

Integrated ATAC-seq Data Analysis Workshop: from Fastq to plots

Kai Hu, Haibo Liu, Lihua Julie Zhu

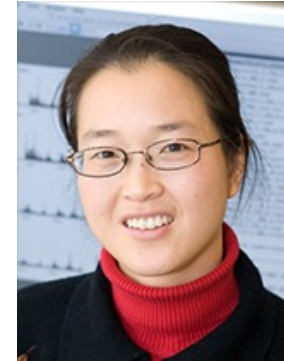
University of Massachusetts Medical School (UMMS)

Acknowledgments

Core developers of
ATACseqQC



Haibo Liu



Lihua Julie Zhu

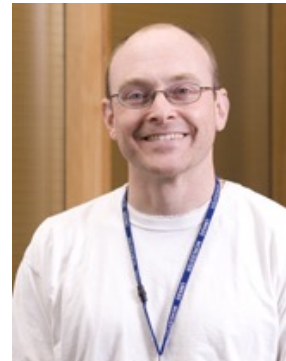


Duke
UNIVERSITY



Jianhong Ou

Other contributors



Nathan Lawson



Michelle Kelliher



Lucio Castilla

Agenda



Mini-lecture (15~20min)

- Introduction to ATAC-seq technology
- A common ATAC-seq workflow
- Best practices in ATAC-seq assays and data analysis



Demo (15~20min)

- ATACseqQC functions
- Demo plots



Q/A (5~10min)

- Bioconductor support site

How to run the workshop

Step1: install Docker engine: <https://docs.docker.com/get-docker/>

Step2: download the docker image:

- `docker pull hukai916/integrated_atacseq_analysis_workshop2021`

Step3: start the container:

- `docker run -e PASSWORD=yourpassword -p 8787:8787
hukai916/integrated_atacseq_analysis_workshop2021`

Step4: in your web browser:

- Enter <http://localhost:8787> using username **rstudio** and password **yourpassword**

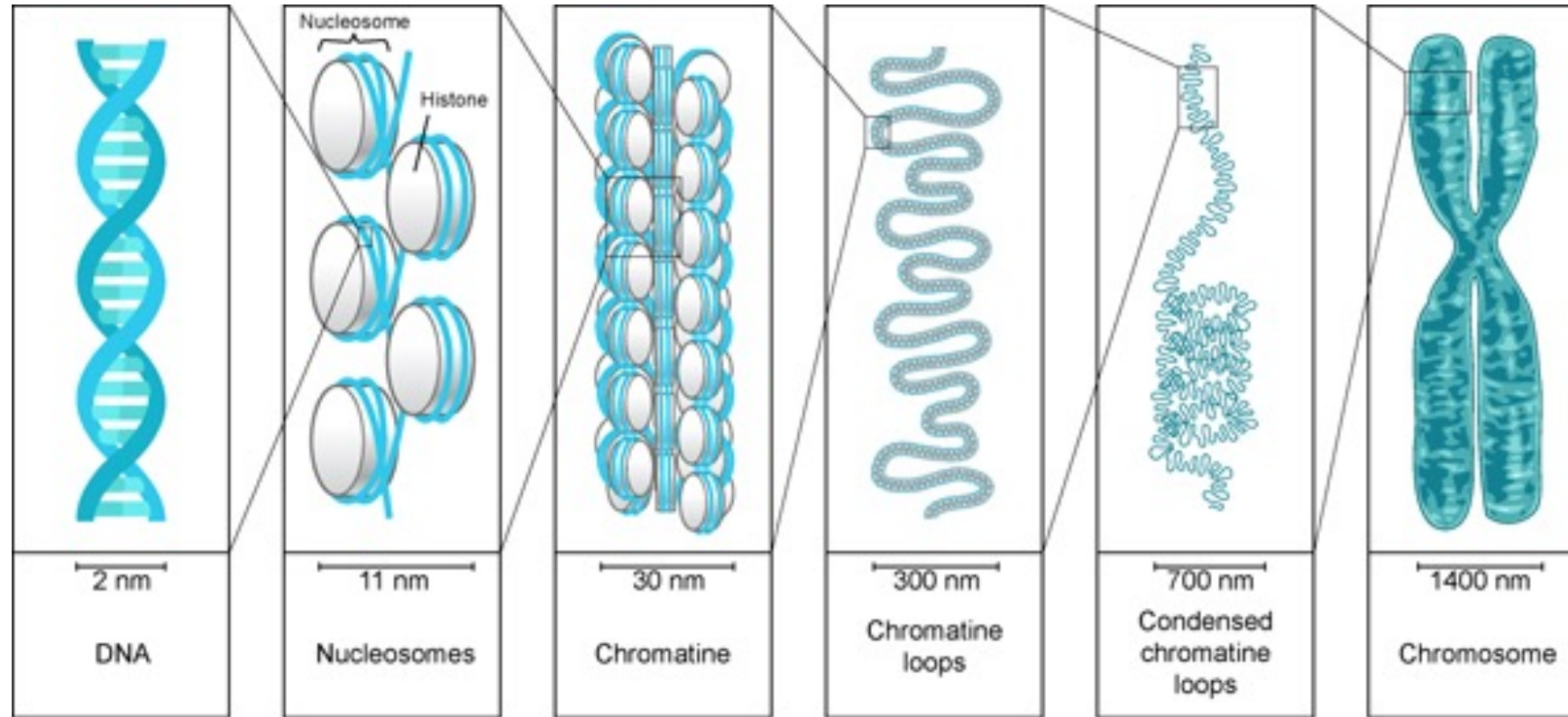
More instructions:

<https://github.com/hukai916/IntegratedATACseqAnalysisWorkshop2021>



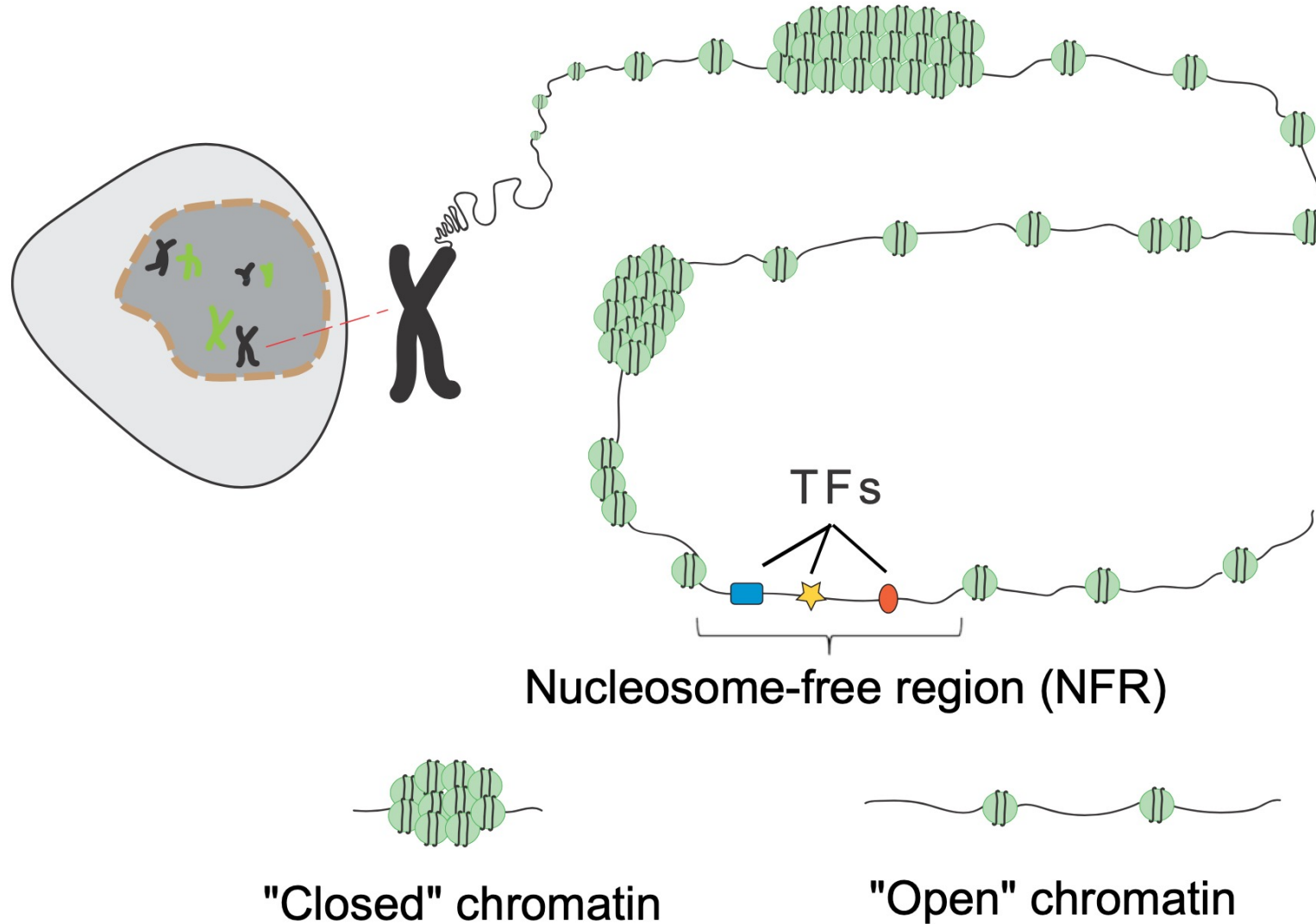
- Introduction to ATAC-seq technology
- A common ATAC-seq workflow
- Best practices in ATAC-seq assays and data analysis

DNA packaging in eukaryotes

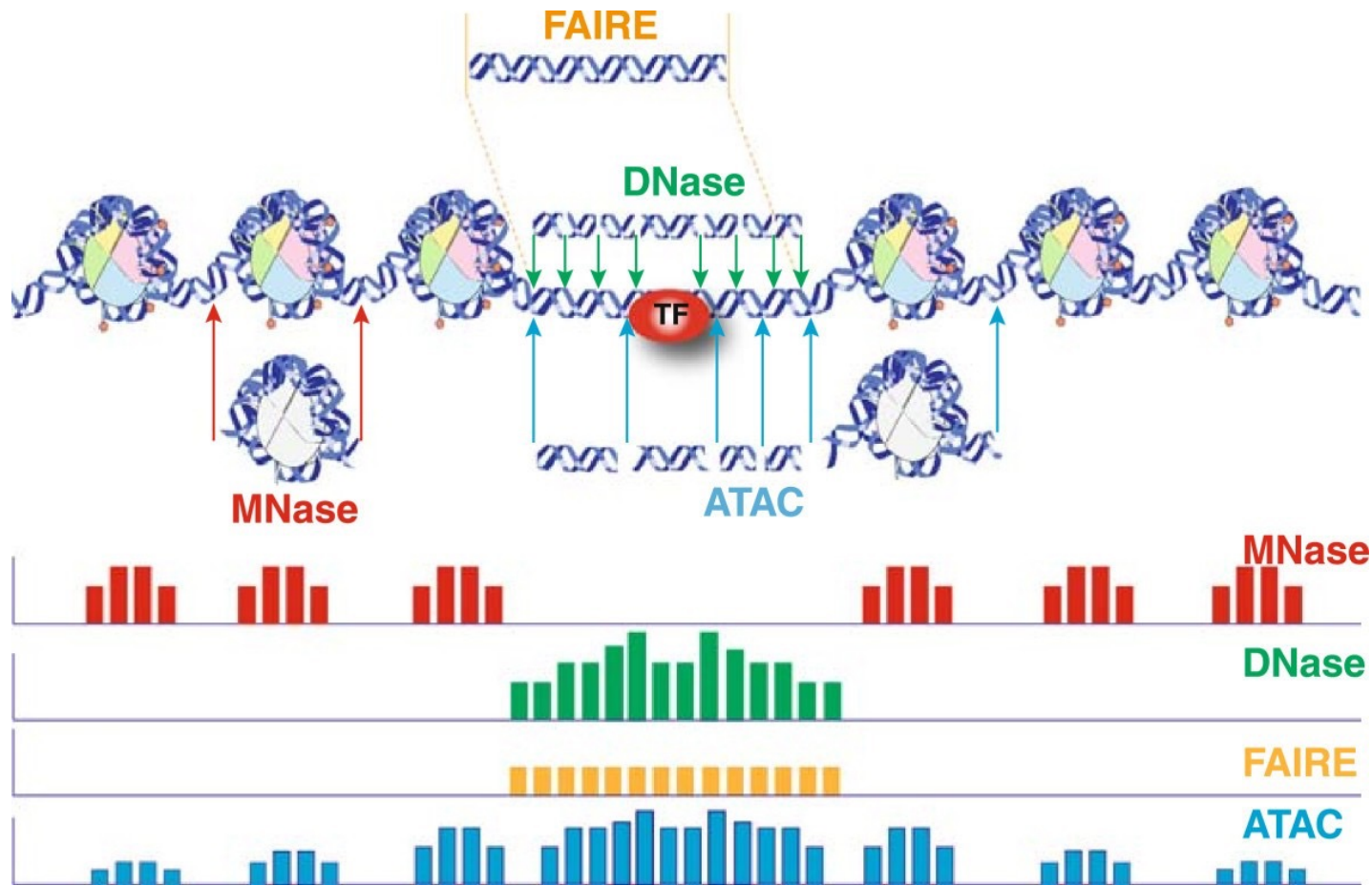


Condensed approximately 10,000 times.

“Open” vs “closed” chromatin



Methods for profiling chromatin accessibility landscape



MNase: endo-exonuclease

DNase: endonuclease, preferentially cut nucleosome free region.

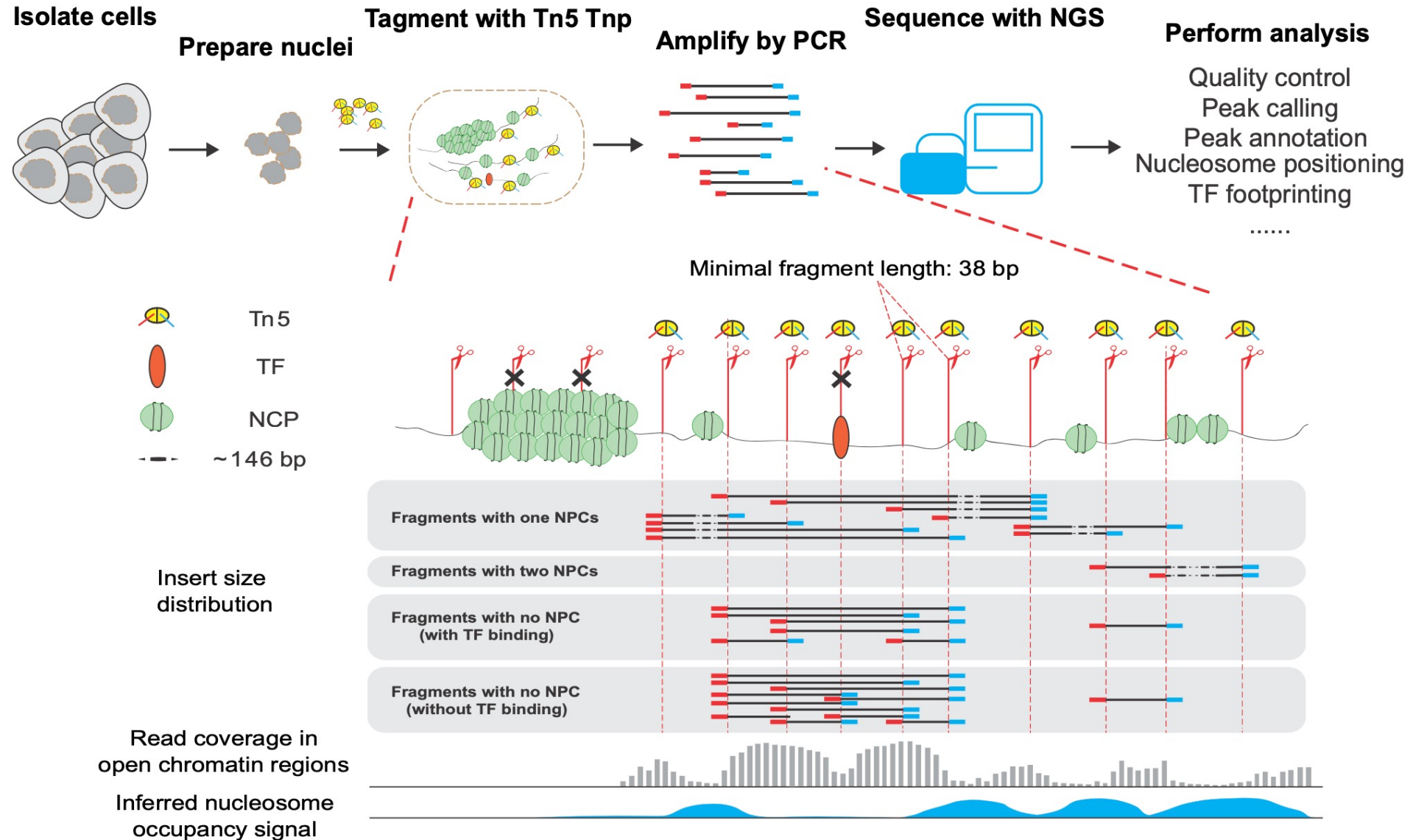
FAIRE: regulatory DNA fragments will be preferentially released to extraction solution after sonication.

ATAC: hypersensitive transposase that cuts accessible regions.



- Introduction to ATAC-seq technology
- **A common ATAC-seq workflow**
- Best practices in ATAC-seq assays and data analysis

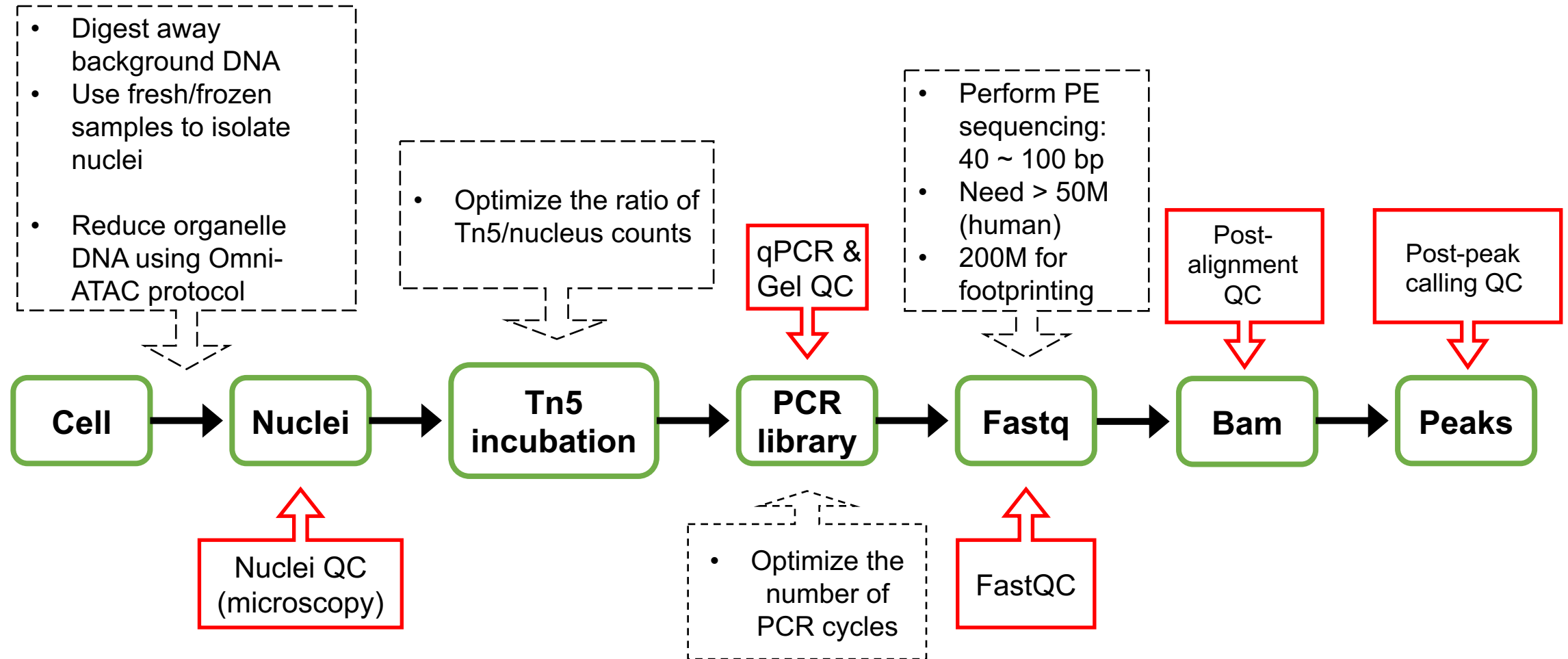
A typical ATAC-seq workflow





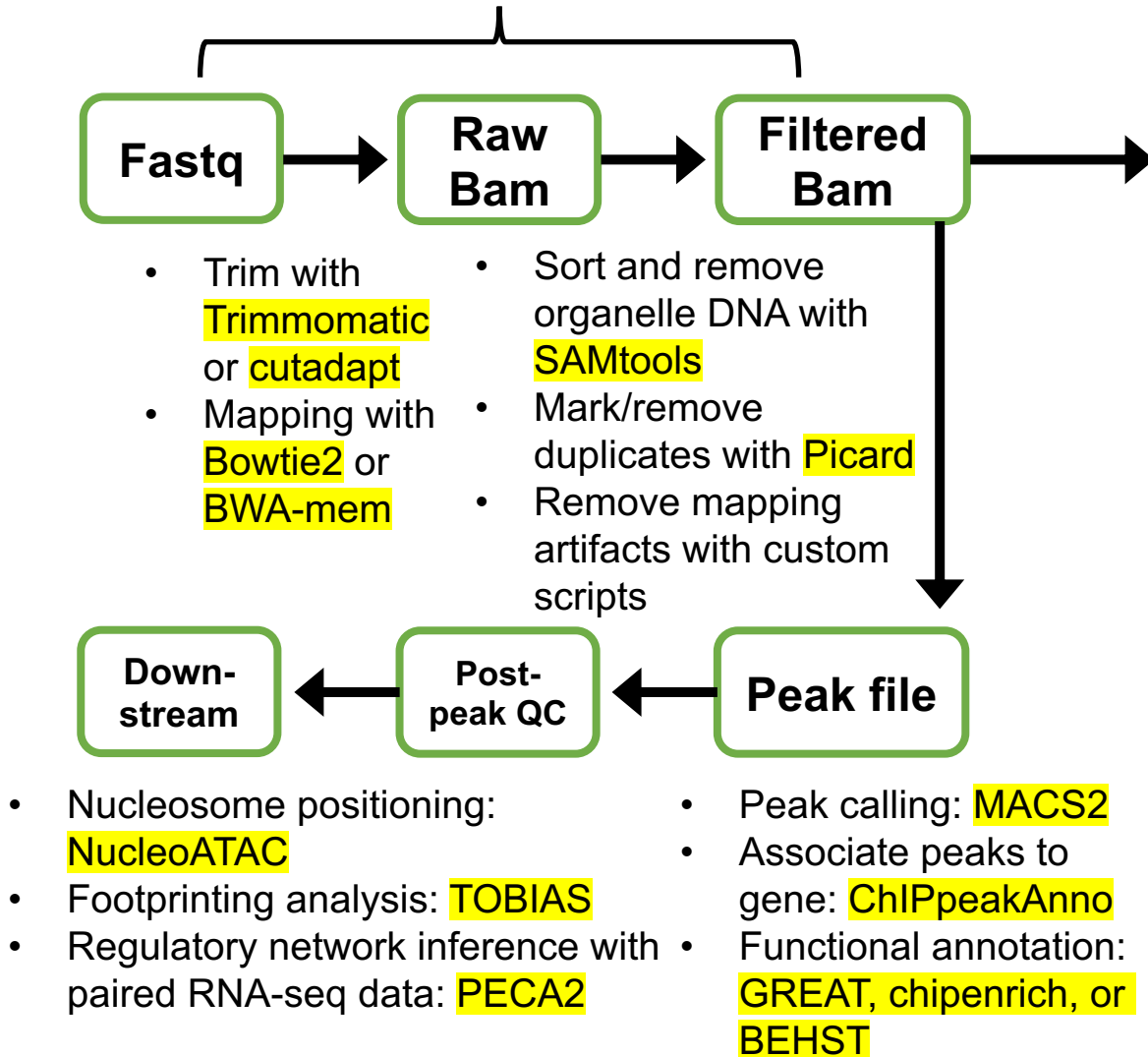
- Introduction to ATAC-seq technology
- A common ATAC-seq workflow
- **Best practices in ATAC-seq assays and data analysis**

Important steps, best practices, and important QCs



Popular software tools in each step

Preprocessing



ATACseqQC workflow

- Assessing mapping status: **bamQC()**
(Alternatives: samtools, picard)
- Assessing sequencing depth and library complexity: **saturationPlot()** & **estimateLibComplexity()**
- Assessing insert size distribution: **fragSizeDist()**
- Assessing similarities of replicates: **plotCorrelation()**
- Shifting aligned reads: **shiftGAlignmentsList()**
- Splitting BAM files: **splitGAlignmentsByCuts()**
- Plotting aggregate signals around TSSs: **featureAlignedHeatmap()**, **featureAlignedDistribution()**
- Streamlining IGV snapshots: **IGVSnapshot()**
- Assessing DNA-binding factor footprints: **factorFootprints()**



- ATACseqQC functions
- Demo plots

Demo I: assessing mapping status

`bamQC()`:

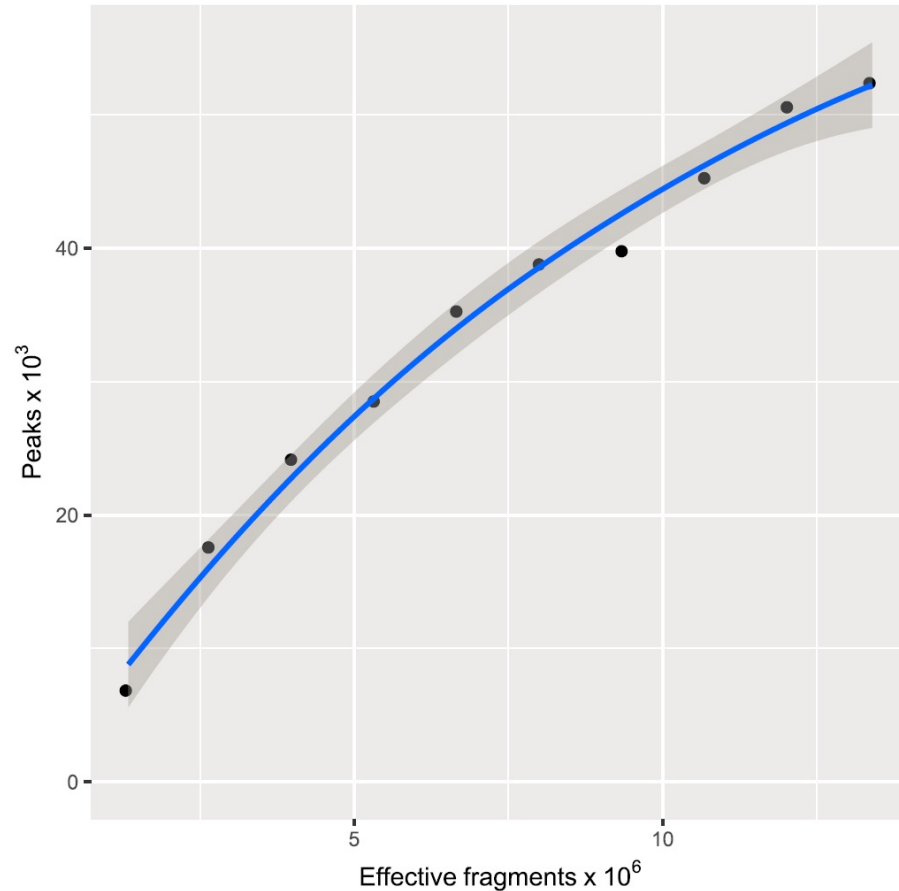
- Input: sorted BAM files with duplicates marked
- Output: mapping rate, duplicate rate, mapping quality, *etc.*

Alternative tools:

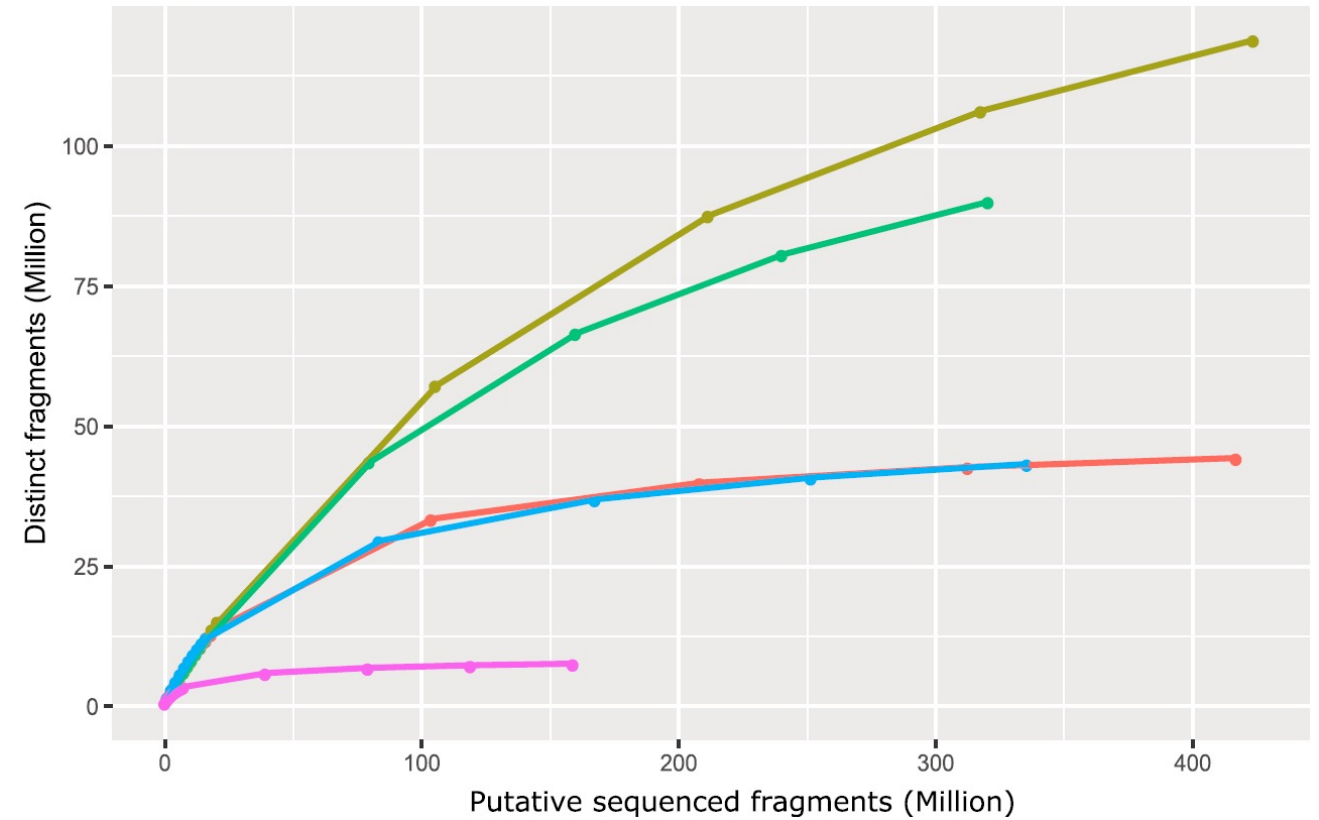
- SAMtools
- picard tools

Demo II: assessing sequencing depth and library complexity

saturationPlot()



estimateLibComplexity()

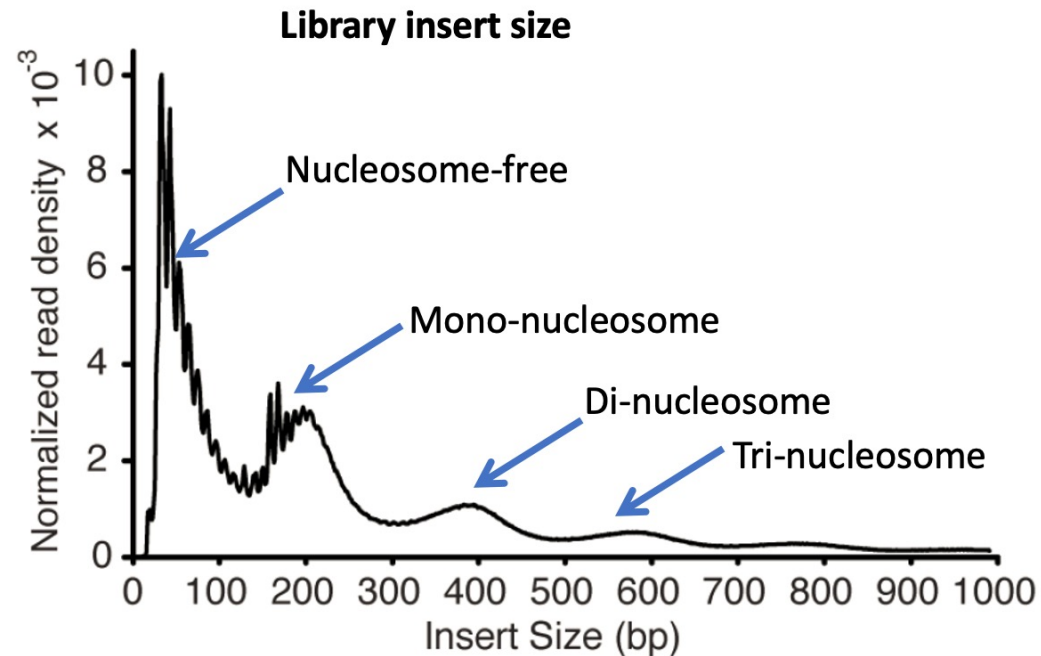


Ou, J., Liu, H., Yu, J. *et al.* ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics* **19**, 169 (2018).

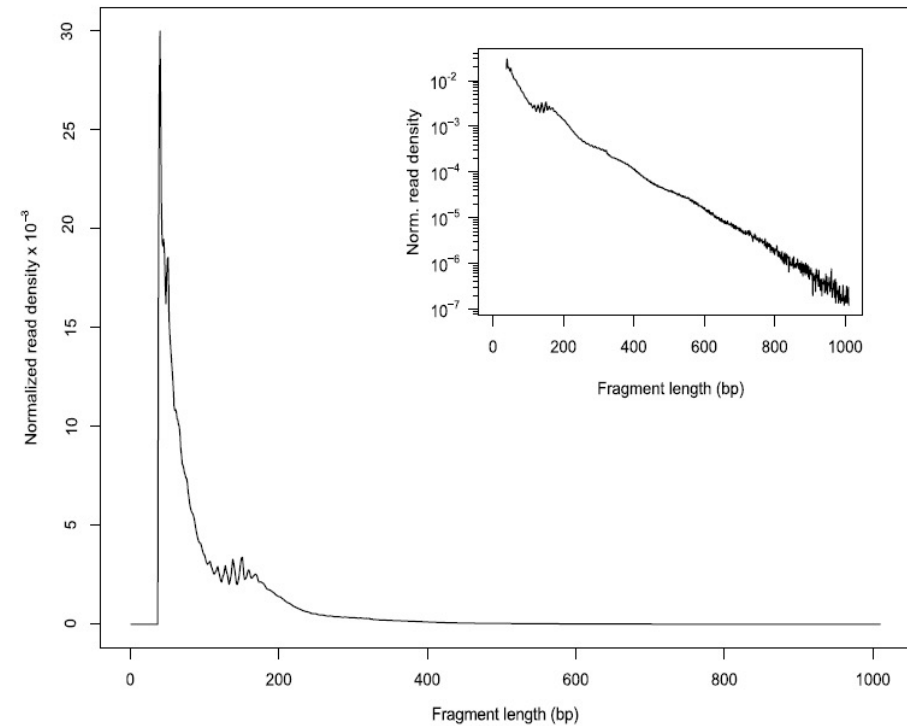
Demo III: assessing insert size distribution

fragSizeDist()

High quality

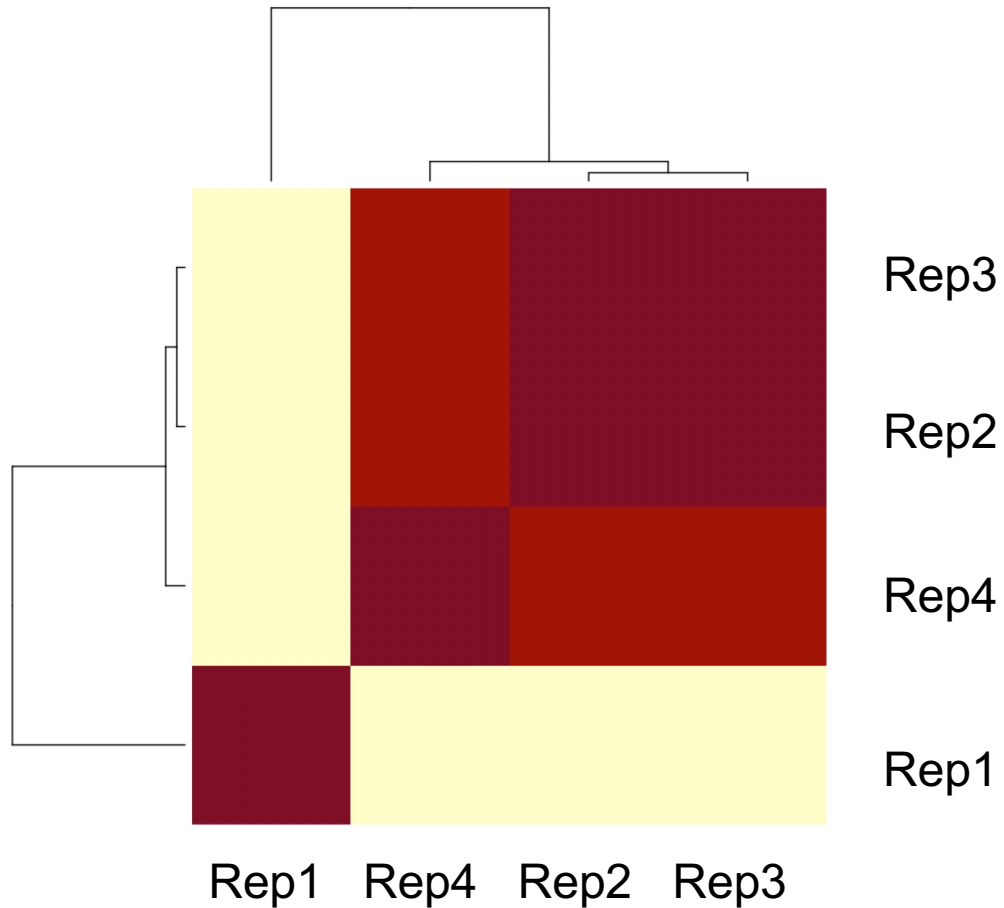


Poor quality



Ou, J., Liu, H., Yu, J. *et al.* ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics* **19**, 169 (2018).

Demo IV: assessing similarities of replicates:



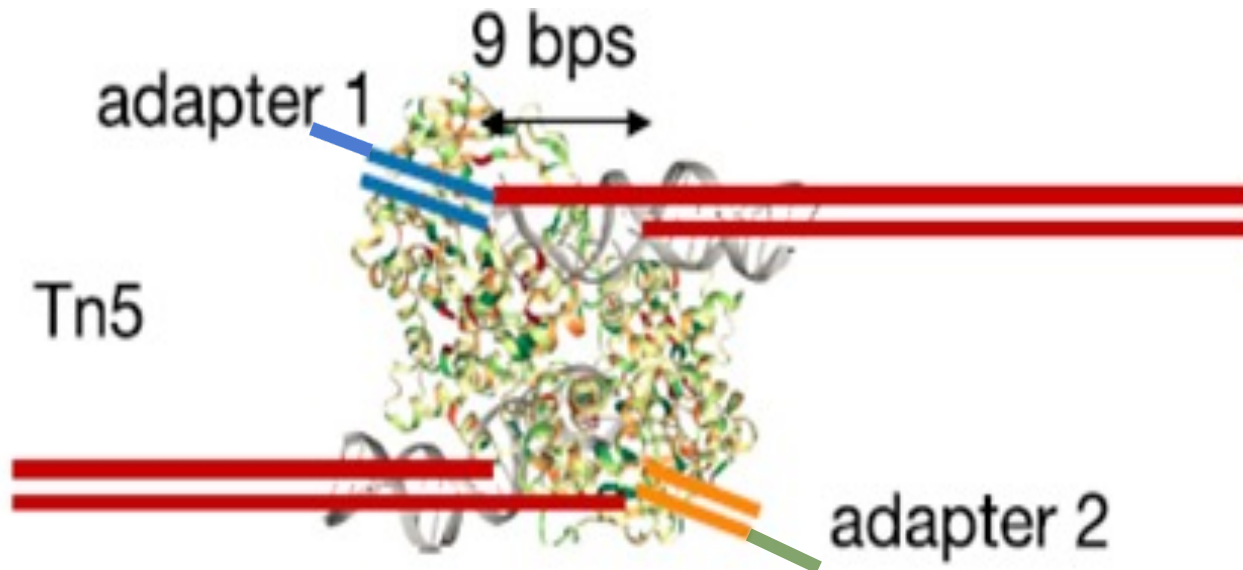
`plotCorrelation()`

- Can plot PCA or heatmap.
- The correlation is calculated by the counts in the promoter regions.

Rep1 is quite different from other replicates.

Demo V: shifting aligned reads

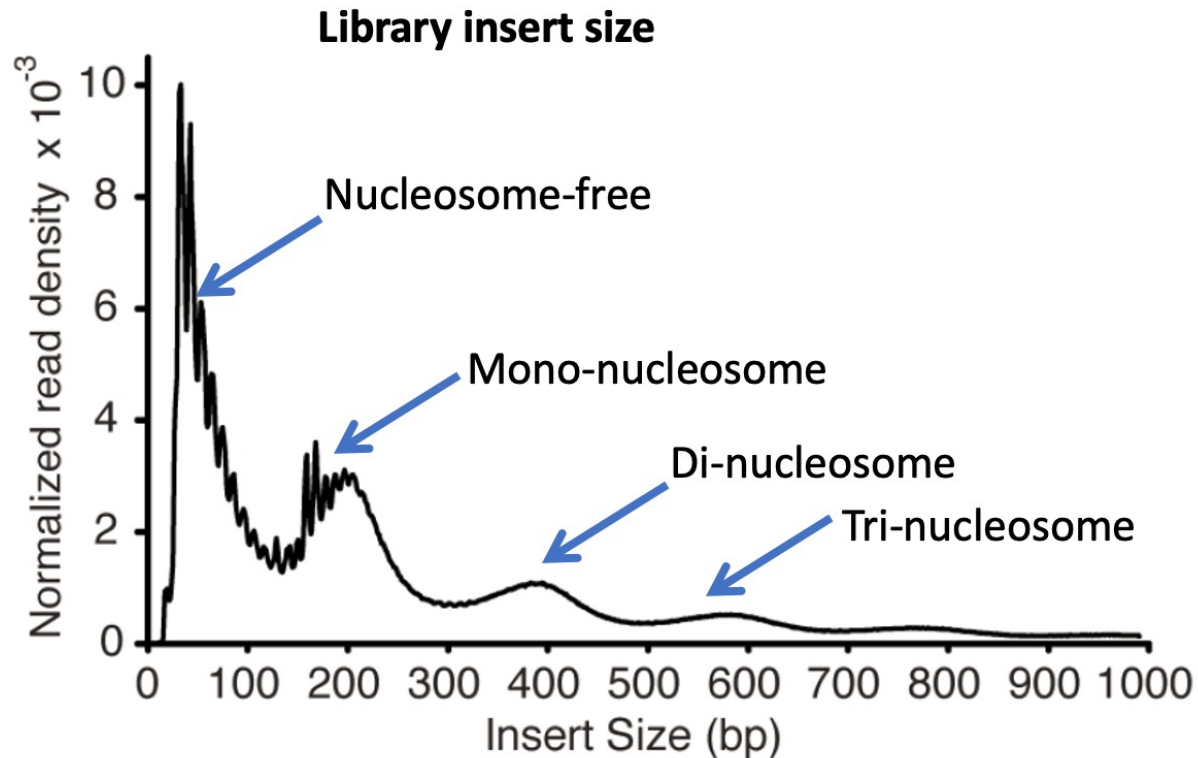
shiftGAlignmentsList()



- Tn5 tagmentation produces fragments with 9bp overhang at the 5' ends
- To center the cleavage events:
 - shift +4 bp for reads mapped to forward strand
 - shift -5 bp for reads mapped to reverse strand

Demo VI: splitting BAM files

`splitGAlignmentsByCut()`



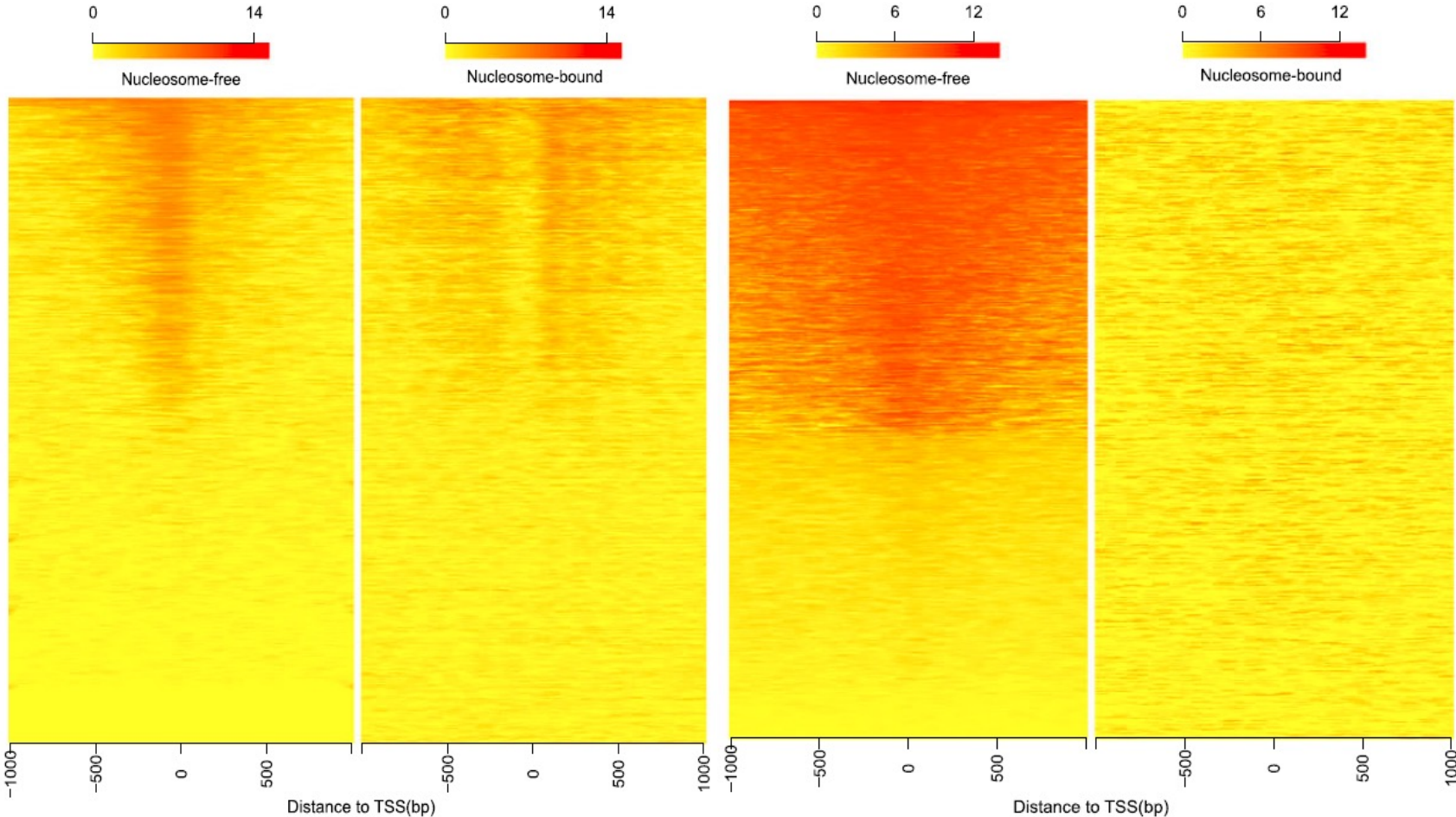
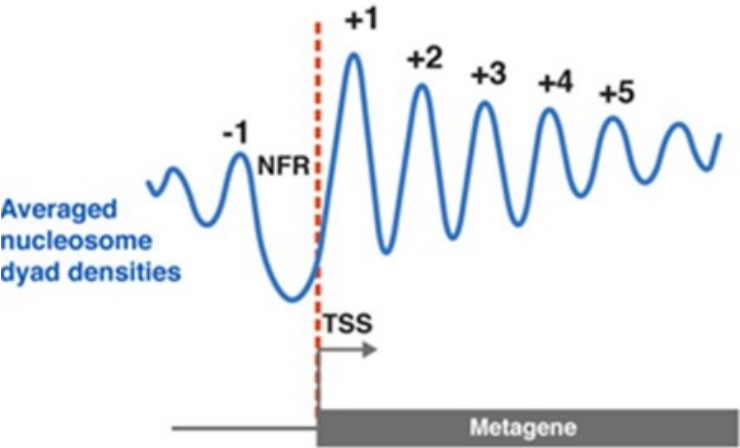
- You can split the Bam into different bins:
 - Nucleosome-free fragments
 - Inter1
 - Mono-nucleosome fragments
 - Inter2
 - Di-nucleosome fragments
 - *etc.*
- The reads from different bins can be used to visualize different signals:
 - promoter/enhancer/insulators are localized in NFR

Demo VII: plotting aggregate signals around TSSs

`featureAlignedHeatmap()`

High quality 

Poor quality 

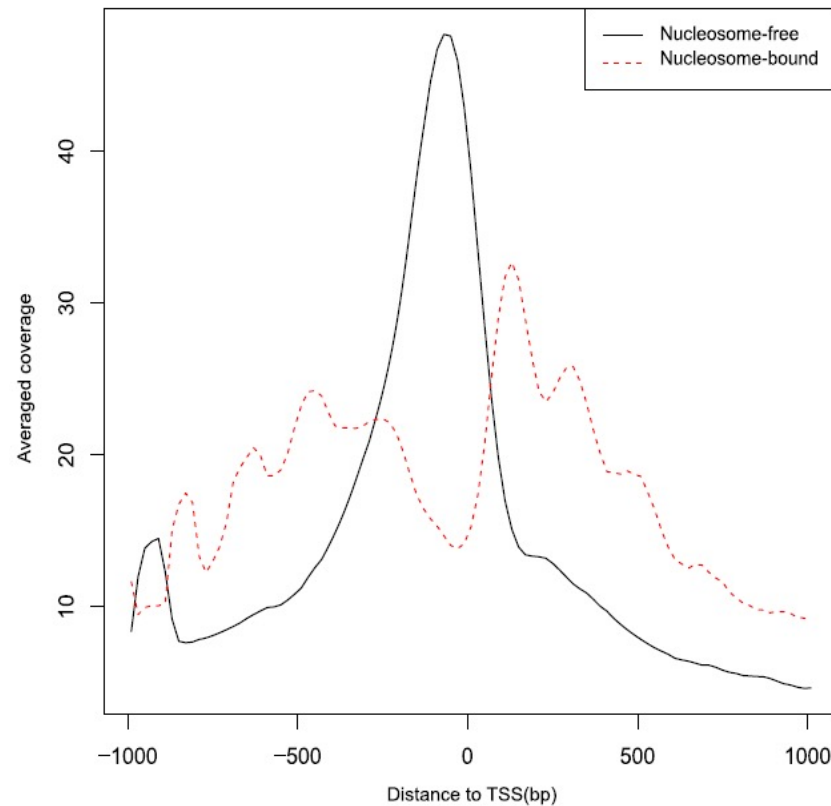


Nucleosome positioning around TSSs

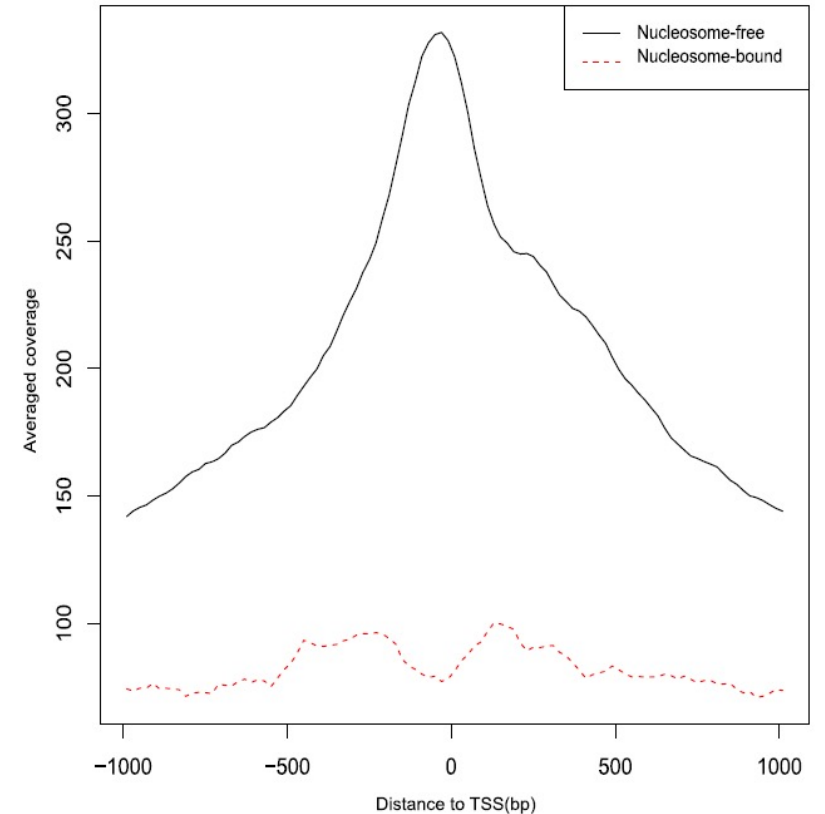
Demo VII: plotting aggregate signals around TSSs

featureAlignedDistribution()

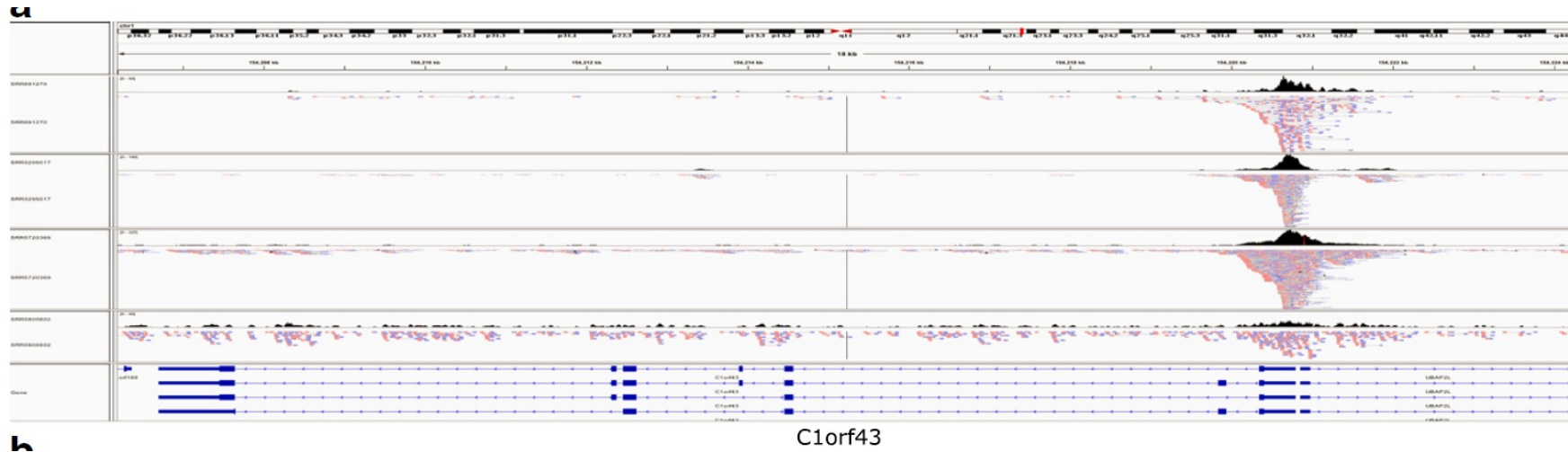
High quality ✓



Poor quality ✗



Demo VIII: streamlining IGV snapshots



IGVSnapshot()

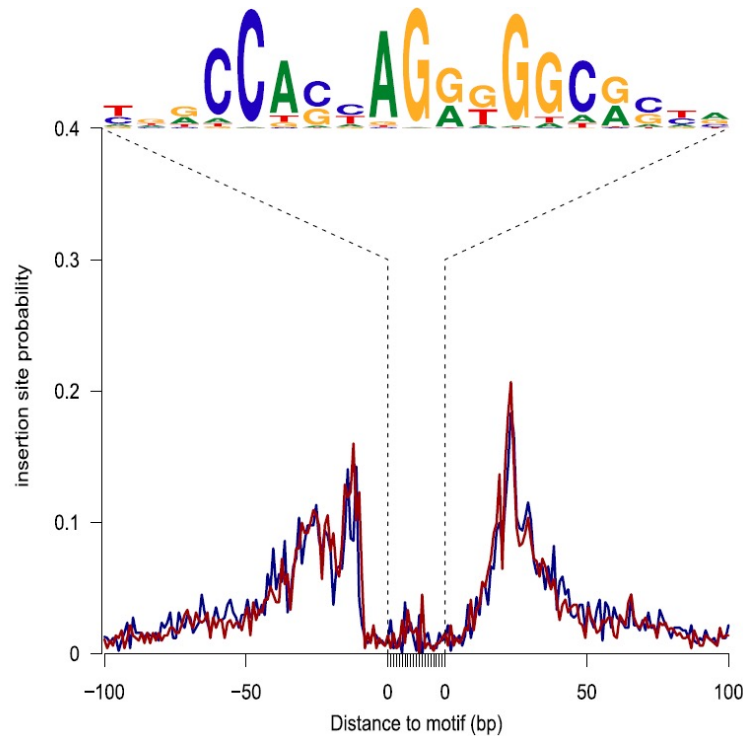


- Need to install IGV
- For high quality data:
 - read signals are enriched to open regions
- For poor quality data:
 - read signals distribute across the genome

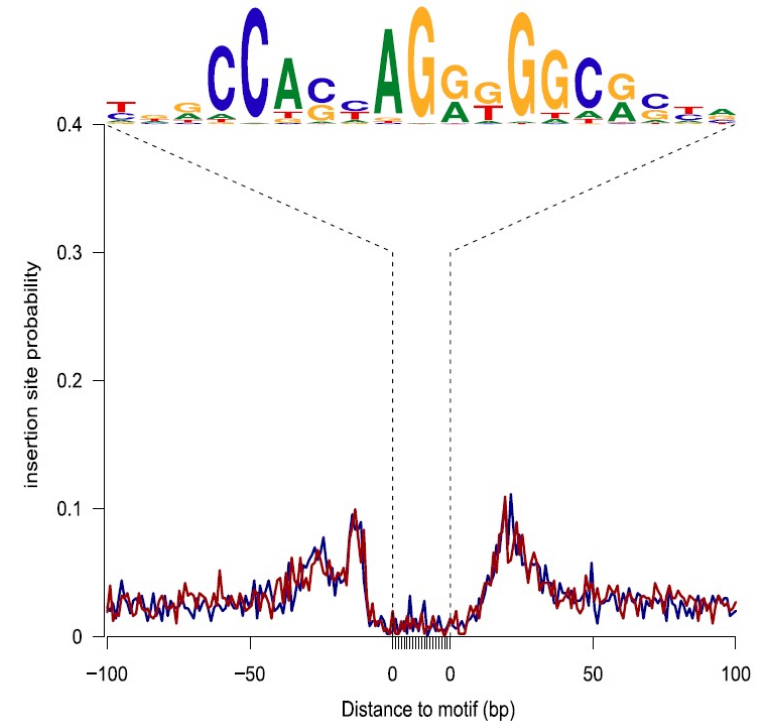
Ou, J., Liu, H., Yu, J. *et al.* ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics* **19**, 169 (2018).

Demo IX: assessing DNA-binding factor footprints

A (High quality)



B (Poor quality)



factorFootprints()

Ou, J., Liu, H., Yu, J. *et al.* ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics* **19**, 169 (2018).



- ATACseqQC functions
- Demo plots

Datasets for demo with ATACseqQC

Dataset1: good quality

SRR891269 and SRR891270 are ATAC-seq data for two biological replicates of 50K cells from EBV-transformed lymphoblastoid cell line GM12878 (Buenrostro *et al.* 2013).

Dataset2: bad quality

SRR5800801 and SRR5800802 are ATAC-seq data for two replicates of 75k cells from a breast cancer cell line T47 (Valles and Izquierd-Bouldstridge, unpublished).

Preprocessing already performed for you:

- FastqQC on raw fastq files
- Mapped to hg38
- Bam file filtered according to the steps aforementioned
- Download link to the preprocessed bam files and preprocessing scripts: [Google Drive](#)
- A small subset of the bam files are included in the workshop package for quick demo

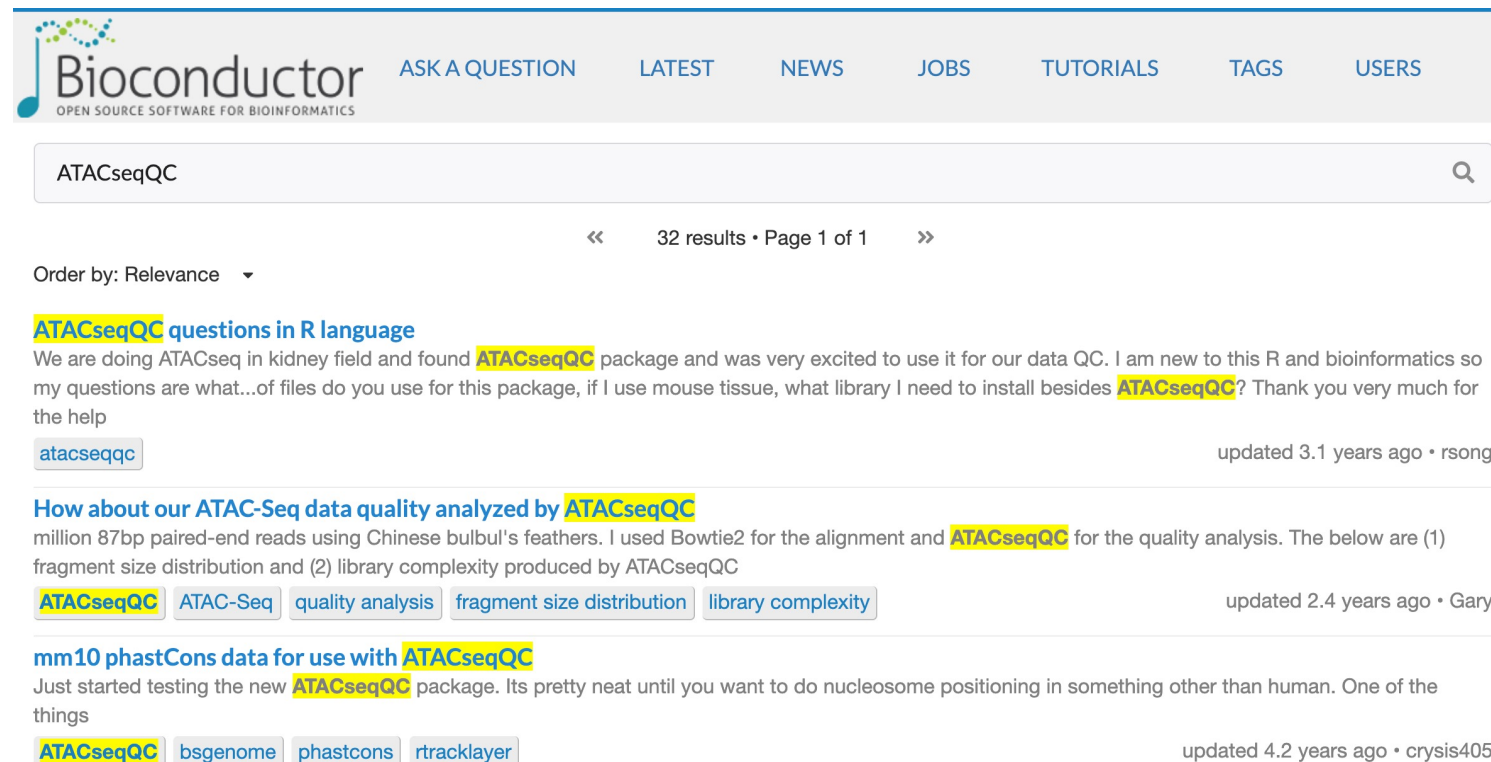


Demo with Rstudio

Q & A

View and submit your questions to **Bioconductor support site**:
<https://support.bioconductor.org/post/search/?query=ATACseqQC>

Developers actively monitor questions posted there.



The screenshot shows the Bioconductor support site interface. At the top is the Bioconductor logo and navigation links: ASK A QUESTION, LATEST, NEWS, JOBS, TUTORIALS, TAGS, and USERS. Below the navigation bar is a search bar containing the text 'ATACseqQC'. Under the search bar, it indicates '32 results • Page 1 of 1'. The results are ordered by Relevance. The first result is titled 'ATACseqQC questions in R language' and describes a user's experience with the ATACseqQC package. The second result is titled 'How about our ATAC-Seq data quality analyzed by ATACseqQC' and discusses data quality analysis. The third result is titled 'mm10 phastCons data for use with ATACseqQC' and mentions testing the ATACseqQC package. Each result includes a brief description, relevant tags, and the user's name and update time.

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

ASK A QUESTION LATEST NEWS JOBS TUTORIALS TAGS USERS

ATACseqQC

32 results • Page 1 of 1

Order by: Relevance

ATACseqQC questions in R language
We are doing ATACseq in kidney field and found **ATACseqQC** package and was very excited to use it for our data QC. I am new to this R and bioinformatics so my questions are what...of files do you use for this package, if I use mouse tissue, what library I need to install besides **ATACseqQC**? Thank you very much for the help
atacseqqc updated 3.1 years ago • rsong

How about our ATAC-Seq data quality analyzed by ATACseqQC
million 87bp paired-end reads using Chinese bulbul's feathers. I used Bowtie2 for the alignment and **ATACseqQC** for the quality analysis. The below are (1) fragment size distribution and (2) library complexity produced by ATACseqQC
ATACseqQC ATAC-Seq quality analysis fragment size distribution library complexity updated 2.4 years ago • Gary

mm10 phastCons data for use with ATACseqQC
Just started testing the new **ATACseqQC** package. Its pretty neat until you want to do nucleosome positioning in something other than human. One of the things
ATACseqQC bsgenome phastcons tracklayer updated 4.2 years ago • crisis405

References

1. ATAC-seq technology: Buenrostro, J., Giresi, P., Zaba, L. *et al.* Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–1218 (2013).
<https://doi.org/10.1038/nmeth.2688>
2. ATACseqQC package: Ou, J., Liu, H., Yu, J. *et al.* ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics* **19**, 169 (2018).
<https://doi.org/10.1186/s12864-018-4559-3>
3. ChIPpeakAnno package: Zhu, L.J., Gazin, C., Lawson, N.D. *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**, 237 (2010). <https://doi.org/10.1186/1471-2105-11-237>
4. ATACseqQC workshop: <https://github.com/haibol2016/ATACseqQCWorkshop>

Read more:

1. ChIPpeakAnno workshop: <https://github.com/hukai916/IntegratedChIPseqWorkshop>