Bioinformatics Lab

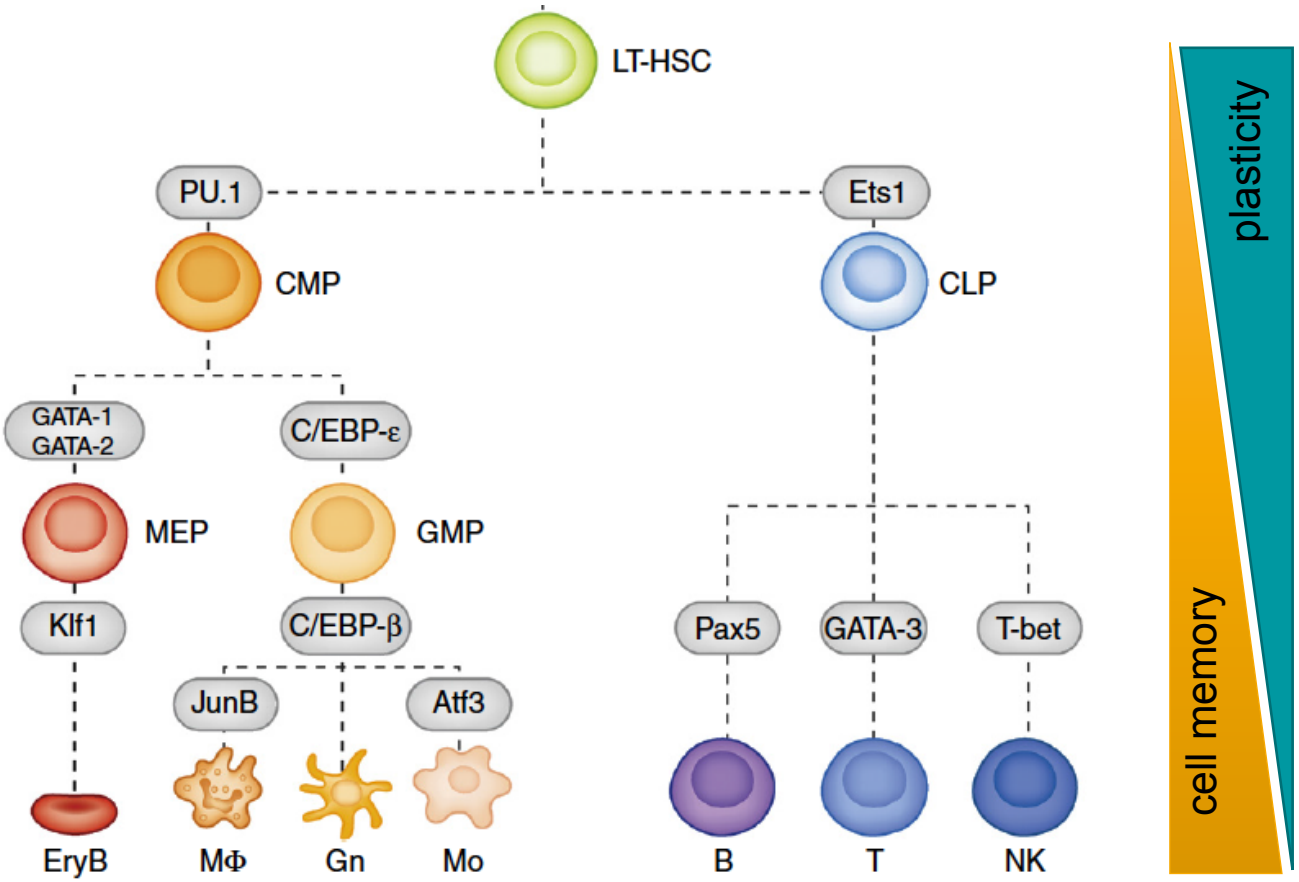
Ivan Gesteira Costa, Mingbo Cheng, Zhijian Li, James Nagai, Mina Shaigan
Institute for Computational Genomics

Institute for Computational Genomics
01011011010
10100100101
RWTH AACHEN UNIVERSITY

Computational Epigenomics

Cell Differentiation

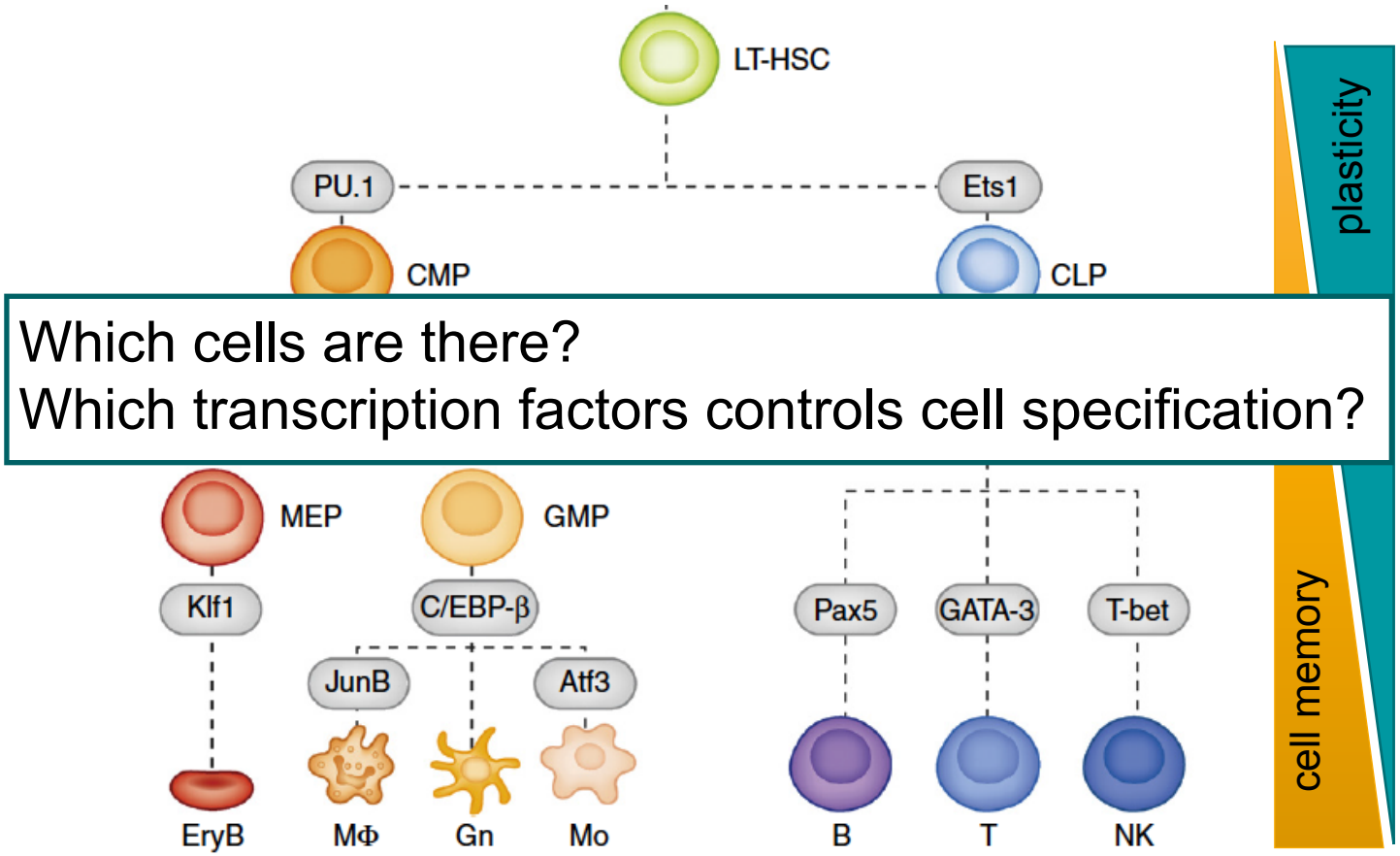
Hematopoiesis



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

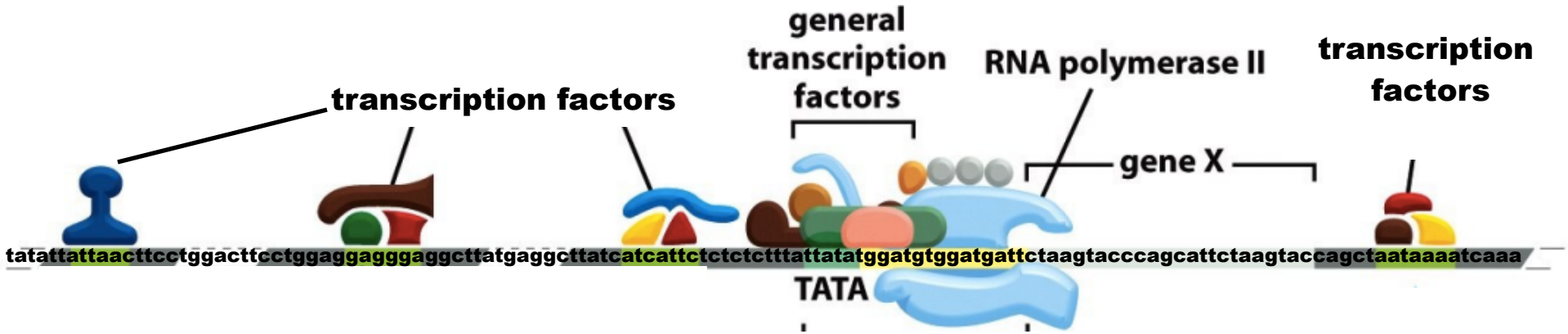
Cell Differentiation

Hematopoiesis



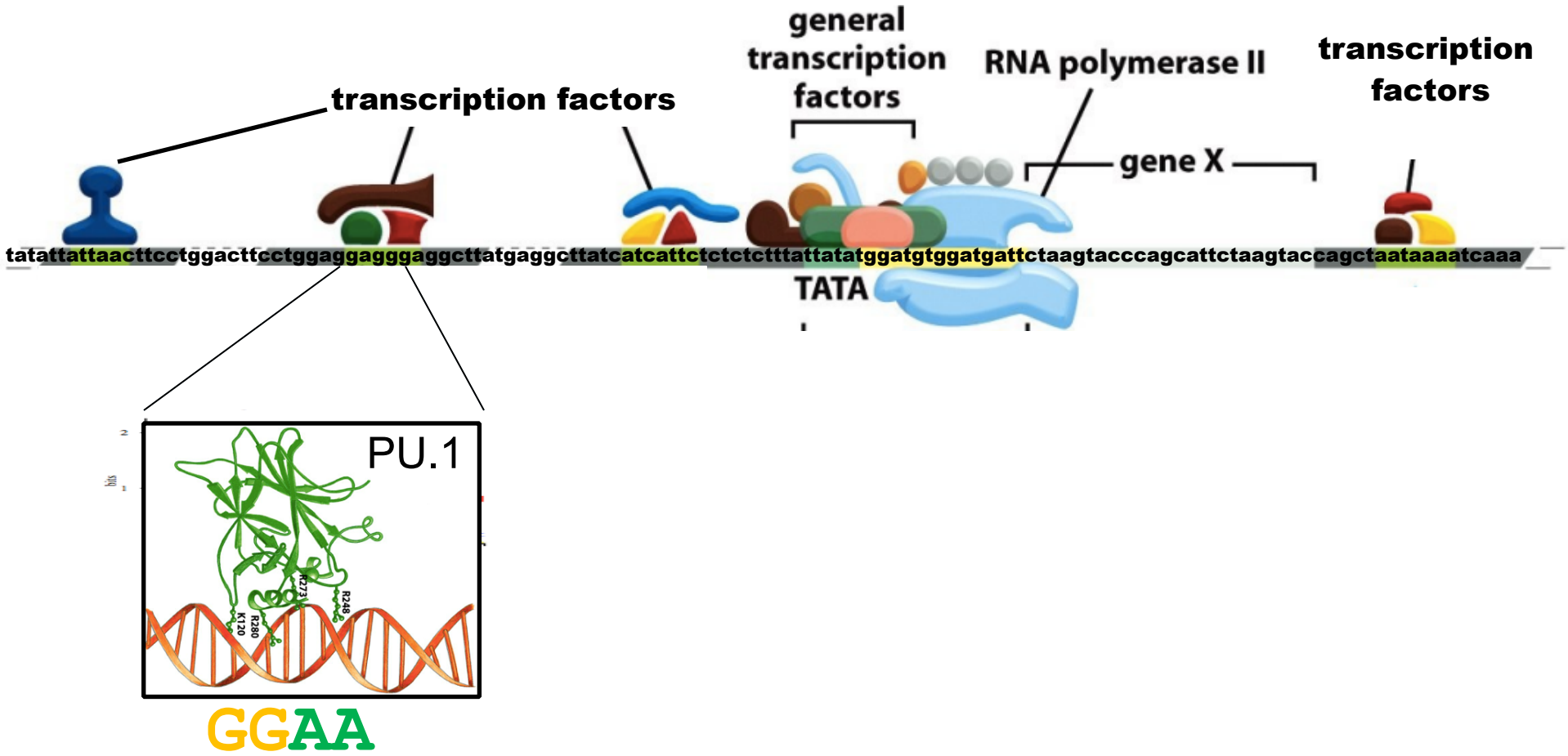
Source: Amit (2016), *Nature Immunoloy.*

Regulatory Control – Transcription Factor Binding



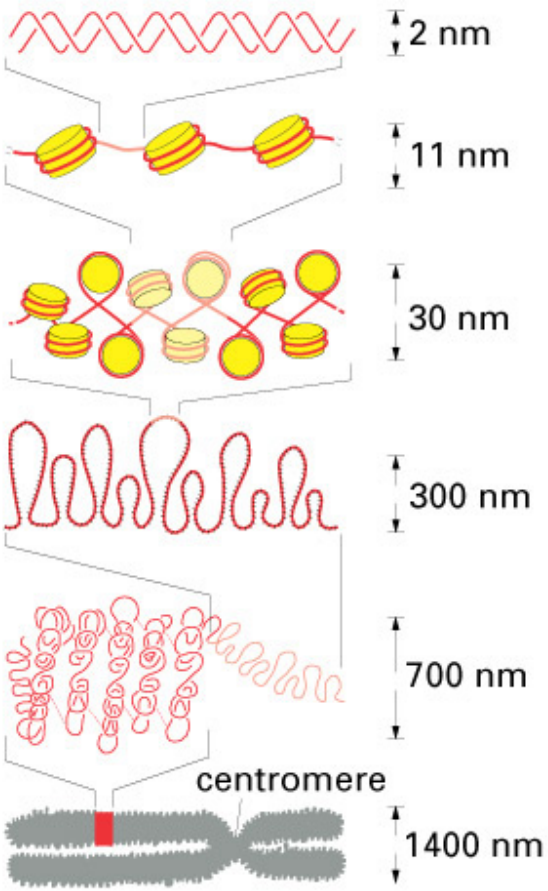
Source: Alberts, B. et al. (2008) Garland Science, 5th ed.

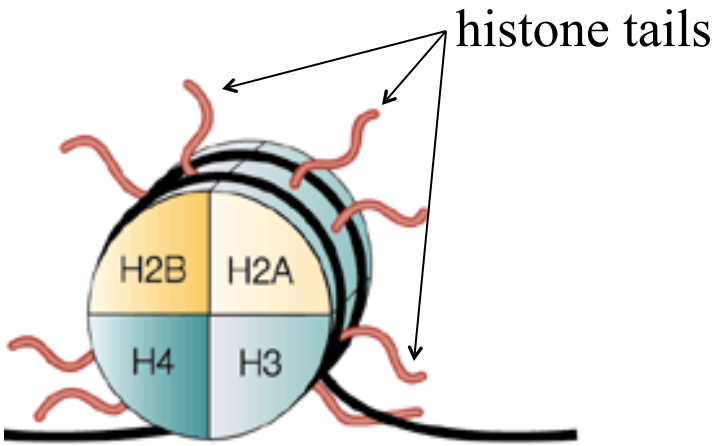
Regulatory Control – Transcription Factor Binding



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

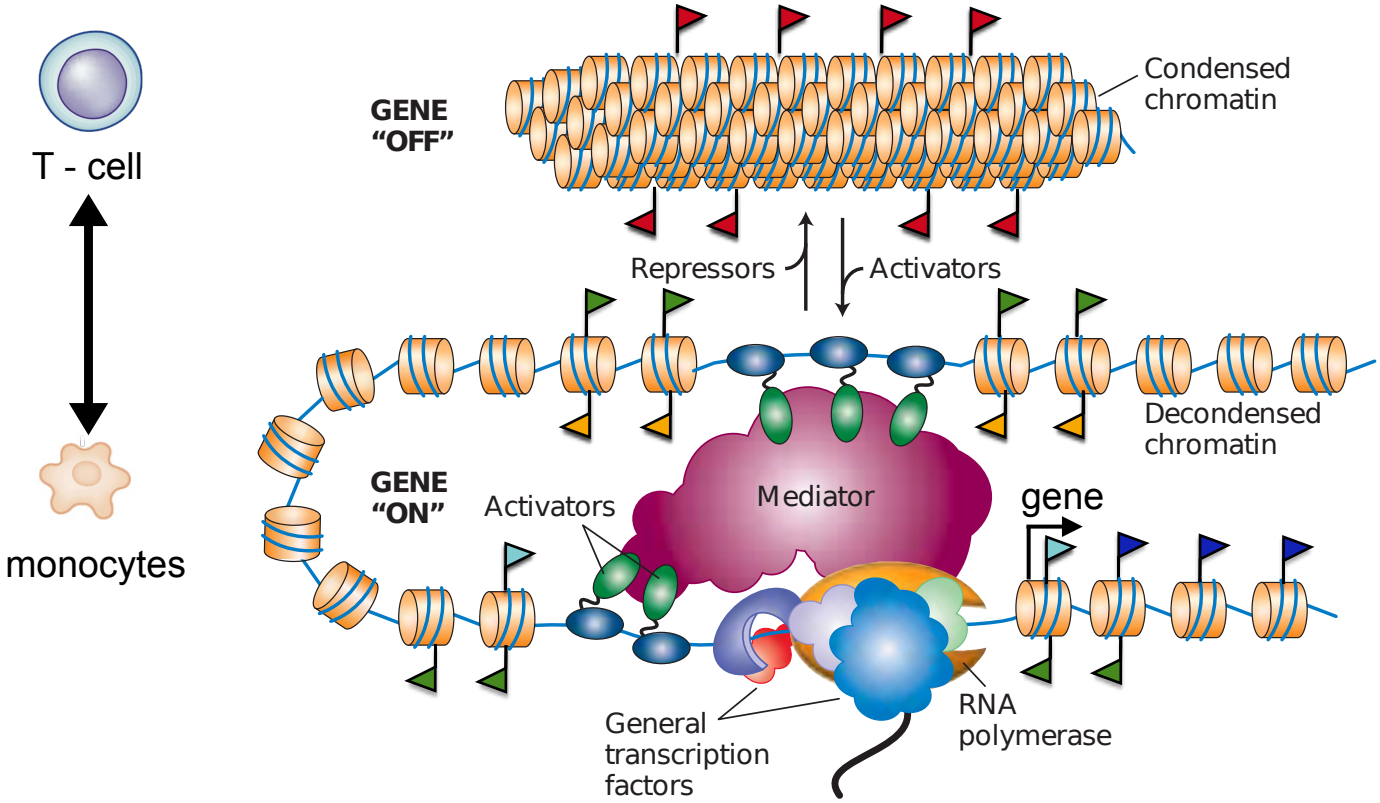
Epigenetics & Histones





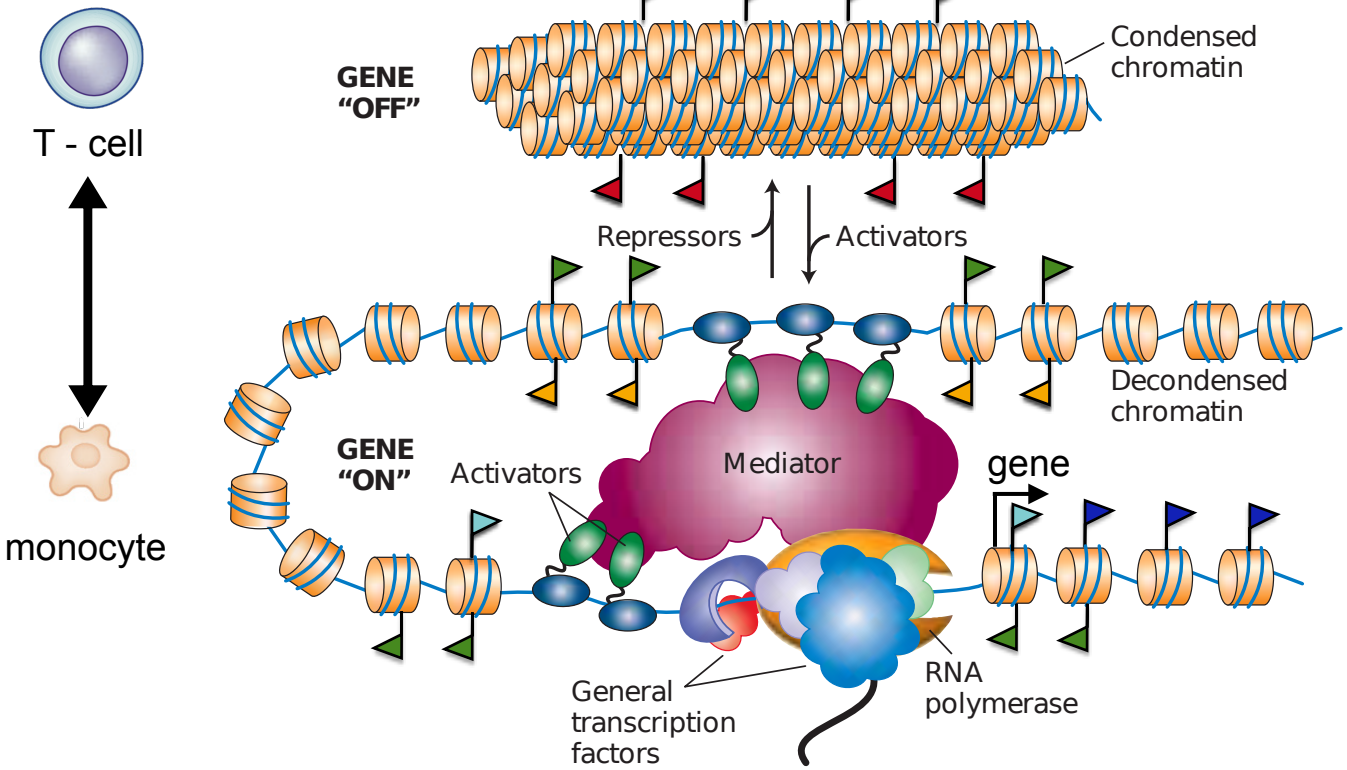
Modification in histone tails
- change strength of DNA binding
- recruit transcription factors

Chromatin, Regulation and Cellular Memory



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

Chromatin & Histone Code



Histone Code

Transcription

H3K79me2, H3k36me3

Active Regions

H3K27ac, H3K9ac

Active Promoters

H3K4me3

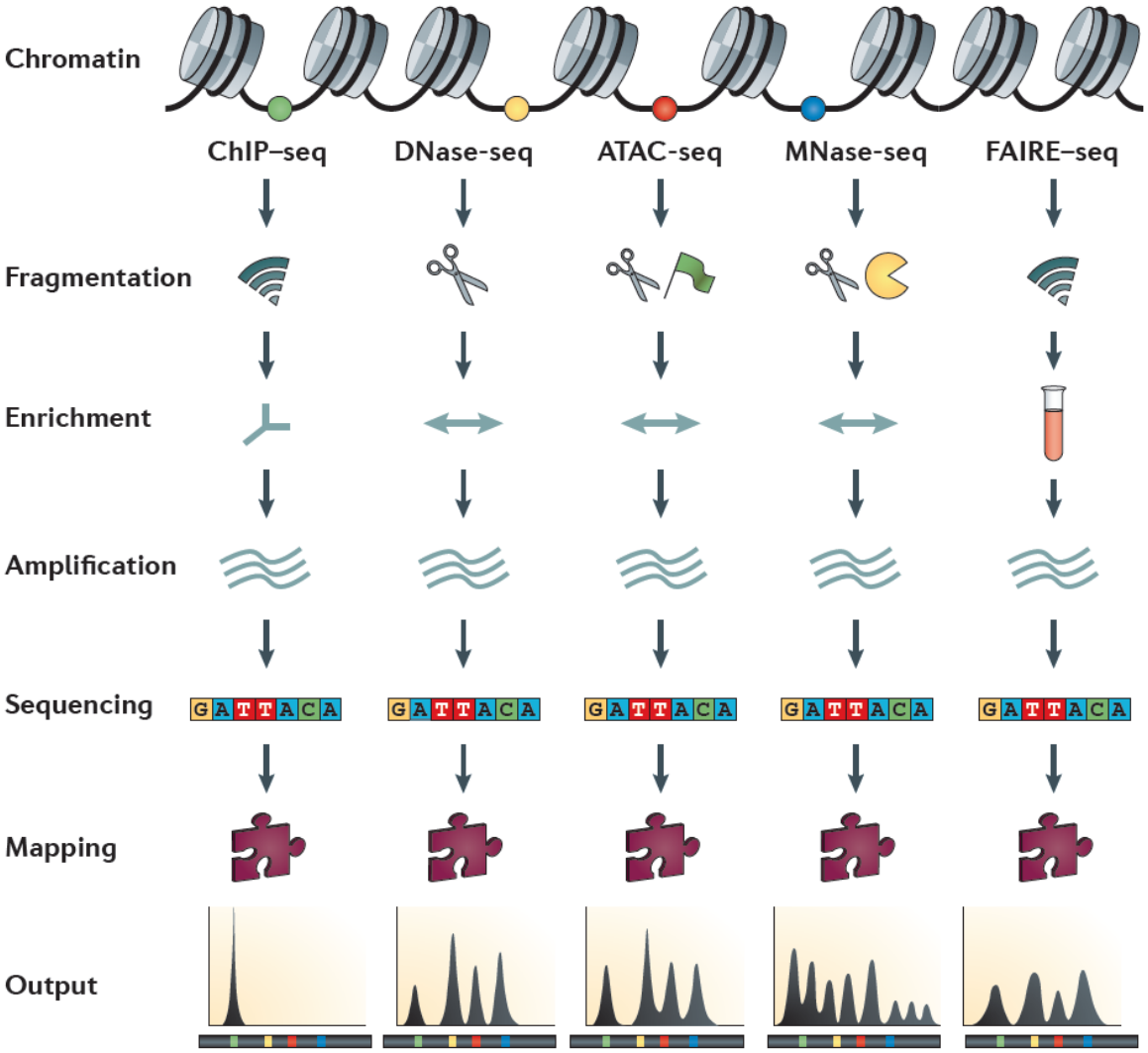
Active Enhancers

H3K4me1

Repressed regions

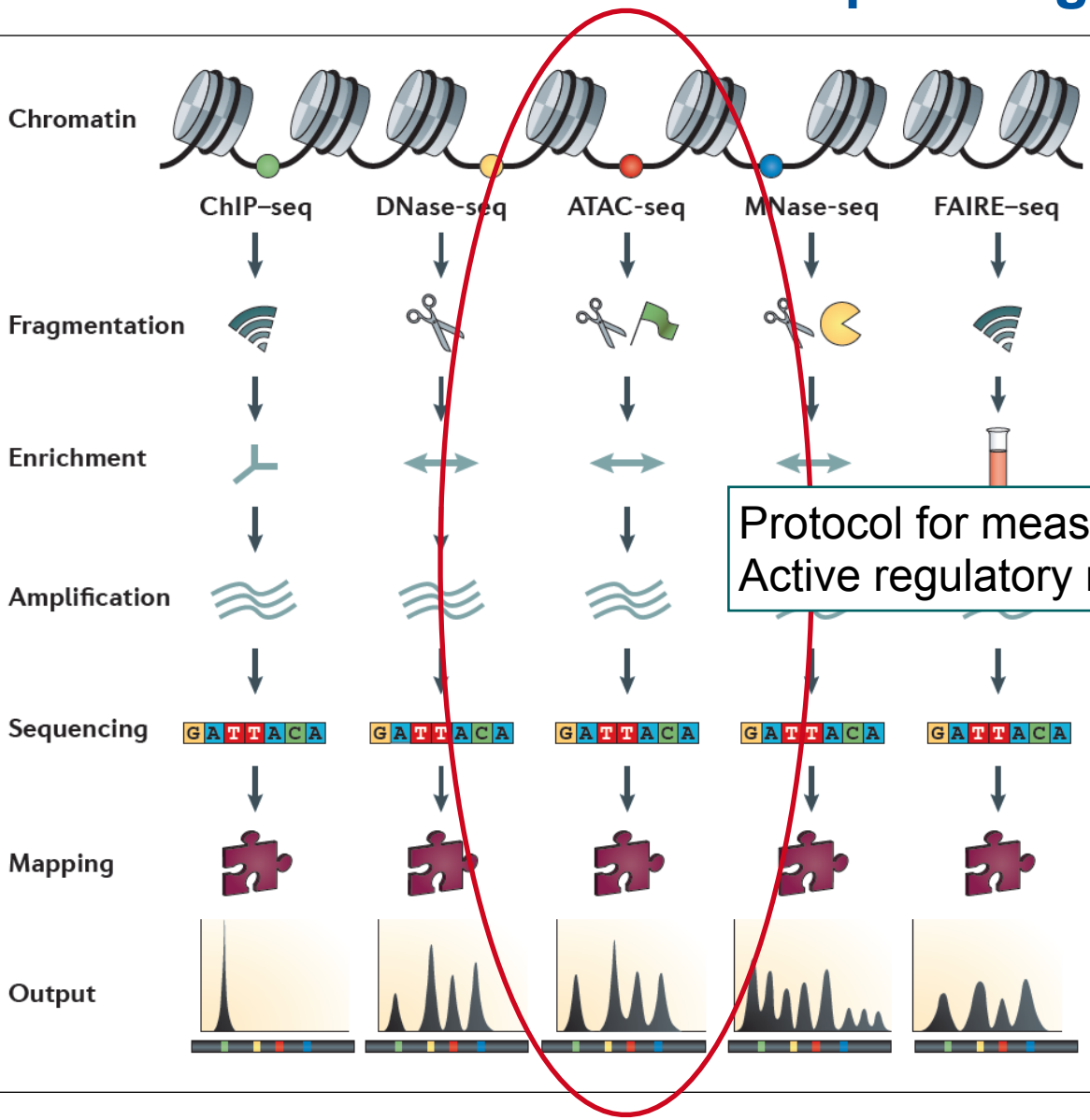
H3K27me3, H3K9me3

Chromatin with Next Generation Sequencing



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

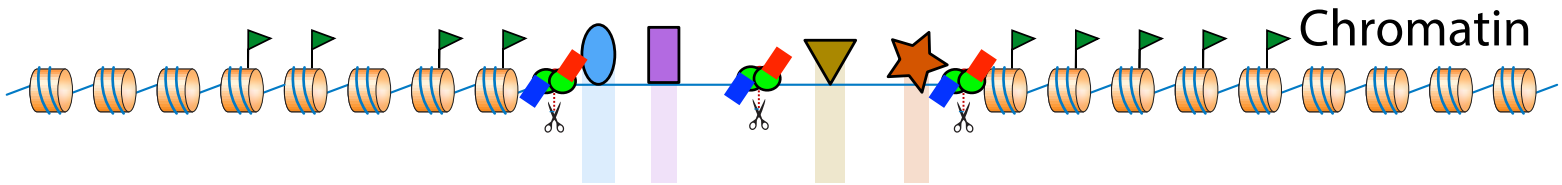
Chromatin with Next Generation Sequencing



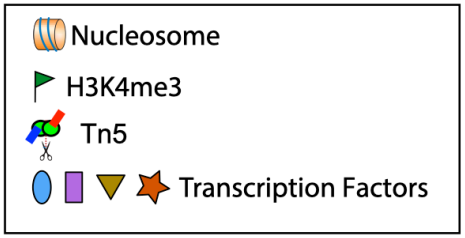
Protocol for measuring open chromatin
Active regulatory regions

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

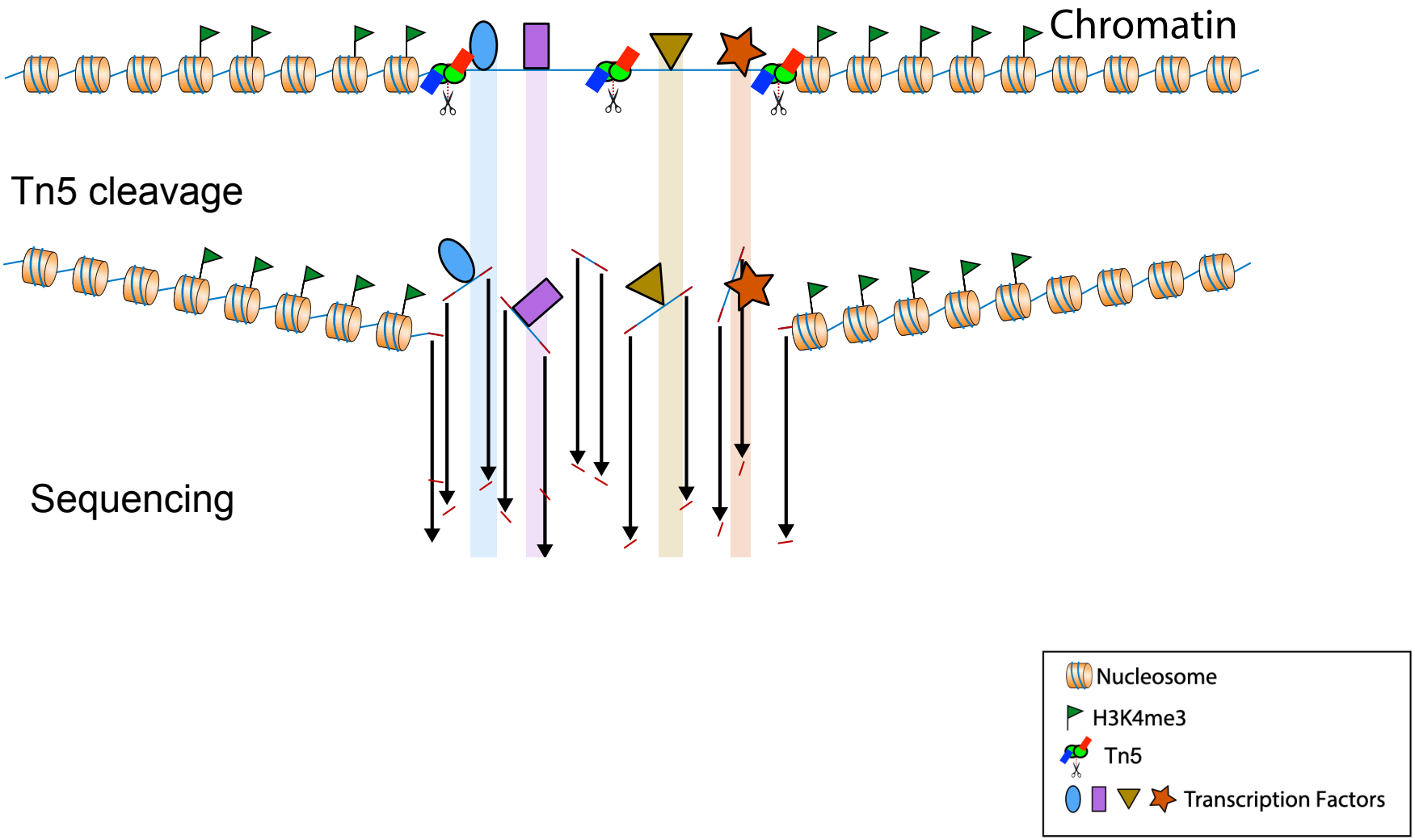
Open Chromatin with ATAC-seq



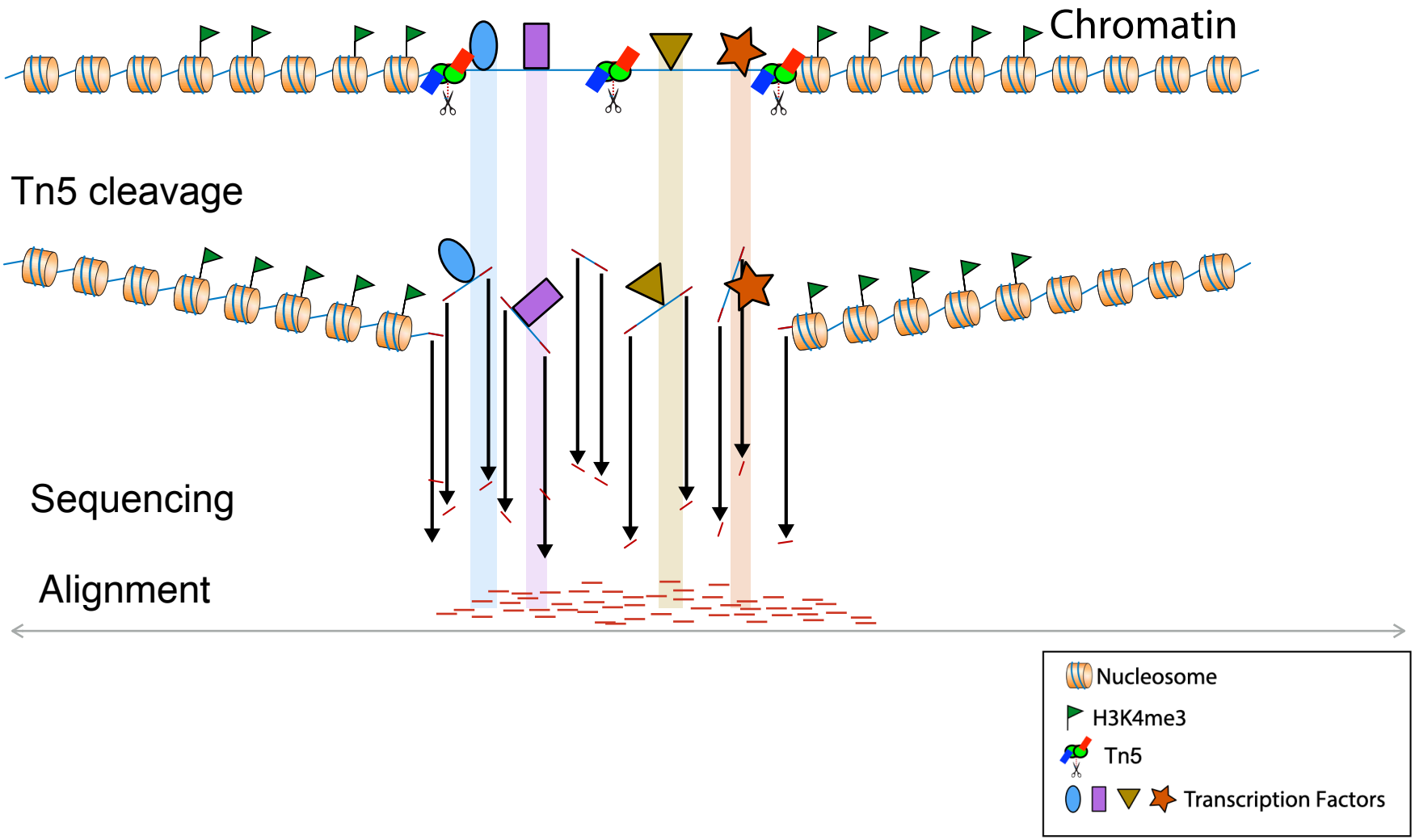
Tn5 cleavage



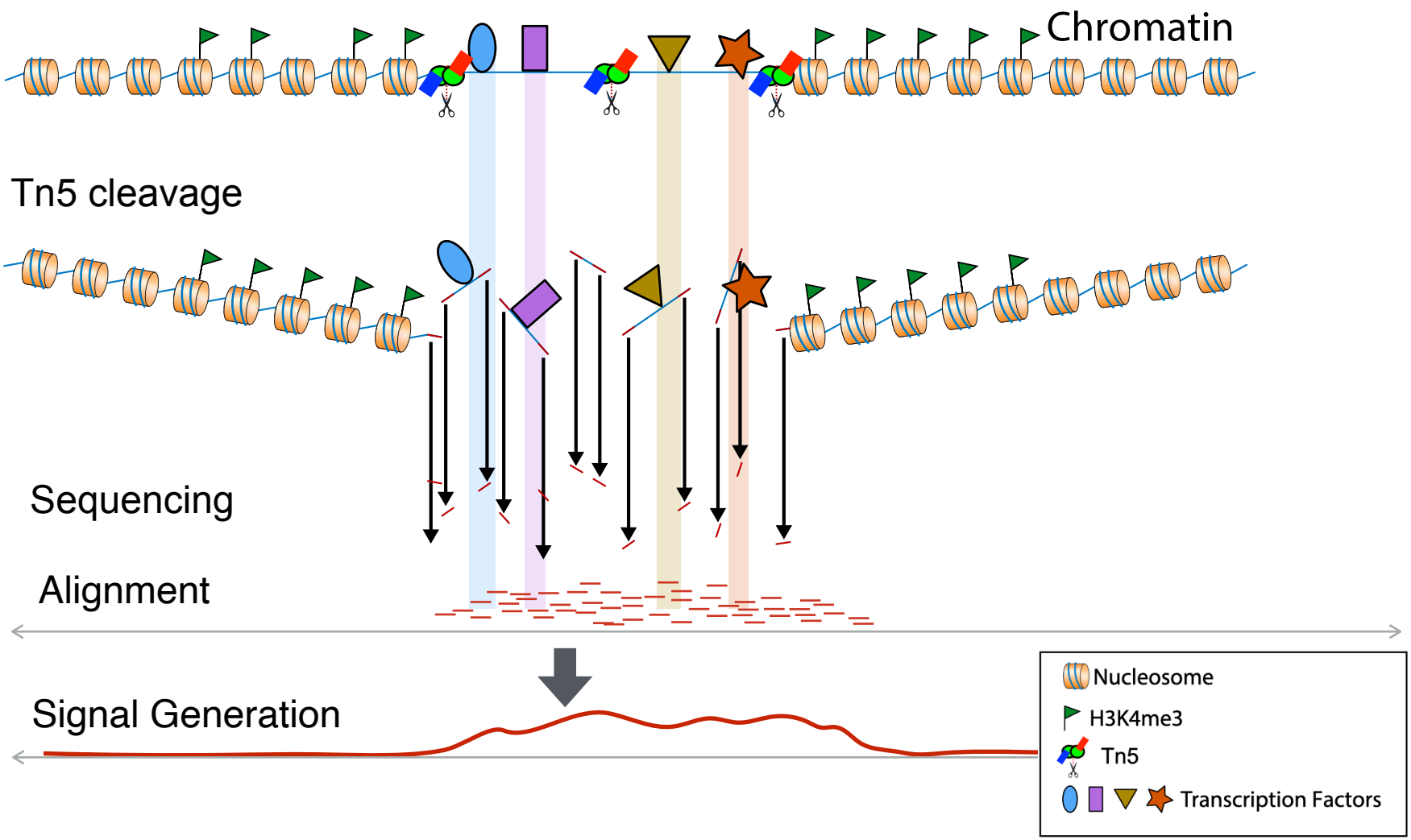
Open Chromatin with ATAC-seq



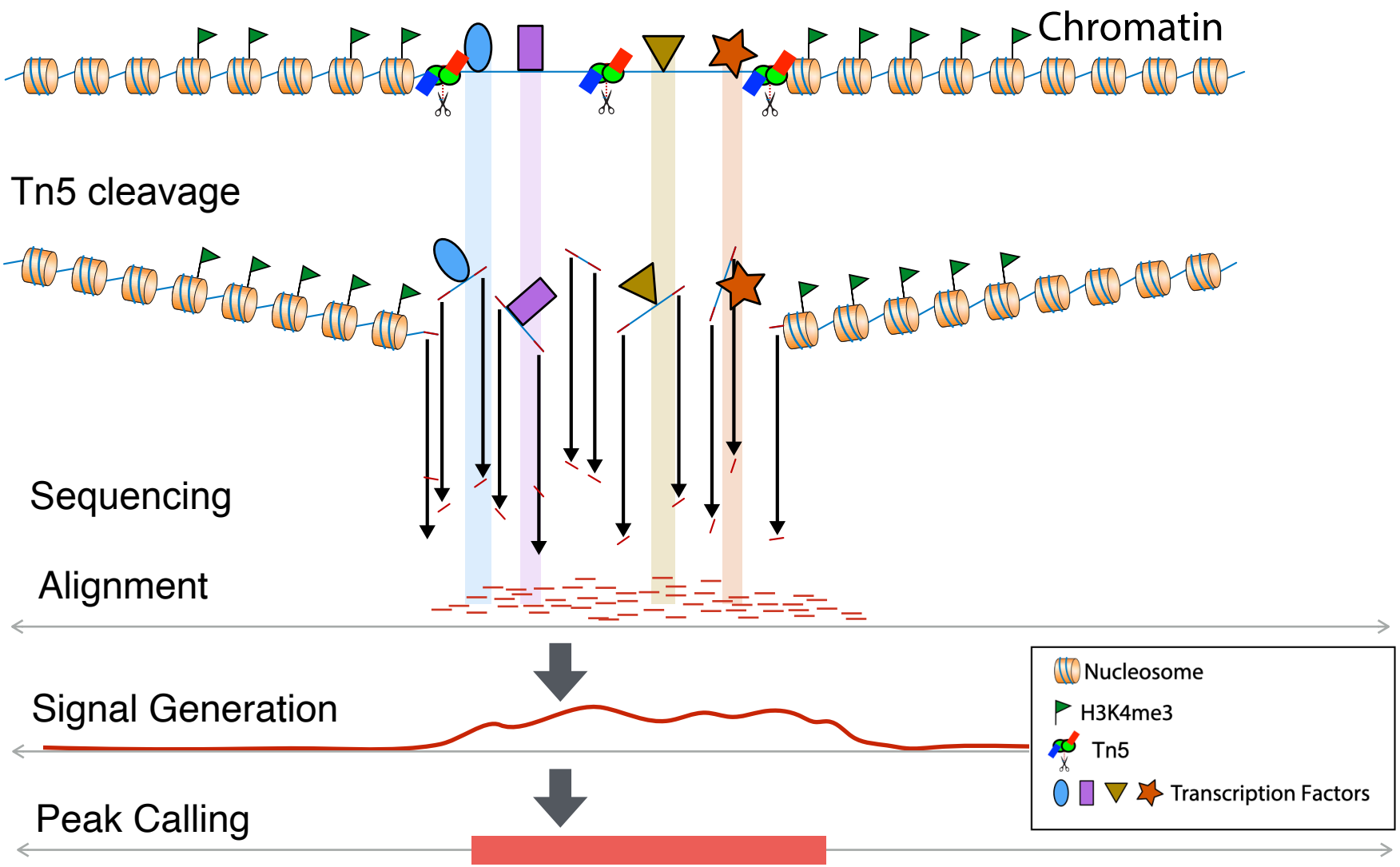
Open Chromatin with ATAC-seq



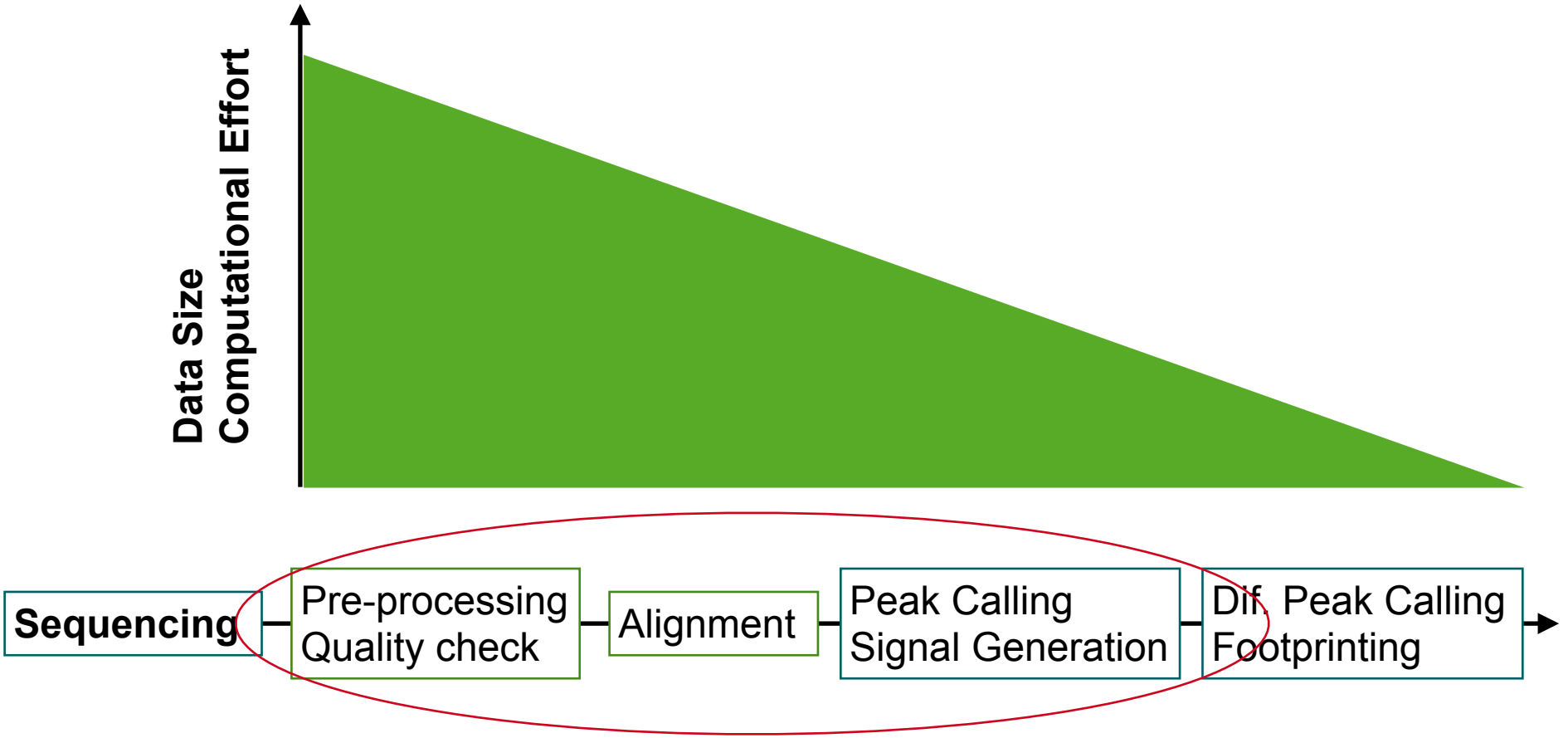
Open Chromatin with ATAC-seq



Open Chromatin with ATAC-seq



Bioinformatics Pipeline / ATAC-seq



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

Peak Calling

Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.

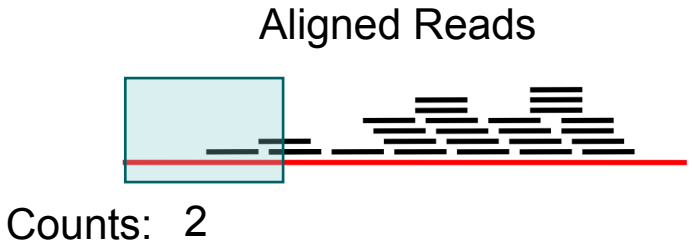
Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change

See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.



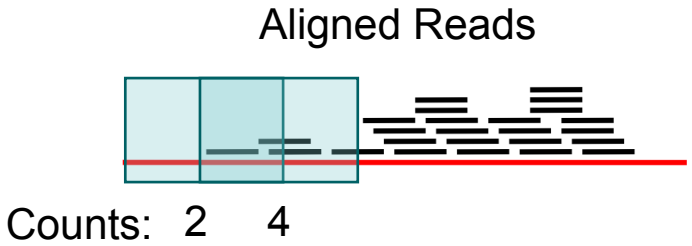
Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change

See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.



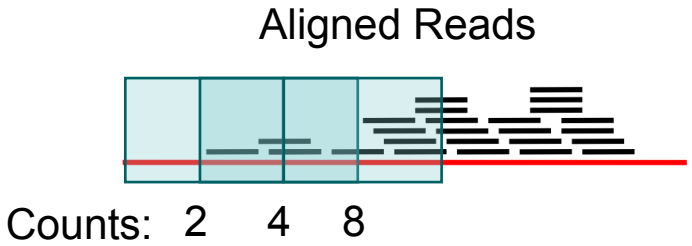
Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change

See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.



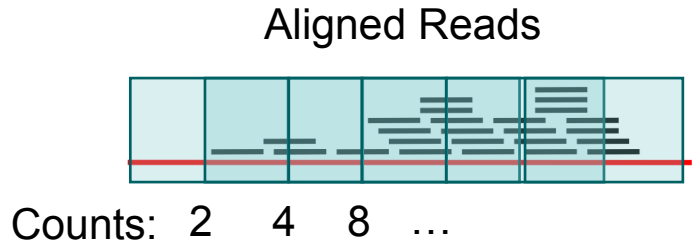
Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change

See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.



Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change

See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

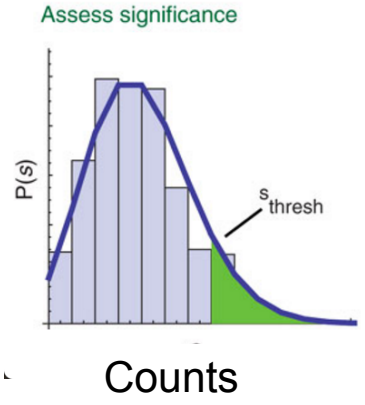
Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.

Aligned Reads

Counts: 2 4 8 …

Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change

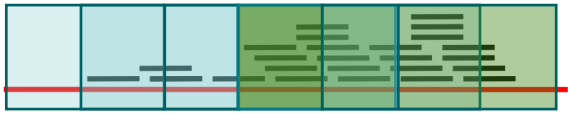


See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.

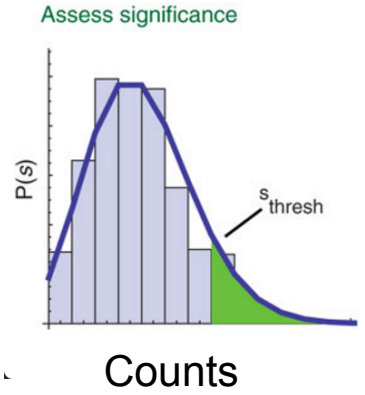
Aligned Reads



Counts: 2 4 8 …

Example of a simple peak caller :

1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads in higher than expected by change



See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

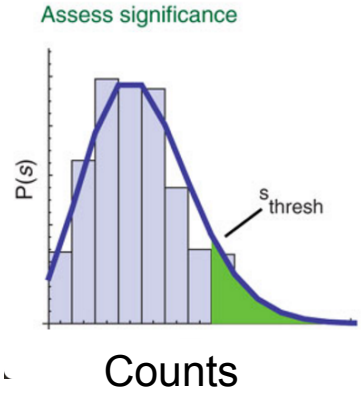
Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.

Problems:

- which window size to use?
- proper quantification of read counts require several further steps: CG bias correction, duplicated reads, mappability, **fragment size**, …

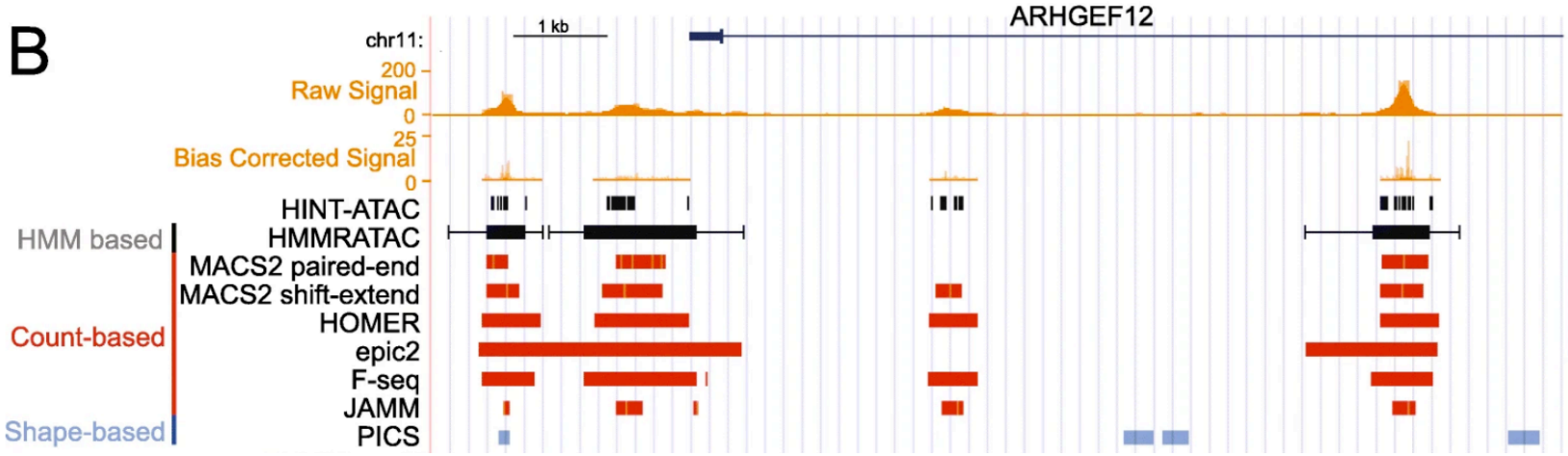
unts: 2 4 8 …

if the number of reads in higher than expected by change



See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

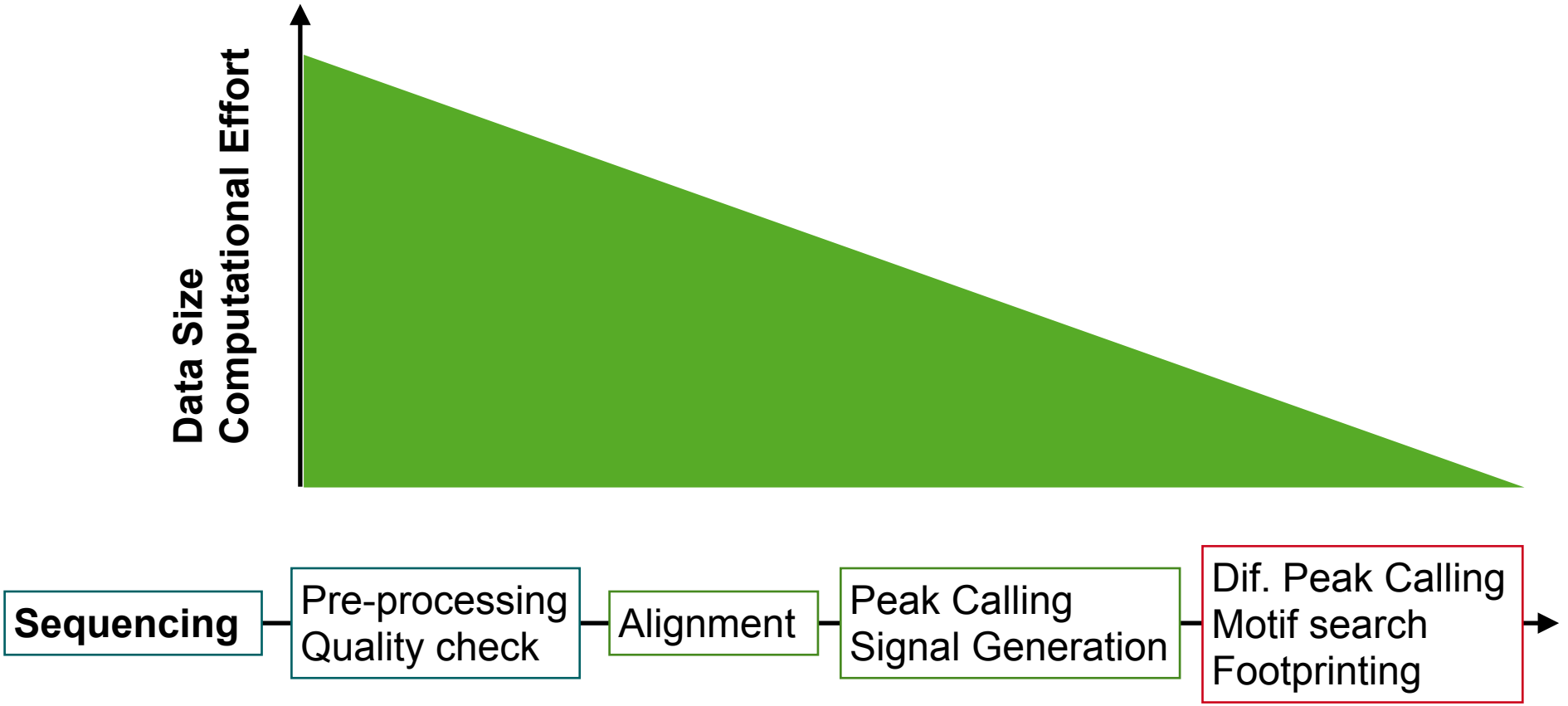
Peak calling in ATAC-seq



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

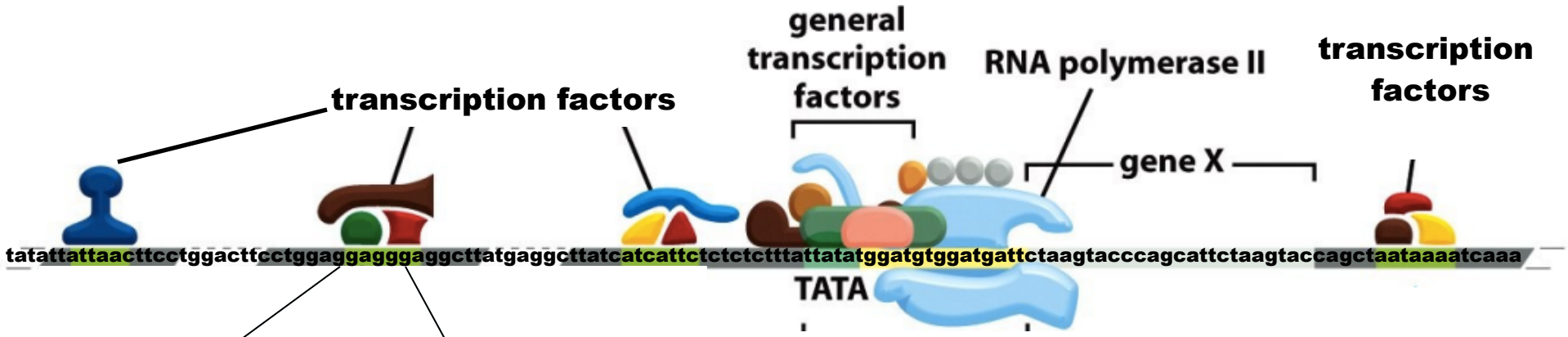
Source: Yan, Genome Biology, 2020.

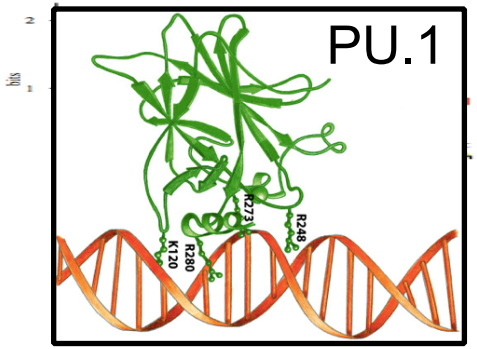
Bioinformatics Pipeline / ATAC-seq



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

Motif Search – Computational Approach





Problems

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

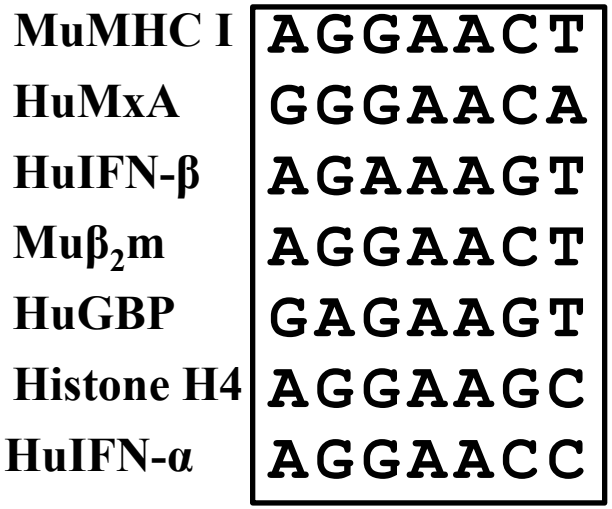
Too many false positive predictions!

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

Model for DNA-protein binding

PU.1 binding sites

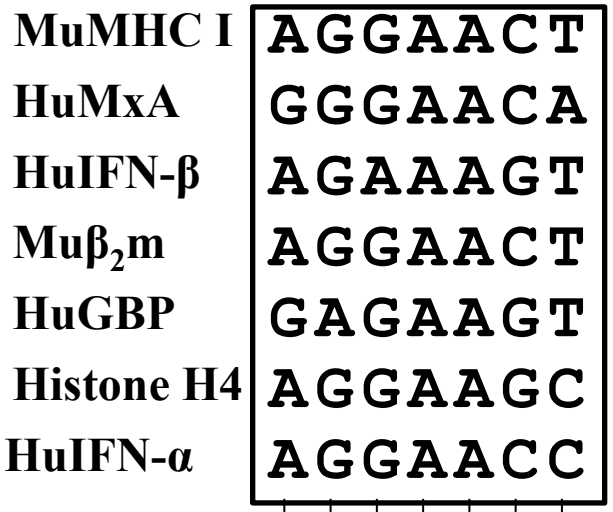
Kanno, Y. et al. (2005) Immune Cell-Specific Amplification of Interferon Signaling by the IRF-4/8-PU.1 Complex.



Model for DNA-protein binding

PU.1 binding sites →

Kanno, Y. et al. (2005) Immune Cell-Specific Amplification of Interferon Signaling by the IRF-4/8-PU.1 Complex.



PU.1 Position Weight Matrix (PWM) →

A	5	1	1	7	7	3	1
C	0	0	0	0	0	0	2
G	2	6	6	0	0	4	0
T	0	0	0	0	0	0	4

Model for DNA-protein binding

PU.1 binding sites →

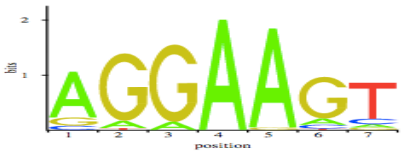
Kanno, Y. et al. (2005) Immune Cell-Specific Amplification of Interferon Signaling by the IRF-4/8-PU.1 Complex.

MuMHC I	AGGAACT
HuMxA	GGGAACA
HuIFN-β	AGAAAGT
Muβ$_2$m	AGGAACT
HuGBP	GAGAAGT
Histone H4	AGGAAGC
HuIFN-α	AGGAACC

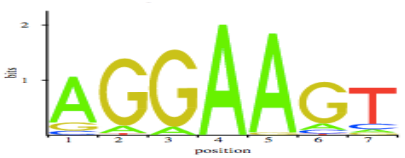
PU.1 Position Weight Matrix (PWM) →

A	5	1	1	7	7	3	1
C	0	0	0	0	0	0	2
G	2	6	6	0	0	4	0
T	0	0	0	0	0	0	4

PU.1 Logo →

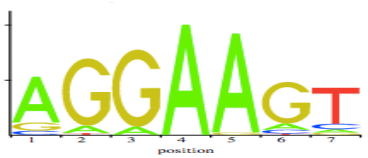


Motif Search

PU.1 PWM 

Genome `TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA`

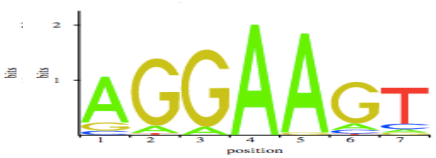
Motif Search

PU.1 PWM 

Genome TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA

Score 10.06

Motif Search

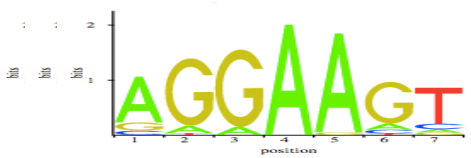
PU.1 PWM 

Genome TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA

Score 10.06

4.81

Motif Search

PU.1 PWM 

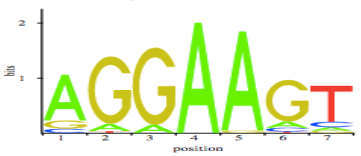
Genome **TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA**

Score 10.06 | 3.19

4.81

Motif Search

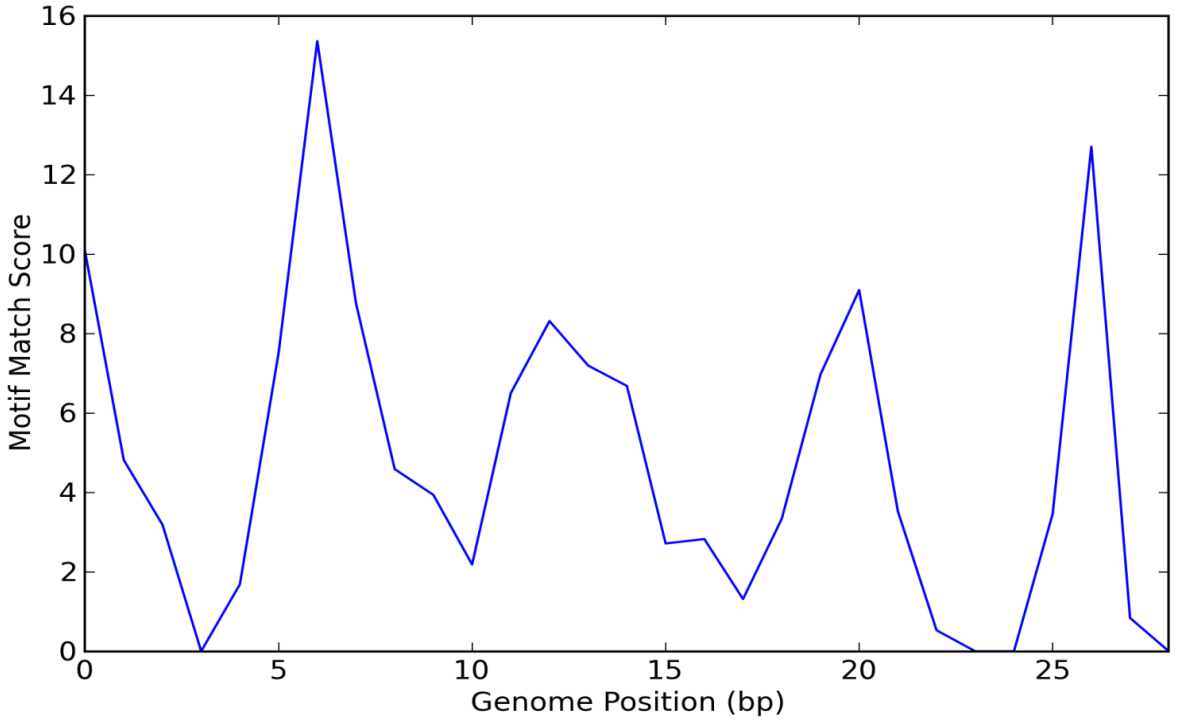
PU.1 PWM



Genome **TATCTTTGGAAGTGAAACTACTATCCTGAAACTCGAA**

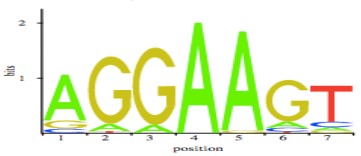
Score 10.06 | 3.19

4.81 . . .



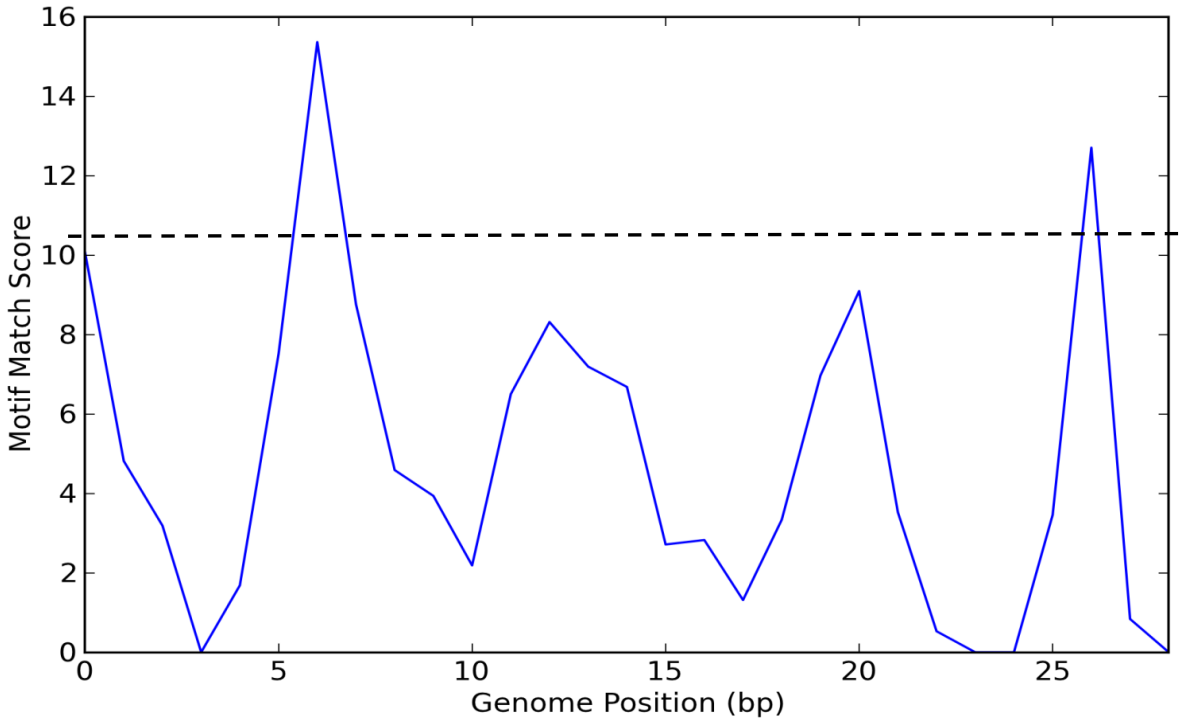
Motif Search

PU.1 PWM



Genome TATCTTTGGAAGTGAAACTACTATCCTGAAAGTCGAA

Score 10.06 | 3.19 | 4.81 ...



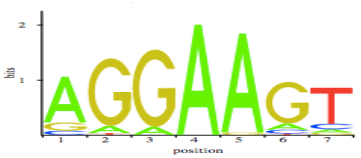
Statistical Test

FDR = 1×10^{-4}

FDR- False Discovery Rate

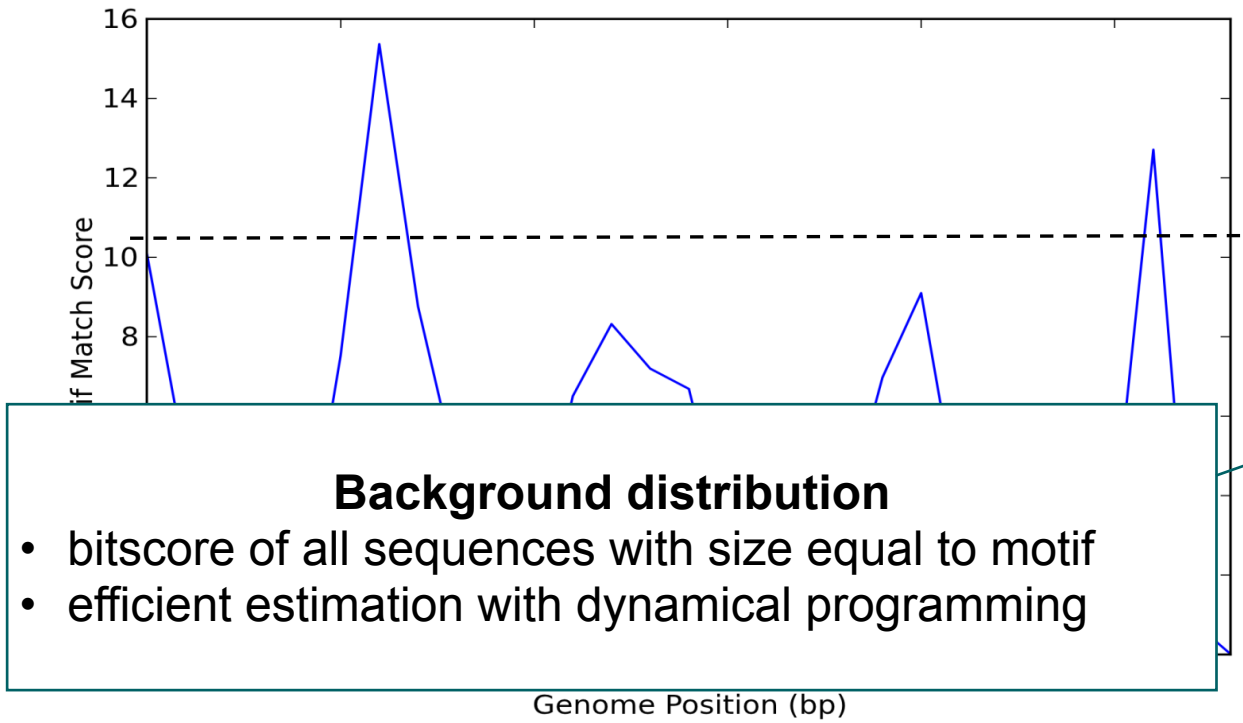
Motif Search

PU.1 PWM



Genome TATCTTTGGAAGTGAAACTACTATCCTGAAAGTCGAA

Score 10.06 | 3.19 | 4.81 ...



Statistical Test

FDR = 1×10^{-4}

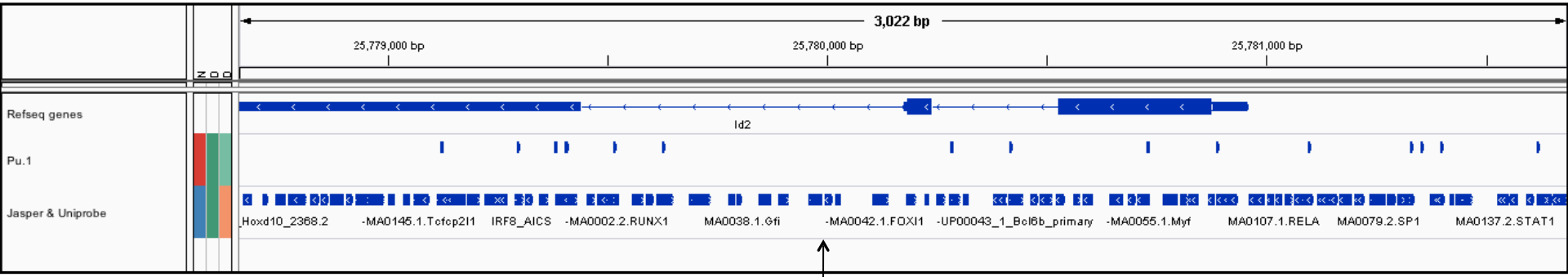
FDR- False Discovery Rate

Background distribution

- bitscore of all sequences with size equal to motif
- efficient estimation with dynamical programming

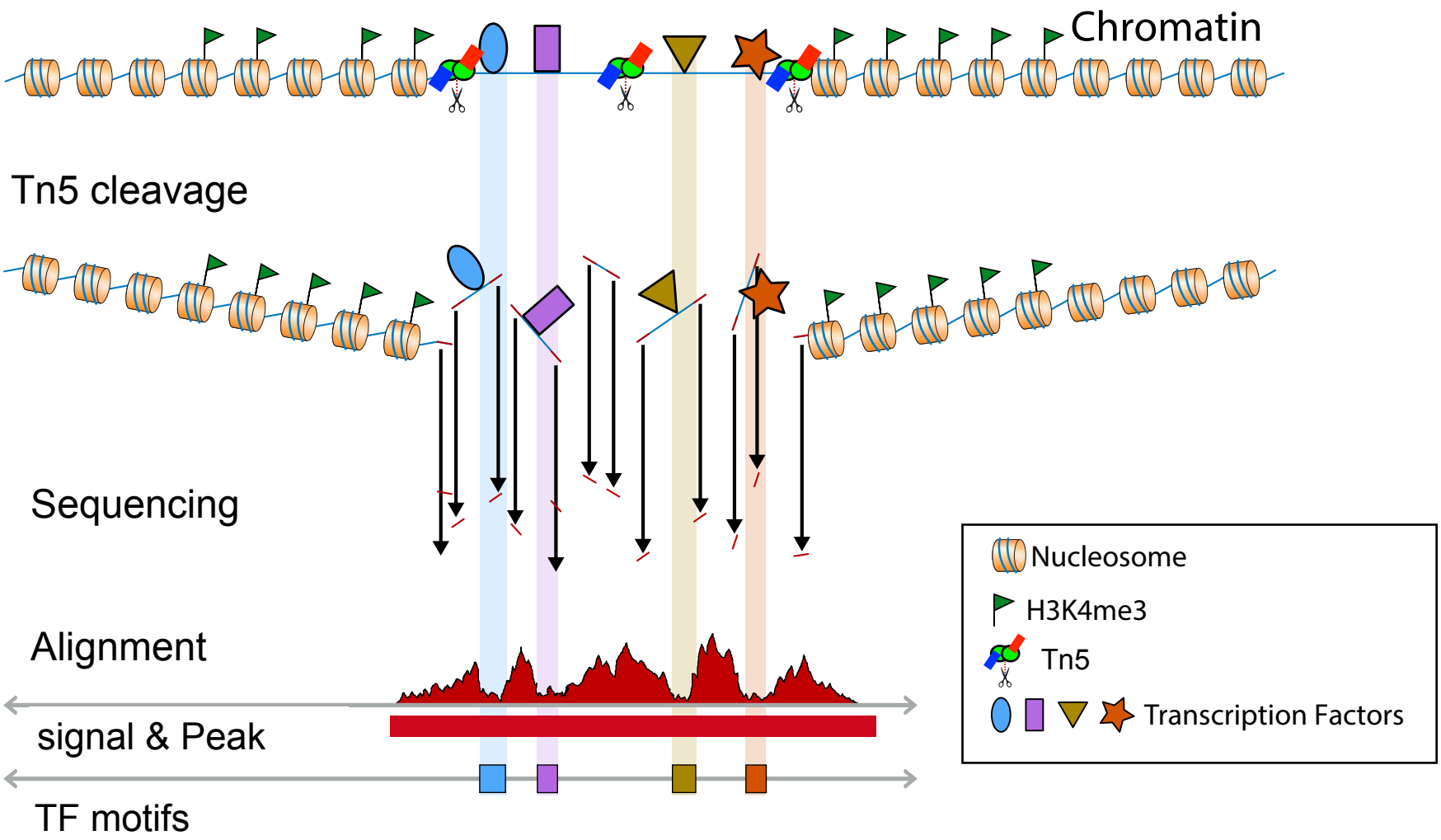
Example: Binding sites in ID2

Motif search for binding sites with 536 PWMs (Jaspar & Uniprobe) and FDR=0,01

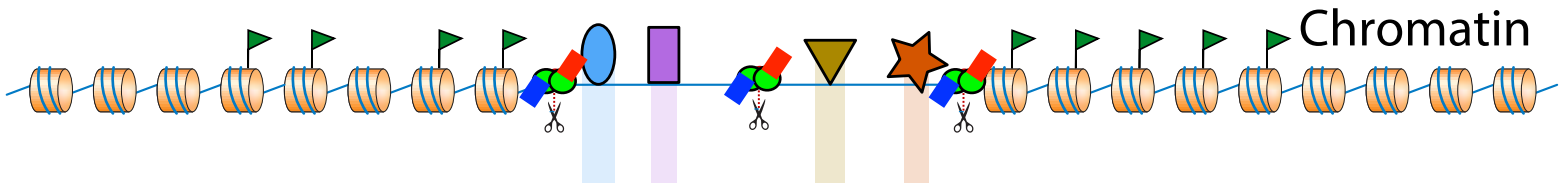


> 3000 predicted binding sites

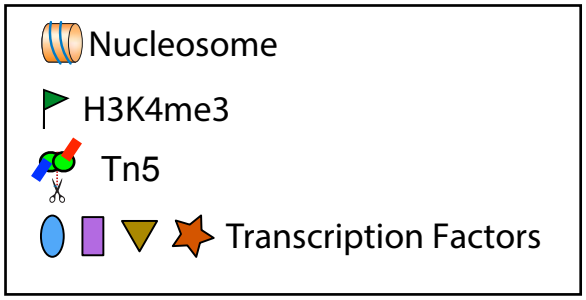
Open Chromatin and TF Binding with ATAC-seq



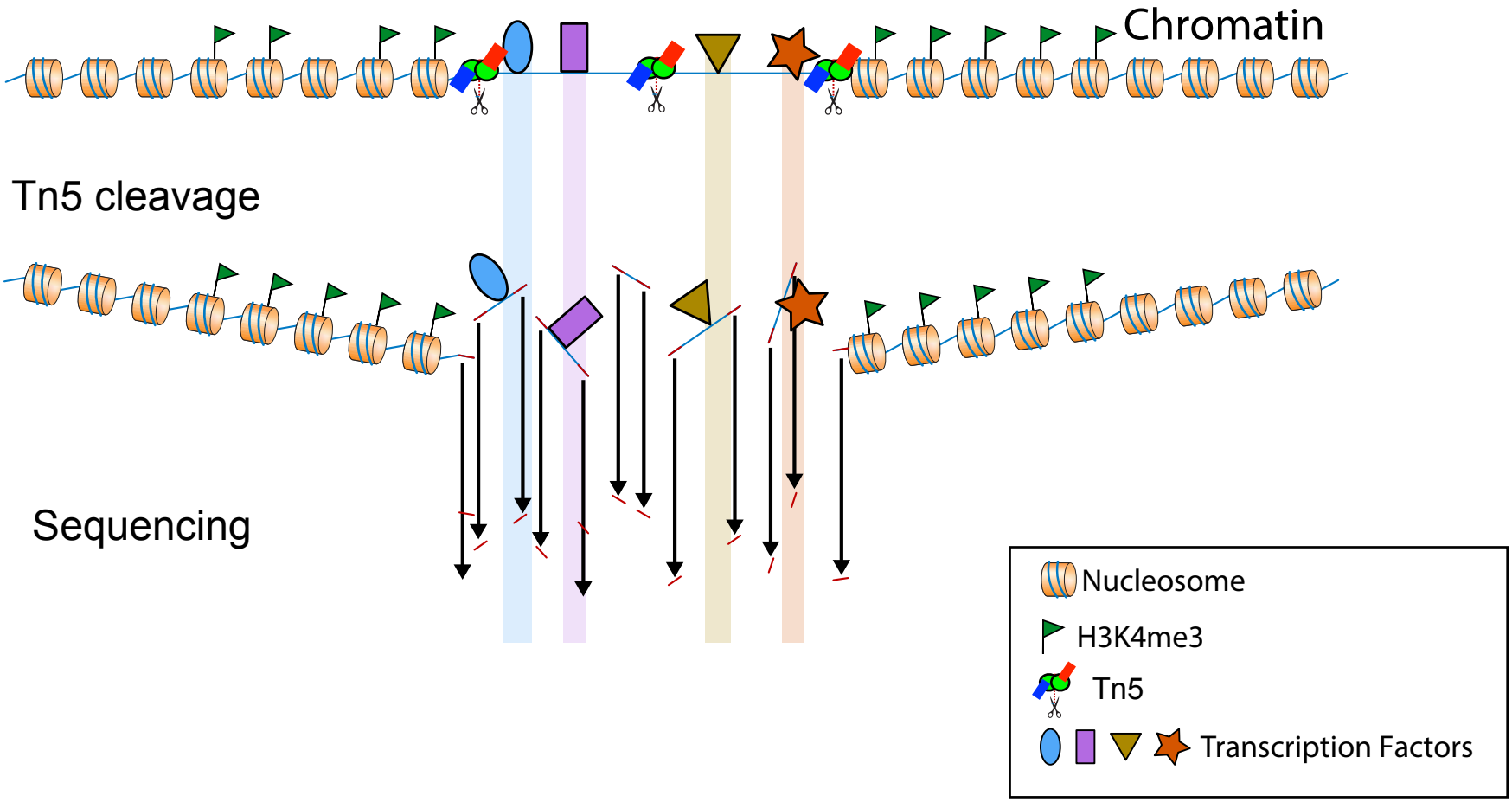
Open Chromatin and TF Binding with ATAC-seq



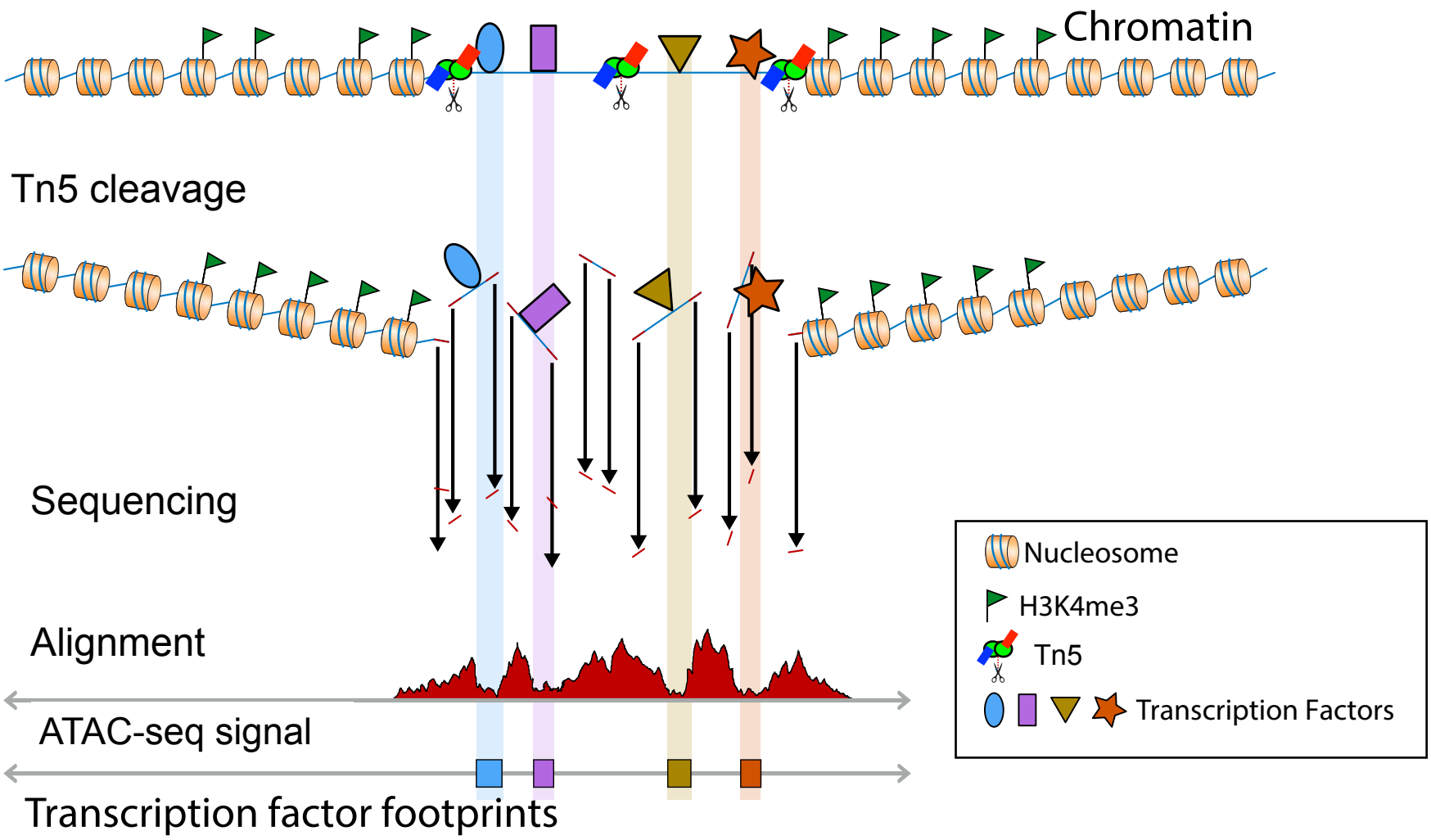
Tn5 cleavage



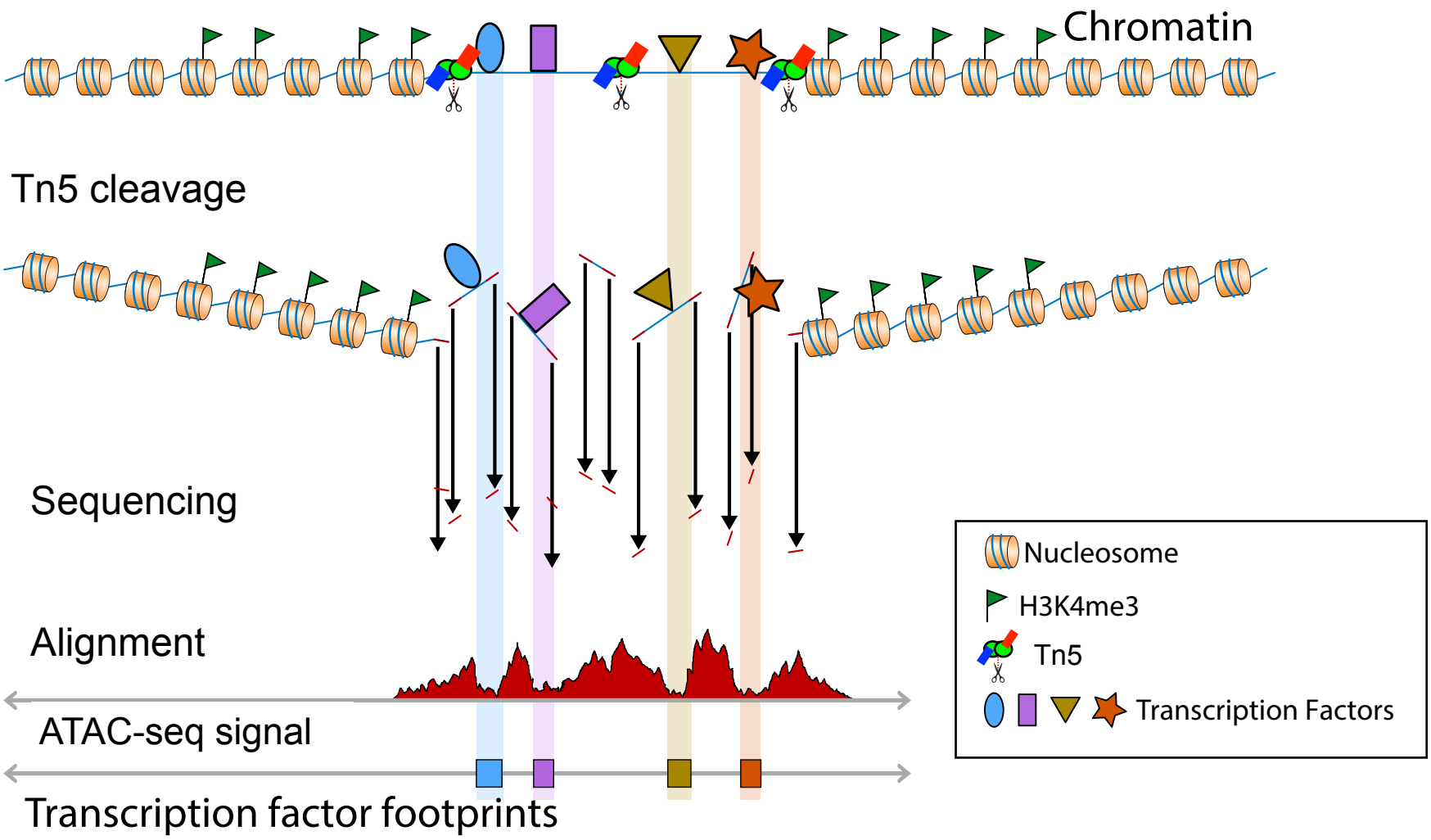
Open Chromatin and TF Binding with ATAC-seq



Open Chromatin and TF Binding with ATAC-seq

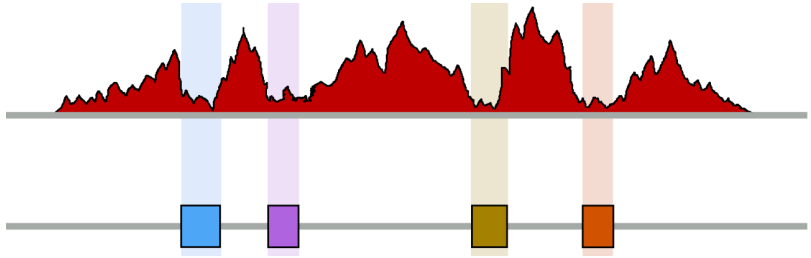


Open Chromatin and TF Binding with ATAC-seq



Digital Footprinting

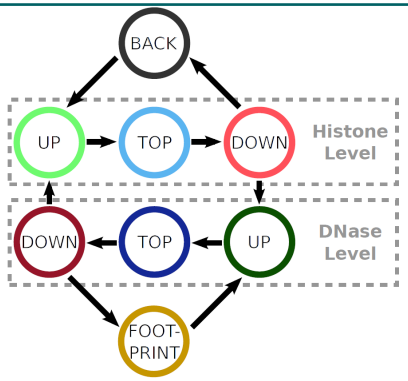
Problem definition: Find genomic regions (of small size) with depletion in DNase-seq signals



Computational Footprinting

HINT (Hmm-based IdeNtification of Transcription factor footprints)

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

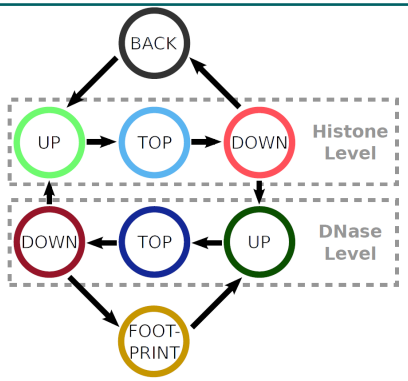


Gusmao EG *et. al*, (2014), Bioinformatics, 30(22):3143-51.
Gusmao EG *et. al,* (2016), Nature Methods, 13, 303–309.
Li et al. (2019), Genome Biology, 20:45.

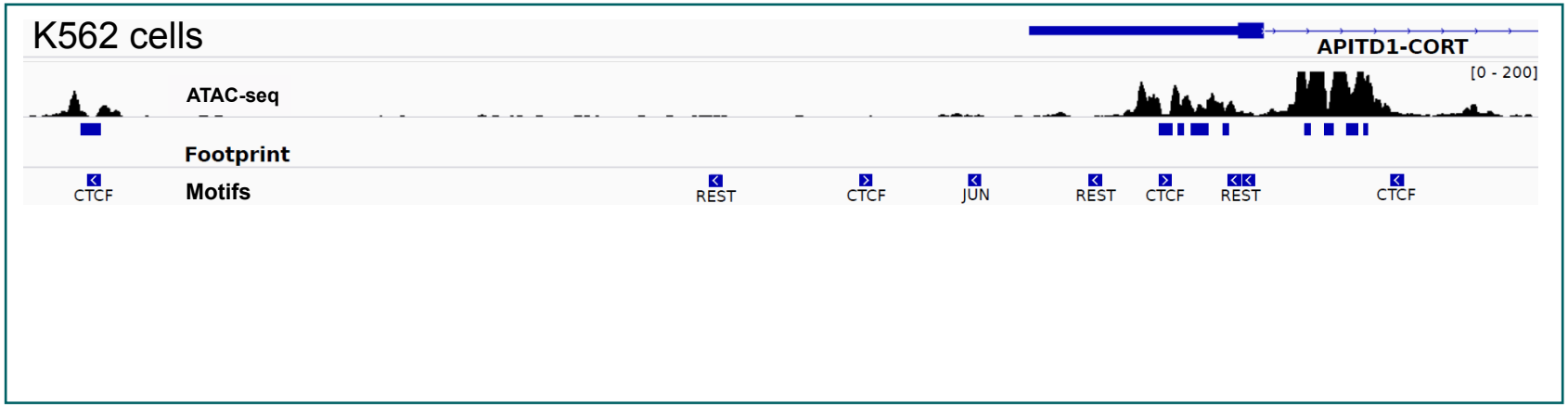
Computational Footprinting

HINT (Hmm-based IdeNtification of Transcription factor footprints)

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



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.

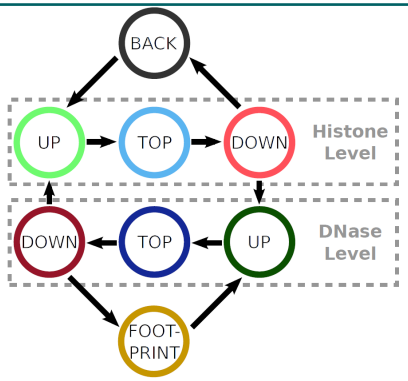
Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

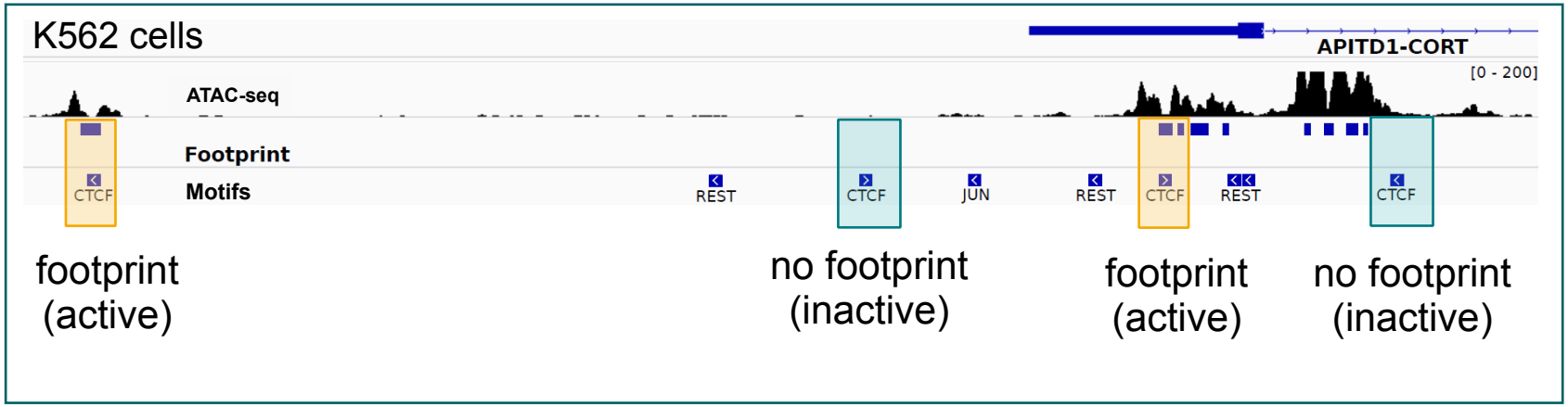
Computational Footprinting

HINT (Hmm-based IdeNtification of Transcription factor footprints)

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



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.

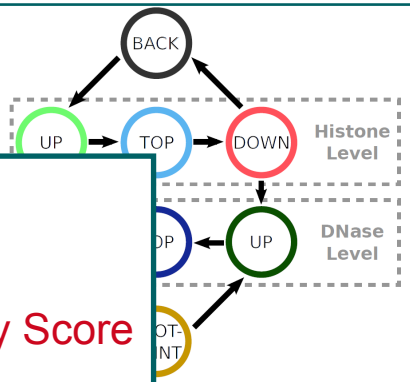
Institute for Computational Genomics
01011011010
10100100101

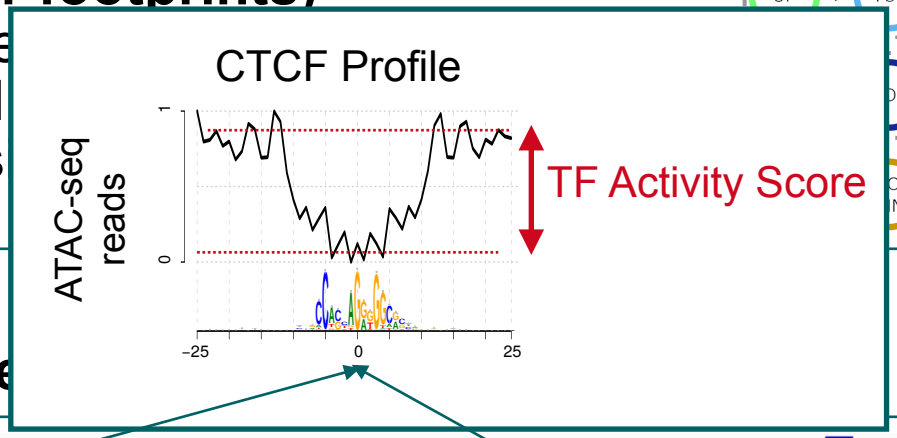
RWTH AACHEN UNIVERSITY

Computational Footprinting

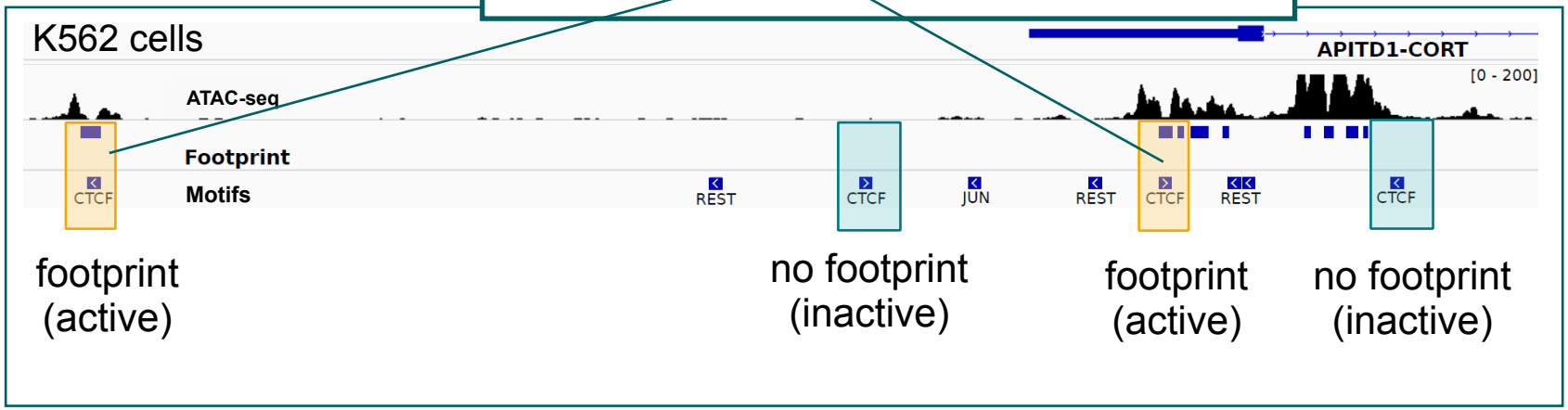
HINT (Hmm-based IdeNtification of Transcription factor footprints)

- generate normalize
- trained with limited
- scan multivariate s





Prediction Example

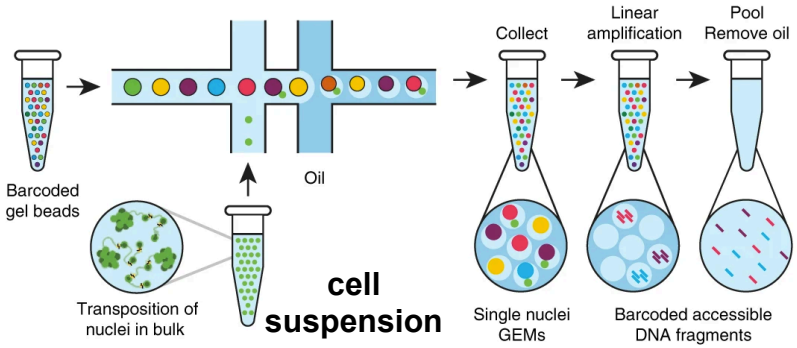


Gusmao EG *et. al*, (2014), Bioinformatics, 30(22):3143-51.
Gusmao EG *et. al,* (2016), Nature Methods, 13, 303–309.
Li et al. (2019), Genome Biology, 20:45.

Institute for Computational Genomics
RWTH AACHEN UNIVERSITY

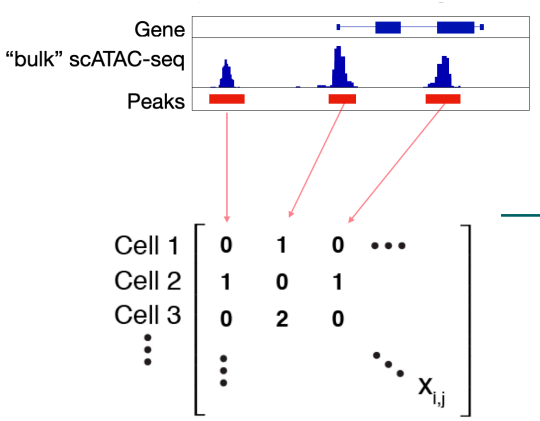
Open chromatin with scATAC-seq

Droplet based scATAC-seq

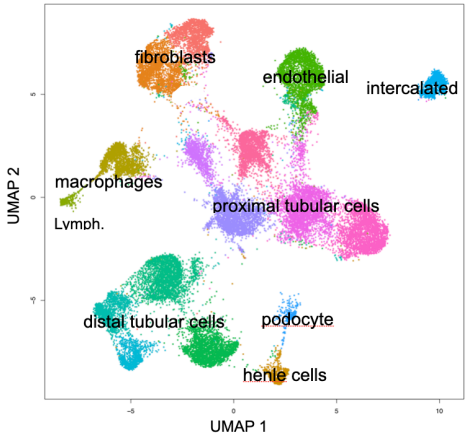


Adapted from Satpathy,
Nature biotechnology, 2019

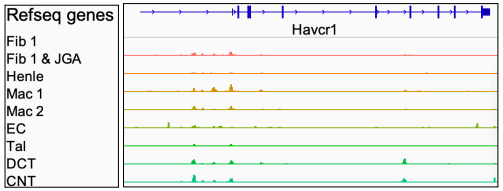
Open Chromatin Matrix

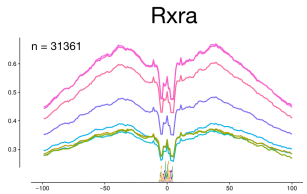


UMAP / Clustering



cluster pseudo bulk ATAC-seq





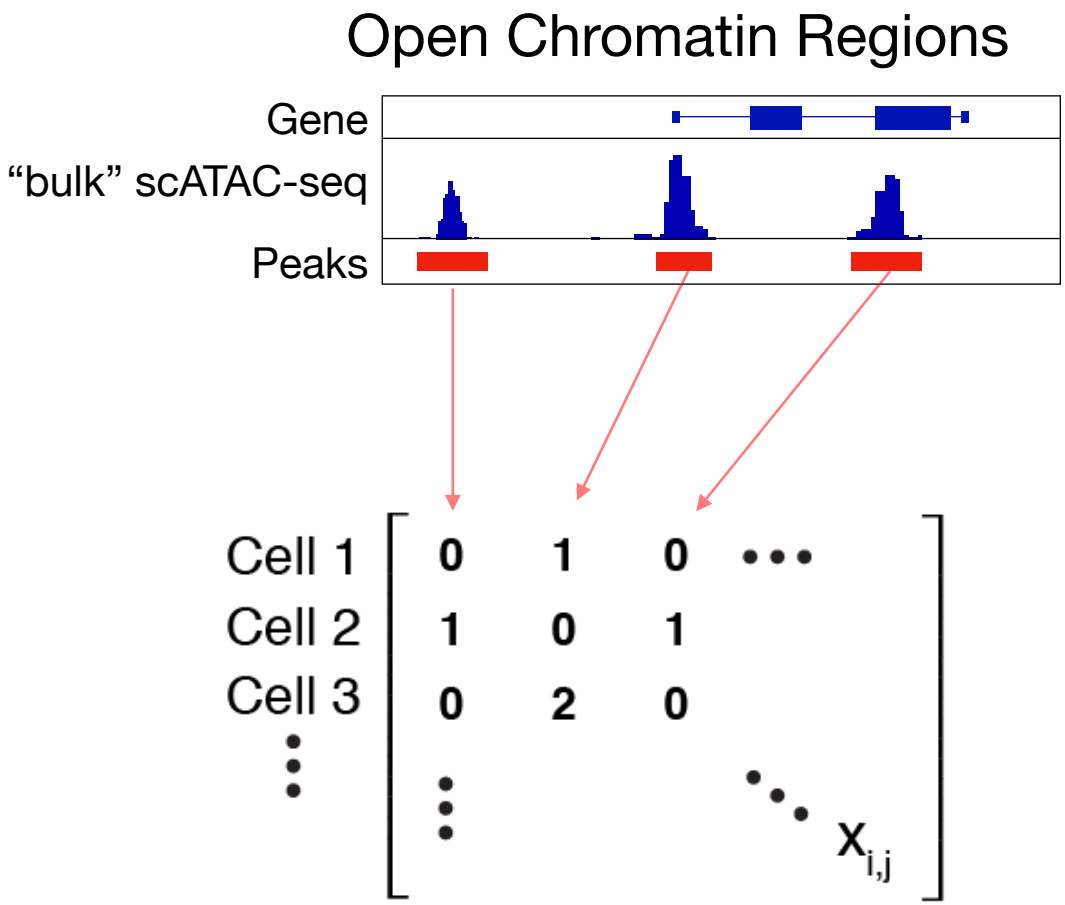
Footprinting & TF Activity

Li, …, **Kramann, Costa**, Biorvx, https://doi.org/10.1101/865931.

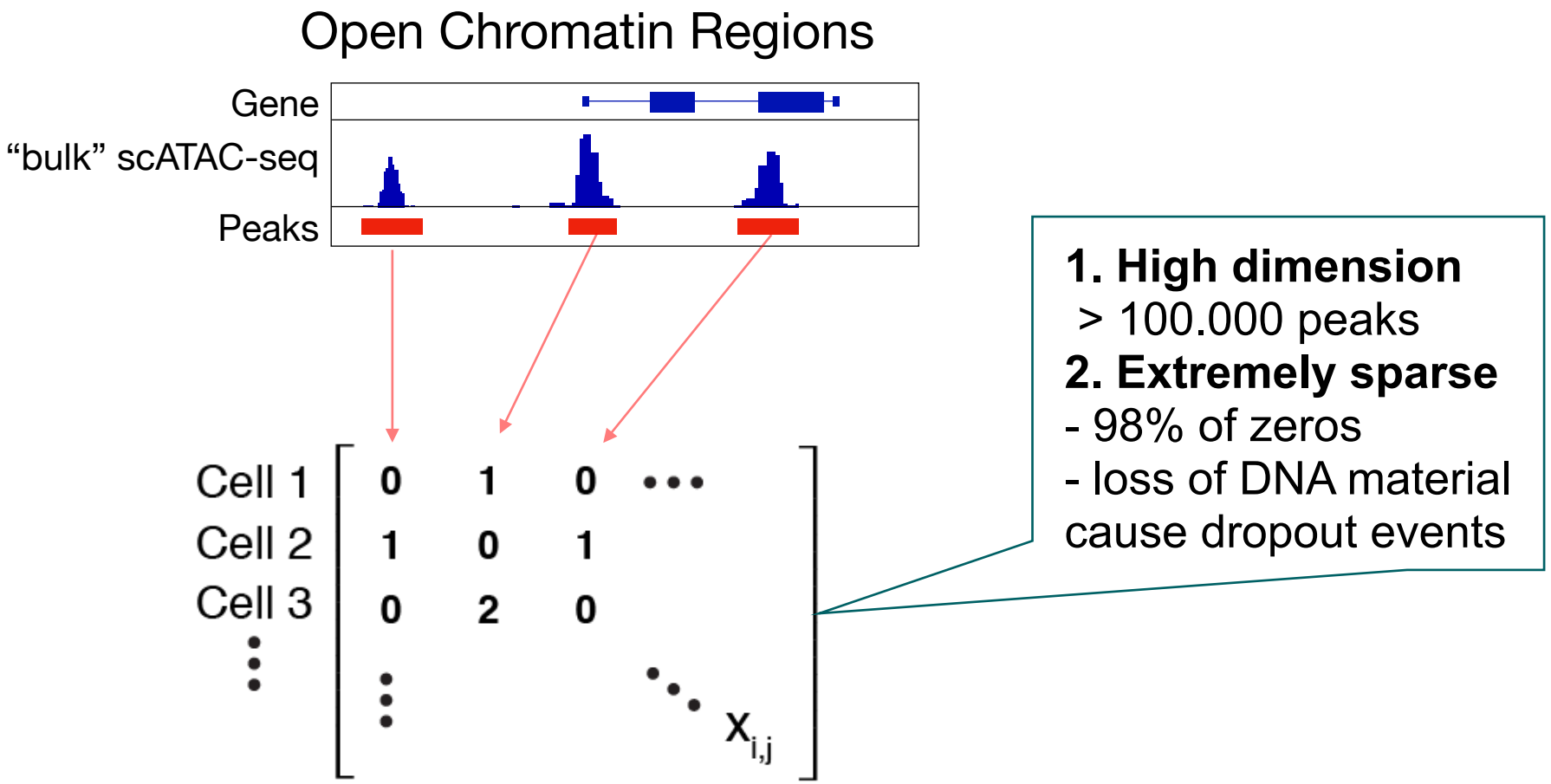
Institute for Computational Genomics
01011011010
10100100101

RWTH AACHEN UNIVERSITY

Computational Challenges - Single Cell ATAC



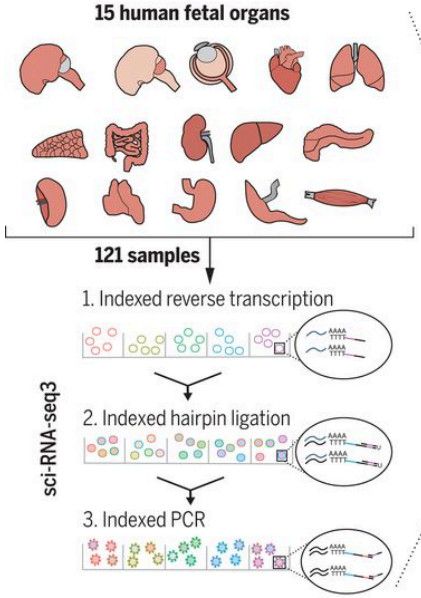
Computational Challenges - Single Cell ATAC



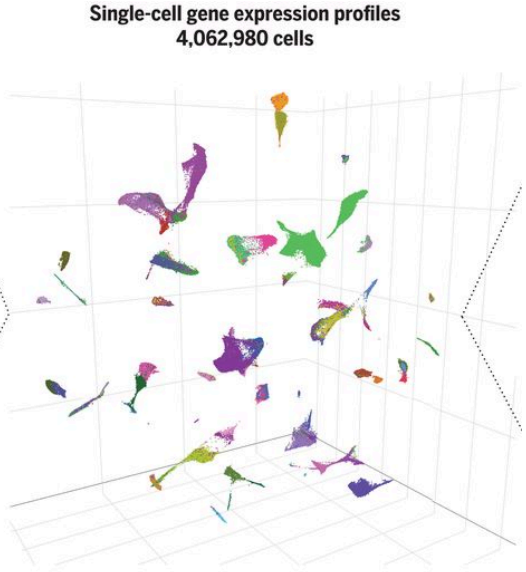
Resume / Single cell clustering

- Finding groups of single cells require complex pipeline:
  - Cell filtering
  - Normalisation
  - Artefact removal
  - Dimension reduction
  - Integration
  - Clustering
  - Cell annotation / visualisation
- Open points:
  - How to deal with large data sets (millions of cells)?
  - How to detect cells of rare populations?
  - How to deal with sparsity of scATAC seq data?

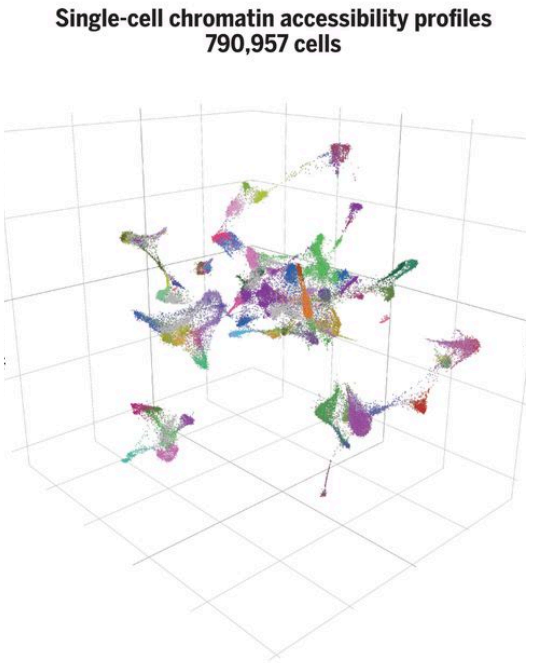
Clustering of cells / Human Fetal Cell Atlas



scRNA-seq



scATAC-seq



https://descartes.brotmanbaty.org/

- Open points:
  - How to deal with large data sets (millions of cells)?
  - How to detect cells of rare populations?
  - How to deal with sparsity of scATAC seq data?

Adapted from Donke et al. 2020.

Resume

- Review basic biological/computational aspects
  1. basics of molecular biology
  2. basics of sequencing
  3. basics bioinformatics problems
     - short sequences read alignment
     - gene expression quantification
     - single cell approaches
     - computational epigenetic (today)

Calendar

Today – Introduction to Bioinformatics, Next Generation Sequencing

2.05.2022 – Single cell sequencing Practical

8.05.2022 – Computational Epigenomics / Theory & Practical / Using RWTH HPC/GPU cluster

15.05.2022 – 4.7.2022 – Project development

11.07.2022 – Project Presentation

Communication/discord channel: https://discord.gg/hmGxznNpZH .

Thank you!