Genome-scale technologies 2/ Algorithmic and statistical aspects of DNA sequencing

Open chromatin (continued)

Ewa Szczurek
University of Warsaw, MIMUW

szczurek@mimuw.edu.pl

Sequencing-based tracking of open chromatin

Open chromatin

←→ repositioning of nucleosomes

←→ regulatory regions

- Open chromatin profiling:
  1. MNase-seq
  2. DNase-seq
  3. Faire-seq
  4. ATAC-seq

- No antibody required

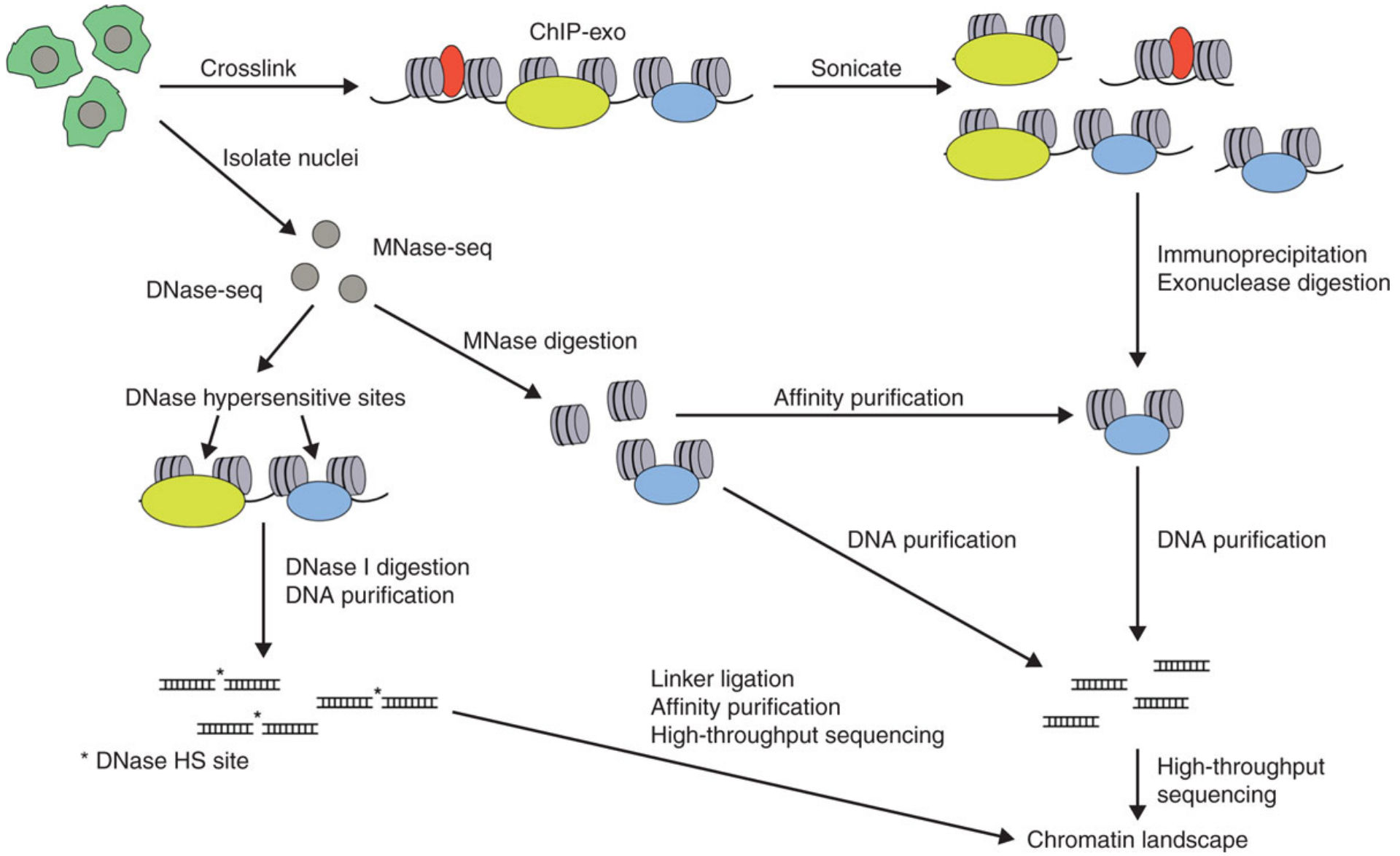
Micrococcal Nuclease (MNase): endo-exonuclease

- preferentially digests single-stranded nucleic acids
- but also double-stranded DNA and RNA
- MNase continues to digest the exposed DNA ends until it reaches an obstruction, such as
  - a nucleosome,
  - a stably bound TF
  - or a refractory DNA sequence

- MNase-seq:
  - Chromatin crosslinked with formaldehyde
  - MNase fragments all accessible chromatin.
  - MNase-protected DNA is sequenced

- Mnase-seq: a nucleosome occupancy assay,
- Dnase-seq: free chromatin assay

MNase-seq *versus* DNase-seq



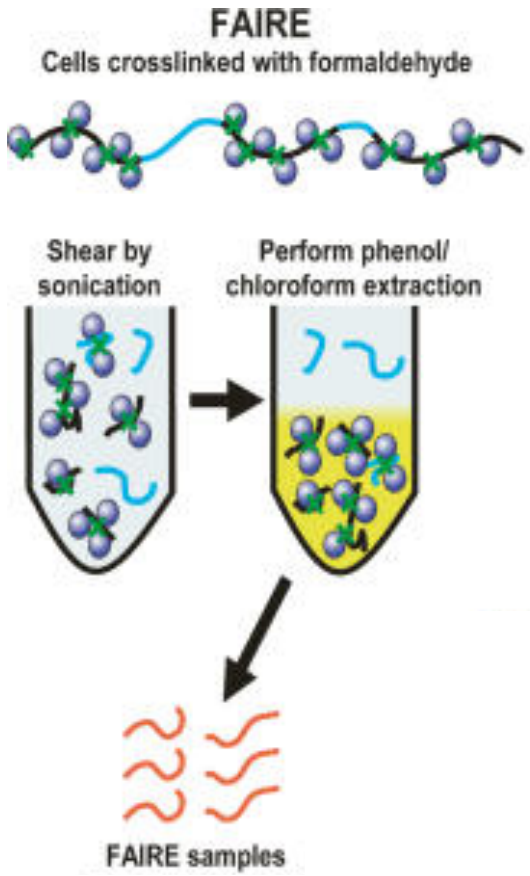
Mnase-seq

- High degree of AT-cleavage specificity
- Differential levels of enzymatic digestion - different outputs
- Affected by MNase *titration* (concentration)
- Affected by linker length
- Ideal if 80-90% mononucleosomes (high digestion level)

Formaldehyde-Assisted Isolation of Regulatory Elements

- FAIRE

1. crosslinking of chromatin with formaldehyde (capturing protein-DNA interactions)
2. shearing of chromatin with sonication
3. phenol-chloroform extraction
4. nucleosome-depleted DNA → the aqueous phase of the solution
5. histone-bound DNA (high crosslinking efficiency) → organic phase
6. sequencing the chromatin-accessible population of fragments



FAIRE-seq

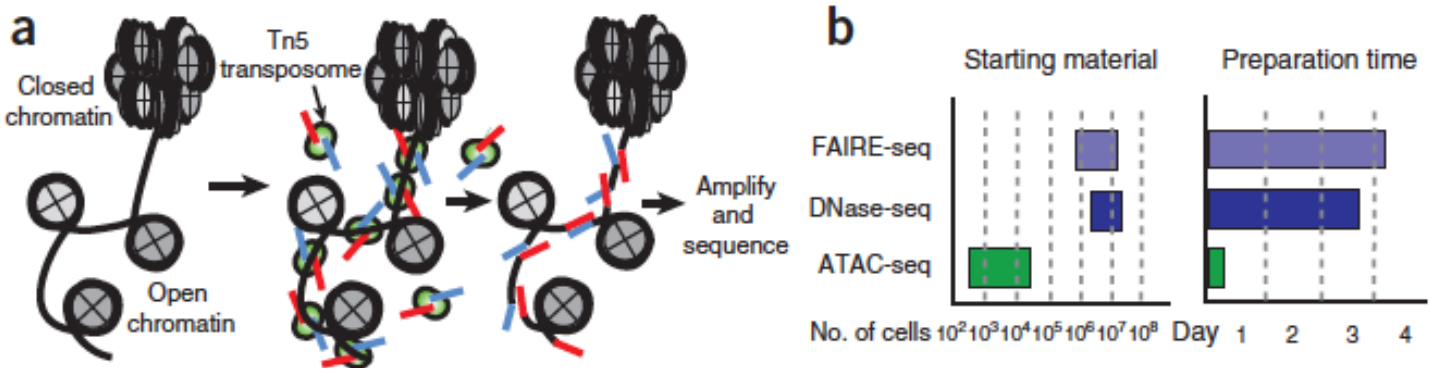
- Pros:
  - negative relationship with nucleosome occupancy
  - overlap with various cell type-specific marks of active chromatin
  - Applicable to any cell type or tissue
  - No laborious preparations
  - No sequence specific cleavage bias

- Cons:
  - The experiment heavily depends on fixation efficiency
  - lower signal-to-noise ratio compared to the other assays.
  - only strong recovered signal informative.

ATAC

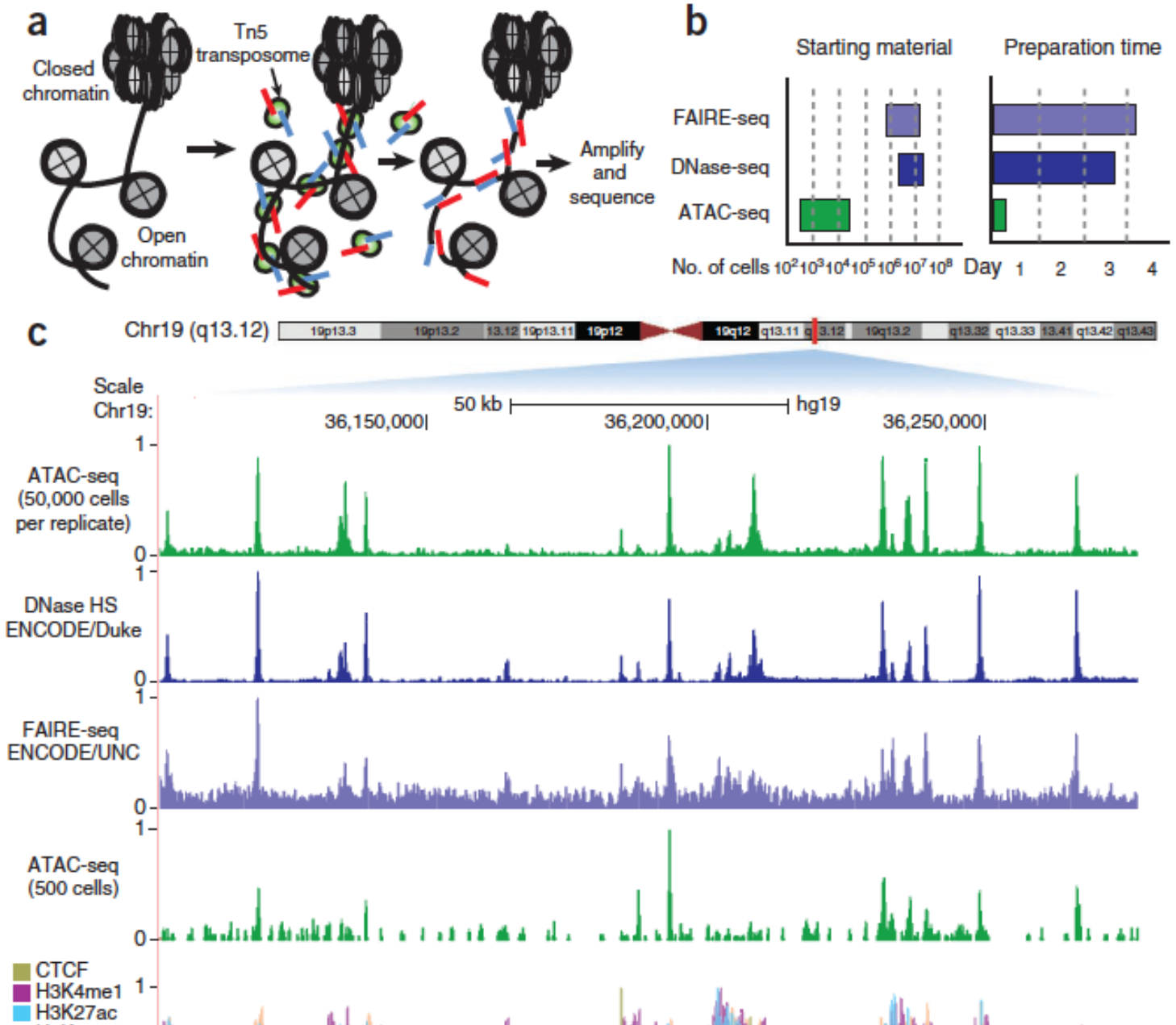
- **A**ssay for **T**ransposase-**A**ccessible **C**hromatin with highthroughput **seq**uencing.
- Transposases are enzymes
  - catalyzing the movement of transposons to other parts in the genome.
  - naturally occurring have a low level of activity
- ATAC-seq employs a mutated hyperactive transposase Tn5.
- The hyperactive transposase Tn5 fragments DNA and integrates into active regulatory regions *in vivo*

ATAC-seq



- The protocol:
  - Transposase (green), loaded with sequencing adaptors (red and blue)
  - inserts only in regions of open chromatin (between nucleosomes in gray)
  - generates sequencing-library fragments that can be PCR-amplified

ATAC-seq


a
Closed chromatin
Tn5 transposome
Open chromatin
Amplify and sequence
b
Starting material
Preparation time
FAIRE-seq
DNase-seq
ATAC-seq
No. of cells 10² 10³ 10⁴ 10⁵ 10⁶ 10⁷ 10⁸
Day 1 2 3 4
c
Chr19 (q13.12)
Scale
Chr19:
50 kb
hg19
36,150,000
36,200,000
36,250,000
ATAC-seq (50,000 cells per replicate)
DNase HS ENCODE/Duke
FAIRE-seq ENCODE/UNC
ATAC-seq (500 cells)
CTCF
H3K4me1
H3K27ac

ATAC-seq

- Sensitivity and specificity of ATAC-seq similar to DNase-seq obtained from approximately three to five orders of magnitude more cells,
- Diminishes only for really small numbers of cells
- The protocol does not involve any size-selection steps
  - Can identify accessible locations and nucleosome positioning simultaneously.
- Ability to map nucleosomes genome-wide limited to regions in close proximity to accessible sites

Chromatin accessibility and nucleosome positioning

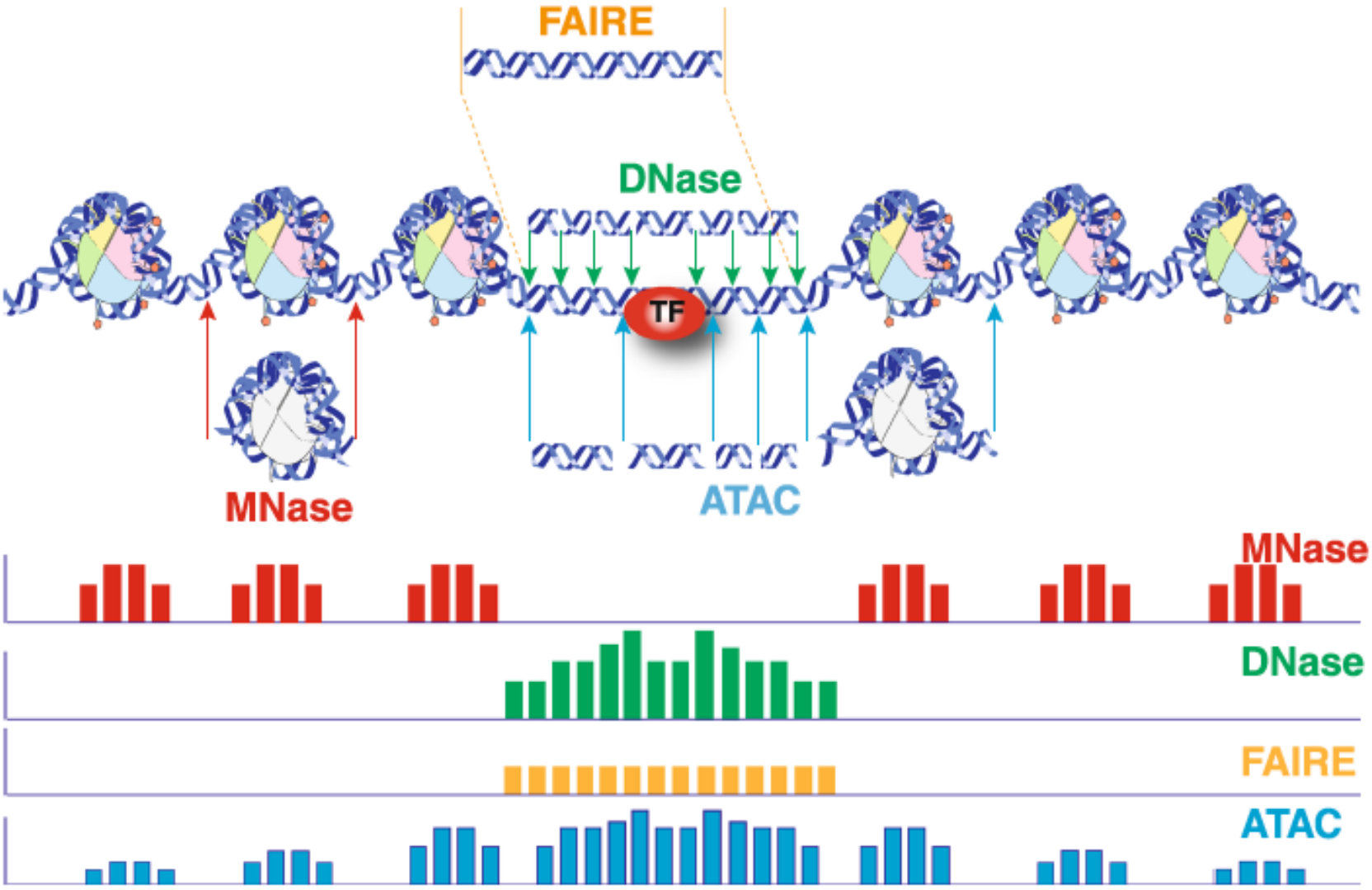


Table 1 Current genome-wide high-throughput chromatin accessibility assays

	Cell type/Number	Sequencing type	Traditional approach	Genomic target	Experimental considerations	Key references
MNase-seq	Any cell type 1 to 10 million cells	Paired-end or Single-end	MNase digests unprotected DNA	Maps the total nucleosome population in a qualitative and quantitative manner	1. Requires many cells. 2. Laborious enzyme titrations. 3. Probes total nucleosomal population, not active regulatory regions only. 4. Degrades active regulatory regions, making their detection possible only *indirectly*. 5. Requires 150 to 200 million reads for standard accessibility studies of the human genome.	[37,46,49]
DNase-seq	Any cell type 1 to 10 million cells	Paired-end or Single-end	DNase I cuts within unprotected DNA	Maps open chromatin	1. Requires many cells. 2. Time-consuming and complicated sample preparations. 3. Laborious enzyme titrations. 4. Requires 20 to 50 million reads for standard accessibility studies of the human genome.	[61,75,76]
FAIRE-seq	Any cell type 100,000 to 10 million cells	Paired-end or Single-end	Based on the phenol-chloroform separation of nucleosome-bound and free sonicated areas of a genome, in the interphase and aqueous phase respectively	Maps open chromatin	1. Low signal-to-noise ratio, making computational data interpretation very difficult. 2. Results depend highly on fixation efficiency. 3. Requires 20 to 50 million reads for standard accessibility studies of the human genome.	[86-90]
ATAC-seq	500 to 50,000 freshly isolated cells	Paired-end	Unfixed nuclei are tagged *in vitro* with adapters for NGS by purified Tn5 transposase. Adapters are integrated into regions of accessible chromatin	Maps open chromatin, TF and nucleosome occupancy	1. Contamination of generated data with mitochondrial DNA. 2. Immature data analysis tools. 3. Requires 60 to 100 million reads for standard accessibility studies of the human genome.	[103]

ATAC: assay for transposase-accessible chromatin; DNase I: deoxyribonuclease I; FAIRE: formaldehyde-assisted isolation of regulatory elements; MNase: micrococcal nuclease.

Sources of bias in chromatin profiling experiments

- A common misconception: the digital readout of NGS read counts gives unbiased results.
- Different sources of bias:

- *Chromatin fragmentation and size selection: sonication.*
  - In ChIP-seq, sonication is required before protein-bound fragments are isolated by immunoprecipitation
  - The mechanical characteristics of chromatin vary across the genome, which creates fluctuations in DNA fragility.
  - a single input sample as a control for ChIP–seq peak calling←→ only if it is not sonicated together with the ChIP sample.
  - Combined control: from many different batches of ChIP–seq experiments produced from the same cell line under consistent conditions

Meyer and Liu, Nat Rev Genet

Sources of bias in chromatin profiling experiments

- *Chromatin fragmentation and size selection: enzymatic cleavage.*
  - MNase cleavage is affected by the cleavage reaction temperature.
  - DNase I cleavage affected by the precise sequence of the three nucleotides on either side of the cleavage site, (strand specific).
  - Also Mnase, and Tn5 transposase cleavage is sequence-specific
  - the enzymes tend to cleave some DNA sequences more efficiently

- *PCR amplification biases and duplications.*
  - Multiple instances of the same sequence read in an NGS data set
  - from sequencing PCR amplicons derived from the same original fragment or
  - from the presence of multiple fragments in the original sample.
  - particularly bad with small amounts of starting material.
  - Due to combination of temperature profile, polymerase and buffer used during PCR
  - a bias towards GC-rich fragments, but not regions with extreme GC
  - bias increases with every PCR cycle

Meyer and Liu, Nat Rev Genet

Sources of bias in chromatin profiling experiments

- ***Read mapping.***
- algorithm-specific biases when finding imperfect or ambiguous matches to the genome.
- algorithm-specific ‘unmappable’ regions of the genome to which no reads can be aligned.

Experimental design

- controls are required to accurately evaluate the effects of biases
  - For ChIP –seq, in input controls, weak TF binding signals may be observed because regions of TF binding also tend to be regions where chromatin is more amenable to fragmentation
  - For Mnase, Dnase and ATAC-seq, controls show cleavage biases
- replicates are needed to make an assessment of data variability
- biologically distinct treatment groups need to be distributed evenly over processing batches so that experimental effects and batch effects can be distinguished
- the experimental protocol needs to be carried out in a highly consistent manner for all samples
- Random barcoding used to distinguish PCR duplicates from duplicates in the unamplified DNA

Nucleosome positioning

- Nucleosome:147-bp DNA wrapping around a histone octamer
- An array of nucleosome units across the genome
- the exact nucleosome positions
  - deviate between different cells
  - center around the most preferred position
- Nucleosome organization characterized by
  - Position (the most preferred)
  - Fuziness (deviation of nucleosome positions within each unit in a cell population)
  - Frequency (with which the position is occupied in a cell population)
- Regulated:
  - By the DNA sequence
  - Dynamically, by environmental factors such as heat shock
  - At the position level (position shifts and fuzziness changes),
  - and/or at the occupancy level.

nucleR: non-parametric nucleosome positioning

- Nucleosomes:
  - well-positioned (phased across different cells)
  - fuzzy (not-phased)
- Non-parametric (no expert knowledge involved) method for nucleosome positioning

- Works with Tiling Array and MNAse-seq data

nucleR: steps

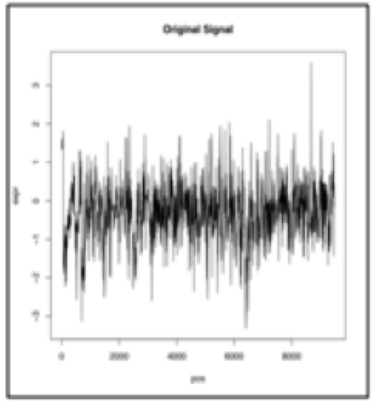
1. Summarize short reads coverage (NGS)
   - correcting the strand bias if working with single-ended sequencing
   - trimming the reads in paired-end cases.
   - Reducing bias due to the sequence preferences of Mnase obtained for nucleosomal DNA are corrected with control (naked DNA).
2. Coverage 'profile cleaning' based on Fourier analysis
   - Transforming the original (complex) profile into the Fourier Space using Fast Fourier Transform (FFT)
   - The signal described as a combination of simple periodic waves
   - Analyzing the contribution of every frequency to the original signal.
   - High frequencies are usually echoes of lower frequencies (sources of noise) → can be removed.
   - 2% of components is left before performing the inverse FFT
   - Smoothens the signal and cleans the distortions at once

nucleR: steps

1. Summarize short reads coverage (NGS)
   - correcting the strand bias if working with single-ended sequencing
   - trimming the reads in paired-end cases.
   - Reducing bias due to the sequence preferences of Mnase obtained for nucleosomal DNA are corrected with control (naked DNA).
2. Coverage 'profile cleaning' based on Fourier analysis
   - Transforming the original (complex) profile into the Fourier Space using Fast Fourier Transform (FFT)
   - The signal described as a combination of simple periodic waves
   - Analyzing the contribution of every frequency to the original signal.
   - High frequencies are usually echoes of lower frequencies (sources of noise) → can be removed.
   - 2% of components is left before performing the inverse FFT
   - Smoothes the signal and cleans the distortions at once
3. Detection of nucleosomes:
   - High score to large and sharp peaks, penalizing fuzziness

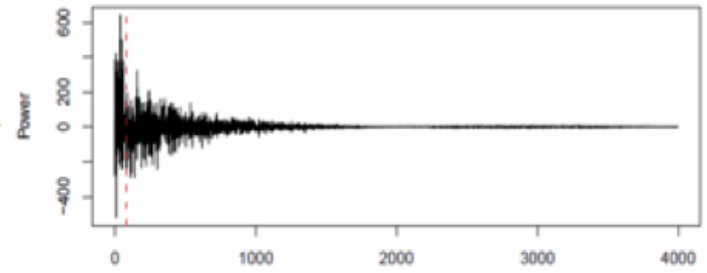
FFT

Original signal



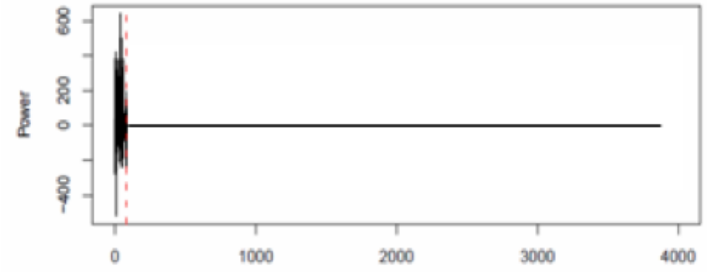
FFT

Frequency spectrum



K.O.

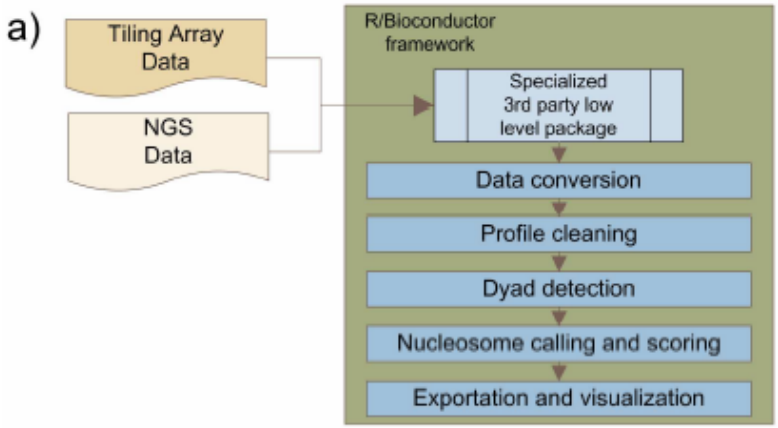
Cutting out the high frequencies



FFT^{-1}

Smoothed signal

nucleR:steps



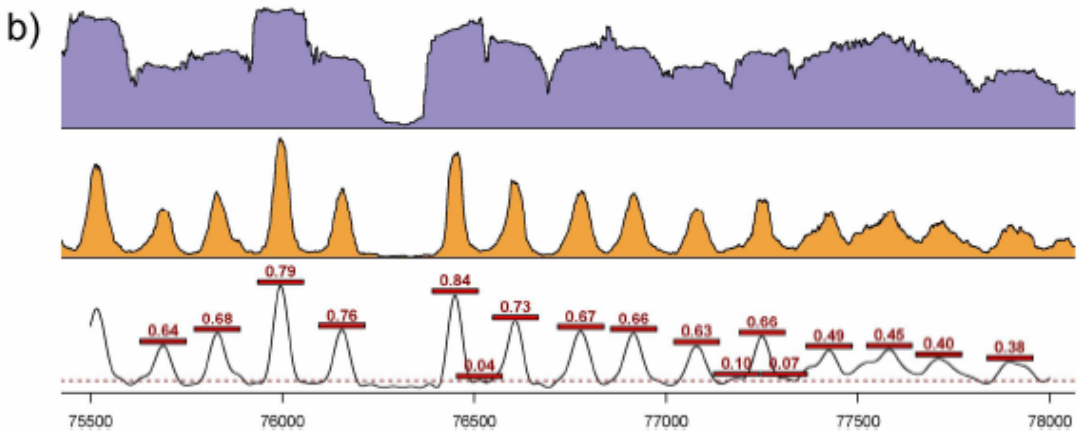
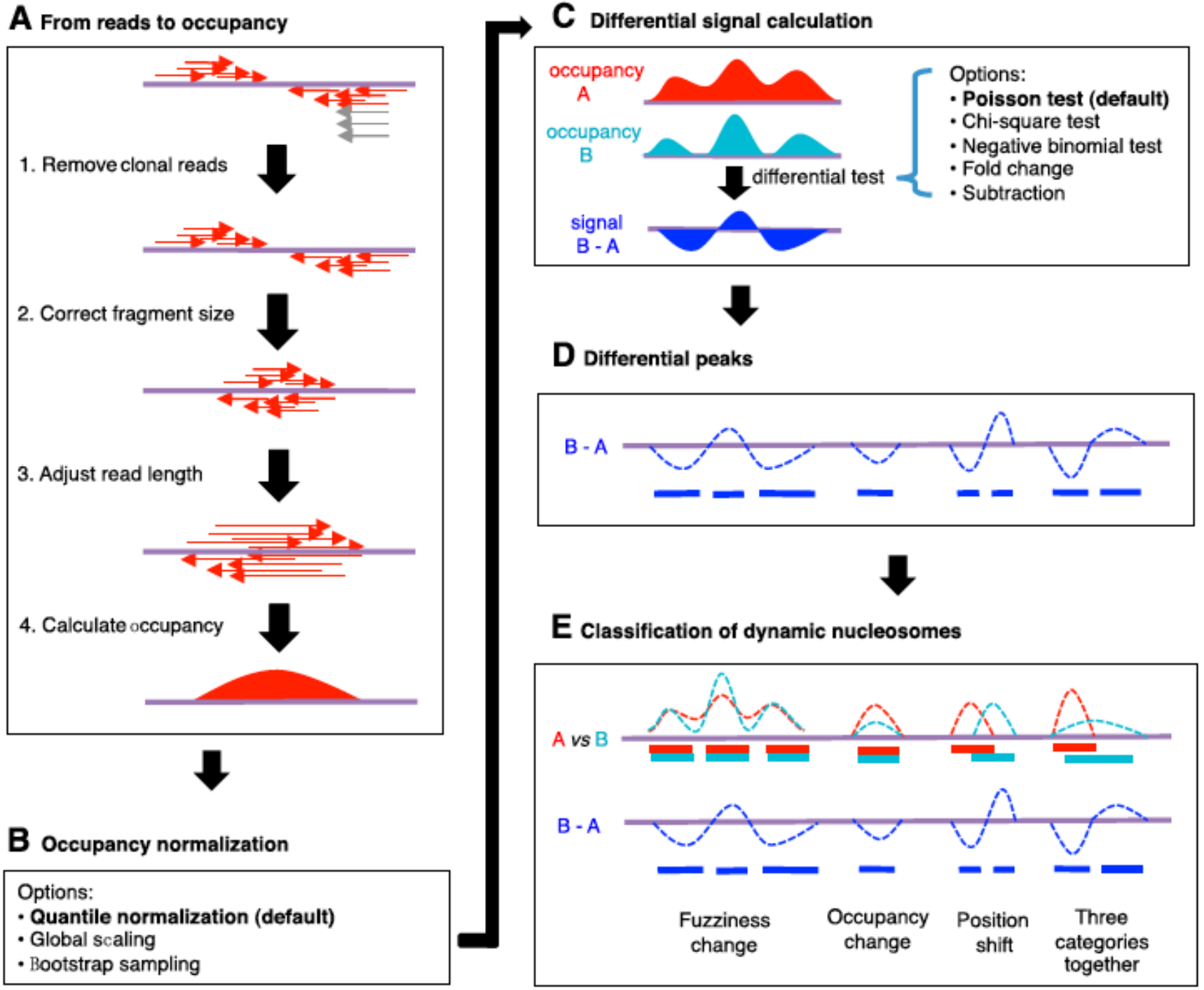


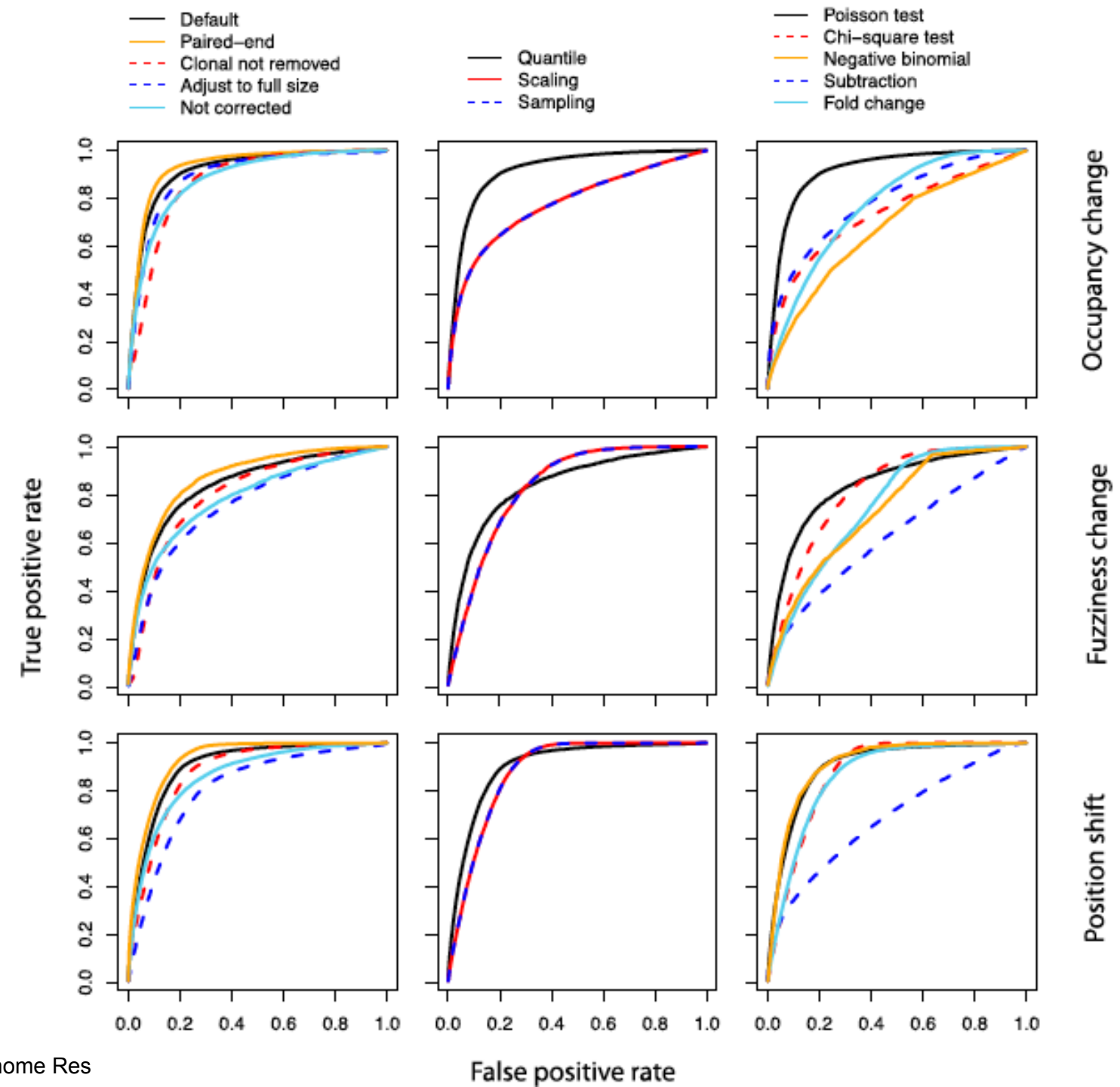
Fig. 1. a) nucleR workflow diagram. b) Top (purple): raw coverage from NGS paired-end reads; middle (orange): coverage using trimmed reads, remarking the dyad location; bottom (white): filtered trimmed coverage and scored nucleosome positions (in red). Detection threshold is marked as a red dashed line.

DANPOS

- comparative analysis of nucleosome physical organization

Chen et al., Genome Res


A From reads to occupancy
1. Remove clonal reads
2. Correct fragment size
3. Adjust read length
4. Calculate occupancy
B Occupancy normalization
Options:
• Quantile normalization (default)
• Global scaling
• Bootstrap sampling
C Differential signal calculation
occupancy A
occupancy B
differential test
signal B - A
Options:
• Poisson test (default)
• Chi-square test
• Negative binomial test
• Fold change
• Subtraction
D Differential peaks
B - A
E Classification of dynamic nucleosomes
A vs B
B - A
Fuzziness change
Occupancy change
Position shift
Three categories together



Chen et al., Genome Res

The project

- http://students.mimuw.edu.pl/~szczurek/TSG2_Project/project.html
- Report deadline: 20.01.2016
- Presentations: 26.01.2016
- Each presentation: 15 min

Bibliography

- Tsompana and Buck. Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin* 2014, **7**:33

- Meyer and Liu. *Identifying and mitigating bias in next-generation sequencing methods for chromatin biology* Nature Reviews Genetics. 2014. 15, 709–721 (2014) .

- Buenrostro et al. (2013) "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." Nature Methods

- Chen et al. DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing *Genome Res. 2013. 23: 341-351 doi:* 10.1101/gr.142067.112