Transcription Factor Activity Analysis Using Multi-Omic Single Cell Data
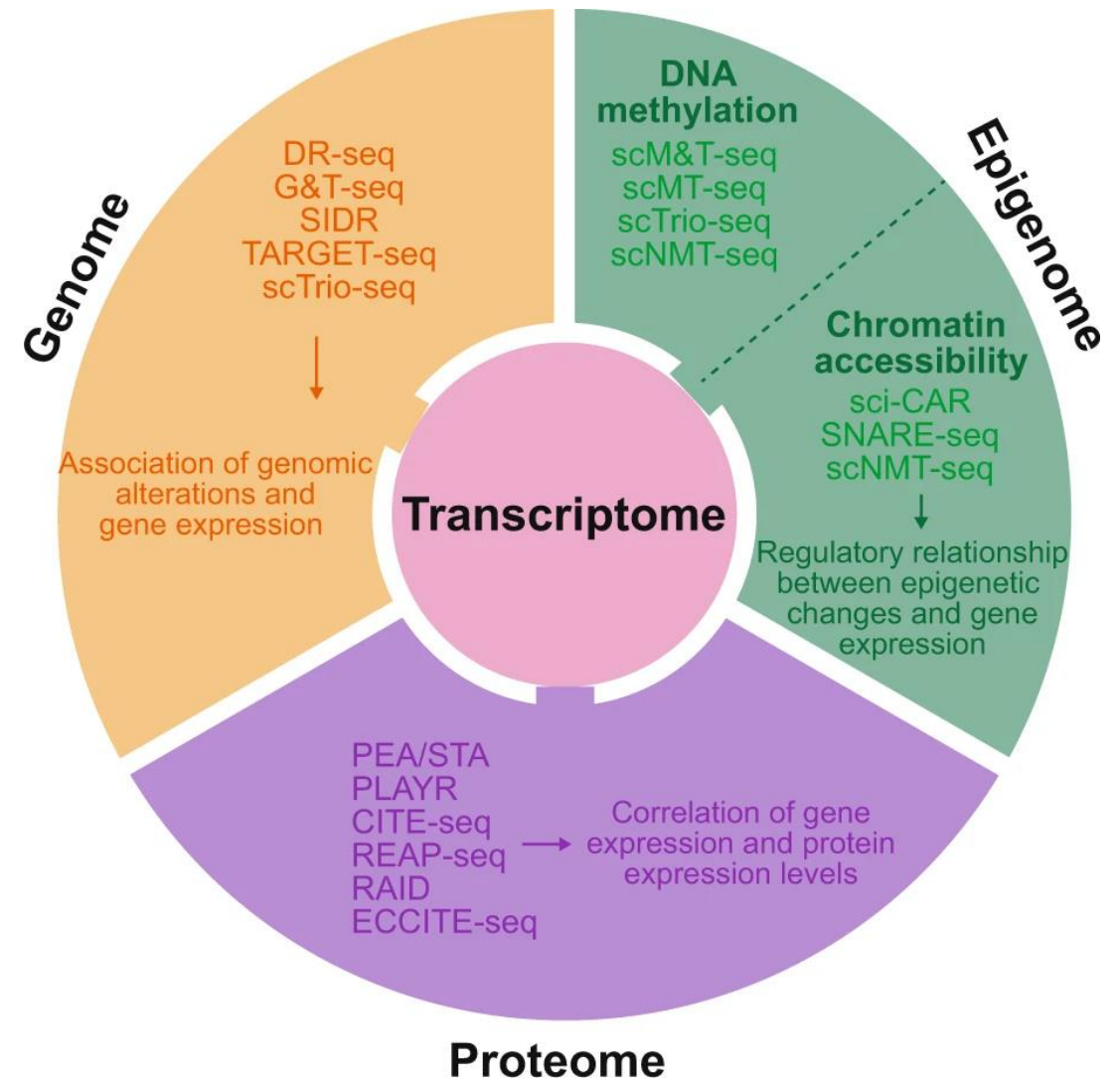
BMI 585 Final Project
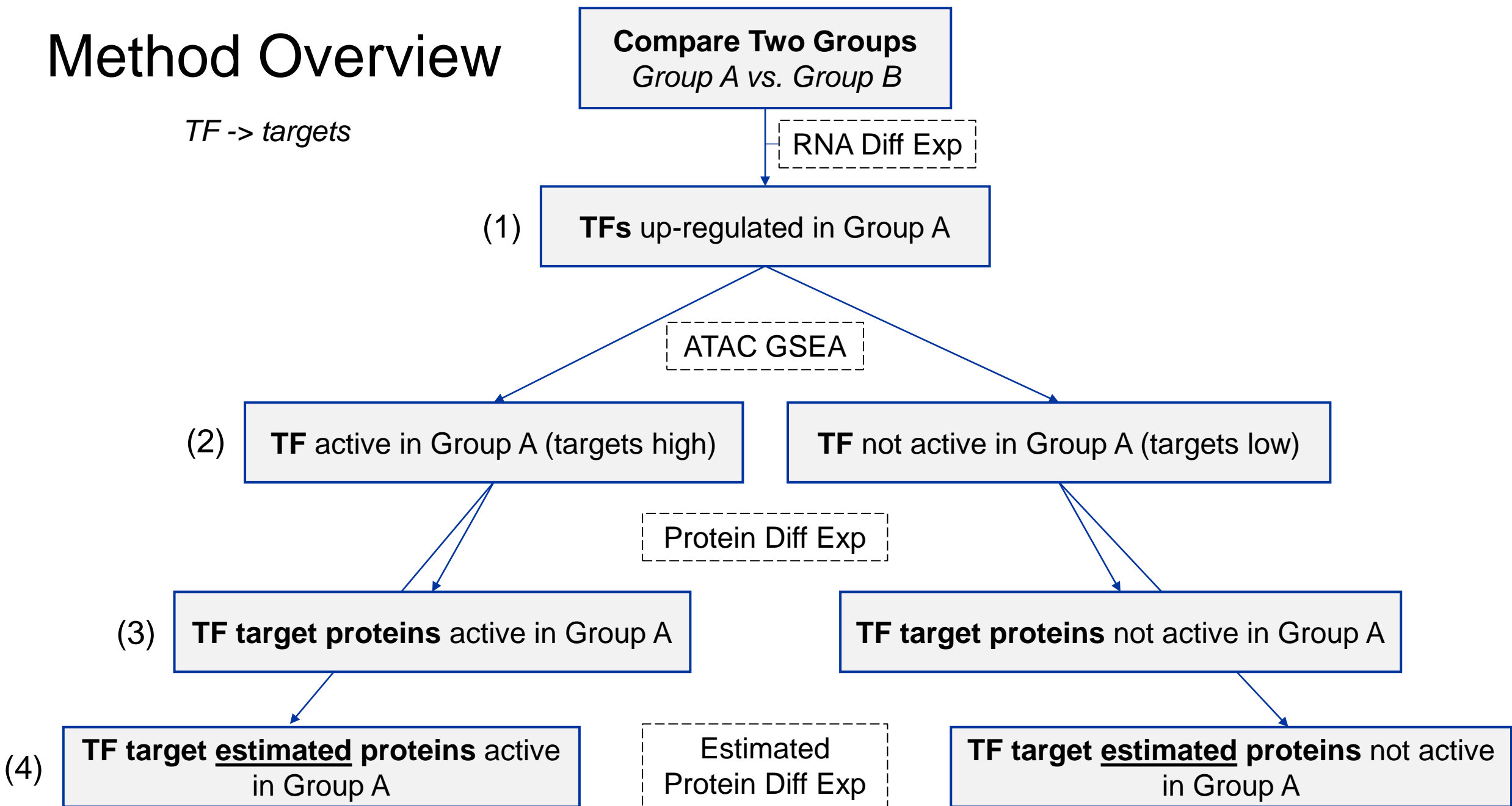
Hope Mumme

Background

Integration of genomic, epigenomic, or proteomic data with transcriptomic data:

- Association of genomic alterations + gene expression (mRNA-genome)
- Regulatory relationship between epigenetic changes and gene expression (mRNA-epigenome)
- Correlation of gene expression and protein expression levels (mRNA-proteome)

[1] Single-cell multiomics: technologies and data analysis methods (Lee, Hyeon, and Hwang – 2020 Experimental & Molecular Medicine)

Method Overview

TF -> targets

Compare Two Groups
Group A vs. Group B

RNA Diff Exp

(1) **TFs** up-regulated in Group A

ATAC GSEA

(2) **TF** active in Group A (targets high)

TF not active in Group A (targets low)

Protein Diff Exp

(3) **TF target proteins** active in Group A

TF target proteins not active in Group A

Estimated Protein Diff Exp

(4) **TF target <u>estimated</u> proteins** active in Group A

TF target <u>estimated</u> proteins not active in Group A

TF Dataset

ENCODE Transcription Factor Targets Dataset:

https://maayanlab.cloud/Harmonizome/dataset/ENCODE+Transcription+Factor+Targets

Contains 181 TFs and all known gene targets,

1,651,393 gene-transcription factor associations

CITE/ATAC Dataset

Mixed Phenotype Acute Leukemia (MPAL) bone marrow from adults at diagnosis (CITE-seq and scATAC-seq)

- 2 T-Myeloid MPAL

Healthy Adult bone marrow (CITE-seq and scATAC-seq)

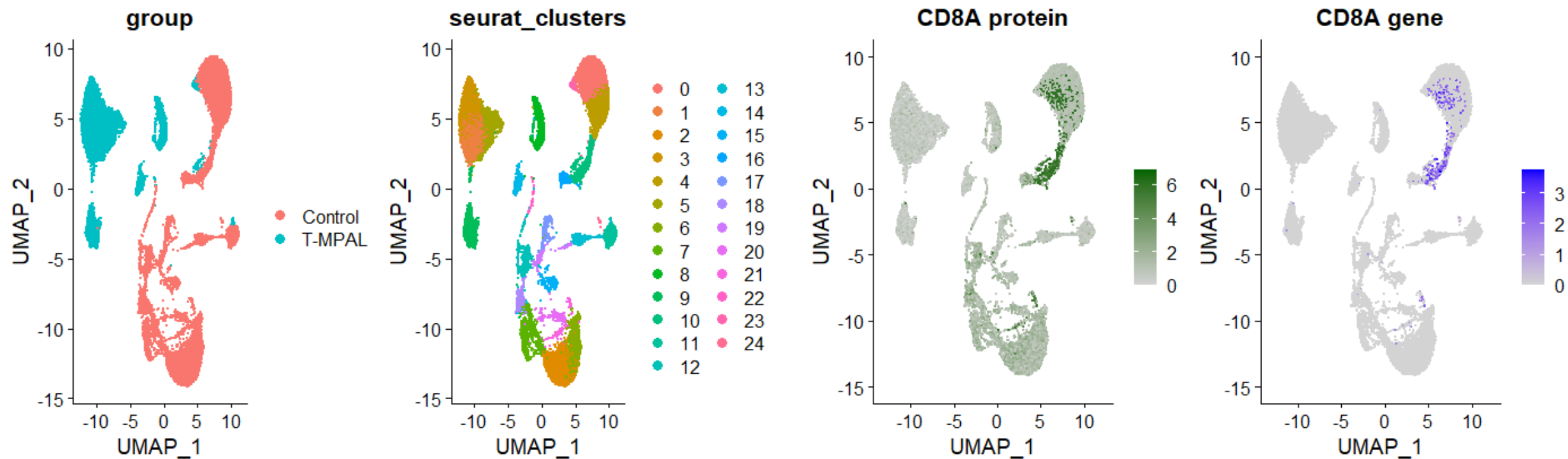- 2 samples

data from GSE139369 by Granja et al [2]

Pre-Processing

- Build Seurat objects using CITE-seq data (scADT + scRNA)
  - ADT assay
  - RNA assay
- Build ArrowProject (ArchR) using scATAC-seq fragment files
  - ATAC assay

Filtering, Normalization, Clustering (RNA-ADT)

Differentially expressed genes and gene activity regions between T-MPAL and Control groups
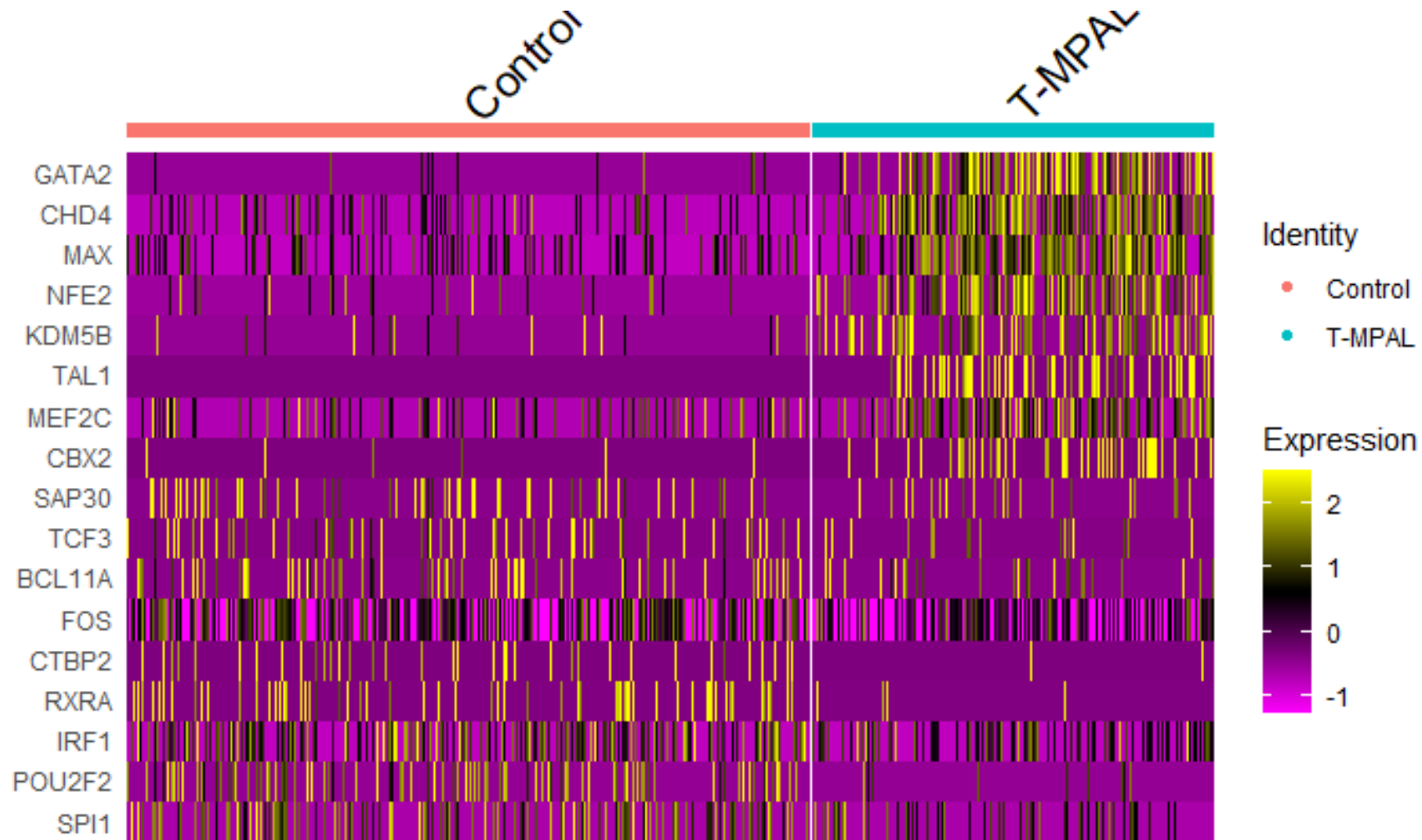
Pre-Processing

Analysis

(1) Identify top up/down regulated TFs for T-MPAL

tfMap(seuratObject, rnaMarkers, tf_list, type = up)
DoHeatmap(obj, features = unlist(topTFs), group.by = group)

Analysis

```
tfs = getTFs(obj, rnaMarkers, db, num = all)[[T-MPAL]]
db.filt = db[names(db) %in% tfs]
en = tfEnrichment(atacMarkers[[T-MPAL]], db.filt)
head(en)
```

pathway <chr>	pval <dbl>	padj <dbl>	log2err <dbl>	ES <dbl>	NES <dbl>	size <int>	leadingEdge <list>
ATF3	9.616904e-37	4.039100e-36	1.5763736	-0.2618345	-2.156421	1977	<chr [1,341]>
BATF	6.596011e-26	2.308604e-25	1.3188888	-0.2846389	-2.250415	1085	<chr [647]>
BDP1	1.234257e-01	1.619962e-01	0.2020717	-0.3139491	-1.320581	29	<chr [25]>
CBX2	3.786982e-01	4.678037e-01	0.1752040	-0.1403863	-1.031677	406	<chr [94]>
CEBPZ	6.007067e-01	7.008245e-01	0.1009906	-0.1465165	-0.945335	161	<chr [107]>
CHD1	NA	NA	NA	-0.1582863	NA	4806	<chr [2,052]>

6 rows

```
posTFs = en %>% filter(NES > 0 & padj < 0.05) %>% select(pathway)
negTFs = en %>% filter(NES < 0 & padj < 0.05) %>% select(pathway)
```

pathway <chr>	NES <dbl>
ATF3	-2.156421
BATF	-2.250415
CHD4	-1.300768
CUX1	-2.251888
FLI1	-1.989872
KDM1A	-1.667133
MAFF	-1.503808
MEF2C	-1.729570
NFE2	-2.211867
SMARCC1	-1.655428

Only significant negatively enriched TF target sets.

Analysis

pathway <chr>	pval <dbl>	padj <dbl>	log2err <dbl>	ES <dbl>	NES <dbl>	size <int>	leadingEdge <list>
ATF3	9.616904e-37	4.039100e-36	1.5763736	-0.2618345	-2.156421	1977	<chr [1,341]>
BATF	6.596011e-26	2.308604e-25	1.3188888	-0.2846389	-2.250415	1085	<chr [647]>
BDP1	1.234257e-01	1.619962e-01	0.2020717	-0.3139491	-1.320581	29	<chr [25]>
CBX2	3.786982e-01	4.678037e-01	0.1752040	-0.1403863	-1.031677	406	<chr [94]>
CEBPZ	6.007067e-01	7.008245e-01	0.1009906	-0.1465165	-0.945335	161	<chr [107]>
CHD1	NA	NA	NA	-0.1582863	NA	4806	<chr [2,052]>

6 rows

Analysis
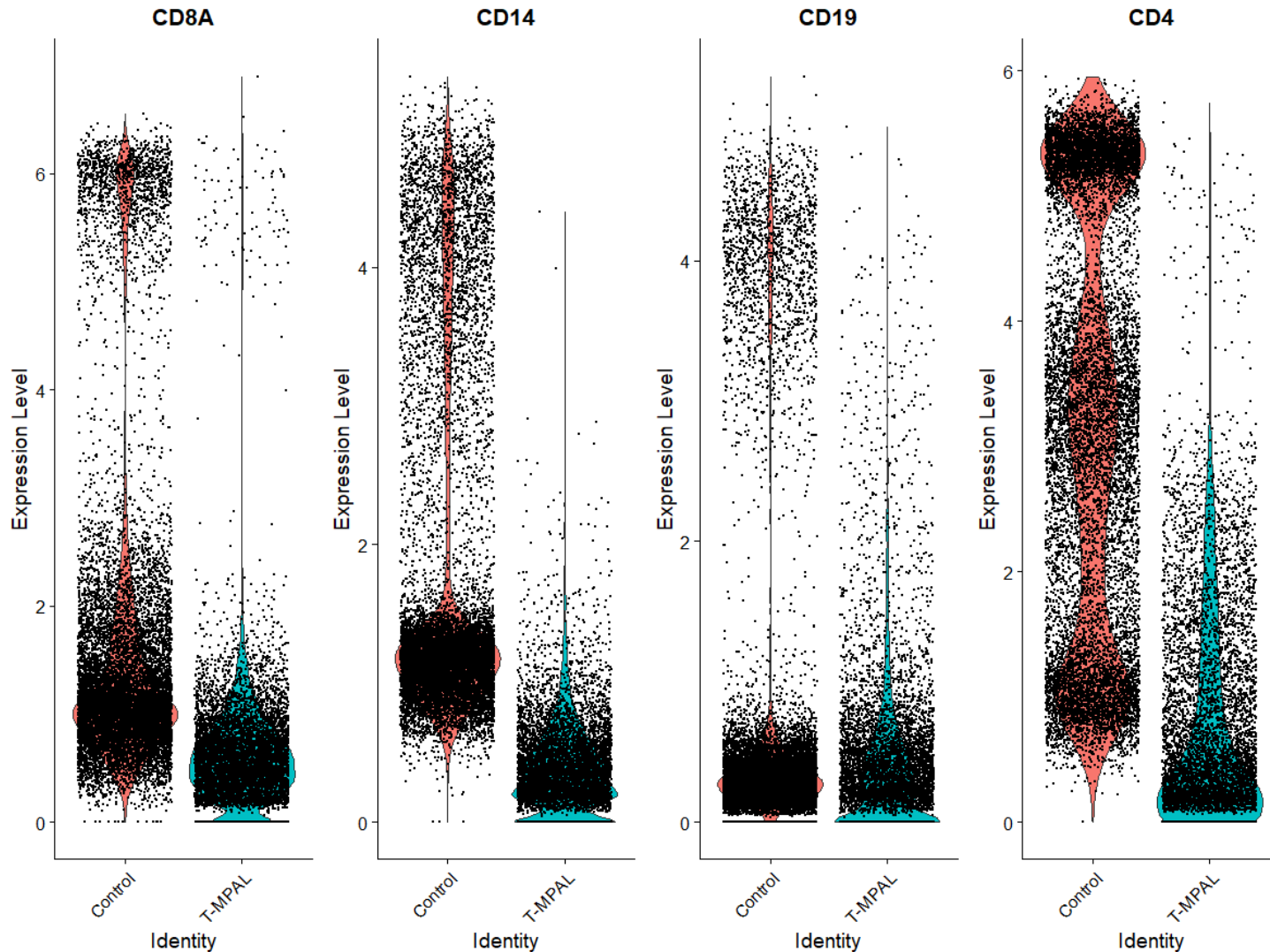
[1] CD8A
 [1] ATF3   CBX2   CHD1   CHD2   CUX1   ELF1   GATA2
GTF2F1 HDAC2 HMGN3 KDM4A MAFF   MAX   MYC
[15] REST   SMC3   STAT5A TAL1   TBL1XR1 WRNIP1 ZKSCAN1
ZMIZ1

[1] CD14
 [1] CHD1   ELF1   HDAC2 MAFF   MAX   MYC   REST SMC3
STAT3 TAL1   TBL1XR1 WRNIP1
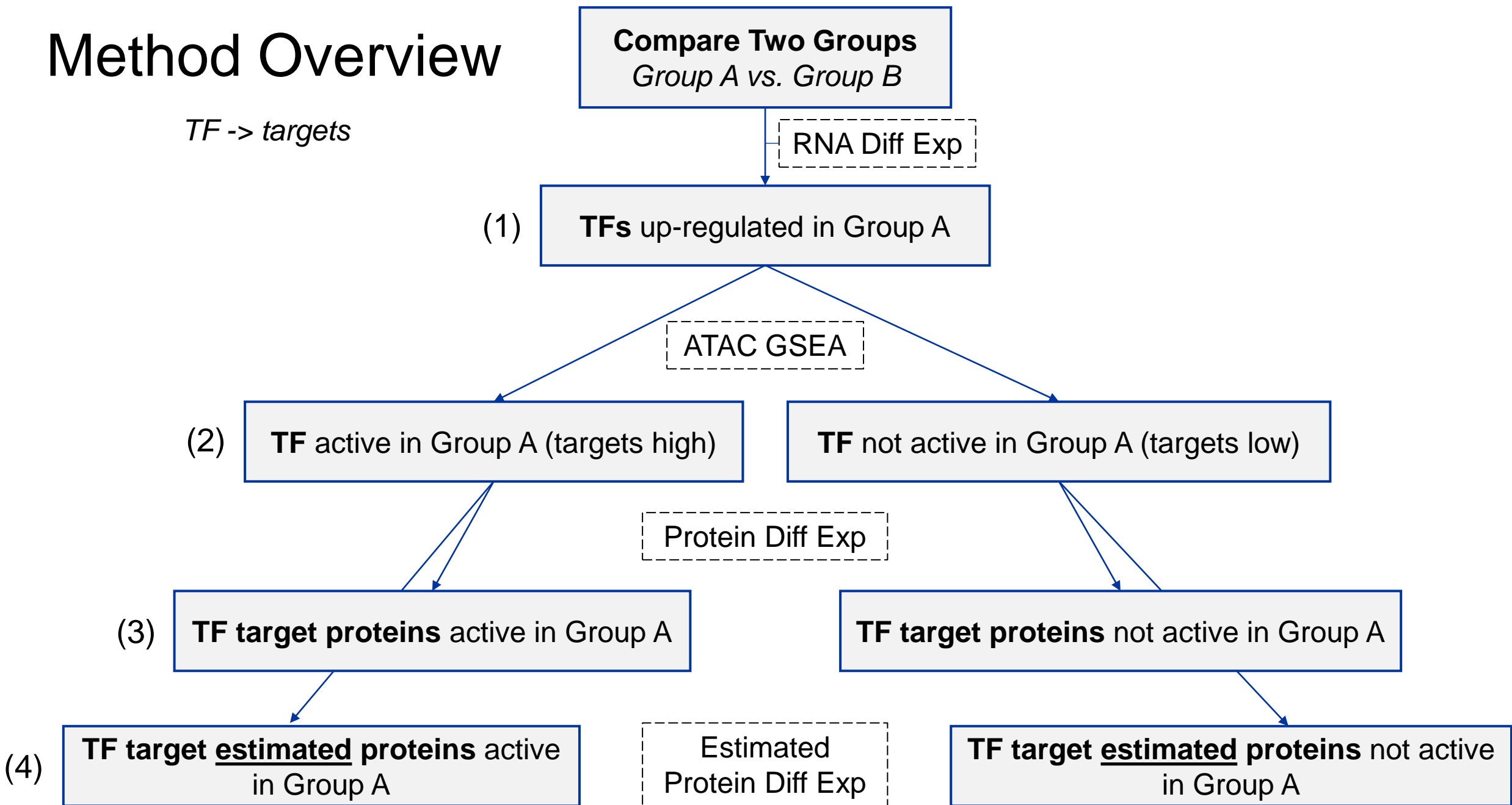
[1] CD19
 [1] CHD1   CHD2   CUX1   ELF1   HDAC2 KDM5B MAX   MYC
REST   SMC3   STAT5A TAL1   TBL1XR1 WRNIP1
[15] ZMIZ1

[1] CD4
[1] CHD1 KDM5A KDM5B MAFF MAX   MYC   REST

Method Overview

TF -> targets

Compare Two Groups
Group A vs. Group B

RNA Diff Exp

(1) **TFs** up-regulated in Group A

ATAC GSEA

(2) **TF** active in Group A (targets high)

TF not active in Group A (targets low)

Protein Diff Exp

(3) **TF target proteins** active in Group A

TF target proteins not active in Group A

Estimated Protein Diff Exp

(4) **TF target <u>estimated</u> proteins** active in Group A

TF target <u>estimated</u> proteins not active in Group A

TF patterns

TF -> targets

(1) **37 TFs** up-regulated in T-MPAL compared to Control

ATAC GSEA

(2) -

15 TFs' targets are less enriched in T-MPAL than in Control

Protein Diff Exp

(3) -

5 TF target proteins down-regulated in T-MPAL compared to Control

Estimated Protein Diff Exp

(4) -

-

TF patterns

TF -> targets

(1) **37 TFs** up-regulated in T-MPAL compared to Control

(2) **15 TFs' targets** are less enriched in T-MPAL than in Control

(3) **5 TF target proteins** down-regulated in T-MPAL compared to Control

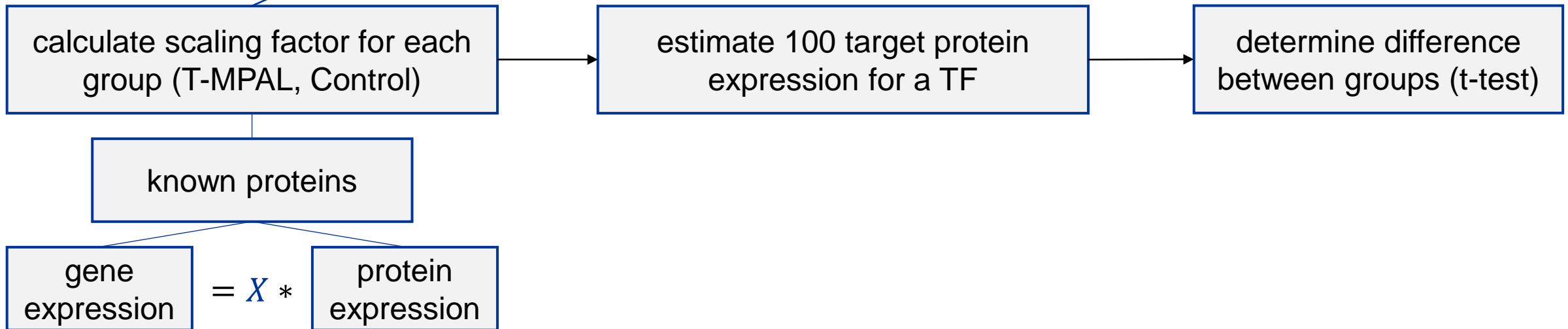(4) **TF target <u>estimated</u> proteins targets** have lower expression in T-MPAL

TFs for 5 target proteins (step 3)
ATF3   CBX2   CHD1   CHD2   CUX1   ELF1   GATA2
GTF2F1   HDAC2   HMGN3   KDM4A   KDM5A   KDM5B
MAFF   MAX   MYC
REST   SMC3   STAT3   STAT5A TAL1   TBL1XR1
WRNIP1 ZKSCAN1 ZMIZ1

Analysis

general formula: $gene_{exp} = X * protein_{exp}$

```
┌─────────────────────────────┐     ┌─────────────────────────────┐     ┌─────────────────────────────┐
│  calculate scaling factor   │     │   estimate 100 target       │     │  determine difference       │
│  for each group             │ ──▶ │   protein expression        │ ──▶ │  between groups (t-test)    │
│  (T-MPAL, Control)          │     │   for a TF                  │     │                             │
└─────────────────────────────┘     └─────────────────────────────┘     └─────────────────────────────┘
```

known proteins

| gene expression | $= X *$ | protein expression |

Analysis

(4) Identify up/down regulated <u>estimated</u> protein targets of TFs in T-MPAL

Analysis

(4) Identify up/down regulated <u>estimated</u> protein targets of TFs in T-MPAL

	tf	high	pval
1	ATF3	T-MPAL	0.284427848040953
2	BATF	T-MPAL	0.585295003037739
3	BDP1	T-MPAL	0.399650074028783
4	CBX2	T-MPAL	0.59595926188664
5	CEBPZ	T-MPAL	0.124757631279817
6	CHD1	T-MPAL	0.241981388120949
7	CHD2	T-MPAL	0.432706330075244
8	CHD4	T-MPAL	0.781212758895344
9	CUX1	Control	0.875679376875177
10	ELF1	T-MPAL	0.223015590755234
11	FLI1	T-MPAL	0.0837432486650766
12	GATA2	T-MPAL	0.500739902794131
13	GTF2F1	T-MPAL	0.248085589172119
14	HDAC2	T-MPAL	0.575673021807084
15	HMGN3	T-MPAL	0.586247576655327
16	KDM1A	T-MPAL	0.280454557205832
17	KDM4A	T-MPAL	0.0561099550135497
18	KDM5A	Control	0.673134904486428
19	KDM5B	T-MPAL	0.168913523538044
20	MAFF	T-MPAL	0.15927565910058
21	MAX	T-MPAL	0.937213276482223
22	MEF2C	T-MPAL	0.797834512842945
23	MYB	T-MPAL	0.0547911258995726
24	MYC	T-MPAL	0.897321965669034
25	NFE2	Control	0.705795721031258
26	REST	T-MPAL	0.394581583936276
27	SMARCA4	T-MPAL	0.679440031307116
28	SMARCC1	T-MPAL	0.692923430324685
29	**SMARCC2**	**T-MPAL**	**0.0459629362102477**
30	SMC3	T-MPAL	0.0659953585436789
31	STAT3	T-MPAL	0.249986995235082
32	STAT5A	Control	0.599413591614296
33	TAL1	T-MPAL	0.487368960307839
34	TBL1XR1	T-MPAL	0.135851522430585
35	WRNIP1	T-MPAL	0.446779079838831
36	ZKSCAN1	T-MPAL	0.57282629752689
37	ZMIZ1	Control	0.980607392815335

SMARCC2 target protein estimated expression
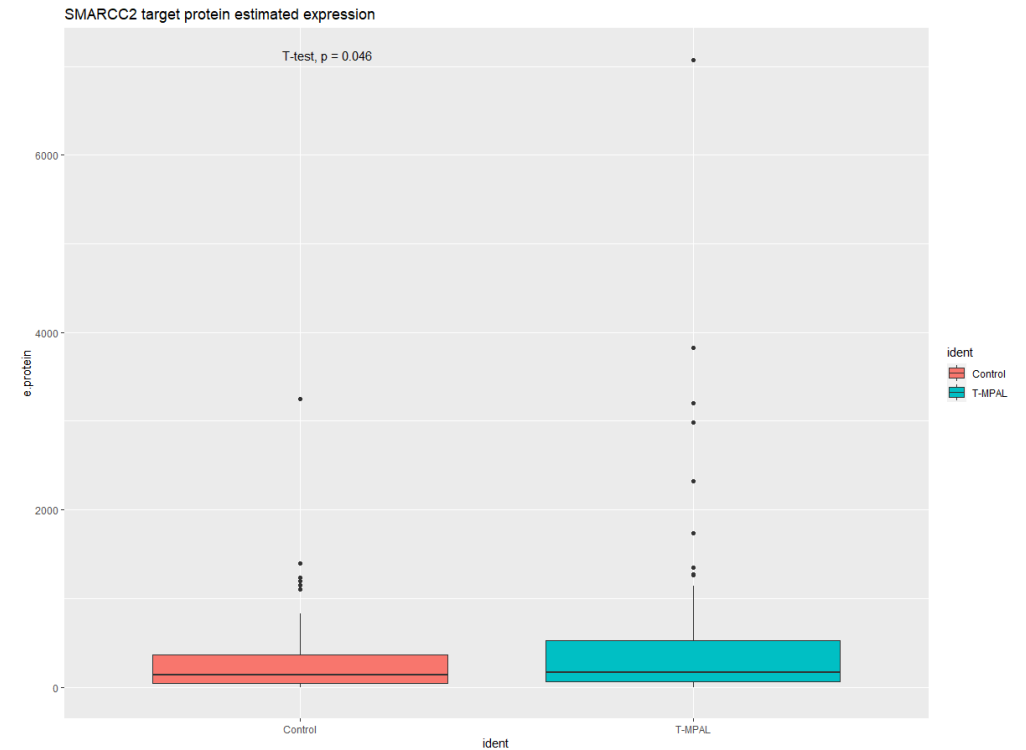
TF patterns

TF -> targets

Goal

Identify transcription factors that meet the following requirements:

1. up-regulated in T-MPAL compared to control
2. have higher enriched targets in T-MPAL versus control
3. have target proteins active in T-MPAL versus control
4. estimated protein targets are significantly higher in T-MPAL versus control

R Package - github.com/hmumme/finalProjectMUMME

Provide functions to predict transcription factor and target activities using Seurat (CITE) and ArchR (ATAC)

Functions:

- *buildCITE(rna, adt)*
- *getTFs(object, rna.markers, tf.db, num = all)*
- *tfEnrichment(atac.markers, db.filt)*
- *plotEn(atac.markers, db.filt, TF.name)*
- *getProteins(adt.markers, db.filt)*
- *getScale(object, ident)*
- *estProtein(x, object, db.filt, TF.name, ident)*
- *tfBar(est, TF.name)*

Demo: demo.Rmd, demo.RData

Future Improvements

- Compare individual cell types
- Separate analysis for inhibitory/promotive targets of TFs
- Refine protein estimation
- Perform analysis on larger dataset

References

[1] Lee, J., Hyeon, D.Y. & Hwang, D. Single-cell multiomics: technologies and data analysis methods. Exp Mol Med 52, 1428–1442 (2020). https://doi.org/10.1038/s12276-020-0420-2

[2] Granja JM, Klemm S, McGinnis LM, Kathiria AS et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat Biotechnol 2019 Dec;37(12):1458-1465. PMID: 31792411