ATAC-seq analysis

Bingbing Yuan

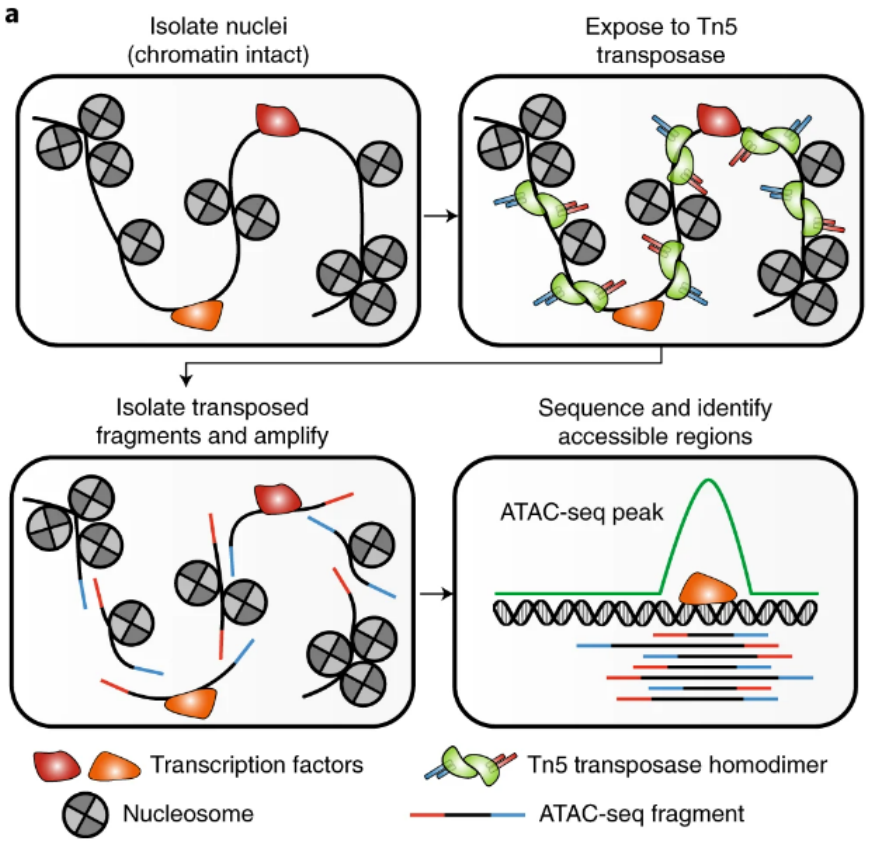
BaRC Hot Topics –April 4th 2024
Bioinformatics and Research Computing
Whitehead Institute

http://barc.wi.mit.edu/hot_topics/

Whitehead Institute

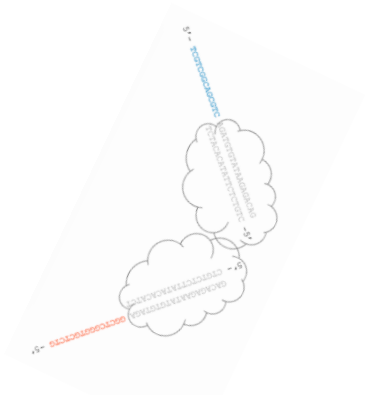
BaRC Hot Topics

Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq)

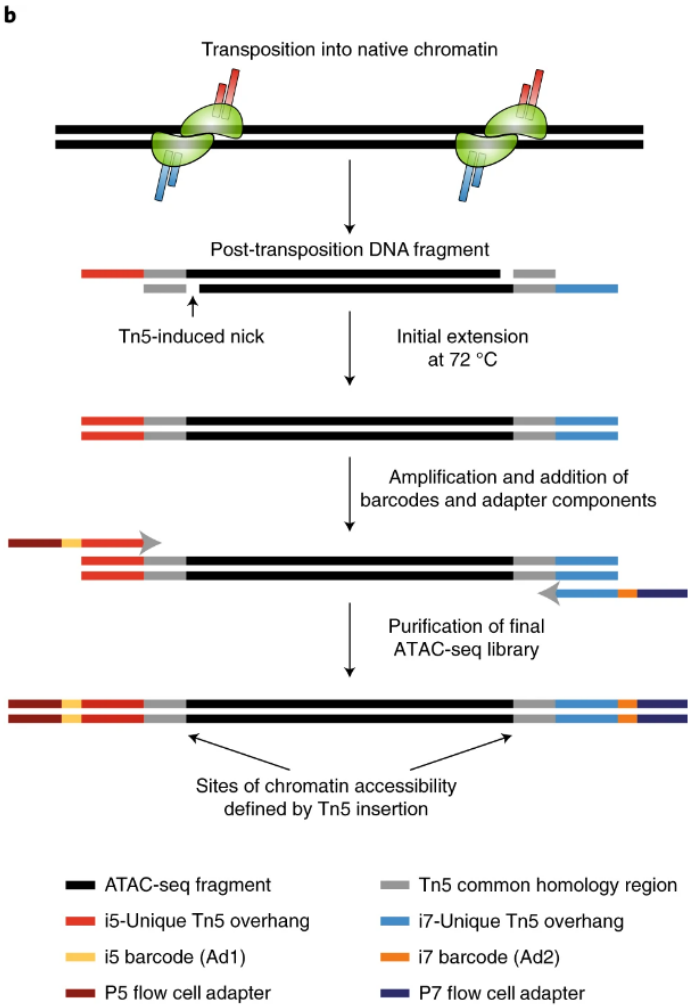


Nat Protoc 17, 1518–1552 (2022)

ATAC-seq



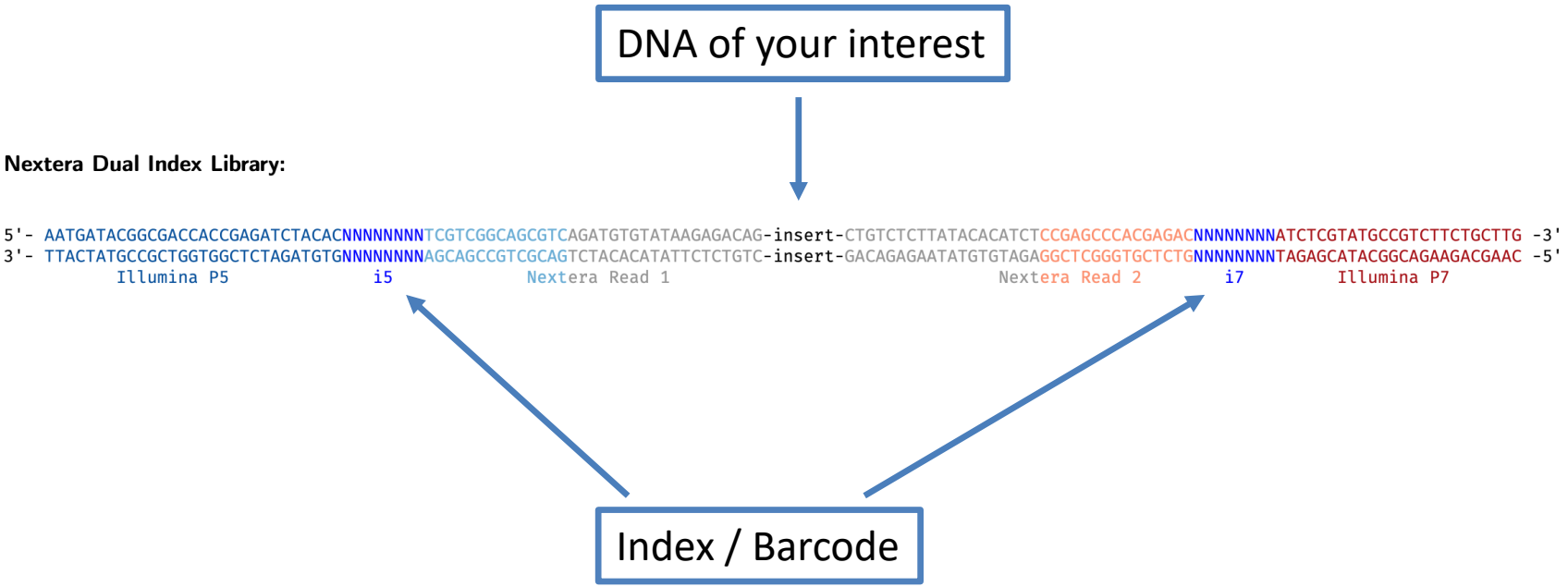
Nextera



Nextera

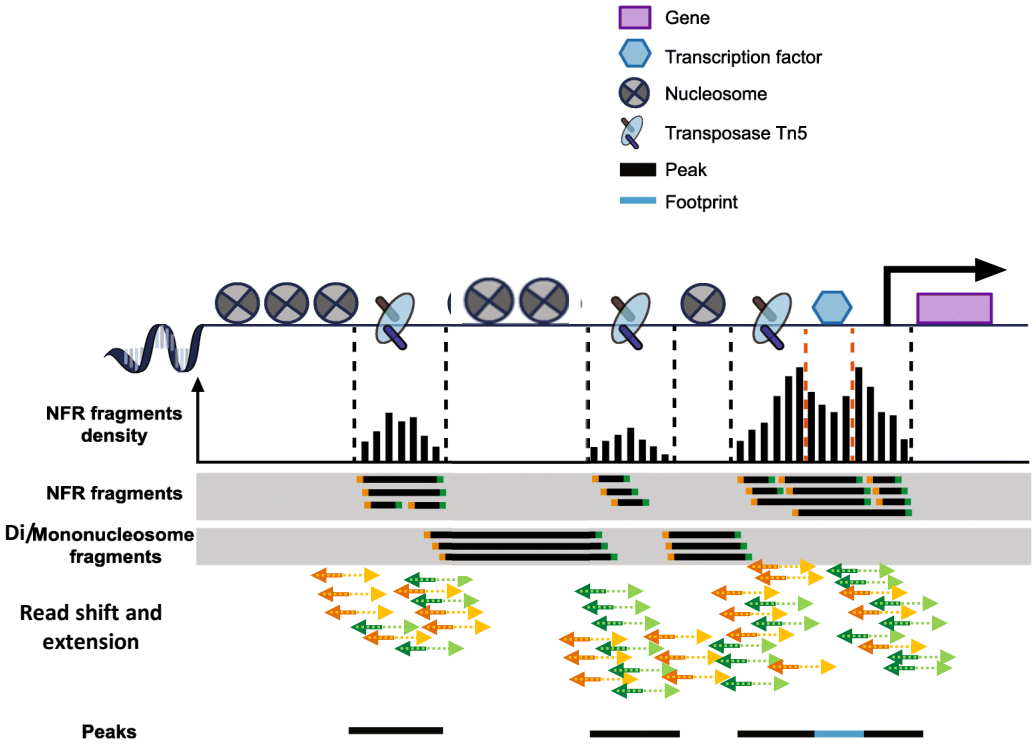
Nat Protoc 17, 1518–1552 (2022)
https://www.plob.org/article/11443.html

Nextera libray



Modified based on https://teichlab.github.io/scg_lib_structs/methods_html/Illumina.html

Profile generation in ATAC-seq



Modified from Feng Yan et.al Genome Biology 21 (2020)

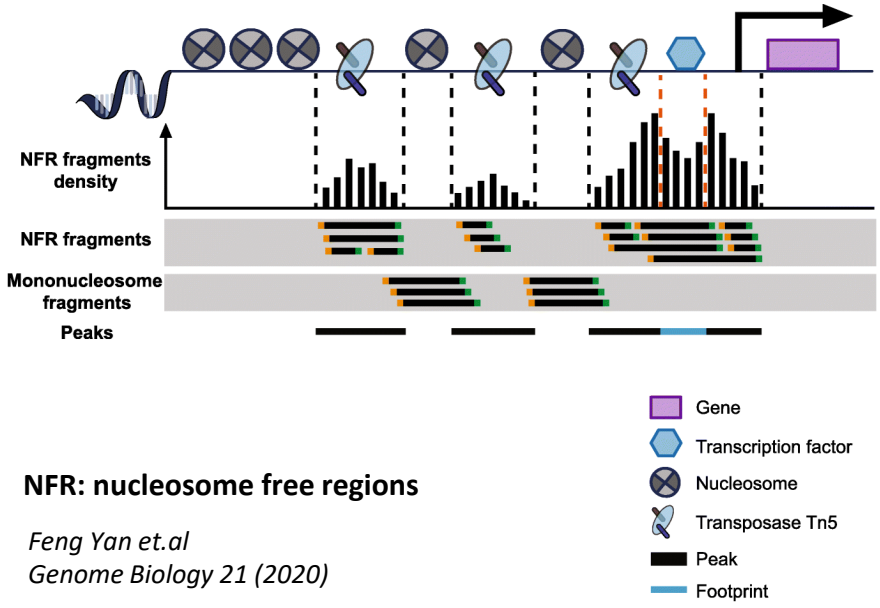
- Open chromatin
- Motif enrichment
- TF footprinting
- Nucleosome positions

ATAC-seq pipeline goal

- Using short-read sequencing, identify genome-wide regions of open chromatin

BaRC

ATAC-seq analysis workflow



NFR: nucleosome free regions

Feng Yan et.al
Genome Biology 21 (2020)

Steps in ATAC-seq data analysis

Goal: Find the open chromatin regions
Peak = open chromatin region

1. Quality control
   Remove adapters if necessary
2. Mapping
   Tailored to paired end reads
3. Post-alignment filtering

4. Peak calling and differential analysis
   i) Read shift and extension and signal profile generation.
   ii) Peak assignment

5. Peak interpretation
   i) Find genes next the open chromatin regions
   ii) Find motifs within peaks

ENCODE ATAC-seq pipeline

- Human and mouse with biological replicates
- QC report (a sample report)
- Includes all the steps in a single run

Instructions to run pipeline using Whitehead server

1. Pre-alignment quality control
2. Aligning reads to genome
3. Post-alignment filtering
4. Post-alignment quality control
5. Peak (accessible regions) calls
6. Assessing Peak Calls with FRiP score (same as ChIP-seq)
7. Blacklist filtering for peaks (same as ChIP-seq)

BaRC

What sequencing works best for ATAC-seq?

- Read depth recommended by ENCODE:
  - 50 million for paired-ended sequencing (25 million fragments)
- No input control sample
- Shorter read lengths (50x50 or 75x75) better than longer reads (100x100 or longer)
  - Ensure NFR
- Pair-end (PE) reads are recommended over single reads
  - Both ends of DNA fragments are considered
  - More detailed information about nucleosome packing and positioning in PE
  - Avoid losing reads due to mapping to repetitive elements

BaRC

Illumina data format

- Fastq format:

  http://en.wikipedia.org/wiki/FASTQ_format

```
@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhhhghhhhhhhehhhedhhhhfhhhhhh
```

/1 or /2 paired-end

→ @seq identifier

→ seq

→ +any description

→ seq quality values

Input qualities	Illumina versions
--solexa-quals	<= 1.2
--phred64	1.3-1.7
--phred33	>= 1.8

BaRC

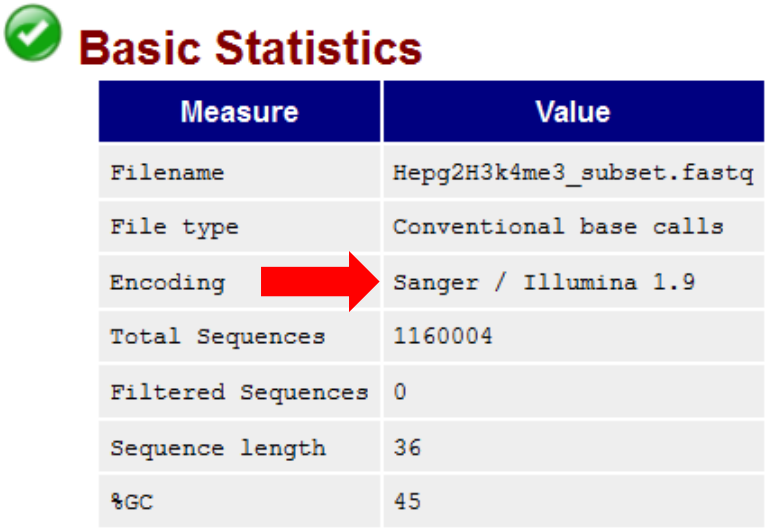
Check read quality with FastQC

(http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

1. Run FastQC to check read quality

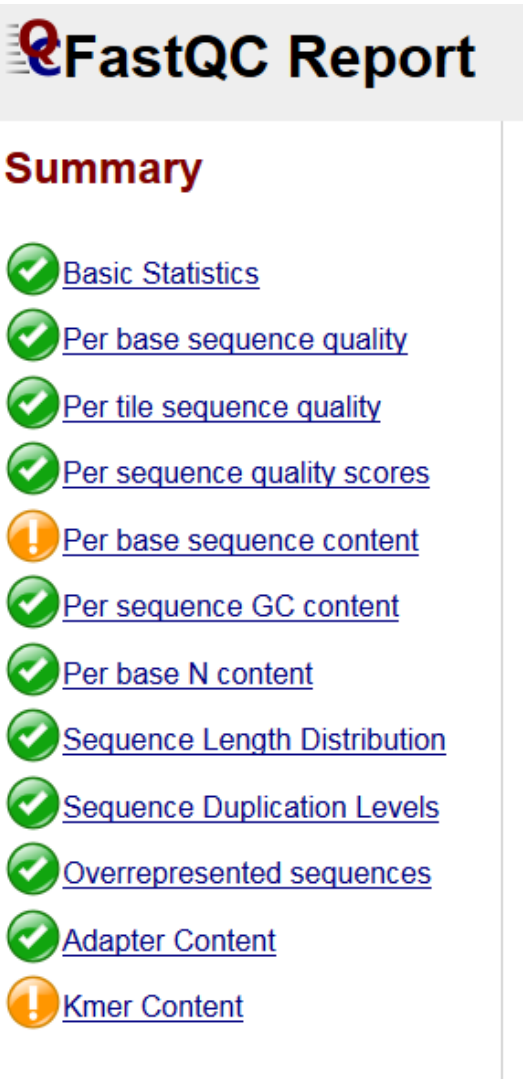
   fastqc sample.fastq

2. Open output file: "fastqc_report.html"



Basic Statistics

Measure	Value
Filename	Hepg2H3k4me3_subset.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1160004
Filtered Sequences	0
Sequence length	36
%GC	45



FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Pre-alignment quality control

1. Check reads quality with fastqc

   fastqc read1.fq read2.fq

2. Remove adapter when necessary:

   trim_galore --fastqc -nextera --paired --length 30 read1.fq read2.fq

   - -nextera: ATAC-seq experiments use the Nextera DNA Library Prep Kit
   - --paired: both reads need to pass or they are both removed.
   - --length: discard trimmed reads shorter than this length.

IO BaRC

Local genomic files needed for mapping

tak: /nfs/genomes/

- Human, mouse, zebrafish, *C.elegans*, fly, yeast, etc.
- Different genome builds
  - mm10: mouse_mm10_dec_11
  - mm39: mouse_mm39_jun20
- human_hg38_dec13 vs human_hg38_dec13_no_random
  - human_hg38_dec13 includes *_random.fa, *hap*.fa, etc.
- Sub directories:
  - bowtie
    - Bowtie1: *.ebwt
    - Bowtie2: *.bt2
  - fasta: one file per chromosome
  - fasta_whole_genome: all sequences in one file
  - gtf: gene models from Refseq, Ensembl, etc.

BaRC

Map reads to reference genome

Map reads with a non-spliced mapping tools: bowtie2 or BWA

bowtie2 --very-sensitive --no-discordant -p 2 -X 2000 -x hg38 -1 read1.fq -2 read2.fq | samtools view -ub - | samtools sort - >| bowtie_out.bam

- --no-discordant:
  - Suppress discordant alignments for paired reads
- -X 2000:
  - Increase maximum fragment length to 2k to include nucleosome distribution
  - Used for plotting fragment size distribution

Post alignment filtering

- Remove reads with low quality score: MAPQ < 30

  alignmentSieve -b file.bam --minMappingQuality 30 --samFlagInclude 2 -o MAPQ30.bam"

- Remove duplicates with Picard's 'MarkDuplicates’

  java -jar picard.jar MarkDuplicates I=foo.bam O=noDups.bam M=foo.marked_dup_metrics.txt REMOVE_DUPLICATES=true

- Remove reads mapped to mitochondria

  samtools view -h file.bam | grep -v chrM | samtools view -b -h -f 0x2 - | samtools sort - > file.sorted.bam

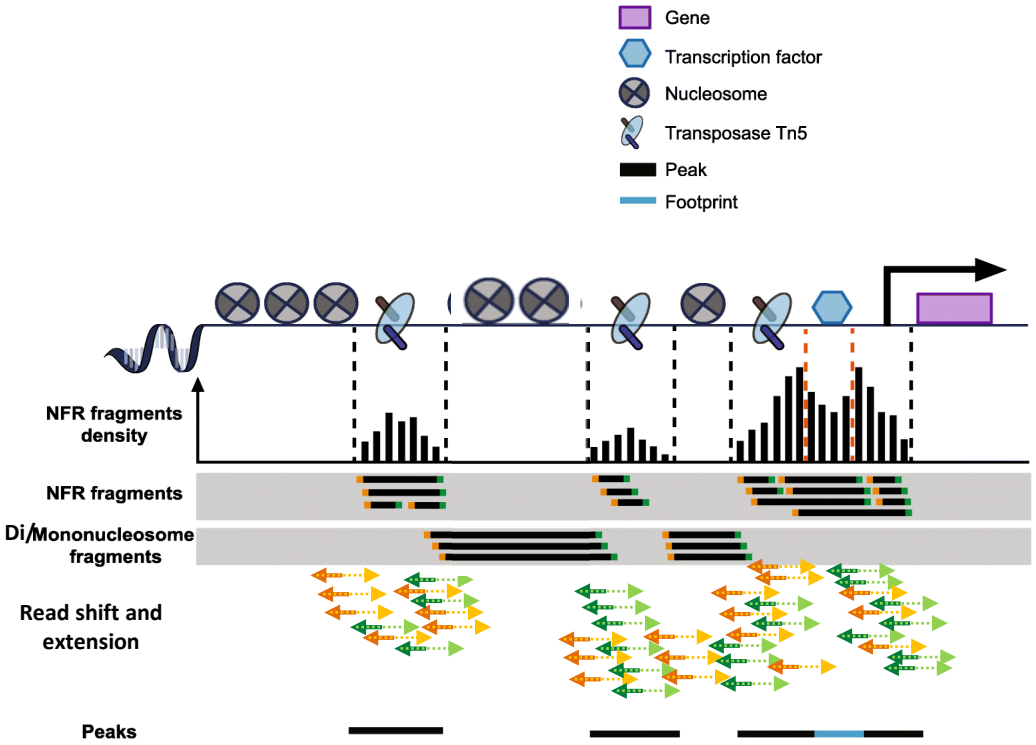
-f 0x2
samFlagInclude 2 } keep only properly paired reads

Post-alignment quality control

- Fragment size distribution
- TSS enrichment score

IO BaRC

Profile generation in ATAC-seq



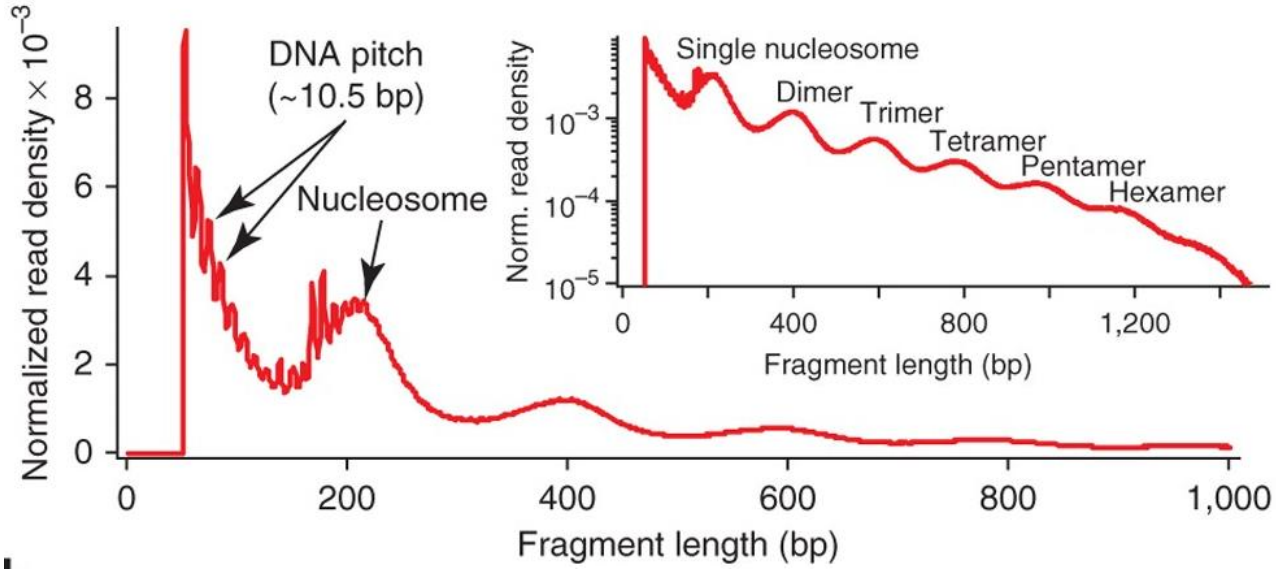
Modified from Feng Yan et.al Genome Biology 21 (2020)

- Open chromatin
- Motif enrichment
- TF footprinting
- Nucleosome positions

Fragment size distribution

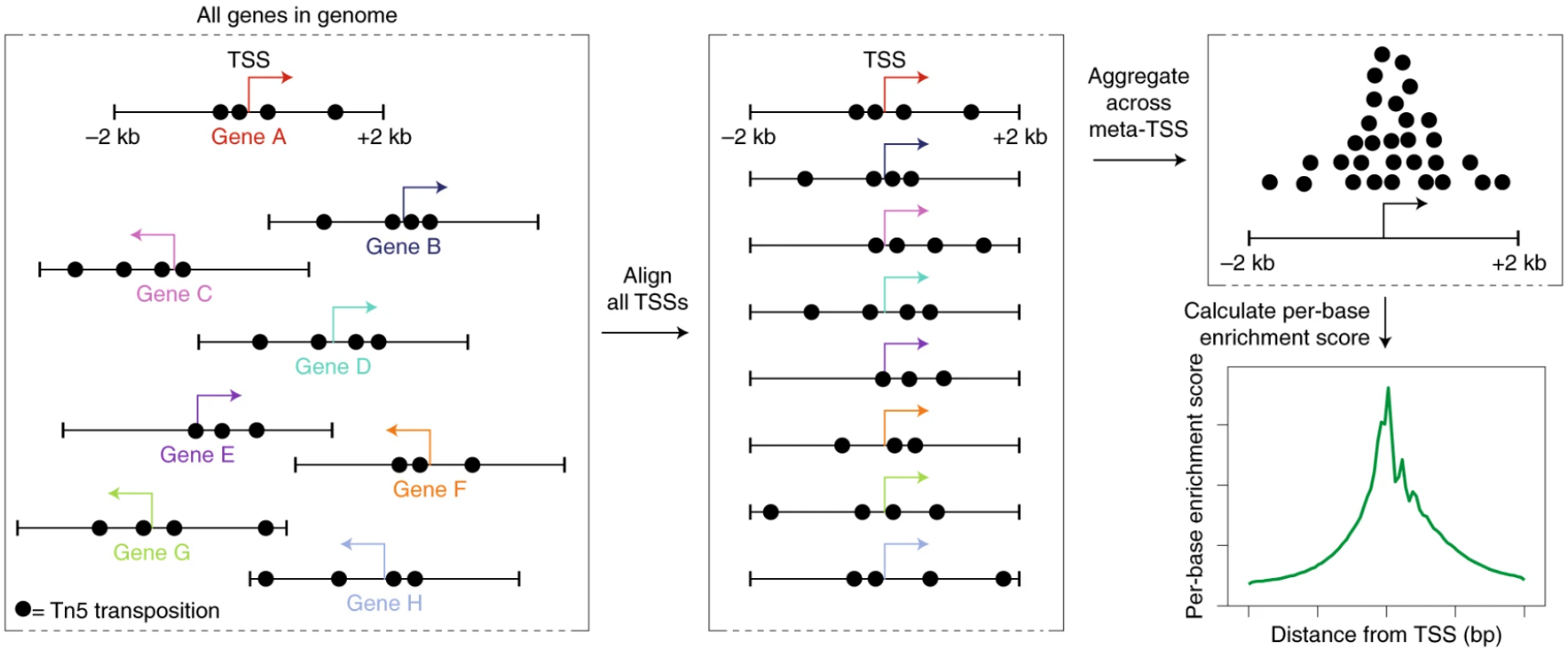
calculate_ATACseq_fragment_size_distribution.R

```
library("ATACseqQC")
pdf("sample.fragment_sizes.pdf", w=11, h=8.5)
fragSizeDist(”sample.bam", "sample")
dev.off()
```



Jason D Buenrostro, et.al. Nature Methods 10 (2013)

(Transcription Start Site) TSS enrichment score



Nat Protoc 17, 1518–1552 (2022)

(Transcription Start Site) TSS enrichment score

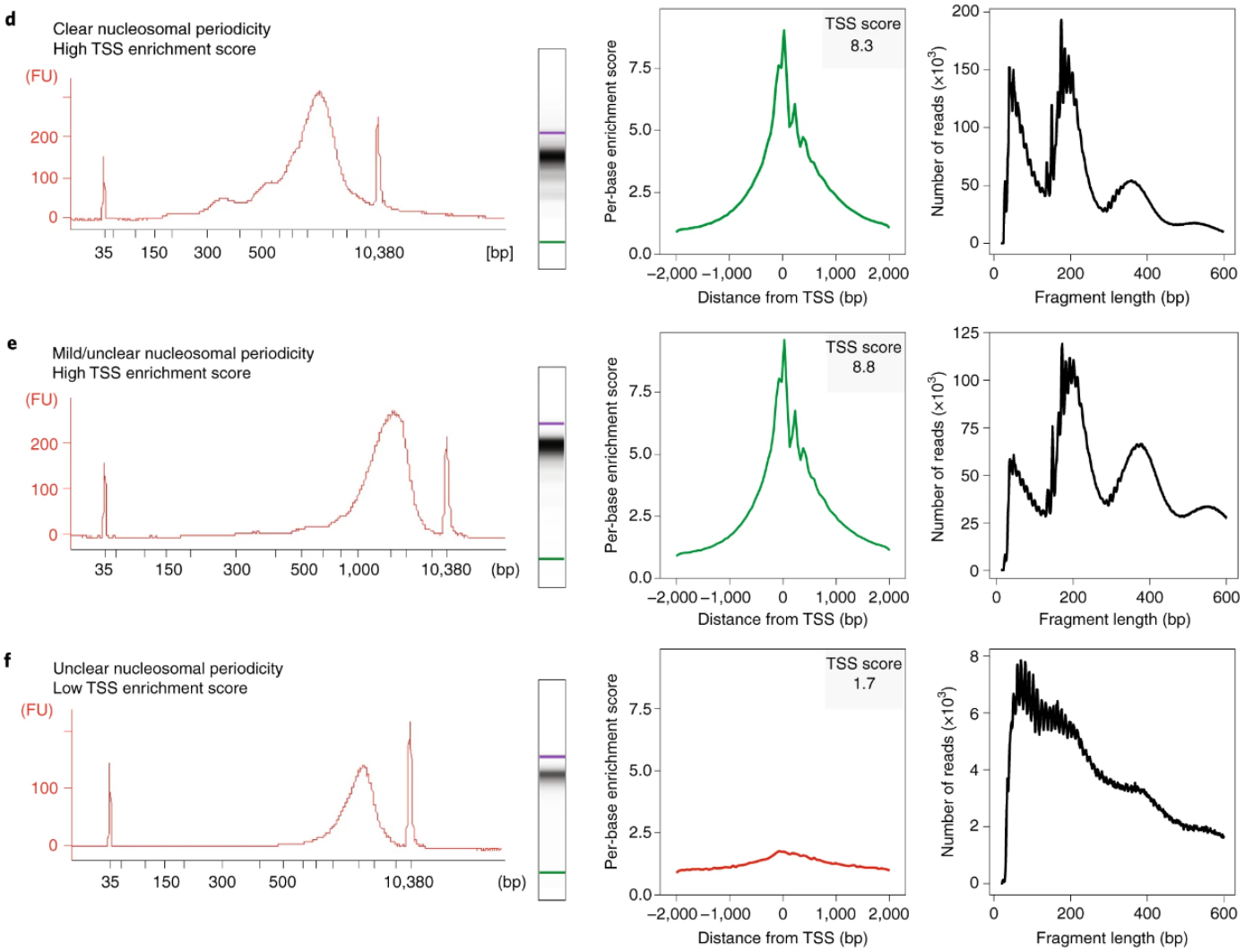
ENCODE standard with one TSS per gene:

- Score with TSS per transcript is smaller than with one TSS per gene

Annotation used	Value	Resulting Data Status
hg19 Refseq TSS annotation	< 6	Concerning
	6 - 10	Acceptable
	> 10	Ideal
GRCh38 Refseq TSS annotation	< 5	Concerning
	5 - 7	Acceptable
	> 7	Ideal
mm9 GENCODE TSS annotation	< 5	Concerning
	5 - 7	Acceptable
	> 7	Ideal
mm10 Refseq TSS annotation	< 10	Concerning
	10 -15	Acceptable
	> 15	Ideal

calculate_TSS_enrichment_score.py

Good ATAC-seq library with a high TSS enrichment score and clear nucleosomal periodicity



Nat Protoc 17, 1518–1552 (2022)

Peak calls

- Call peaks with MACS

1. Call peaks centering around the cutting sites
2. Limited to nucleosome free region (NFR) with BAMPE mode

```
macs2 callpeak -f BAMPE -t NFR.bam  --keep-dup all -B -q 0.01 -g mm -n MACS_ATACSeq_Peaks
```

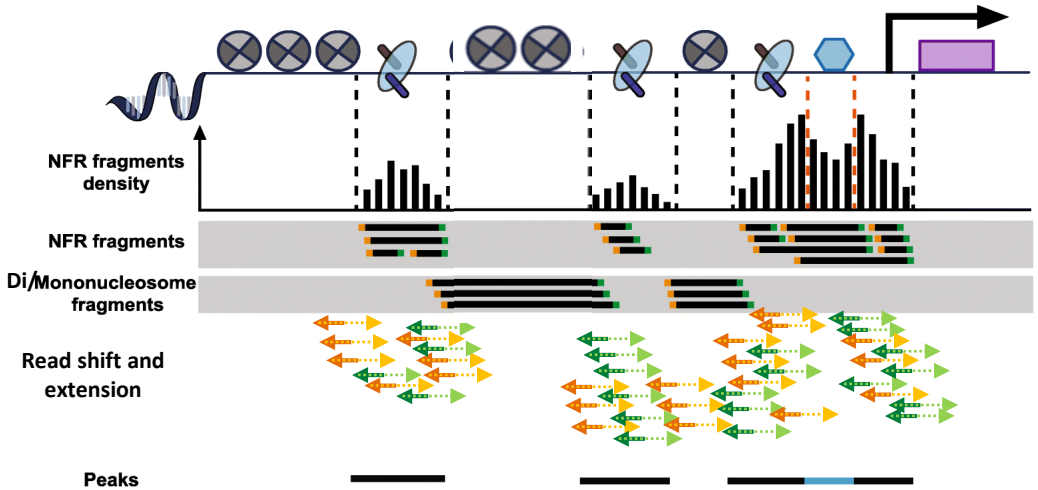
- Which method to use?
  - Read length?
  - Verify with genome browser

- Call peaks with Genrich, HMMRATAC

BaRC

Identify potential open chromatin regions (peaks)

regions with reads piled up to a greater extent than the background read coverage



Modified from Feng Yan et.al Genome Biology 21 (2020)

Create a signal profile centered around the cutting sites

Convert bam to bed

- Consider both reads of a pair:

  bedtools bamtobed -i foo.bam > foo.bed

- Convert an input paired-end alignment:

  K00168:88:HFF7YBBXX:1:1203:28371:20304 163 chr1 3199165 42 51M = 3199703 589 ACTAAAAACAGACAAATGCTCAACATTTACATGAAATGTAAGACTAAATAT AA-FFFJA-FAJJJ<--FA-7AJ7-AJJJF-<77AJ-A<F<FJA7FJJJJA MD:Z:51 PG:Z:MarkDuplicates XG:i:0 NM:i:0 XM:i:0 XN:i:0 XO:i:0 AS:i:0 YS:i:0 YT:Z:CP

  K00168:88:HFF7YBBXX:1:1203:28371:20304 83 chr1 3199703 42 51M = 3199165 -589 ACTTTGAAAAAAATGAGTAACAGACTTCTGTTAAAATGACCACAGTGTACT JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA MD:Z:51 PG:Z:MarkDuplicates XG:i:0 NM:i:0 XM:i:0 XN:i:0 XO:i:0 AS:i:0 YS:i:0 YT:Z:CP

- To bed output file:

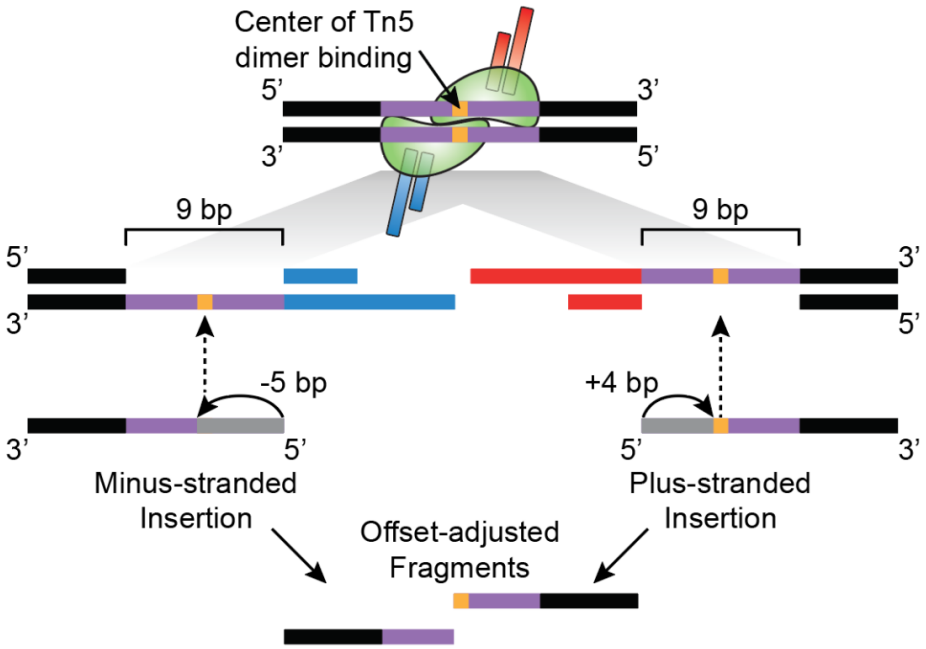
  chr1 3199164 3199215 K00168:88:HFF7YBBXX:1:1203:28371:20304/2 42 +

  chr1 3199702 3199753 K00168:88:HFF7YBBXX:1:1203:28371:20304/1 42 -

BaRC

Adjusted the read start sites to represent the center of the transposon binding event

- The Tn5 adapters are inserted in a staggered manner into the 5' ends of target sequence strands with a 9-bp gap between them
- The center of the Tn5 binding is 4 bp to the right of the edge on positive-strand reads, or 5 bp to the left on negative-strand reads.



Nat Protoc 17, 1518–1552 (2022)

Shift reads

- Reads should be shifted + 4 bp and − 5 bp for positive and negative strand respectively, to account for the 9-bp duplication created by DNA repair of the nick by Tn5 transposase

  cat foo.bed | awk -F $'\t' 'BEGIN {OFS = FS}{ if ($6 == "+") {$2 = $2 + 4} else if ($6 == "-") {$3 = $3 - 5} print $0}' >| foo_tn5.bed

Bed format:

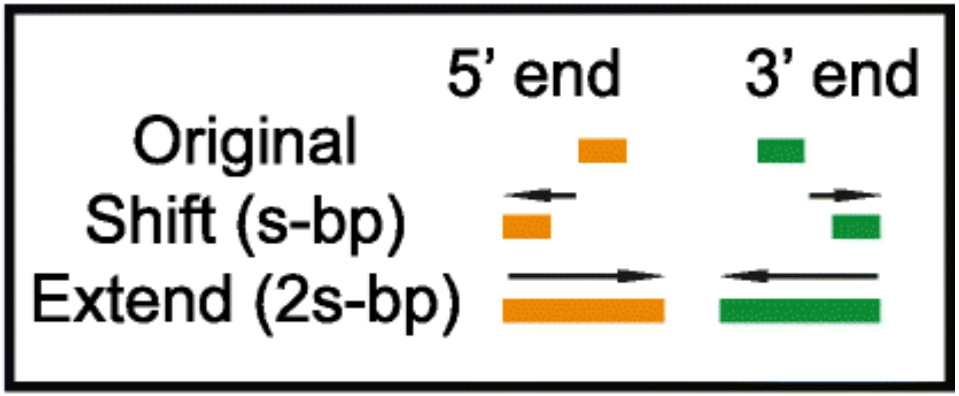
Chr	start	end	read_name	score	strand
chr1	3199164	3199215	K00168:88:HFF7YBBXX:1:1203:28371:20304/2	42	+
chr1	3199702	3199753	K00168:88:HFF7YBBXX:1:1203:28371:20304/1	42	-

BaRC

Call peaks

macs2 callpeak -t foo_tn5.bed -n foo -f BED -g mm -q 0.01 --nomodel --shift -75 --extsize 150 --call-summits --keep-dup all

Shift the reads to create a signal profile centered around the cutting sites



Feng Yan et.al Genome Biology 21 (2020)

MACS output

Output files:

1. Excel peaks file ("_peaks.xls") contains the following columns

   ```
   Chr, start, end, length, abs_summit, pileup,
   -LOG10(pvalue), fold_enrichment, -
     LOG10(qvalue), name
   ```

2. "_summits.bed": contains the peak summits locations for every peaks. The 5th column in this file is -log10qvalue
3. "_peaks.narrowPeak" is BED6+4 format file. Contains the peak locations together with peak summit, fold-change, pvalue and qvalue.

BaRC

Visualize peaks in IGV

Refseq Genes

peaks

summit

ATAC-seq

FRiP (Fraction of reads in peaks) score

- Fraction of all mapped reads that fall into the called peak regions
- The higher the score, the better

  According to ENCODE, score is preferably over 0.3, values greater than 0.2 are acceptable.

calculate_FRiP_score.py Sample.bam Sample_peaks.narrowPeak

BaRC

Blacklist filtering for peaks

- Anomalous, unstructured, or high signal in next-generation sequencing experiments independent of cell line or experiment

- Often in repeats (centromeres, telomeres, satellite repeats)
  - 0.5% of genome, but could account for >10% total signal

- BaRC_datasets -> ENCODE_blacklists
  - ENCODE: human, mouse, fly, or C. elegans

```
bedtools intersect -v -a foo_peaks.narrowPeak -b
blacklist.bed > bfiltered_peaks.narrowPeak
  #  -v option: only report those “A” peaks with no overlaps in “B”
```

Compare open chromatin between different experimental conditions

- Without replicates, you can use bedtools to compare two samples:

  intersectBed (finds the subset of peaks **common** in 2 samples **or unique** to one them)

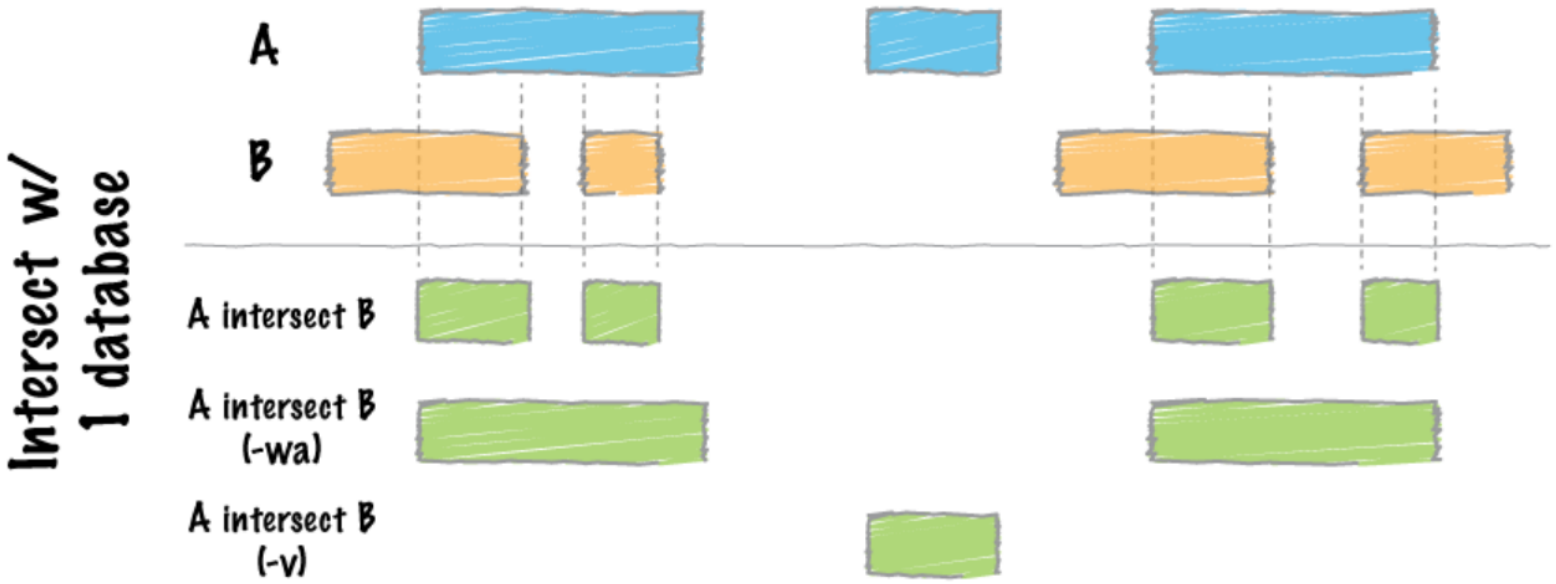
- If you have replicates, you can use:
  1) bedtools merge to merge the peaks
  2) bedtools coverage to count the number of reads in peaks
  3) DESeq2/EdgeR on reads assigned the peaks to get differentially open peaks

BaRC

bedtools

- intersectBed

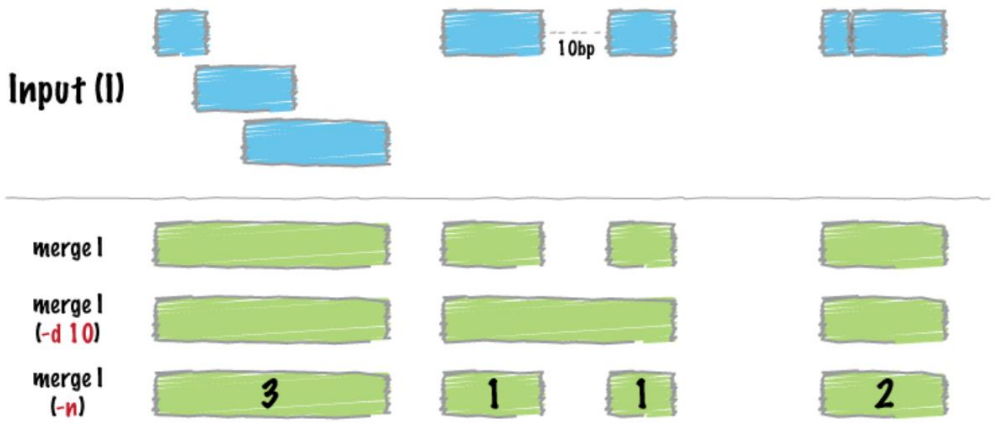
intersect



https://bedtools.readthedocs.io/

bedtools

merge



coverage

Below are the number of features in A (N=...) overlapping B and fraction of bases in B with coverage.

```
Chromosome  ========================================================================
BED File B  ===============     ================     ======     =============
BED File A  ==== ====               ==               =========    === == ====
              ========                                        ===== ===== ==


Result      [  N=3, 10/15 ]     [   N=1, 2/16  ]    [N=1,6/6]  [N=5, 11/12 ]
```

Linking peaks to genes: Bedtools

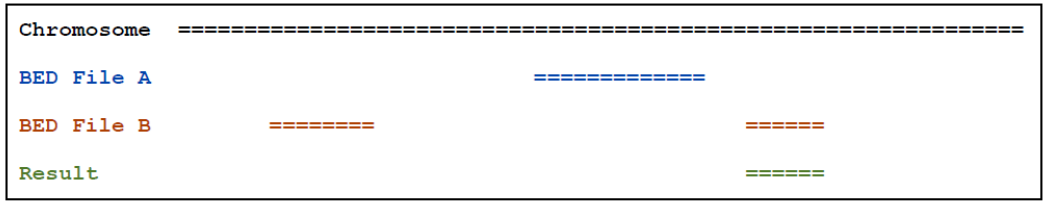
slopBed: extend a feature by a user-defined number of bases

```
$ cat A.bed
chr1  5     100
chr1  800   980

$ cat my.genome
chr1  1000

$ slopBed -i A.bed -g my.genome -b 5
chr1  0     105
chr1  795   985

$ slopBed -i A.bed -g my.genome -l 2 -r 3
chr1  3     103
chr1  798   983
```

groupBy

It groups rows based on the value of a given column/s and it summarizes the other columns

closestBed



Linking peaks to nearby genes:

Get all the genes at a certain distance (*i.e.* 3Kb) of the peak.
The distance we use depends on the area where we want to find regulatory interactions.

1. Take all genes and add 3Kb up and down with slopBed
    1. slopBed -b 3000 -i GRCh37.p13.HumanENSEMBLgenes.bed -g /nfs/genomes/human_gp_feb_09_no_random/anno/chromInfo.txt > HumanGenesPlusMinus3kb.bed
2. Intersect the slopped genes with peaks and get the list of unique genes overlapping

    intersectBed -wa -a HumanGenesPlusMinus3kb.bed -b peaks.bed | awk '{print $4}' | sort -u > Genesat3KborlessfromPeaks.txt

intersectBed -wa -a HumanGenesPlusMinus3kb.bed -b peaks.bed | head -3

```
chr1    45956538    45968751    ENSG00000236624_CCDC163P
chr1    45956538    45968751    ENSG00000236624_CCDC163P
chr1    51522509    51528577    ENSG00000265538_MIR4421
```


IO BaRC

Link peaks to closest gene

For each region find the closest gene and filter based on the distance to the gene

The commands below are an example where we are looking for interactions at 3Kb or less.

```
closestBed -d -a  peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed |head
chr1    20870   21204   H3k4me3_chr1_peak_1     5.77592         chr1    14363   29806   ENSG00000227232_WASH7P          0
chr1    28482   30214   H3k4me3_chr1_peak_2     374.48264       chr1    29554   31109   ENSG00000243485_MIR1302-10      0
chr1    28482   30214   H3k4me3_chr1_peak_2     374.48264       chr1    14363   29806   ENSG00000227232_WASH7P          0
#the next two steps can also be done on excel
closestBed -d -a  peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed | groupBy -g 9,10  -c
    6,7,8, -o distinct,distinct,distinct | head -3
ENSG00000227232_WASH7P              0       chr1    14363   29806
ENSG00000243485_MIR1302-10          0       chr1    29554   31109
ENSG00000227232_WASH7P              0       chr1    14363   29806

closestBed -d -a  peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed | groupBy -g 9,10  -c 6,7,8, -o
    distinct,distinct,distinct | awk  'BEGIN {OFS="\t"}{ if ($2<3000)  {print $3,$4,$5,$1,$2} } ' | head -5
chr1    14363   29806   ENSG00000227232_WASH7P  0
chr1    29554   31109   ENSG00000243485_MIR1302-10      0
chr1    14363   29806   ENSG00000227232_WASH7P  0
chr1    134901  139379  ENSG00000237683_AL627309.1      0
chr1    135141  135895  ENSG00000268903_RP11-34P13.15   0
```


BaRC

Link peaks to closest gene (1 command)

For each region find the closest gene and filter based on the distance to the gene

The command below is an example where we are looking for interactions at 3Kb or less.

```
closestBed -d -a  peaks.bed -b
GRCh37.p13.HumanENSEMBLgenes.bed | groupBy -g 9,10  -c 6,7,8,
-o distinct,distinct,distinct | awk  'BEGIN {OFS="\t"}{ if ($2<3000)
{print $3,$4,$5,$1,$2} }'  > closestGeneAt3KborLess.bed
```

closestBed

-d print the distance to the feature in -b

groupBy

-g columns to group on

-c columns to summarize

-o operation to use to summarize

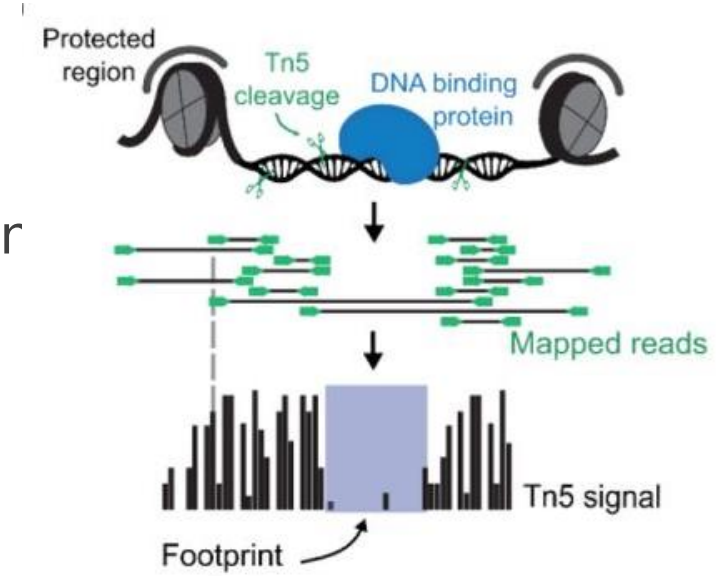
Identify footprints

- DNA sequences directly bound by TFs
- High depth of coverage:
  At least ~200 million of reads
- Shift reads to account for the 9-bp duplication

- Tools:

  HINT: (**H**mm-based **I**de**N**tificatior footprints)
  Genome Biol **20**, 45 (2019)

  TOBIAS:
  Nat Commun **11**, 4267 (2020)



Bentsen, M., Goymann, P., Schultheis, H. *et al. Nat Commun* 11 (2020)

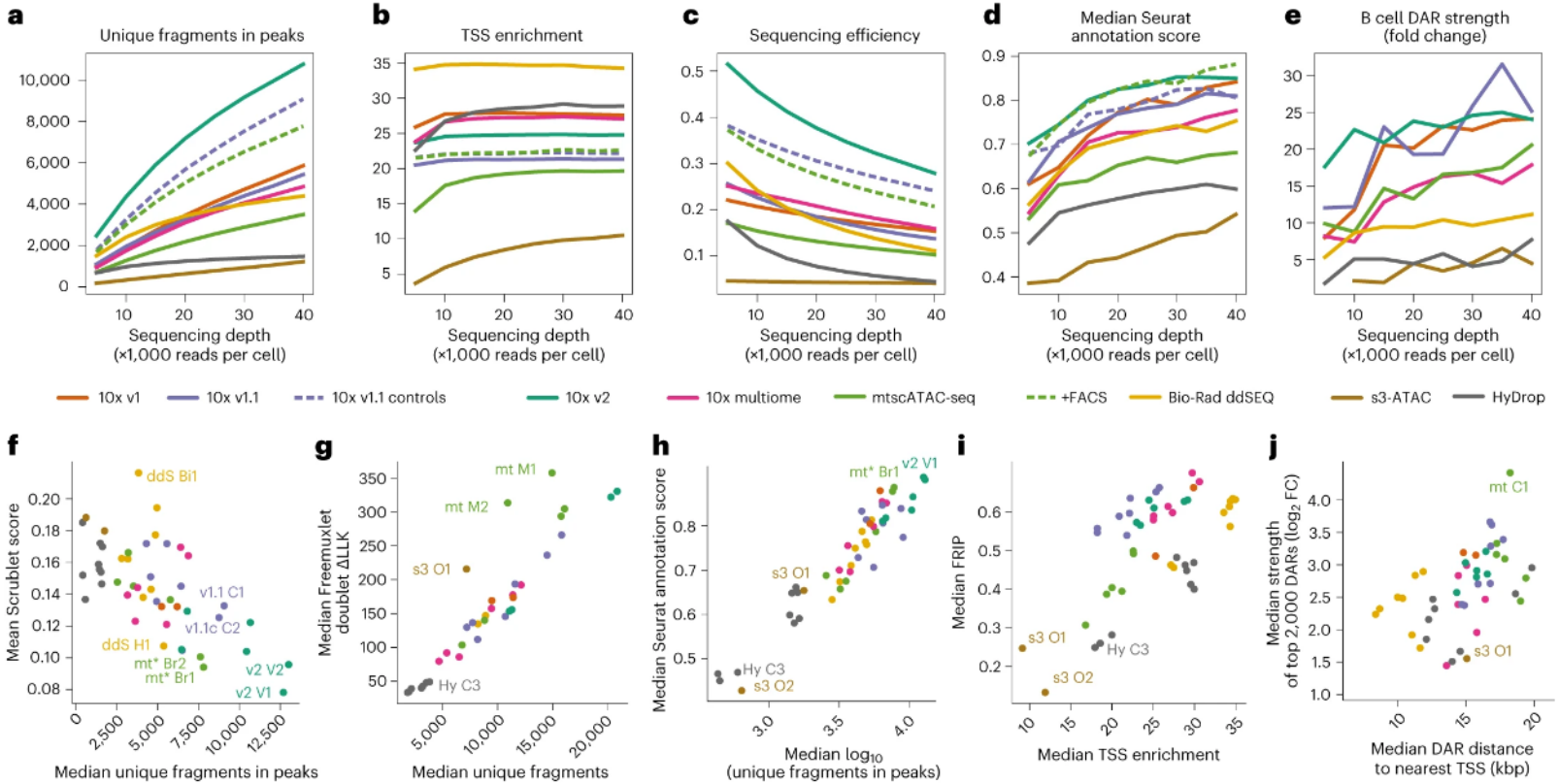
Single cell ATAC-seq (scATAC-seq)

- scATAC-seq is good for heterogeneous samples with multiple cell types or cell states, where bulk ATAC-seq considers average cell signals and could miss signals from a subset of cells
- Signal is sparse in scATAC-seq
- Remove cells with low quality or doublets
- Make cell-by-feature matrix
- Generate and annotate cell clusters

BaRC

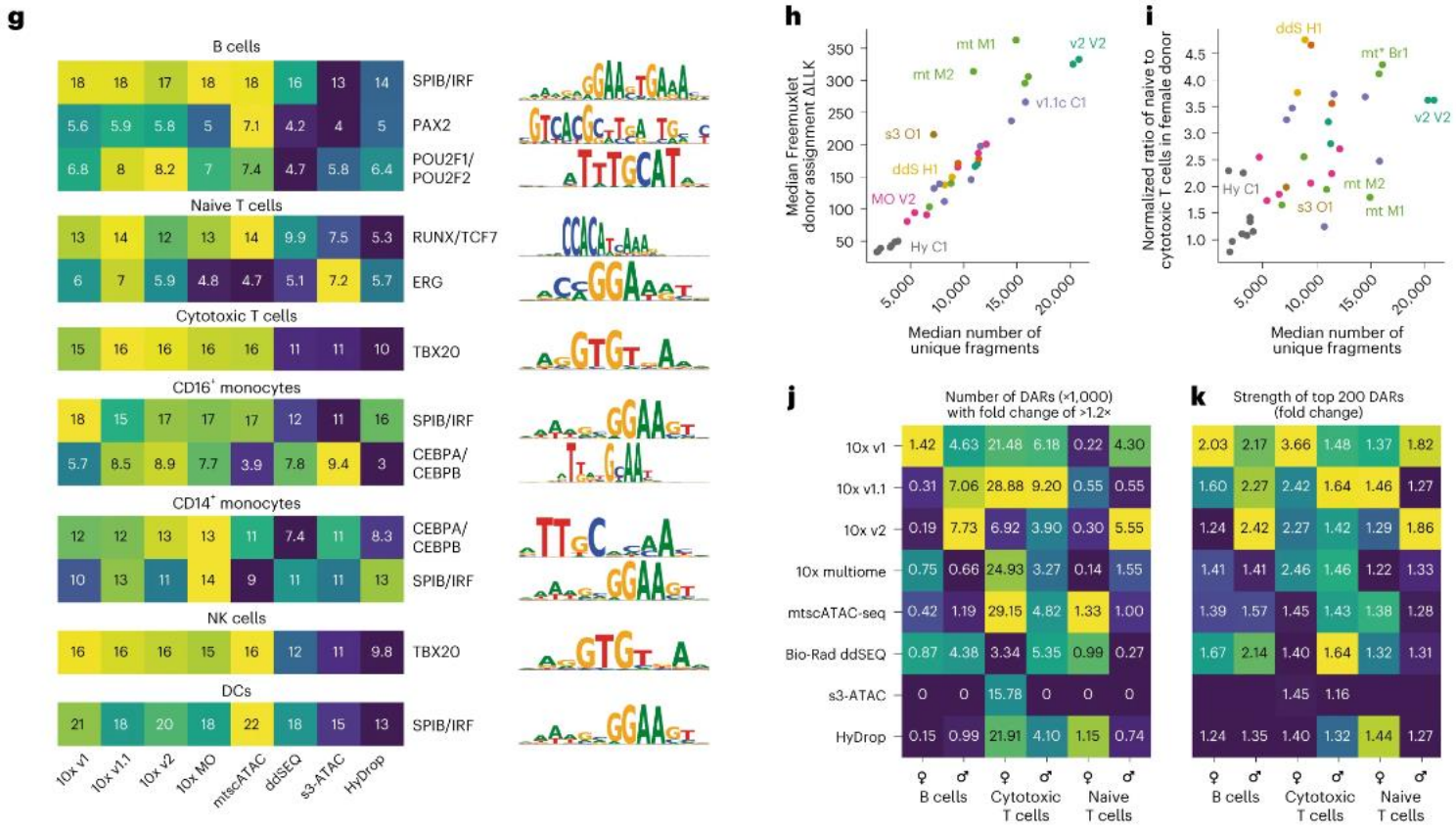
Fig. 2: Differences in automated cell-type annotation accuracy and differential region calling between techniques.

From: Systematic benchmarking of single-cell ATAC-sequencing protocols



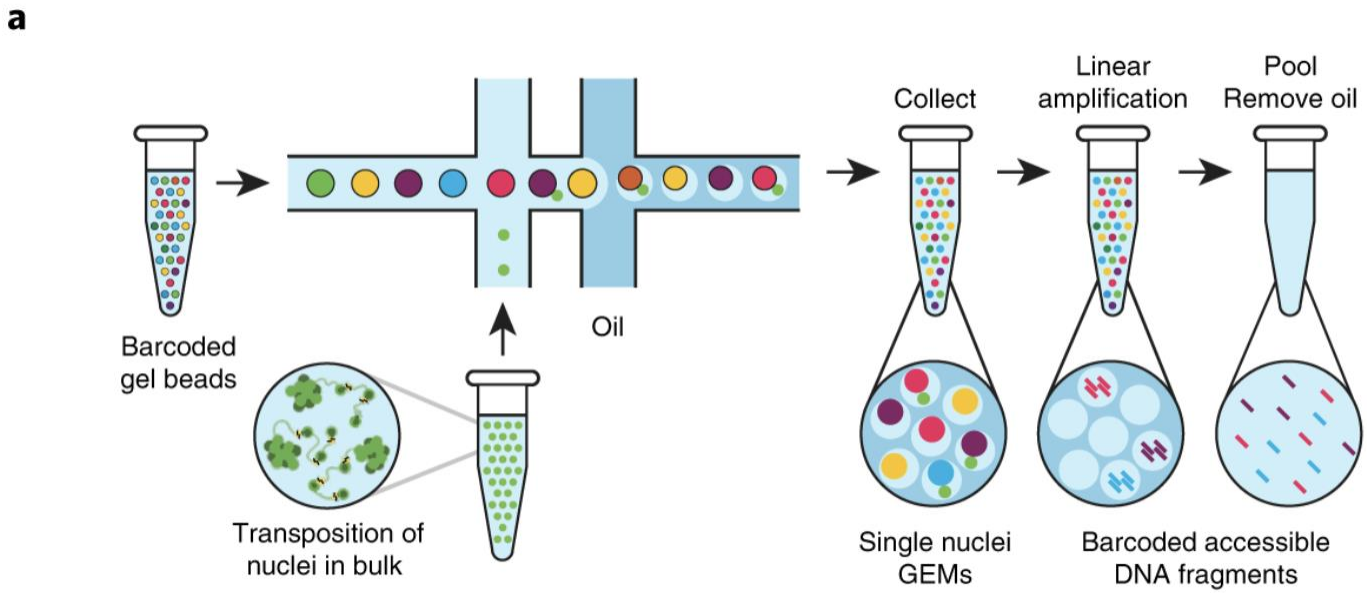
De Rop, F.V., Hulselmans, G., Flerin, C. et al. Nat Biotechnol (2023)

Performance differences



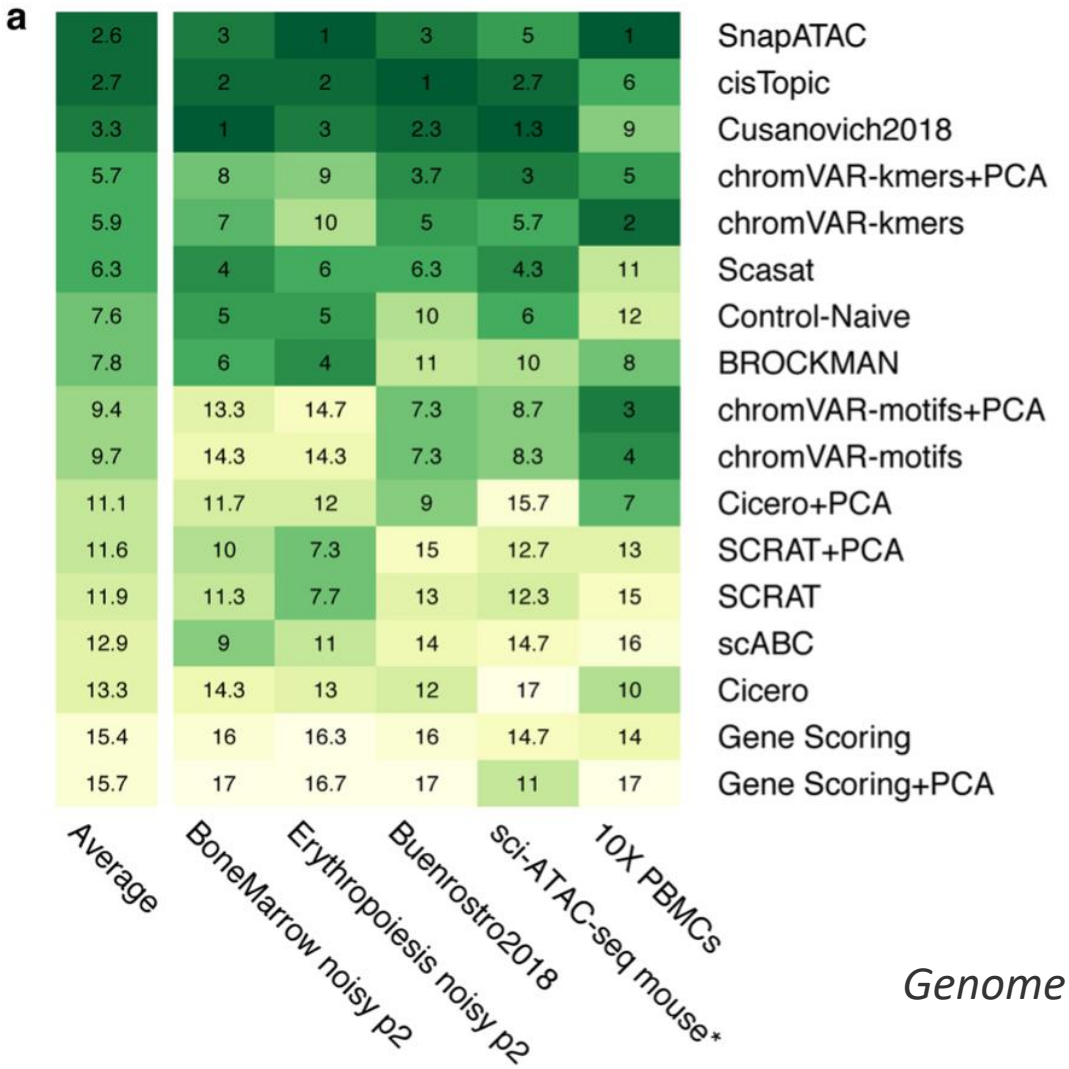
De Rop, F.V., Hulselmans, G., Flerin, C. et al. Nat Biotechnol (2023)

Single cell ATAC-seq (scATAC-seq)



Nat Biotechnol **37**, 925–936 (2019).

scATAC-seq analysis tools



Genome Biol ***20****, 241 (2019)*

The darker shades of green indicates the better performance in clustering

scATAC-seq analysis tools

Table 1
Summary of scATAC-seq analysis software packages.

Tool	Platform	Feature Matrix	Preprocessing	Clustering	DAR	Motif/k-mer	Gene activity	Co-accessibility	Trajectory	Pathway	Enrichment analysis	scRNA integration	Reference
ChromVAR	R	TF motifs, k-mer	O	O	X	O	X	X	X	X	X	X	[17]
SCRAT	R/Web	Selectable feature	O	O	O	X	X	X	X	X	X	X	[18]
scABC	R	Peak	O	O	X	O (ChromVAR)	X	X	X	X	X	X	[19]
Cicero	R	TSS	O	O	O	X	O	O	O	X	X	X	[20]
Scasat	Python/R	Peak	O	O	O	X	X	X	X	O (GREAT)	X	X	[21]
cisTopic	R	Peak	O	O	X	X	O	X	X	O	O	X	[22]
snapATAC	Python/R	Bin, peak	O	O	O	O (ChromVAR, Homer)	O	X	X	O (GREAT)	X	O (Seurat)	[23]
epiScanpy	Python	Peak	O	O	X	X	X	X	X	X	X	X	[24]
Destin	R	Peak	O	O	O	X	X	X	X	X	O	X	[25]
SCALE	Python	Peak	O	O	O	O (ChromVAR)	X	X	X	X	X	X	[26]
scATAC-pro	Python/R	Peak	O	O	O	O (ChromVAR)	O	O (Cicero)	X	O (GREAT)	X	X	[27]
Signac	R	Peak	O	O	O	O (ChromVAR)	O	X	X	X	X	O (Seurat)	[7]
ArchR	R	Bin, peak	O	O	O	O (ChromVAR), TF footprinting	O	O	O	X	O	O (Seurat)	[28]

Tools used in junction are indicated in parentheses.

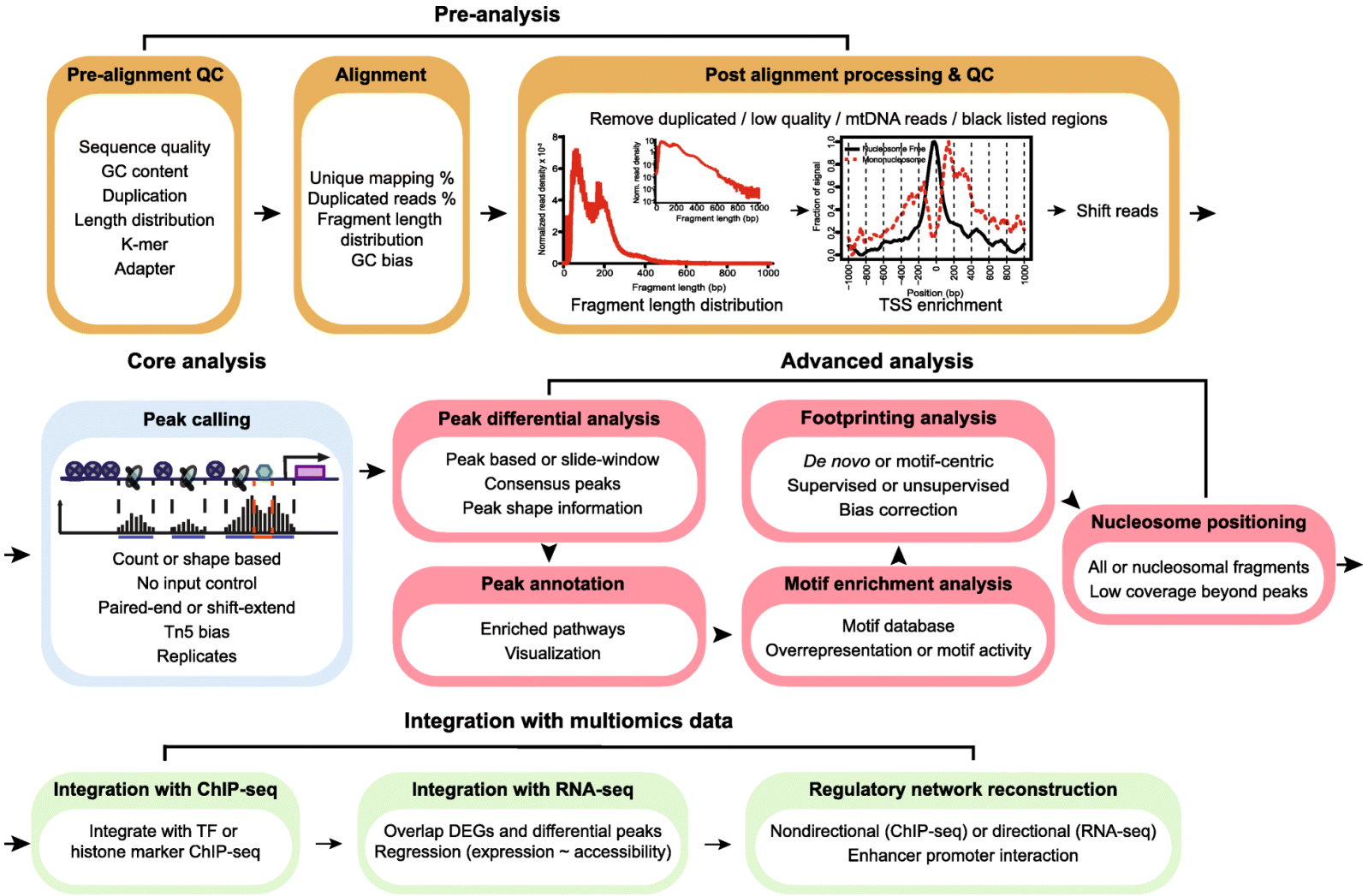
Computational and Structural Biotechnology Journal 18 (2020)

Seurat scATAC-seq analysis pipeline:

scRNA integration

implements packages (MACS2, monocle, Cicero, etc)

ATAC-seq analysis



Feng Yan et.al
Genome Biology 21 (2020)

Exercises:

- Mouse lung tissue postnatal (0 days) from ENCODE

  Randomly chosen 100,000 pair-end reads from the 1st replicate

- \\wi-files1\BaRC_Public\Hot_Topics\ ATAC-seq\ ATAC-seq_2024_commands.txt

References

- From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* 21, 2020
- Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10, 2013
- ENCODE ATAC-seq guideline: *https://github.com/ENCODE-DCC/atac-seq-pipeline*
- MACS:
  - Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9, 2008 *https://liulab-dfci.github.io/software/*
  - Using MACS to identify peaks from ChIP-Seq data. *Curr Protoc Bioinformatics*. 2011
- Picard Tools: https://broadinstitute.github.io/picard/
- Bedtools: *https://code.google.com/p/bedtools/*
  - BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics 15, 2010*
- Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37**, 925–936 (2019)
- Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation. *Computational and Structural Biotechnology Journal, 18, 2020*
- Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol* **20**, 241 (2019).
- Identification of transcription factor binding sites using ATAC-seq. *Genome Biol* **20**, 45 (2019)
- ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun* **11**, 4267 (2020)
- *Grandi, F.C., Modi, H., Kampman, L. et al. Chromatin accessibility profiling by ATAC-seq. Nat Protoc 17, 1518–1552 (2022).*

Other resources

- Previous Hot Topics

  Quality Control

  http://barc.wi.mit.edu/education/hot_topics/NGS_QC_2017/slides4perPage.pdf

- Best Practices:

  http://barcwiki.wi.mit.edu/wiki/SOPs/atac_Seq

Upcoming Hot Topics sessions

- CUT & Tag analysis April 25th

Let us know if you're interested:

- Protein structure visualization
- CRISPR pooled screen analysis
- Enrichment analysis
- Other suggestions