Analysis of Open Chromatin Data

Zhijian Li
Institute for Computational Genomics
Joint Research Center for Computational Biomedicine
RWTH Aachen University Hospital

Institute for Computational Genomics
01011011010101
10100100101

RWTH AACHEN UNIVERSITY

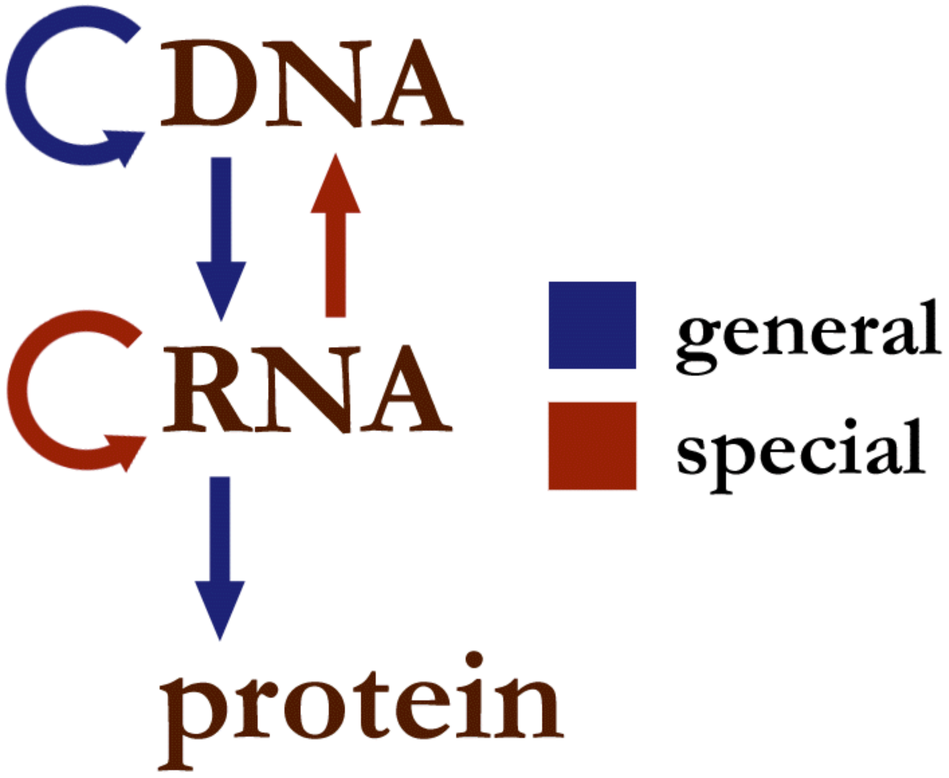
Objectives

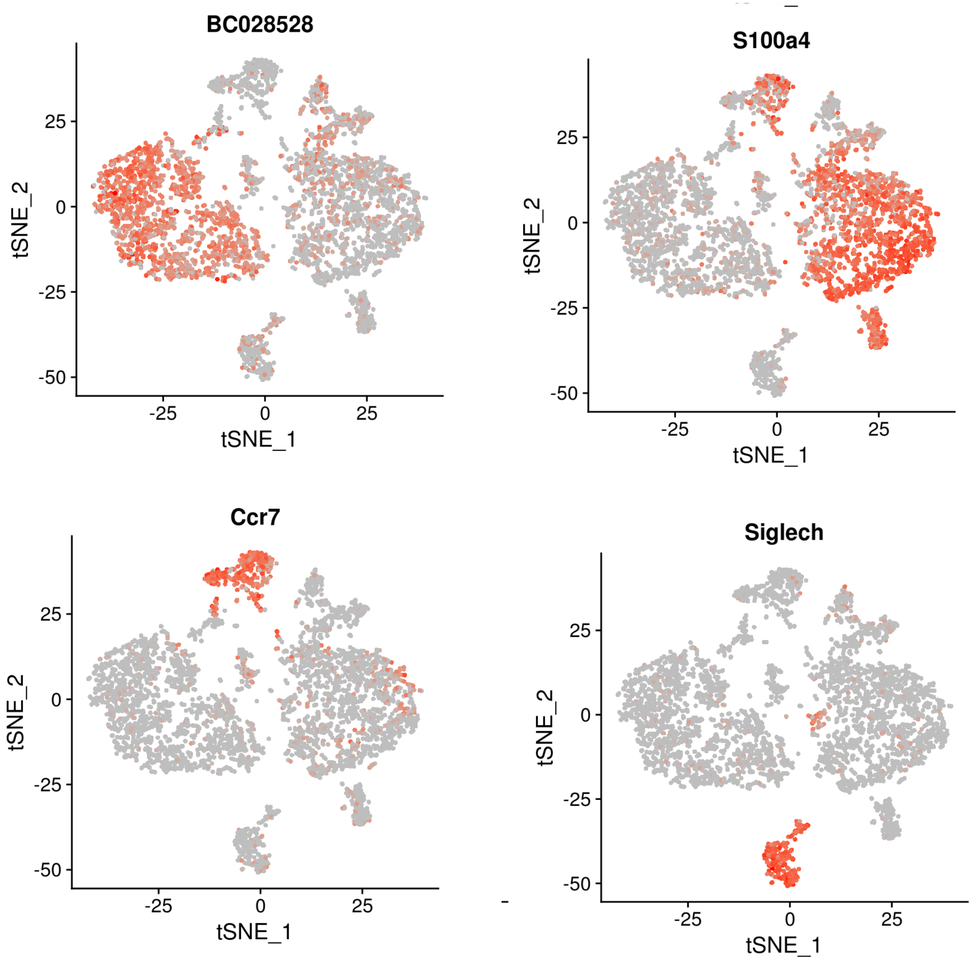
- Understand open chromatin from biological point of view
- Analyse bulk open chromatin data
- Visualise the result using IGV
- Extend the analysis to single cell open chromatin data

Open Chromatin

Institute for Computational Genomics
01011011010101
10100100101
RWTH AACHEN UNIVERSITY

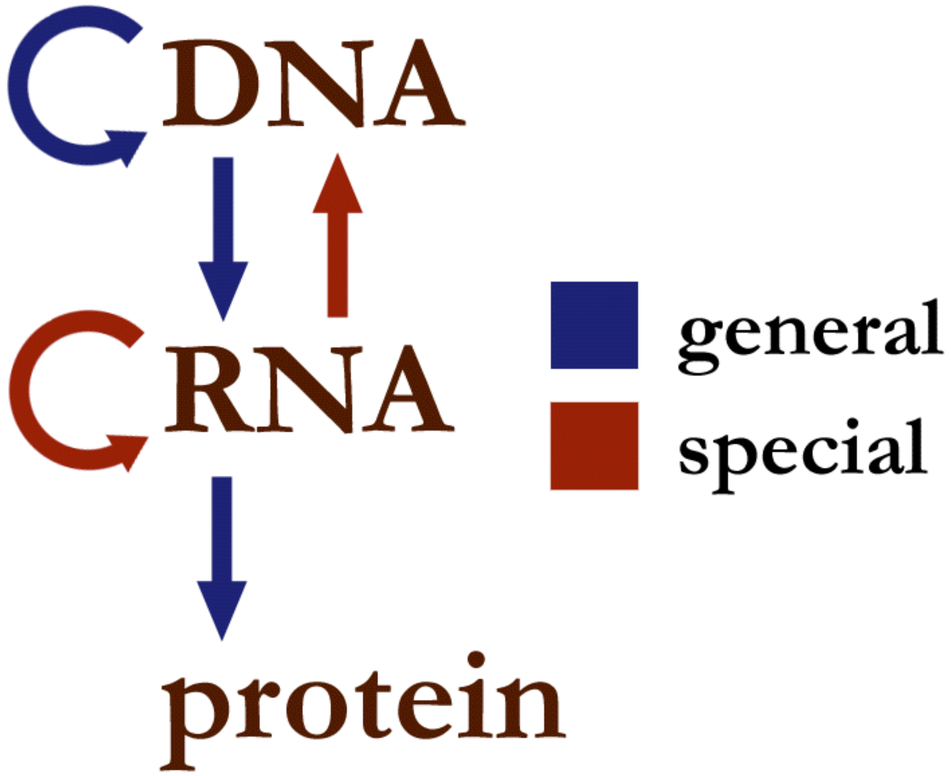
Central dogma of molecular biology





https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

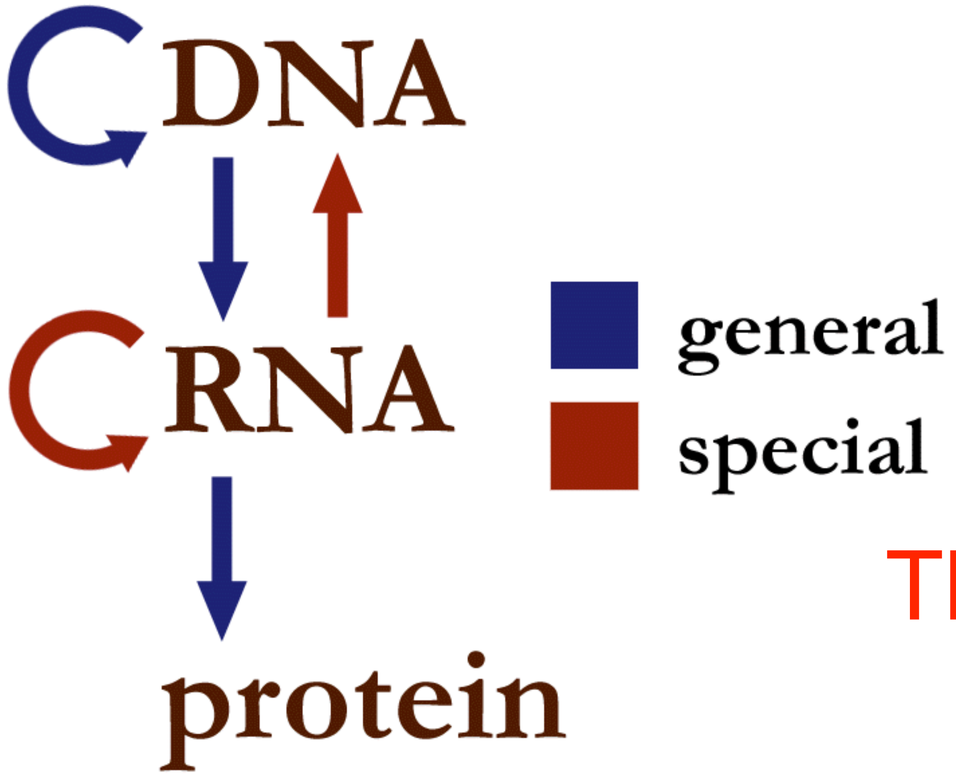
One genome vs. many cell types



Why do the cells have different gene expression, given that they have the exactly same genome?

https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

One genome vs. many cell types

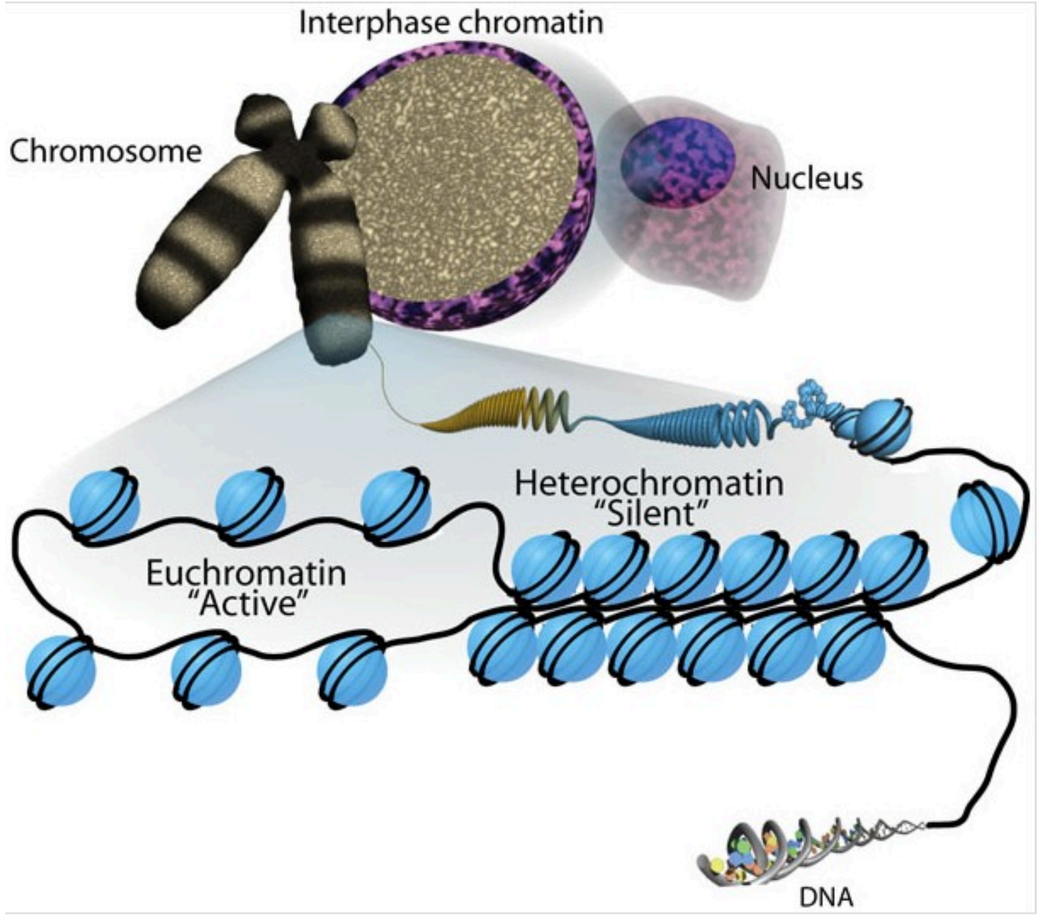


Why do the cells have different gene expression, given that they have the exactly same genome?

Their chromatin states are different!

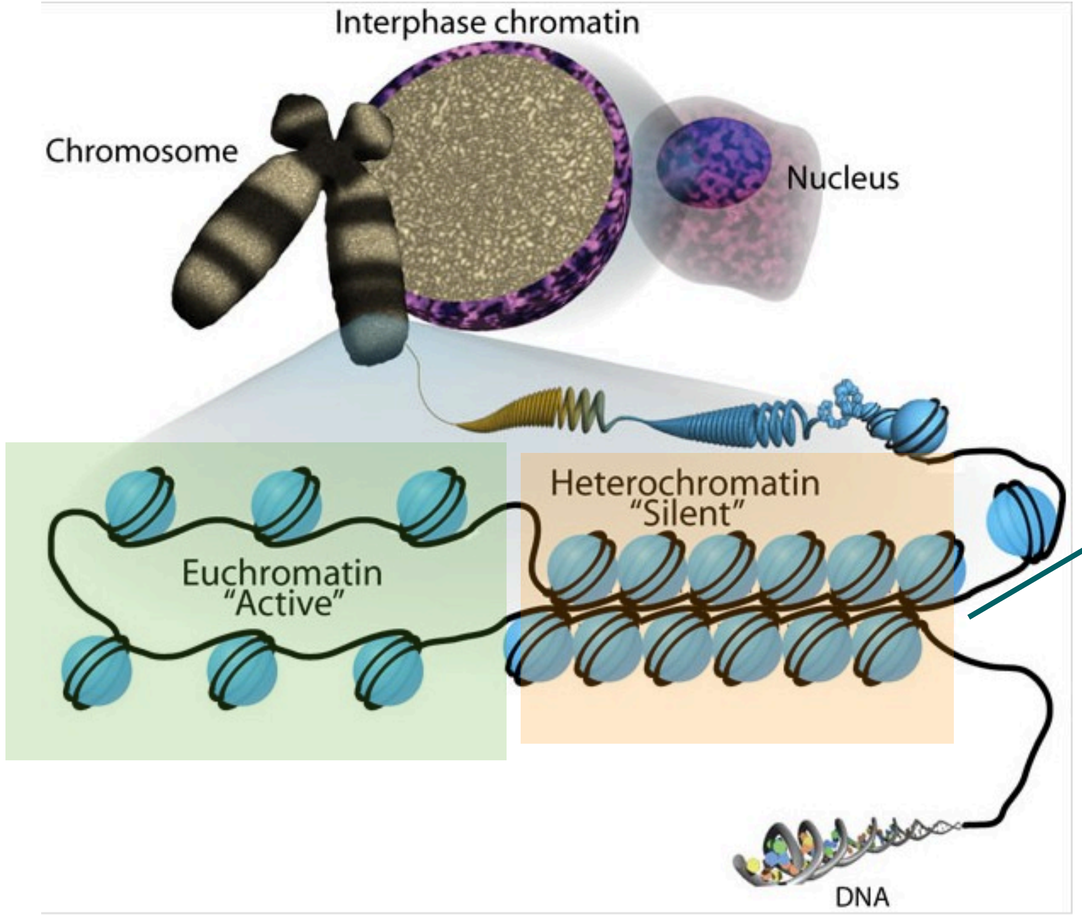
https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

Open chromatin vs. closed chromatin



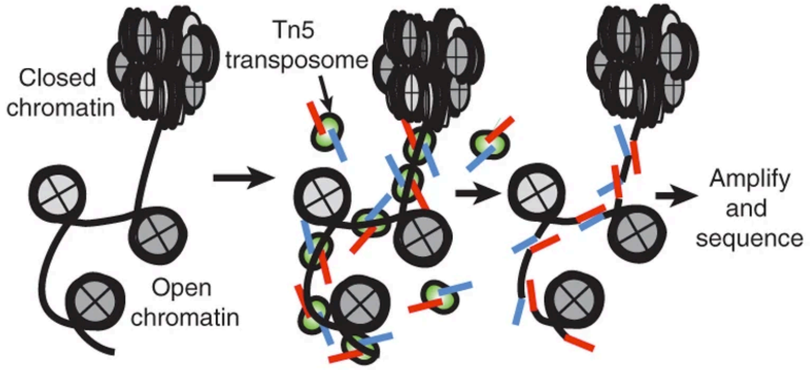
Open chromatin vs. closed chromatin

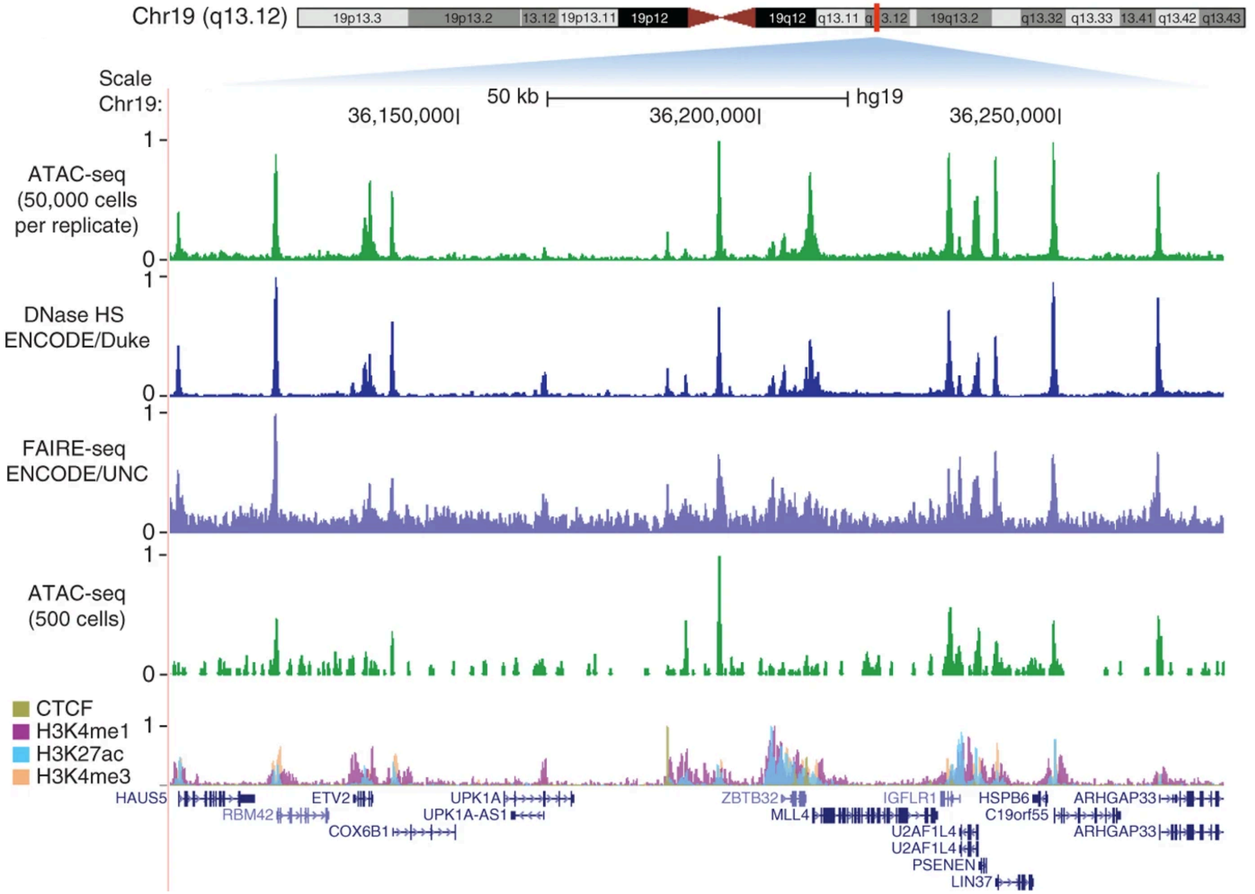
Open chromatin



Close chromatin

ATAC-seq probes open chromatin state





Buenrostro, Jason D., et al. *Nat. Method* 2013

Institute for Computational Genomics
01011011010101
10100100101

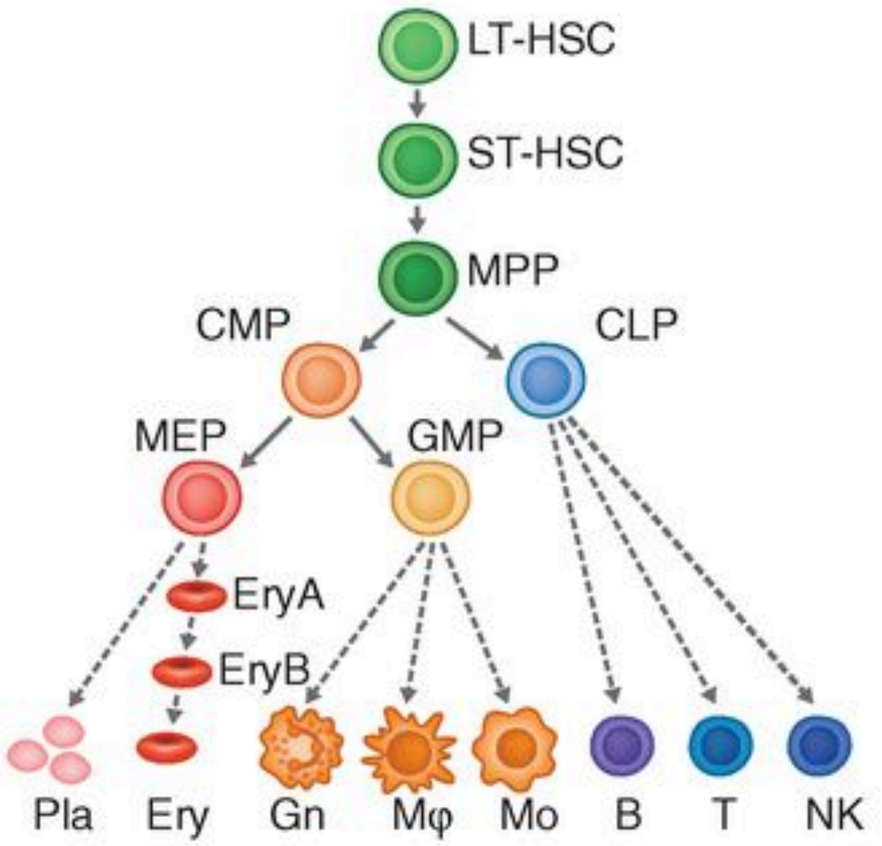
RWTH AACHEN UNIVERSITY

Analysis of ATAC-seq Data

Institute for Computational Genomics
01011011010101
10100100101
RWTH AACHEN UNIVERSITY

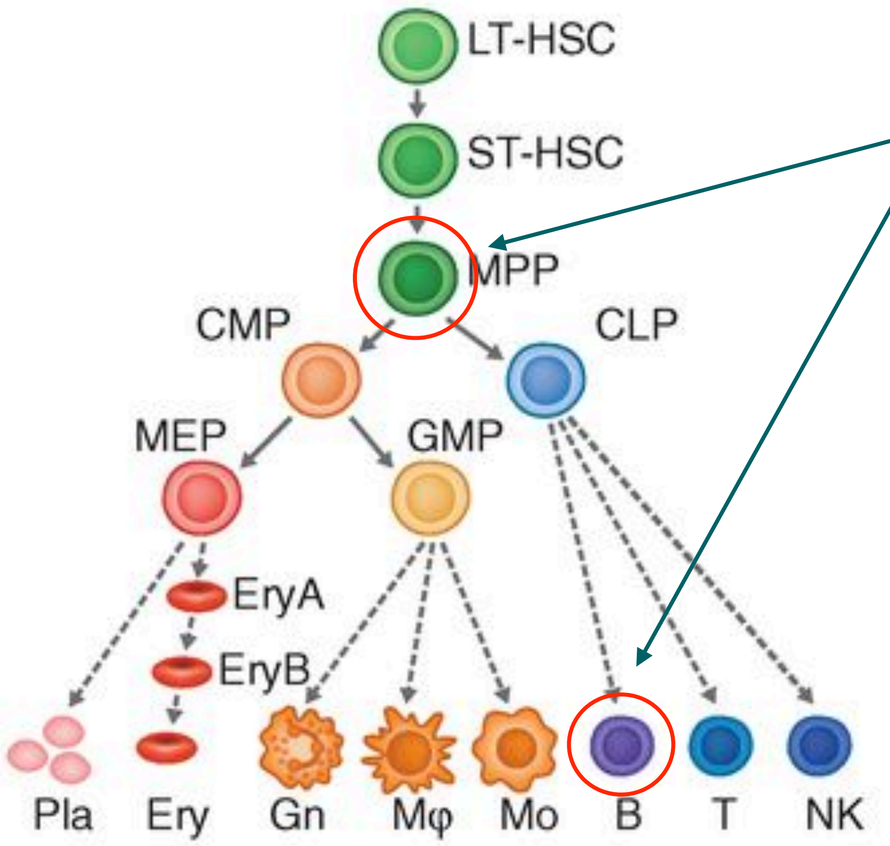
Chromatin dynamic during blood formation

Hematopoietic differentiation



Chromatin dynamic during blood formation

Hematopoietic differentiation



Any differences at chromatin level?

Analysis pipeline

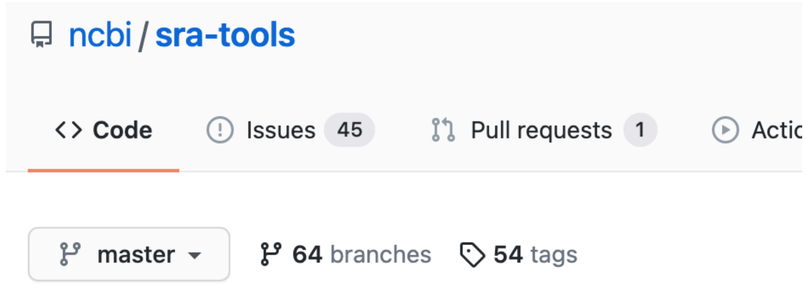
- Download data (SRA toolkit)
- Sequence alignment (Bowtie2)
- Peak calling (MACS2)
- Motif matching (RGT)

https://github.com/CostaLab/SOSE2022/blob/main/Practical_ATAC.md

1. Download data

SRA toolkit

The SRA Toolkit is a collection of tools and libraries for using data in the NCBI Sequence Read Archives.



Two common sub commands:

- *prefetch*
- *fastq-dump*

Download data

10 minutes

2. Short DNA Sequence Alignment

Institute for Computational Genomics
01011011010101
10100100101
RWTH AACHEN UNIVERSITY

Sequence alignment

Input data

- A large reference genome (chr19.fa)
- Millions of short DNA reads (MPP.fastq, B.fastq)

Sequence alignment

- Find most probable position for each read in the genome (allow insertion and deletion)

Output data

- Aligned file (MPP.sam, B.sam)

Bowtie 2

Align reads to reference genome

- Extract ‘seed’ substrings from the reads
- Align the substrings to the reference
- Calculate the position information
- Extend the seeds to full alignment using dynamic programming

More information

- Paper: https://www.nature.com/articles/nmeth.1923
- Website: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

Alignment

10 minutes

3. Peak Calling

Peak calling

Problem definition: Find genomic regions with more aligned reads than expected by chance.

Peak calling

Problem definition: Find genomic regions 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 is higher than expected by change

Peak calling

Problem definition: Find genomic regions 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 is higher than expected by change

Peak calling

Problem definition: Find genomic regions 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 is higher than expected by change

Peak calling

Problem definition: Find genomic regions with more aligned reads than expected by chance.

Example of a simple peak caller :

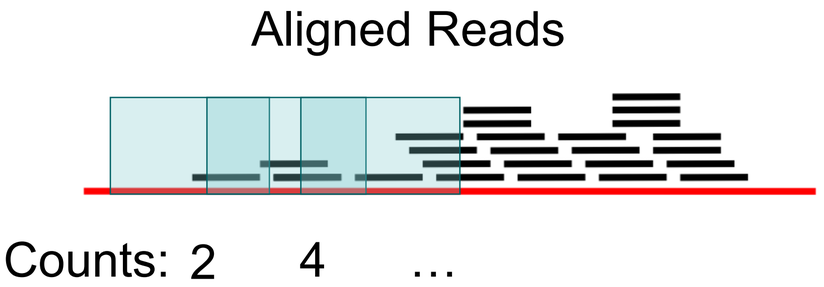
1. use a fix window to scan through the genome and obtain a distribution of counts per bin
2. define a statistical test to evaluate if the number of reads is higher than expected by change

Peak calling

Problem definition: Find genomic regions 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 is higher than expected by change



Peak calling

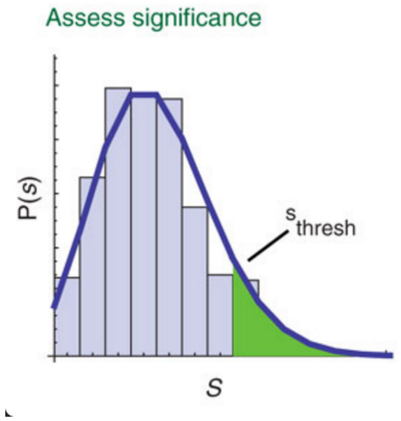
Problem definition: Find genomic regions with more aligned reads than expected by chance.

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 is higher than expected by change

Aligned Reads

Counts: 2 4 …



Peak calling

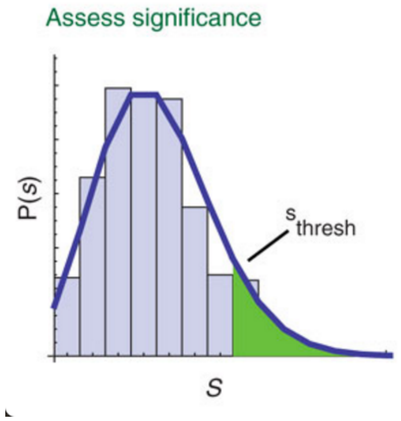
Problem definition: Find genomic regions with more aligned reads than expected by chance.

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 is higher than expected by change

Aligned Reads

Counts: 2 4 …



Peak calling

Problem definition: Find genomic regions with more aligned reads than expected by chance.

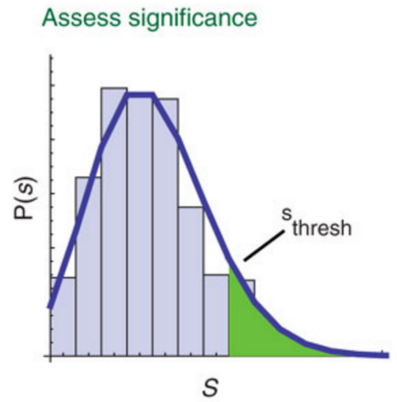
Example of a s

1.use a fix win
obtain a distri
2. define a stat
reads is higher than expected by change

Problems:
- which window size to use?
- distinct proteins have distinct peak sizes
- proper quantification of read counts require several further steps: fragment size estimation, CG bias correction, mappability, …

gned Reads

4 …



Peak calling with MACS2

MACS2

- Models the reads count using a Poisson distribution
  - Only one parameter λ which models mean and variance
  - Estimate a dynamic background reads distribution to capture local biases in the genome, allowing for more robust identification.
- Peaks are defined given a p-value on the Poisson model

More information

- Paper: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-9-r137

Peak calling

2 minutes

4. Motif Matching

Institute for Computational Genomics
01011011010101
10100100101
RWTH AACHEN UNIVERSITY

Motif matching

Problem definition: Matches a set of transcription factor motifs against a set of genomic regions.

Regulatory Genomic Toolbox (RGT)

- is an open-source python library and set of tools for the integrative analysis of high throughput regulatory genomic data
- provides a flexible framework to perform operations related to motif analyses

More information

- Website: https://www.regulatory-genomics.org/

Motif matching

2 minutes

5. Visualizaiton

Institute for Computational Genomics
01011011010101
10100100101
RWTH AACHEN UNIVERSITY

Single cell ATAC-seq

R : https://satijalab.org/signac/index.html
Python: https://episcanpy.readthedocs.io/en/latest/