Bioinformatics Lab

Ivan Gesteira Costa, Mingbo Cheng, James Nagai, Mina Shaigan, Martin Manolov
Institute for Computational Genomics

Institute for Computational Genomics
01011011010
10100100101
RWTH AACHEN UNIVERSITY

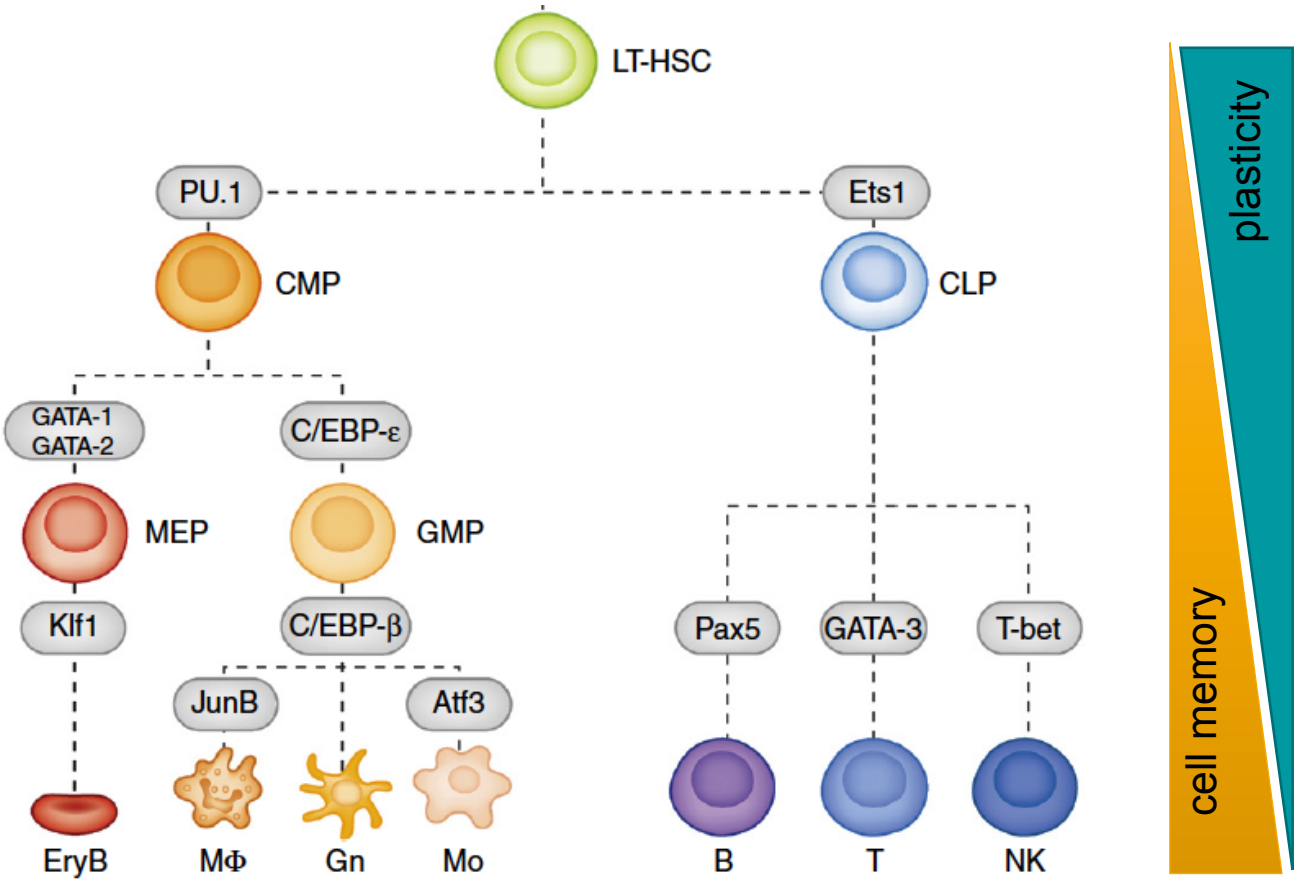
Resume

- 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)**

Computational Epigenomics

Cell Differentiation

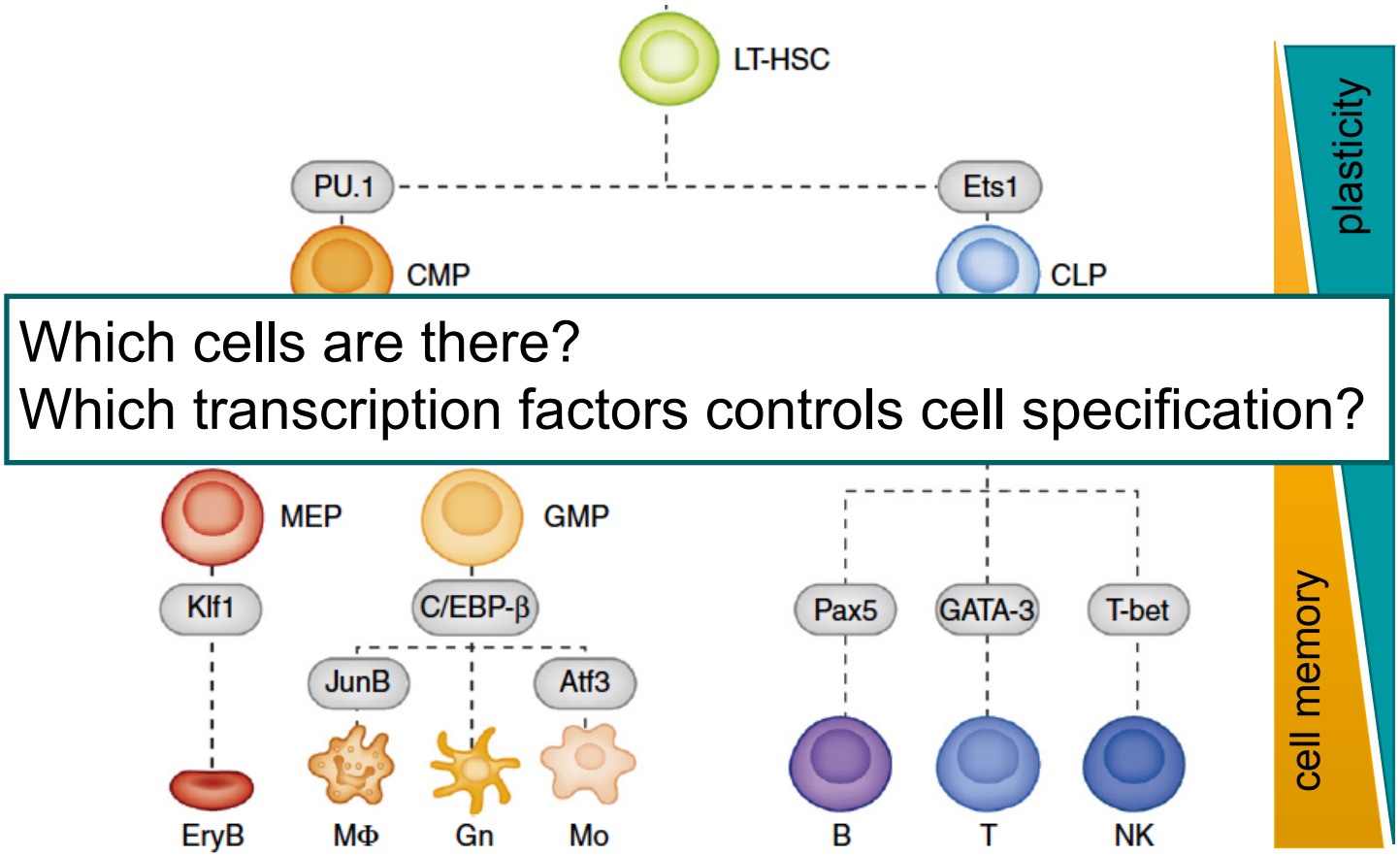
Hematopoiesis



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

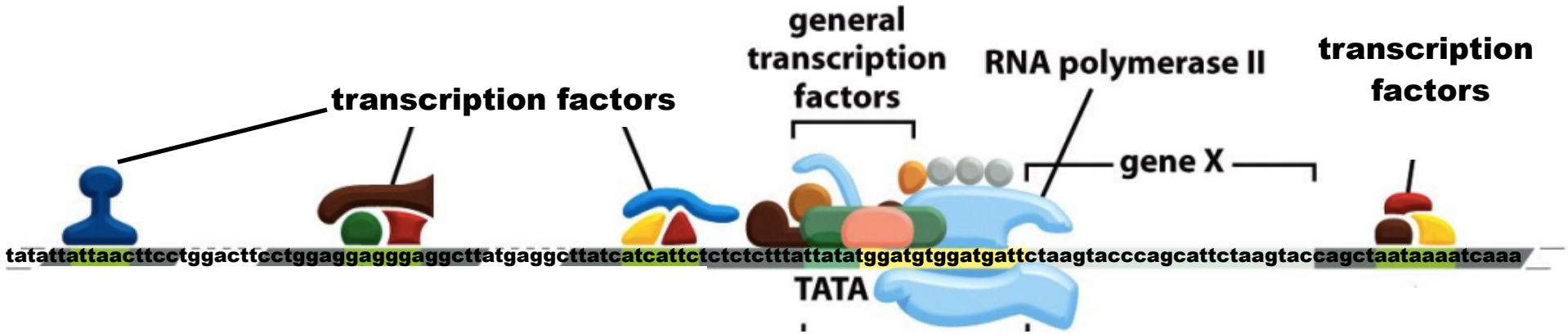
Cell Differentiation

Hematopoiesis



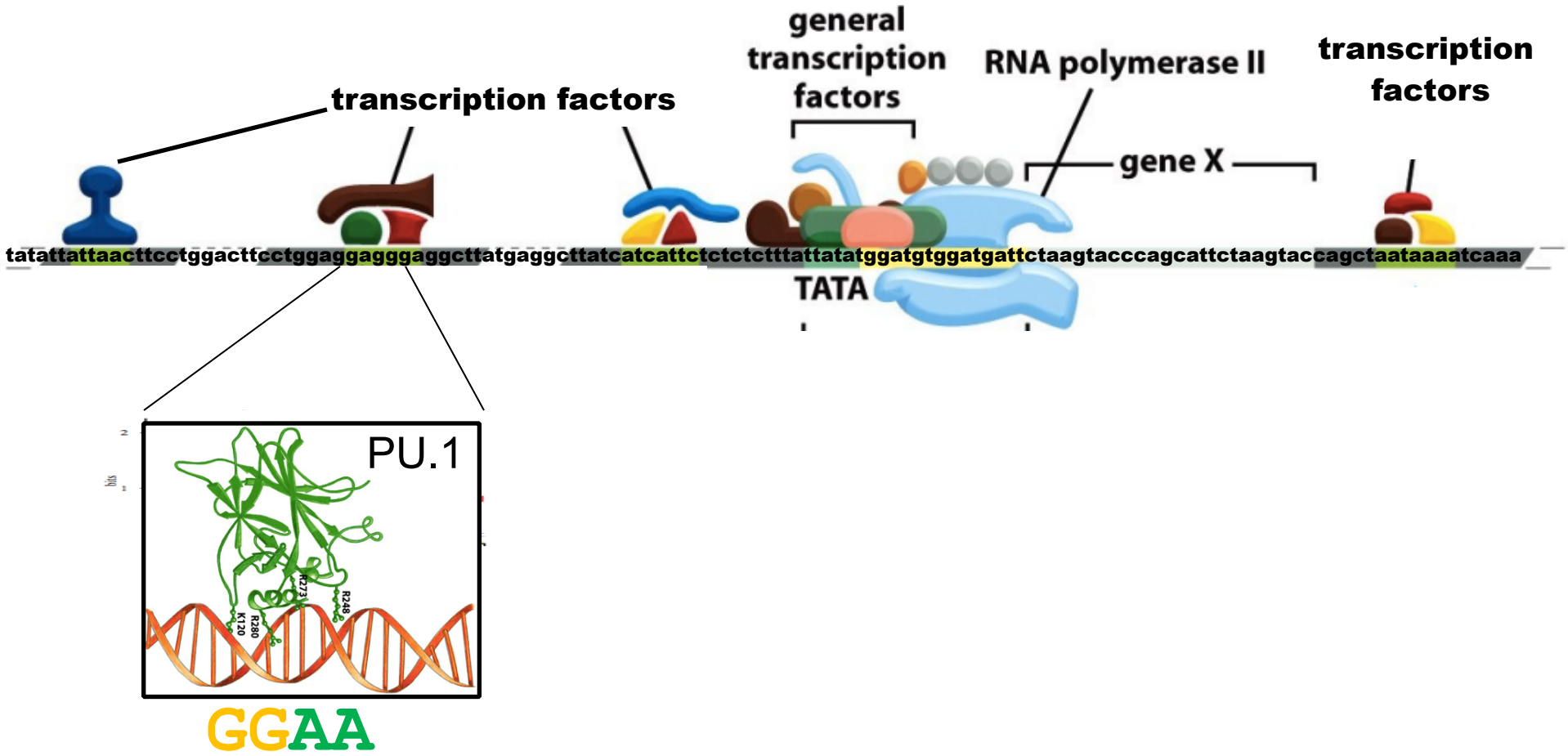
Source: Amit (2016), *Nature Immunoloy.*

Regulatory Control – Transcription Factor Binding



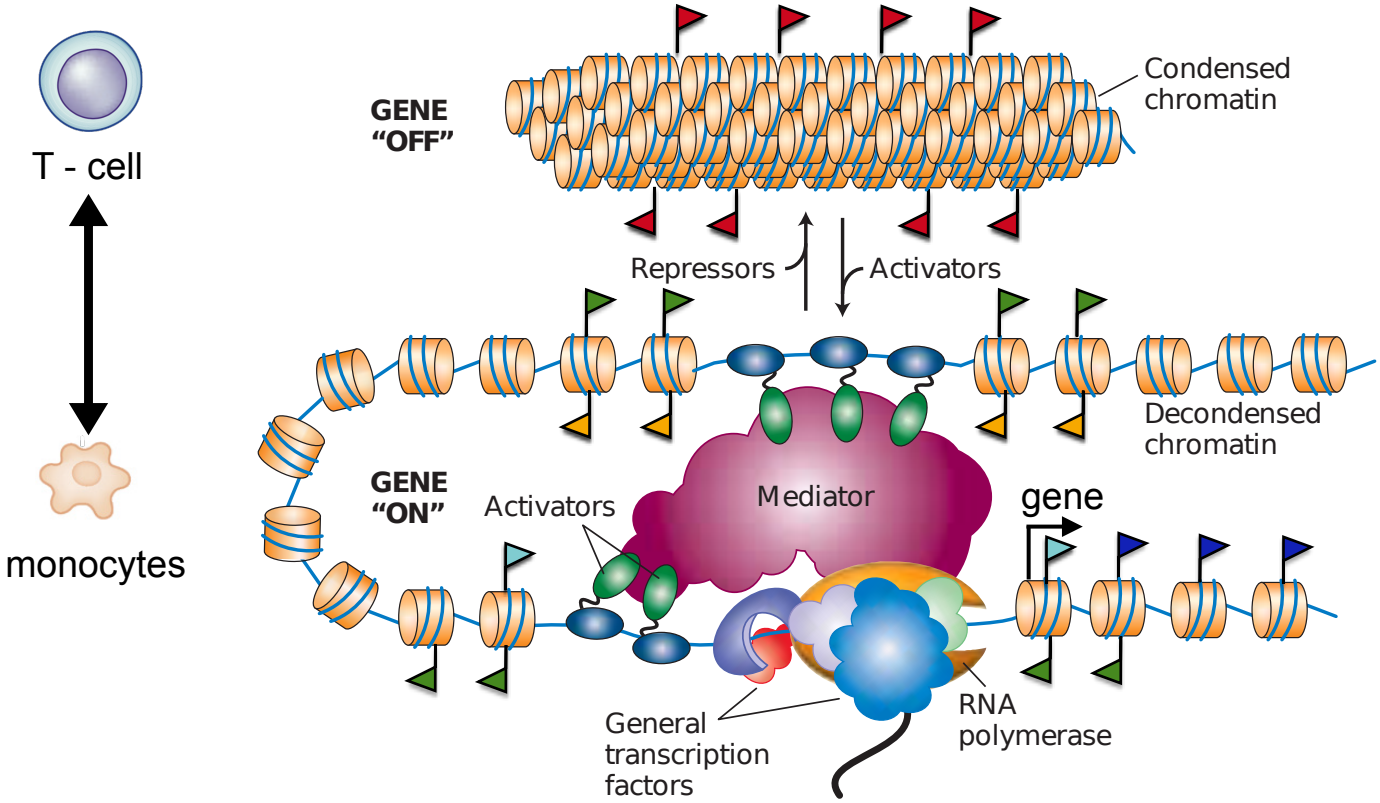
Source: Alberts, B. et al. (2008) Garland Science, 5th ed.

Regulatory Control – Transcription Factor Binding



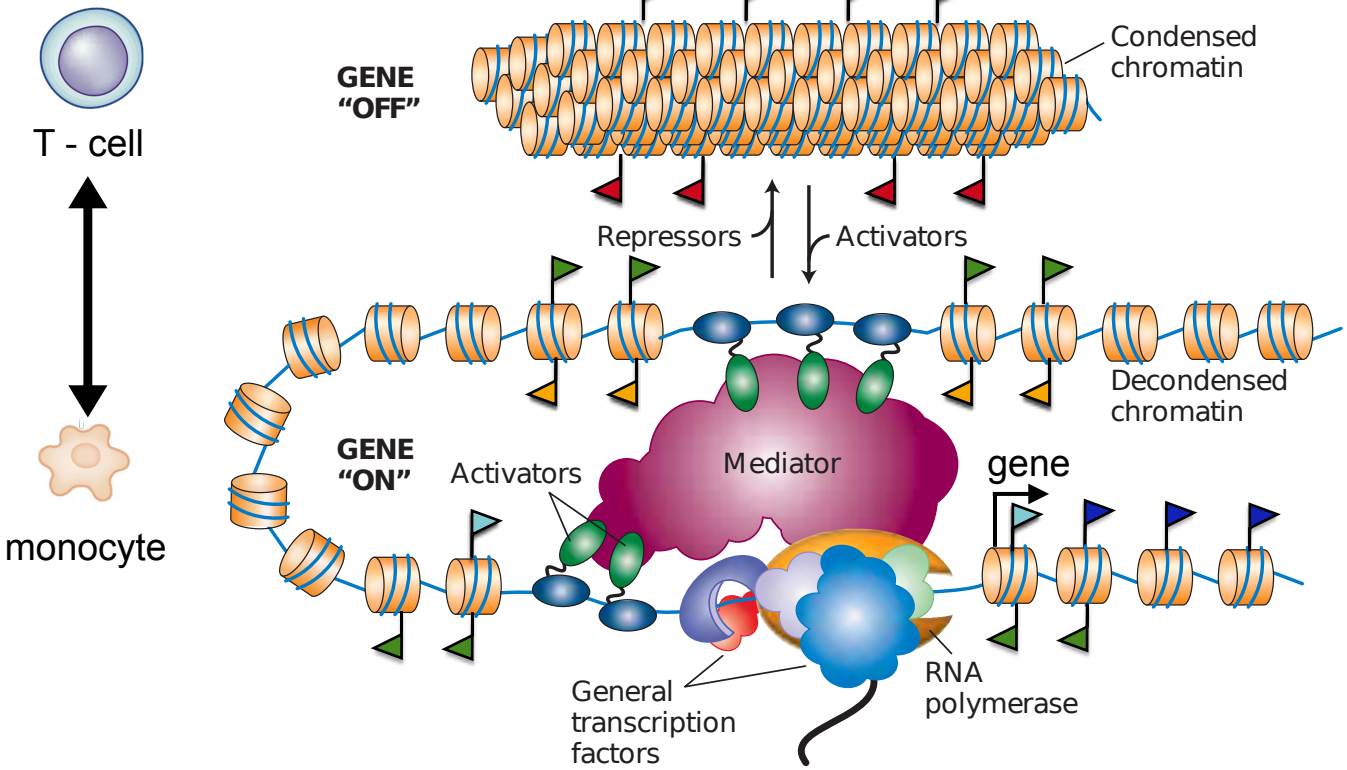
Source: Alberts, B. et al. (2008) Garland Science, 5th ed.

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

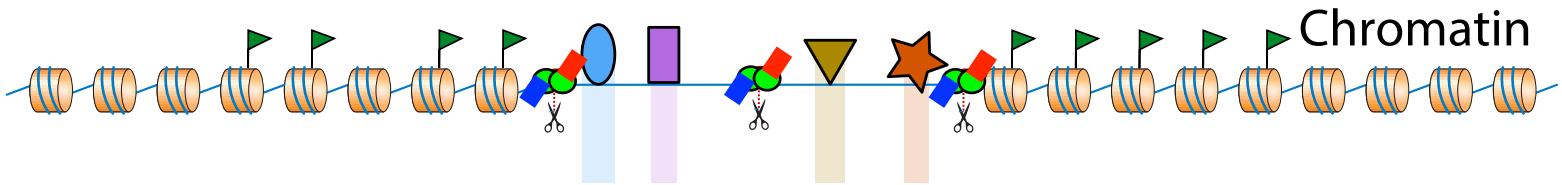
Active Enhancers

H3K4me1

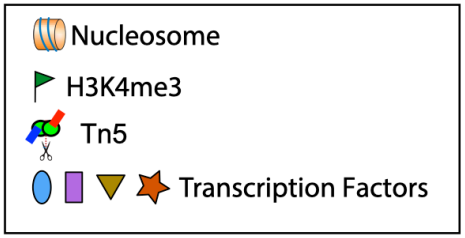
Repressed regions

H3K27me3, H3K9me3

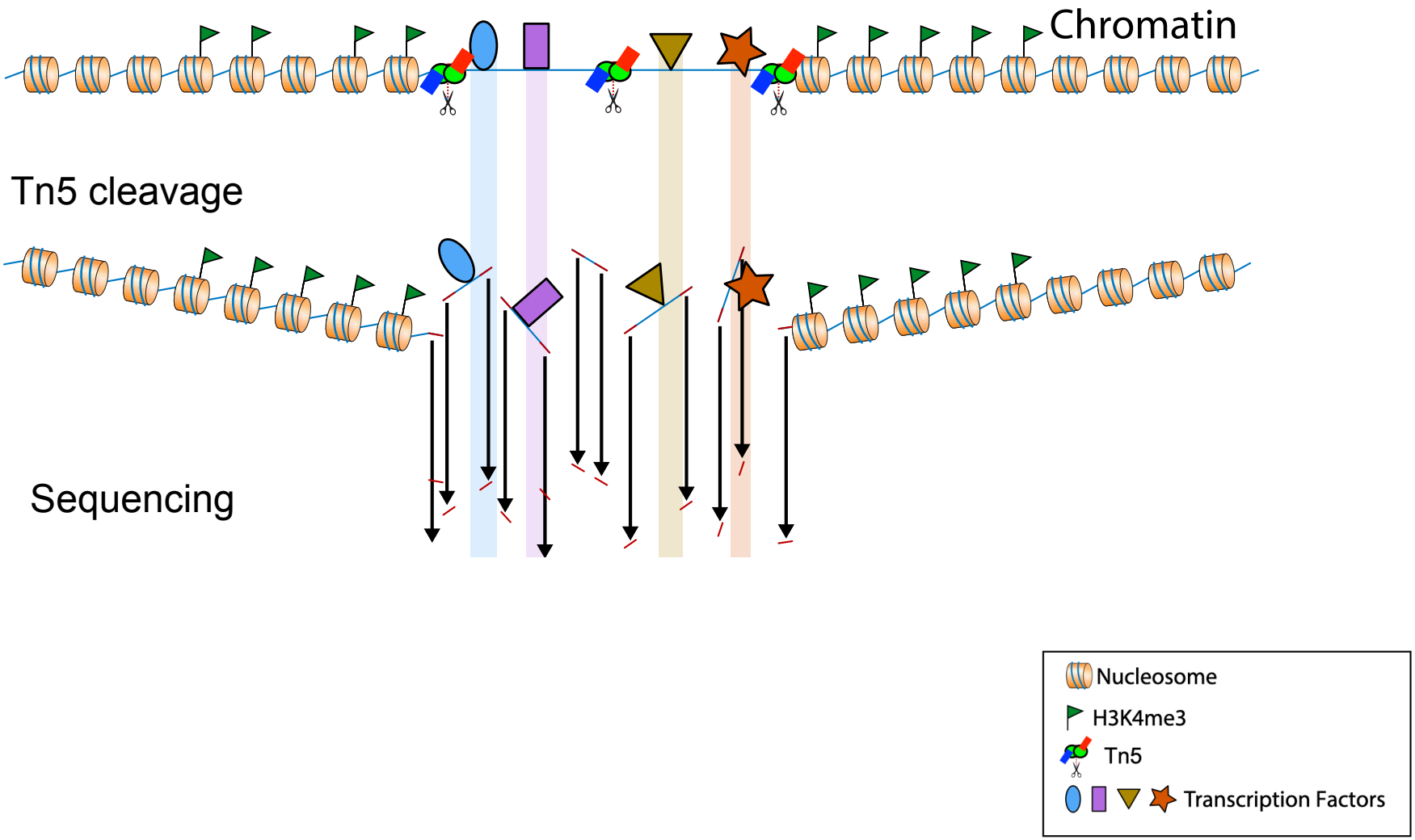
Open Chromatin with ATAC-seq



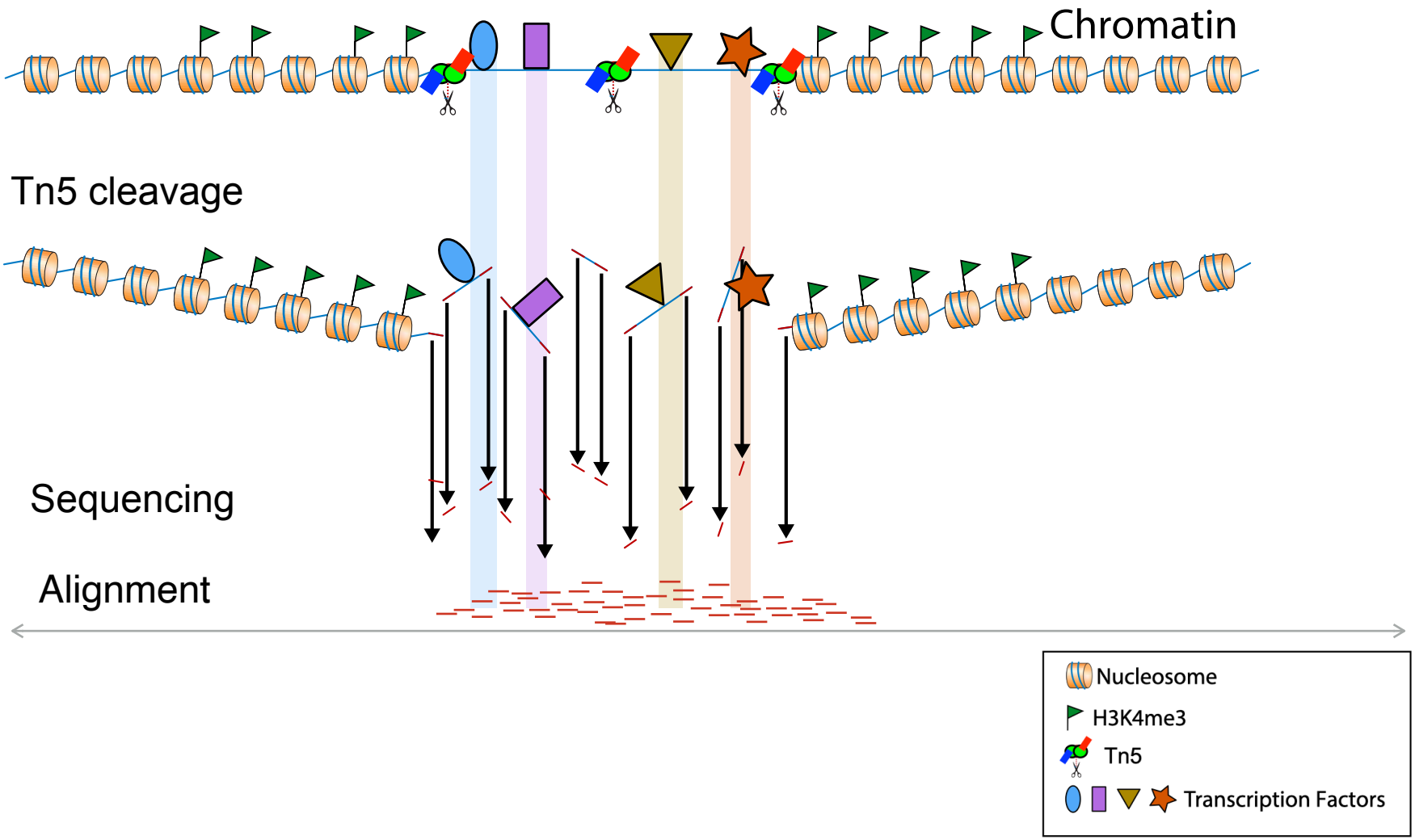
Tn5 cleavage



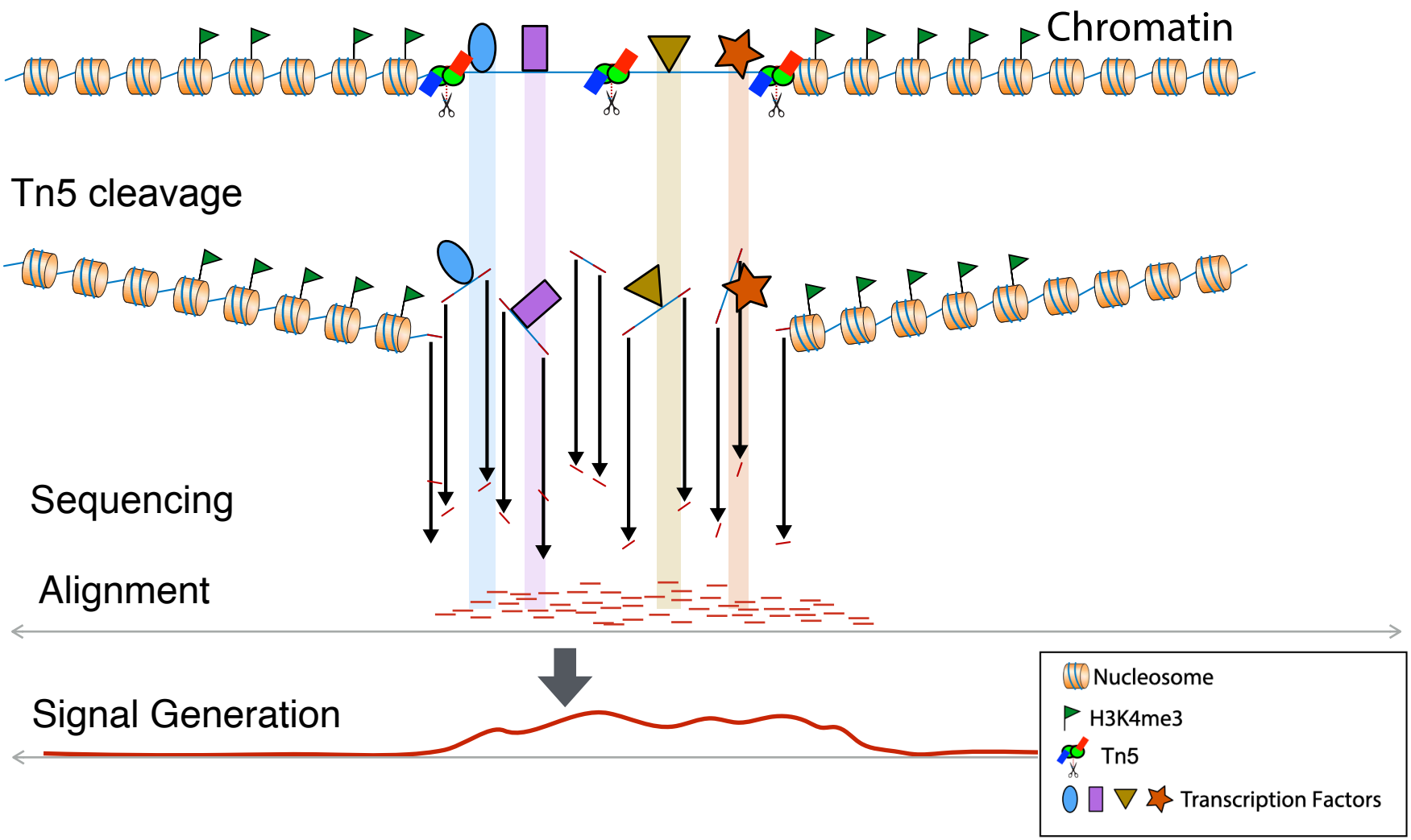
Open Chromatin with ATAC-seq



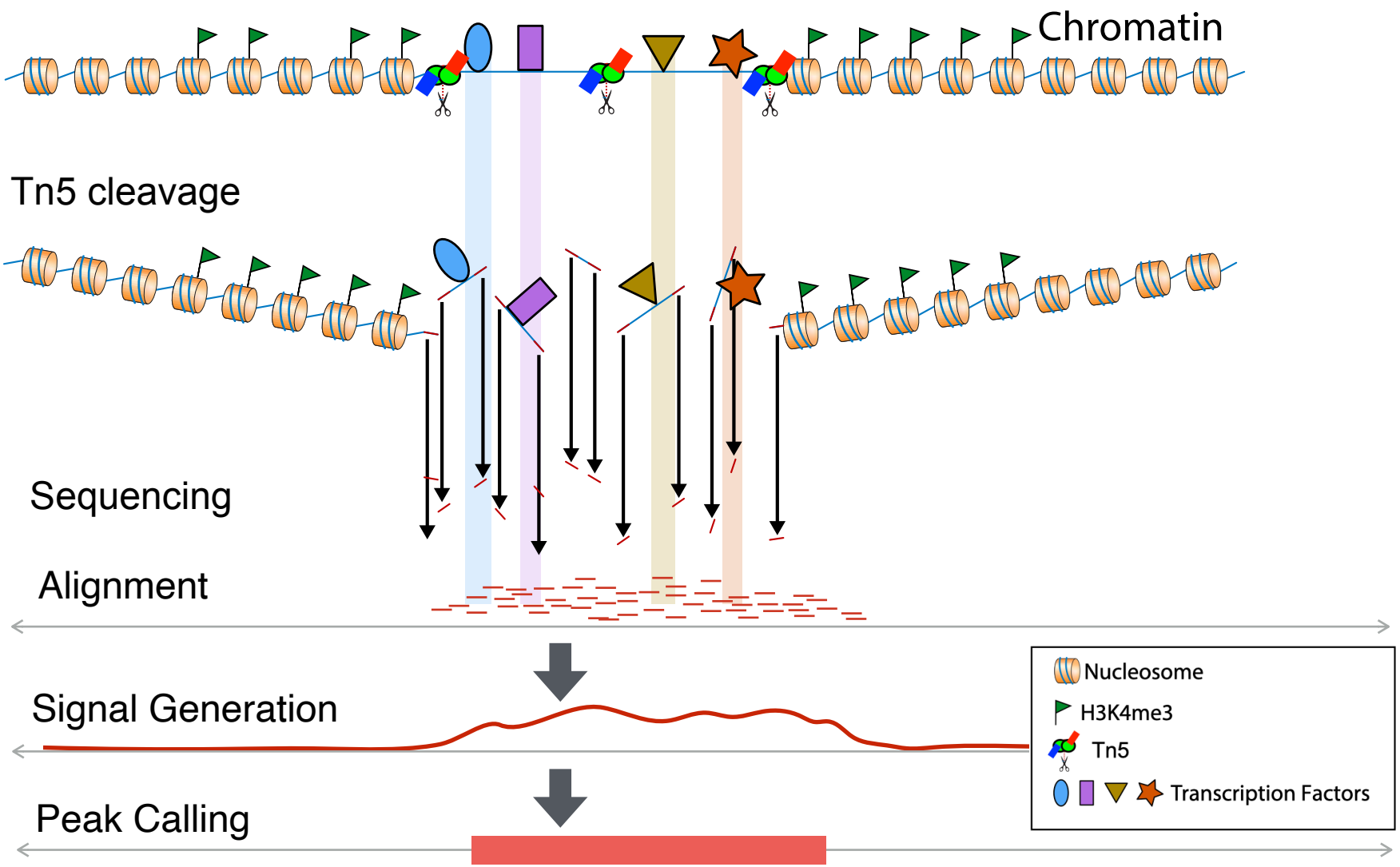
Open Chromatin with ATAC-seq



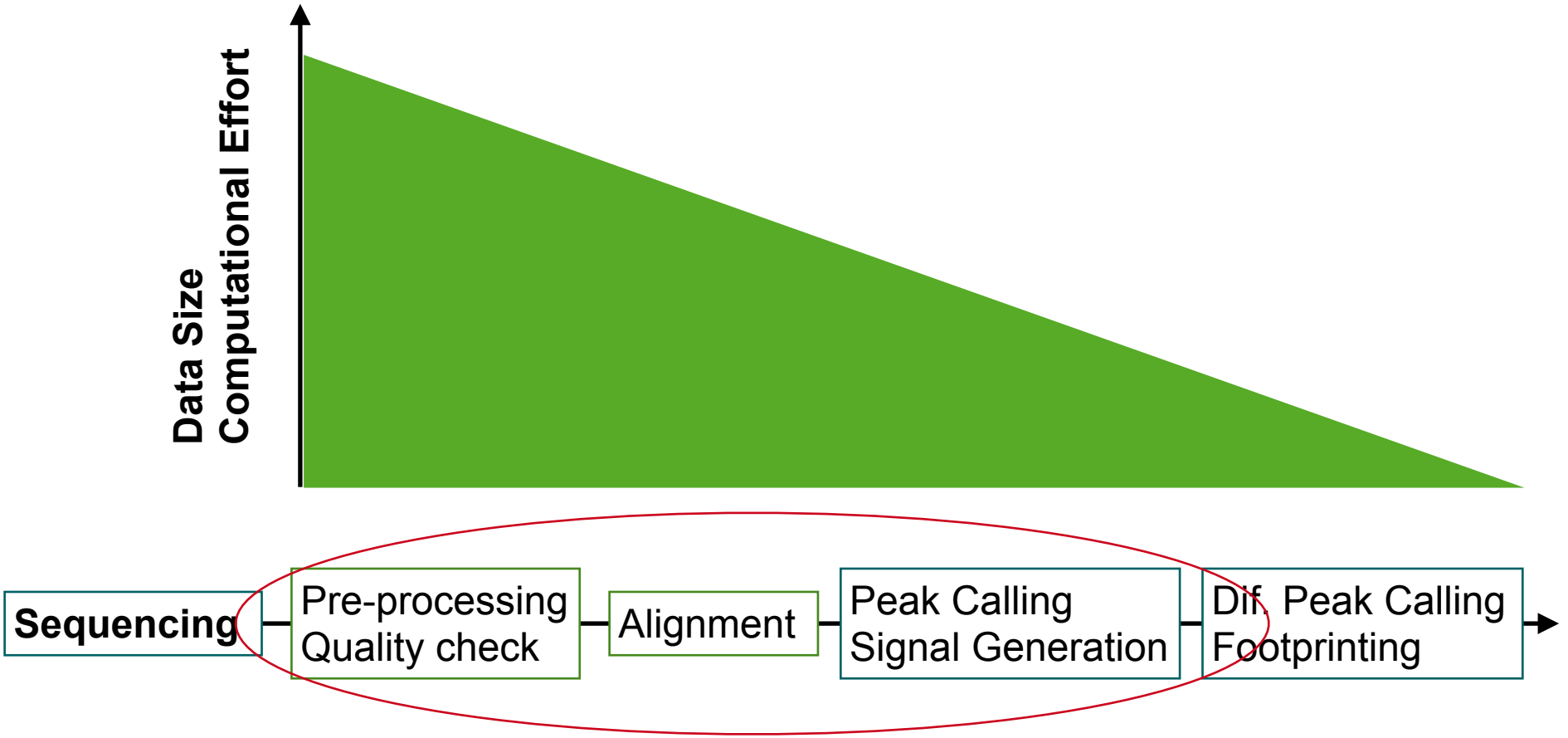
Open Chromatin with ATAC-seq



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.

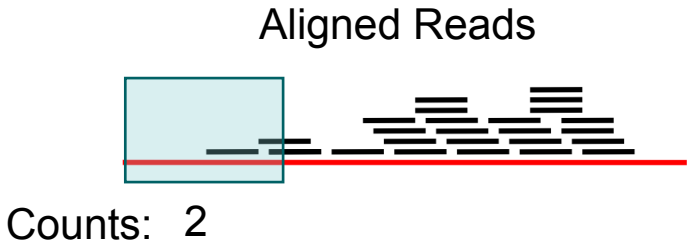
Example of a simple peak caller :

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.



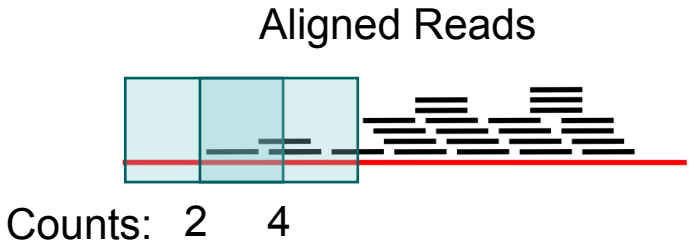
Example of a simple peak caller :

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.



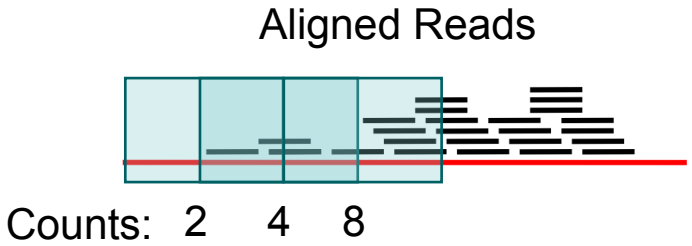
Example of a simple peak caller :

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.



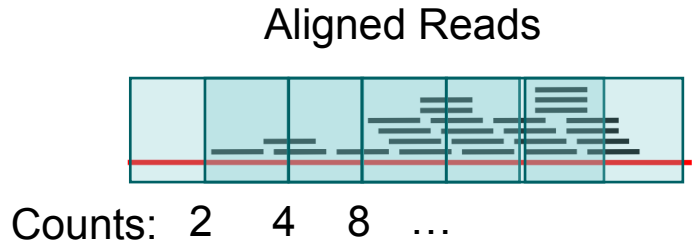
Example of a simple peak caller :

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

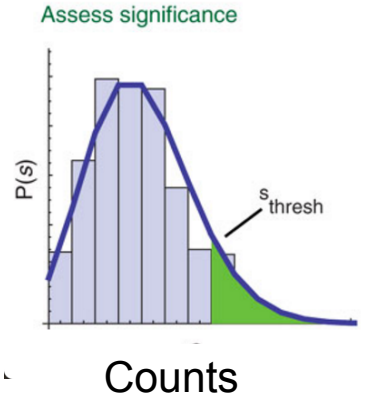
Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.

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

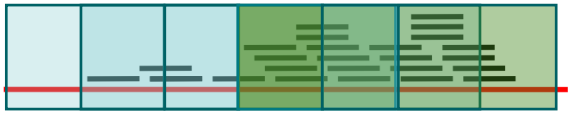


See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

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

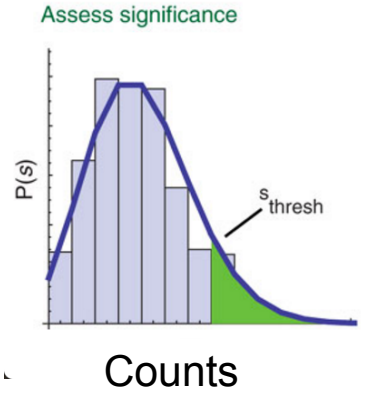
Aligned Reads



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

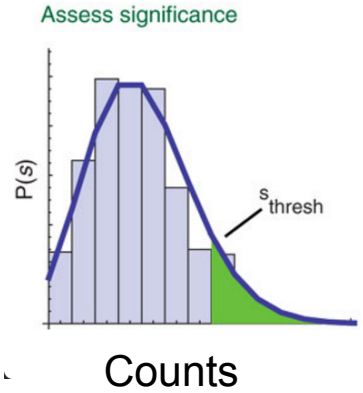
Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.

Problems:

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

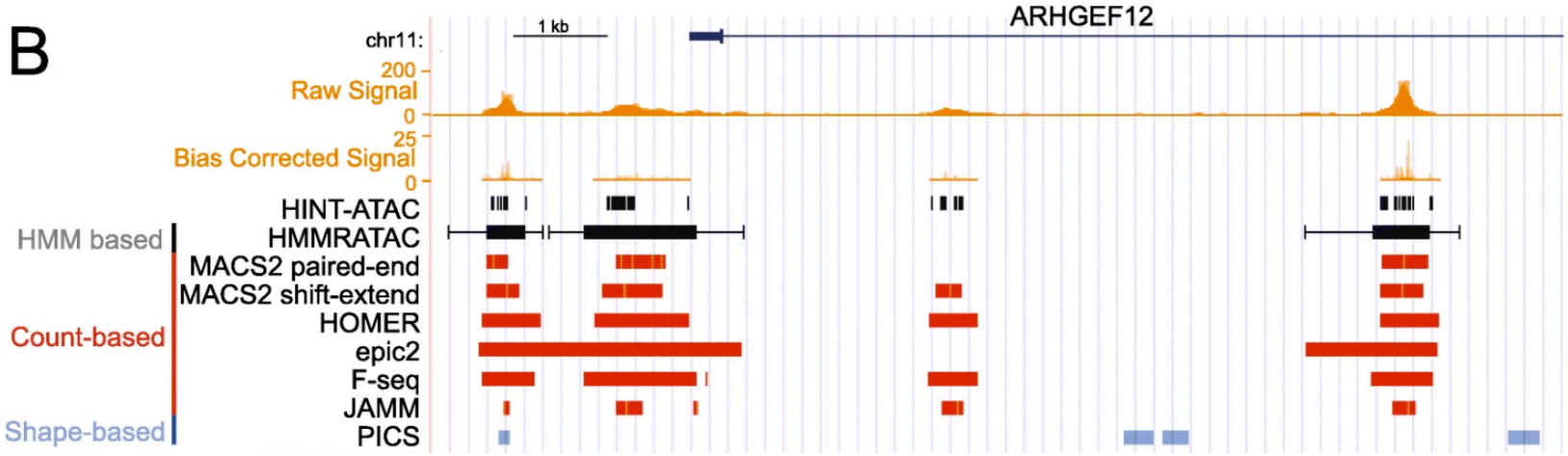
unts: 2 4 8 …

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/

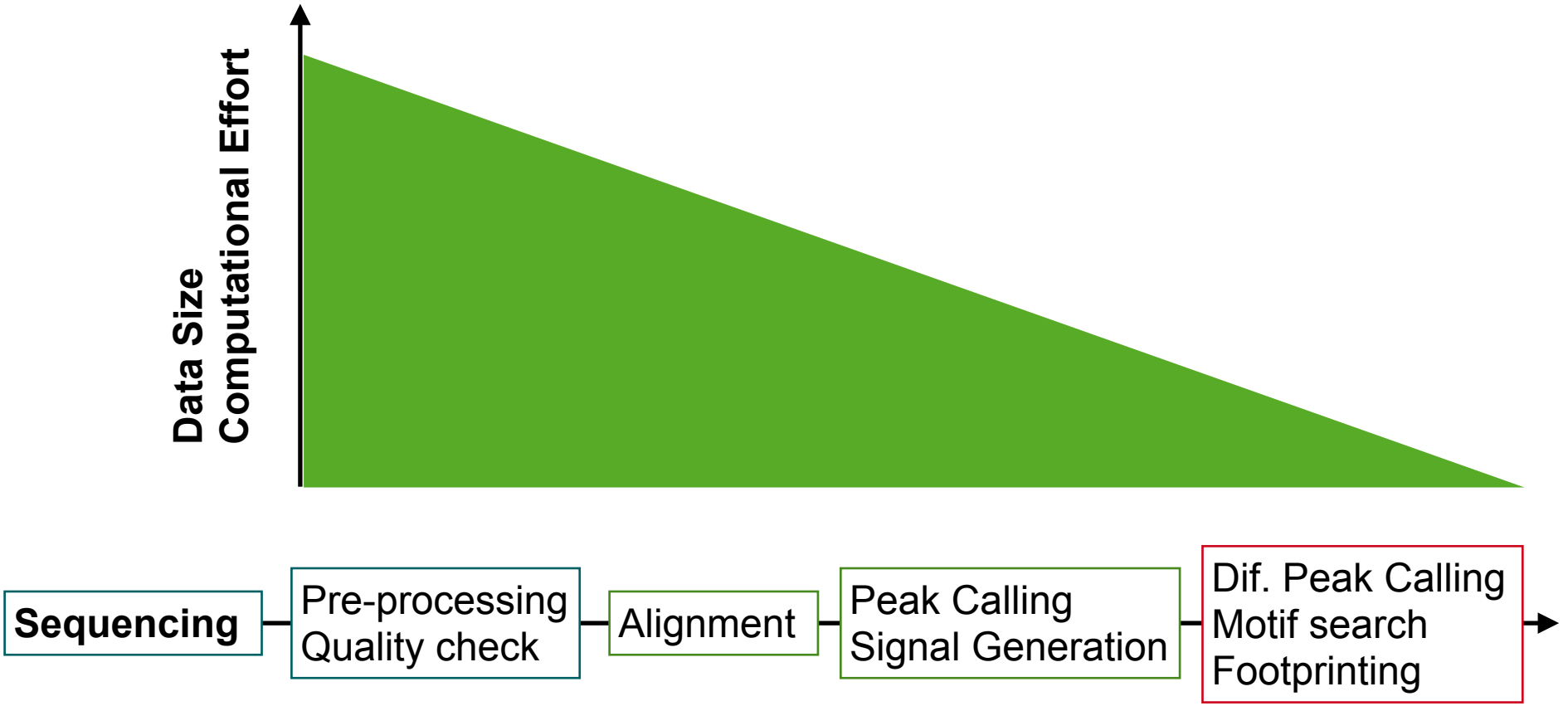
Peak calling in ATAC-seq



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

Source: Yan, Genome Biology, 2020.

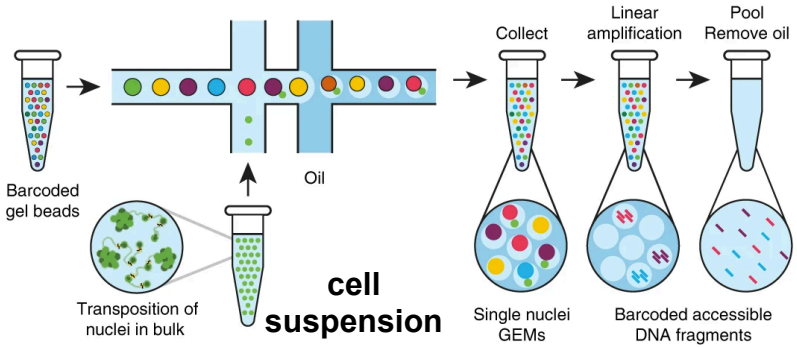
Bioinformatics Pipeline / ATAC-seq



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

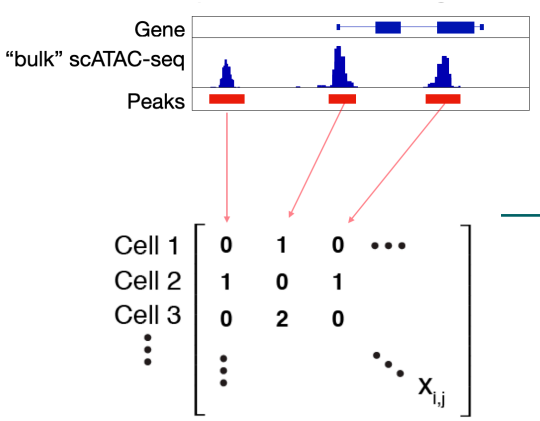
Open chromatin with scATAC-seq

Droplet based scATAC-seq

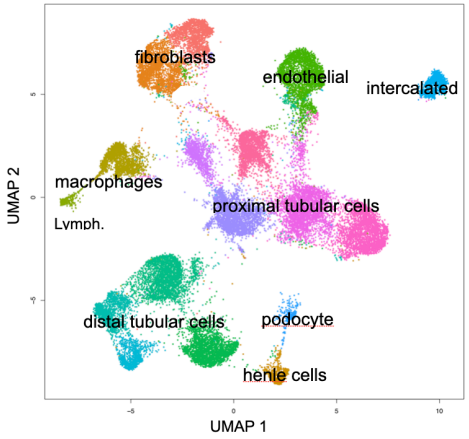


Adapted from Satpathy,
Nature biotechnology, 2019

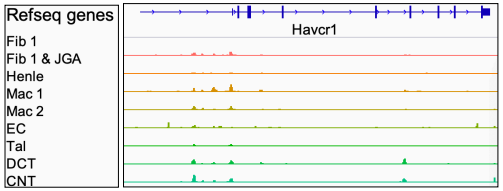
Open Chromatin Matrix

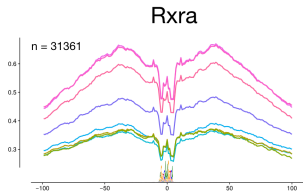


UMAP / Clustering



cluster pseudo bulk ATAC-seq





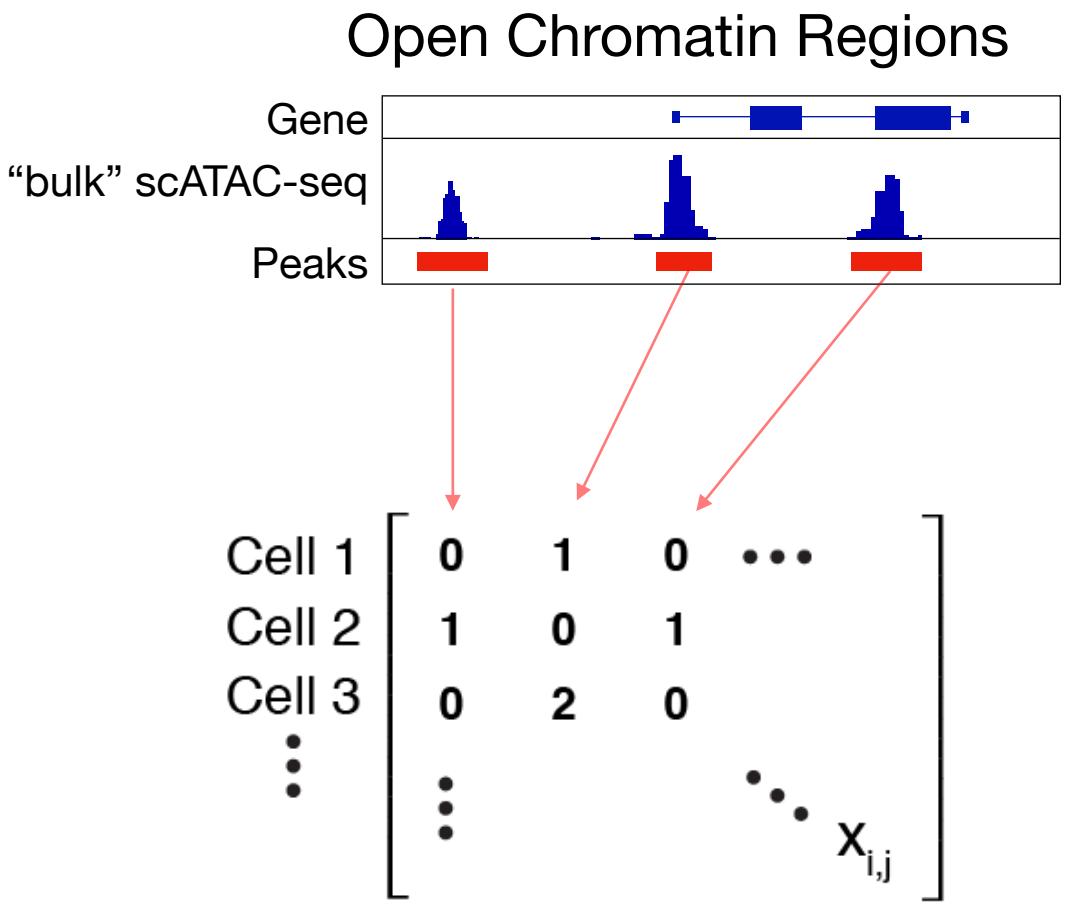
Footprinting & TF Activity

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

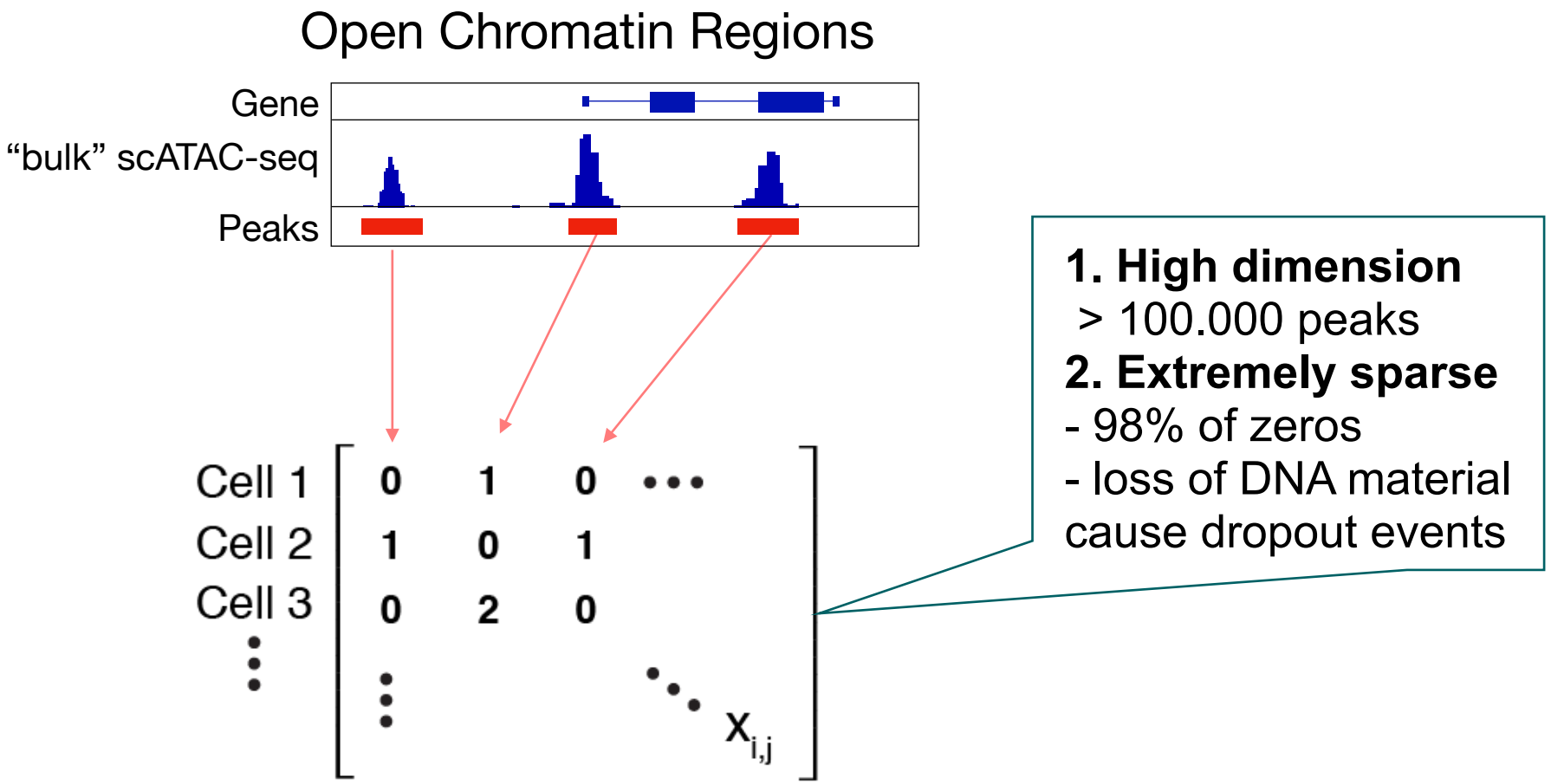
Institute for Computational Genomics
01011011010
10100100101

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 sparsity of single cell (scRNA-seq or scATAC-seq) data?

Thank you!