Practical Example: Analysis of Open Chromatin Data

Ivan Gesteira Costa & Zhijian Li
Institute for Computational Genomics

Institute for Computational Genomics
RWTH AACHEN UNIVERSITY

Contact Information

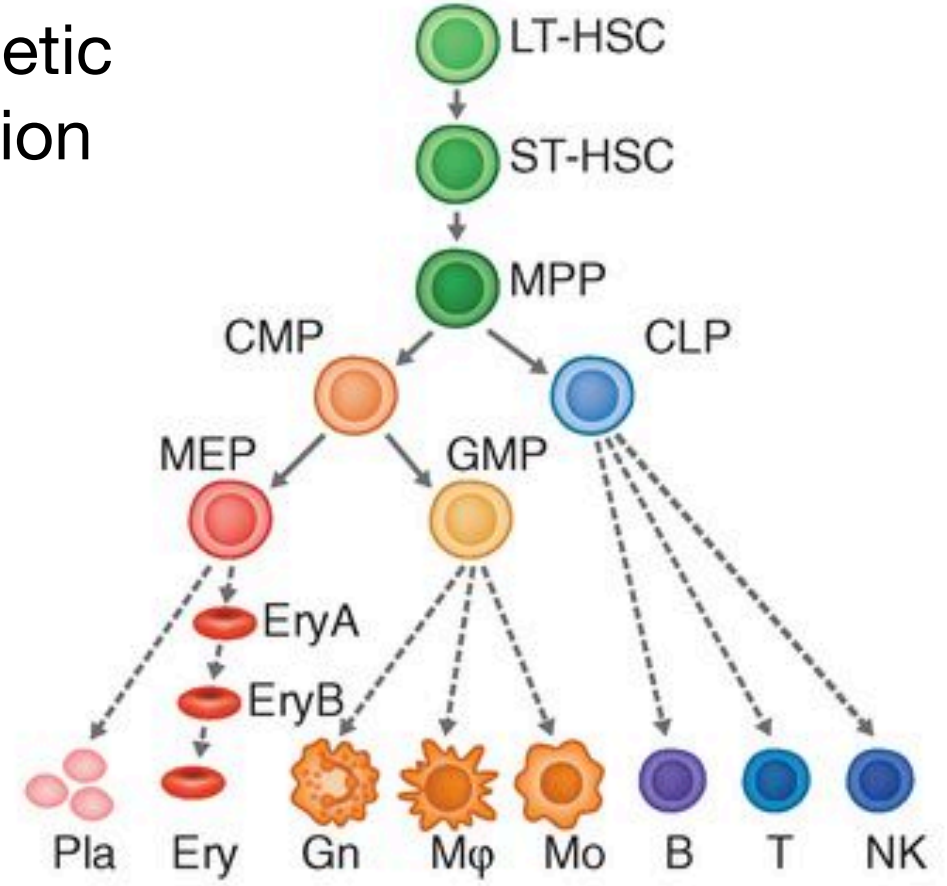
Zhijian Li

zhijian.li@rwth-aachen.de

MTZ, Room 3.02

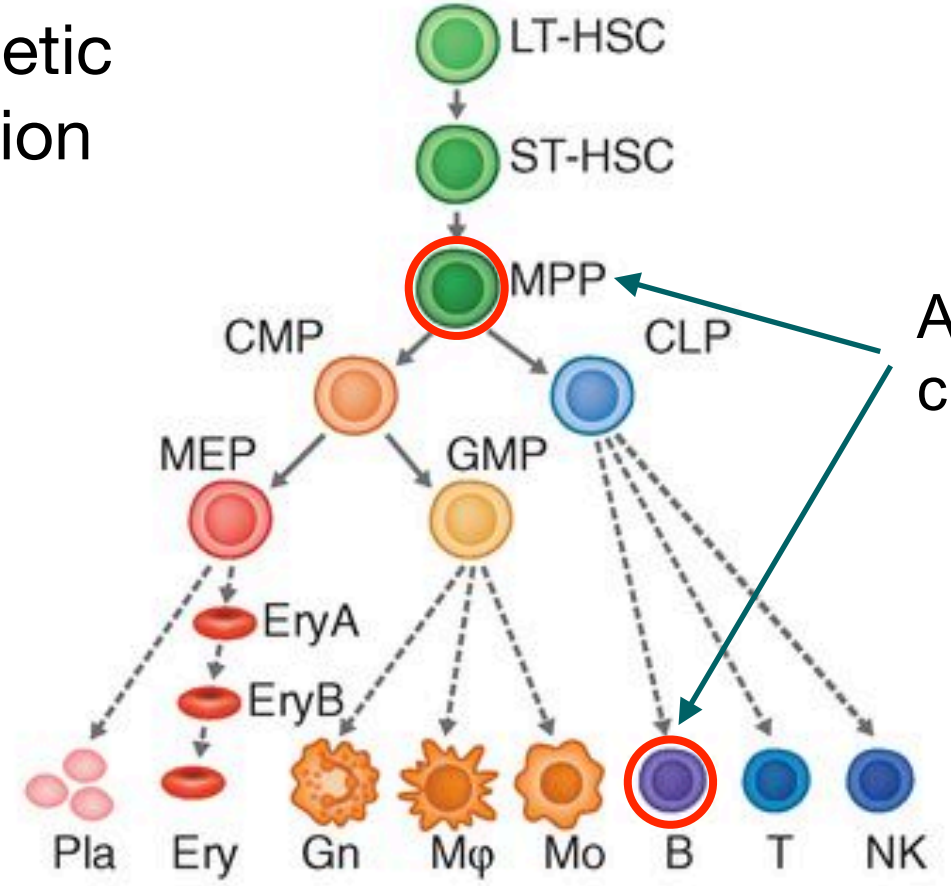
Chromatin dynamics during blood formation

hematopoietic differentiation



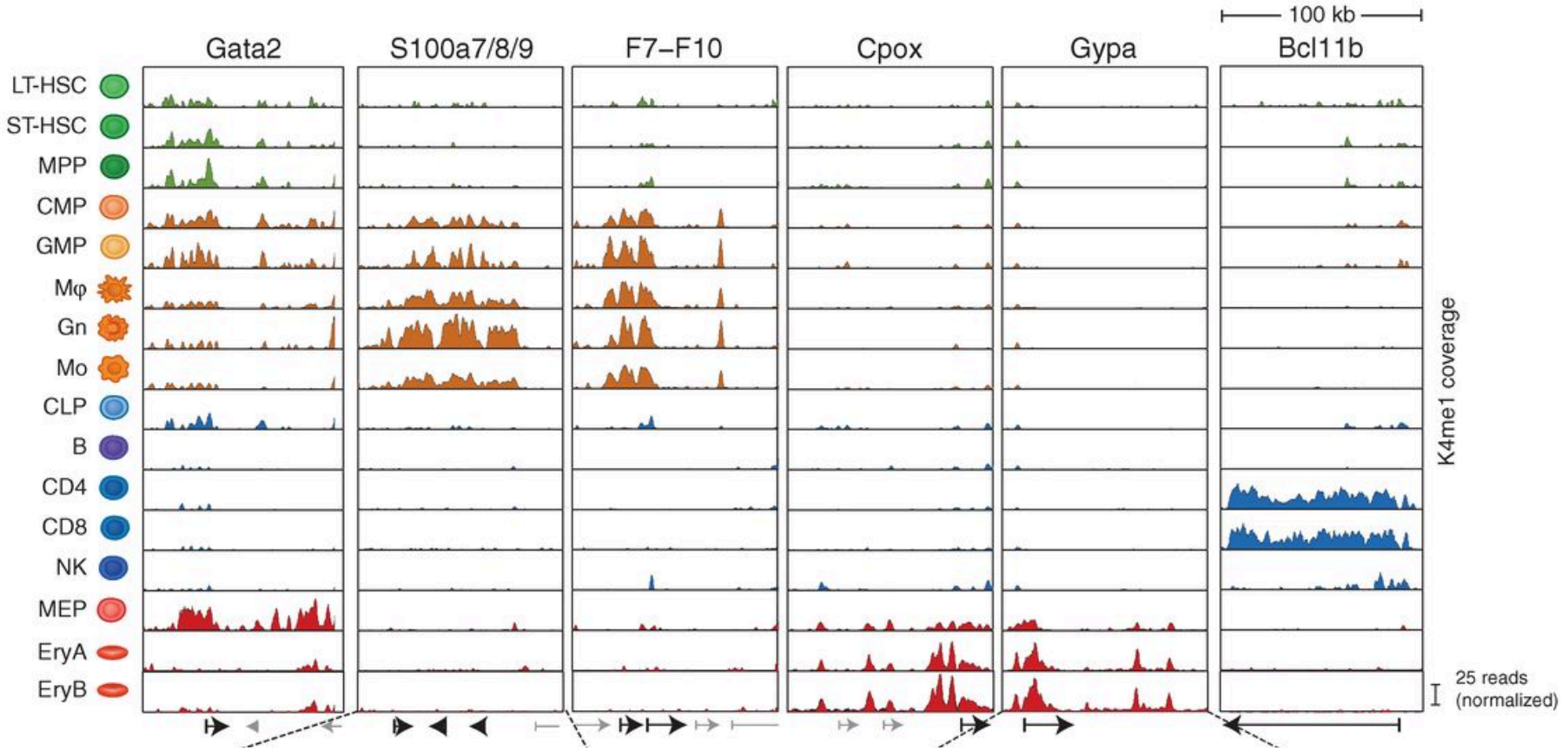
Chromatin dynamics during blood formation

hematopoietic differentiation

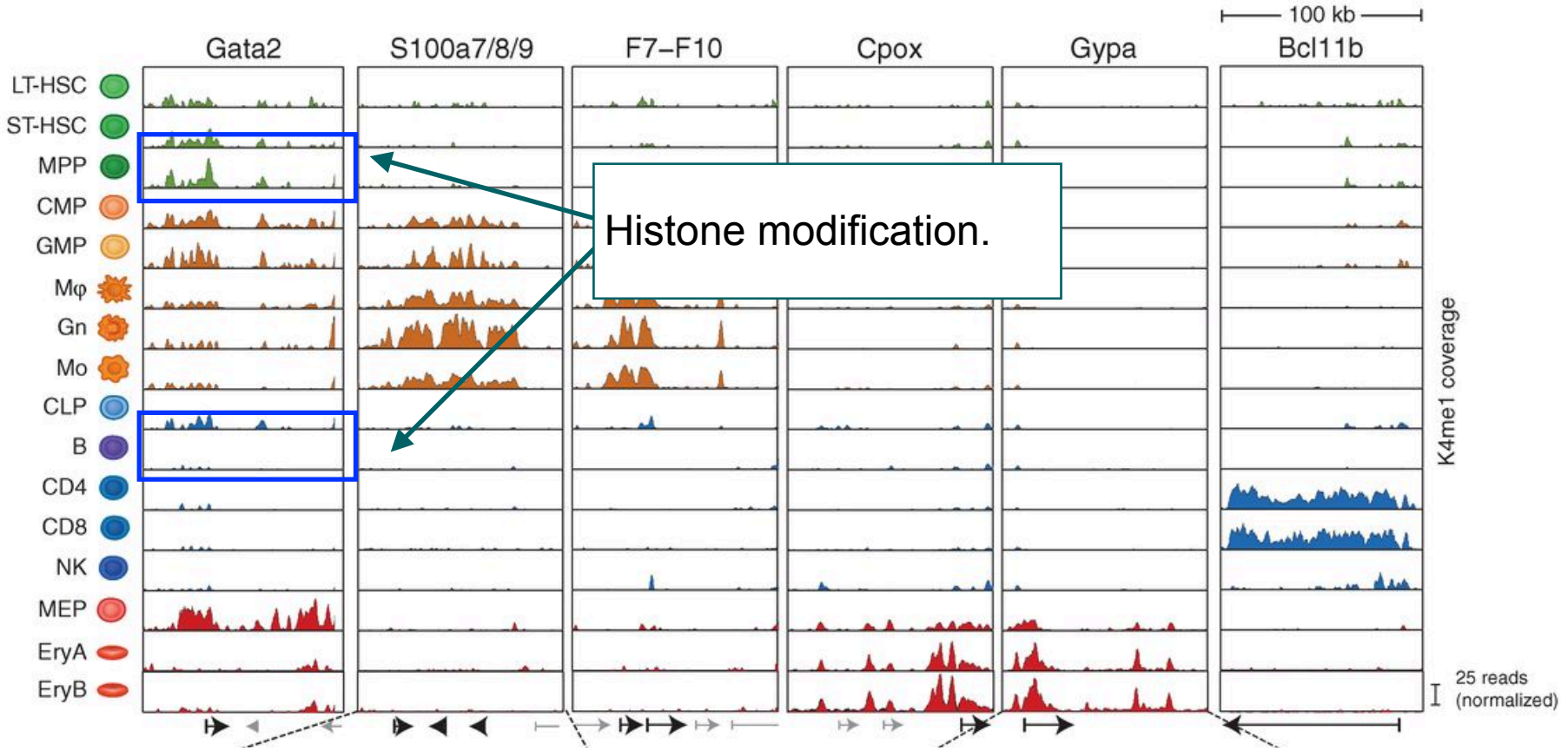


Any differences at chromatin level?

Chromatin dynamics during blood formation



Chromatin dynamics during blood formation



Analysis pipeline

- Download sequencing data
  - SRA toolkit, FASTA, FASTQ and SRA file
- Sequence alignment
  - Bowtie2, samtools, SAM and BAM file
- Peak calling
  - MACS
- Footprinting and motif analysis
  - RGT, bed file

1. Download Data

SRA tools

- A collection of tools and libraries for using data in NCBI Sequence Read Archives

- More information:
  - https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=std

FASTA File

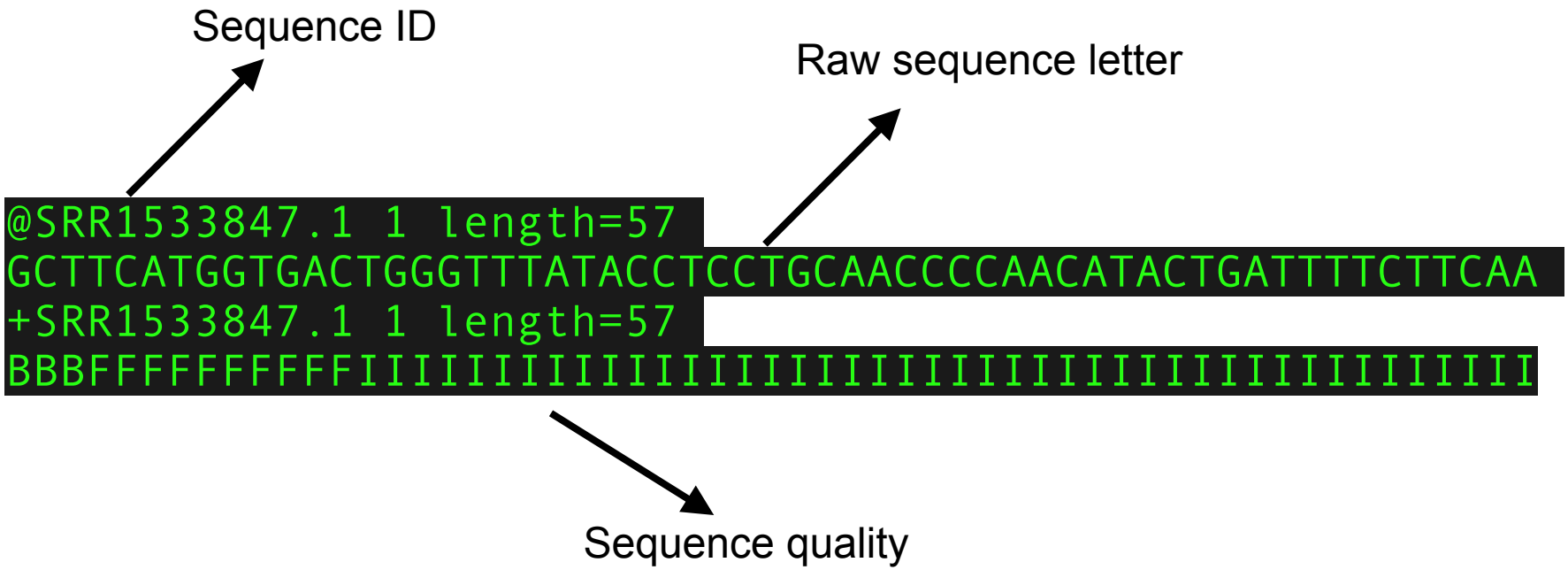
- Store DNA sequences in a text-based file
- Mainly used to store large genomic sequences
- Header (lines that start with ‘>’) + DNA sequence
- DNA alphabeta: A, C, G, T, N

```
>SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
```

FASTQ File

- Also text-based file
- Mainly used to store short DNA sequences
- Normally use four lines per sequence
  - Line 1 begins with '@' and is followed by a sequence ID
  - Line 2 is the raw sequence letter
  - Line 3 begins with a '+' character and is followed by the same sequence ID
  - Line 4 encodes the sequencing quality values

FASTQ File: example



SRA File

- A compressed version of FASTQ file

Download data

- Download and unpack our reference sequence
  - wget http://hgdownload.soe.ucsc.edu/goldenPath/mm10/chromosomes/chr19.fa.gz
  - gunzip chr19.fa.gz
- Download sequencing reads
  - prefetch SRR1533863 SRR1533847
- Use SRA toolkit to convert SRA to FASTQ
  - fastq-dump ~/ncbi/public/sra/SRR1533847.sra
  - fastq-dump ~/ncbi/public/sra/SRR1533863.sra

Practice for data download

5 minutes

2. Short DNA Sequence Alignment

Sequence Alignment

- Input data
  - A large reference sequence (chr19.fa)
  - Millions of short DNA reads (SRR1533863.fastq, SRR1533847.fastq)
- Sequence alignment:
  - Find most probable position for each read in the genome (allow insertion and deletion)

Bowtie2

- Align reads to the genome:
  - Extract ‘seed’ substrings from the read
  - Align the substrings to the reference
  - Calculate the position information
  - Extend the seeds to full alignments using dynamic programming
- More information:
  - Paper: https://www.nature.com/articles/nmeth.1923
  - Website: http://bowtie-bio.sourceforge.net/bowtie2

Samtools

- Provides various utilities for manipulating alignments in SAM format
- More information:
  - Paper: https://www.ncbi.nlm.nih.gov/pubmed/19505943
  - Website: http://www.htslib.org/doc/samtools.html

Perform alignment

- Build genome's index:
  - bowtie2-build chr19.fa chr19
- Align reads to the genome:
  - bowtie2 -x ./chr19 - U MPP.fastq -S MPP.sam
  - bowtie2 -x ./chr19 - U B.fastq -S B.sam

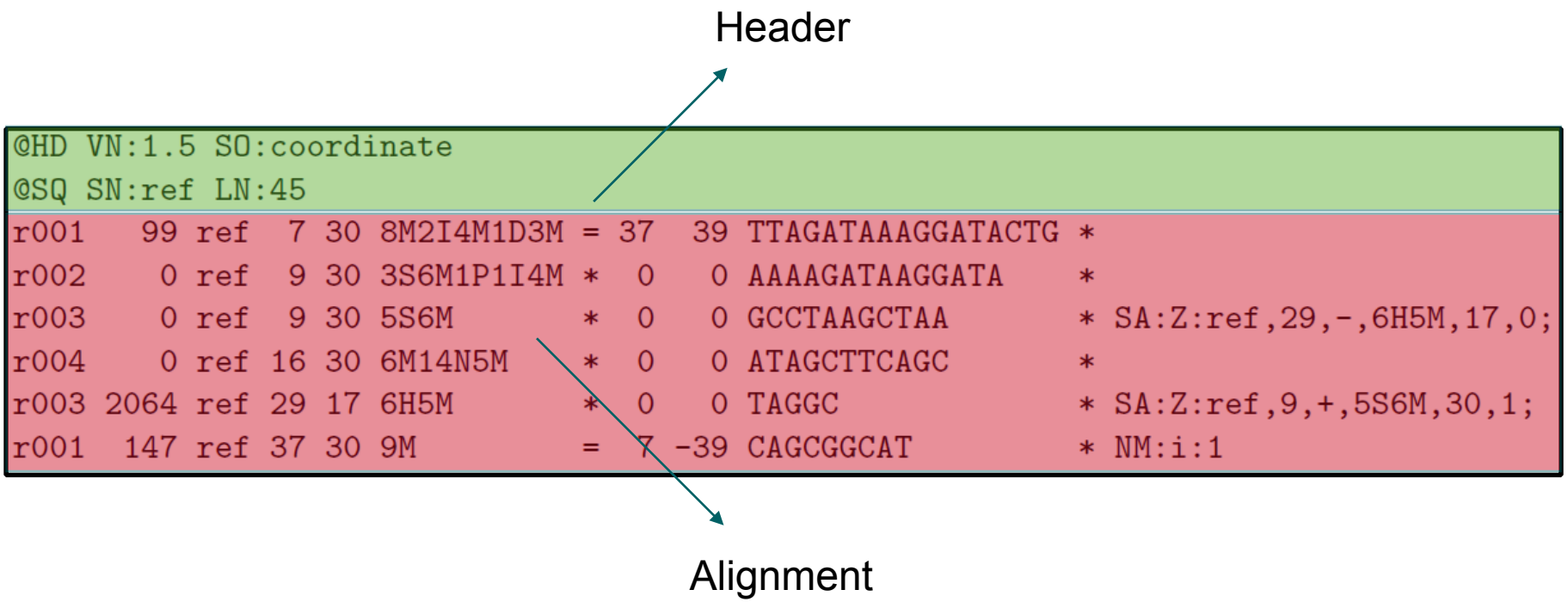
Perform alignment

- Build genome's index:
  - bowtie2-build chr19.fa chr19
- Align reads to the genome:
  - bowtie2 -x ./chr19 - U MPP.fastq -S MPP.sam
  - bowtie2 -x ./chr19 - U B.fastq -S B.sam

Use -p to speed up your alignment!

SAM file

- Store DNA sequences in a text-based file
- Mainly used to store aligned sequences
- Consists of a header and an alignment section



SAM Header

- Begins with character '@' followed with some tags
  - @HD - Header line
  - @SQ - Reference genome information.
  - @RG - Group information
  - @PG - Program (software) information.
  - @CO – Commentary line.

SAM Alignment

- Includes mandatory and optional fields

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Mandatory

Optional

- More information:
  - https://samtools.github.io/hts-specs/SAMv1.pdf

BAM file

- Binary Alignment/Map format - compressed version of SAM
- Efficient random access
- BAI index files

Manipulating alignments

- Convert SAM to BAM
  - samtools view -bS MPP.sam > MPP.bam
  - samtools view -bS B.sam > B.bam
- Sort BAM
  - samtools sort MPP.bam MPP.sorted
  - samtools sort B.bam B.sorted
- Remove low map quality reads
  - samtools view -bq 30 MPP.sorted.bam > MPP.final.bam
  - samtools view -bq 30 B.sorted.bam > B.final.bam

Manipulating alignments

- You can download the results here:
  - http://134.130.18.8/open_data/bioinfolab_2018/Practices/MPP.final.bam
  - http://134.130.18.8/open_data/bioinfolab_2018/Practices/B.final.bam
- Create index files
  - samtools index MPP.final.bam
  - samtools index B.final.bam

IGV visualization

IGV visualization

Practice for short DNA sequence alignment

10 minutes

3. Peak Calling

Peak Calling

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

Aligned Reads

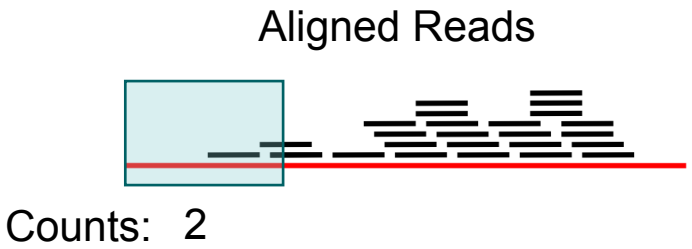
Example of a simple peak caller :

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.



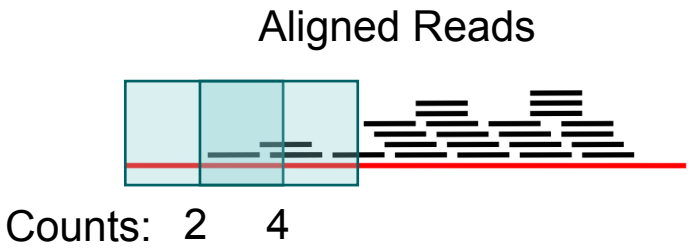
Example of a simple peak caller :

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.



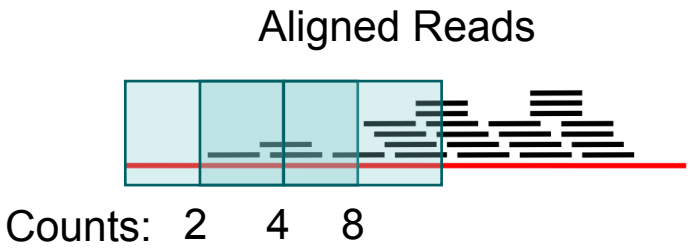
Example of a simple peak caller :

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.



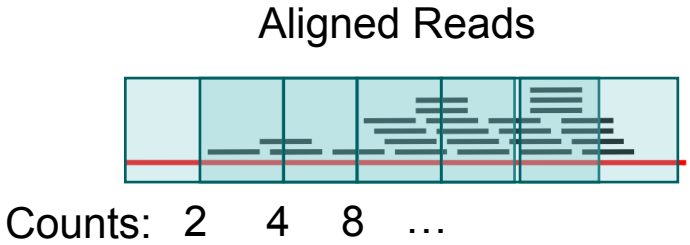
Example of a simple peak caller :

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

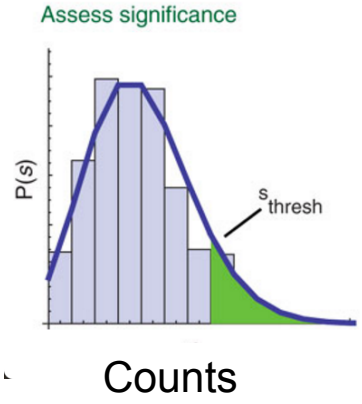
Problem definition: Find genomic regions (of arbitrary size) with more aligned reads than expected by chance.

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

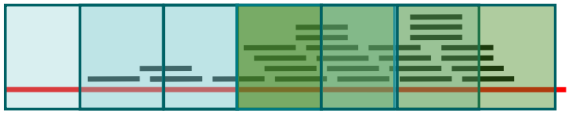


See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

Peak Calling

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

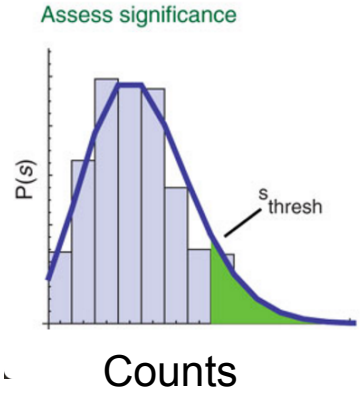
Aligned Reads



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/

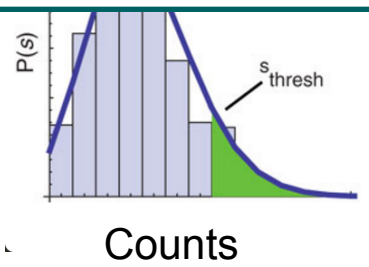
Peak Calling

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

Aligned Reads

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, …

if the number of reads in higher than expected by change

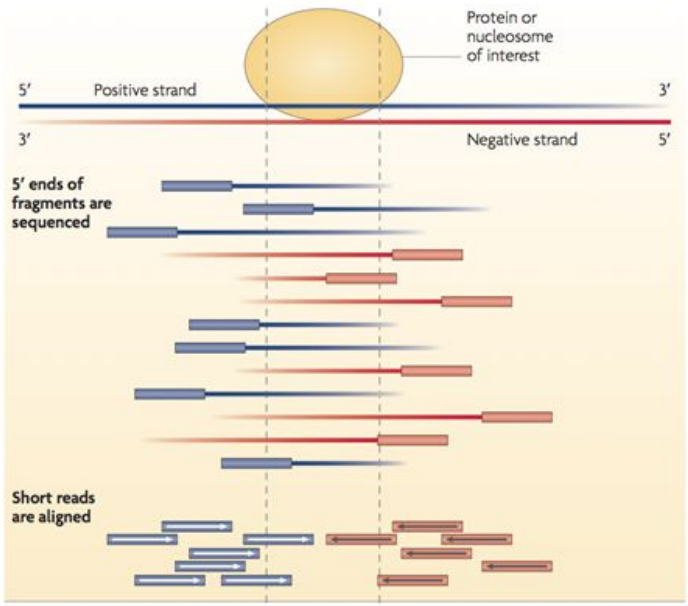


See for an example of a code for a peak caller
http://www.regulatory-genomics.org/rgt/tutorial/implementing-your-own-peak-caller/

MACS Peak Caller

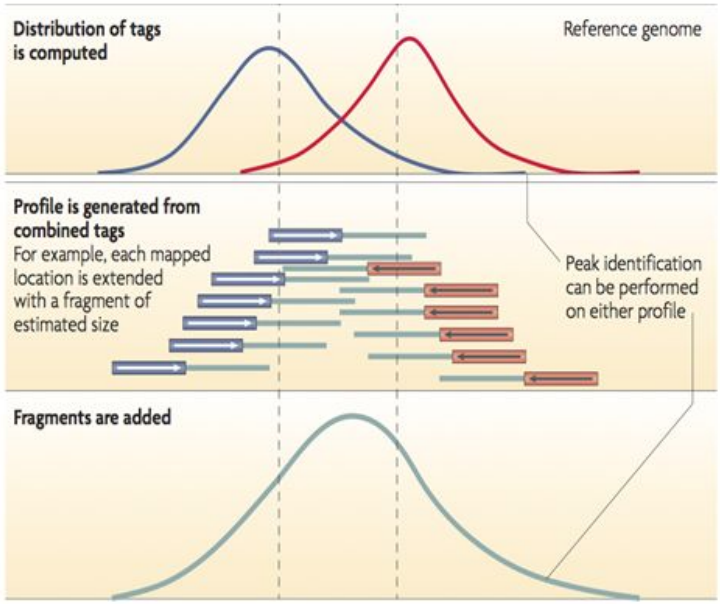
- Model-based Analysis of ChIP-seq
- Two important steps
  - models the shift size of ChIP-seq reads and uses it to improve the spatial resolution of inferred TF binding sites
  - estimates a dynamic background reads distribution to effectively capture local biases in the genome, allowing for more robust identifications

MACS Peak Caller

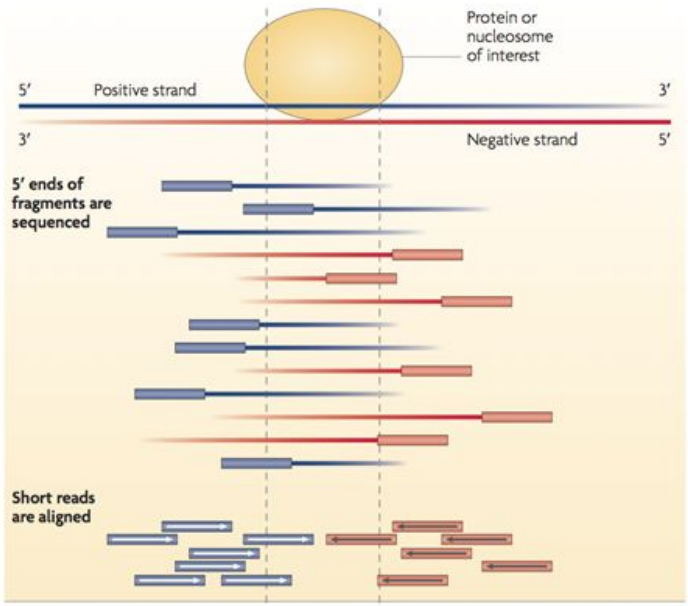


ChIP-seq peaks

- Only 5’ end of fragments are sequenced
- Tags from both + and - strand aligned to reference genome

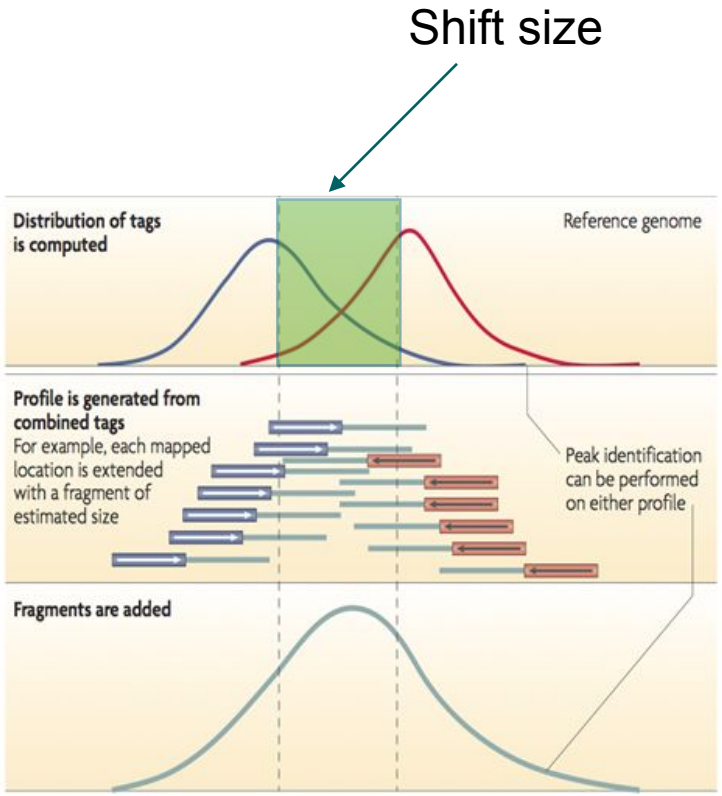


MACS Peak Caller



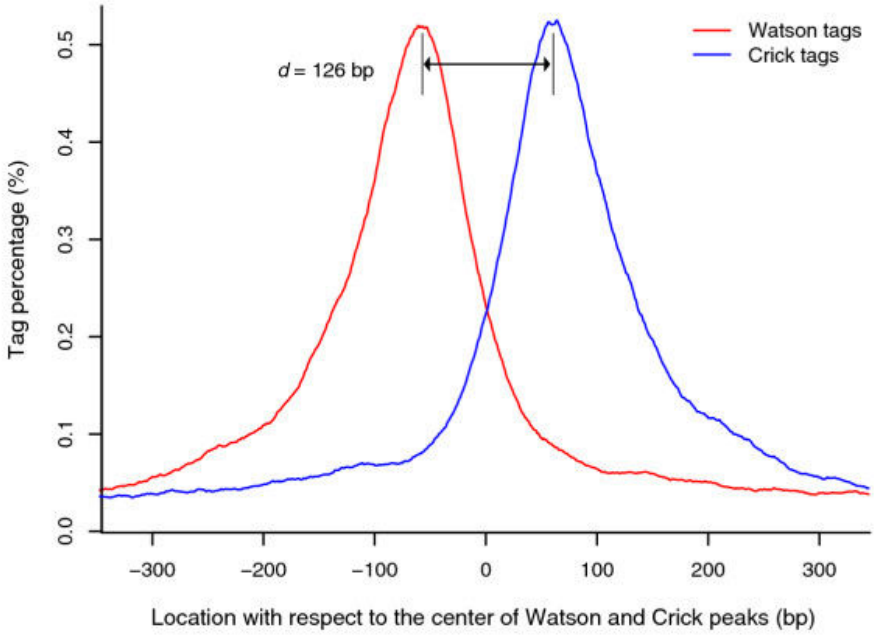
- Only 5’ end of fragments are sequenced
- Tags from both + and - strand aligned to reference genome

ChIP-seq peaks

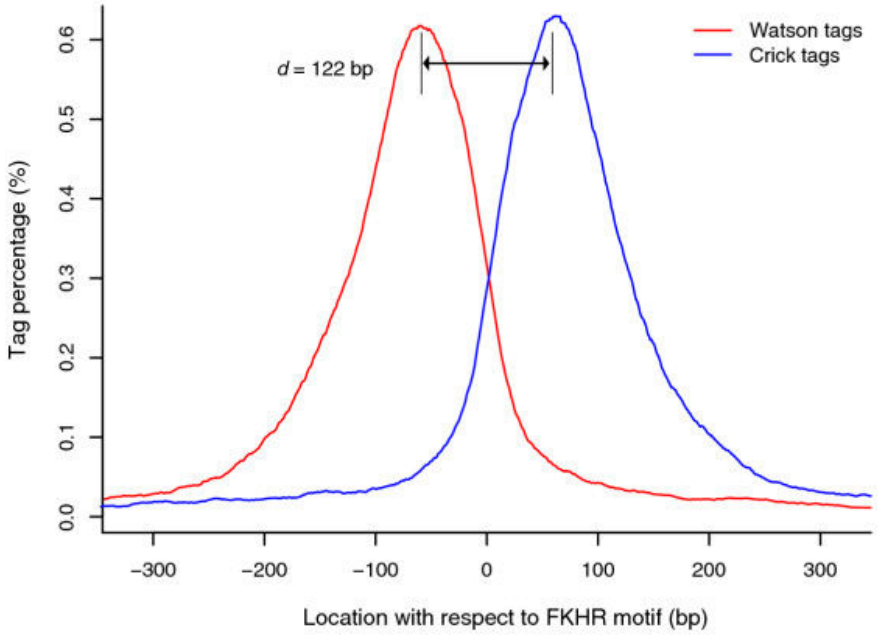


MACS Peak Caller

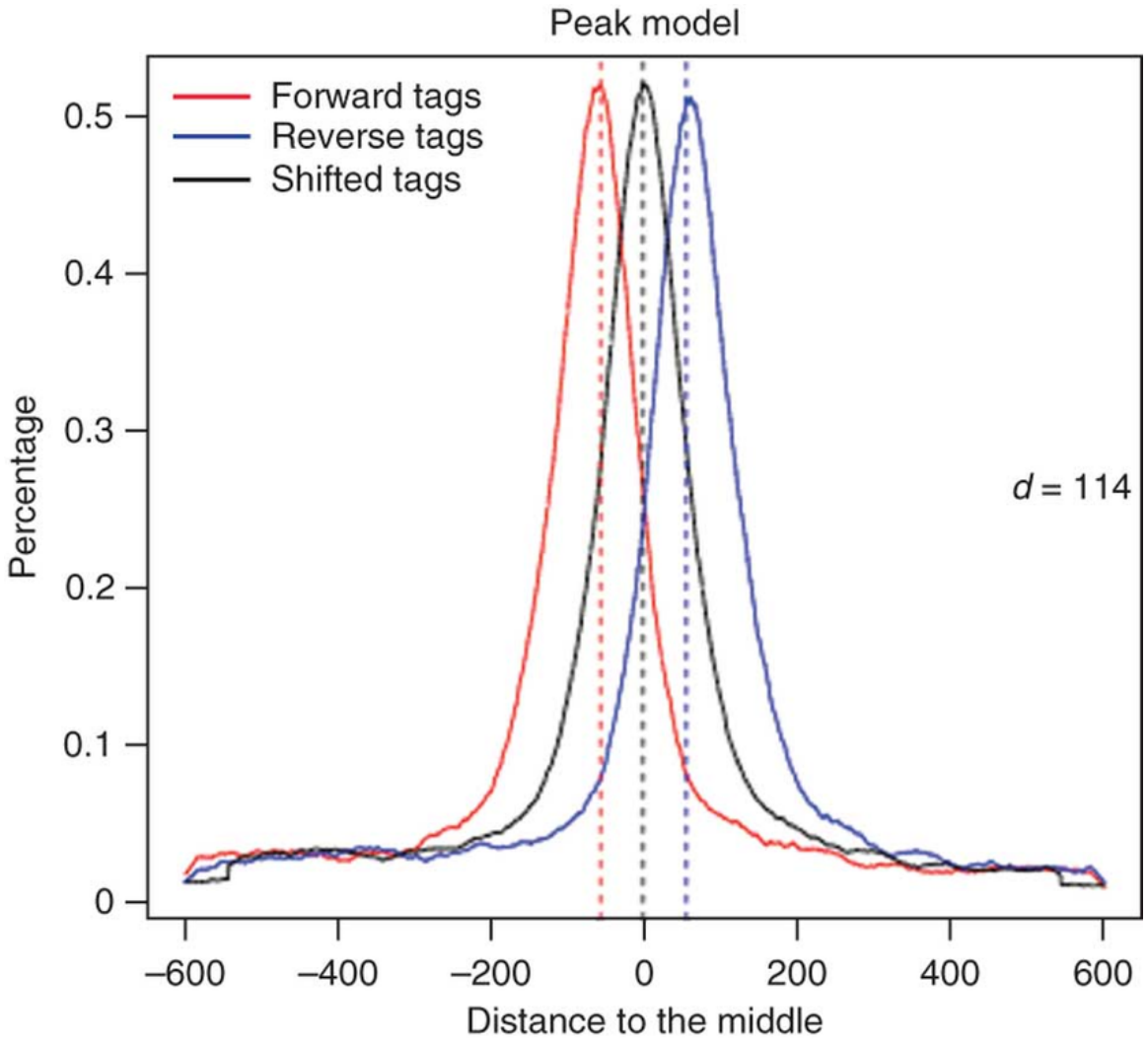
Estimated shift size



shift size based on motif



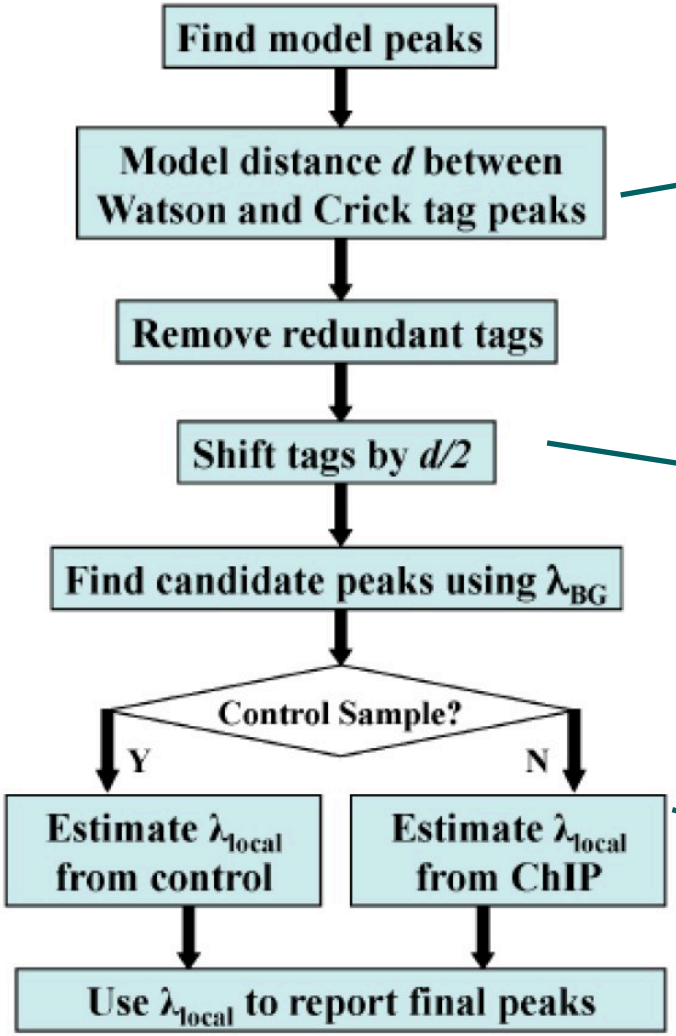
MACS Peak Caller

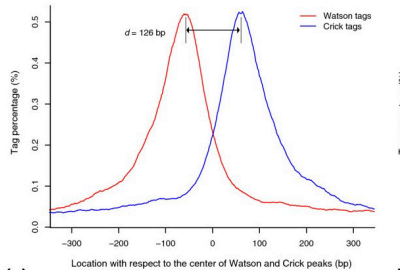


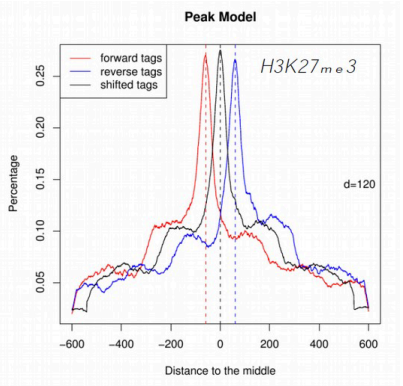
MACS Peak Caller

- Model the reads using a Poisson distribution
- Advantage: only one parameter (λ) which models mean and variance
- Peaks are defined given a p-value on the Poisson model

MACS Peak Caller







$$\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k},] \lambda_{5k}, \lambda_{10k})$$

Perform peak calling with MACS

```
mkdir PeaksMPP
macs2 callpeak -t MPP.final.bam -n MPP --outdir PeaksMPP -g mm
```

Treatment file

Experiment name

Output directory

Genome (necessary to calculate length)

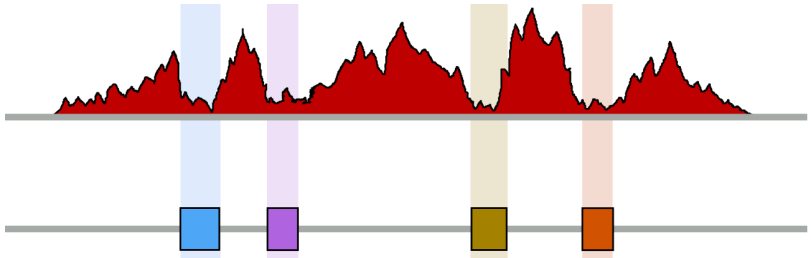
```
mkdir PeaksB
macs2 callpeak -t B.final.bam -n B --outdir PeaksB -g mm
```

4. Footprint & Motif Analysis

Institute for
Computational Genomics
RWTH AACHEN
UNIVERSITY

Digital Footprinting

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



BED file: Storing genomic regions

- Text-based tab-delimited file to store genomic signals.
- Fields:
  - chrom: The name of chromosome
  - chromStart: The starting position of the coordinate (start = 0)
  - chromEnd: The ending position of the coordinate
  - name: Label of the coordinate.
  - score: A score between 0 and 1000.
  - Strand: either '+' or '-'
- Example:

chr1	714057	714099	chr1:714057-714099	424	+
chr1	714102	714120	chr1:714102-714120	463	-
chr1	714121	714135	chr1:714121-714135	473	+
chr1	714137	714148	chr1:714137-714148	429	-
chr1	714220	714228	chr1:714220-714228	419	+

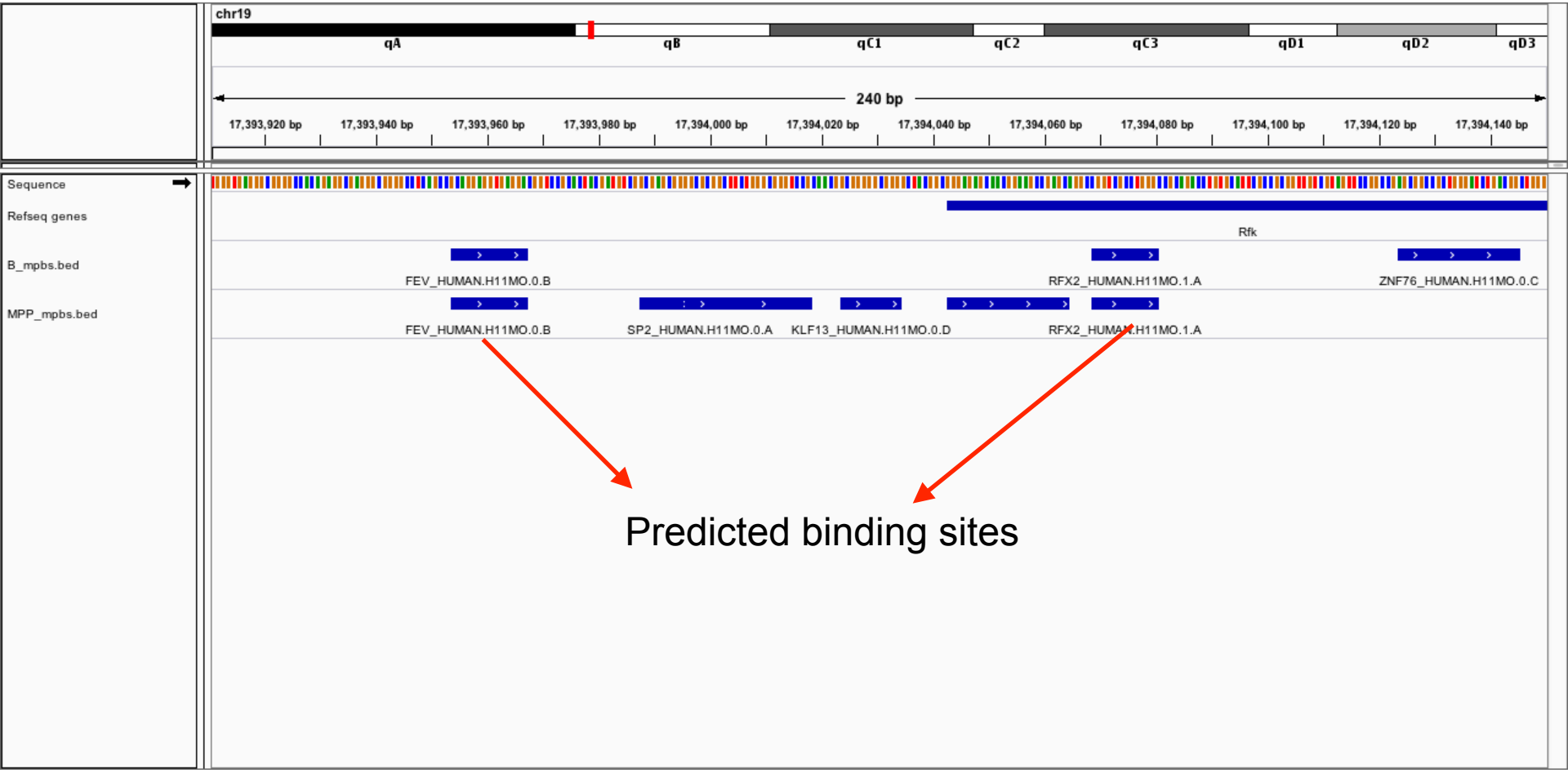
Detect Footprints with HINT

- Detect footprints using HINT:
  - rgt-hint footprinting --atac-seq --organism mm10 --output-prefix=MPP MPP.final.bam ./PeaksMPP/ MPP_peaks.narrowPeak
  - rgt-hint footprinting --atac-seq --organism mm10 --output-prefix=B B.final.bam ./PeaksB/ B_peaks.narrowPeak

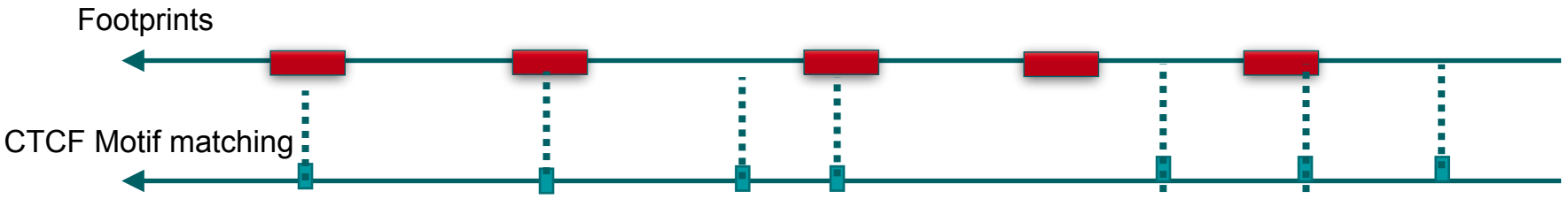
Motif matching

- Motif matching using RGT:
  - rgt-motifanalysis matching --organism mm10 --rand-proportion 10 --input-files B.bed MPP.bed

Visualization

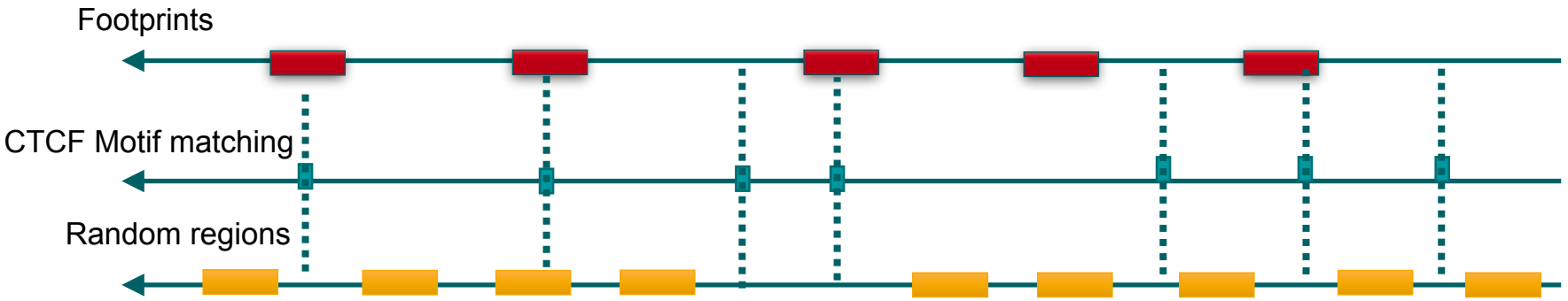


Motif enrichment

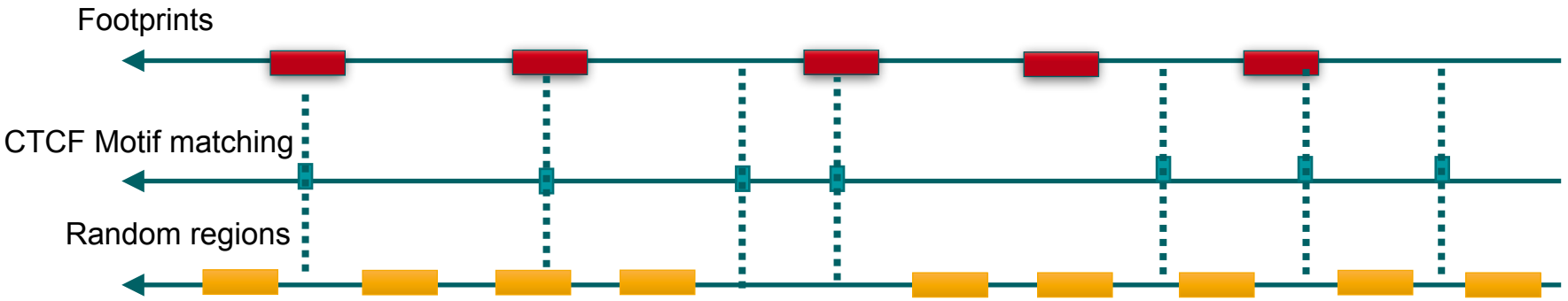


Is CTCF more likely to bind in those regions than in background regions?

Motif enrichment

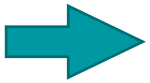


Motif enrichment

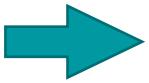


- Performs Fisher's exact test in order to verify if a set of genomic regions are enriched for particular transcription factors.

	With	w/o
Footprints	**a**	**b**
Random	**c**	**d**



	With	w/o
Footprints	**4**	**1**
Random	**1**	**8**



Fisher's exact test

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

P = 0.022

Perform motif enrichment

- rgt-motifanalysis enrichment --organism mm10 ./match/random_regions.bed B.bed MPP.bed

Perform motif enrichment

Regulatory Genomics Toolbox - Motif Enrichment Analysis

B MPP

Results for **B** region **Site Test*** using all input regions

* This test considered all input regions against background regions

FACTOR	MOTIF	P-VALUE	CORRECTED P-VALUE	A	B	C	D	FREQUENCY	BACKGROUND FREQUENCY	GO
KLF12_HUMAN.H11MO.0.C		4.8159e-40	3.7131e-37	72	4214	65	45309	1.68%	0.14%	View
SP1_HUMAN.H11MO.1.A		9.7376e-35	3.7538e-32	56	4230	40	45334	1.31%	0.09%	View
MAZ_HUMAN.H11MO.1.A		1.6060e-31	4.1274e-29	85	4201	151	45223	1.98%	0.33%	View
SP2_HUMAN.H11MO.1.B		4.0513e-29	7.6943e-27	70	4216	107	45267	1.63%	0.24%	View
KLF9_HUMAN.H11MO.0.C		4.9898e-29	7.6943e-27	52	4234	48	45326	1.21%	0.11%	View
E2F4_HUMAN.H11MO.0.A		7.0187e-25	9.0190e-23	33	4253	14	45360	0.77%	0.03%	View
KLF1_HUMAN.H11MO.0.A		6.9452e-24	7.6497e-22	56	4230	84	45290	1.31%	0.19%	View

Institute for Computational Genomics

RWTH AACHEN UNIVERSITY

The data is available on:
http://134.130.18.8/open_data/bioinfolab_2018/Practice.tar.gz

Thank you!