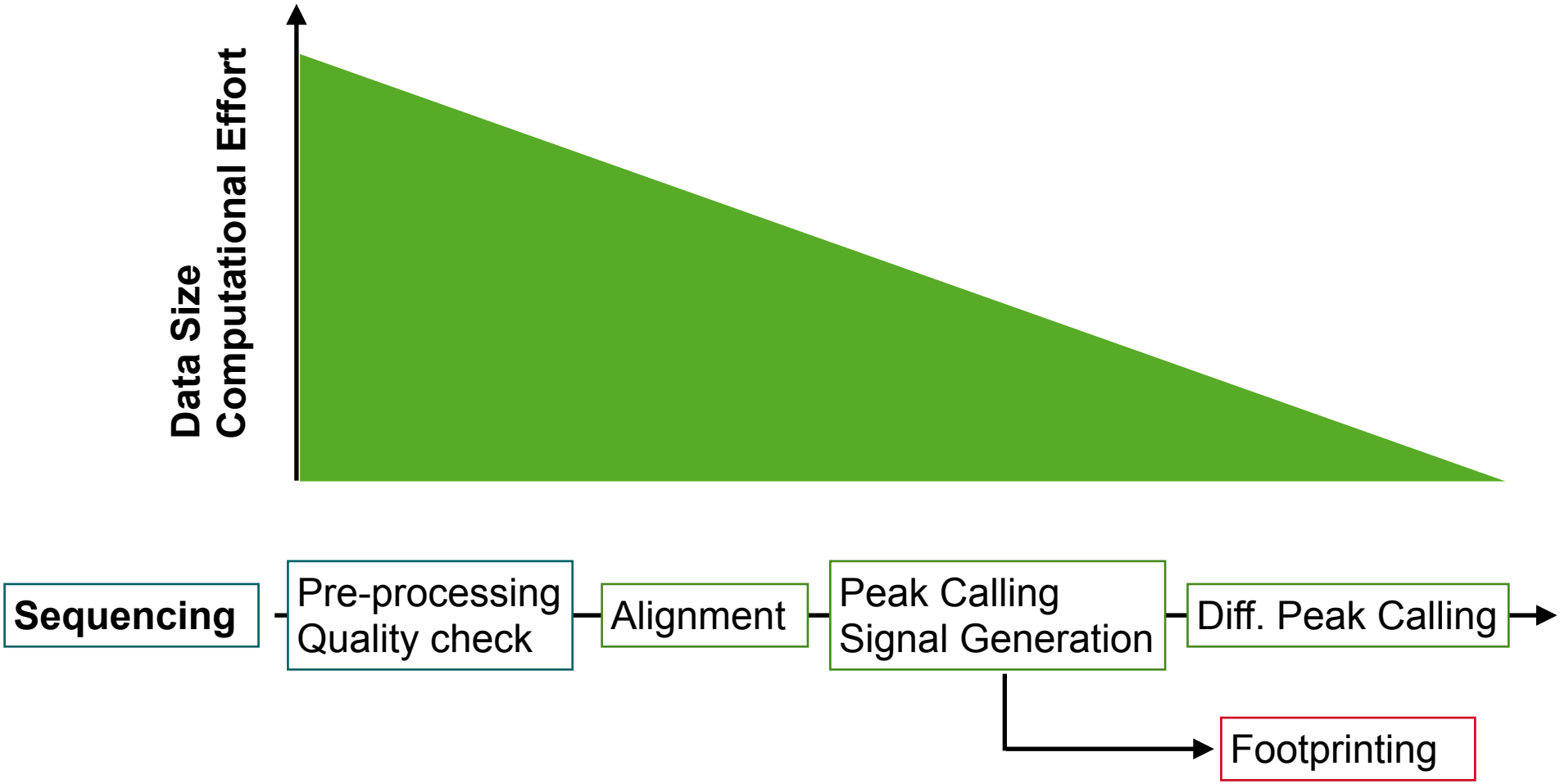
- Two-stage diff. peak caller
  - (1) MAC2 peak calling;
  - (2) Intersect peaks;
  - (3) DEseq2 on merged peaks;

Bioinformatics pipeline for ATAC-seq



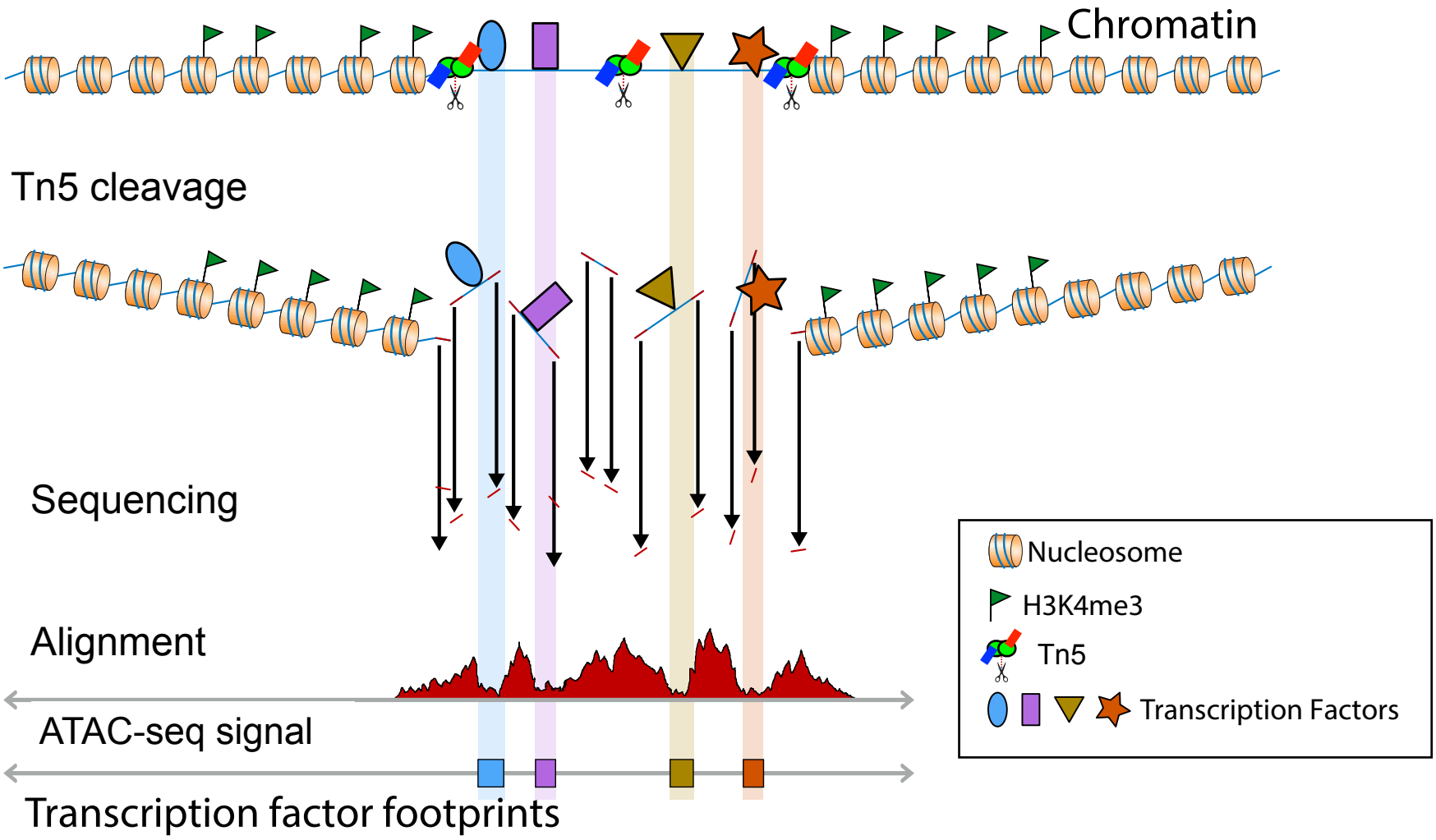
Adapted from Rasmussen:
http://www.cbs.dtu.dk/courses/27626/programme.php

Bioinformatics pipeline for ATAC-seq



Adapted from Rasmussen:
http://www.cbs.dtu.dk/courses/27626/programme.php

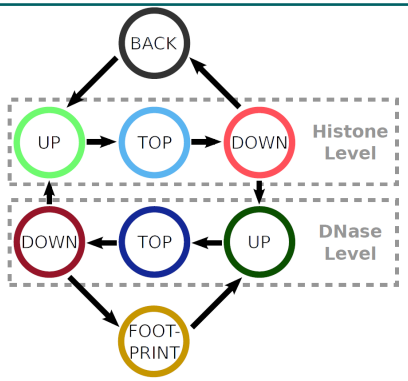
Computational footprinting with ATAC-seq



HMM-based identification of transcription factor footprints

HINT (Hmm-based IdeNtification of Transcription factor footprints)

- generate normalised cleavage signals
- trained with limited supervision
- scan multivariate signals to predict footprints



Gusmao EG *et. al*, (2014), Bioinformatics, 30(22):3143-51.
Gusmao EG *et. al,* (2016), Nature Methods, 13, 303–309.
Li et al. (2019), Genome Biology, 20:45.

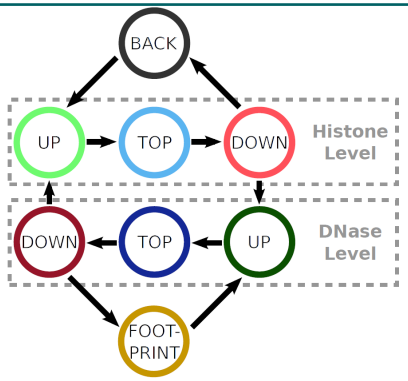
Institute for Computational Genomics
01011011010101
10100100101

RWTH AACHEN UNIVERSITY

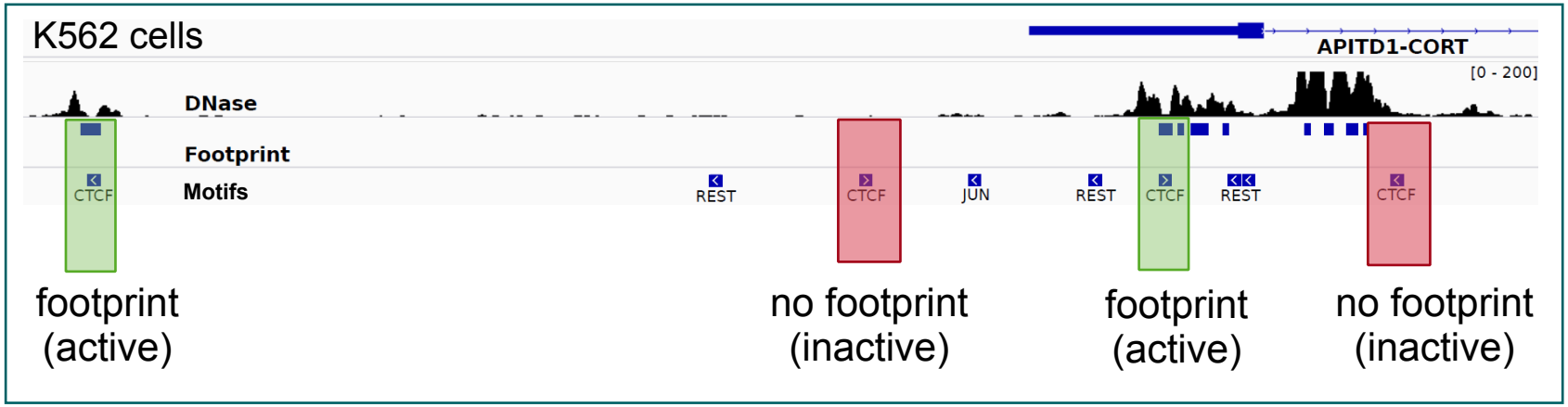
HMM-based identification of transcription factor footprints

HINT (Hmm-based IdeNtification of Transcription factor footprints)

- generate normalized cleavage signals
- trained with limited supervision
- scan multivariate signals to predict footprints



Prediction and Evaluation



Gusmao EG *et. al*, (2014), Bioinformatics, 30(22):3143-51.
Gusmao EG *et. al,* (2016), Nature Methods, 13, 303–309.
Li et al. (2019), Genome Biology, 20:45.

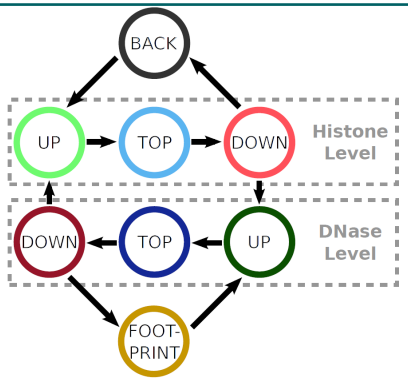
Institute for Computational Genomics
01011011010
10100100101

RWTH AACHEN UNIVERSITY

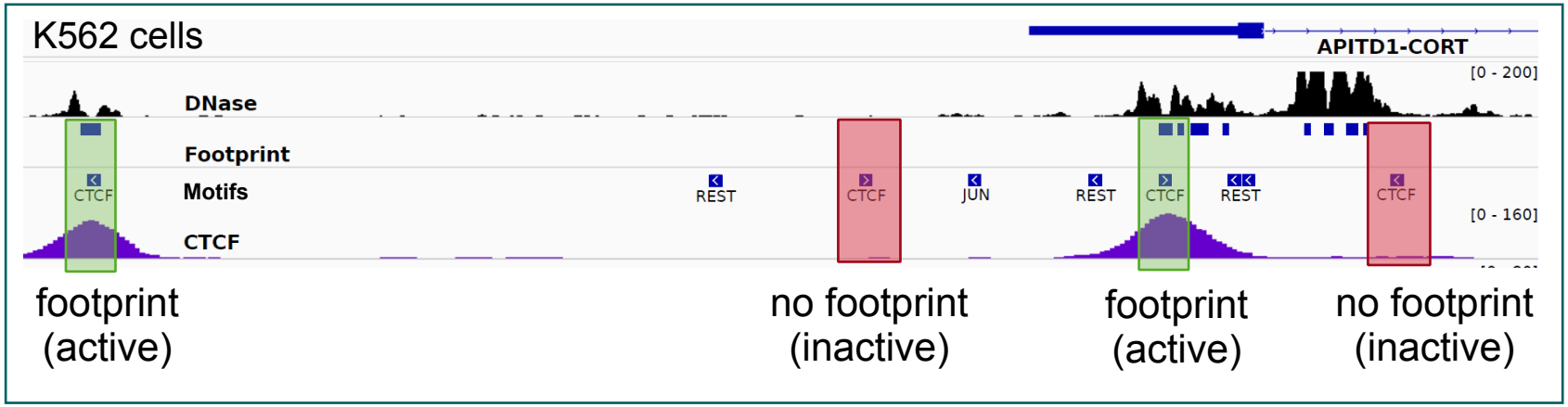
HMM-based identification of transcription factor footprints

HINT (Hmm-based IdeNtification of Transcription factor footprints)

- generate normalized cleavage signals
- trained with limited supervision
- scan multivariate signals to predict footprints



Prediction and Evaluation



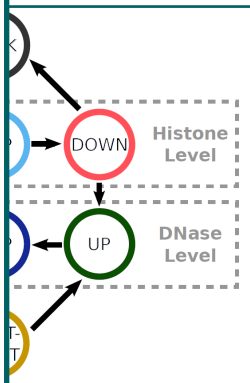
Gusmao EG *et. al*, (2014), Bioinformatics, 30(22):3143-51.
Gusmao EG *et. al,* (2016), Nature Methods, 13, 303–309.
Li et al. (2019), Genome Biology, 20:45.

Institute for Computational Genomics
01011011010
10100100101

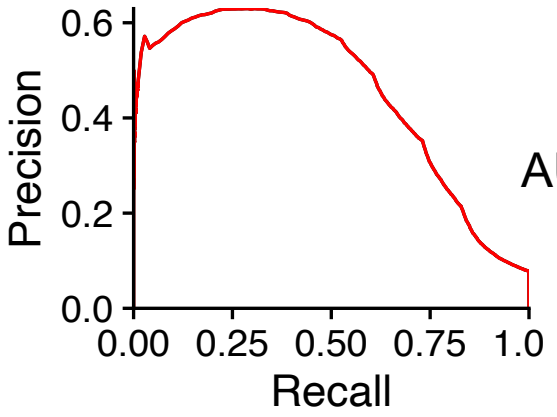
RWTH AACHEN UNIVERSITY

HMM-based identification of transcription factor footprints

HINT (Hmm
Transcriptio
- generate r
- trained wit
- scan multi

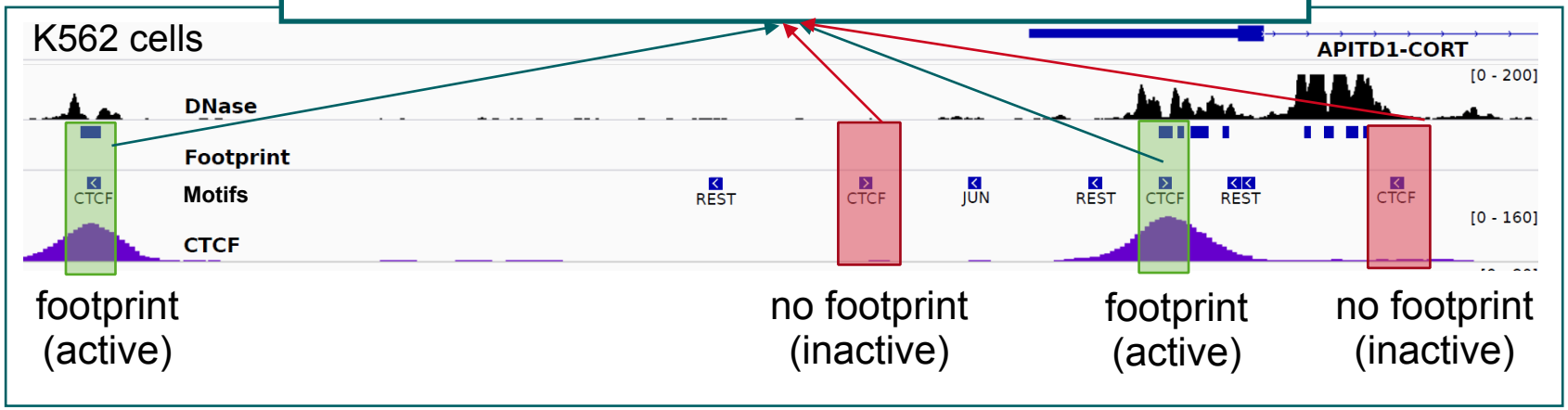


AUPR (Area Under Precision Recall)



AUPR=0.49

Prediction



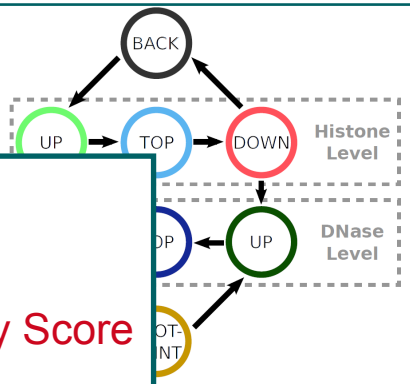
Gusmao EG *et. al*, (2014), Bioinformatics, 30(22):3143-51.
Gusmao EG *et. al,* (2016), Nature Methods, 13, 303–309.
Li et al. (2019), Genome Biology, 20:45.

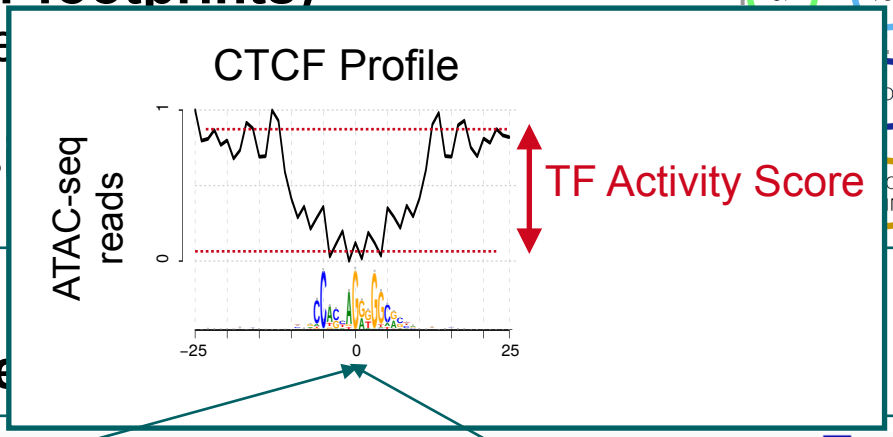
Institute for Computational Genomics
01011011010
10100100101
RWTH AACHEN UNIVERSITY

HMM-based identification of transcription factor footprints

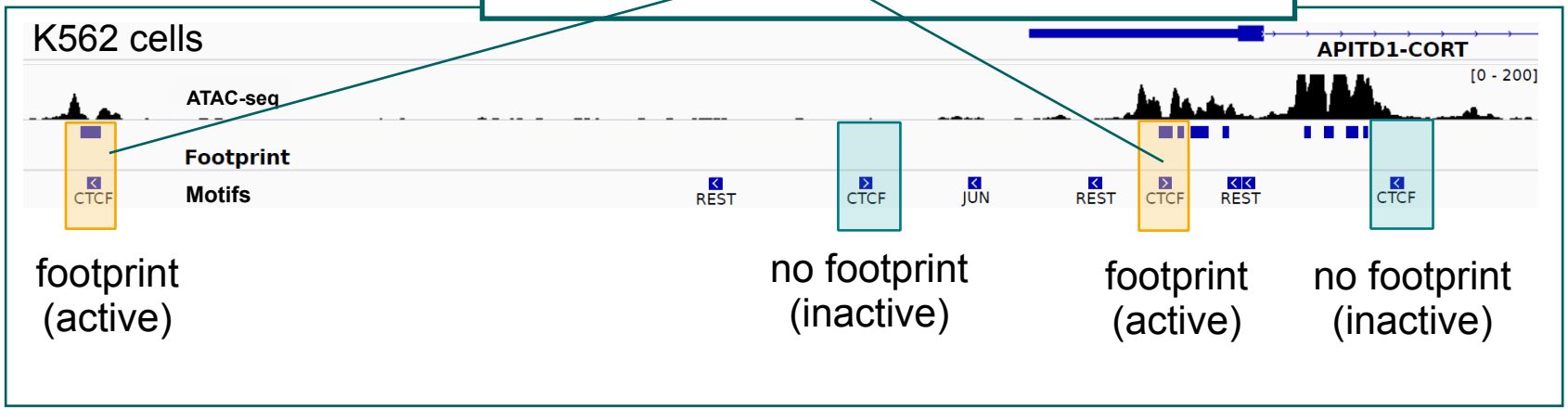
HINT (Hmm-based IdeNtification of Transcription factor footprints)

- generate normalize
- trained with limited
- scan multivariate s





Prediction Example

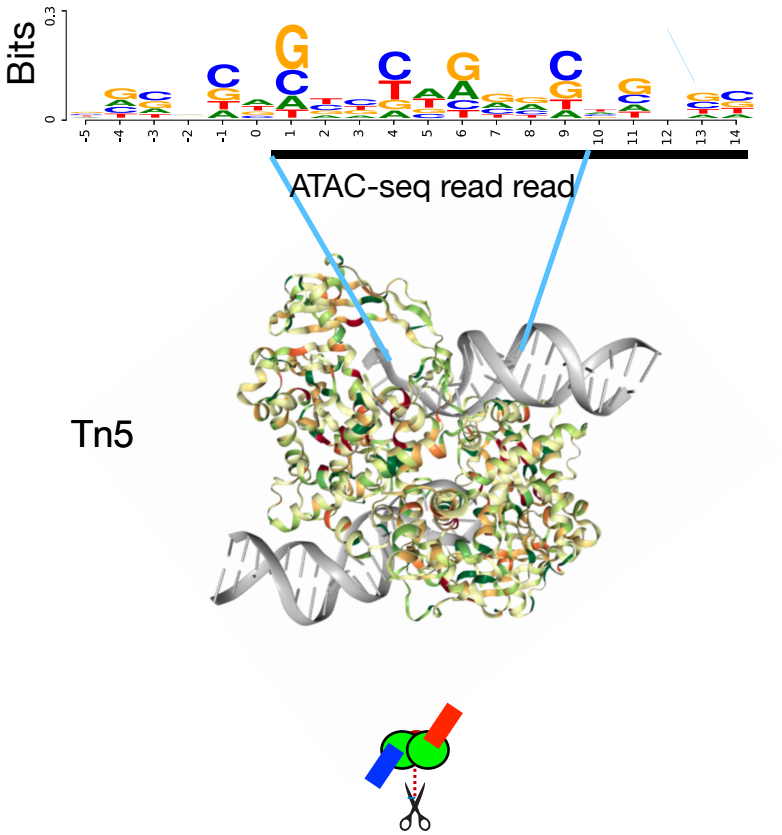


Gusmao EG *et. al*, (2014), Bioinformatics, 30(22):3143-51.
Gusmao EG *et. al,* (2016), Nature Methods, 13, 303–309.
Li et al. (2019), Genome Biology, 20:45.

Institute for Computational Genomics
01011011010
10100100101

RWTH AACHEN UNIVERSITY

ATAC-seq cleavage bias correction



- Some sequences have more/less ATAC-seq cuts than expected

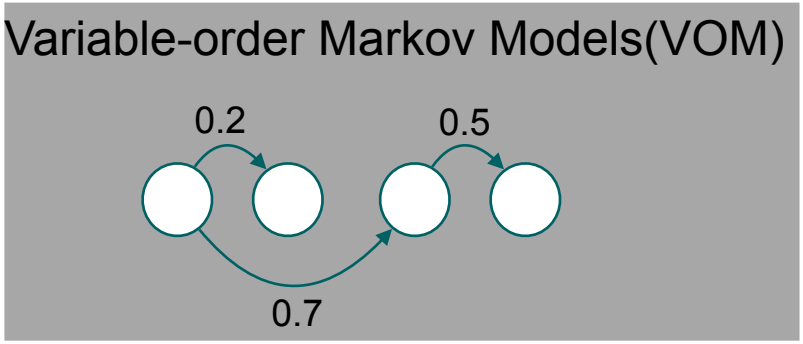
...ACCGGG...

3.5 times more

...TCAATT...

3.4 times less

- We use Variable-order Markov models to capture ATAC-seq cleavage bias

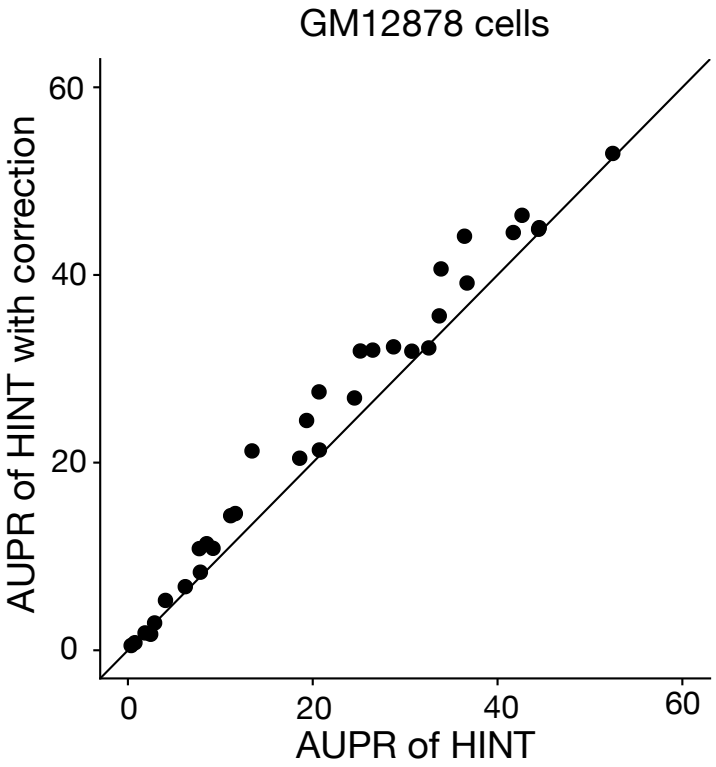


Gusmao EG *et. al*, (2014), Bioinformatics, 30(22):3143-51.
Gusmao EG *et. al,* (2016), Nature Methods, 13, 303–309.
Li et al. (2019), Genome Biology, 20:45.

Institute for Computational Genomics
01011011010
10100100101

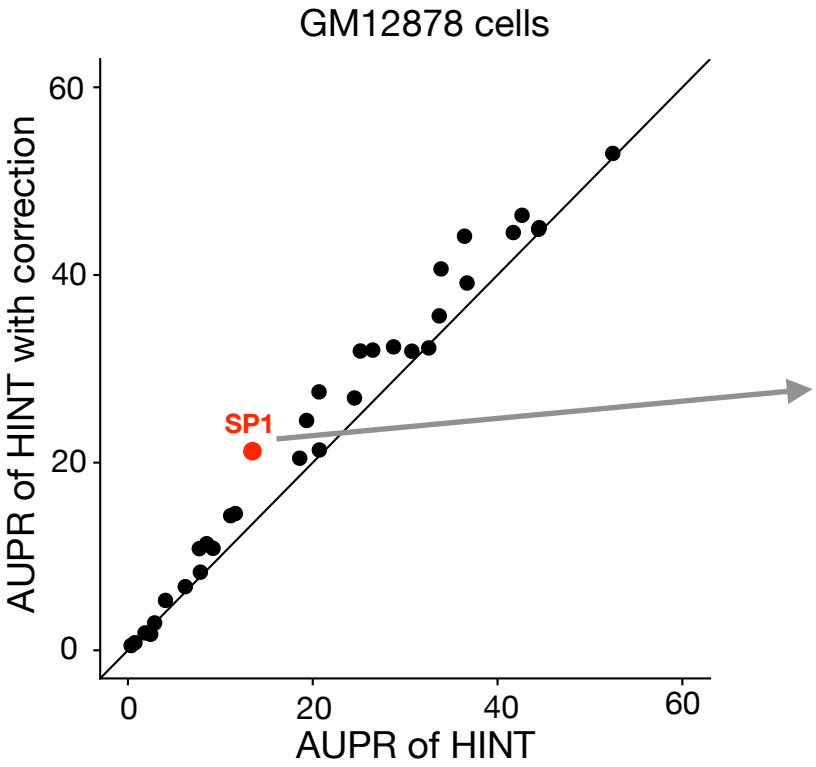
RWTH AACHEN UNIVERSITY

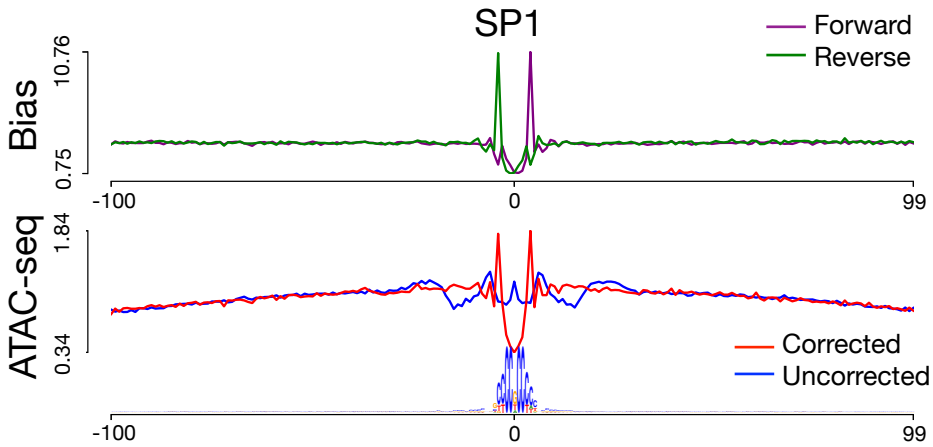
Results: cleavage bias correction method



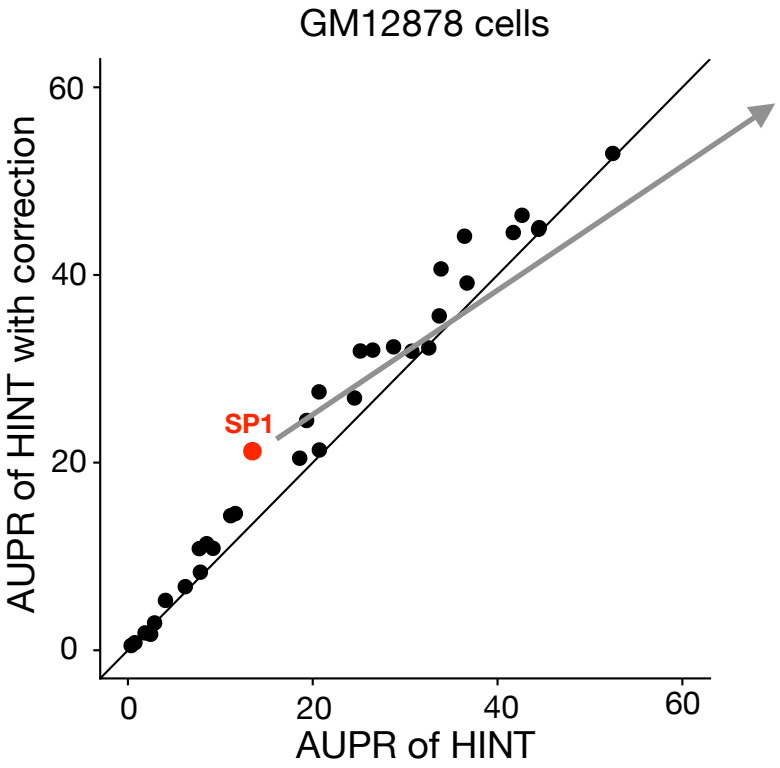
Li et al. (2019), Genome Biology, 20:45.

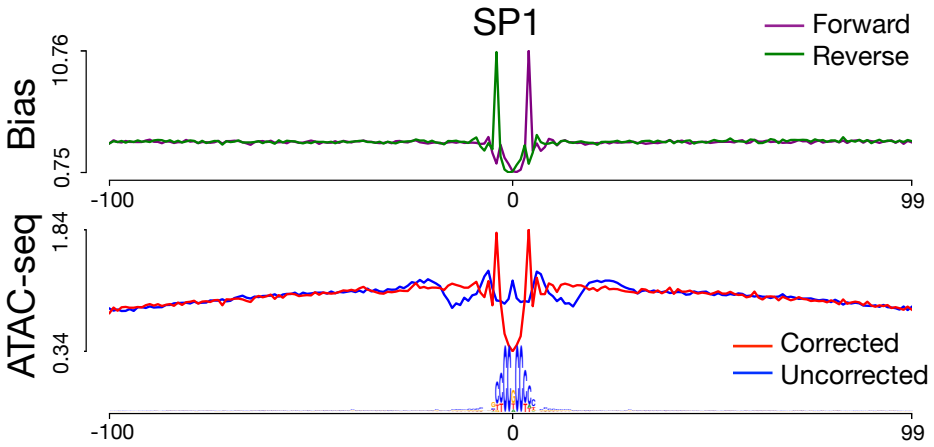
Results: cleavage bias correction method

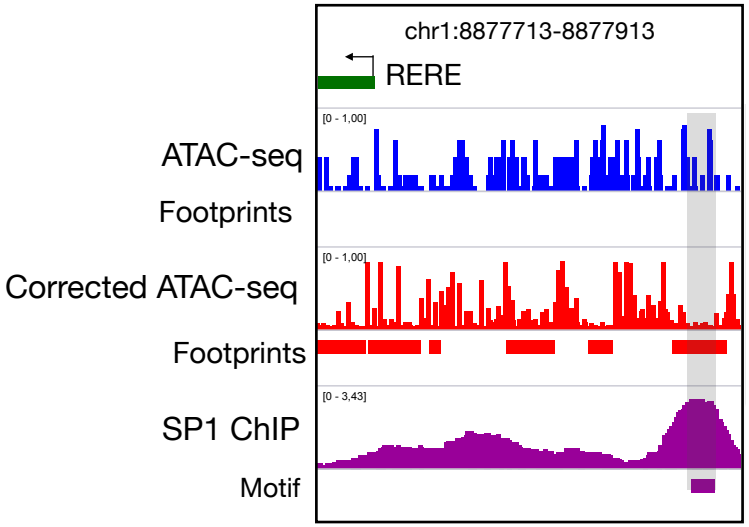




Results: cleavage bias correction method





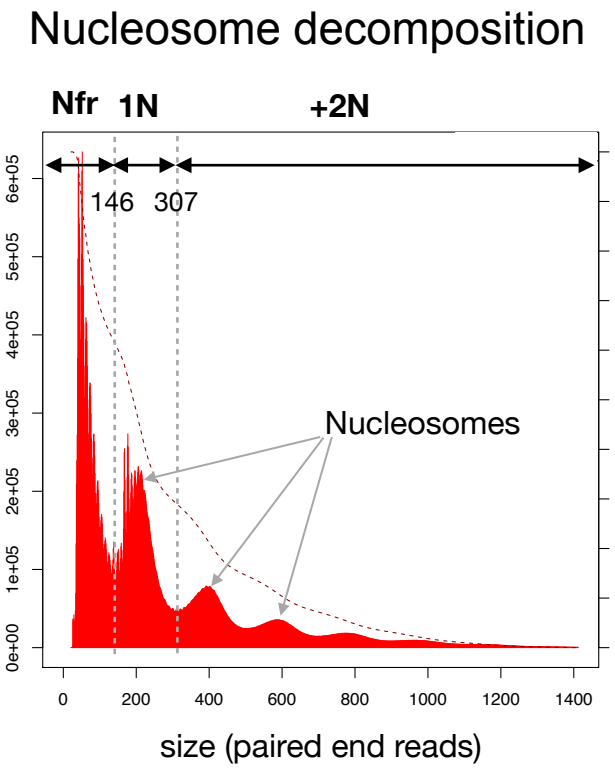


Li et al. (2019), Genome Biology, 20:45.

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

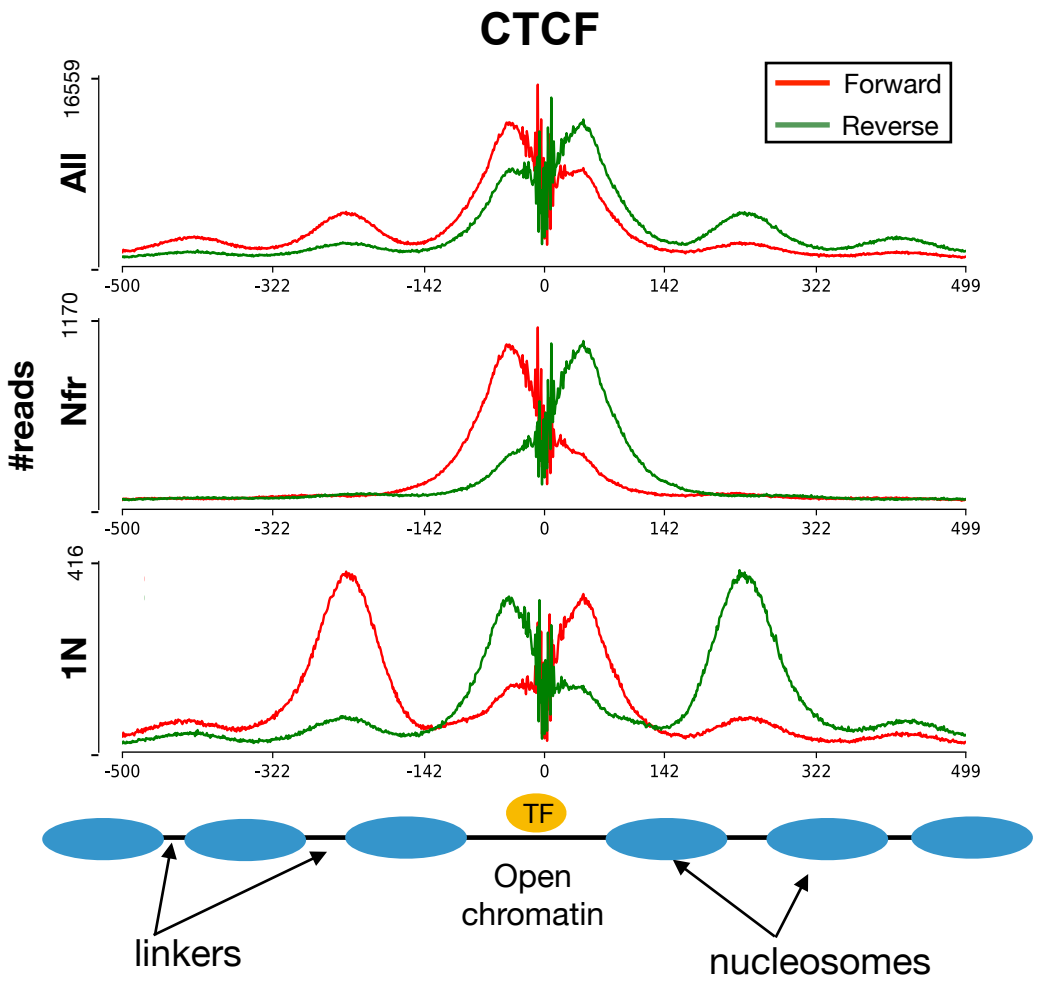
Nucleosome decomposition



Nfr - nucleosome free reads

1N - reads with 1 nucleosome

+2N - reads with 2 or more nucleosomes



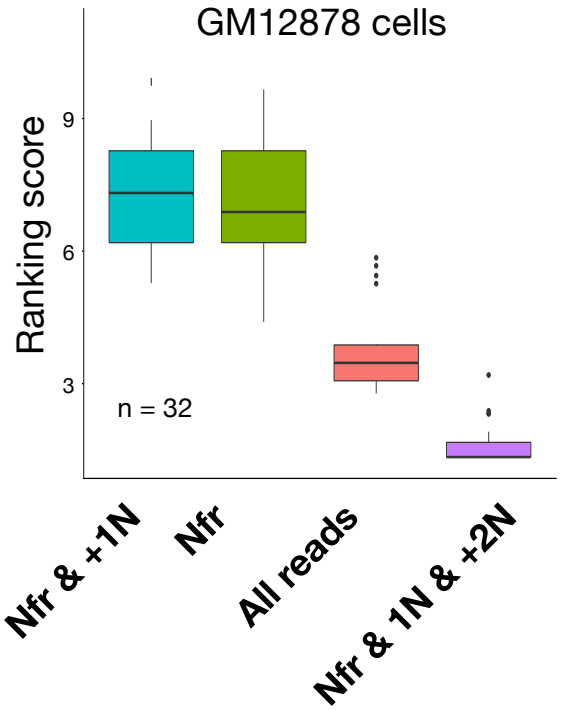
Li et al. (2019), Genome Biology, 20:45.

Institute for Computational Genomics

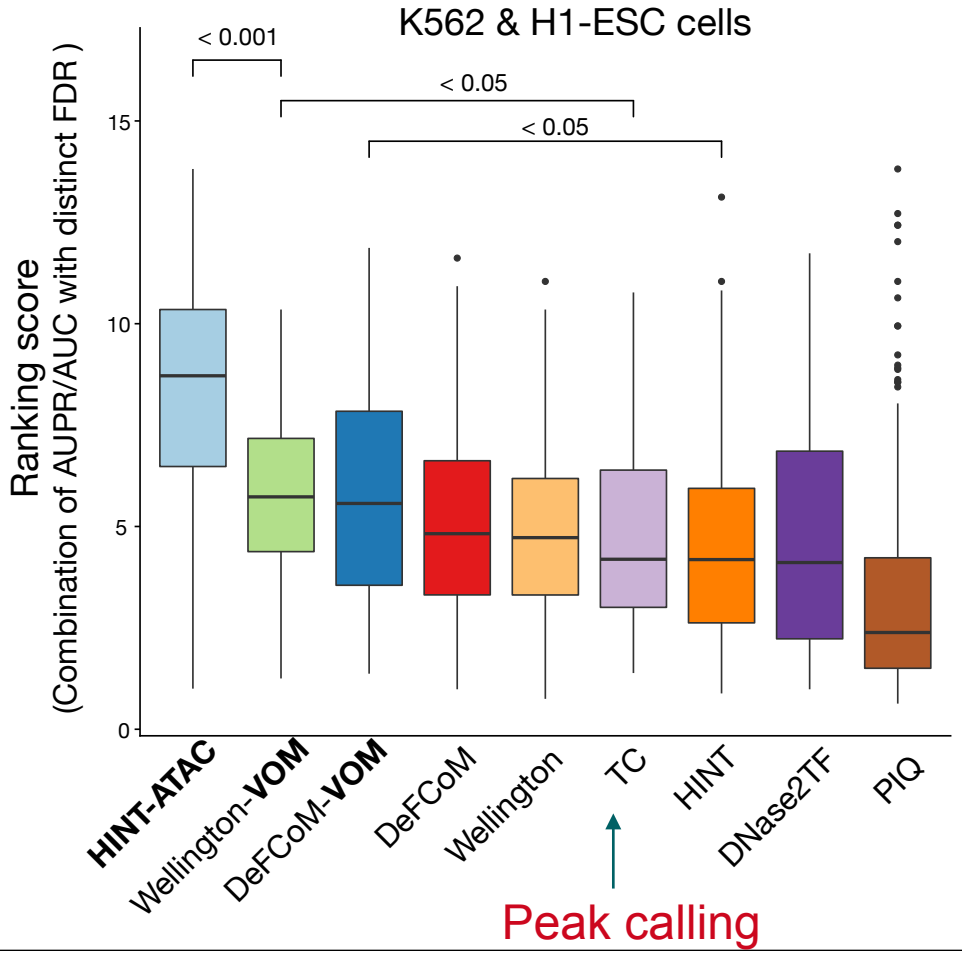
RWTH AACHEN UNIVERSITY

Results: nucleosome decomposition and footprinting

- HINT-ATAC with nucleosome decomposed signals

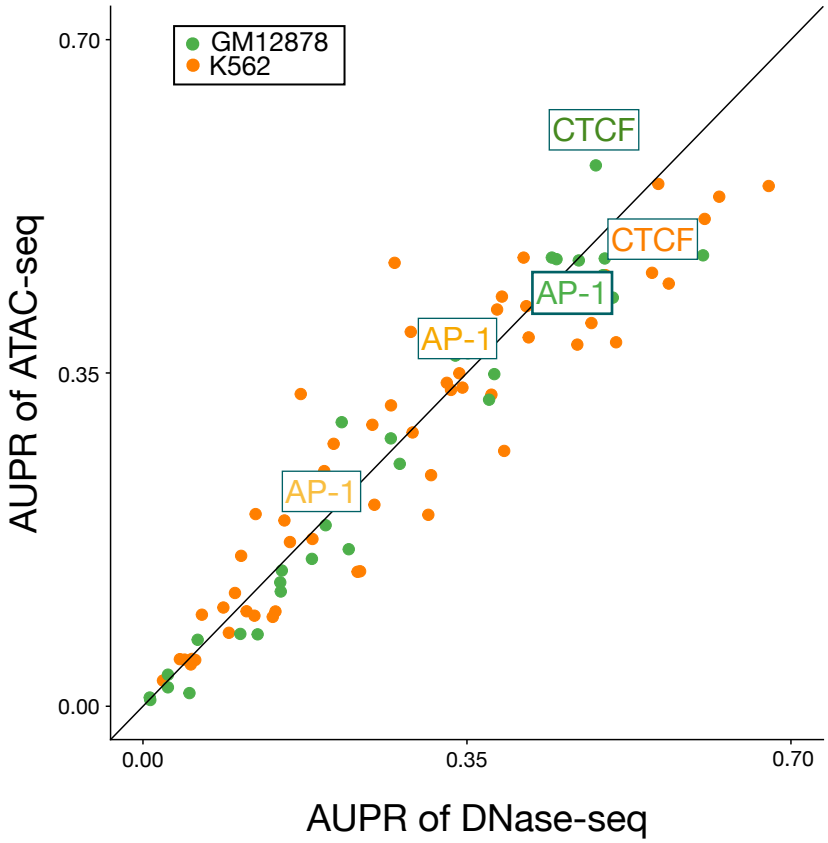


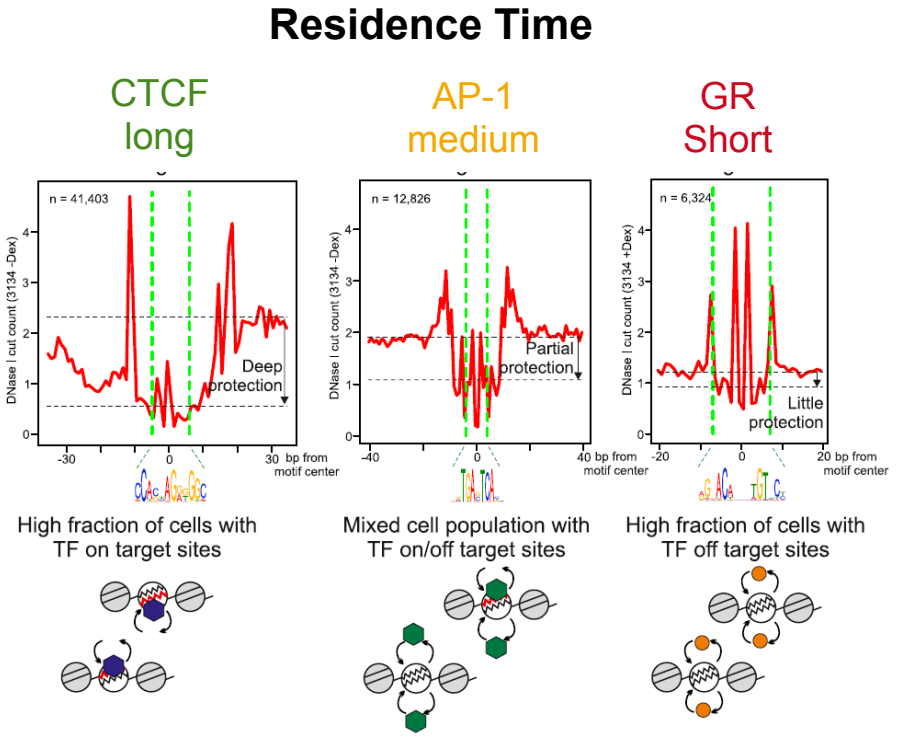
- Comparative analysis of footprinting methods



Caveats of footprinting - residence time

- TFs with low residence have lower AUPR values





Sung, M.H. et al. (2014). *Mol Cell 56(2):275-85.*

Li et al. (2019), Genome Biology, 20:45.

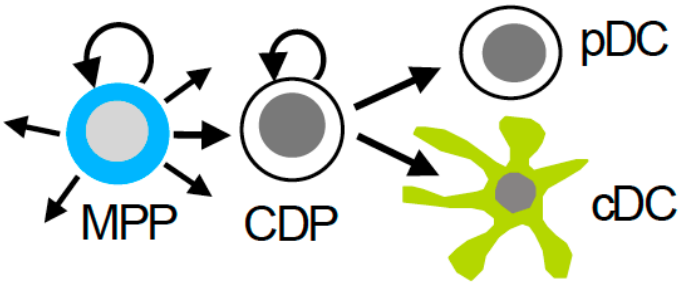
Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

Regulatory control of dendritic cell differentiation

Which transcription factors control dendritic cell differentiation? How is chromatin regulated?

- *transcription factor ChIP-Seq*
  - *PU.1/Sfpi1 – master regulator*
- *gene expression (RNA-Seq)*
- histone modifications H3K4me3, H3K4me1,H3K27ac,H3K27me3
- **open chromatin (ATAC-seq)**

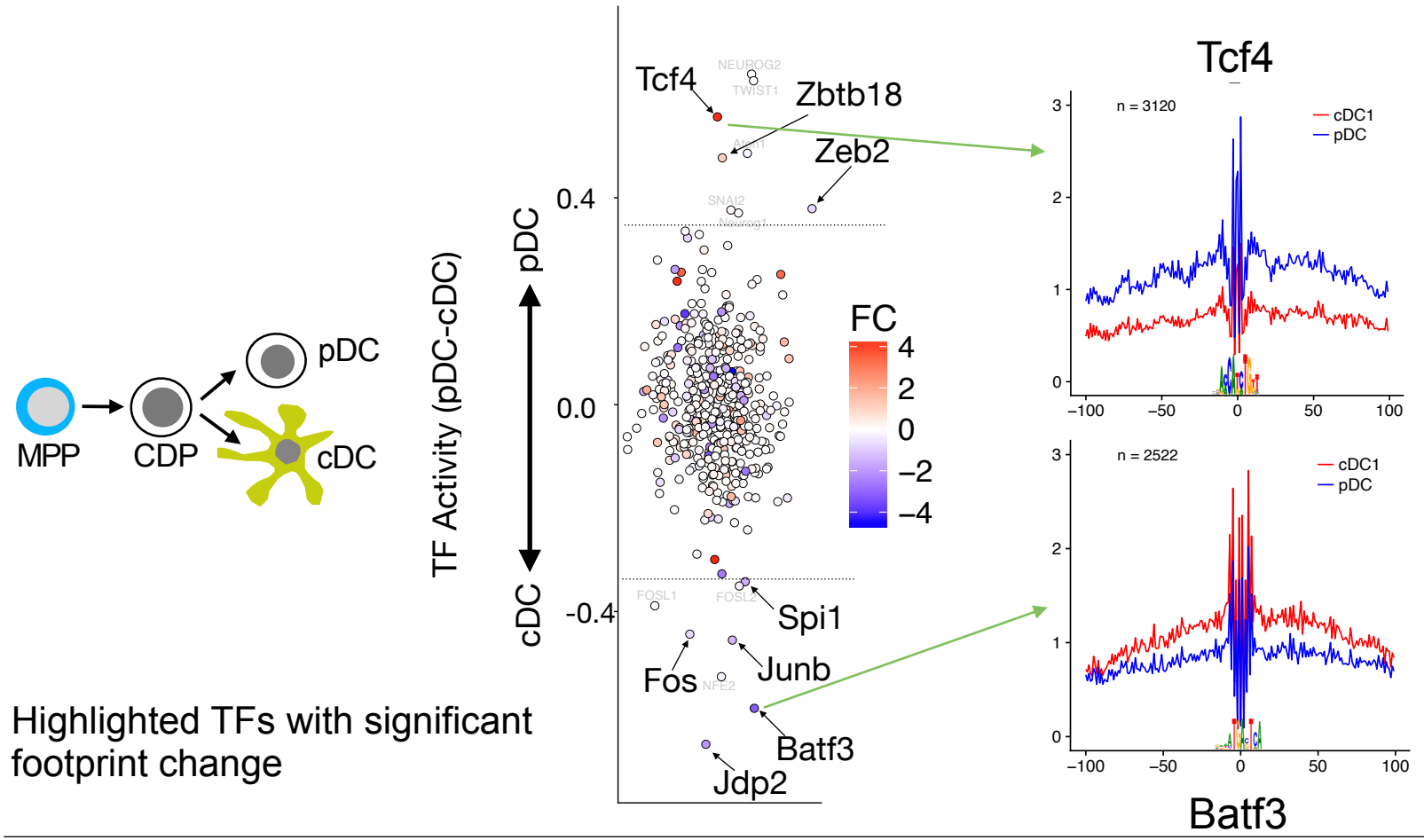


col. with M. Zenke, & K. Sere, Cell Biology Department, UK Aachen

Lin Q *et al.* (2015), NAR, Allhof et al. (2016), NAR,
Li et al. (2019), Genome Biology, 20:45.

Differential footprinting between cDC and pDC

We performed footprinting in cDC and pDC ATAC-seq data and measured differences in TF activity scores for > 500 motifs



Highlighted TFs with significant footprint change

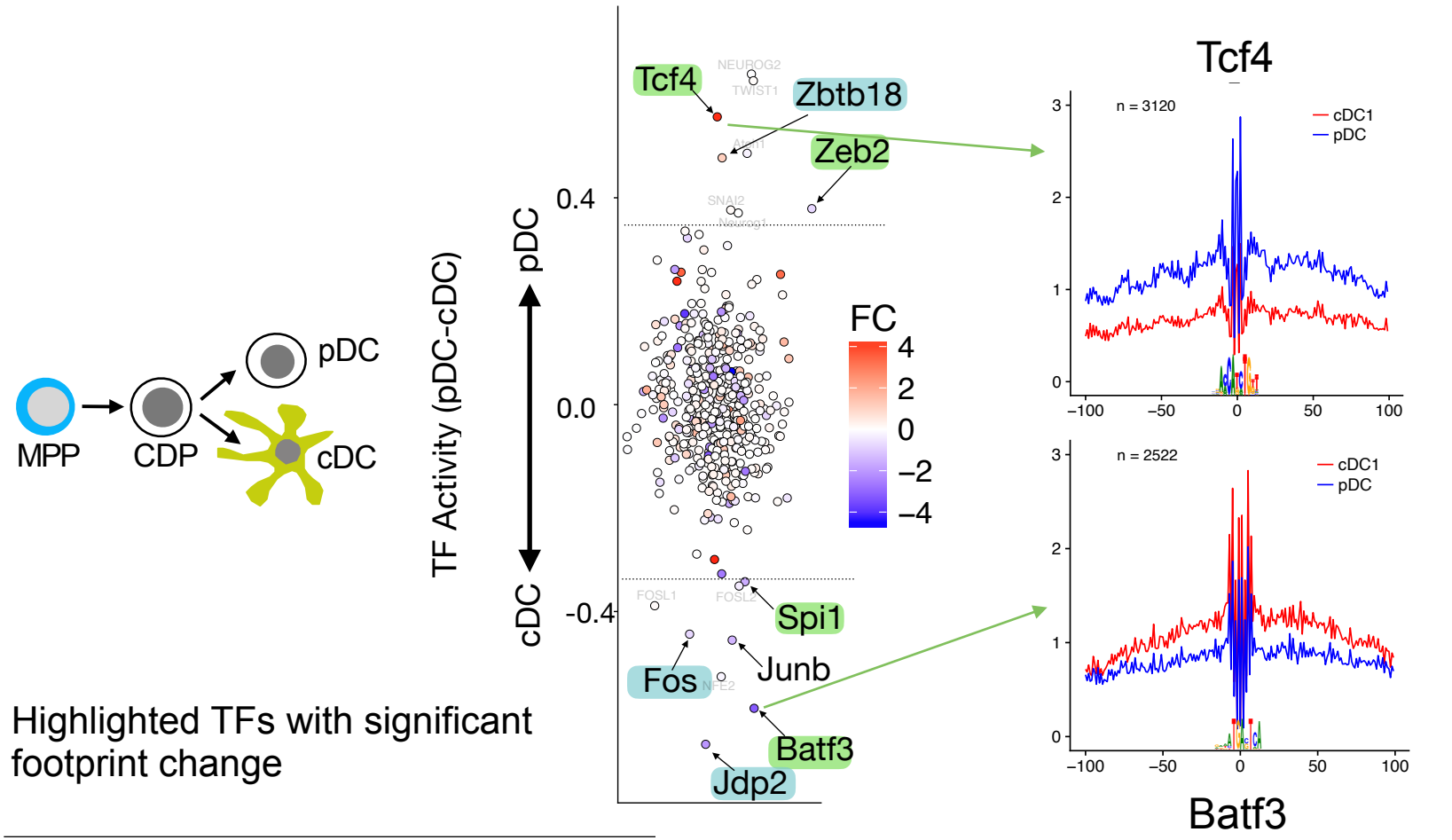
Li et al. (2019), Genome Biology, 20:45.

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

Differential footprinting between cDC and pDC

We performed footprinting in cDC and pDC ATAC-seq data and measured differences in TF activity scores for > 500 motifs



Highlighted TFs with significant footprint change

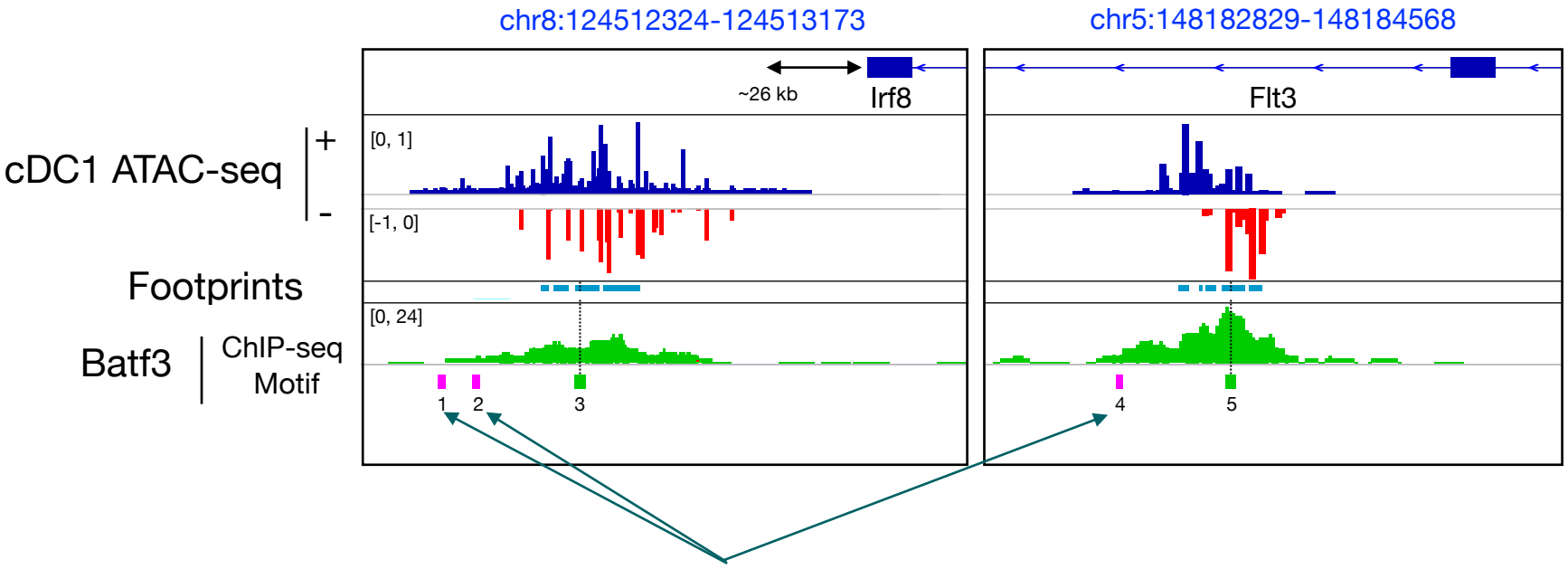
Li et al. (2019), Genome Biology, 20:45.

known DC TFs

uncharacterized TFs

Results: immunological dendritic cell differentiation

Example: footprints with Batf3 motifs close to DC genes



False positives from the use of peaks

Resume

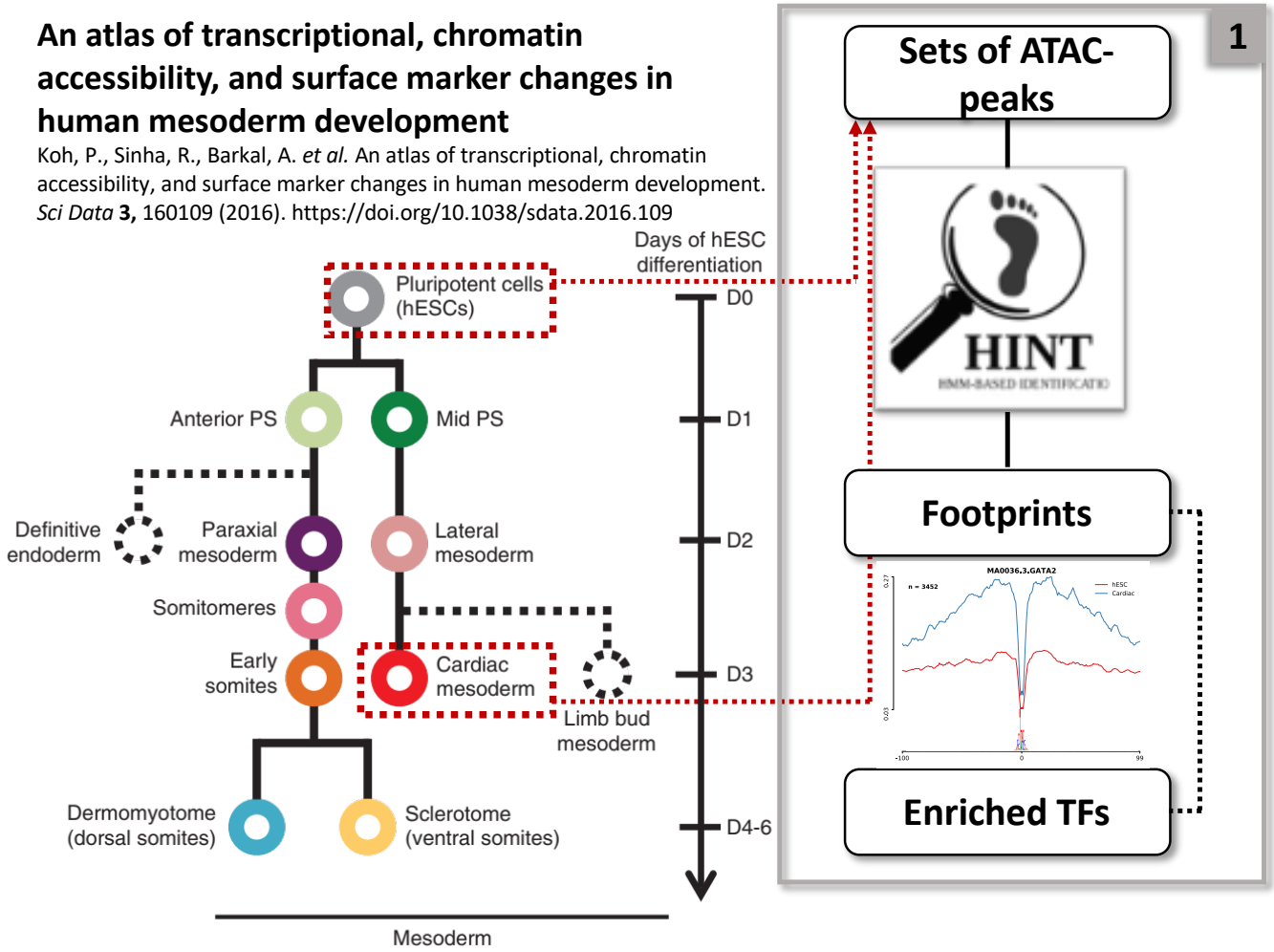
- Overview of relevant aspects of basic ATAC-seq analysis
  - quality check, alignment, peak calling

- Footprinting
  - allow detection of cell specific TF binding sites
  - differential TF activity analysis

Practical session 1

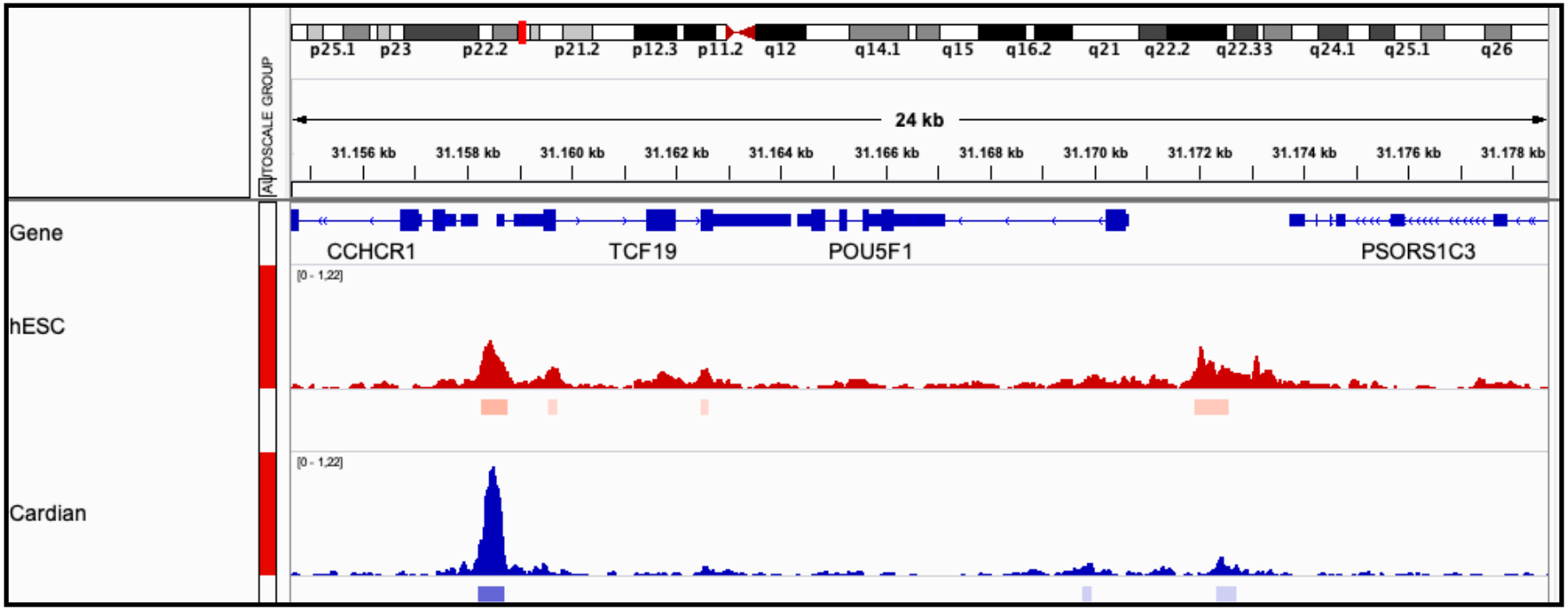
An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development

Koh, P., Sinha, R., Barkal, A. *et al.* An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development. *Sci Data* **3,** 160109 (2016). https://doi.org/10.1038/sdata.2016.109

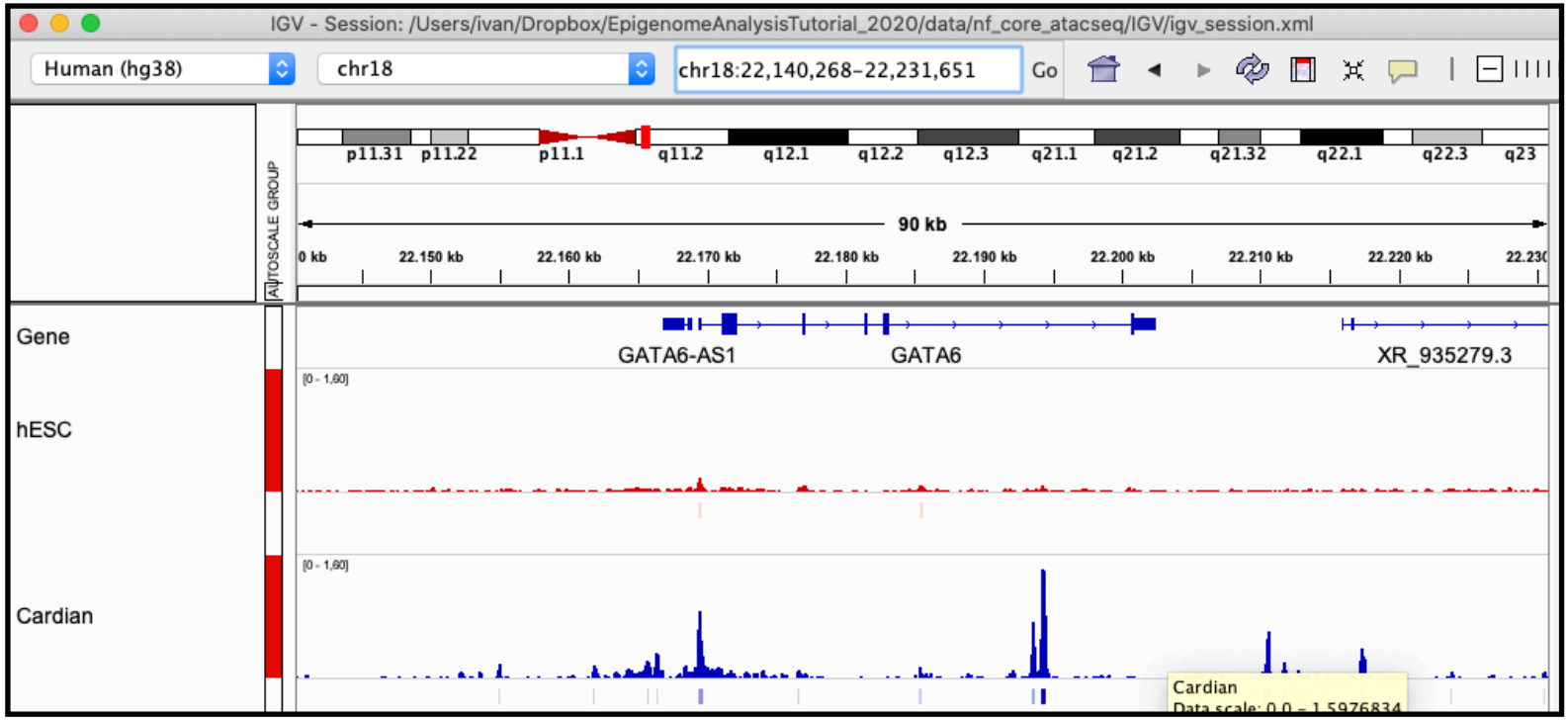


https://epigenomeanalysistutorial-2020.readthedocs.io/en/latest/Practical1.html

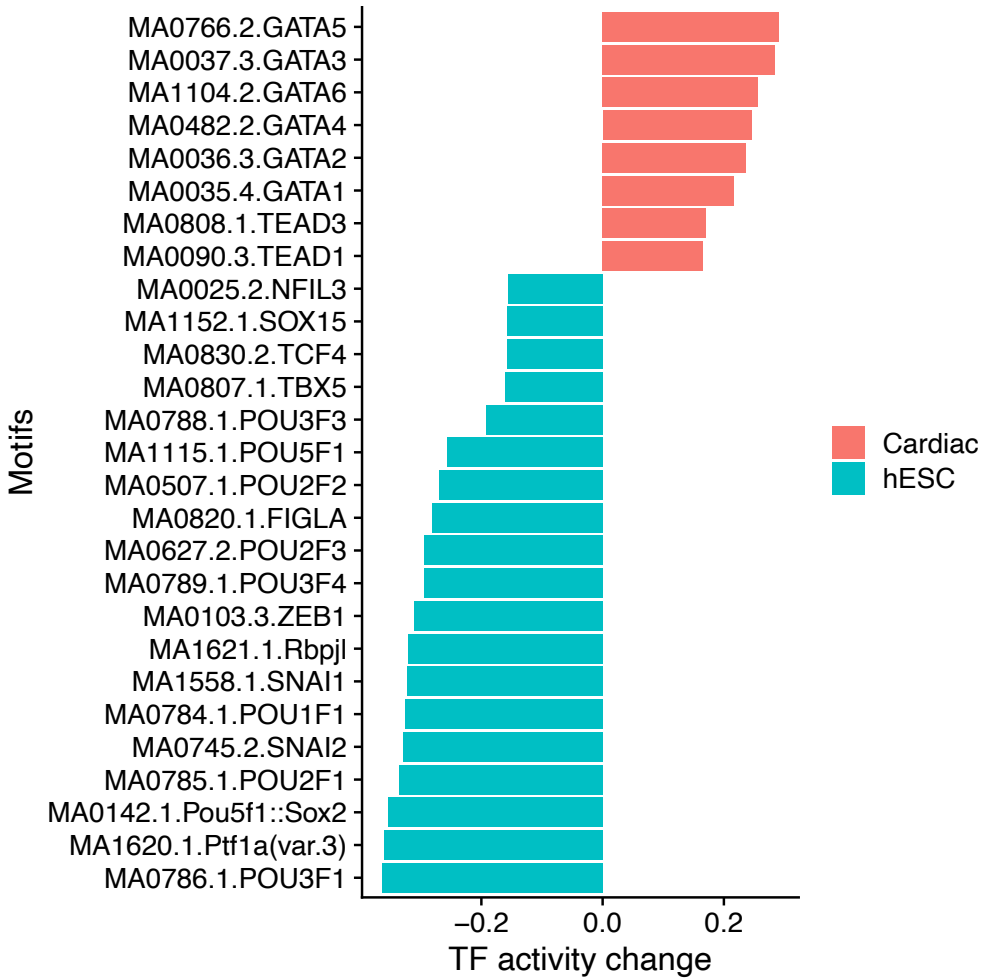
Genome Browser - ES cell gene

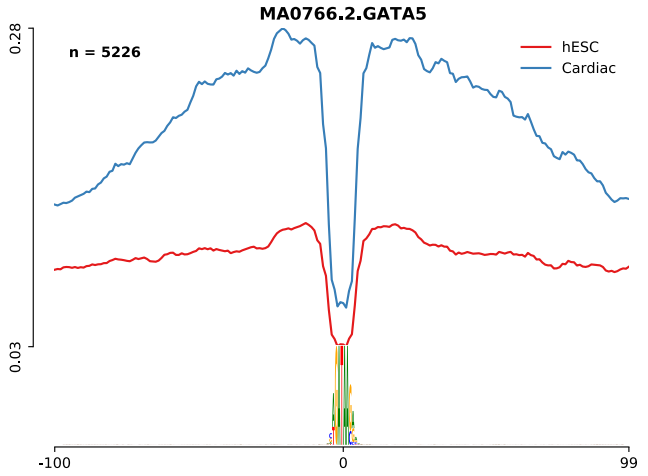


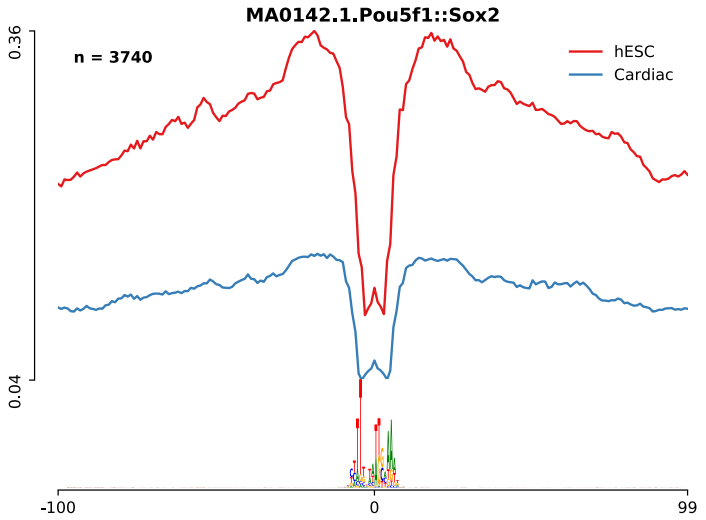
Genome Browser - Cardiac Genes



Results - TF Activity Analysis





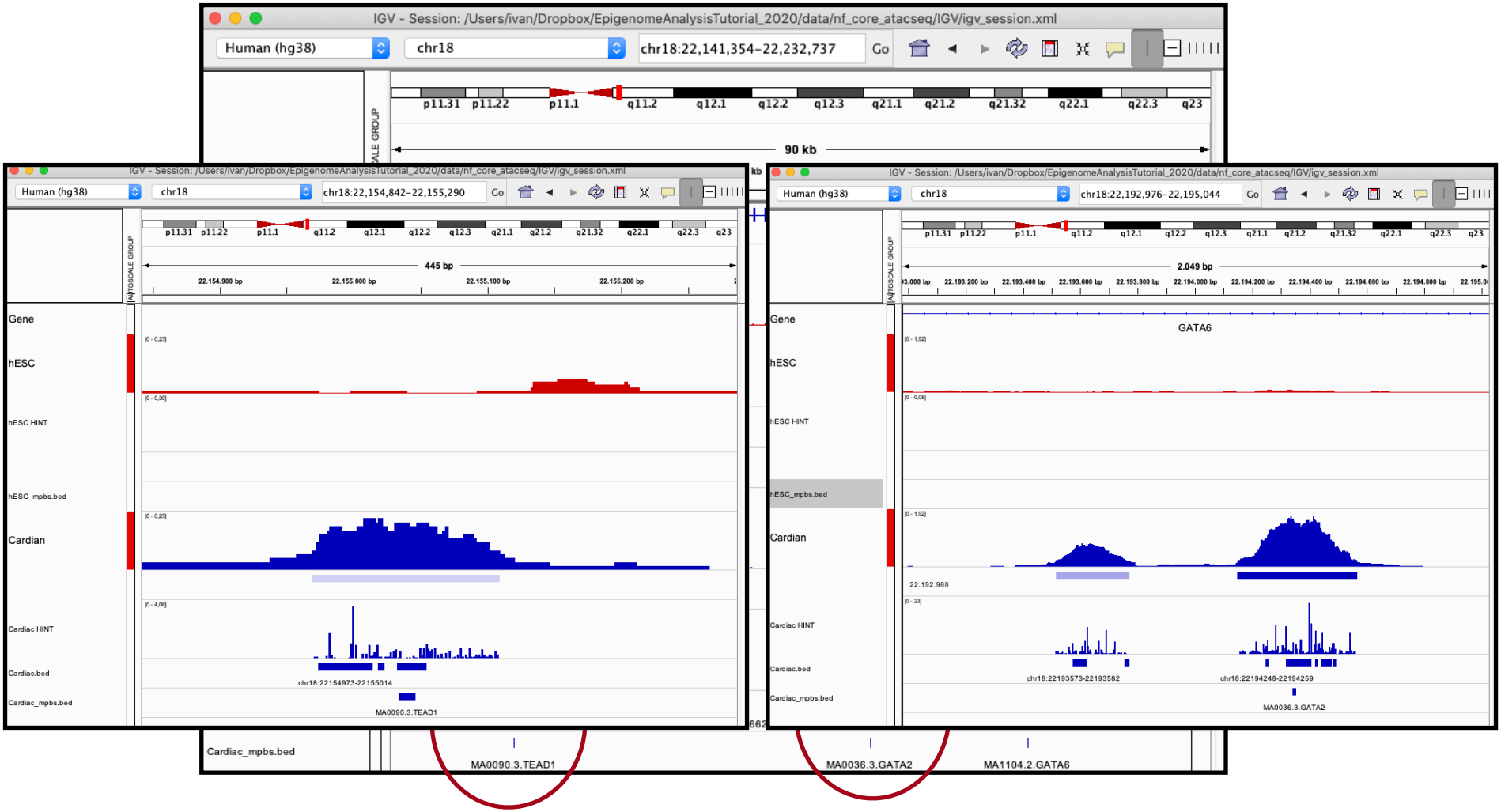


Institute for Computational Genomics
01011011010101
10100100101

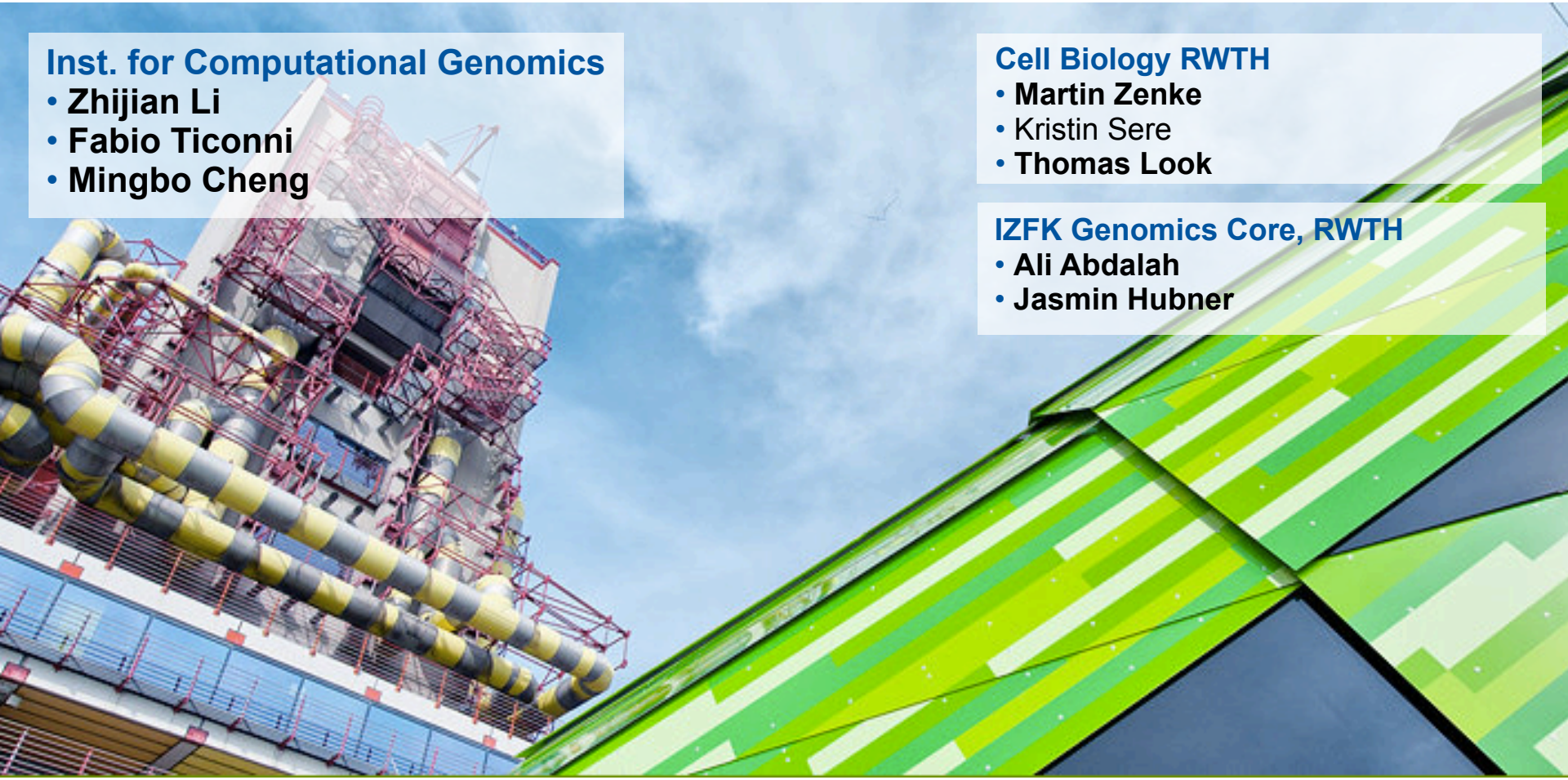
RWTH AACHEN UNIVERSITY

Footprints - Genome Browser

Footprints - Genome Browser



www.costalab.org



Inst. for Computational Genomics
- **Zhijian Li**
- **Fabio Ticonni**
- **Mingbo Cheng**

Cell Biology RWTH
- **Martin Zenke**
- Kristin Sere
- **Thomas Look**

IZFK Genomics Core, RWTH
- **Ali Abdalah**
- **Jasmin Hubner**

Funding:

IZKF Interdisziplinäres Zentrum für Klinische Forschung

Federal Ministry of Education and Research

DFG

Tools: regulatory-genomics.org/hint