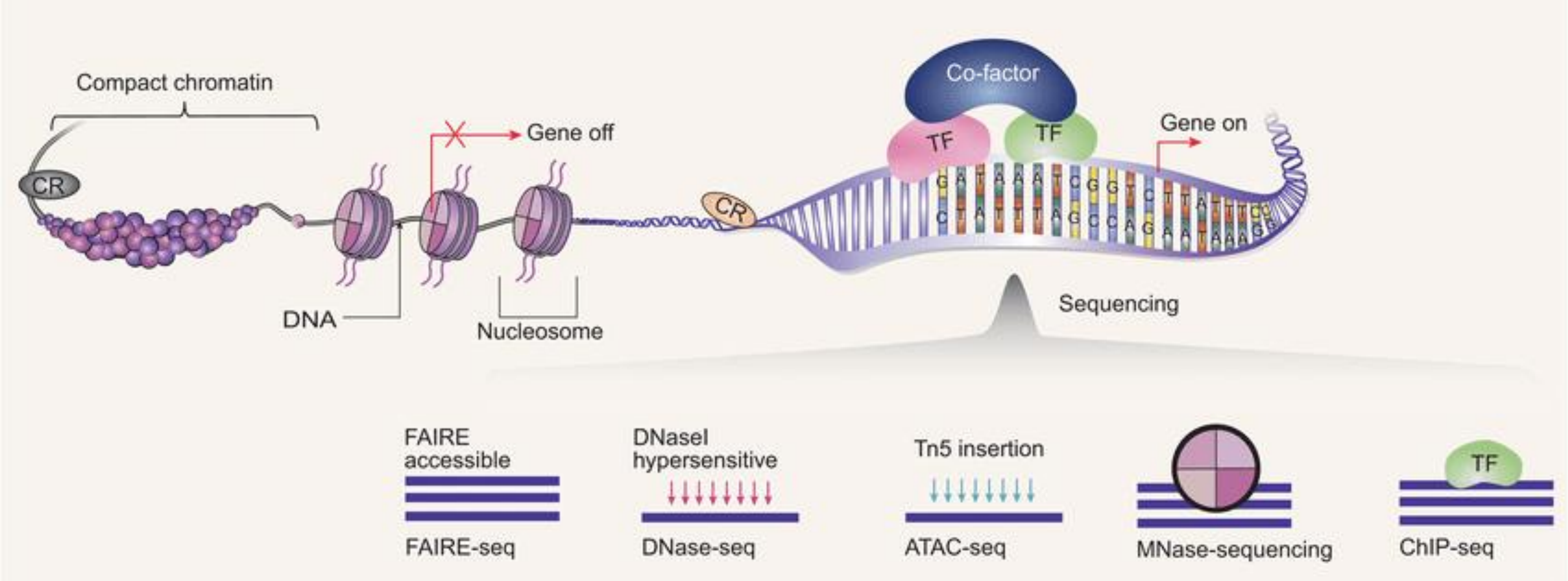
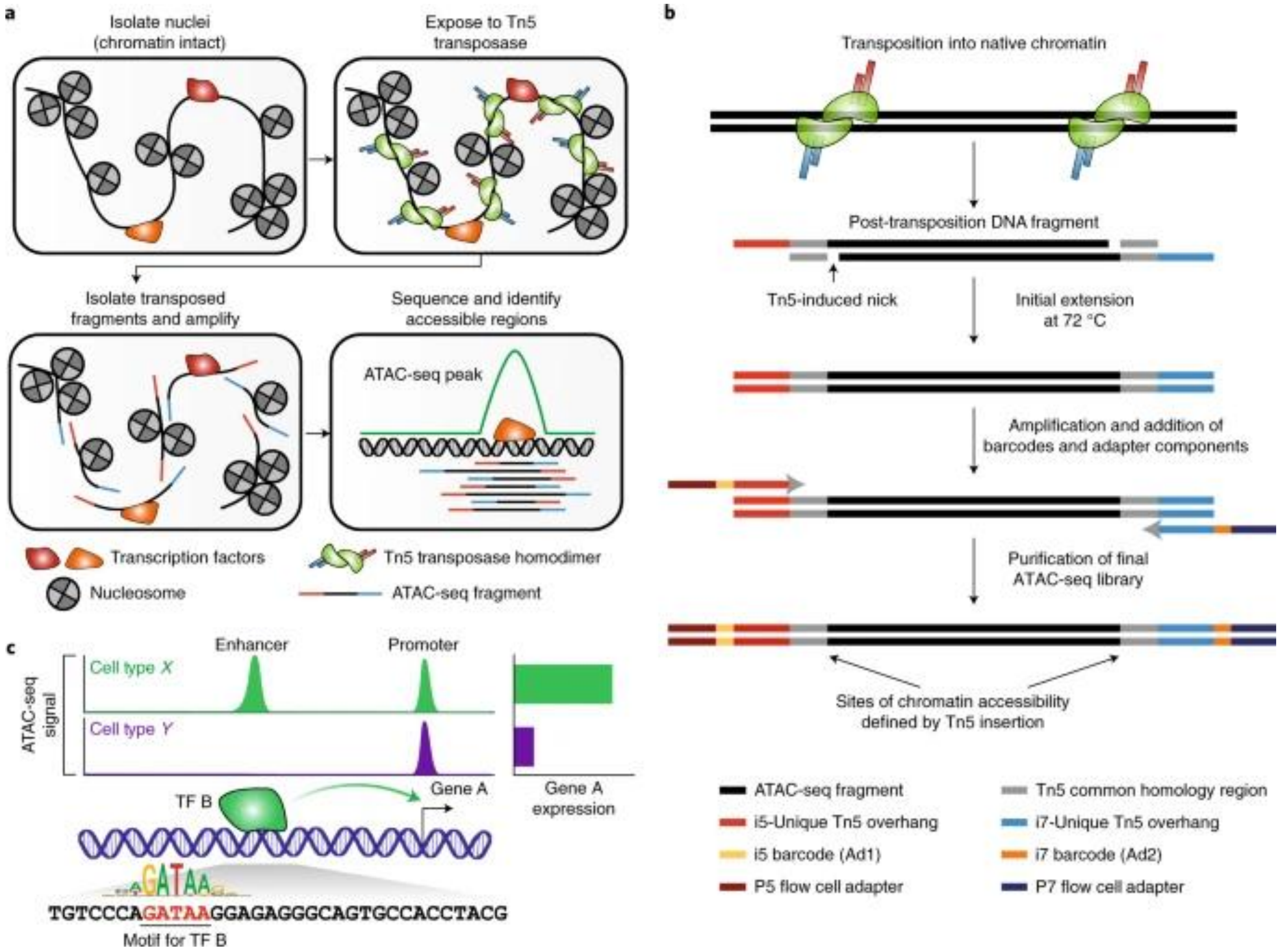
scATAC-seq

- Open chromatin region



-Only open regions can be accessible to TF + etc …
-Epigenetic modulation to regulate gene expression

- Open chromatin region


a
Isolate nuclei (chromatin intact)
Expose to Tn5 transposase
Isolate transposed fragments and amplify
Sequence and identify accessible regions
ATAC-seq peak
Transcription factors
Tn5 transposase homodimer
Nucleosome
ATAC-seq fragment
b
Transposition into native chromatin
Post-transposition DNA fragment
Tn5-induced nick
Initial extension at 72 °C
Amplification and addition of barcodes and adapter components
Purification of final ATAC-seq library
Sites of chromatin accessibility defined by Tn5 insertion
ATAC-seq fragment
Tn5 common homology region
i5-Unique Tn5 overhang
i7-Unique Tn5 overhang
i5 barcode (Ad1)
i7 barcode (Ad2)
P5 flow cell adapter
P7 flow cell adapter
c
Enhancer
Promoter
ATAC-seq signal
Cell type X
Cell type Y
TF B
Gene A
Gene A expression
GATAA
TGTCCCAGATAAGGAGAGGGCAGTGCCACCTACG
Motif for TF B

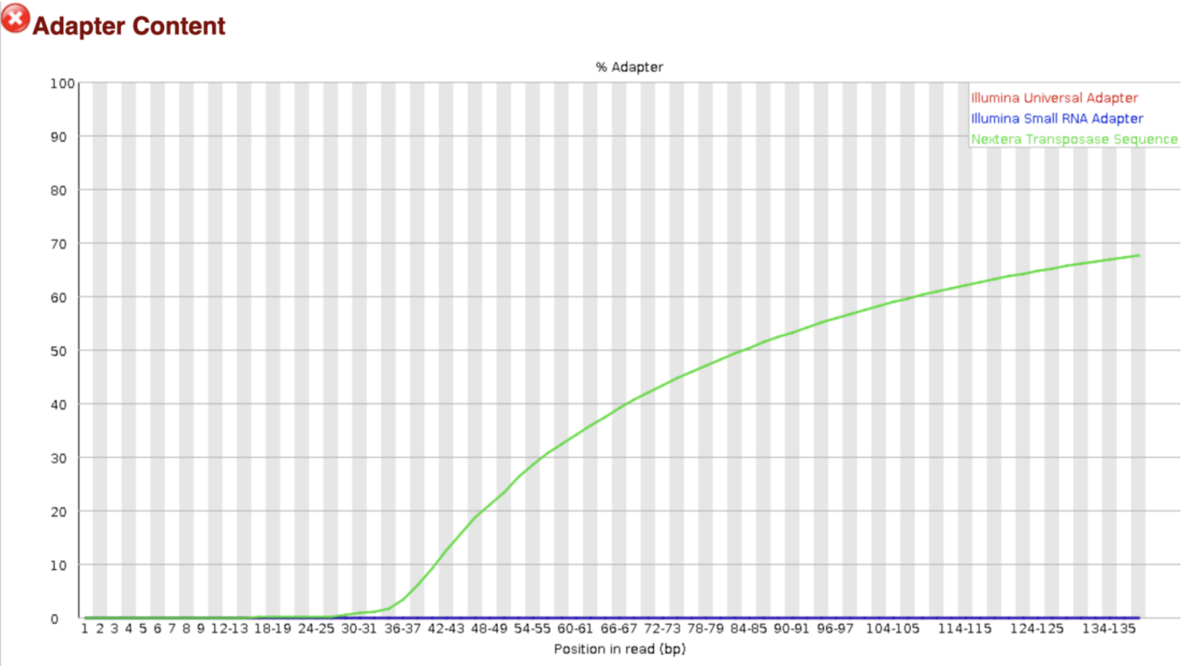
-Tn5 transposase
→ Insert sequencing adaptor
→ Sequencing
→ Captures open regions

Chromatin accessibility profiling by ATAC-seq

• Preprocessing

https://github.com/CebolaLab/ATAC-seq

-FASTQ → FASTQC



-fastp: adapter trimming

-Alignment: bowtie2

-Remove mitochondrial reads

-Remove duplicates

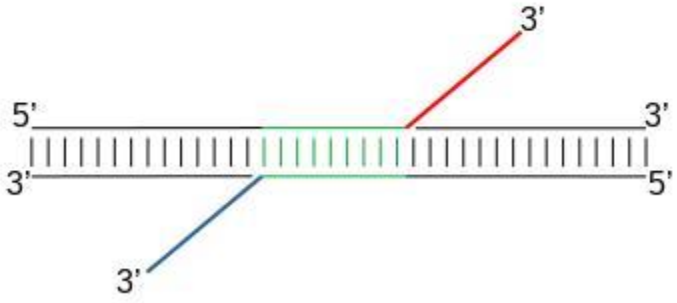
-remove multiple mapping

-remove ENCODE blacklist regions

-shift read coordinates

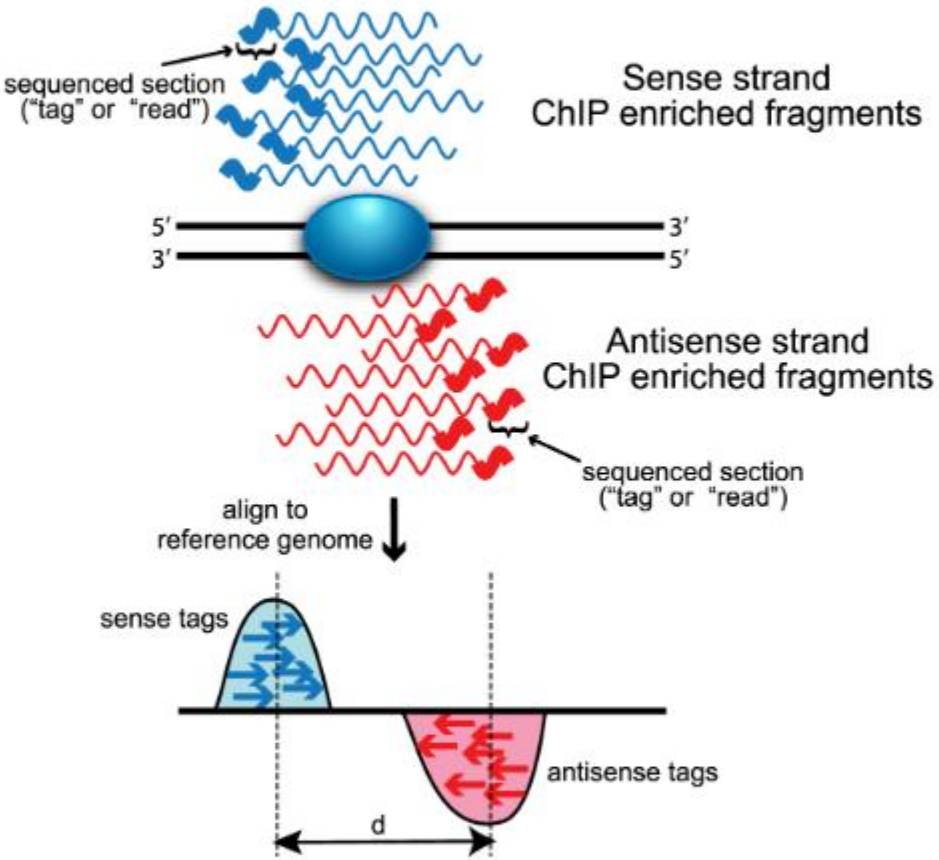
Tn5 small DNA insertion (introducted as repair of the transposase-induced nick introduces a 9bp insertion)

+ strand: offset by +4bp, - strand: -5bp



- # Peak calling

-MACS3



Cf) Chip-seq

→ +/- strand will be sequenced from TF

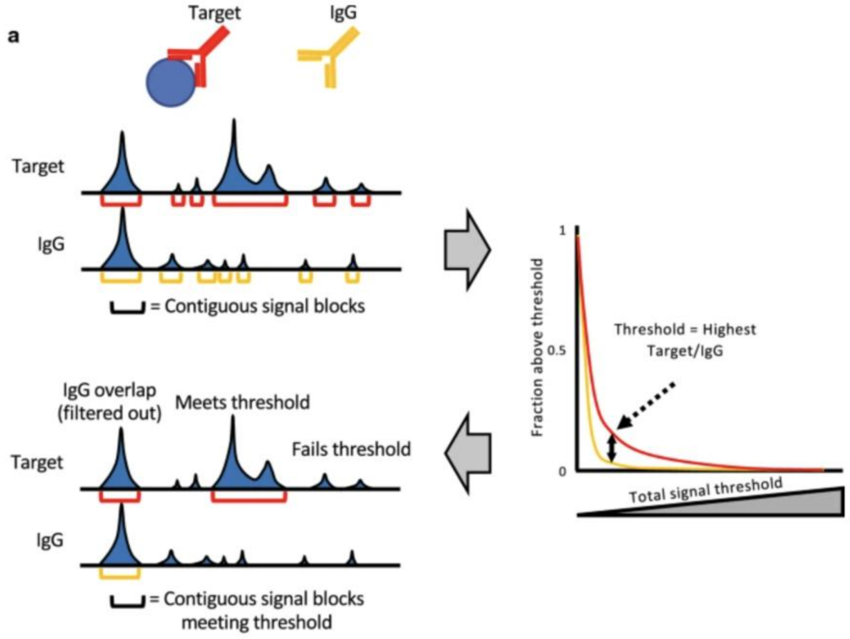
SEACR (Sparse Enrichment Analysis for CUT&RUN)

→ Due to Sparse signal

→ Calibration of background from global distribution

→ define peak threshold

Fig. 1



- Peak calling

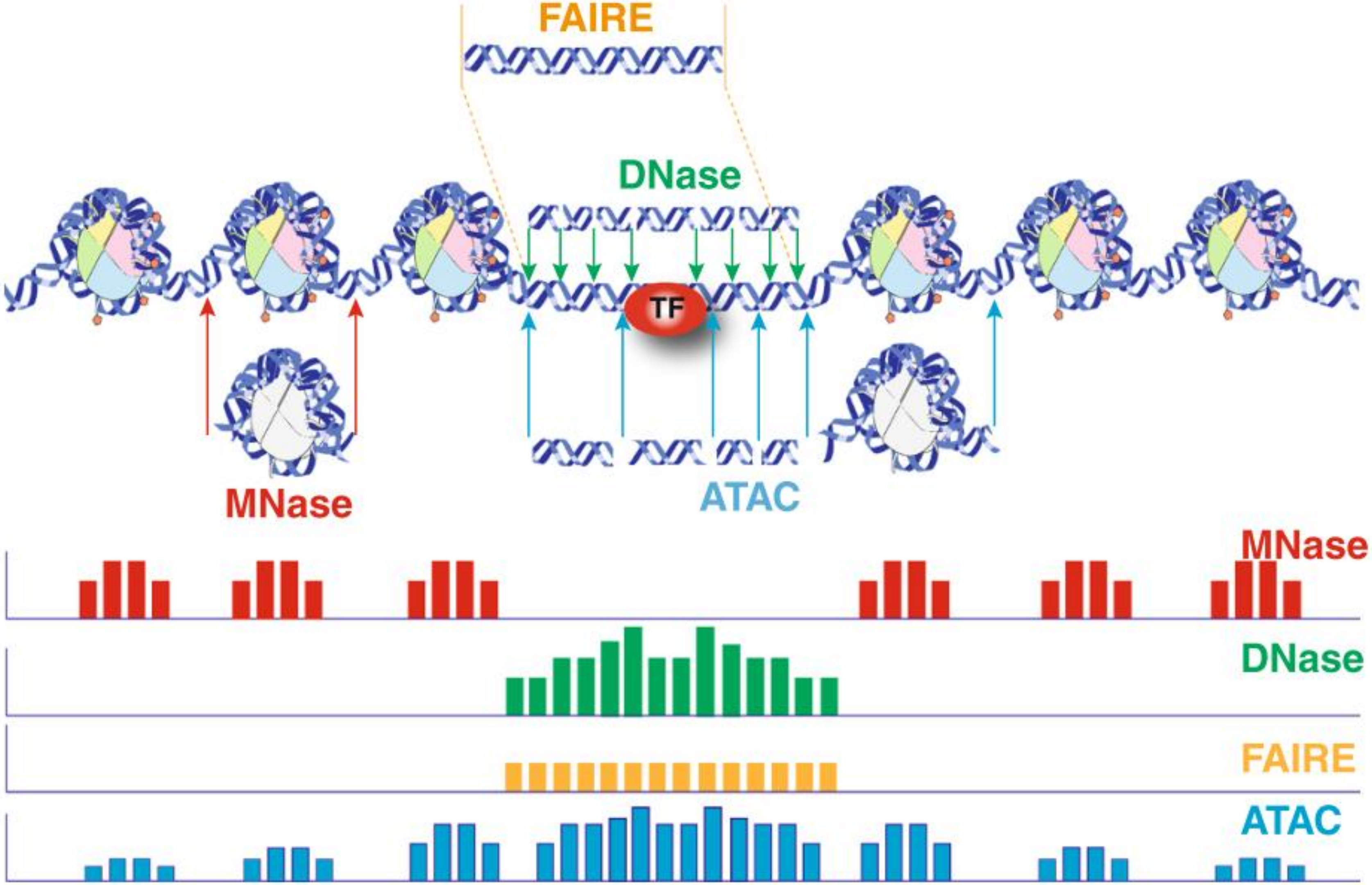
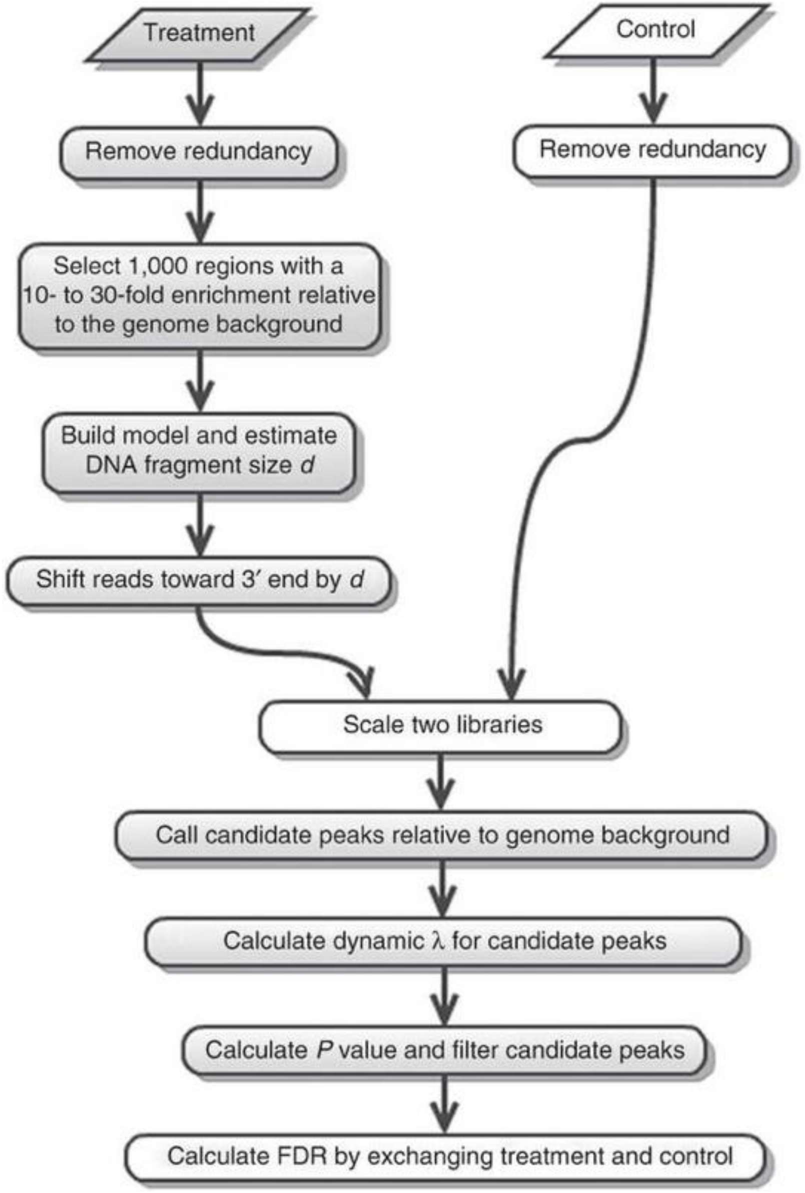


Figure 1 Schematic diagram of current chromatin accessibility assays performed with typical experimental conditions. Representative DNA fragments generated by each assay are shown, with end locations within chromatin defined by colored arrows. Bar diagrams represent data signal obtained from each assay across the entire region. The footprint created by a transcription factor (TF) is shown for ATAC-seq and DNase-seq experiments.

• Peak calling



Treatment

Remove redundancy

Select 1,000 regions with a 10- to 30-fold enrichment relative to the genome background

Build model and estimate DNA fragment size *d*

Shift reads toward 3′ end by *d*

Control

Remove redundancy

Scale two libraries

Call candidate peaks relative to genome background

Calculate dynamic λ for candidate peaks

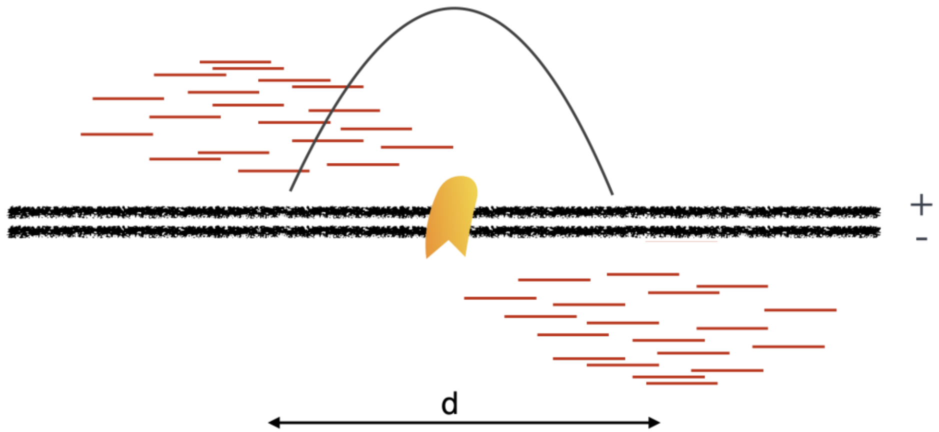
Calculate *P* value and filter candidate peaks

Calculate FDR by exchanging treatment and control

1: removing redundancy

- Duplicated tags, same seq at the same coordinate based on **binomial distribution**

Alignment generates a **bimodal pattern** on the plus and minus strands around binding sites



Peak calling algorithms use this pattern to estimate the relative strand shift

2: 600bp window → find enriched seq (red read)

3: d estimate → d/2: protein binding position

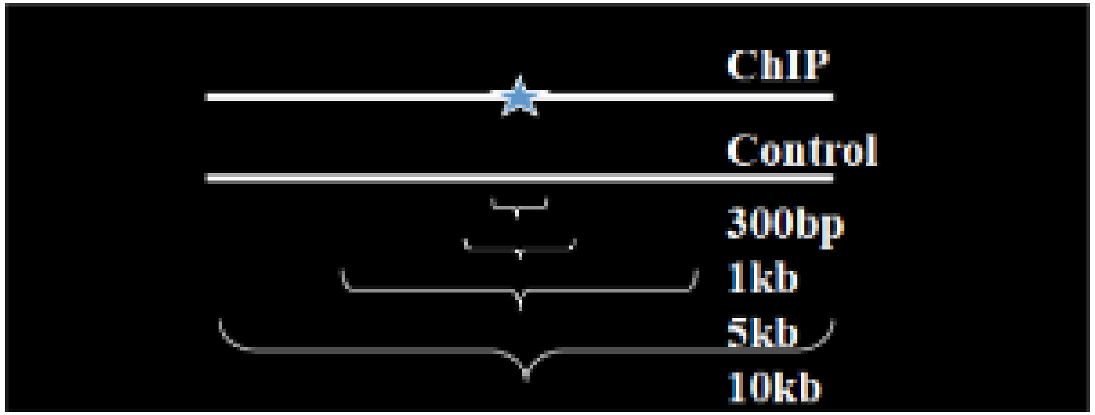
For ChIP-seq, but not ATAC → skip!

- # Peak calling

4: scaling the libraries
→ normalized by total tag count

5: effective genome length: remove low mappability repetitive region

6: peak calling
lambda: Number of reads in that window → follows poisson distribution
Evaluate lambda with multiple window size → optimize
Bg: whole mappable genome: effective genome)



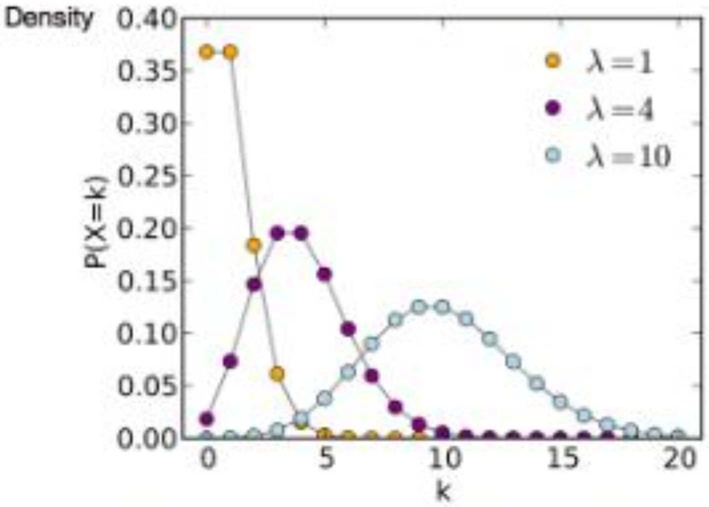
! Peak calling by reads vs background by Poisson distribution

$$P_\lambda(X=k) = \frac{\lambda^k}{k! * e^{-\lambda}}$$

λ = mean = expectated value = variance

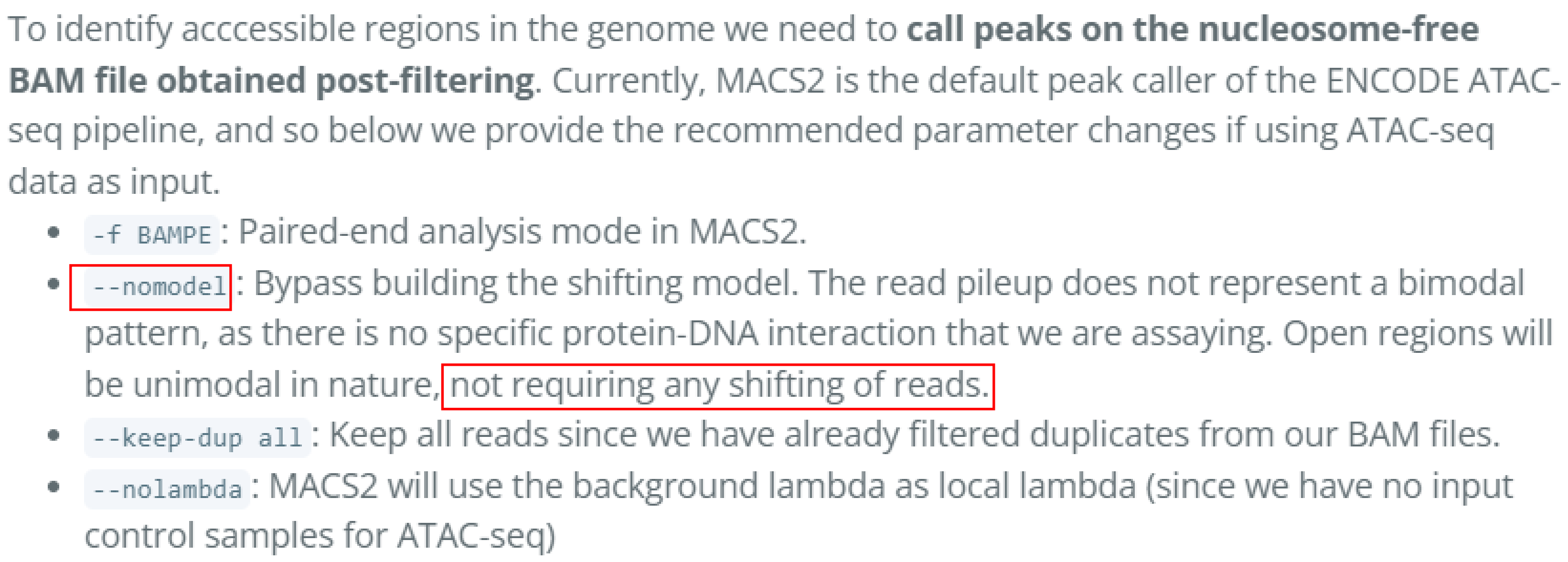
$$\lambda = \frac{\text{total number of events (k)}}{\text{number of units (n) in the data}}$$

$$= \frac{\text{Read length (nt) * Total read number}}{\text{Effective genome length (nt)}}$$



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

- Peak calling

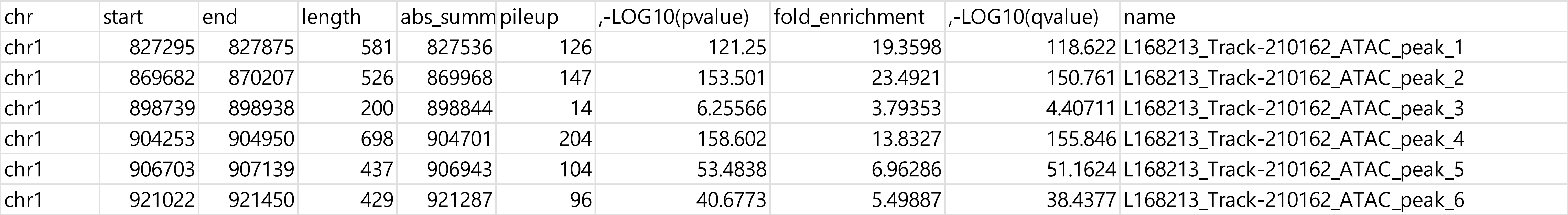


To identify acccessible regions in the genome we need to **call peaks on the nucleosome-free BAM file obtained post-filtering**. Currently, MACS2 is the default peak caller of the ENCODE ATAC-seq pipeline, and so below we provide the recommended parameter changes if using ATAC-seq data as input.

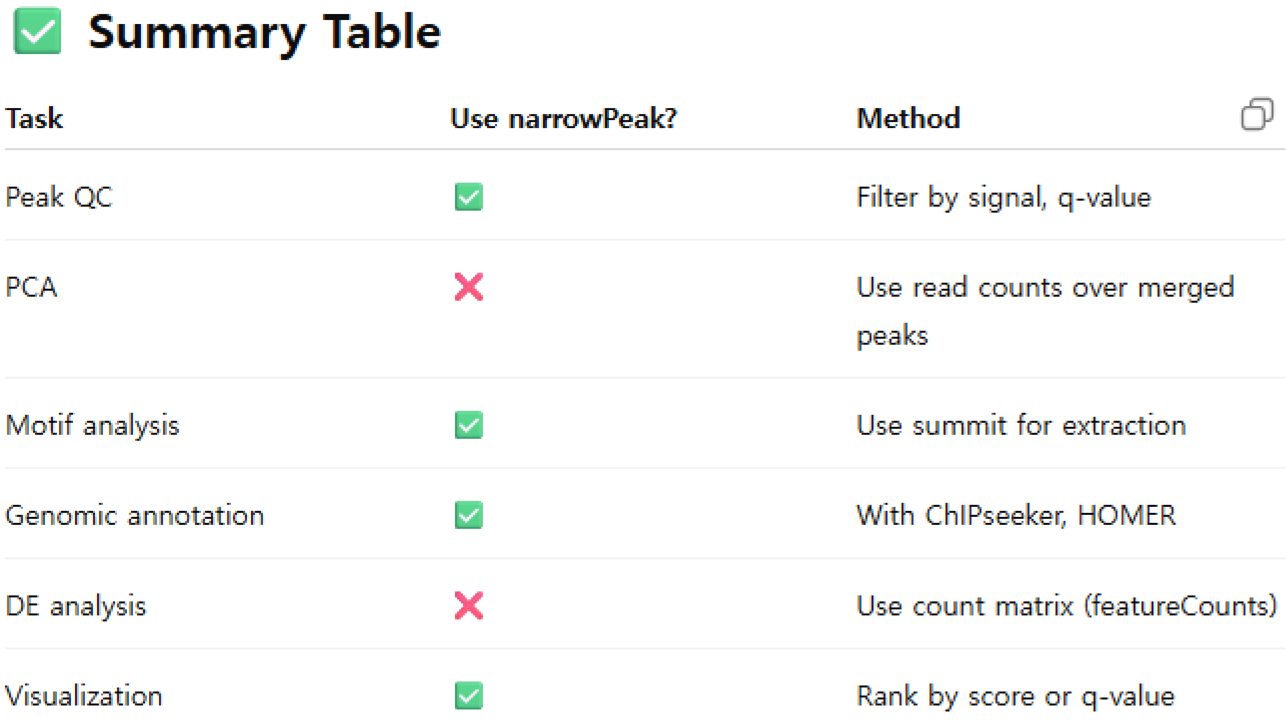
- `-f BAMPE`: Paired-end analysis mode in MACS2.
- `--nomodel`: Bypass building the shifting model. The read pileup does not represent a bimodal pattern, as there is no specific protein-DNA interaction that we are assaying. Open regions will be unimodal in nature, not requiring any shifting of reads.
- `--keep-dup all`: Keep all reads since we have already filtered duplicates from our BAM files.
- `--nolambda`: MACS2 will use the background lambda as local lambda (since we have no input control samples for ATAC-seq)

- # Peak calling

Peakcall → narrowPeak → not for the DAG but annotation, etc



chr	start	end	length	abs_summ	pileup	,-LOG10(pvalue)	fold_enrichment	,-LOG10(qvalue)	name
chr1	827295	827875	581	827536	126	121.25	19.3598	118.622	L168213_Track-210162_ATAC_peak_1
chr1	869682	870207	526	869968	147	153.501	23.4921	150.761	L168213_Track-210162_ATAC_peak_2
chr1	898739	898938	200	898844	14	6.25566	3.79353	4.40711	L168213_Track-210162_ATAC_peak_3
chr1	904253	904950	698	904701	204	158.602	13.8327	155.846	L168213_Track-210162_ATAC_peak_4
chr1	906703	907139	437	906943	104	53.4838	6.96286	51.1624	L168213_Track-210162_ATAC_peak_5
chr1	921022	921450	429	921287	96	40.6773	5.49887	38.4377	L168213_Track-210162_ATAC_peak_6

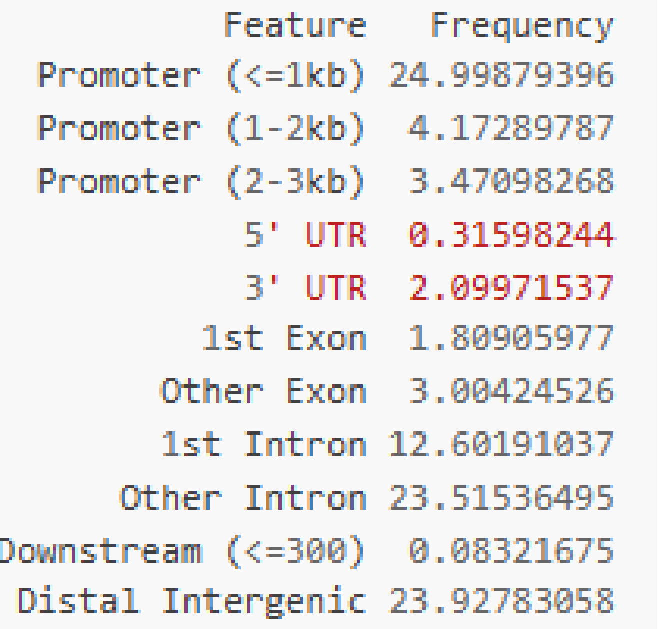


✅ Summary Table

Task	Use narrowPeak?	Method
Peak QC	✅	Filter by signal, q-value
PCA	❌	Use read counts over merged peaks
Motif analysis	✅	Use summit for extraction
Genomic annotation	✅	With ChIPseeker, HOMER
DE analysis	❌	Use count matrix (featureCounts)
Visualization	✅	Rank by score or q-value

- Preprocessing

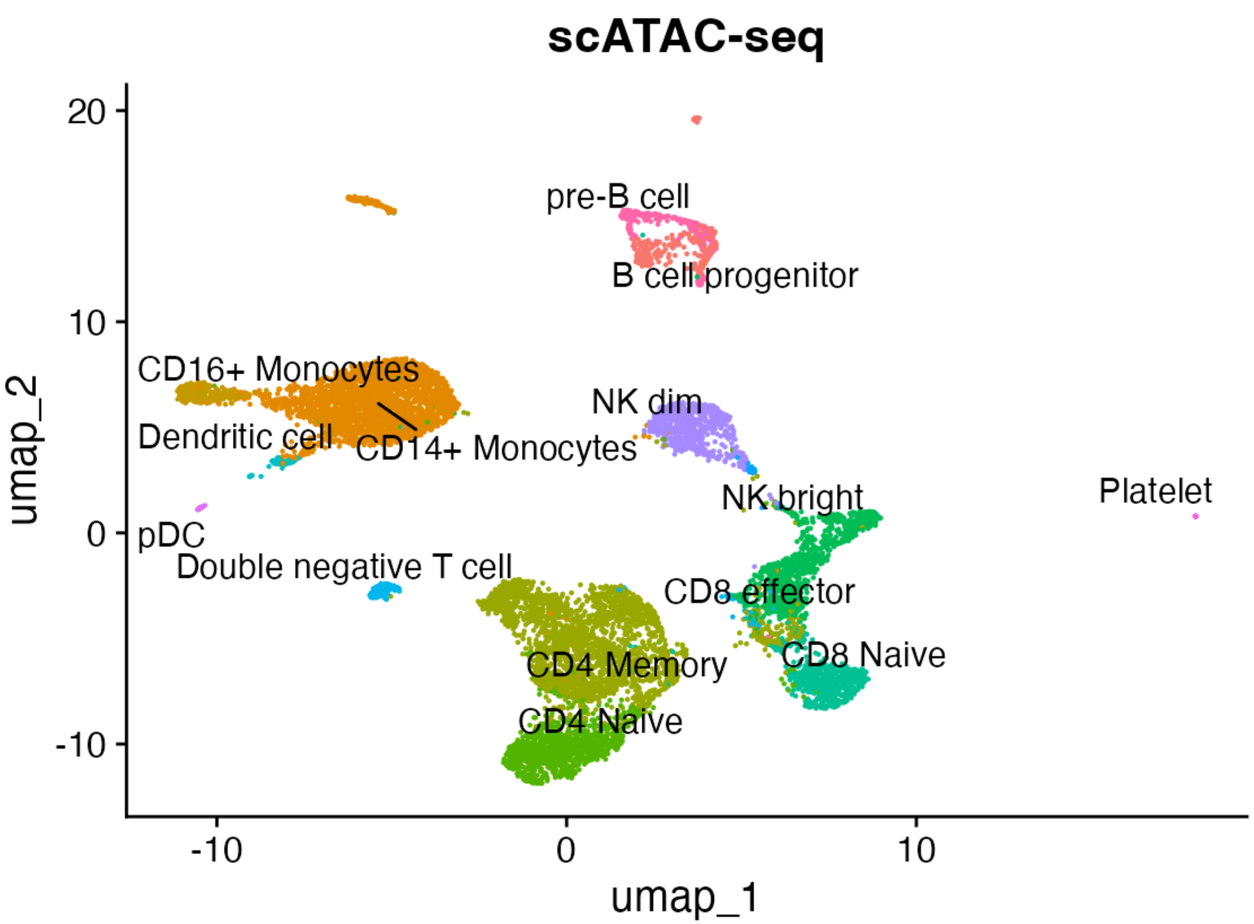
ChIPseeker → Annotate peaks

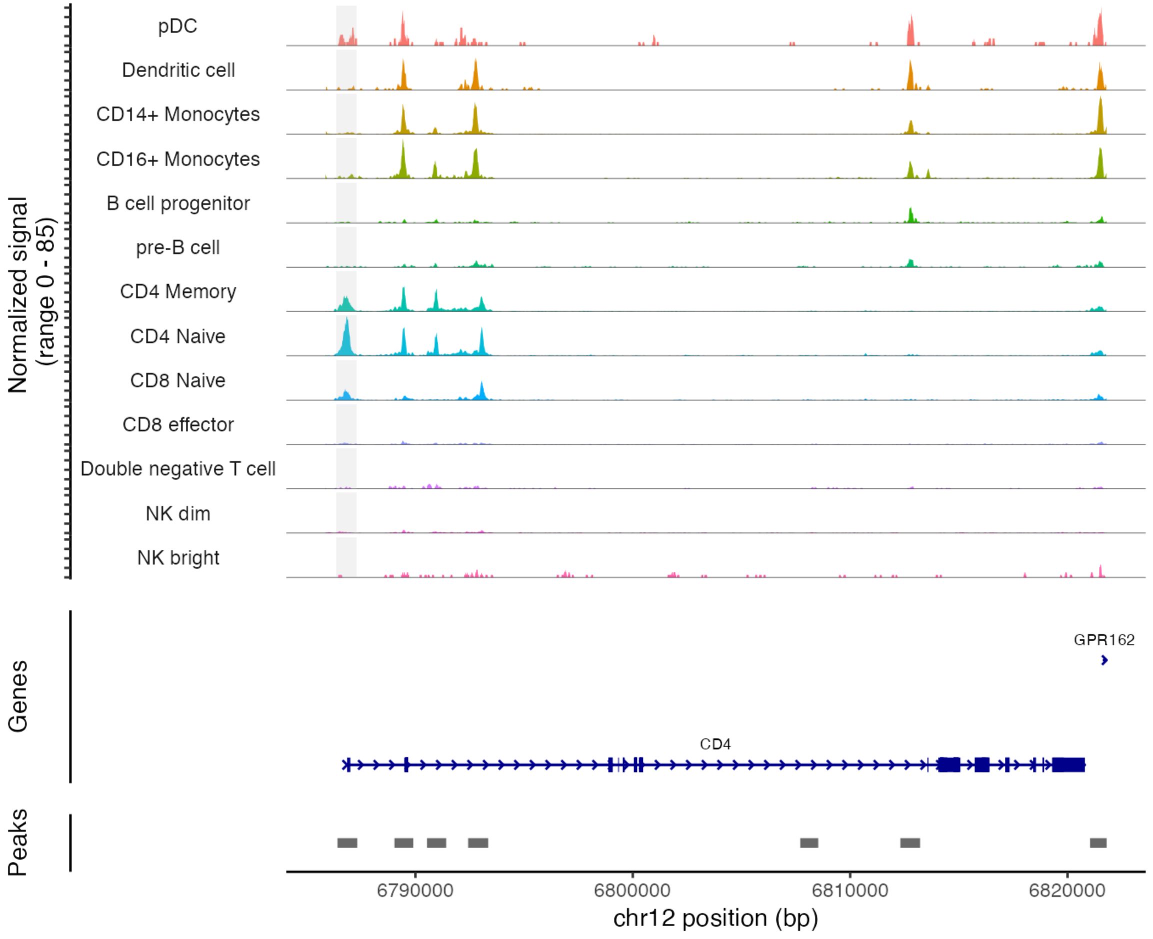


```
           Feature   Frequency
 Promoter (<=1kb) 24.99879396
 Promoter (1-2kb)  4.17289787
 Promoter (2-3kb)  3.47098268
           5' UTR  0.31598244
           3' UTR  2.09971537
         1st Exon  1.80905977
       Other Exon  3.00424526
       1st Intron 12.60191037
     Other Intron 23.51536495
Downstream (<=300)  0.08321675
Distal Intergenic 23.92783058
```

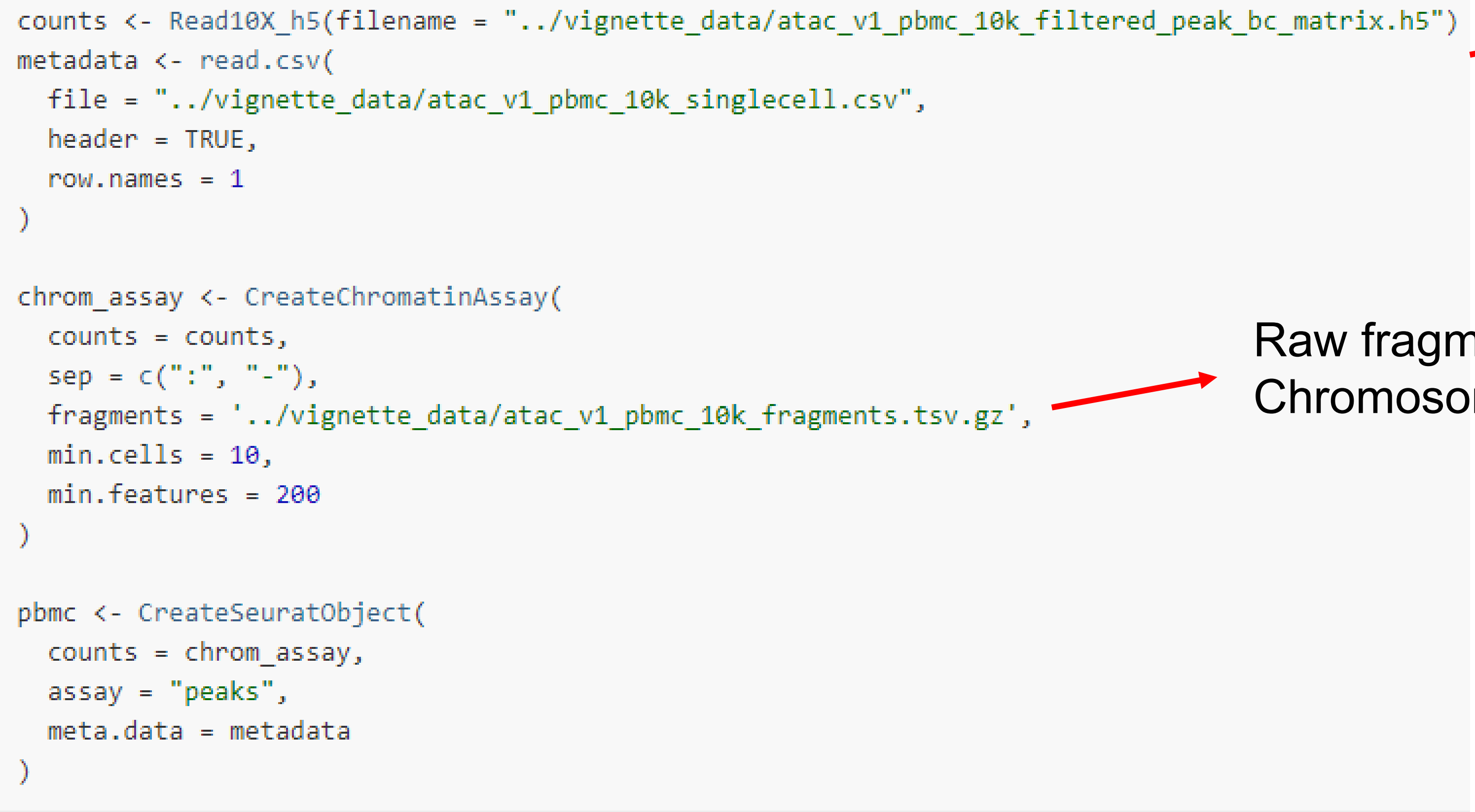
Enhance region: FANTOM5, ENCODE …

- Single-cell ATAC





- Single-cell ATAC



```r
counts <- Read10X_h5(filename = "../vignette_data/atac_v1_pbmc_10k_filtered_peak_bc_matrix.h5")
metadata <- read.csv(
  file = "../vignette_data/atac_v1_pbmc_10k_singlecell.csv",
  header = TRUE,
  row.names = 1
)

chrom_assay <- CreateChromatinAssay(
  counts = counts,
  sep = c(":", "-"),
  fragments = '../vignette_data/atac_v1_pbmc_10k_fragments.tsv.gz',
  min.cells = 10,
  min.features = 200
)

pbmc <- CreateSeuratObject(
  counts = chrom_assay,
  assay = "peaks",
  meta.data = metadata
)
```

Count matrix

Raw fragment file (peak information)
Chromosome, position, cell barcode

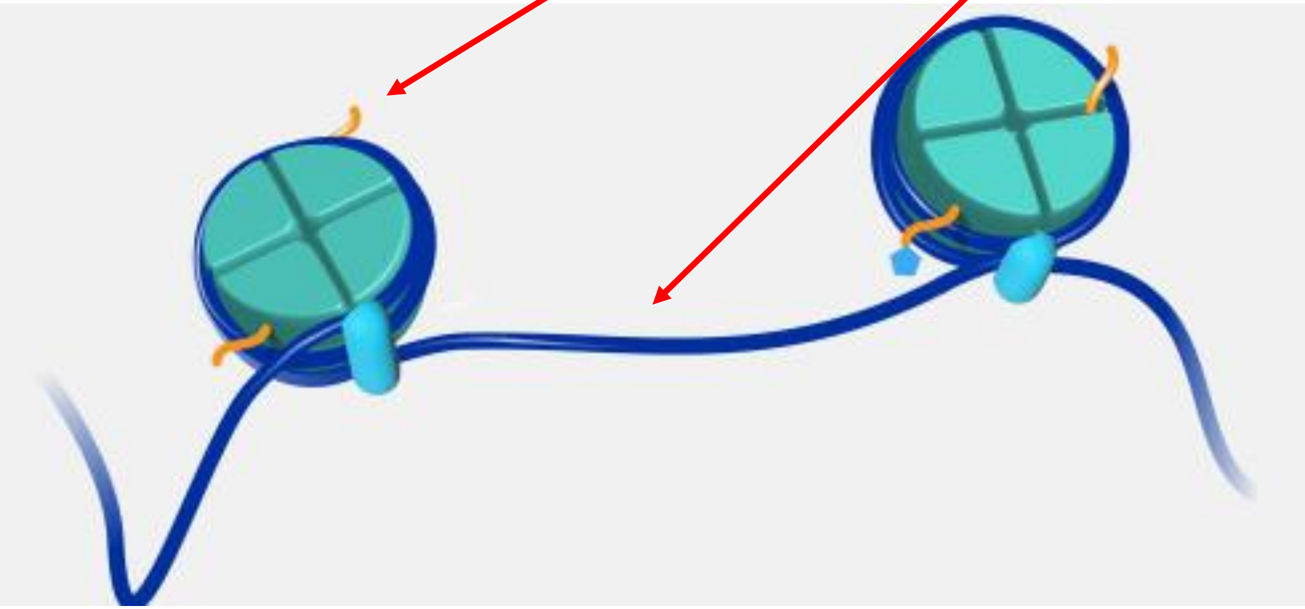
- Peak calling

MACS software → align reads into k-mer bin or known peak region

- QC

-Nucleosome banding pattern: histogram of DNA fragment sizes → should be similar to the length of DNA wrapped around a single nucleosome (147~294 bp)

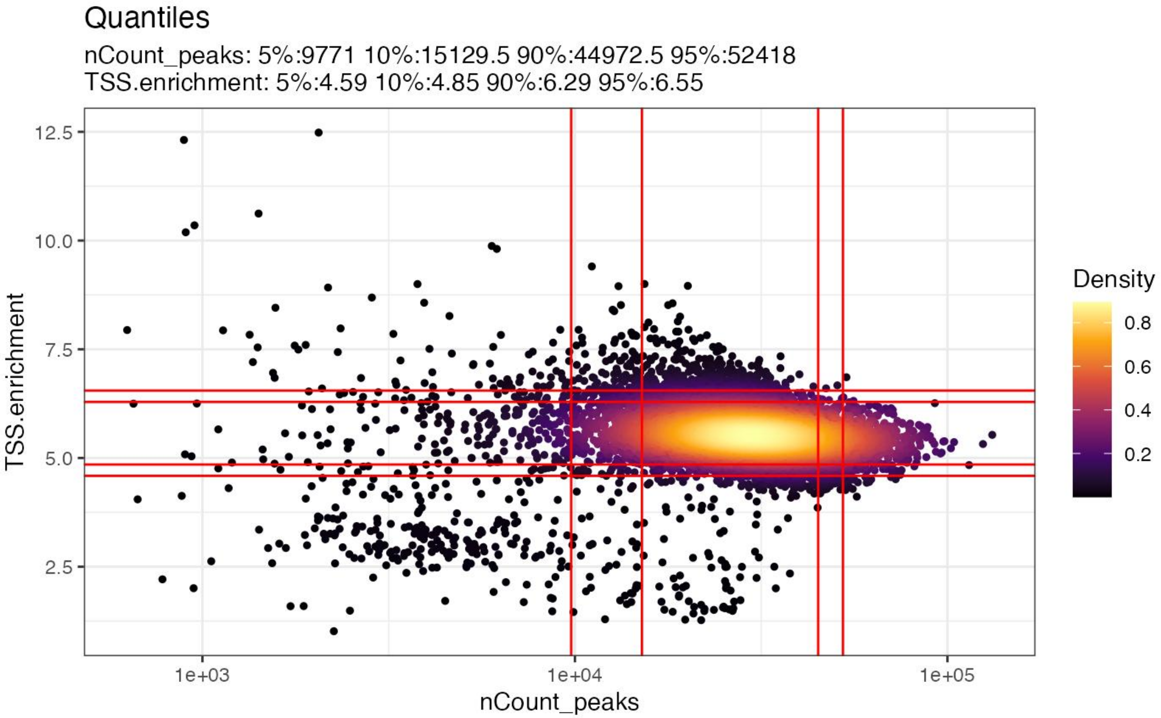
Ratio of mononucleosomal to nucleosome-free (< 147 bp) → mononucleosomal / nuc-free
Good quality: Ratio < 4



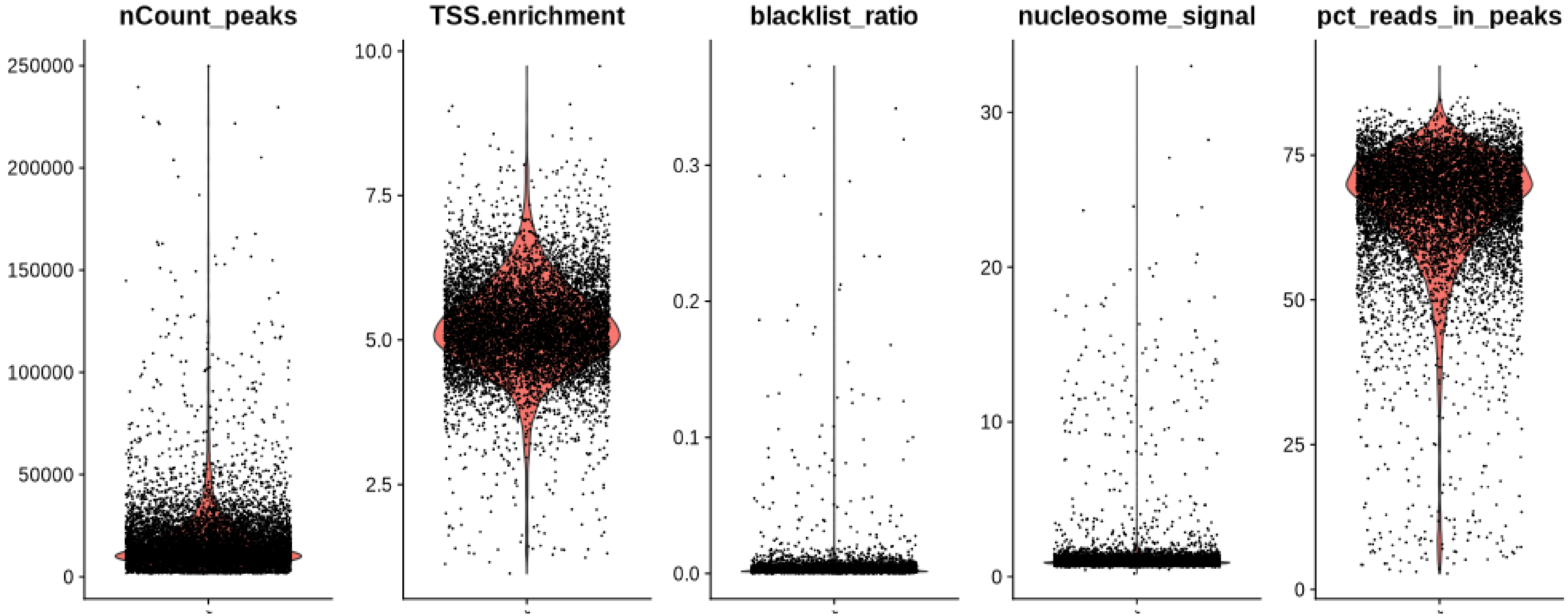
- QC

-TSS (Transcription start site): high enrichment
→ Usually, TSS is opened → reflect experimental sensitivity
→ Mean number of Tn5 insertion event +- 500bp (of TSS) / TSS flanking region (+900~+1000 & -900~-1000)

-Total number of fragments in peaks: too high → doublet
-Fragments in peak fraction: remove <15~20 % → low quality cell
-Black list: ENCODE [experimental artefact prone region] or house-keeping gene



- QC



```
pbmc <- subset(
  x = pbmc,
  subset = nCount_peaks > 3000 &
    nCount_peaks < 30000 &
    pct_reads_in_peaks > 15 &
    blacklist_ratio < 0.05 &
    nucleosome_signal < 4 &
    TSS.enrichment > 3
)
```

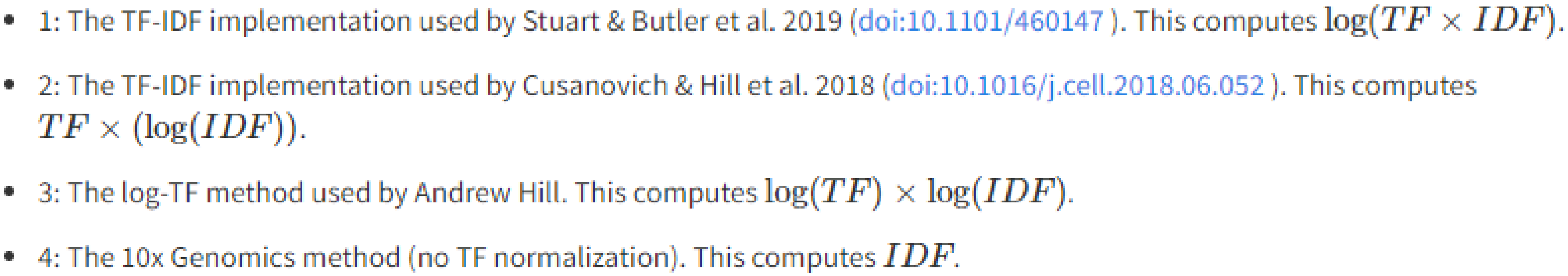
• Normalization

-High sparsity & 0/1 binary data structure for the read
TF-IDF: term frequency-inverse document frequency: seq depth norm across cell + across peak (more weight for rarer peaks)

TF = Cij/Fj where Cij is the total number of counts for peak i in cell j and Fj is the total number of counts for cell j. (~ read depth norm)

IDF = log(1 + N/ni) where N is the total number of cells in the dataset and ni is the total number counts for peak i across all cells. (given peak rareness)

Signac: log(1 + (TF × IDF) × 10^4)
→ nonzero, mean not close to zero, variable across celltype



- 1: The TF-IDF implementation used by Stuart & Butler et al. 2019 (doi:10.1101/460147). This computes $\log(TF \times IDF)$.
- 2: The TF-IDF implementation used by Cusanovich & Hill et al. 2018 (doi:10.1016/j.cell.2018.06.052). This computes $TF \times (\log(IDF))$.
- 3: The log-TF method used by Andrew Hill. This computes $\log(TF) \times \log(IDF)$.
- 4: The 10x Genomics method (no TF normalization). This computes IDF.

A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility

- Feature selection & Dimension reduction

-FeatureSelection (FindTopFeatures): Variable feature + common feature across cells (q5, 95 %)

-Dimension reduction: SVD
→ Skip 1st dimension: LSI_1 highly correlated with seq depth

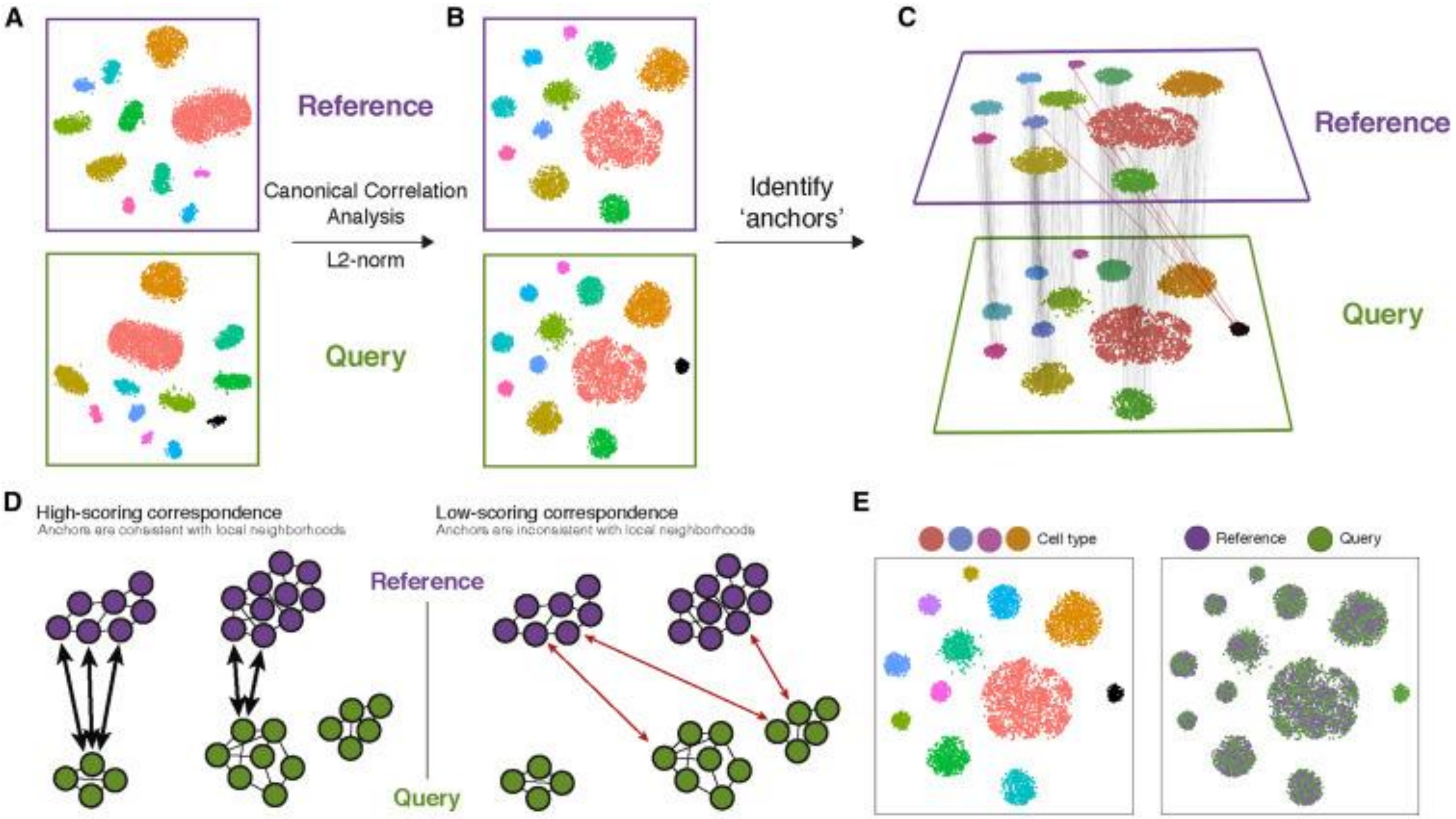
- Characteristic of ATAC

-Only 2 copies from DNA → higher drop-out → higher sparsity
-Require binning analysis rather than single-cell analysis
-ATAC-seq → detect more regions than transcriptome → higher complexity
→ delicate gene expression regulation

Promoter region (or Transcriptional start site) + Enhancer regions
→ Enhancer regions provide delicate and complex regulation for gene expression

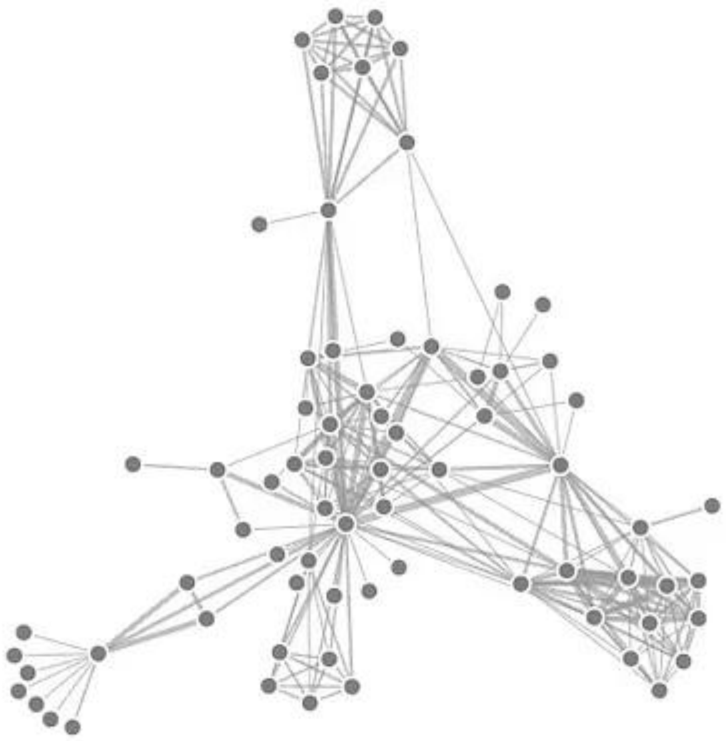
- Adapt from scRNA-seq

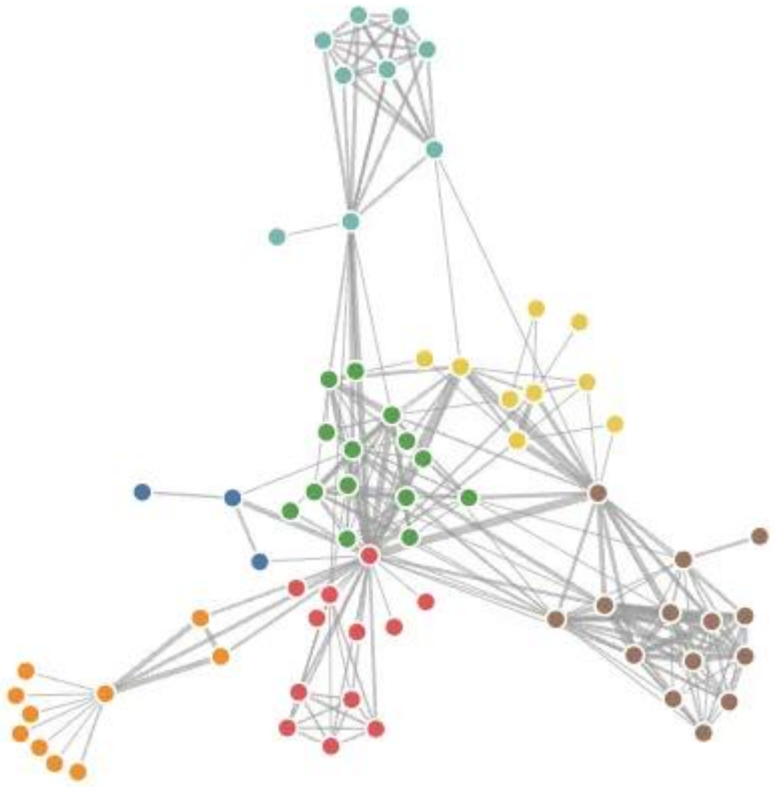
- Batch correction (Seurat)


A
Reference
Query
Canonical Correlation Analysis
L2-norm
B
Identify 'anchors'
C
Reference
Query
D
High-scoring correspondence
Anchors are consistent with local neighborhoods
Reference
Query
Low-scoring correspondence
Anchors are inconsistent with local neighborhoods
E
Cell type
Reference
Query

- # Clustering

- SLM clustering: based on SNN (shared-nearest neighbor) graph → modularity optimization





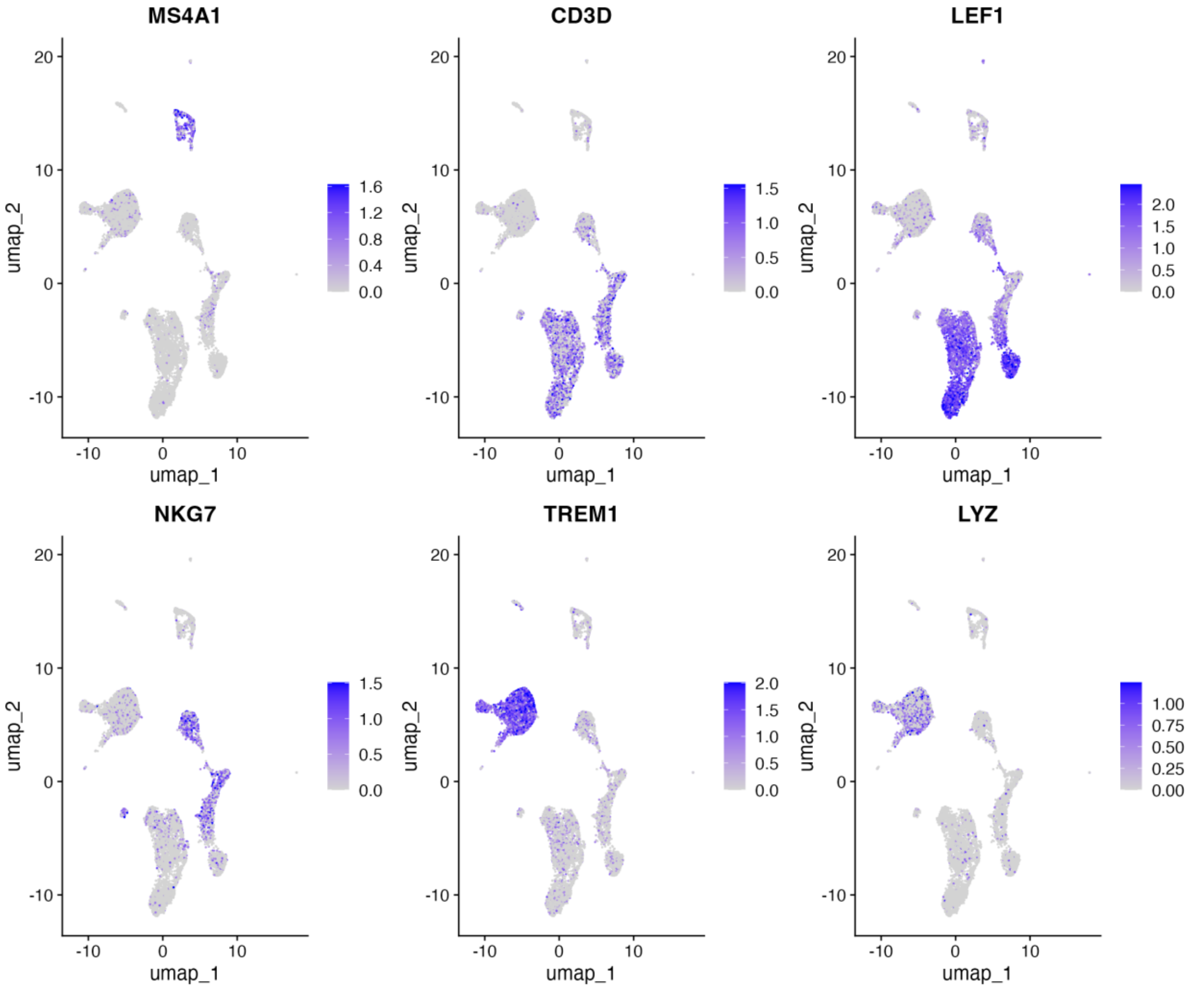
- Gene activity

-Open region → gene expression

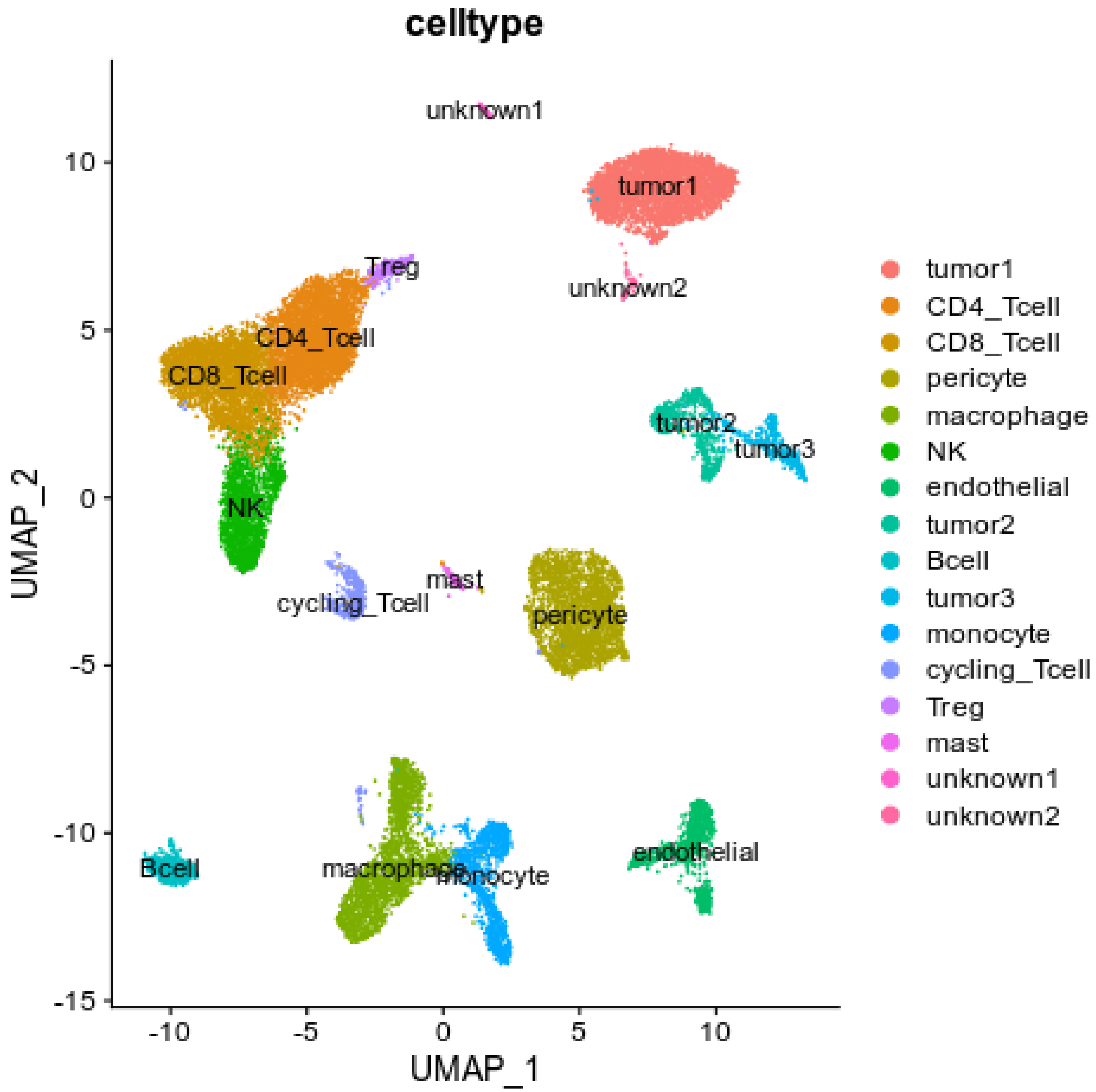
(inference: it is not always that open regions correlate with gene expression)

-upstream of TSS: 2000bp + genebody: compute counts per cell

-Normalization: logNorm, scale factor: median counts

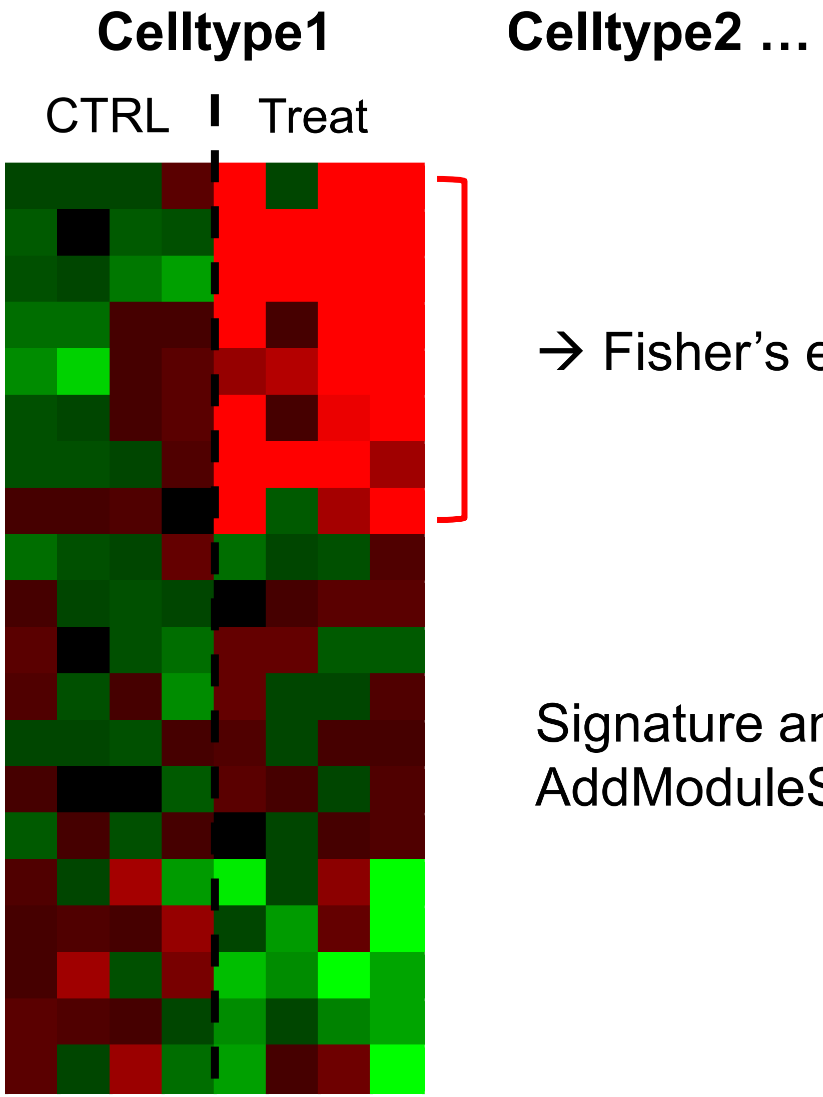


- Celltype annotation



-Based on gene activity

- Geneset analysis

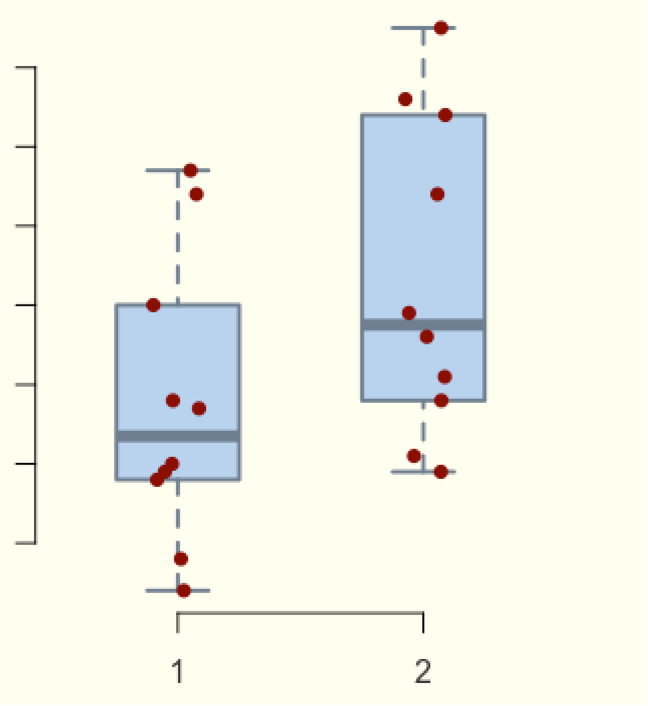


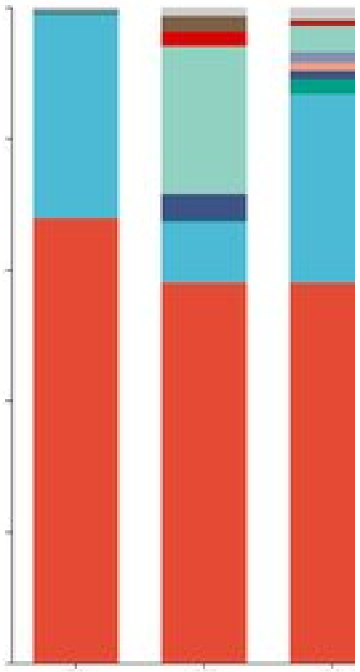
→ Fisher’s exact test or GSEA

Signature analysis
AddModuleScore, AUCell, ssGSEA, cNMF

- Cell abundance

*T-test, Wilcoxon

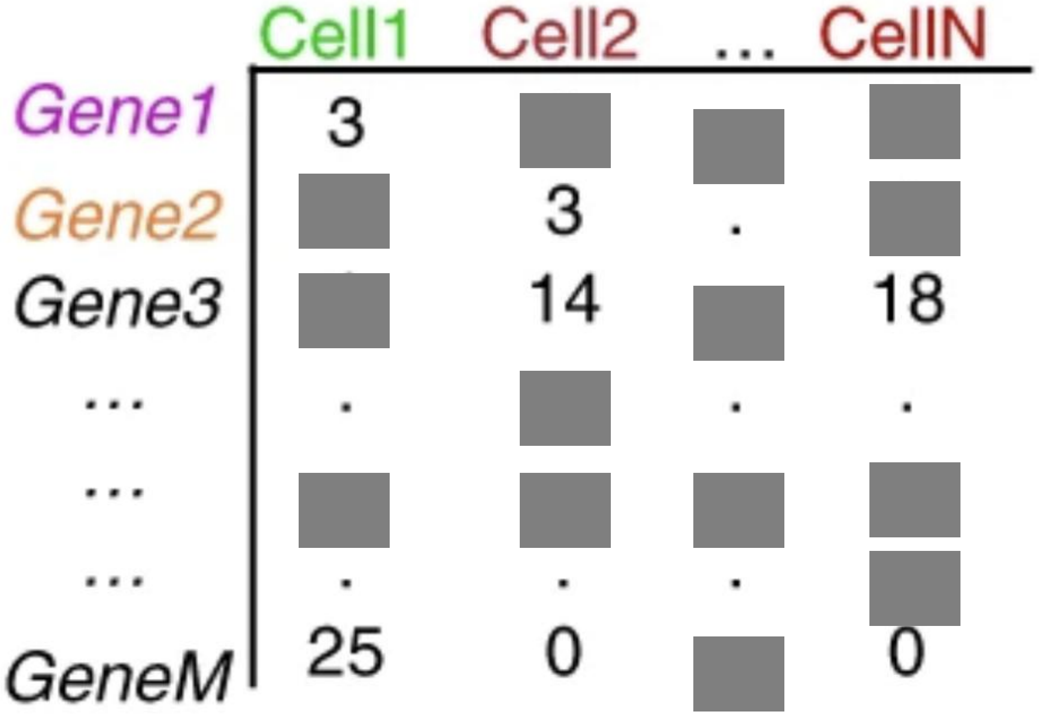




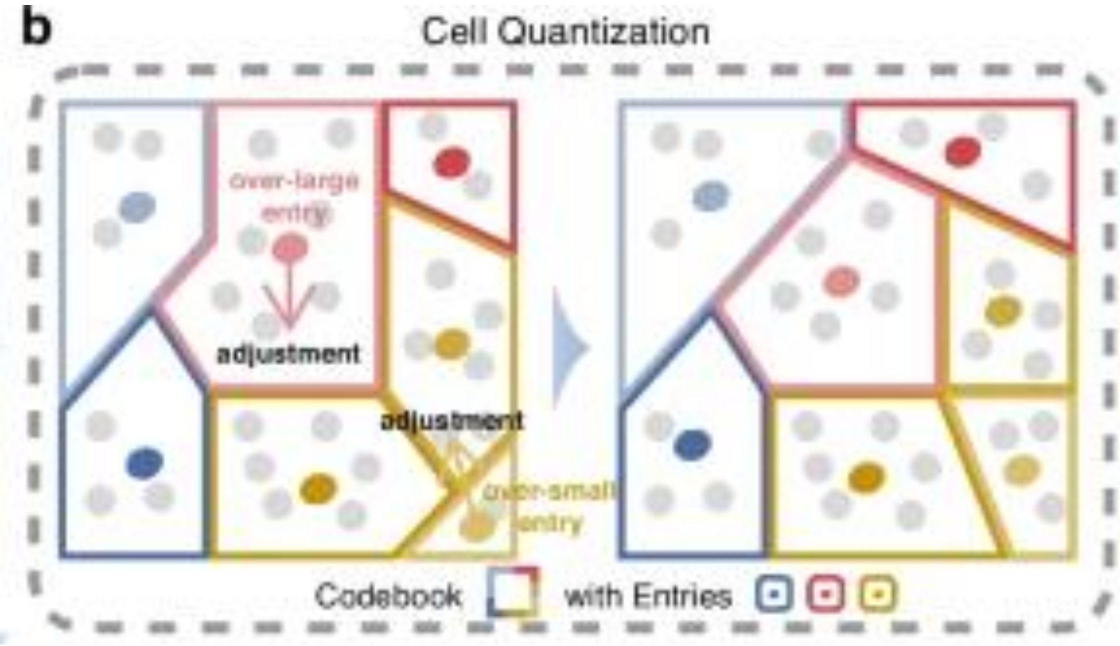
- Higher sparsity

Be cautious during running those analysis
→ Likely to mislead the result

• Cell-pooling



	Cell1	Cell2	…	CellN
Gene1	3	■	■	■
Gene2	■	3	.	■
Gene3	■	14	■	18
…	.	■	.	.
…	■	■	■	■
…	.	.	.	■
GeneM	25	0	■	0

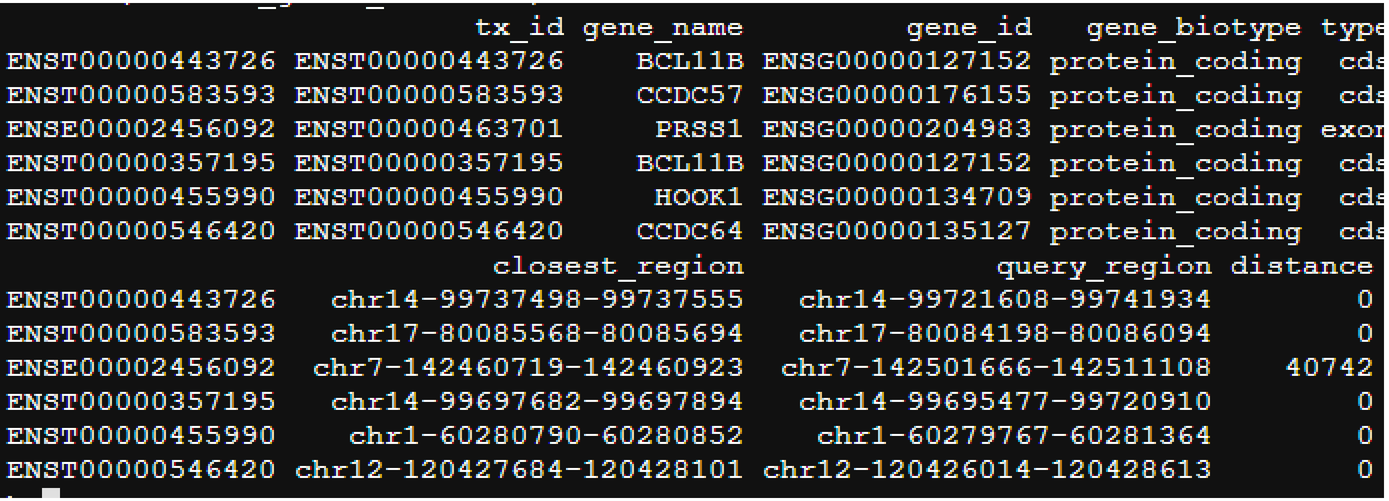


-High drop-out rate: zero count ↑
-merge cells → pseudo cell → averaging → overcome drop-out!

- Differential accessible peak analysis

-Based on FindMarkers
differential testing is to utilize **logistic regression** for, as suggested by Ntranos et al. 2018 for scRNA-seq data, and add the **total number of fragments as a latent variable**
→ Read depth adjustment

ClosestFeature: closest gene from peak



```
                           tx_id gene_name         gene_id   gene_biotype type
ENST00000443726 ENST00000443726    BCL11B ENSG00000127152 protein_coding  cds
ENST00000583593 ENST00000583593    CCDC57 ENSG00000176155 protein_coding  cds
ENSE00002456092 ENST00000463701     PRSS1 ENSG00000204983 protein_coding exon
ENST00000357195 ENST00000357195    BCL11B ENSG00000127152 protein_coding  cds
ENST00000455990 ENST00000455990     HOOK1 ENSG00000134709 protein_coding  cds
ENST00000546420 ENST00000546420    CCDC64 ENSG00000135127 protein_coding  cds
                         closest_region                query_region distance
ENST00000443726   chr14-99737498-99737555   chr14-99721608-99741934        0
ENST00000583593   chr17-80085568-80085694   chr17-80084198-80086094        0
ENSE00002456092  chr7-142460719-142460923  chr7-142501666-142511108    40742
ENST00000357195   chr14-99697682-99697894   chr14-99695477-99720910        0
ENST00000455990    chr1-60280790-60280852    chr1-60279767-60281364        0
ENST00000546420 chr12-120427684-120428101 chr12-120426014-120428613        0
```

- Peak analysis procedure

-Peak → which motif→ motif enrichment test (using JASPAR DB)
Using ChromVar (which TF motif is enriched)

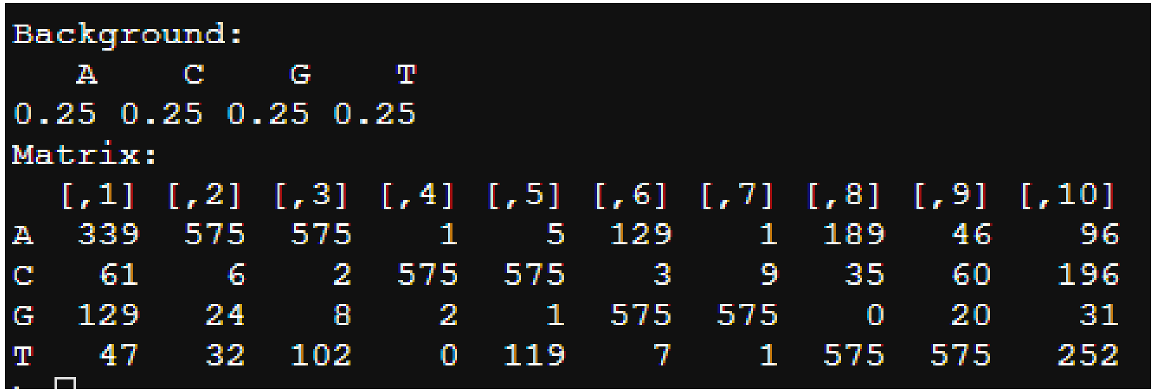
-FIMO: which motif will be used for a given TF or DNA binding protein from a given genome (open region)

-Annotation of a given region: GREAT

- Motif analysis

```
# Get a list of motif position frequency matrices from the JASPAR database
pfm <- getMatrixSet(
  x = JASPAR2020,
  opts = list(collection = "CORE", tax_group = 'vertebrates', all_versions = FALSE)
)

# Add motif information
mouse_brain <- AddMotifs(
  object = mouse_brain,
  genome = BSgenome.Mmusculus.UCSC.mm10,
  pfm = pfm
)
```

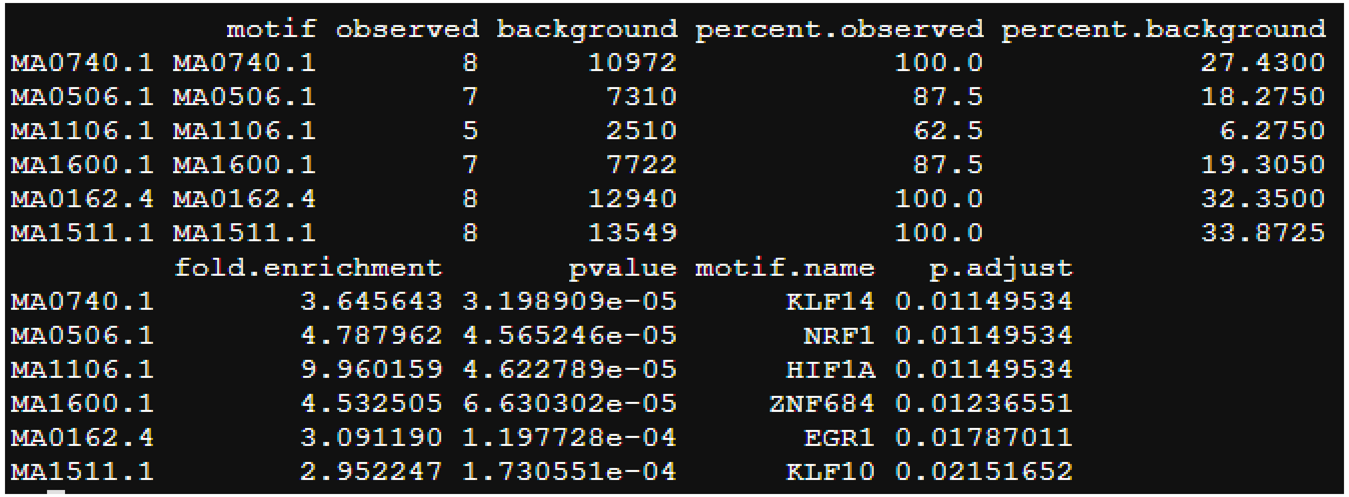


```
Background:
   A    C    G    T
0.25 0.25 0.25 0.25
Matrix:
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
A  339  575  575    1    5  129    1  189   46    96
C   61    6    2  575  575    3    9   35   60   196
G  129   24    8    2    1  575  575    0   20    31
T   47   32  102    0  119    7    1  575  575   252
```

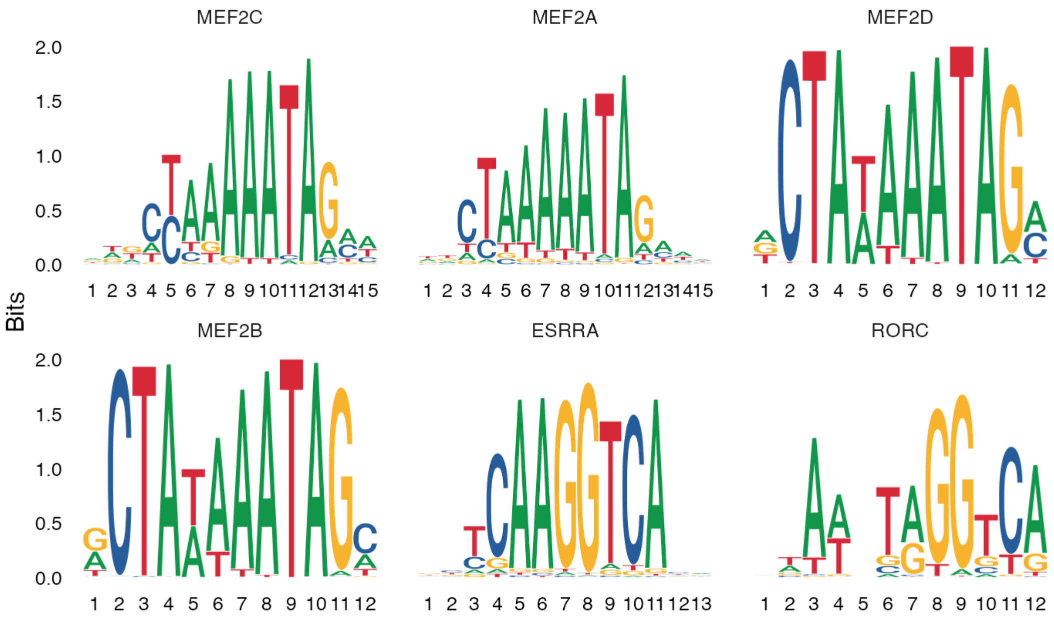
Motif analysis

```
enriched.motifs <- FindMotifs(
  object = pbmc,
  features = peak_list
)
```

→ Find enriched motif from a given peak (by hypergeometric test)



```
            motif observed background percent.observed percent.background
MA0740.1 MA0740.1        8      10972            100.0            27.4300
MA0506.1 MA0506.1        7       7310             87.5            18.2750
MA1106.1 MA1106.1        5       2510             62.5             6.2750
MA1600.1 MA1600.1        7       7722             87.5            19.3050
MA0162.4 MA0162.4        8      12940            100.0            32.3500
MA1511.1 MA1511.1        8      13549            100.0            33.8725
         fold.enrichment       pvalue motif.name   p.adjust
MA0740.1        3.645643 3.198909e-05      KLF14 0.01149534
MA0506.1        4.787962 4.565246e-05       NRF1 0.01149534
MA1106.1        9.960159 4.622789e-05      HIF1A 0.01149534
MA1600.1        4.532505 6.630302e-05     ZNF684 0.01236551
MA0162.4        3.091190 1.197728e-04       EGR1 0.01787011
MA1511.1        2.952247 1.730551e-04      KLF10 0.02151652
```



Open region → enriched Motifs + TF
(or DNA-binding protein)

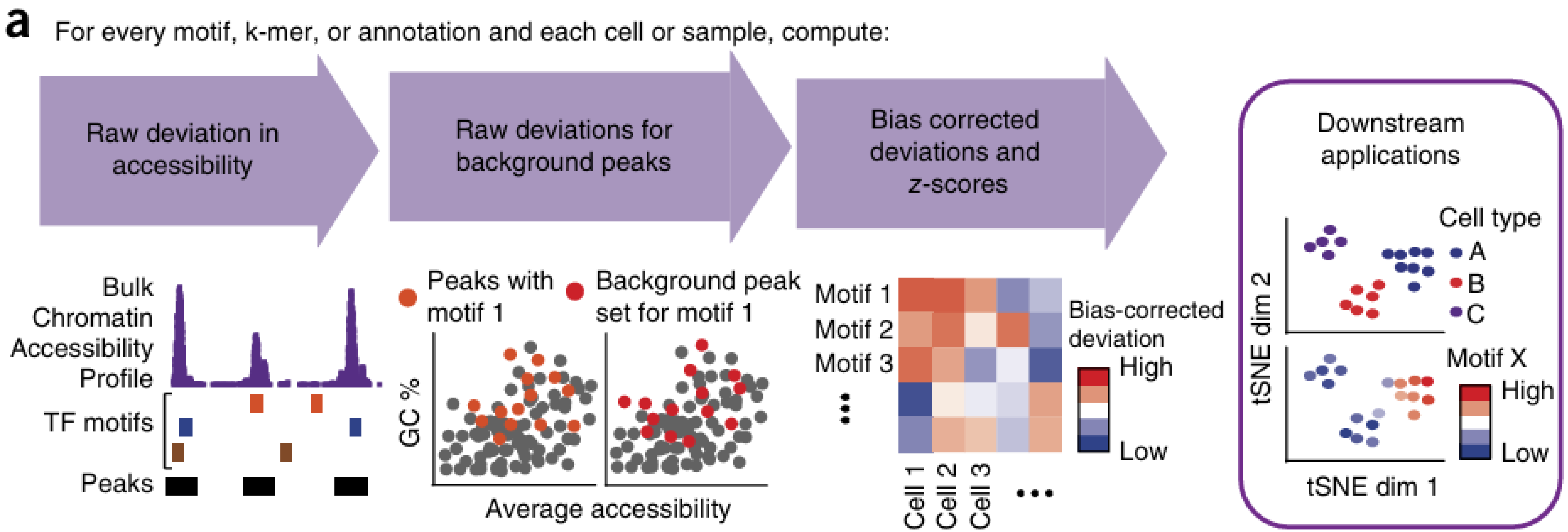
- # Motif analysis

FindMotifs
-Should match for overall GC, accessibility, peak width

1: Bias in PCR amplification from GC-rich region
2: Variable Tn5 tagmentation
3: Accessiblity bias: more reads due to “open region” → Does not mean genome has more motif
4: Peak width bias: Longer peak → more Motif (similar to gene length normalization)

• Motif analysis

-ChromeVar: which TF motif is enriched
→ background corrected peak (motif) signal

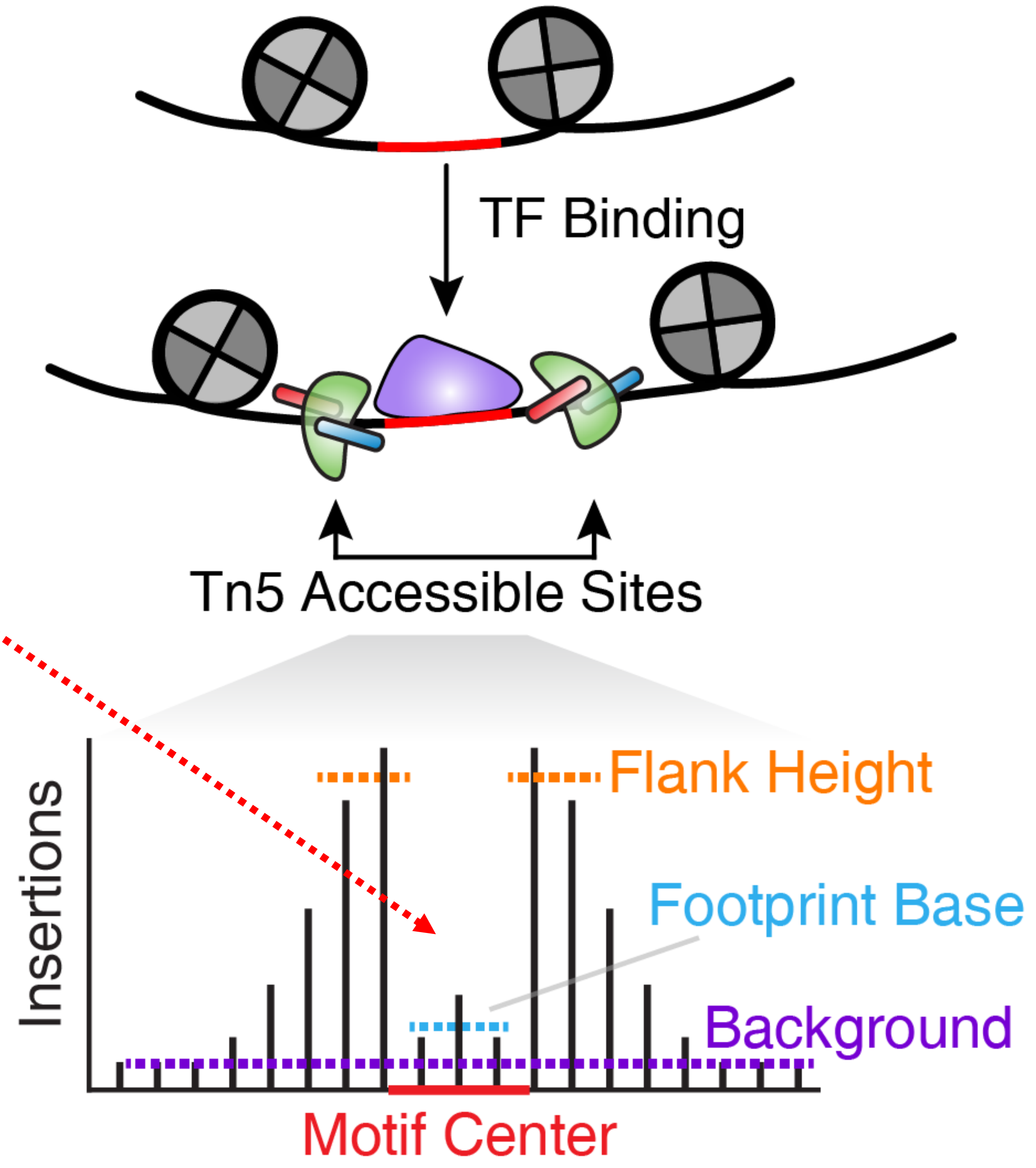

a
For every motif, k-mer, or annotation and each cell or sample, compute:
Raw deviation in accessibility
Raw deviations for background peaks
Bias corrected deviations and z-scores
Downstream applications
Bulk Chromatin Accessibility Profile
TF motifs
Peaks
Peaks with motif 1
Background peak set for motif 1
GC %
Average accessibility
Motif 1
Motif 2
Motif 3
Cell 1
Cell 2
Cell 3
Bias-corrected deviation
High
Low
Cell type
A
B
C
Motif X
High
Low
tSNE dim 2
tSNE dim 1

```
# Motif activity
mouse_brain <- RunChromVAR(
  object = mouse_brain,
  genome = BSgenome.Mmusculus.UCSC.mm10
)
```

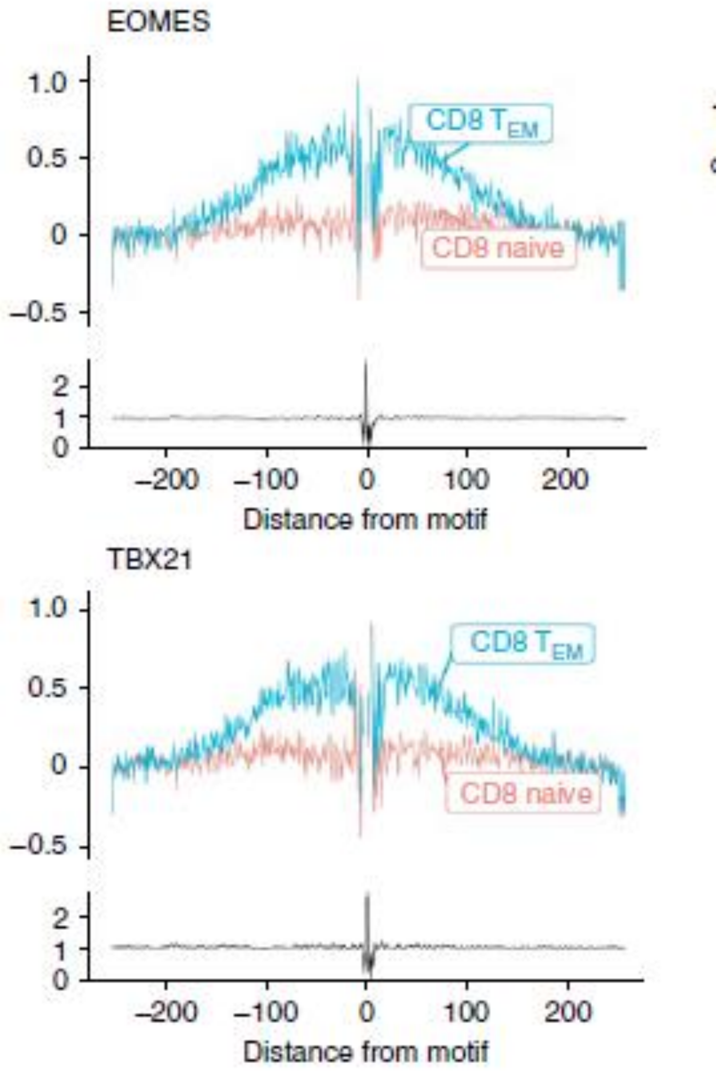
- Motif analysis

-TF footprinting

```
pbmc <- Footprint(
  object = pbmc,
  motif.name = c("GATA3", "TBX21"),
  genome = BSgenome.Hsapiens.UCSC.hg19
)
```



Motif analysis



The reason for vacancy at the center
→ TF binding → cannot be sequenced

- Motif analysis

FIMO

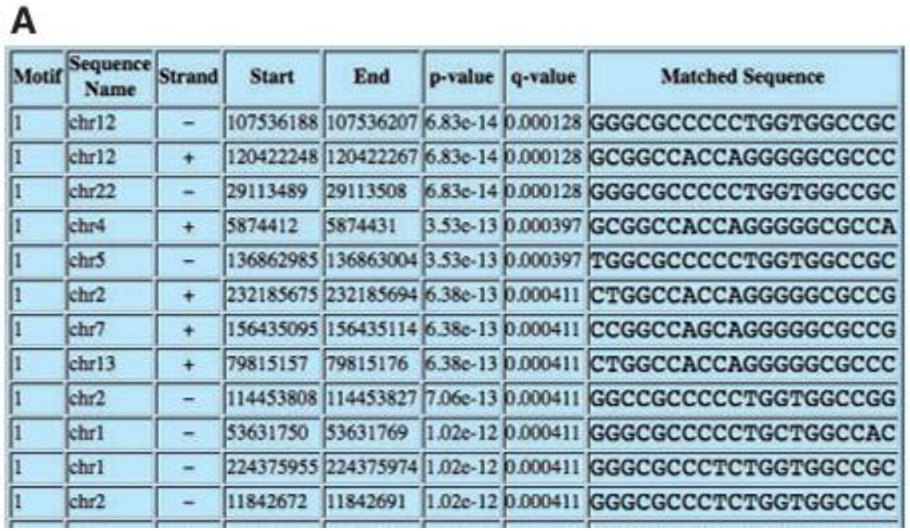
Find individual motif occurrences

→Calculate motif occurrence at the Genome (open chromatin)

-Position-specific freq matrix → log-likelihood ratio

-Pvalue by random seq (user-defined ATGC ratio)

-Bootstrap → FDR



A

Motif	Sequence Name	Strand	Start	End	p-value	q-value	Matched Sequence
1	chr12	–	107536188	107536207	6.83e-14	0.000128	GGGCGCCCCCTGGTGGCCGC
1	chr12	+	120422248	120422267	6.83e-14	0.000128	GCGGCCACCAGGGGGCGCCC
1	chr22	–	29113489	29113508	6.83e-14	0.000128	GGGCGCCCCCTGGTGGCCGC
1	chr4	+	5874412	5874431	3.53e-13	0.000397	GCGGCCACCAGGGGGCGCCA
1	chr5	–	136862985	136863004	3.53e-13	0.000397	TGGCGCCCCCTGGTGGCCGC
1	chr2	+	232185675	232185694	6.38e-13	0.000411	CTGGCCACCAGGGGGCGCCG
1	chr7	+	156435095	156435114	6.38e-13	0.000411	CCGGCCAGCAGGGGGCGCCG
1	chr13	+	79815157	79815176	6.38e-13	0.000411	CTGGCCACCAGGGGGCGCCC
1	chr2	–	114453808	114453827	7.06e-13	0.000411	GGCCGCCCCCTGGTGGCCGG
1	chr1	–	53631750	53631769	1.02e-12	0.000411	GGGCGCCCCCTGCTGGCCAC
1	chr1	–	224375955	224375974	1.02e-12	0.000411	GGGCGCCCTCTGGTGGCCGC
1	chr2	–	11842672	11842691	1.02e-12	0.000411	GGGCGCCCTCTGGTGGCCGC

Ex: CTCF binding site (motif) from a given region

- Motif analysis

-**GREAT**

Genomic regions enrichment of **annotations** tool

→ Annotation enrichment for a given region
(ex: gene ontology)

TSS → -5k, +1k (proximal region)

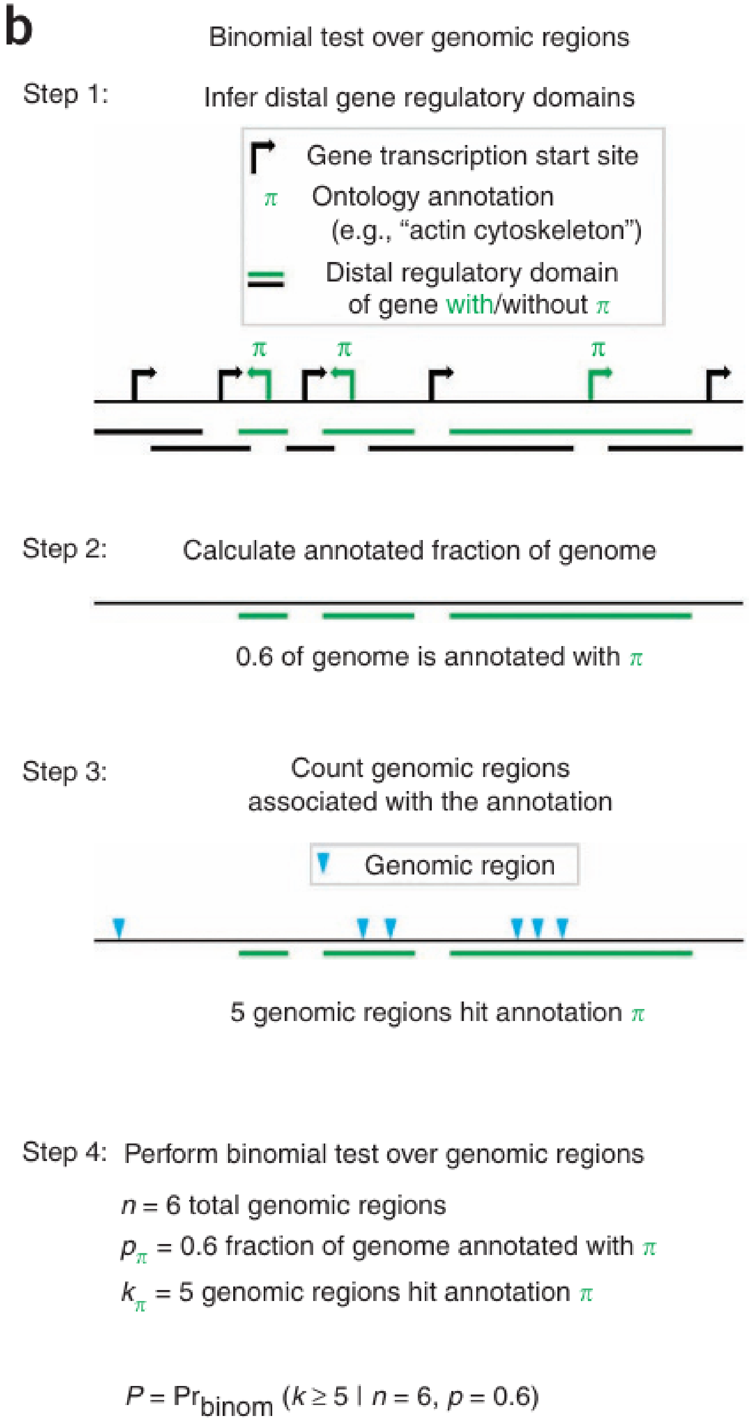
→ +/- 1MB (Distal regulation)

Binomial distribution: B(n,p)

→ Target Annotated region vs (n)

total annotated genomic region / genomic region (p)

→pbinom


b
Binomial test over genomic regions
Step 1:
Infer distal gene regulatory domains
Gene transcription start site
π Ontology annotation
(e.g., "actin cytoskeleton")
Distal regulatory domain
of gene with/without π
π
π
π
Step 2:
Calculate annotated fraction of genome
0.6 of genome is annotated with π
Step 3:
Count genomic regions
associated with the annotation
Genomic region
5 genomic regions hit annotation π
Step 4: Perform binomial test over genomic regions
n = 6 total genomic regions
p_π = 0.6 fraction of genome annotated with π
k_π = 5 genomic regions hit annotation π
$P = \mathrm{Pr}_{\mathrm{binom}}(k \geq 5 \mid n = 6, p = 0.6)$