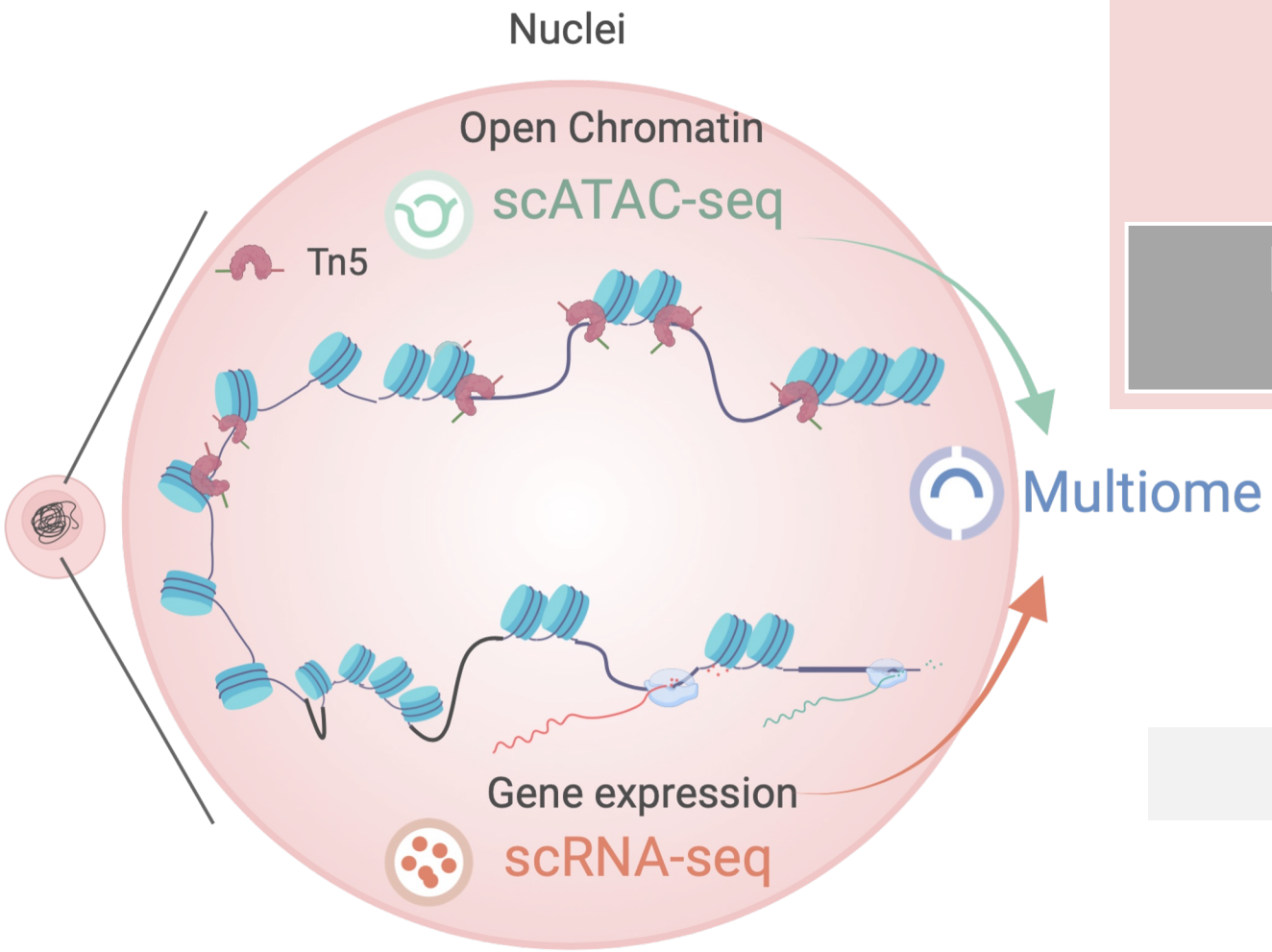


Workshop
Single-cell ATAC-seq

Part II: Clustering, Peak calling and Peak Links analysis,

María Lucía Romero
PhD student in Bioinformatics, Epigenetics Biomedical, IDIBAPS

cnag centre nacional d'anàlisi genòmica centro nacional de análisis genómico
CRG Centre for Genomic Regulation
IDIBAPS Institut D'Investigacions Biomèdiques August Pi i Sunyer

Objective

What are we going to learn?

Overview of last session
Modality integration (WNN)

Clustering
B-cell type selection

Peak Calling
Focused on scATAC-seq

Peak-Gene Linkage analysis
Graphic explanation
Filtering Links by DARs and DEGs

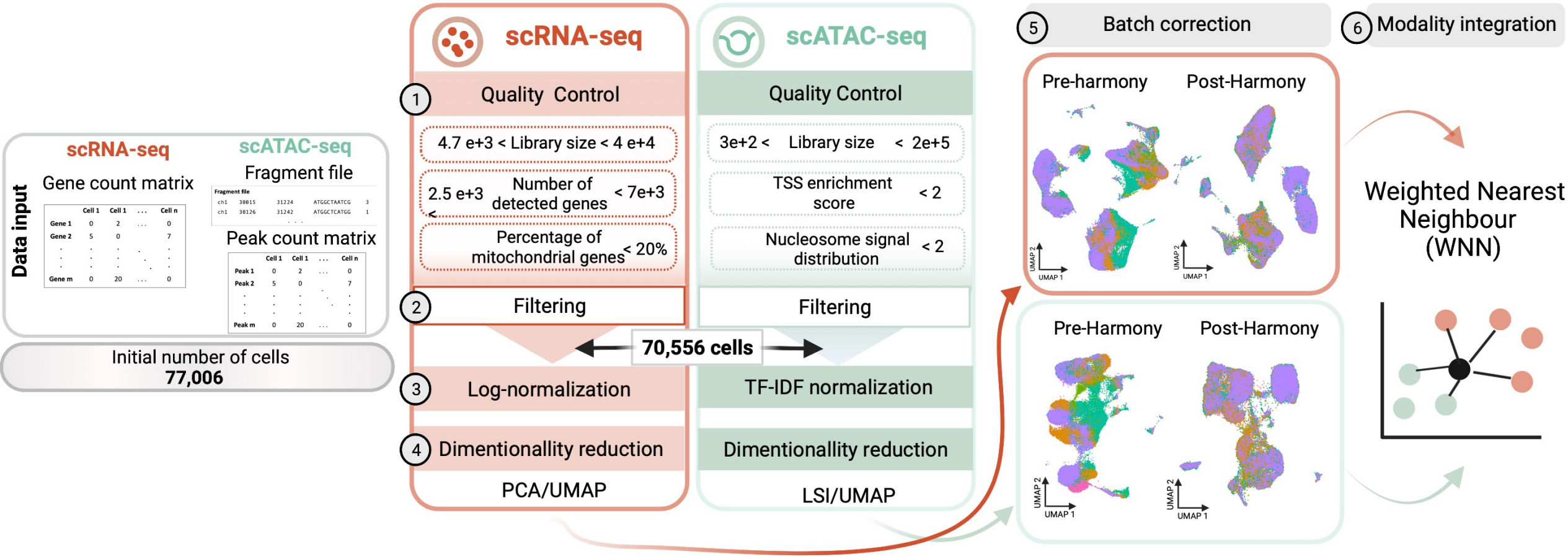
Peak-Gene Links Classification by chromatin states
Localization of BCL6 super-enhancer

0 1 2 4 5

In the last session...

Multiome analysis workflow

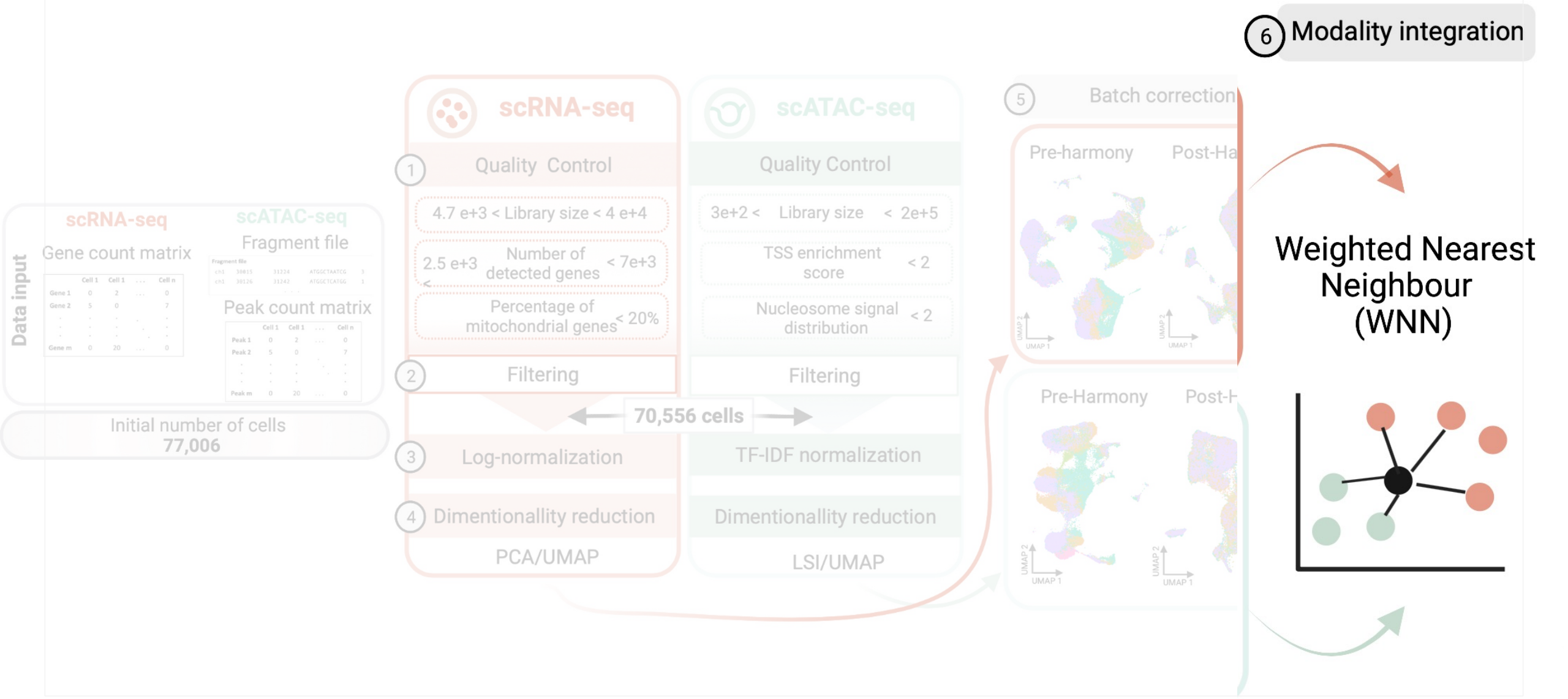
(Tools: Seurat, Signac, Harmony, Scrublet)



Multiome analysis workflow

Modality integration

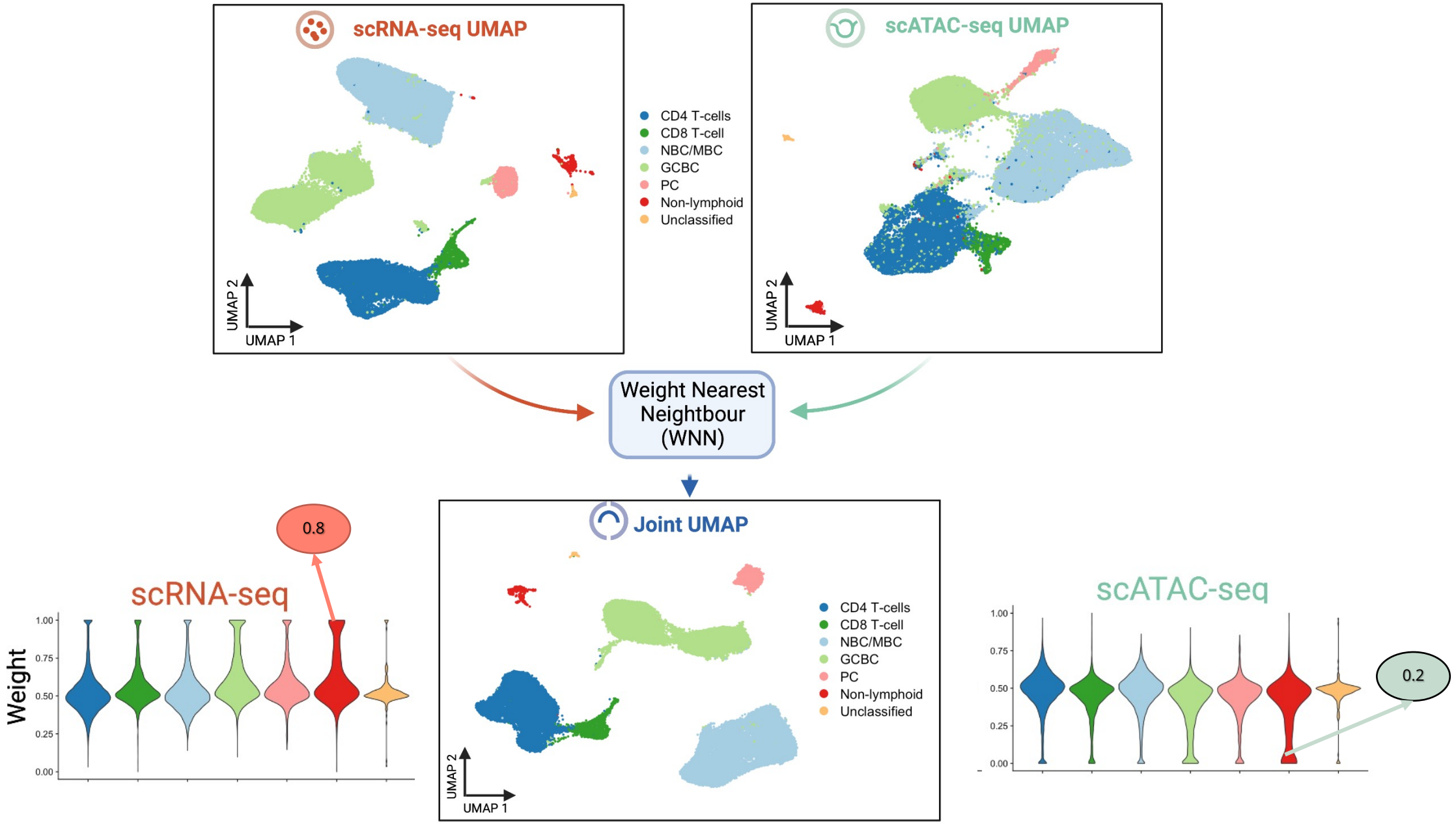
(scRNA-seq + scATAC-seq)



Modality integration

(scRNA-seq + scATAC-seq)

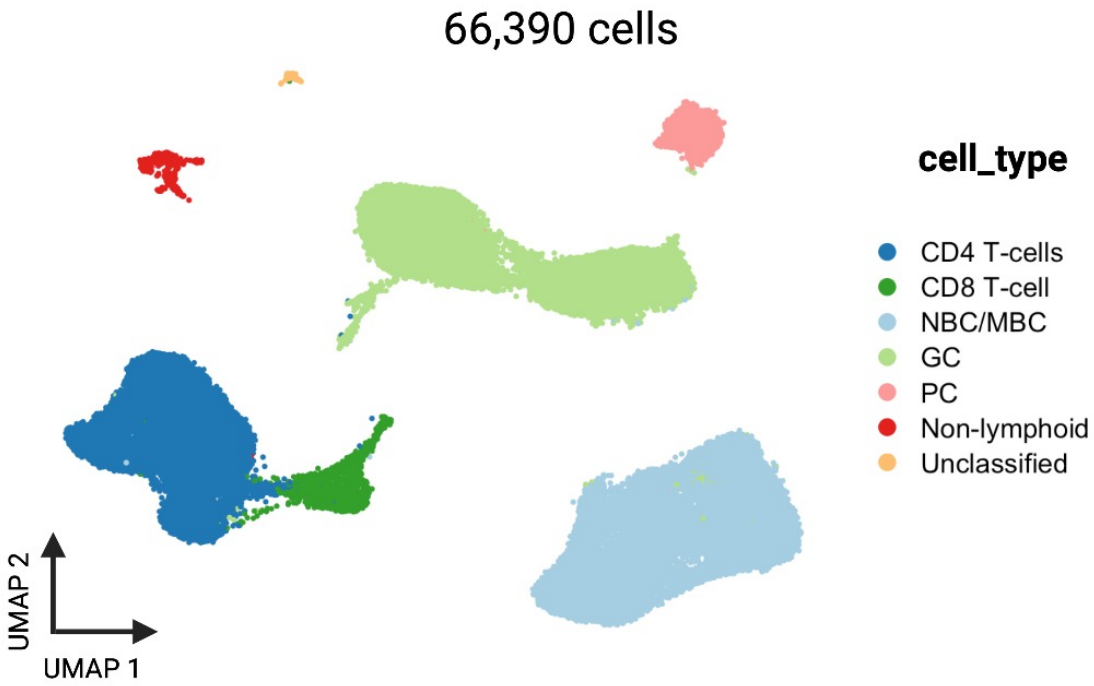
Multiome analysis workflow



Modality integration

(scRNA-seq + scATAC-seq)

Joint UMAP

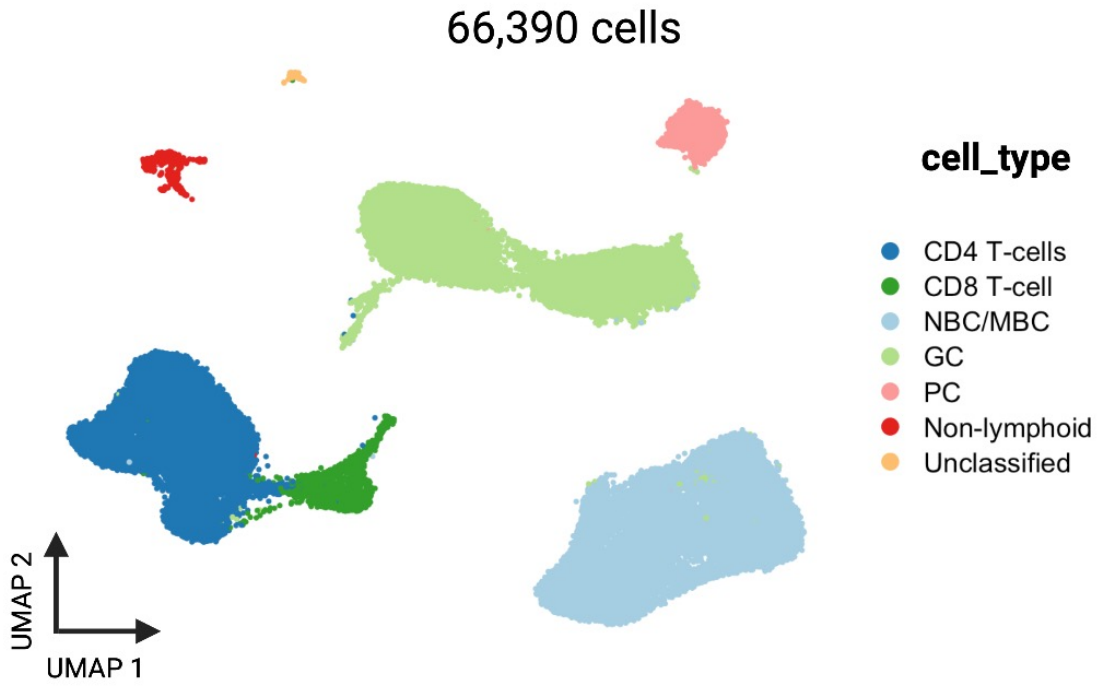


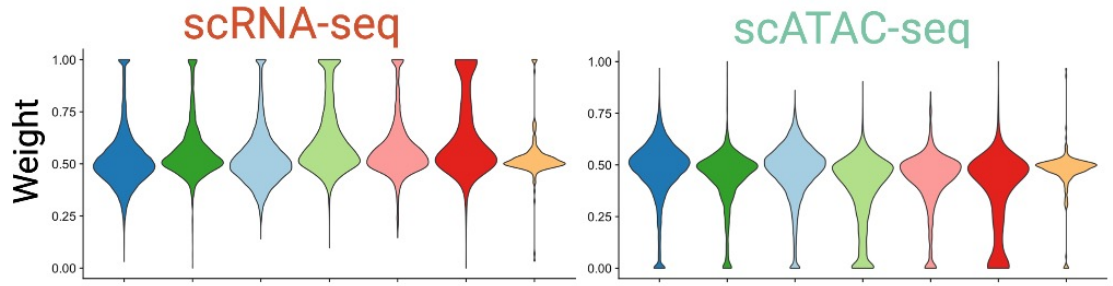
1° ***FindModalNeighbors***
This function will construct a weighted nearest neighbor (WNN) graph.

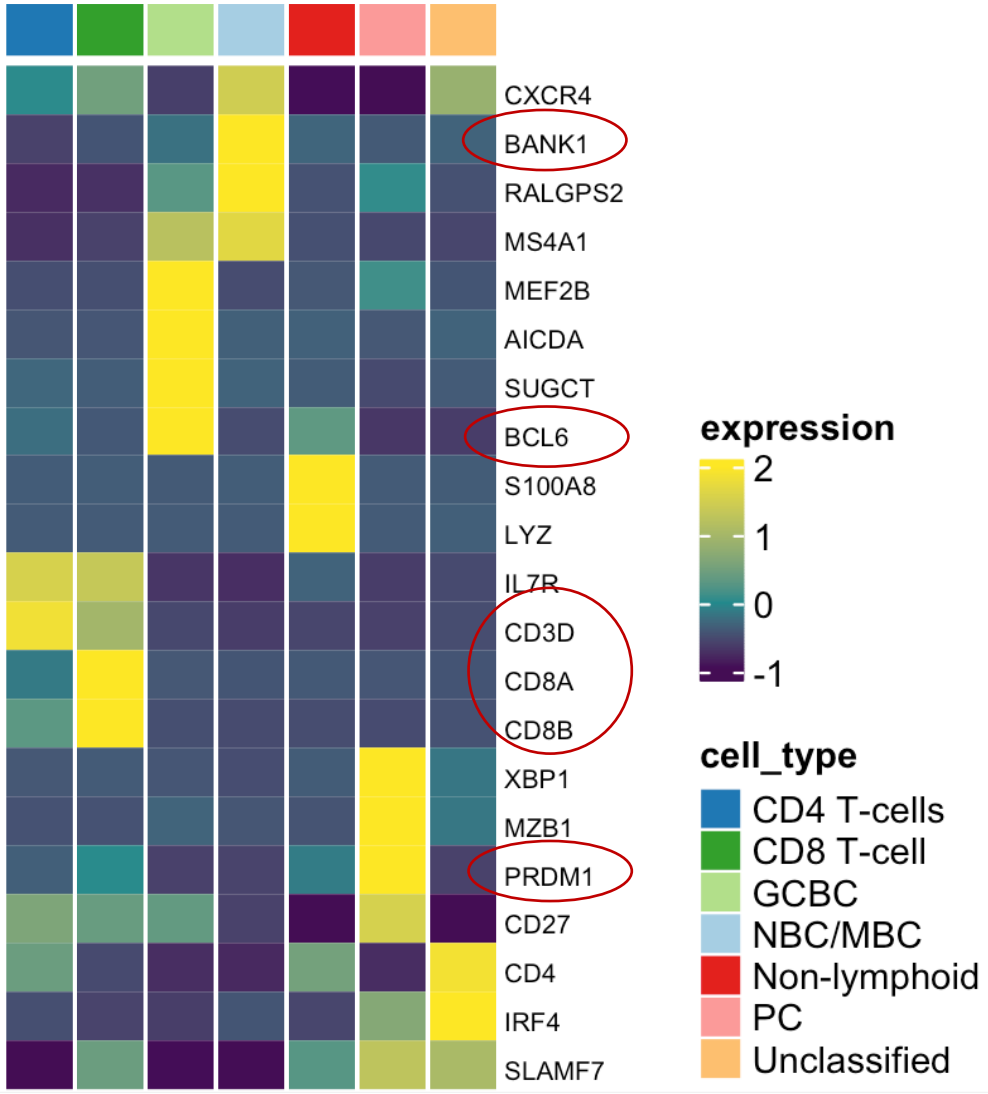
2° ***RunUMAP:*** *as we do in* scRNA-seq

Cell type identification

Curated list of bibliographic biomarkers

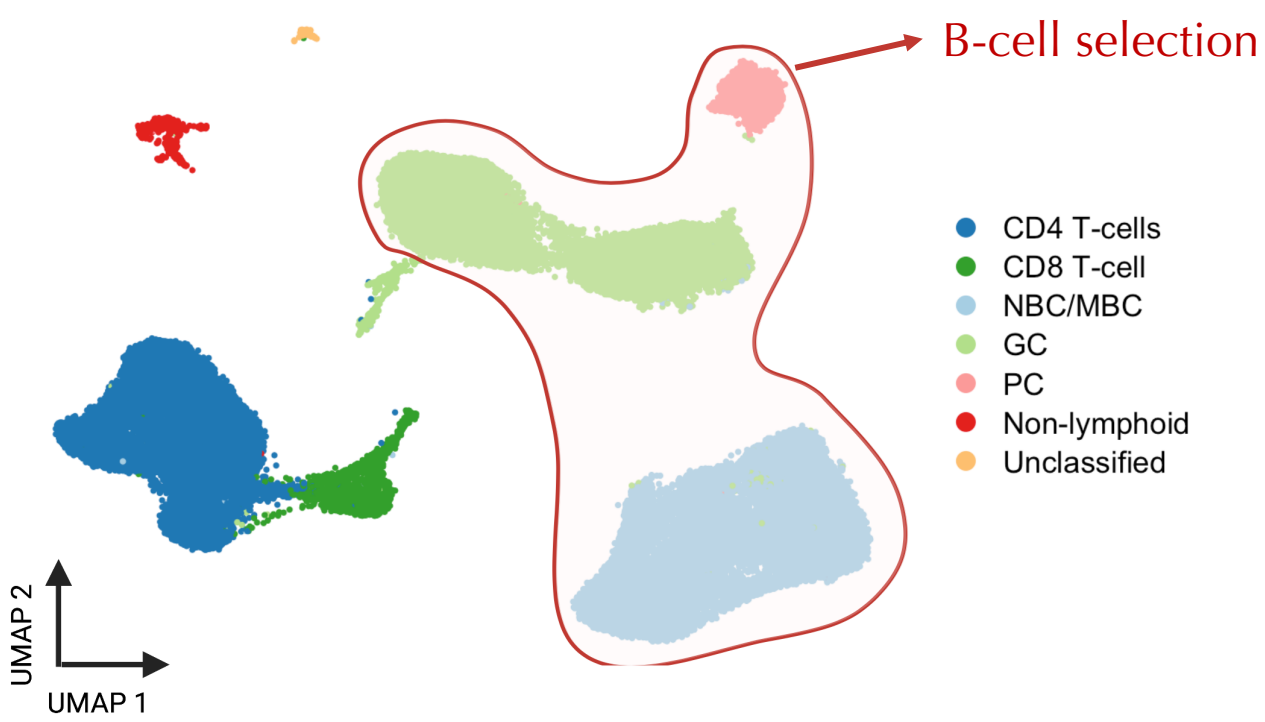


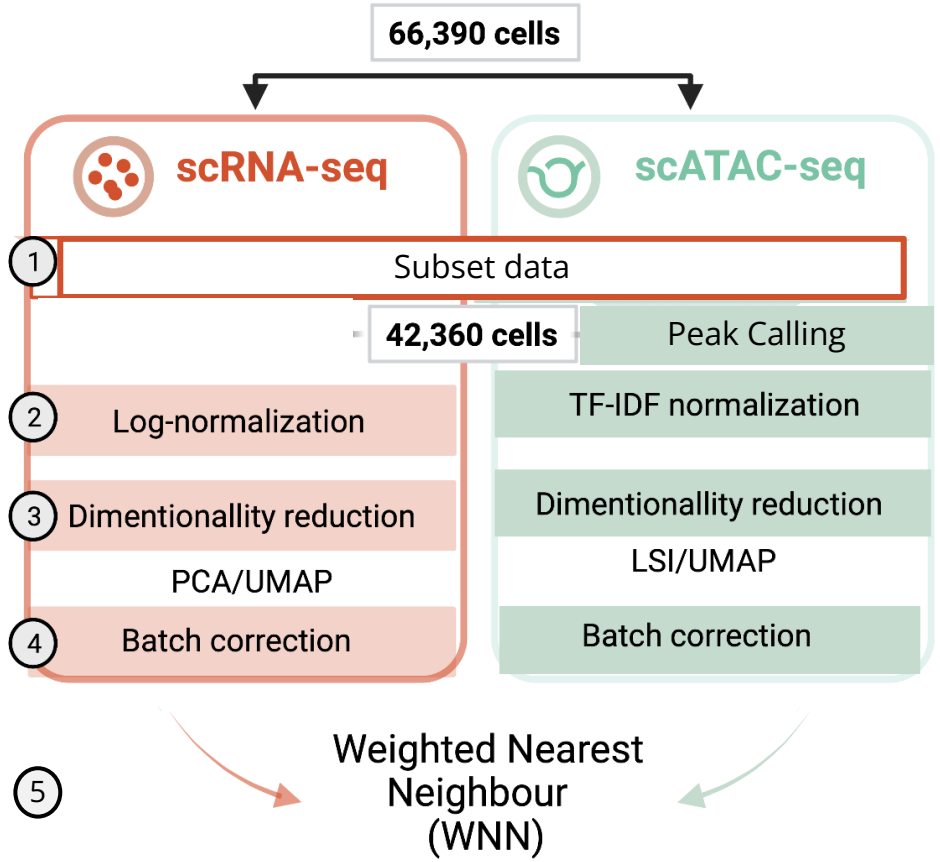




B-Cell type selection

Pre-processing B-cell data set

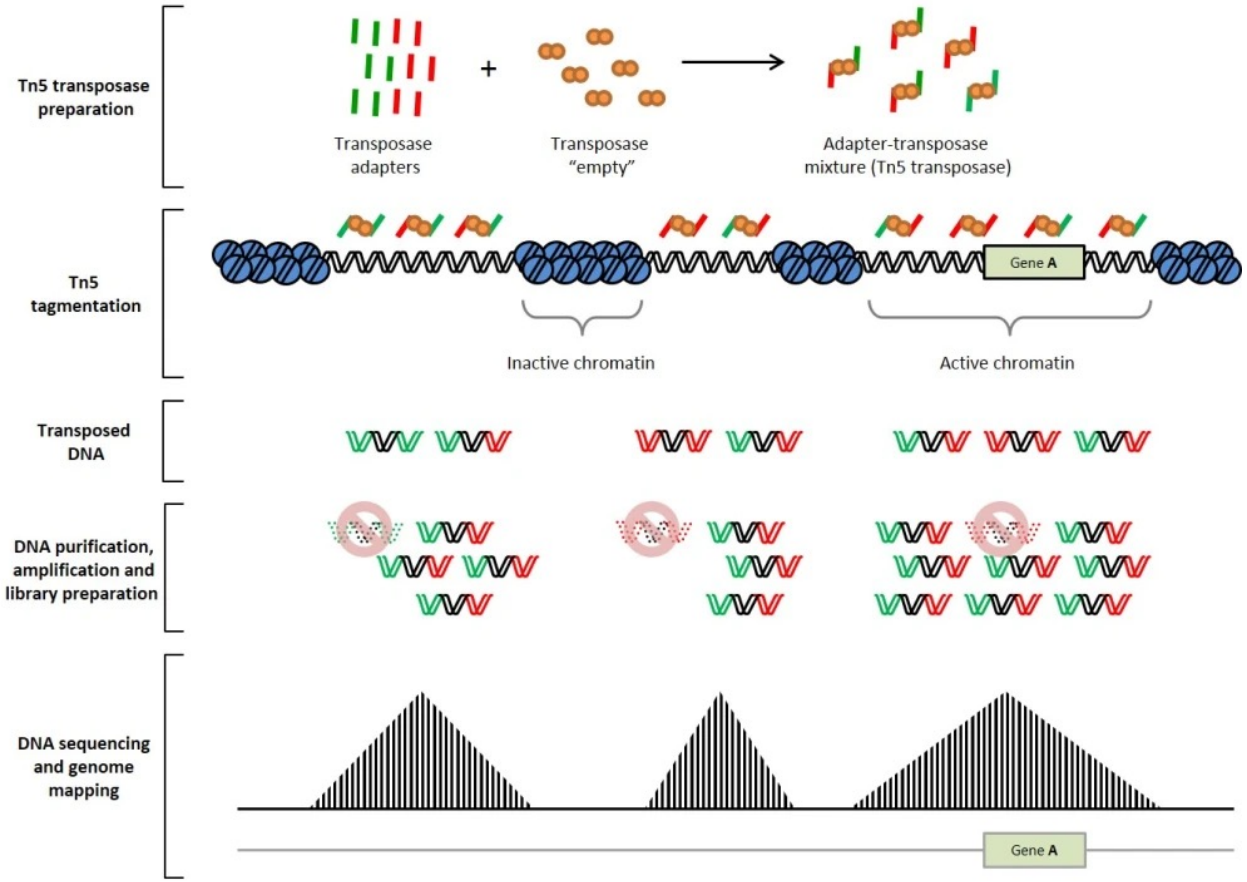




Peak calling

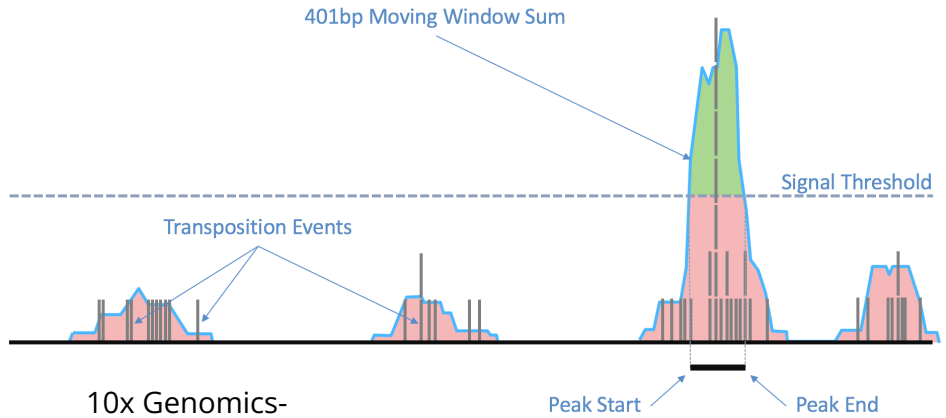
Peak calling is one of the most common analyses to identify areas in the genome that have been enriched with aligned reads which are known as **differentially accessible regions (DARs) in a data set.**

The goal of the peak calling algorithm in the scATAC assay is to identify which distinct regions of the genome, known as peaks (open chromatin), are the key features of interest.



Since both peak signal and background noise can vary across different datasets and locally across the genome, the algorithm generates a global peak threshold to identify specific peaks.

'CallPeaks' function of Signac uses MACS2 algorithm. We can call peaks separately for different groups of cells, or performed using data from all the cells.



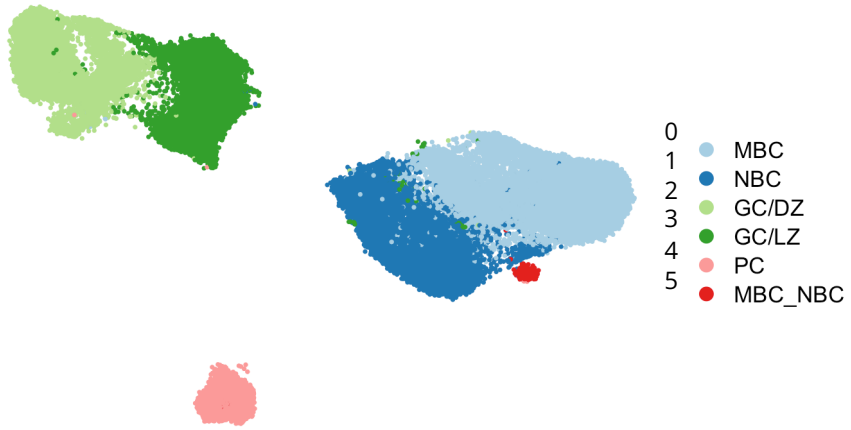
10x Genomics- https://stuartlab.org/signac/articles/peak_calling.html

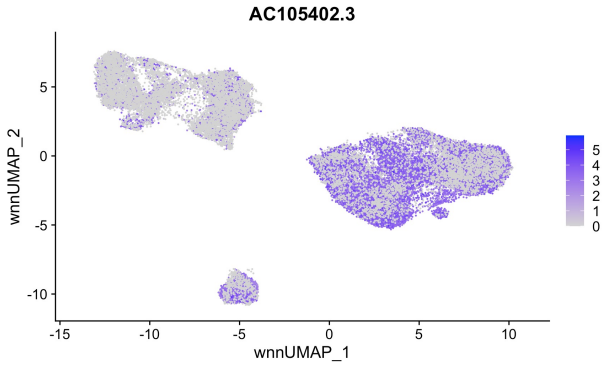
Image by Manuel Rosa Garrido

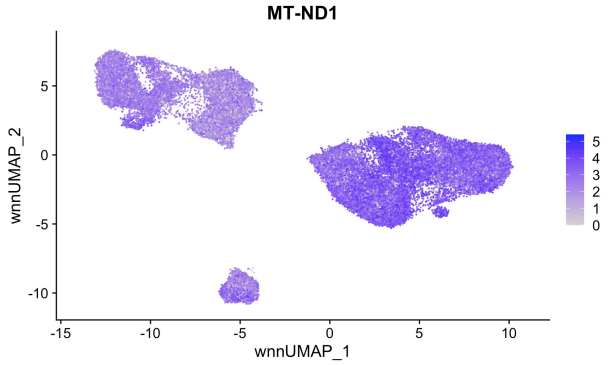
Joint UMAP – main B-cell population

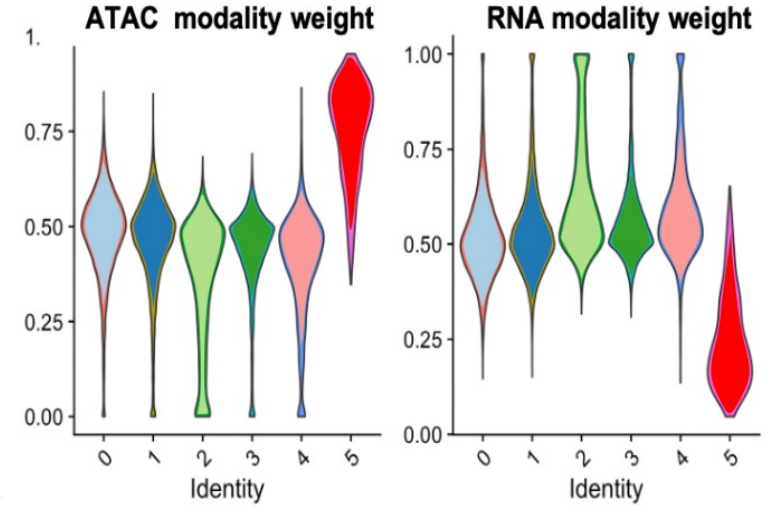
Analysing MBC_NBC Cluster

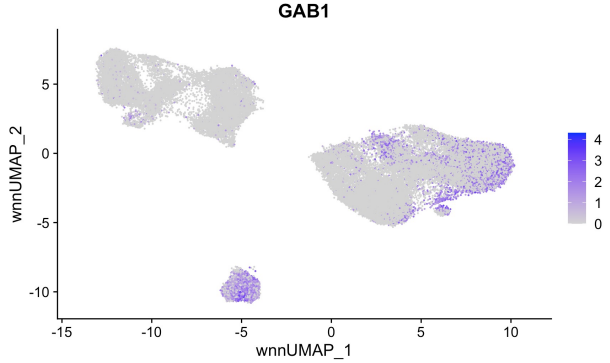
Joint UMAP





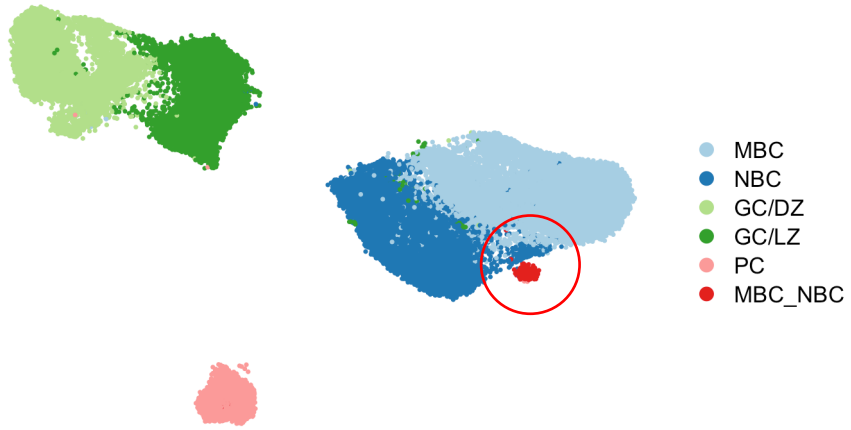




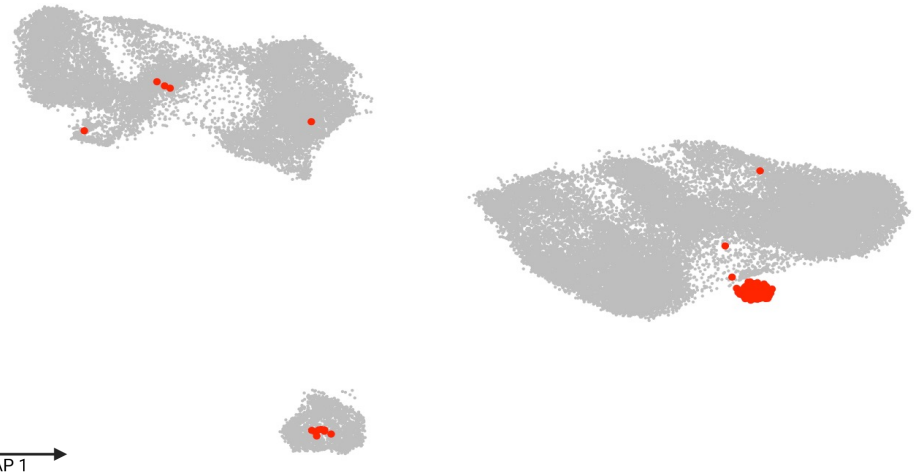


Analysing MBC_NBC Cluster

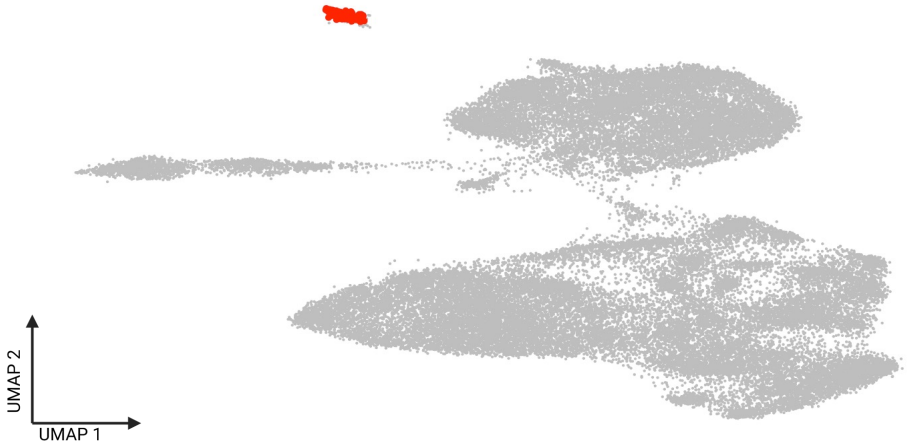
Joint UMAP



Joint UMAP



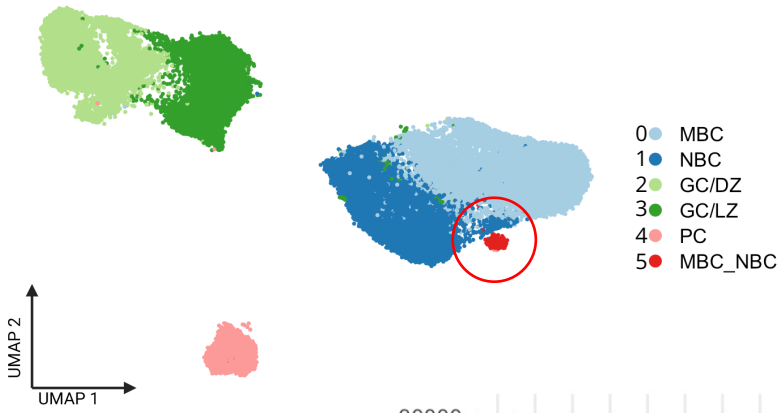
ATAC UMAP

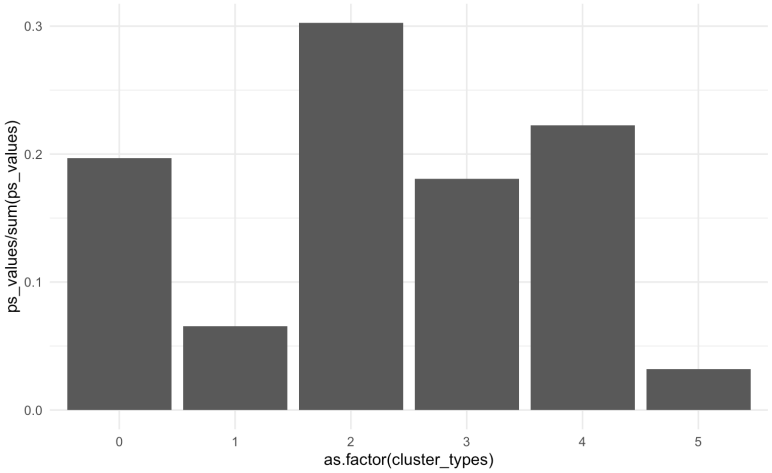


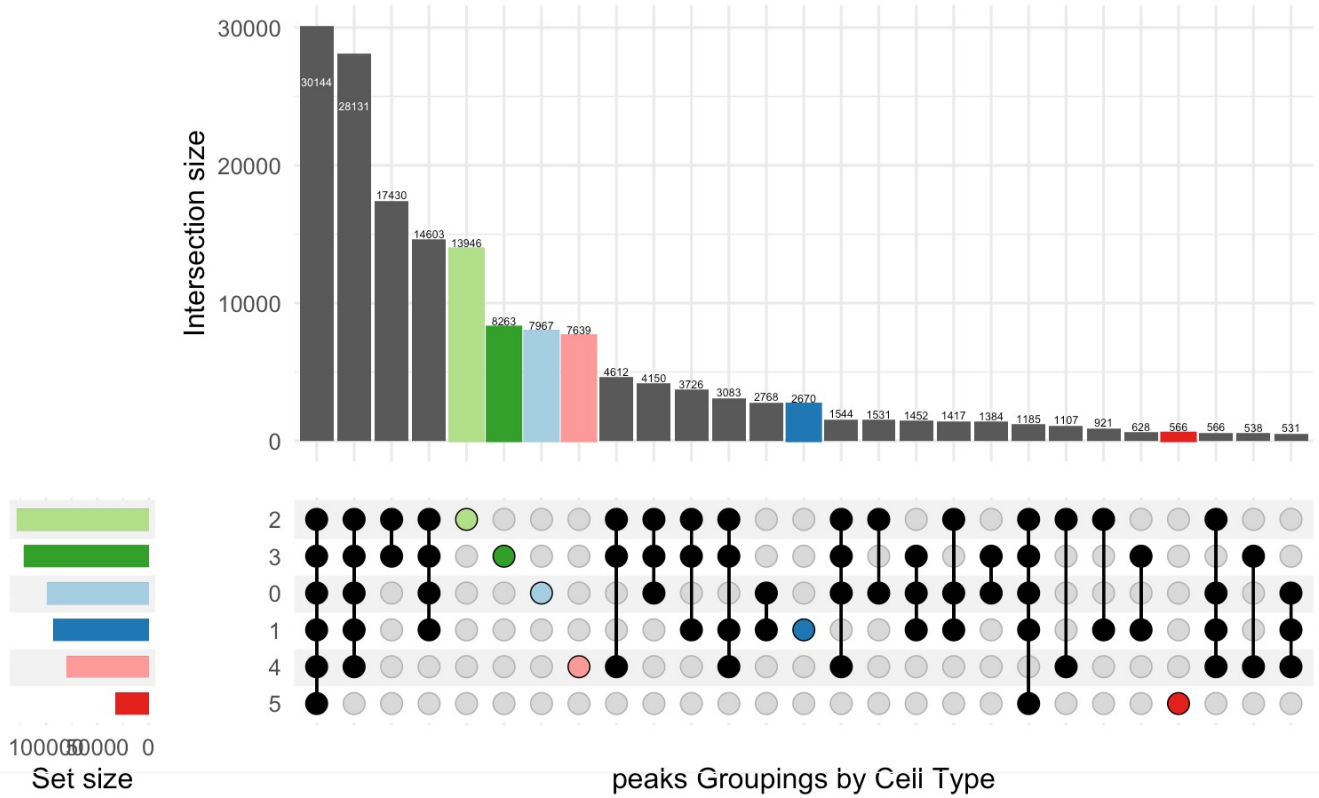
RNA UMAP

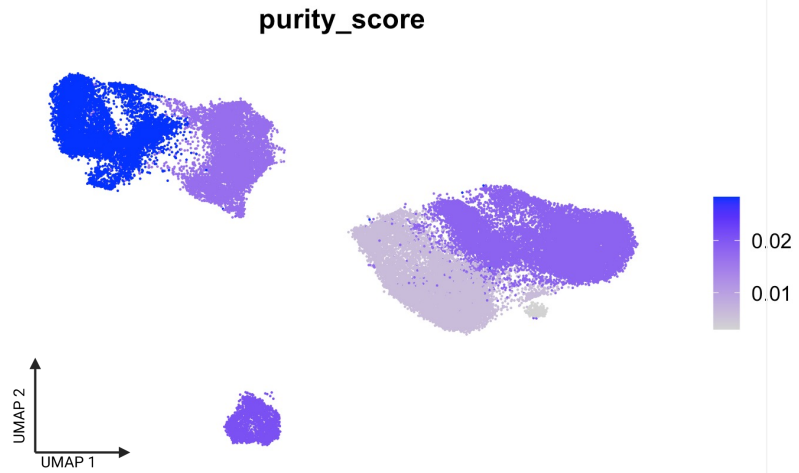
Analysing MBC_NBC Cluster

Joint UMAP





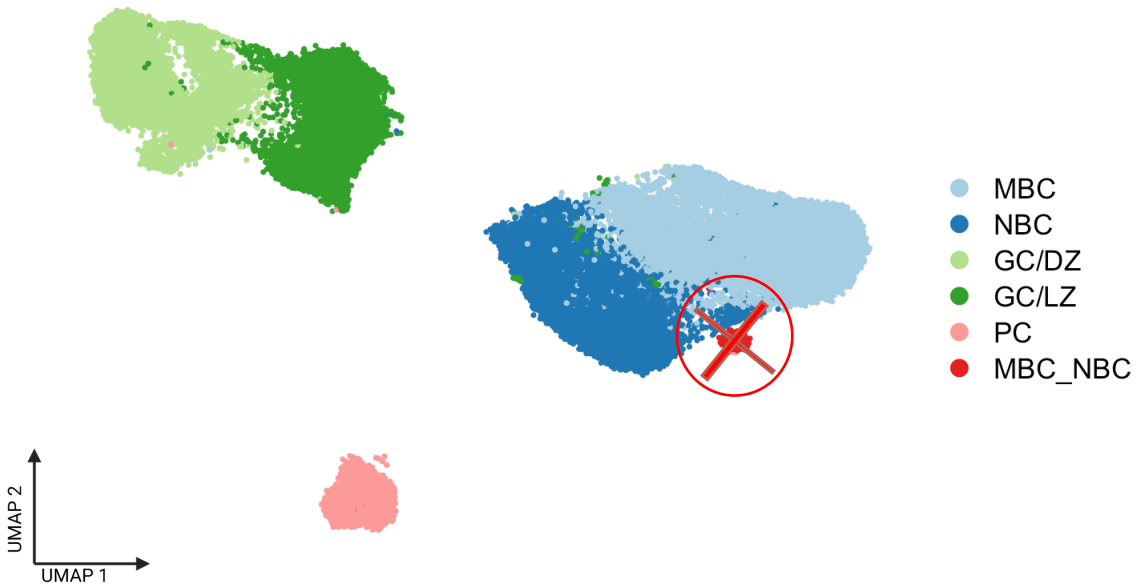


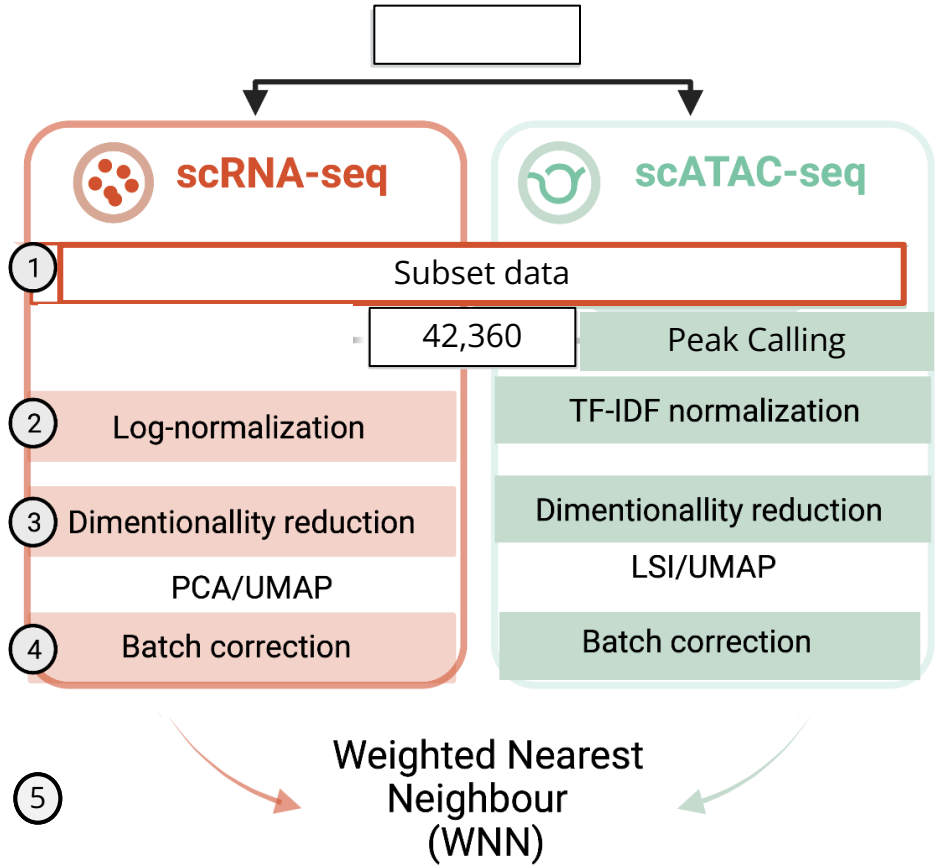


B-Cell type selection

Pre-processing B-cell data set

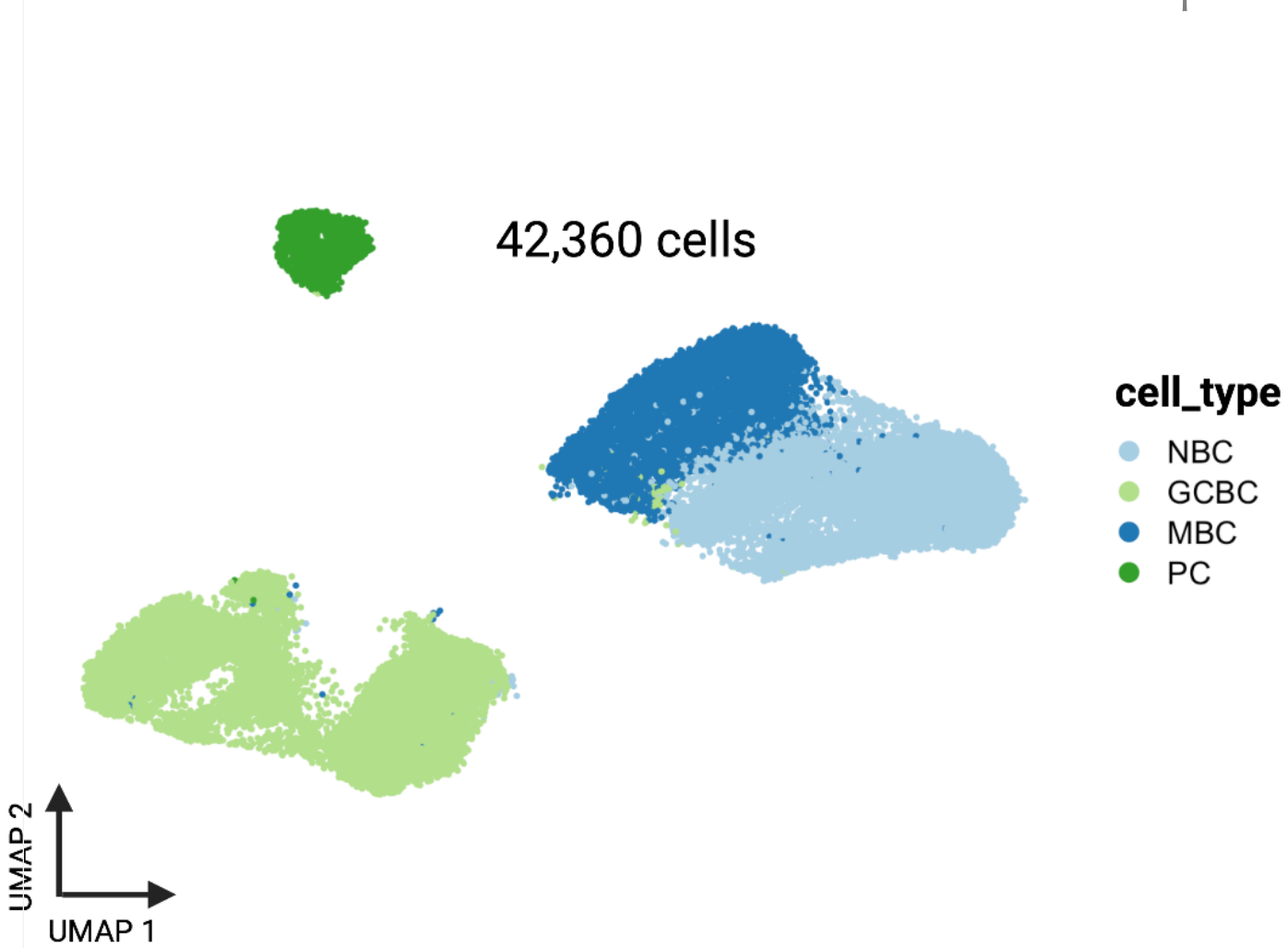
Joint UMAP

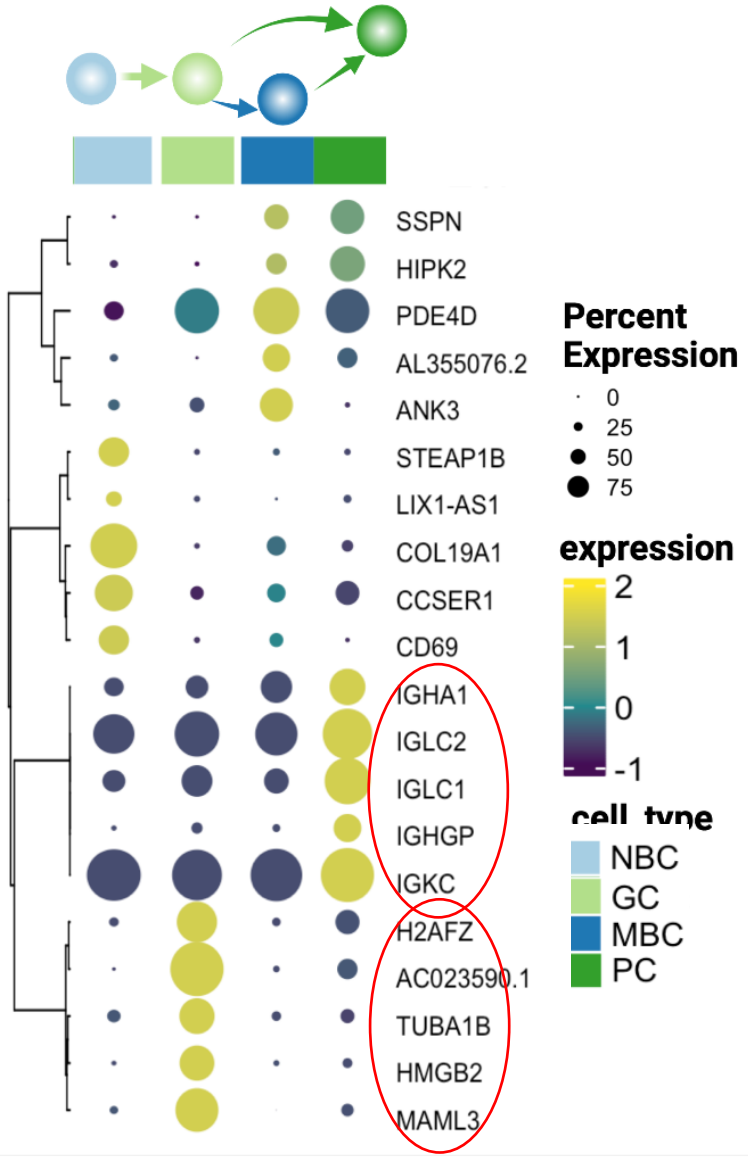




B-Cell type identification

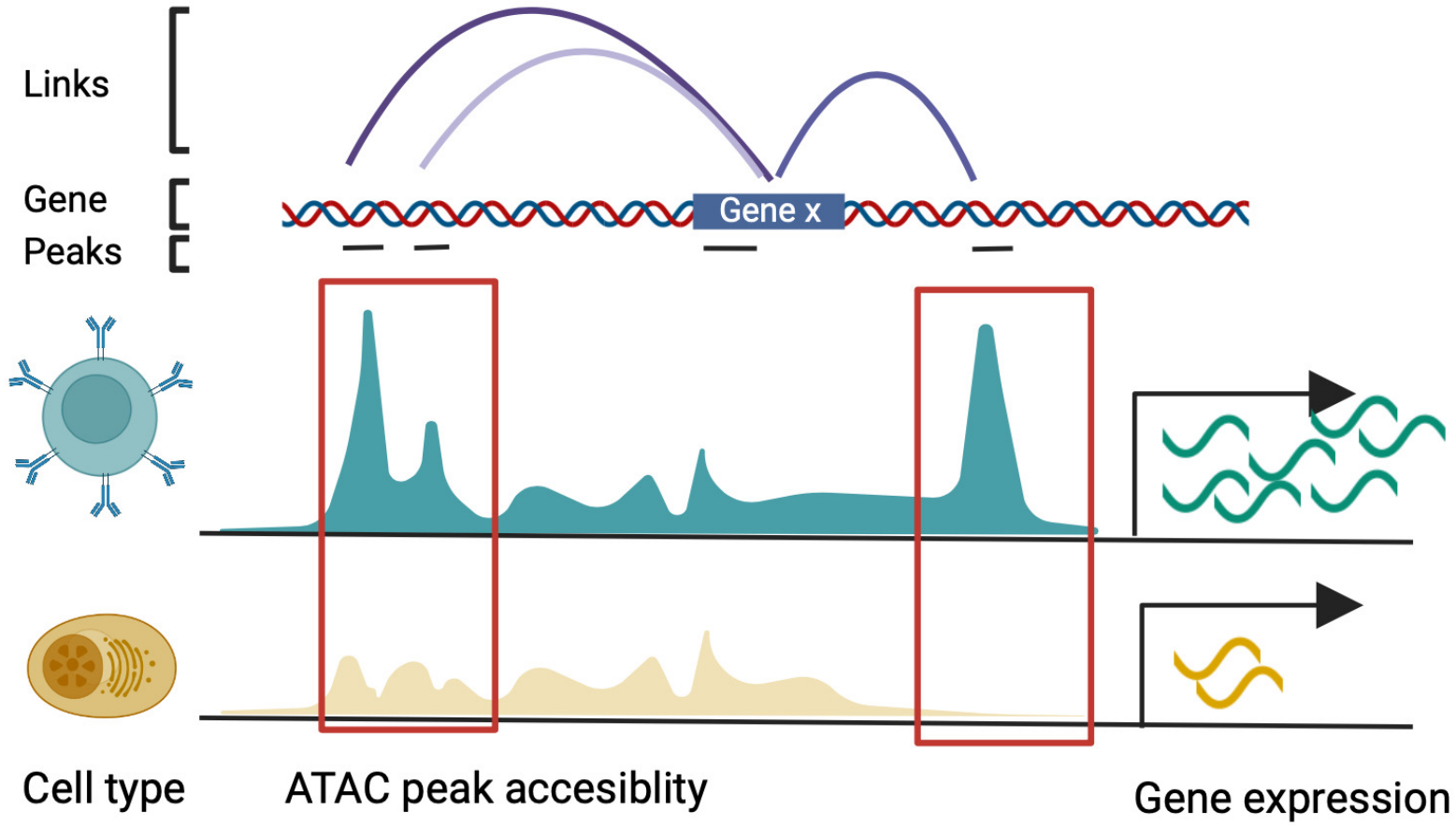
Curated list of top biomarkers





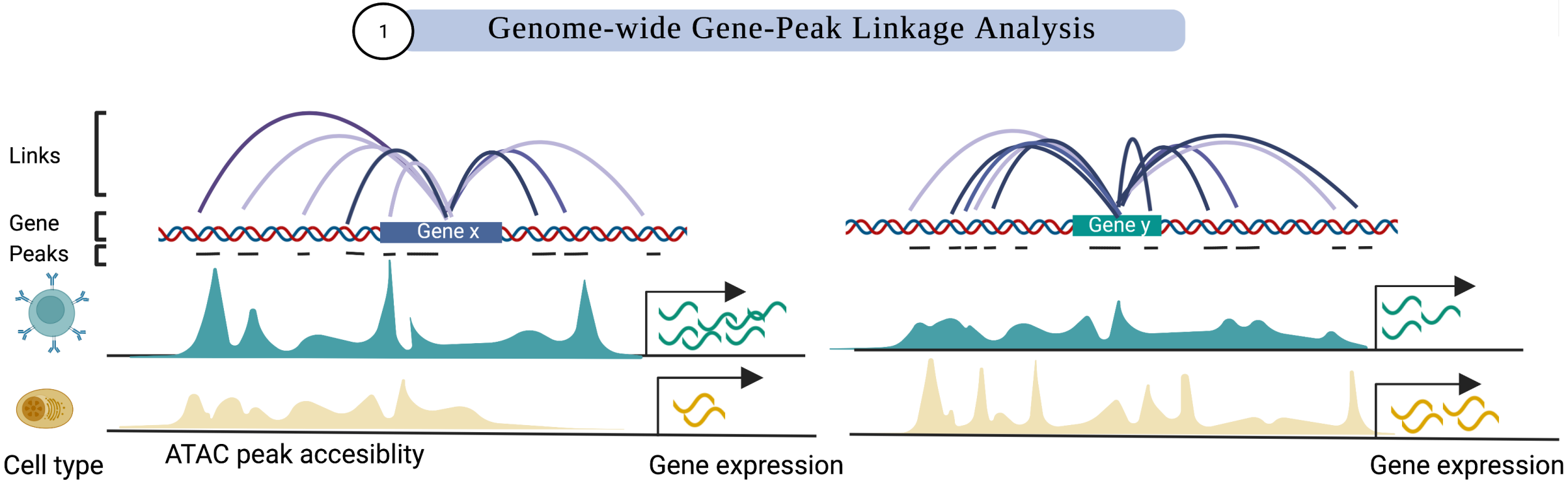
Peak-gene Linkage

Coverage plot explanation



Peak-gene Linkage

Gene-Peak Links

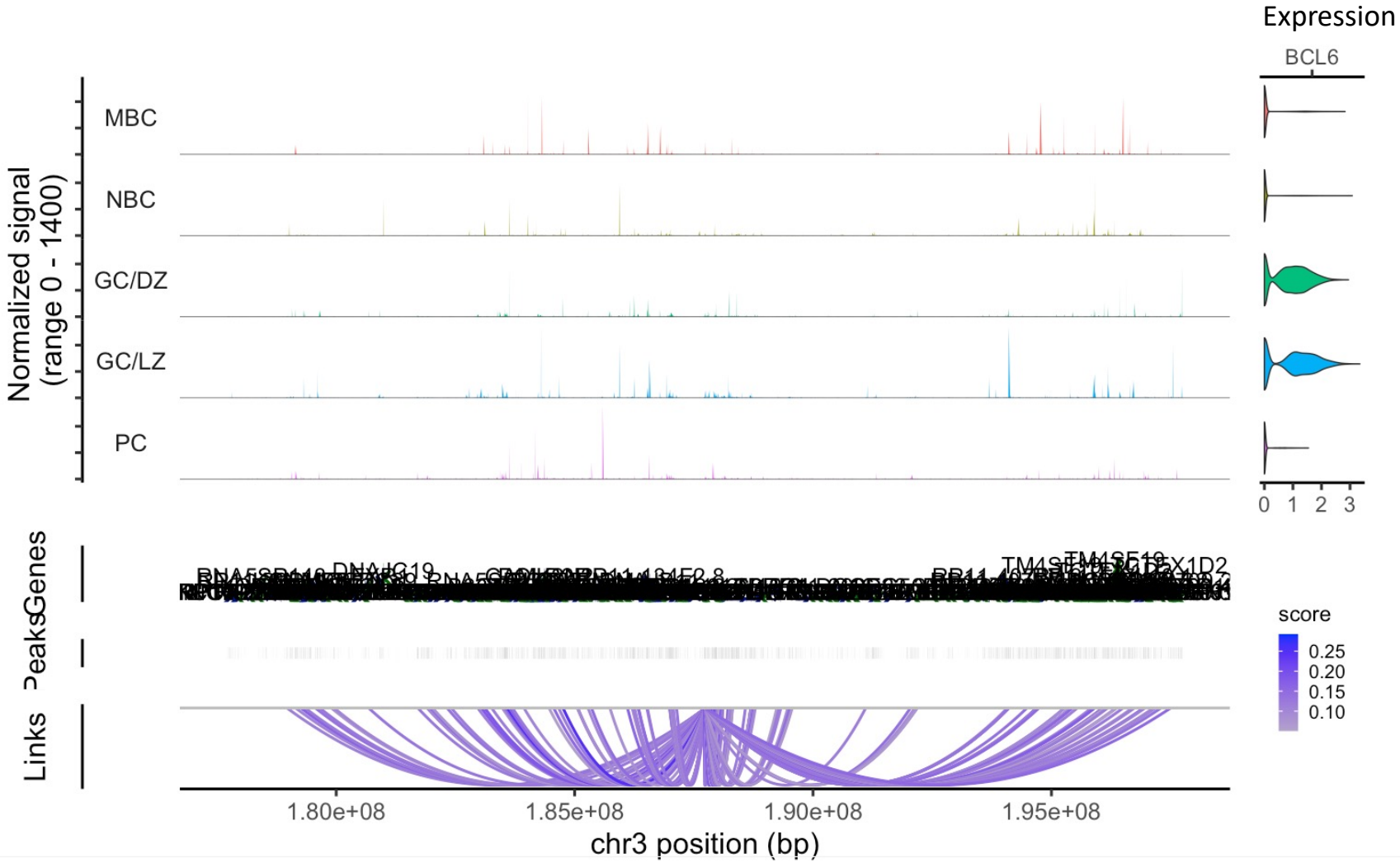


- **1e+7 bp** upstream and downstream from the gene TSS.
- **Limitation**: we cannot perform peak- gene linkage analysis at the inter-chromosomal level.

Peak-gene Linkage

Genome-wide Gene-Peak Links in BLC6 gene

Genome-wide Link at 1e+7 bp up and downstream distance from the TSS of BCL6 gene

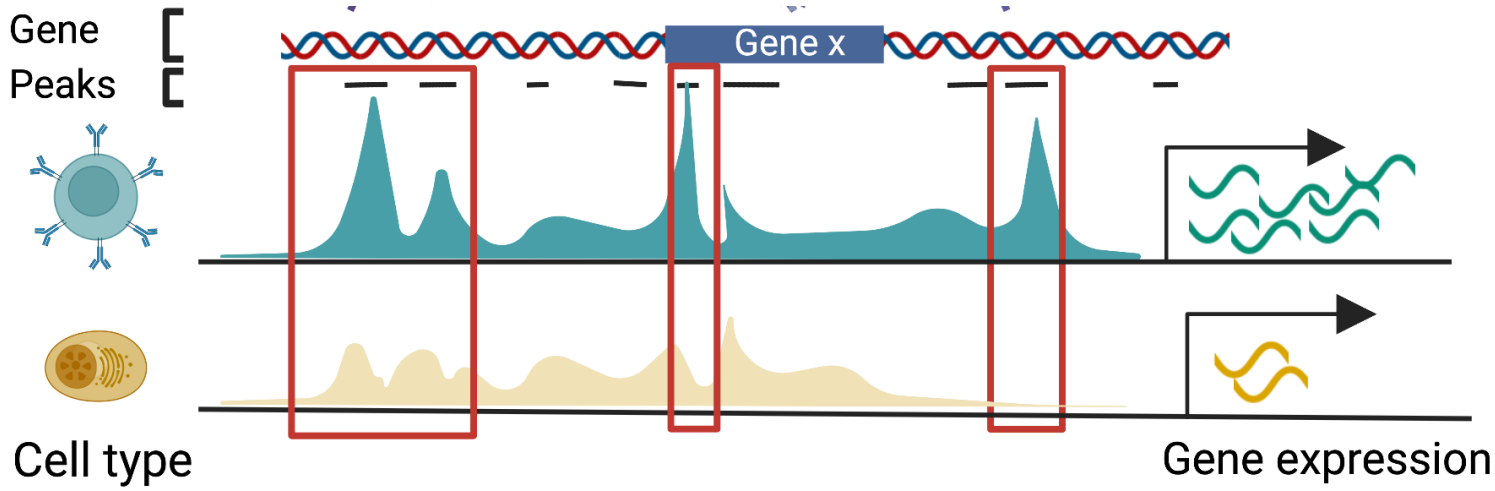


Peak-gene Linkage

Performing Differentially Accessible Region (DARs) Analysis

In this step we identify open chromatin regions (peaks) that are differentially accessible in each B-cell subtype compared to the rest.

- ***FindAllCluster:*** *assay="ATAC"*



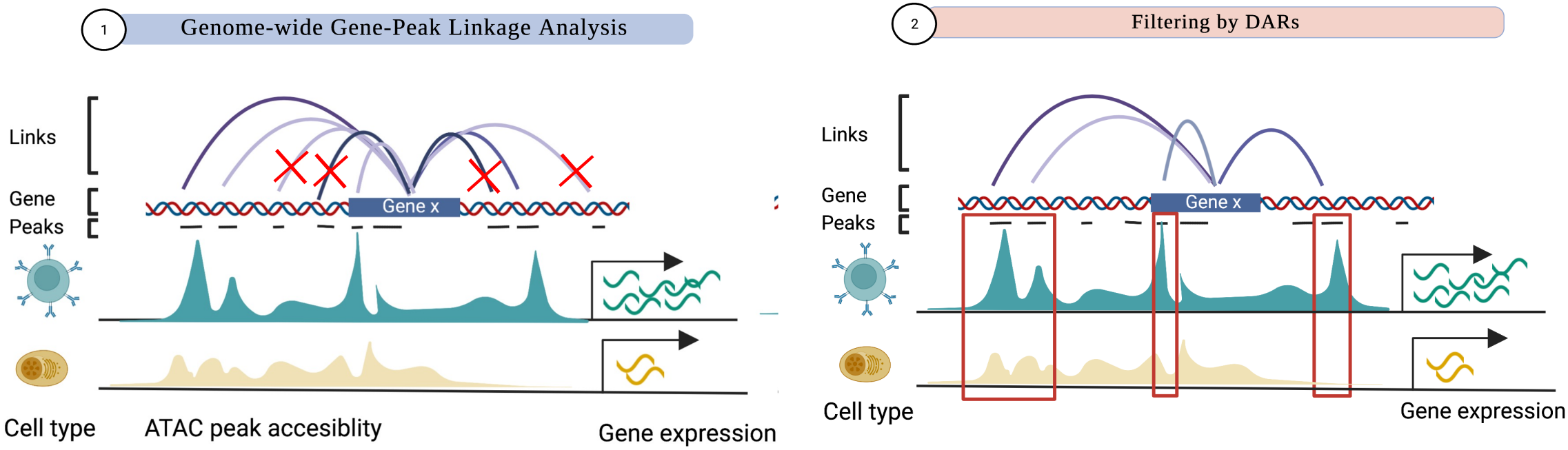
This returns a data frame with all the DNA coordinates location of these differentially accessible regions of each cell type.

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	peak	seqnames	start	end
chr19-56476517-56478318	0	0.4877339	0.209	0.057	0	MBC	chr19-56476517-56478318	chr19	56476517	56478318
chr19-56507186-56508386	0	0.4016254	0.206	0.079	0	MBC	chr19-56507186-56508386	chr19	56507186	56508386

Peak-gene Linkage

Links filtering by DARs peaks

In this step we remove all the links coordinates that overlap
with differentially accessible region coordinates



Peak-gene Linkage

Joining gene-peaks links coordinates and DARs coordinates

PRDM1 and BCL6

Gene-peak links table 1326

Table of all Link peaks of tonsil leve 1 object

	seqnames	start	end	width	strand	score	gene	peak	zscore	pvalue
chr3-277802-279075	chr3	278439	187745727	187467289	*	0.1242044	BCL6	chr3-277802-279075	2.659344	0.0039147
chr3-1380485-1381960	chr3	1381223	187745727	186364505	*	0.1209302	BCL6	chr3-1380485-1381960	2.012666	0.0220749
chr3-4100706-4101604	chr3	4101155	187745727	183644573	*	0.1007473	BCL6	chr3-4100706-4101604	2.325074	0.0100340
chr3-4492144-4494162	chr3	4493153	187745727	183252575	*	0.1216546	BCL6	chr3-4492144-4494162	2.044232	0.0204653
chr3-4502868-4503806	chr3	4503337	187745727	183242391	*	0.1653158	BCL6	chr3-4502868-4503806	3.089958	0.0010009
chr3-4977216-4979817	chr3	4978517	187745727	182767211	*	-0.1018364	BCL6	chr3-4977216-4979817	-3.332075	0.0004310
chr3-4983777-4984537	chr3	4984157	187745727	182761571	*	-0.0560302	BCL6	chr3-4983777-4984537	-2.279728	0.0113119

DARs 7837

Table of the DARs

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	peak	seqnames	start	end
chr19-56476517-56478318	0	0.4877339	0.209	0.057	0	MBC	chr19-56476517-56478318	chr19	56476517	56478318
chr19-56507186-56508386	0	0.4016254	0.206	0.079	0	MBC	chr19-56507186-56508386	chr19	56507186	56508386
chr6-81751450-81754199	0	0.3958804	0.295	0.153	0	MBC	chr6-81751450-81754199	chr6	81751450	81754199
chr11-114178964-114179911	0	0.3611183	0.116	0.028	0	MBC	chr11-114178964-114179911	chr11	114178964	114179911
chr4-123443156-123444674	0	0.3756049	0.154	0.052	0	MBC	chr4-123443156-123444674	chr4	123443156	123444674
chr10-36534247-36535094	0	0.3537690	0.112	0.029	0	MBC	chr10-36534247-36535094	chr10	36534247	36535094
chr7-63899675-63901274	0	0.3561356	0.150	0.051	0	MBC	chr7-63899675-63901274	chr7	63899675	63901274

Join by="peak"

Peak-gene Linkage

Filtering links by the DARs

560 Links overlaping in DARs

Table of the join of the all links peaks and the DARs

seqnames	start	end	width	strand	score	gene	peak	zscore	pvalue	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	start.peak	end.peak
chr3	1381223	187745727	186364505	*	0.1209302	BCL6	chr3-1380485-1381960	2.012666	0.0220749	0	0.2983296	0.176	0.069	0	GC/LZ	1380485	1381960
chr3	4978517	187745727	182767211	*	-0.1018364	BCL6	chr3-4977216-4979817	-3.332075	0.0004310	0	0.2900502	0.371	0.245	0	MBC	4977216	4979817
chr3	8501630	187745727	179244098	*	0.1294057	BCL6	chr3-8501296-8501964	1.792125	0.0365565	0	0.2607253	0.312	0.125	0	GC/DZ	8501296	8501964
chr3	8658427	187745727	179087301	*	0.1187620	BCL6	chr3-8657833-8659020	1.686345	0.0458647	0	0.2907075	0.123	0.029	0	GC/DZ	8657833	8659020
chr3	9653316	187745727	178092412	*	0.1859457	BCL6	chr3-9652401-9654230	3.503141	0.0002299	0	0.3467498	0.313	0.103	0	GC/DZ	9652401	9654230
chr3	9653316	187745727	178092412	*	0.1859457	BCL6	chr3-9652401-9654230	3.503141	0.0002299	0	0.2631839	0.242	0.120	0	GC/LZ	9652401	9654230

BCL6

	MBC	NBC	GC/DZ	GC/LZ	PC
FALSE	0	0	15	1	79
TRUE	1	0	334	121	9

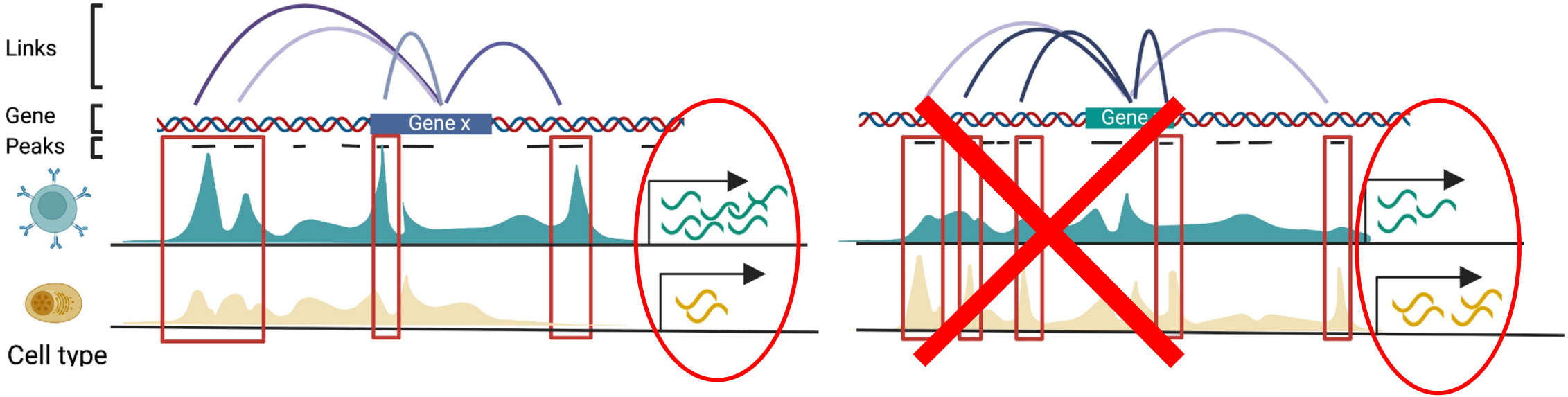
PRDM1

	MBC	NBC	GC/DZ	GC/LZ	PC
FALSE	1	0	334	121	9
TRUE	0	0	15	1	79

Peak-gene Linkage

Links filtering by DEGs and Links classification

3 Filtering by DEGs



Peak-gene Linkage

Filtering links by the DEG, so we remove all the links that are differentially accessible

560 Links overlapping in DARs

Table of the join of the all links peaks and the DARs

seqnames	start	end	width	strand	score	gene	peak	zscore	pvalue	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	start.peak	end.peak
chr3	1381223	187745727	186364505	*	0.1209302	BCL6	chr3-1380485-1381960	2.012666	0.0220749	0	0.2983296	0.176	0.069	0	GC/LZ	1380485	1381960
chr3	4978517	187745727	182767211	*	-0.1018364	BCL6	chr3-4977216-4979817	-3.332075	0.0004310	0	0.2900502	0.371	0.245	0	MBC	4977216	4979817
chr3	8501630	187745727	179244098	*	0.1294057	BCL6	chr3-8501296-8501964	1.792125	0.0365565	0	0.2607253	0.312	0.125	0	GC/DZ	8501296	8501964
chr3	8658427	187745727	179087301	*	0.1187620	BCL6	chr3-8657833-8659020	1.686345	0.0458647	0	0.2907075	0.123	0.029	0	GC/DZ	8657833	8659020
chr3	9653316	187745727	178092412	*	0.1859457	BCL6	chr3-9652401-9654230	3.503141	0.0002299	0	0.3467498	0.313	0.103	0	GC/DZ	9652401	9654230
chr3	9653316	187745727	178092412	*	0.1859457	BCL6	chr3-9652401-9654230	3.503141	0.0002299	0	0.2631839	0.242	0.120	0	GC/LZ	9652401	9654230

DEGs

Table of the markers

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluste	gene
COL19A1	0.0000000	2.4074748	0.796	0.211	0.0000000	MBC	COL19A1
STEAP1B	0.0000000	2.1946661	0.520	0.118	0.0000000	MBC	STEAP1B
CD69	0.0000000	1.6809713	0.517	0.168	0.0000000	MBC	CD69
CCSER1	0.0000000	1.5955518	0.648	0.294	0.0000000	MBC	CCSER1
LIX1-AS1	0.0000000	1.4879506	0.270	0.095	0.0000000	MBC	LIX1-AS1
PTPRK	0.0000000	1.4630285	0.471	0.122	0.0000000	MBC	PTPRK
GAB1	0.0000000	1.4321868	0.307	0.100	0.0000000	MBC	GAB1
KHDRBS2	0.0000000	1.3803476	0.707	0.382	0.0000000	MBC	KHDRBS2
ST6GALNAC3	0.0000000	1.3096820	0.416	0.124	0.0000000	MBC	ST6GALNAC3
PDE7B	0.0000000	1.2796047	0.313	0.072	0.0000000	MBC	PDE7B
FOXP1	0.0000000	1.2734718	0.941	0.727	0.0000000	MBC	FOXP1
ABCB4	0.0000000	1.2462463	0.404	0.113	0.0000000	MBC	ABCB4
LINC00926	0.0000000	1.2245089	0.580	0.267	0.0000000	MBC	LINC00926
FCER2	0.0000000	1.2215059	0.373	0.112	0.0000000	MBC	FCER2
AUTS2	0.0000000	1.2214382	0.777	0.524	0.0000000	MBC	AUTS2

Join by=c("gene","cluster")

Removing the no-matches links

Peak-gene Linkage

Filtering links by the DEG

534 Links overlapping in DARs and are related to DEG

Table of the filtered links peaks by DARs and GEx

seqnames	start	end	width	strand	score	gene	peak	zscore	pvalue	p_val.dars	avg_log2FC.dars	pct.1.dars	pct.2.dars	p_val_adj.dars	cluster	start.peak	end.peak	p_val.rna	avg_log2FC.rna	pct.1.rna	pct.2.rna	p_val_adj.rna
chr3	1381223	187745727	186364505	*	0.1209302	BCL6	chr3-1380485-1381960	2.012666	0.0220749	0	0.2983296	0.176	0.069	0	GC/LZ	1380485	1381960	0	1.1745827	0.658	0.197	0
chr3	8501630	187745727	179244098	*	0.1294057	BCL6	chr3-8501296-8501964	1.792125	0.0365565	0	0.2607253	0.312	0.125	0	GC/DZ	8501296	8501964	0	0.8223499	0.731	0.175	0
chr3	8658427	187745727	179087301	*	0.1187620	BCL6	chr3-8657833-8659020	1.686345	0.0458647	0	0.						8659020	0	0.8223499	0.731	0.175	0
chr3	9653316	187745727	178092412	*	0.1859457	BCL6	chr3-9652401-9654230	3.503141	0.0002299	0	0.3467498	0.313	0.103	0	GC/DZ	9652401	9654230	0	0.8223499	0.731	0.175	0
chr3	9653316	187745727	178092412	*	0.1859457	BCL6	chr3-9652401-9654230	3.503141	0.0002299	0	0.2631839	0.242	0.120	0	GC/LZ	9652401	9654230	0	1.1745827	0.658	0.197	0

kable_paper(kable_input, lightable_options = "basic", html_font = "\"Arial

BCL6

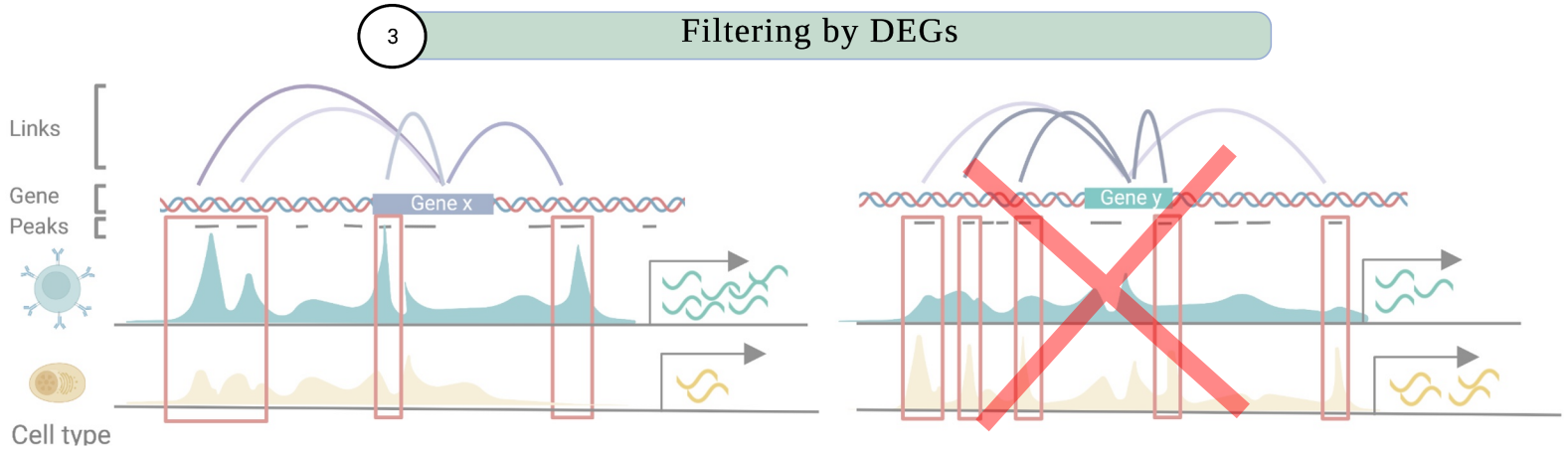
	MBC	NBC	GC/DZ	GC/LZ	PC
FALSE	0	0	0	0	79
TRUE	0	0	334	121	0

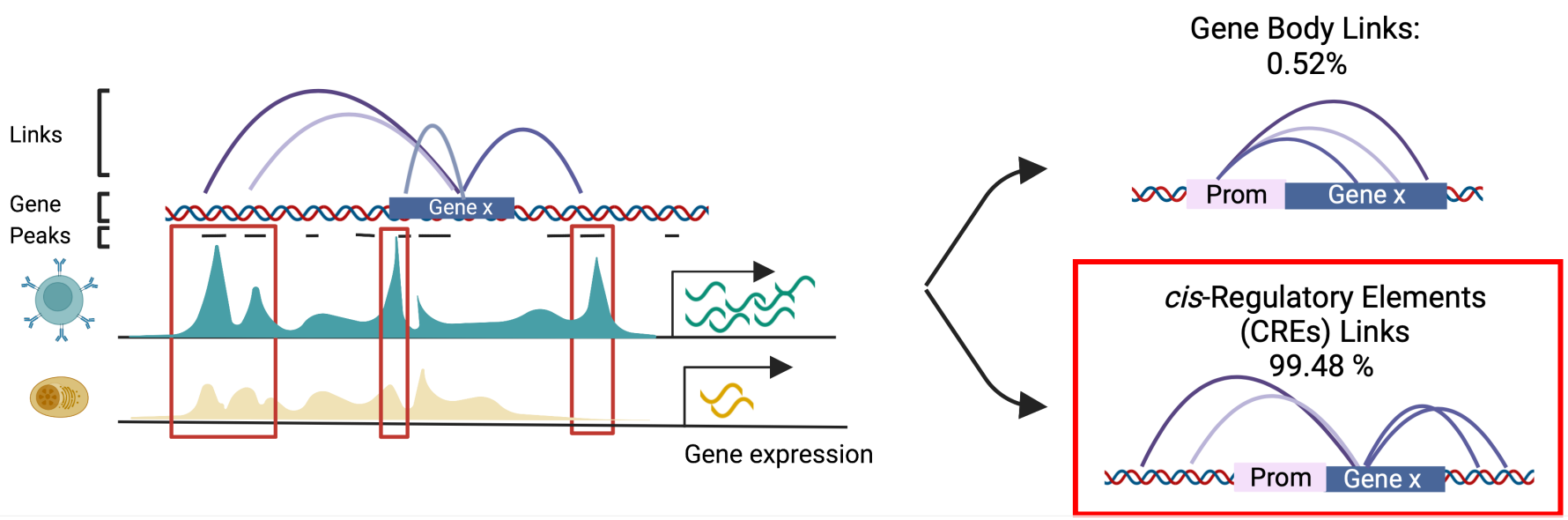
PRDM1

	MBC	NBC	GC/DZ	GC/LZ	PC
FALSE	0	0	334	121	0
TRUE	0	0	0	0	79

Peak-gene Linkage

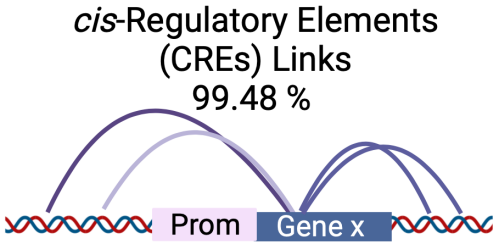
Links filtering by DEGs and Links classification

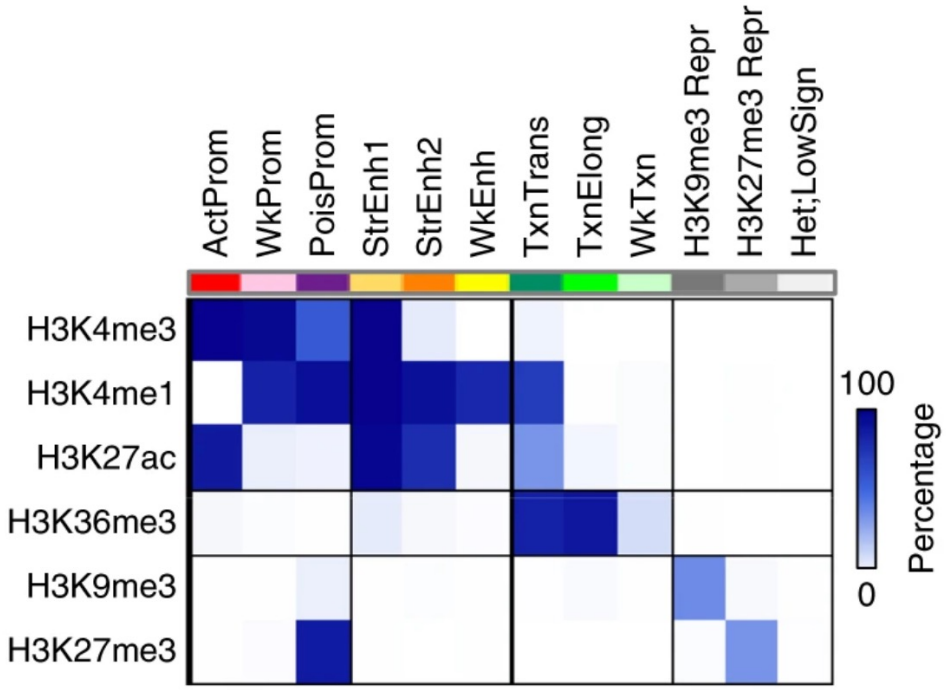




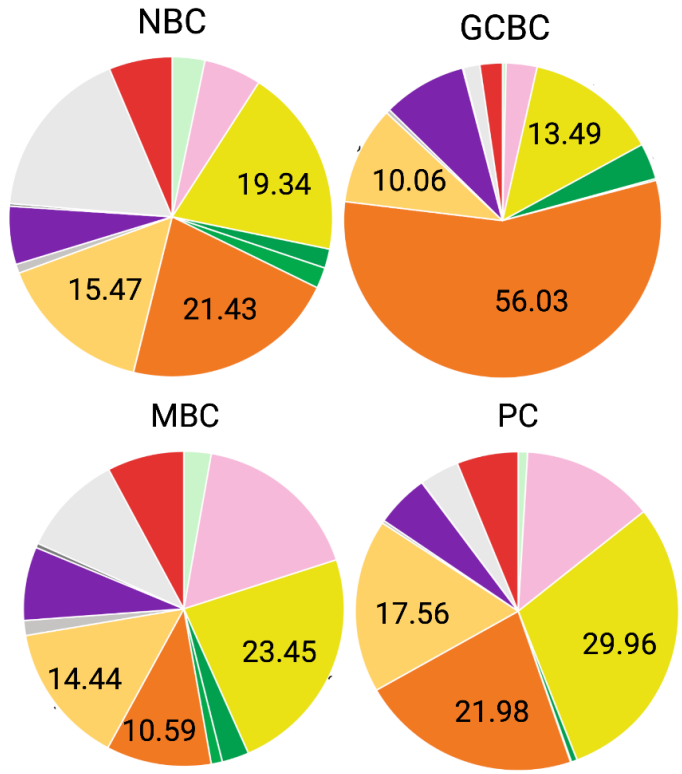
cis-Regulatory Elements (CREs) identification of target B-cell genes

Peak-gene Links filtering & classification by chromatin states



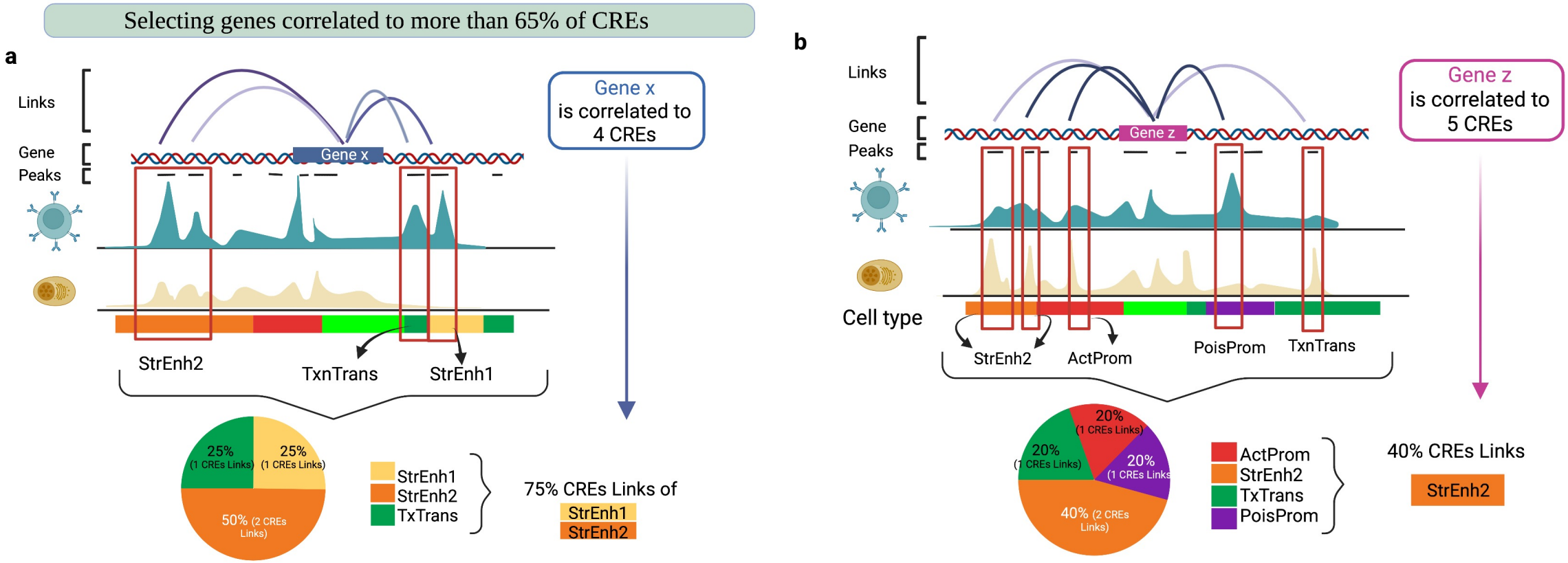


Beekman R. Nat. Med. 2018



CREs identification of target B-cell genes

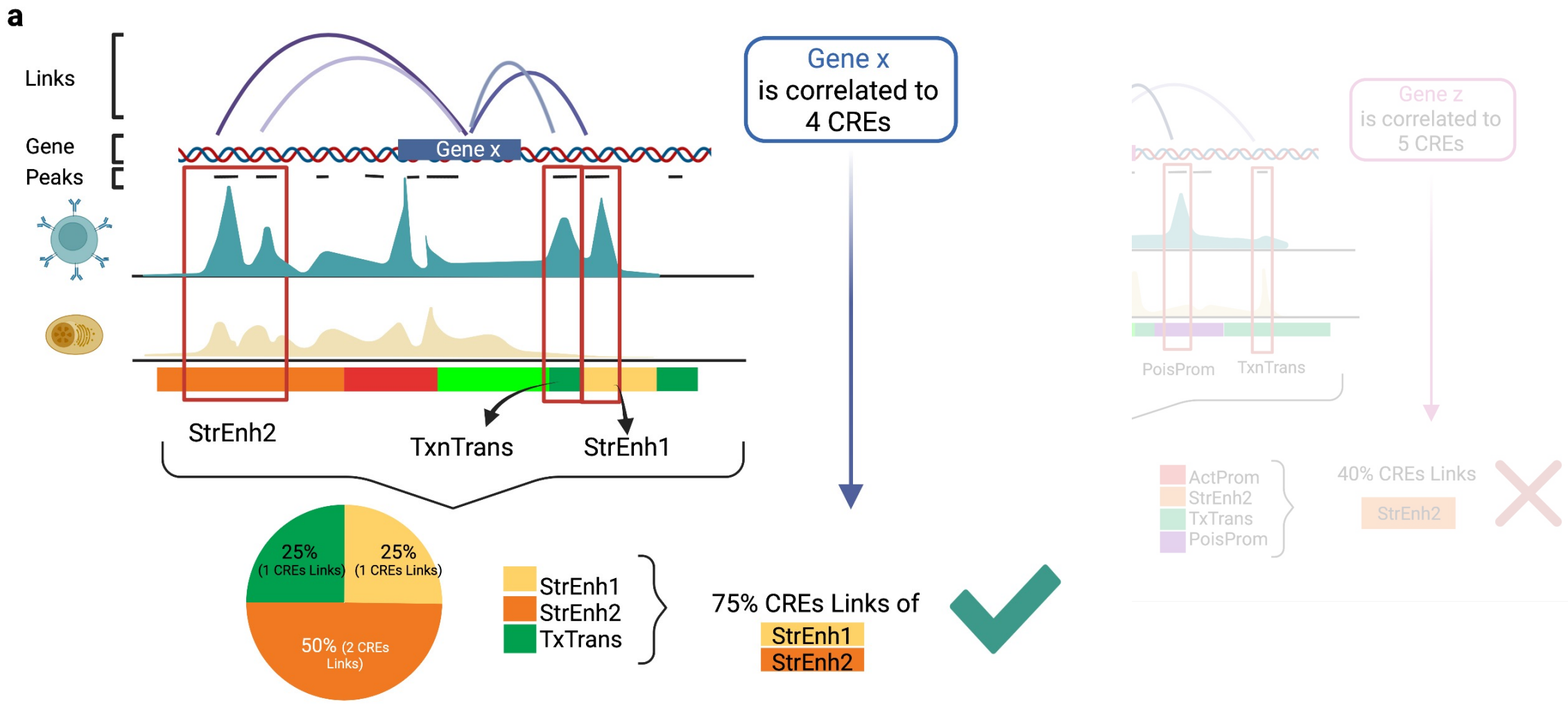
Peak-gene Links filtering & classification by chromatin states



CREs identification of target B-cell genes

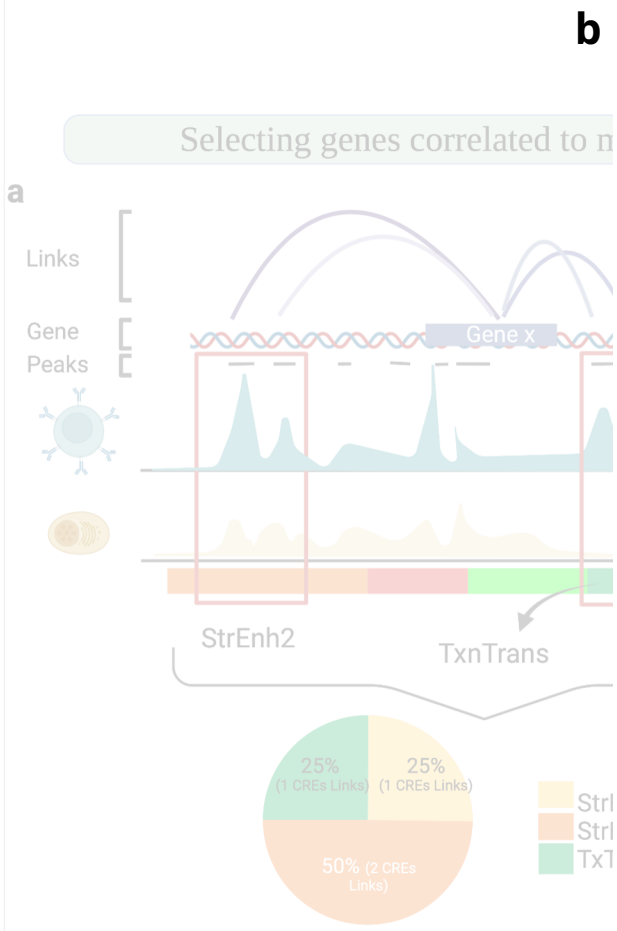
Peak-gene Links filtering & classification by chromatin states

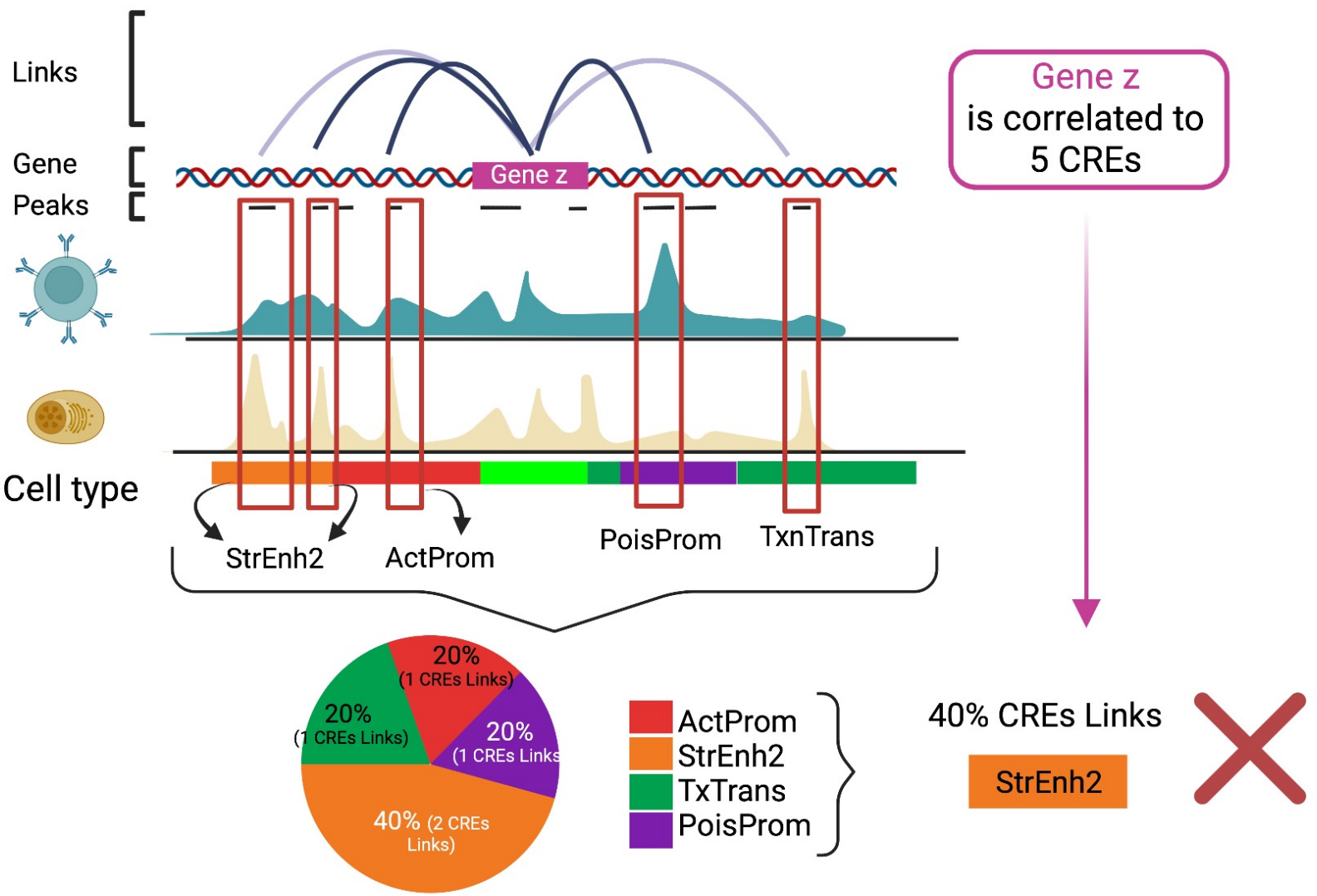
Selecting genes correlated to more than 65% of CREs



CREs identification of target B-cell genes

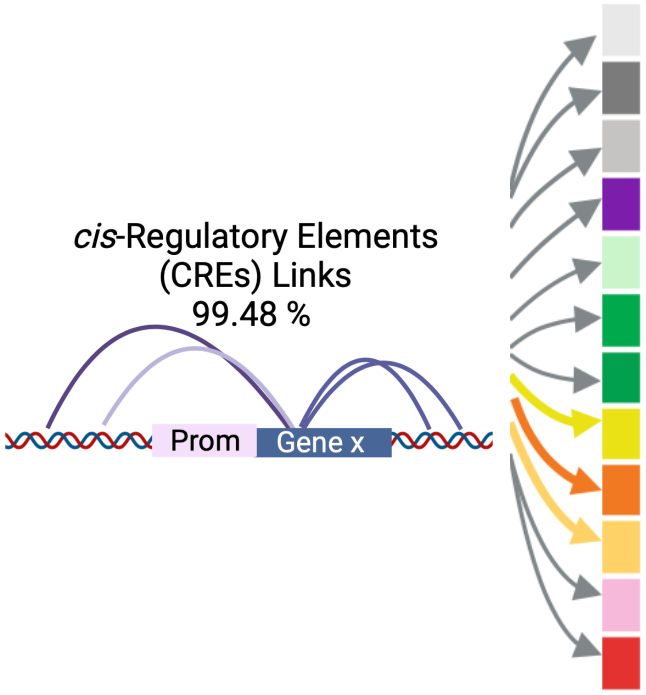
Peak-gene Links filtering & classification by chromatin states

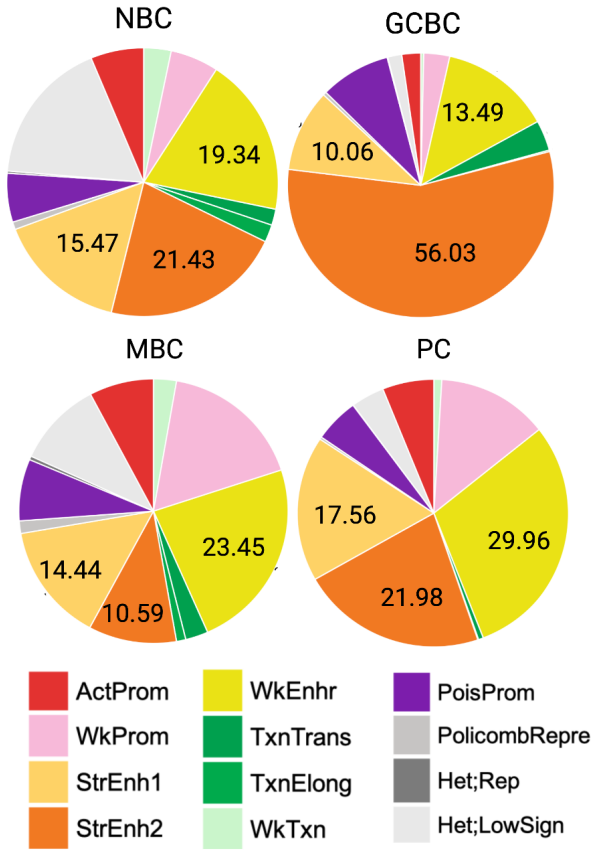




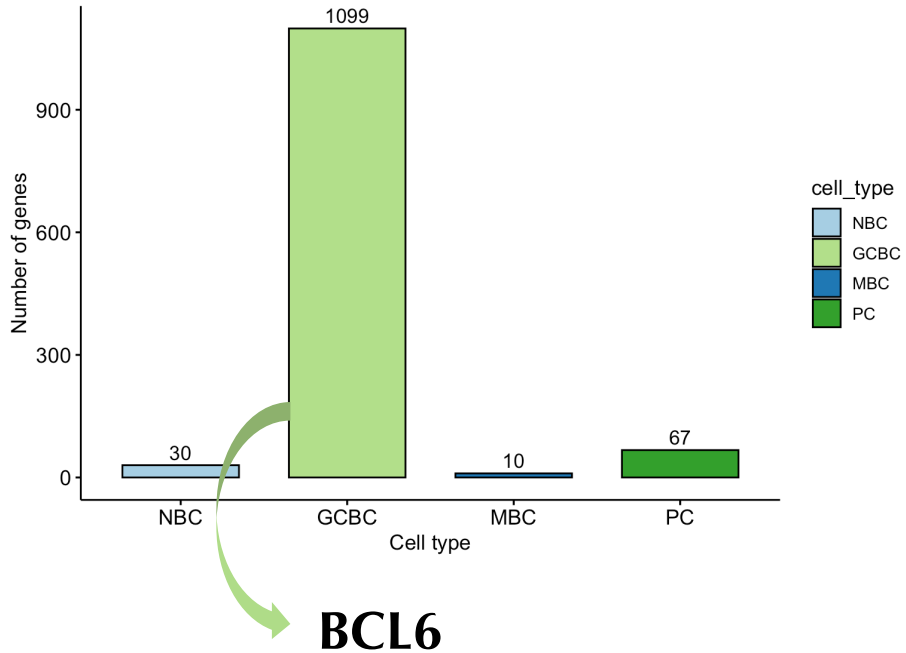
CREs identification of target B-cell genes

Peak-gene Links filtering & classification by chromatin states





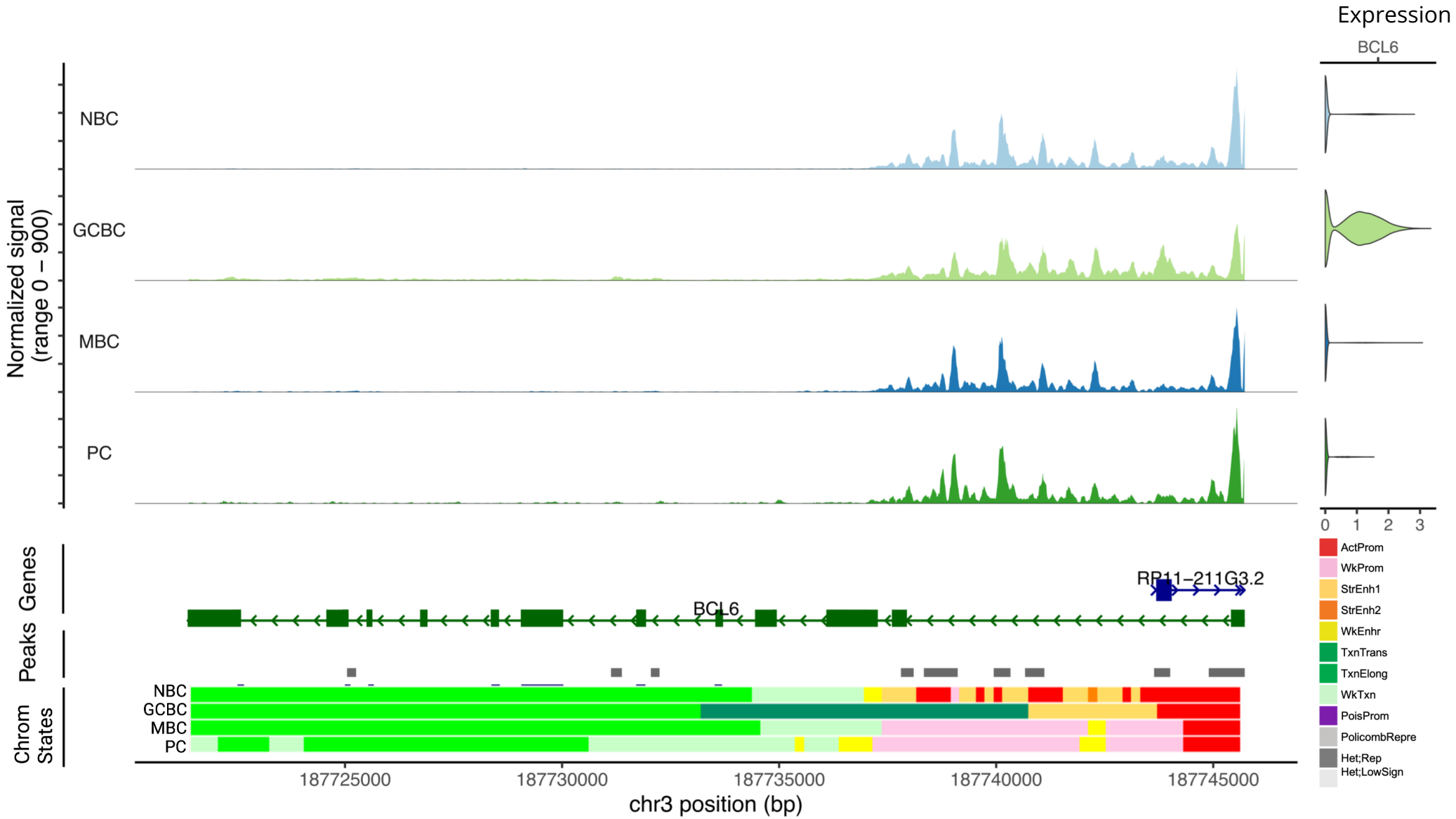
Genes with > 65% of CREs Links in
Strong Enhancer 1 and
Strong Enhancer 2



BCL6

CREs identification of BCL6 gene

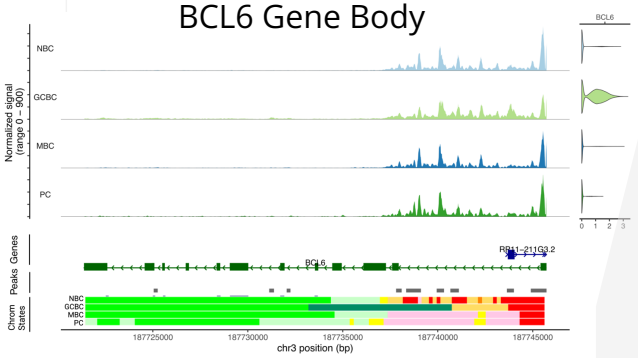
Links within the gene of BLC6

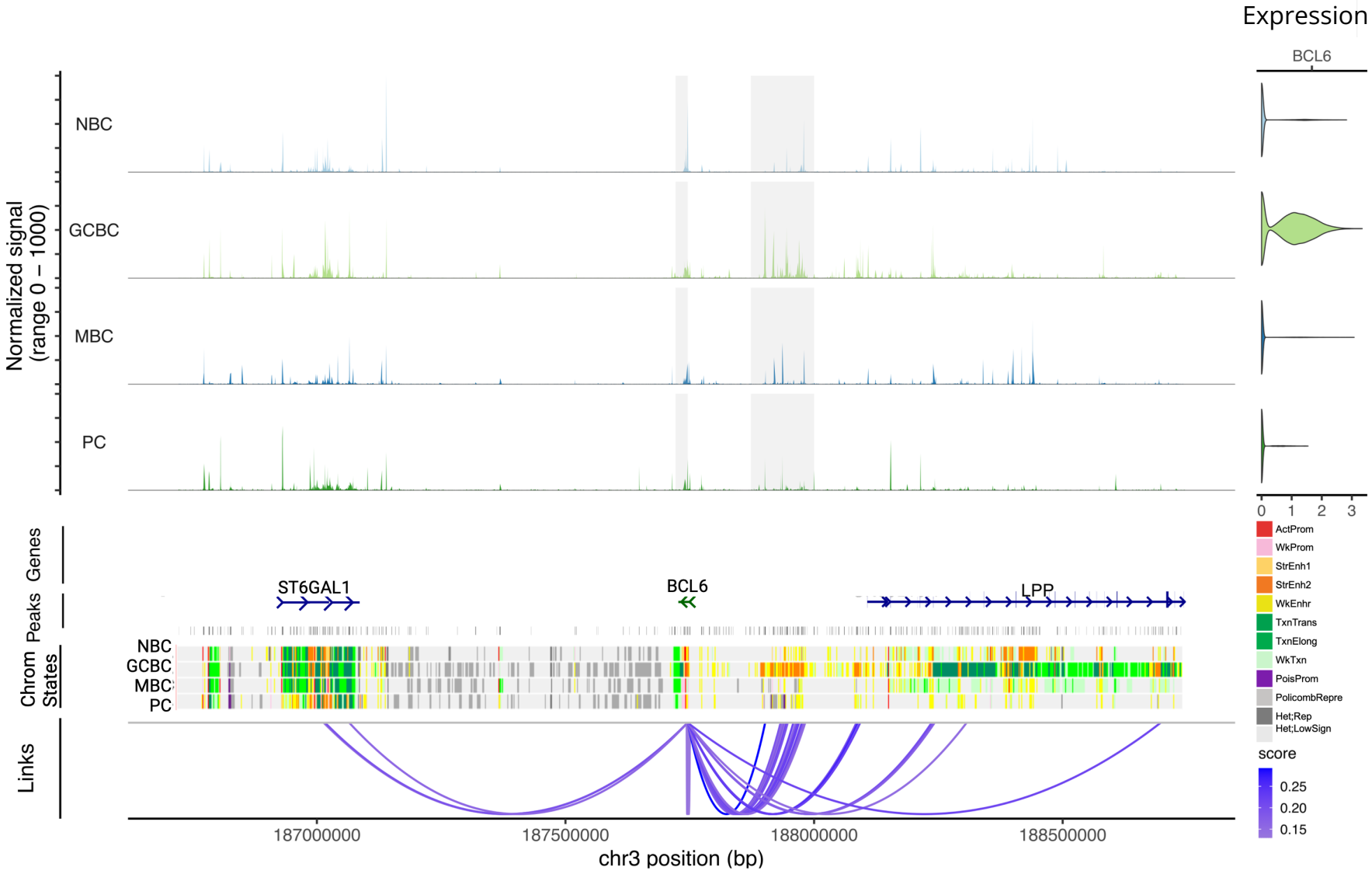


CREs identification of BCL6 gene

Links from BCL6 at 1e+6 bp upstream and downstream of the TSS

BCL6 Gene Body

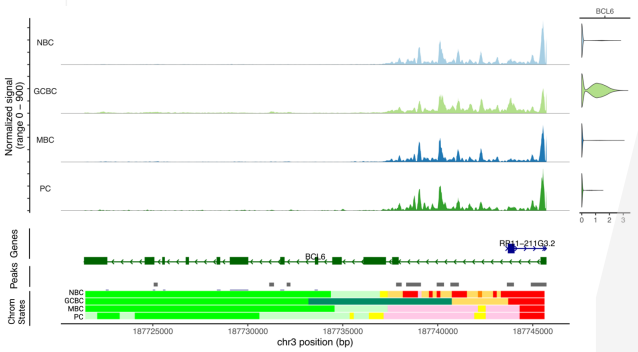


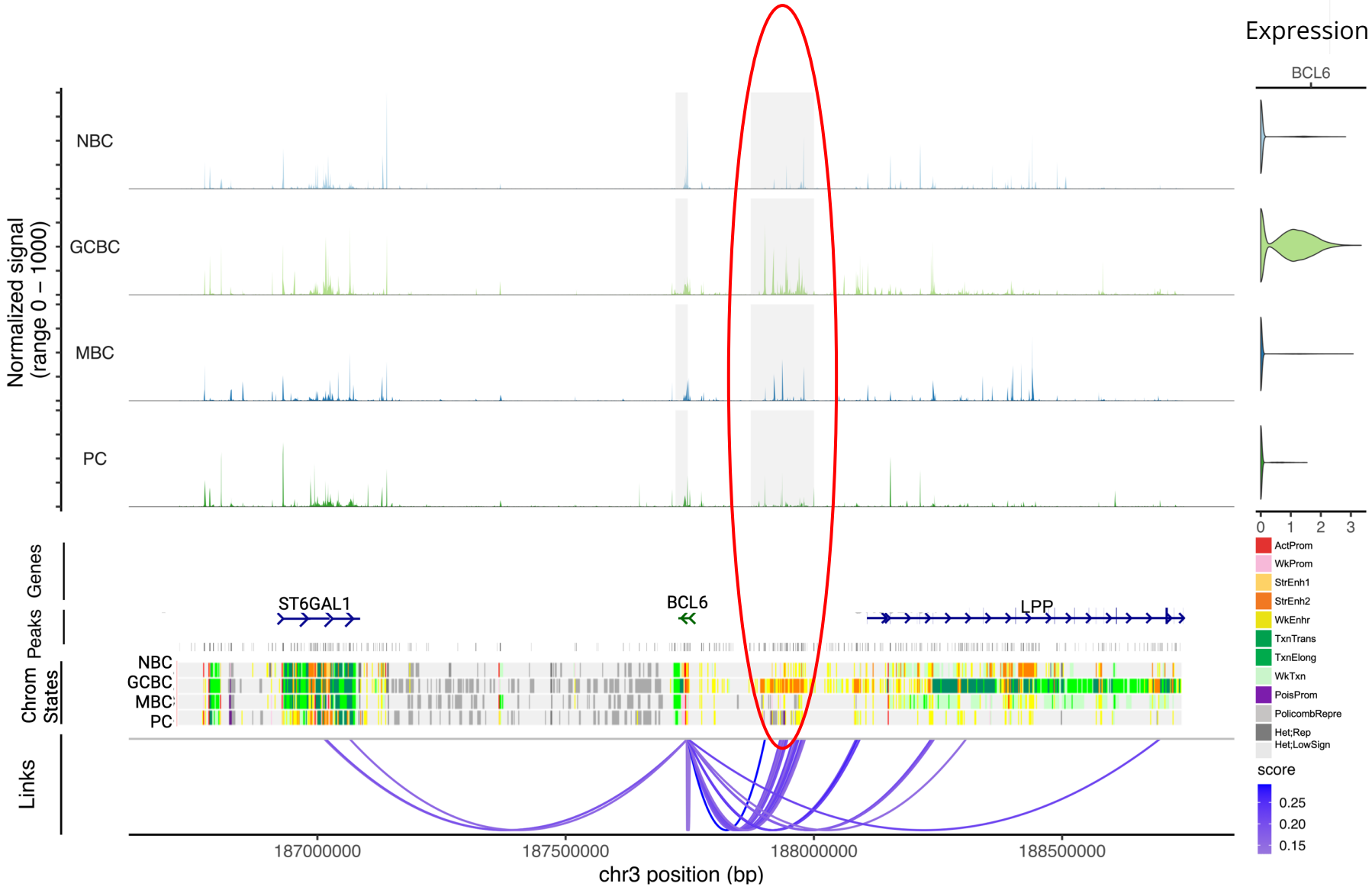


CREs identification of BCL6 gene

Links from BCL6 at 1e+6 bp upstream and downstream of the TSS

BCL6 Gene Body

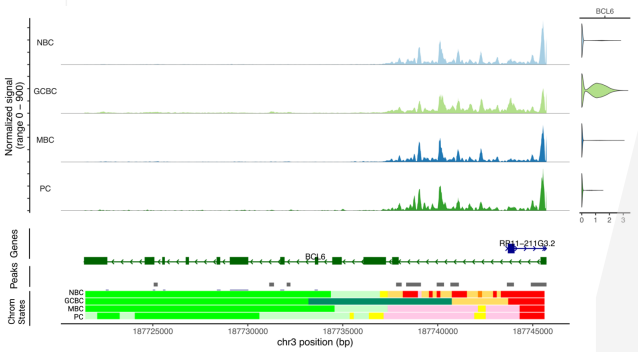


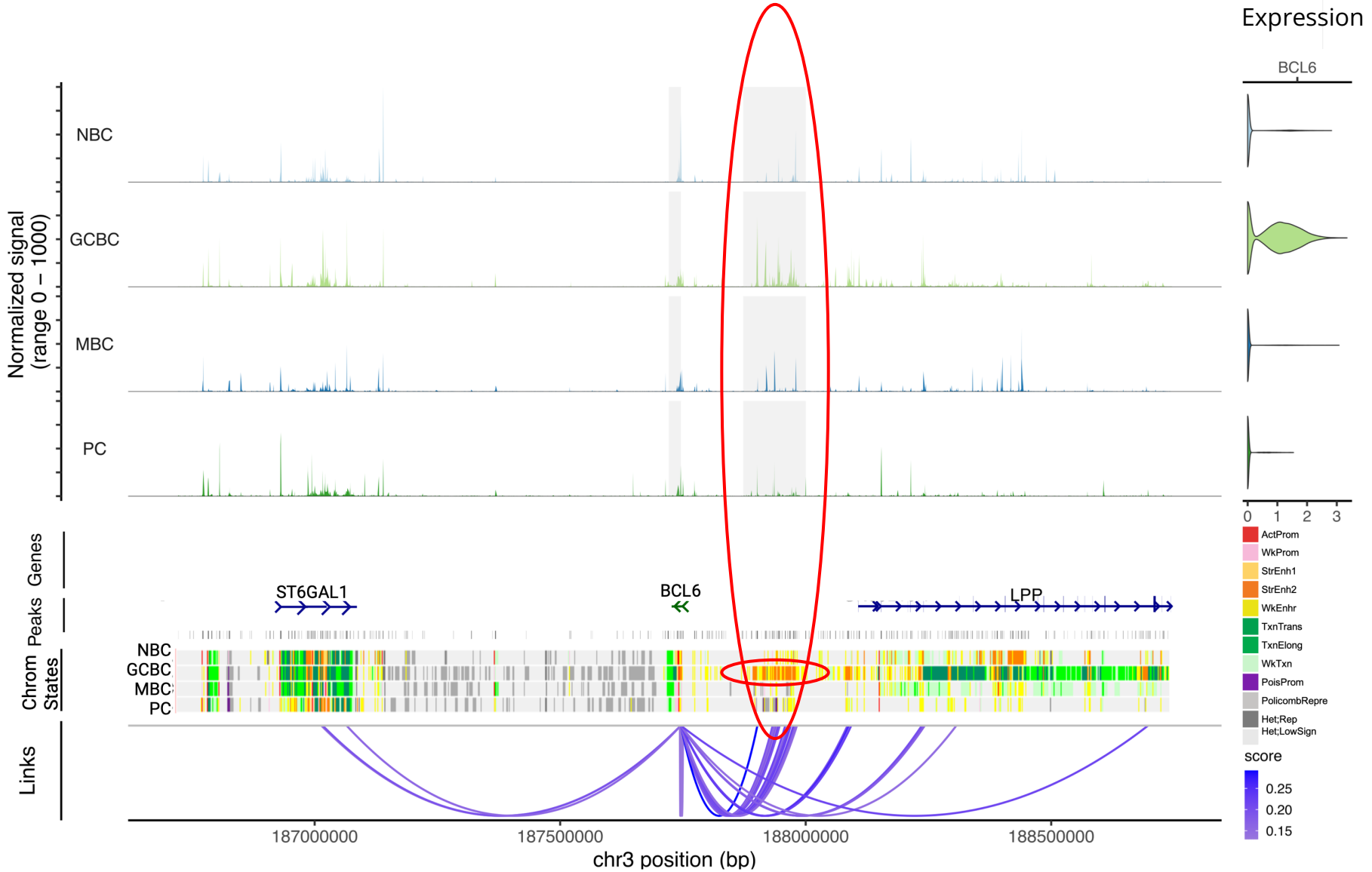


CREs identification of BCL6 gene

Links from BCL6 at 1e+6 bp upstream and downstream of the TSS

BCL6 Gene Body

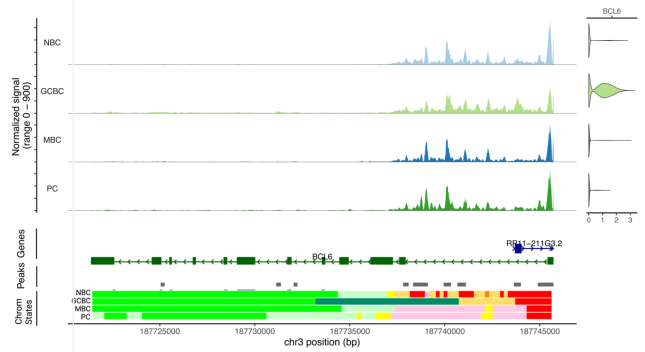


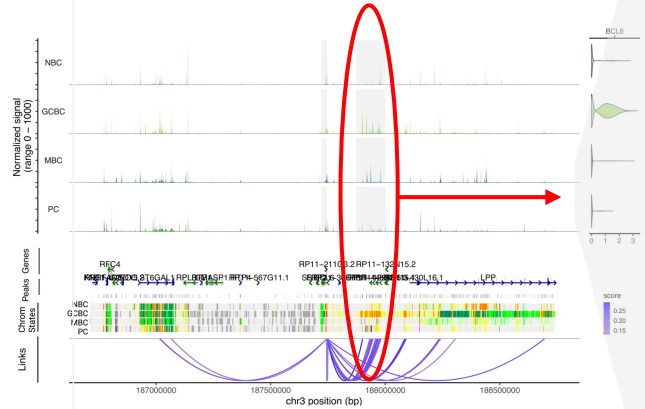


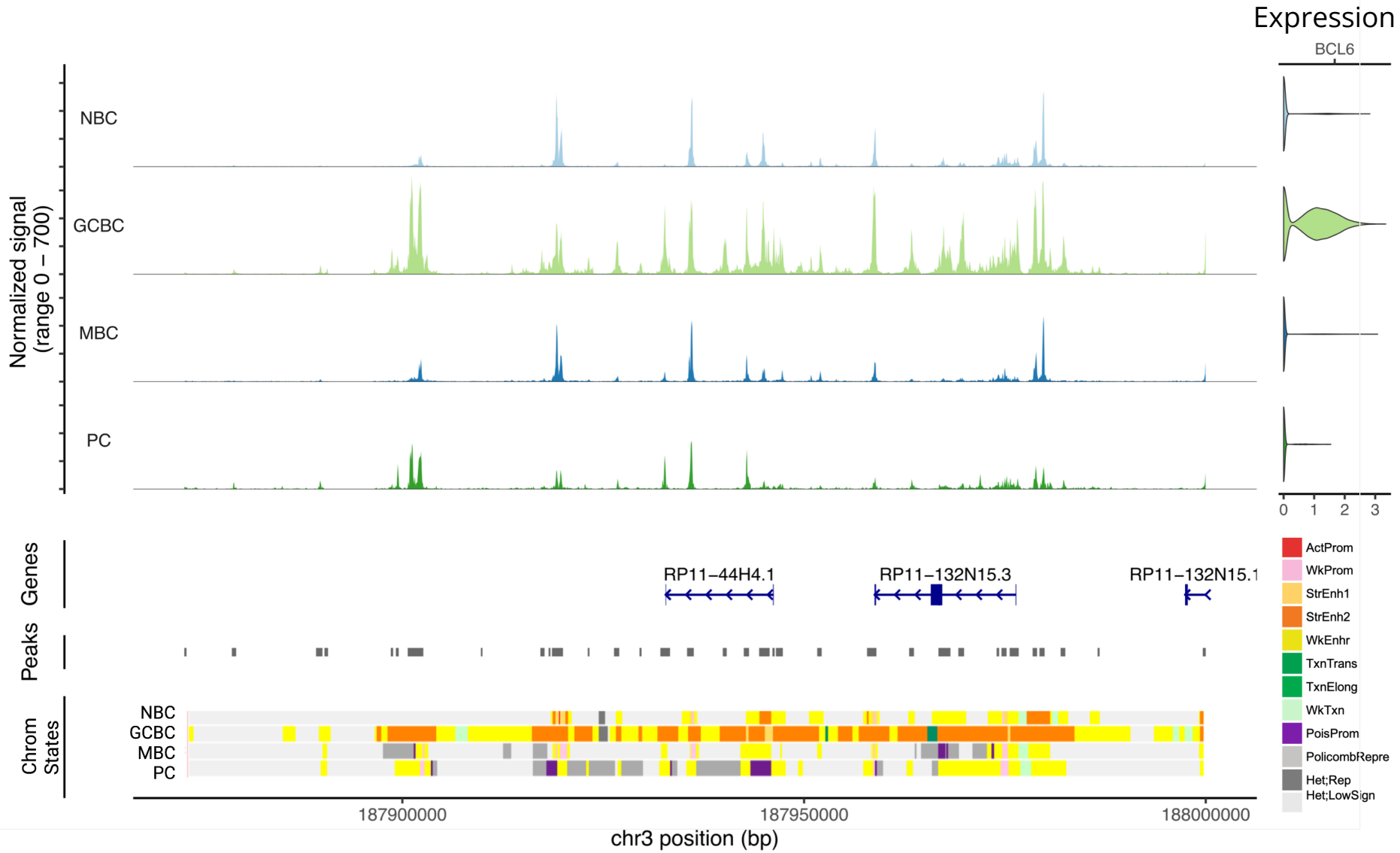
CREs identification of BCL6 gene

BCL6 Potential Super-enhacer

BCL6 Gene Body







Conclusions

Level of ATAC:

- Study the **single cell epigenetic states**:
  - Providing information on chromatin accessibility as an additional level of information beside gene expression.
  - Possibility to make indirect correlation between open chromatin and gene expression
- Increase our understanding of complex **regulatory networks.**
- Reveal **regulatory patterns** that are complementary to gene expression

Level of Multiome:

- Correlate gene expression and chromatin accessibility of the very same cell (**Peak-Gene Linkage analysis**) without inferring both methodology information.
- Discriminate cell type from distinct perspective, and avoid biases related to strong gene expression signatures, such as proliferation signatures (i.e. Germinal Center).
- Discover cells with **similar transcriptional** profiles but functionally **different chromatin landscapes (i.e. MBC and NBC)**
- Discover new potential ***cis*-regulatory elements (CREs)** of target genes involved in cell differentiation, development and disease (Need use chromatin state information)

Conclusions

Limitations in ATAC:

- Limited cell subtype discriminatory power.
- More sensitive to number of principal component in lineal dimensionality reduction (LSI). (The first component correlate with sequencing depth).
- Difficulties to combine with gene expression data

Limitation in Multiome:

- Clustering tend to be more enriched by scRNA-seq, due to the higher power of cell subtype differentiation than scATAC-seq.
- Gene-peak linkages is based on the correlation of chromatin accessibility and gene expression but we need more data to filter links
- We need the histone marks bulk data to validate potential cis-Regulatory elements (CREs).
- It needs more data input than each methodology analysis.
- In each iteration of reprocessing data we usually find new clusters enriched by one methodology. It is not easy to be sure that is a biological or a technical signal.

in the following session

Homeworks

You will find 2 pipeline vignette we you will use a data set of 2 tonsil Multiome data and you will see all the functions and outputs to perform step by step all the downstream analysis of Multiome data .

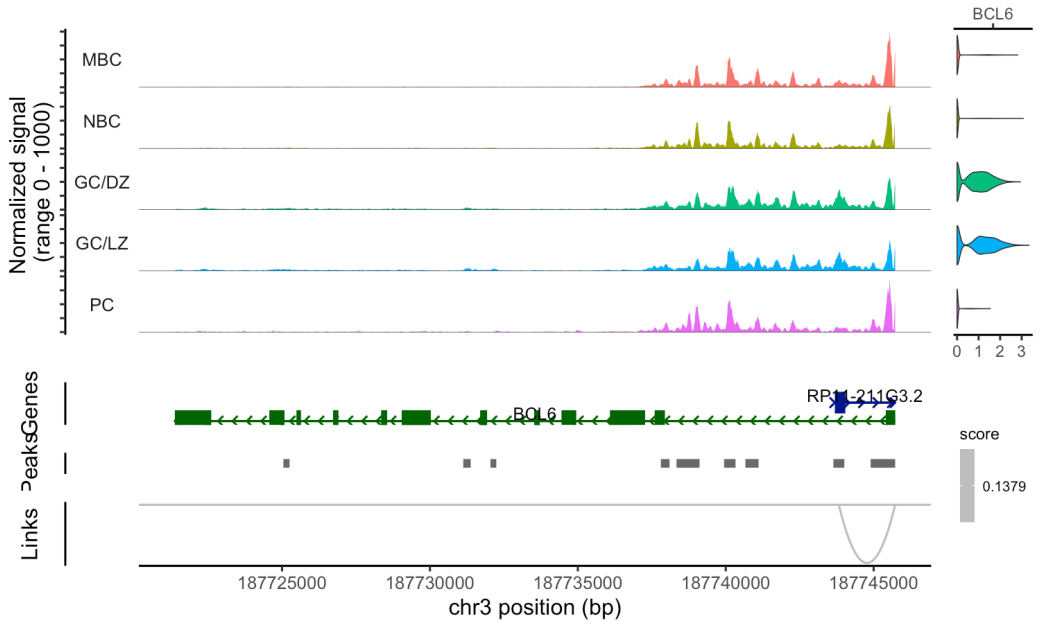
- **Seminar_II_Clustering_PeakCalling_integration.Rmd:** You will perform the clustering, main B cell subtype identification, the subsetting of B-cell clusters and how to reprocess the data in order to do the integration by WNN. You will also perform the peak calling ayou will perfom the quality control and visualization of the results, the filtering step and the merging of both methodologies (scATAC-seq and scRNA-seq)
- **Seminar_II_GenePeakLinkages .Rmd :** You will perform the Gene-Peak Linkages analysis of the B –cell data and to localize que

For your attention

Link Peaks

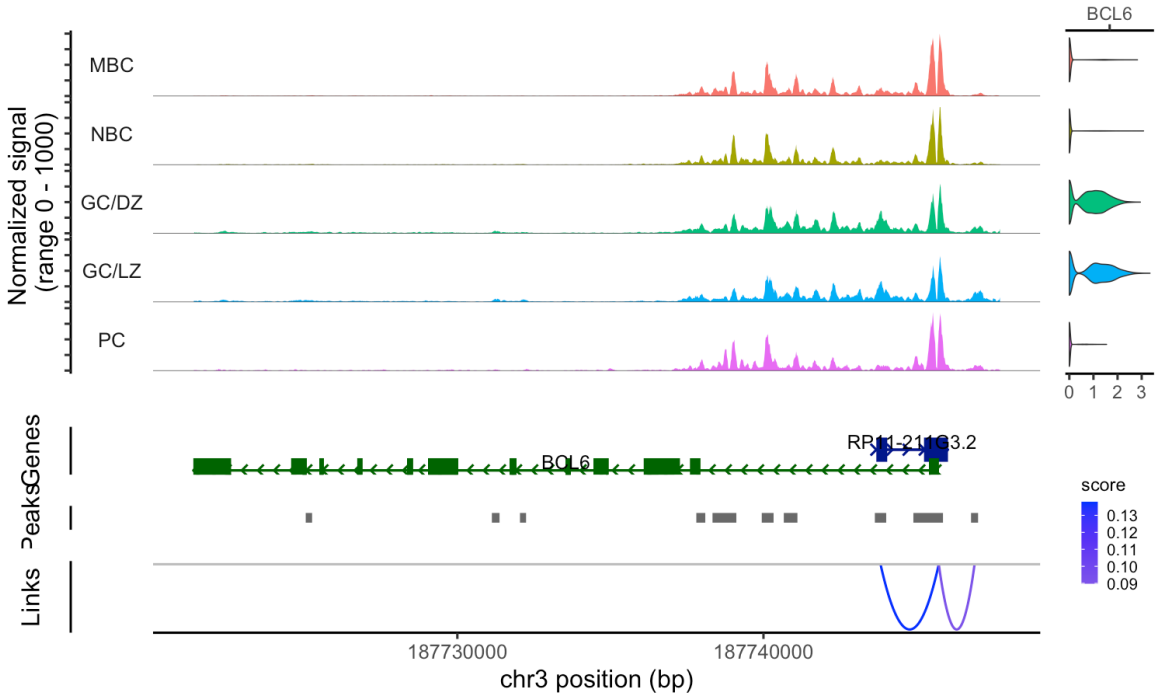
BCL6

Gene Body



Promoter

2000 bp upstream



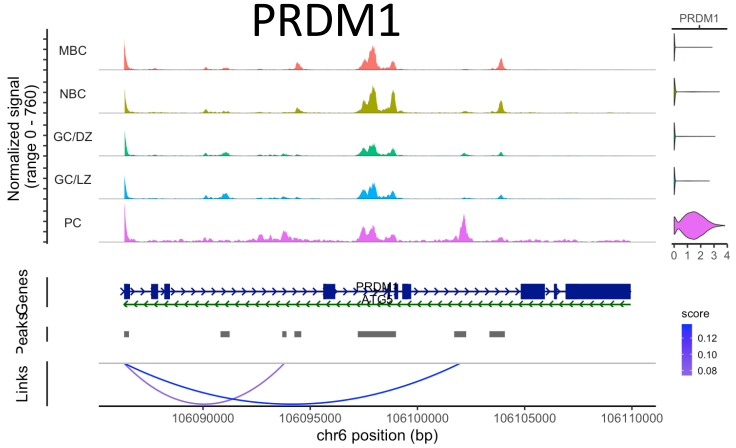
Link Peaks

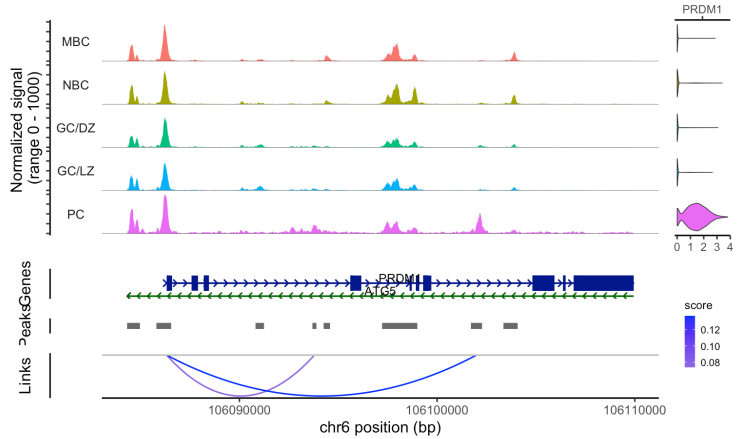
Gene Body

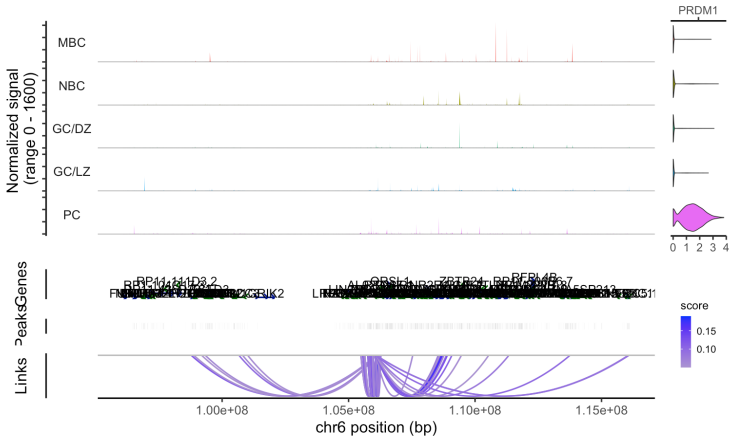
Promoter
2000 bp upstream

CREs
1e+7 bp up and downstream

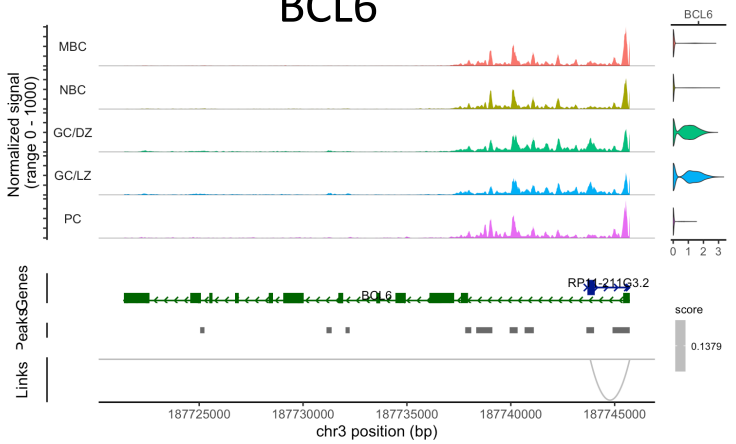
PRDM1

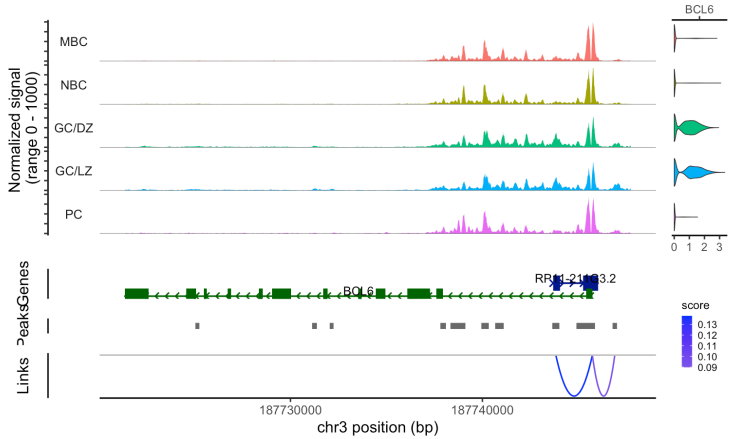


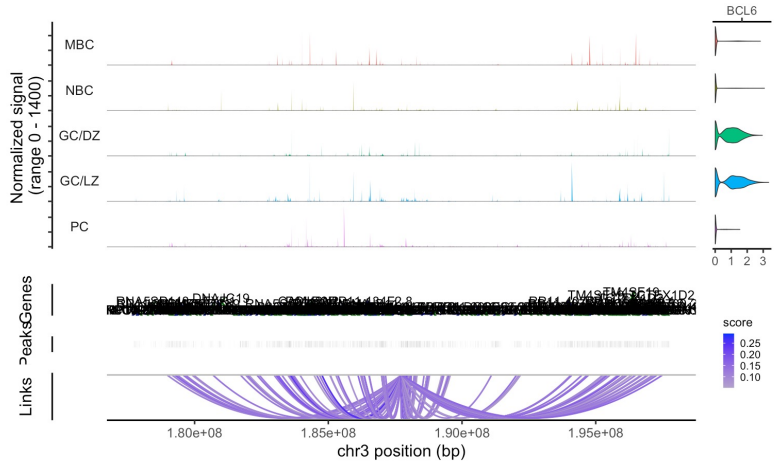




BCL6



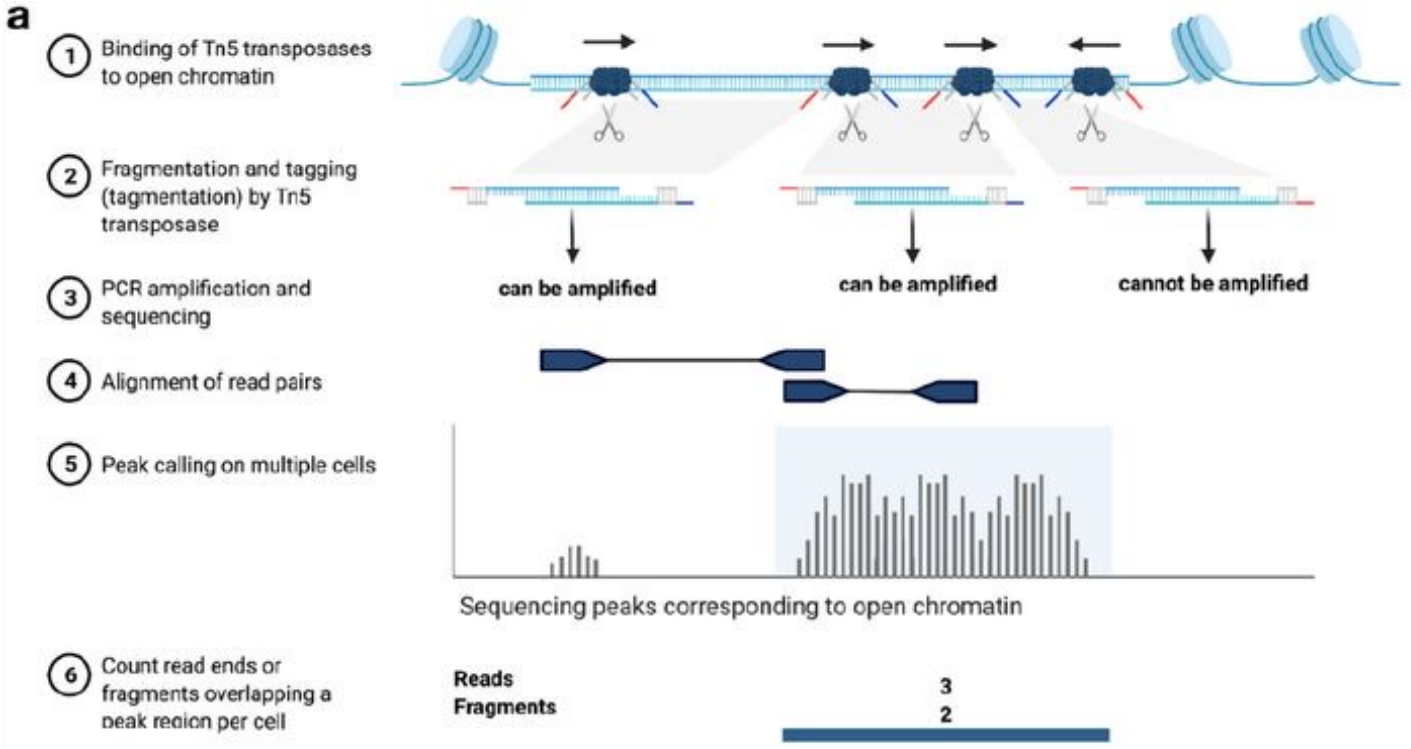




Peak calling

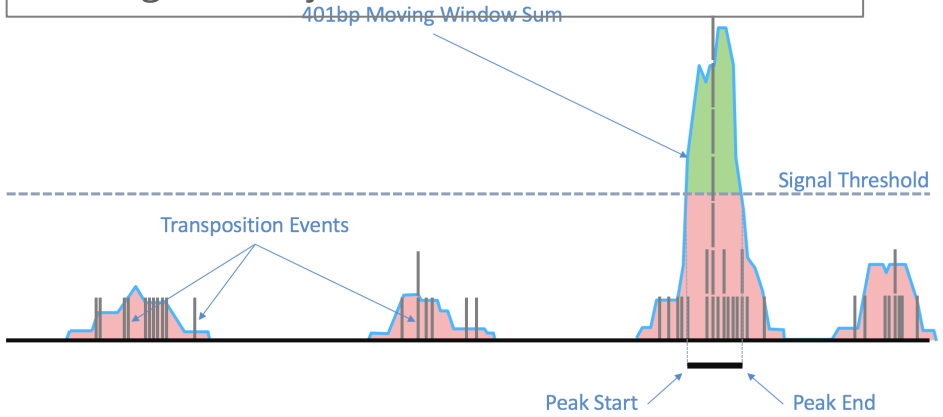
Peak calling is one of the most common analyses to identify areas in the genome that have been enriched with aligned reads to identify **differentially accessible regions (DARs) in a data set.**

The goal of the peak calling algorithm in the scATAC assay is to identify which distinct regions of the genome, known as peaks (open chromatin), are the key features of interest.



Martens, Laura D. 2022, bioRxiv 2022.05.04.490536

- ***'CallPeaks'*** function of Signac uses MACS2 algorithm, which is a count we will re-quantify the counts for each data set and it will create a new assay storing the Fragment object for each dataset.



10x Genomics- Cell Ranger ARC- more info:
https://support.10xgenomics.com/single-cell-multiome-atac-gex/software/pipelines/latest/algorithms/overview#atac_peaks

B-Cell type identification

Curated list of bibliographic biomarkers