# Single-cell multi-omics analysis identifies context specific gene regulatory gates and mechanisms

Seyed Amir Malekpour[1], Laleh Haghverdi[2], and Mehdi Sadeghi[3]

[1]School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), 19395-5746, Tehran, Iran., [2]Berlin Institute for Medical Systems Biology, Max Delbrück Center (BIMSB-MDC) in the Helmholtz Association, Berlin, Germany., [3]Department of Medical Genetics, National Institute of Genetic Engineering and Biotechnology, 1497716316, Tehran, Iran.

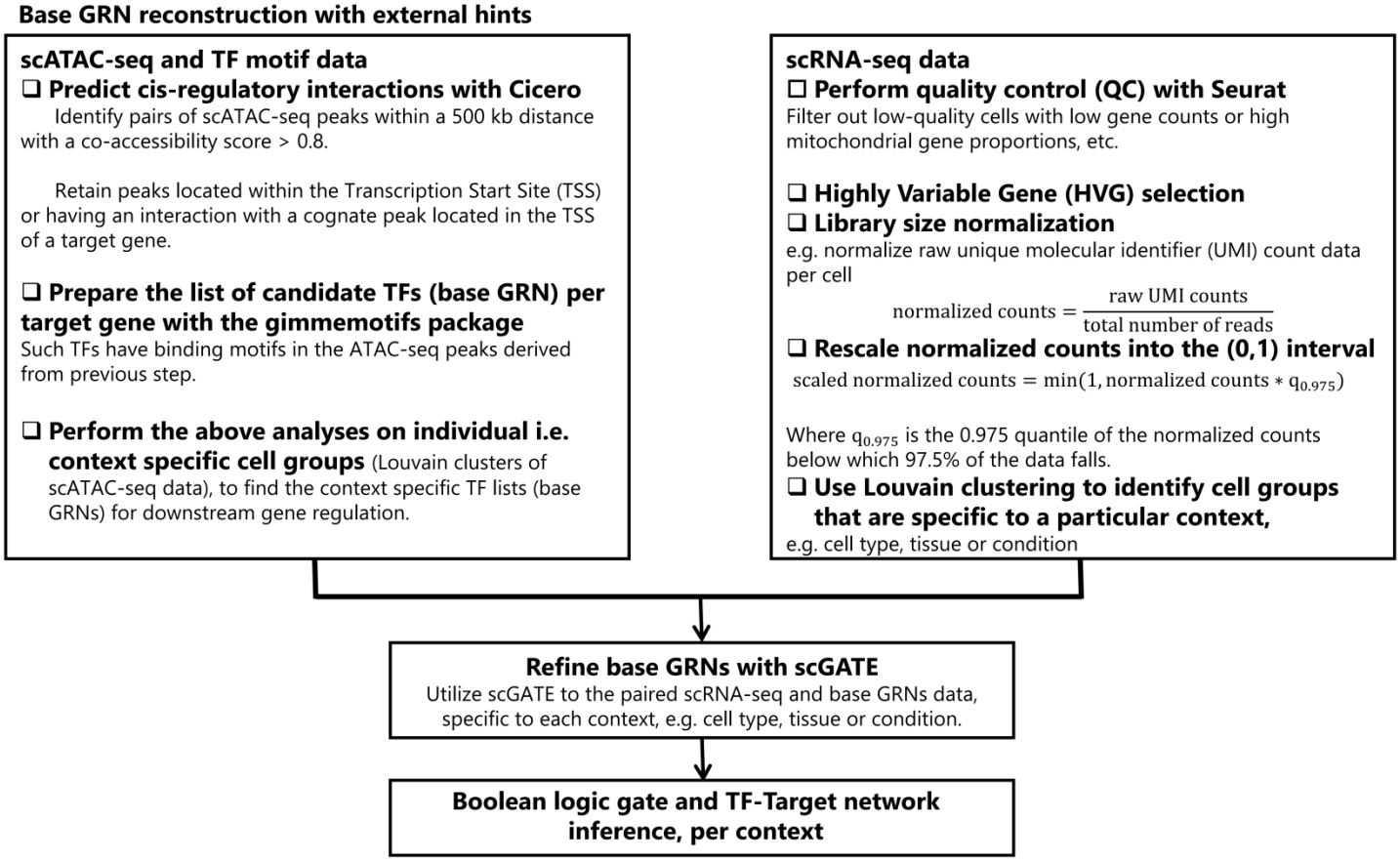


**Figure S1: Data processing pipeline in scGATE:**
(I) The scATAC-seq analysis involves the Cicero package to predict cis-regulatory interactions based on co-accessibility scores between scATAC-seq peaks. Peaks within a 500 kb distance with a co-accessibility score > 0.8 were retained, with a focus on peaks located within the Transcription Start Site (TSS) or having an interaction with a cognate peak located in the TSS of a target gene. Candidate TF lists, base gene regulatory networks (GRNs), were identified using the gimmemotifs package, based on TF binding motifs in the ATAC-seq peaks.

(II) The scRNA-seq analysis involves quality control (QC) and highly variable gene (HVG) selection using Seurat, library size normalization, and rescaling normalized counts with quantile techniques [1, 2]. Louvain clustering was used to identify context specific cell groups, such as cell type, tissue, or condition.

(III) The scGATE tool is employed to refine the base GRNs by integrating the scRNA-seq data and the context specific base GRNs derived from scATAC-seq analysis. This joint analysis allowed for the refinement of the base GRNs specific to each biological context, such as cell type, tissue, or condition.

**Table S1:** Datasets analyzed in scGATE. The table includes the GSE codes and links to the scRNA-seq and scATAC-seq datasets. It also provides additional metadata, such as information about the sequencing platform, the origin tissue or cell type of the samples, and the number of cells sequenced.

| Dataset | Sequencing Platform | Tissue or cell type | Channel | Cell numbers | Accession number/Link |
|---|---|---|---|---|---|
| Mouse haematopoiesis scRNA-seq<br><br>Joakim S. Dahlin et al. 2018[3] | 10X Genomics droplet experiments | Bone marrow Hematopoietic stem and progenitor cells (HSPC) | Lin- c-Kit+ (LK) and Lin- Sca-1+ c-Kit+ (LSK) | 44,802 | https://gottgens-lab.stemcells.cam.ac.uk/adultHSPC10X/ |
| Mouse scRNA-seq<br><br>Tabula Muris Consortium<br><br>Nicholas Schaum et al. 2018[4] | 10X Genomics droplet experiments | Spleen | 10X_P7_6 | 6,115 | GSE109774<br><br>https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109774<br><br>https://github.com/czbiohub-sf/tabula-muris |
| | | | 10X_P4_7 | 3,458 | |
| | | Lung | 10X_P7_9 | 1,525 | |
| | | | 10X_P7_8 | 625 | |
| | | Liver | 10X_P7_1 | 322 | |
| | | Kidney | 10X_P4_6 | 908 | |
| | | | 10X_P4_5 | 610 | |
| | | Heart and Aorta | 10X_P7_4 | 654 | |
| Mouse scATAC-seq<br><br>Darren A Cusanovich et al. 2018[5] | Illumina HiSeq 2500 | Spleen | 62016_P2 | 4,338 | GSE111586<br><br>https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111586 |
| | | Lung | 62216_P1 | 6,119 | |
| | | Liver | 62016_P1 | 7,023 | |
| | | Kidney | 62016_P1 | 7,266 | |
| | | Heart and Aorta | 62816_P1 | 8,991 | |
| human haematopoiesis scATAC-seq<br><br>Jason D Buenrostro et al. 2018[6] | Illumina NextSeq 500 | CD34+ bone marrow | - | 2,034 | GSE96772<br><br>https://www.dropbox.com/sh/8o8f0xu6cvr46sm/AAB6FMIDvHqnG6h7athgcm5-a/Buenrostro_2018.tar.gz?dl=0 |
| human haematopoiesis scRNA-seq<br><br>Jason D Buenrostro et al. 2018[6] | 10X Genomics droplet experiments | CD34+ bone marrow | - | 14,432 | Data S2 of Buenrostro<br><br>https://ars.els-cdn.com/content/image/1-s2.0-S009286741830446X-mmc4.zip |

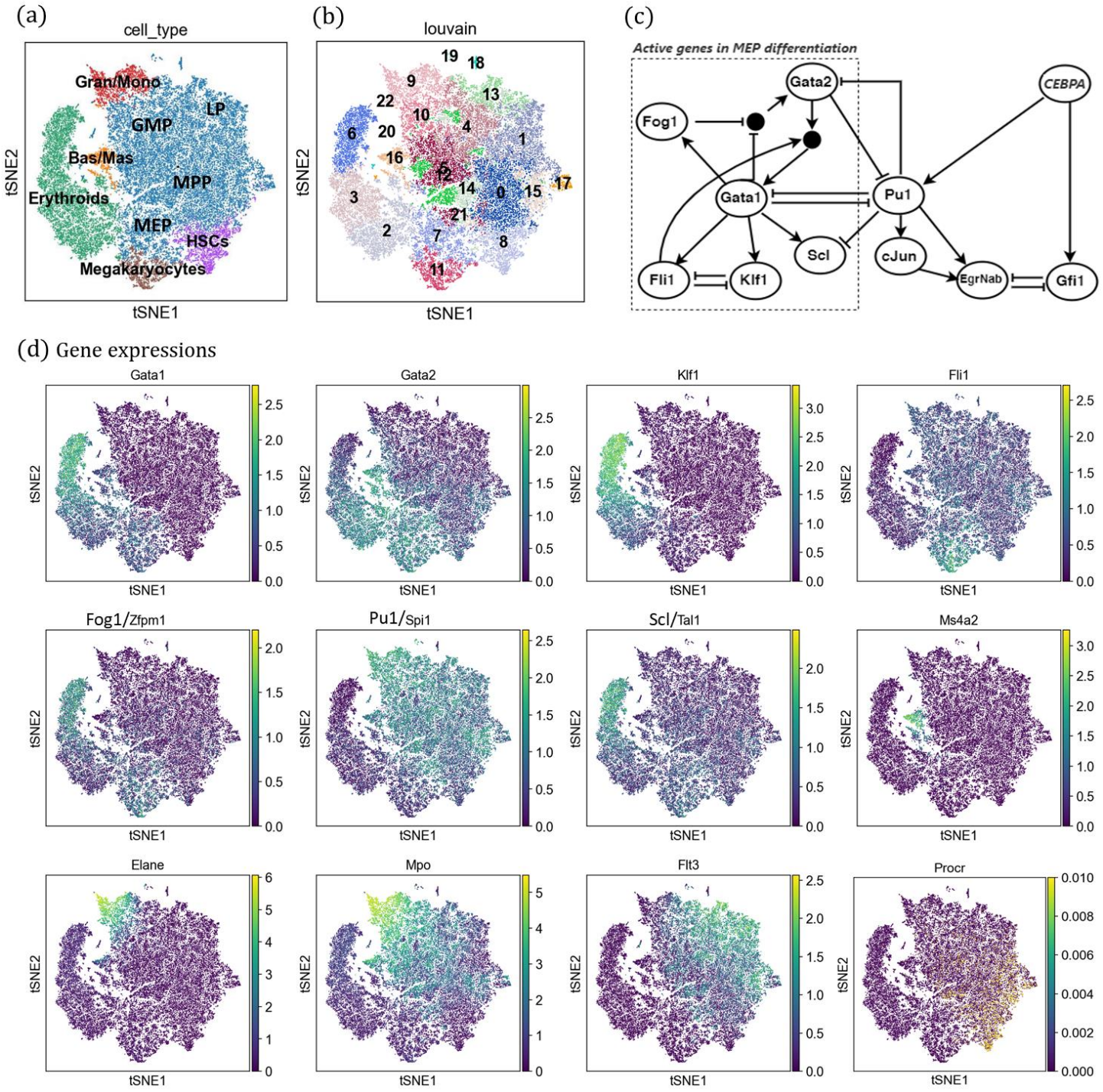


**Figure S2**: Cell-type specific gene expression profile and regulatory network in the mouse haematopoiesis scRNA-seq data [3].

(a) A tSNE plot visualizes distinct trajectories of HSCs (hematopoietic stem cells) as they differentiate into MegE cells (Megakaryocytes and Erythrocytes) and Gran/Mono cells (Granulocytes and Monocytes). The plot also annotates other cell types such as MPP (Multipotent Progenitor), GMP (Granulocyte-Monocyte Progenitor), LP (Lymphoid Progenitor), MEP (Megakaryocyte-Erythrocyte Progenitor), Bas (Basophil), and Mas (Mast). For marker gene expression projected onto the tSNE plot, please refer to panel (d). The specific markers mentioned are Procr for MPP, Gata1 for Erythrocyte, Fli1 for Megakaryocyte, Flt3 for LP, Elane and Mpo for Granulocyte/Monocyte, and Ms4a2 for Basophil/Mast cells.

(b) Louvain cell clusters were identified along the differentiation trajectories.

Context specific gene regulatory gates

(c) A regulatory network is shown, incorporating Boolean update rules that control the cell differentiation process. Black circles connecting edges represent multiple possible update rules (OR relationships) between genes.

(d) The expression profiles of Gata1, Gata2, Klf1, Fli1, Fog1, Pu1, Scl (genes involved in MEP differentiation), Ms4a2 (Bas/Mas marker), Elane, Mpo (Gran/Mono markers), Flt3 (LP marker) and Procr (MPP marker) are depicted.

**Table S2**: Reference and predicted logic gates for the MegE cell differentiation in the mouse haematopoiesis scRNA-seq data [3].

| Target | Krumsiek gates* | scGATE predictions | Cells used for predictions |
|---|---|---|---|
| Fli1 | $\mathrm{Gata1} \wedge \overline{\mathrm{Klf1}}$ | $\mathrm{Gata1} \wedge \overline{\mathrm{Klf1}}$ | Ery and Meg |
| Klf1 | $\mathrm{Gata1} \wedge \overline{\mathrm{Fli1}}$ | $\mathrm{Gata1} \wedge \overline{\mathrm{Fli1}}$ | Ery and Meg |
| Fog1 | $\mathrm{Gata1}$ | $\mathrm{Gata1}$ | Ery and Meg |
| Gata2 | $(\overline{\mathrm{Gata1}} \wedge \overline{\mathrm{Pu1}}) \vee (\overline{\mathrm{Fog1}} \wedge \overline{\mathrm{Pu1}})$ | $(\overline{\mathrm{Gata1}} \wedge \overline{\mathrm{Pu1}}) \vee (\overline{\mathrm{Fog1}} \wedge \overline{\mathrm{Pu1}})$ | Ery and Meg |
| Gata1† | $(\mathrm{Gata2} \wedge \overline{\mathrm{Pu1}}) \vee (\mathrm{Fli1} \wedge \overline{\mathrm{Pu1}})$ | $\mathrm{Gata2} \wedge \overline{\mathrm{Pu1}}$<br>$\mathrm{Gata2} \wedge \overline{\mathrm{Pu1}}$<br>$\overline{\mathrm{Fli1}} \wedge \overline{\mathrm{Pu1}}$<br>$\mathrm{Fli1} \wedge \overline{\mathrm{Pu1}}$ | MegE progenitor cells, Cluster 7<br>Early Ery, Cluster 2<br>Early Ery, Cluster 3<br>Meg, Cluster 11 |
| Scl | $\mathrm{Gata1} \wedge \overline{\mathrm{Pu1}}$ | $\mathrm{Gata1} \wedge \overline{\mathrm{Pu1}}$ | Ery and Meg |

* Krumsiek gates are derived from existing scientific literature [7].

† Gata1 is a key TF that plays a central regulatory role in the specification and differentiation of the MegE lineage [8]. Consistent with other studies [9], the activatory effect of the Gata2 on the Gata1 during early haematopoiesis is also predicted by the scGATE in the MegE progenitor cells (Cluster 7) and early erythroid cells (Cluster2).

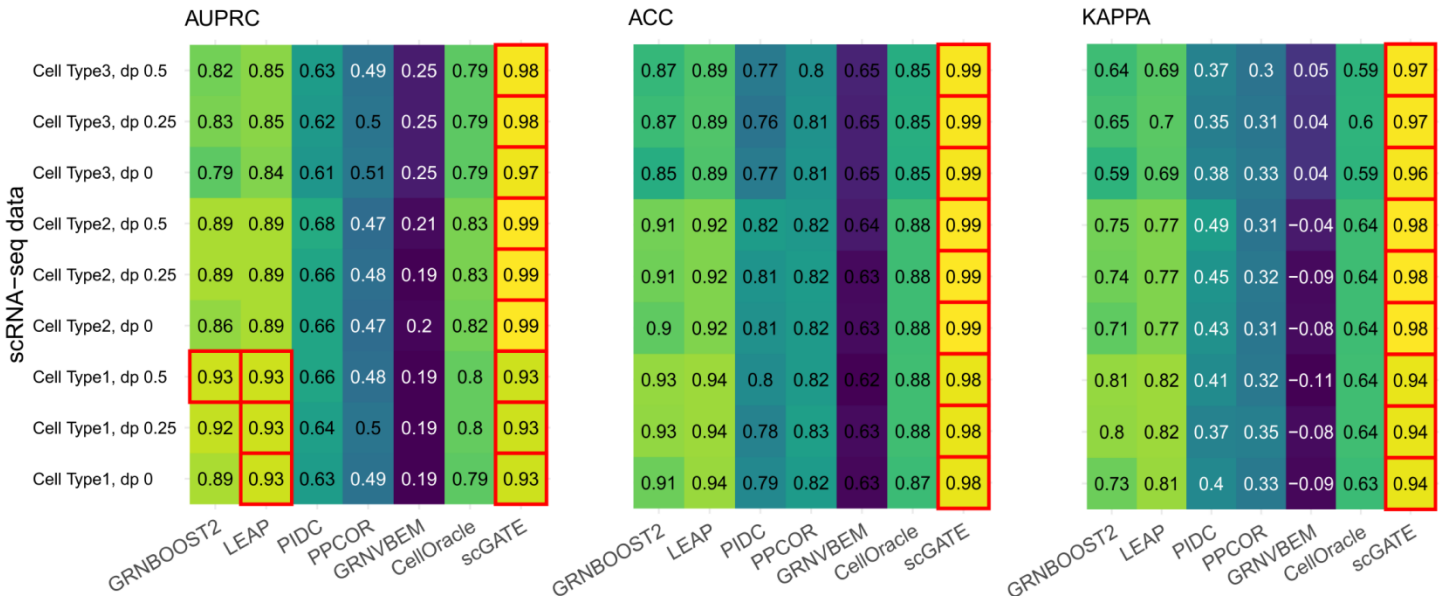


**Figure S3:** Benchmarking the performance of scGATE against other algorithms in terms of AUPRC, ACC, and Kappa-coefficient metrics for cell-type specific GRN inference. Datasets are synthesized with BoolODE package for three GRNs consisting of 15 TFs and 65 target genes, at three dropout (dp) levels 0%, 25%, and 50%.

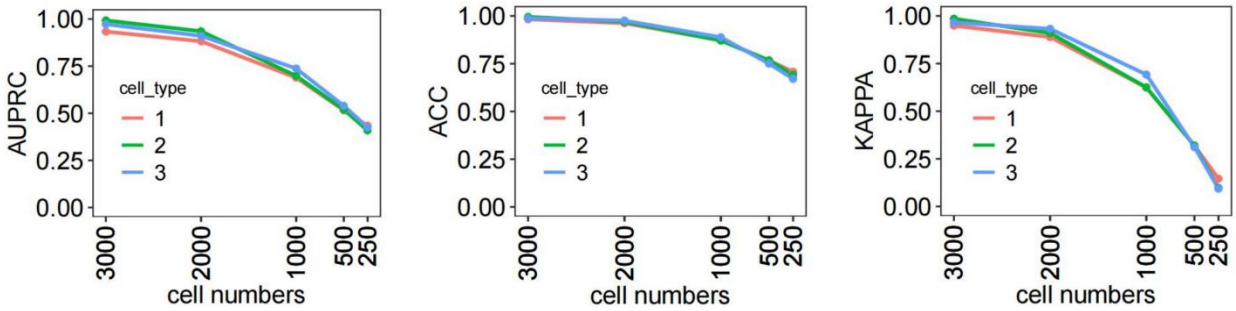


**Figure S4:** scGATE is evaluated in terms of AUPRC, ACC, and Kappa-coefficient metrics on the downsampled datasets with cell numbers reduced to 2000, 1000, 500, and 250. Datasets are synthesized with BoolODE package for three GRNs consisting of 15 TFs and 65 target genes, with 0% dropout (similar results for other dropouts).

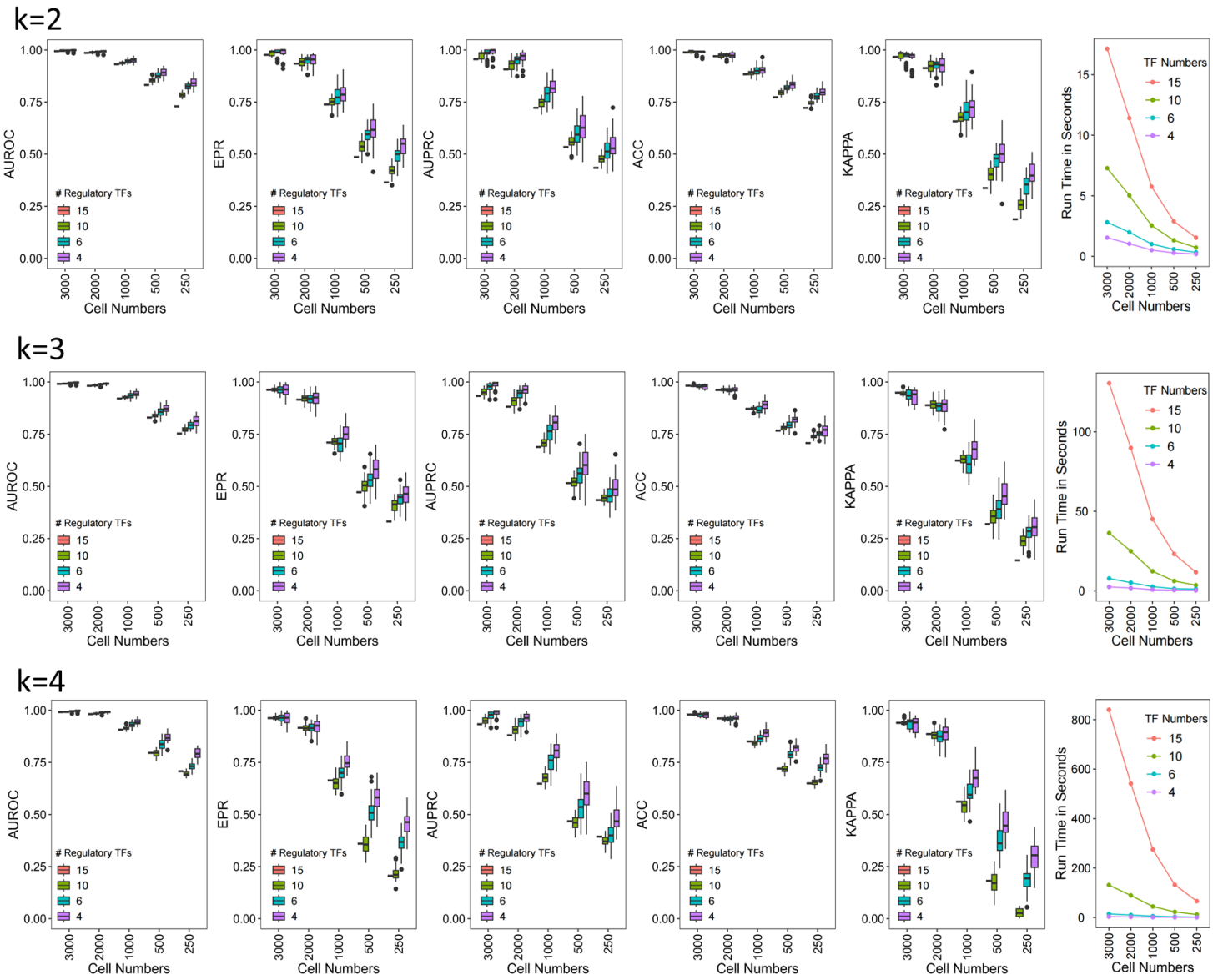


**Figure S5:** scGATE is evaluated considering different numbers of cells and regulatory TFs in the network (non-functional (decoy) TFs are not included). Four, six, ten, and fifteen regulatory TFs, and 3,000, 2,000, 1,000, 500, and 250 cells are considered. The top row represents the evaluation by fitting Boolean logic gates with up to two (k=2) factors from the candidate TF list. The middle and bottom rows correspond to the evaluation with up to three (k=3) and four (k=4) factors,

respectively. The last column displays the scGATE runtime in seconds per target gene when fitting Boolean logic gates including up to two, three, and four factors from the candidate TF list. Datasets are synthesized with BoolODE package for a GRN consisting of 15 TFs and 65 target genes, with 0% dropout (similar results for other dropouts).

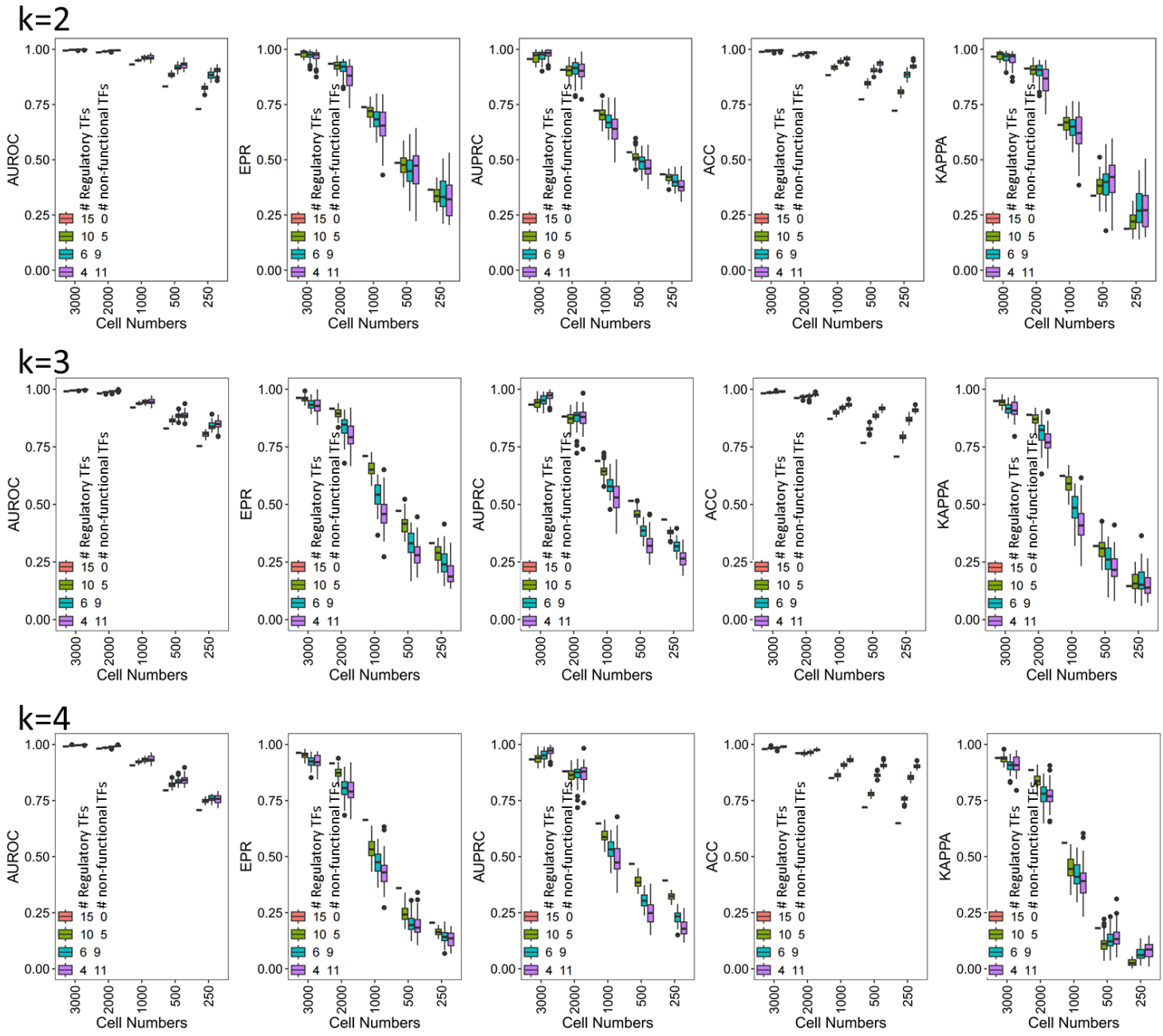


**Figure S6:** scGATE is evaluated considering different numbers of cells, and including different numbers of regulatory and non-functional (decoy) TFs in the network. The top row represents the evaluation by fitting Boolean logic gates with up to two (k=2) factors from the candidate TF list. The middle and bottom rows correspond to the evaluation with up to three (k=3) and four (k=4) factors, respectively. Datasets are synthesized with BoolODE package for a GRN consisting of 15 TFs and 65 target genes, with 0% dropout (similar results for other dropouts).

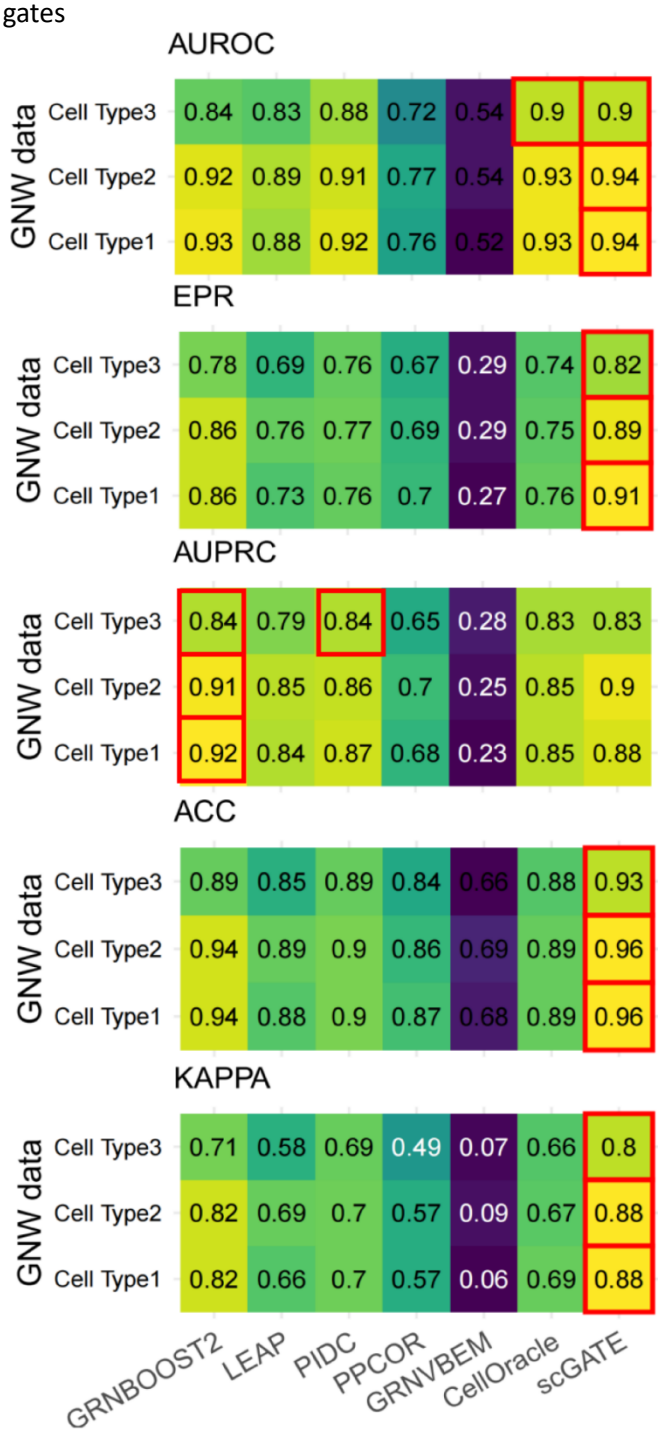


**Figure S7:** Benchmarking the performance of scGATE against other well-known algorithms on the cell-type specific datasets synthesized with GNW package. Performance is evaluated in terms of AUROC, EPR, AUPRC, ACC, and Kappa-coefficient.

Context specific gene regulatory gates

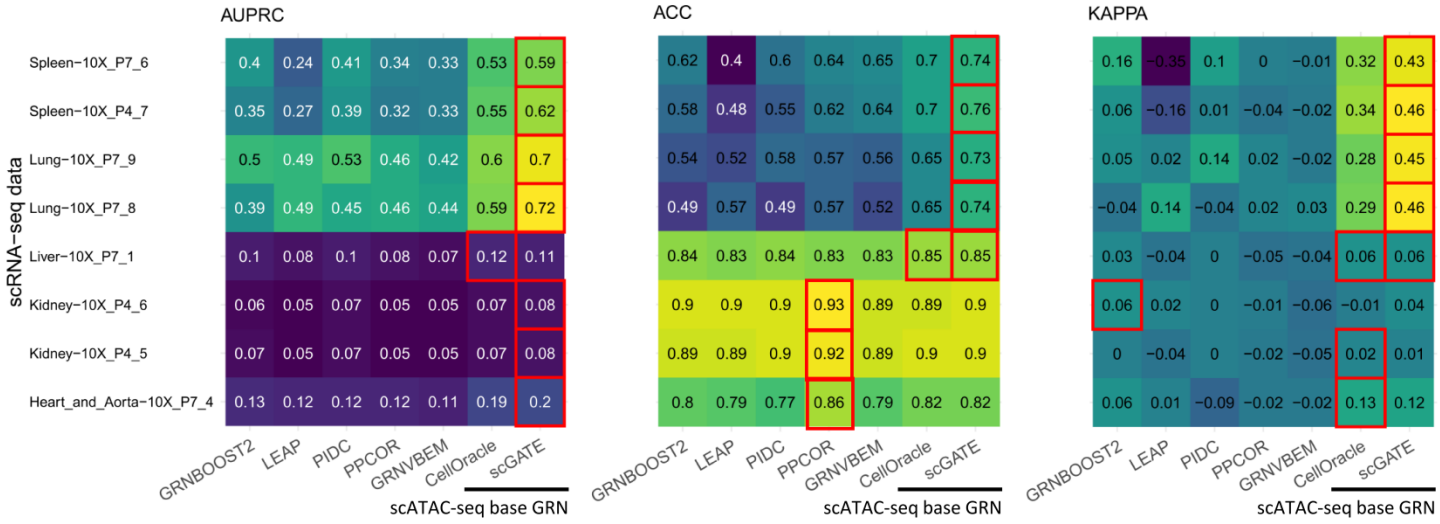


**Figure S8:** Benchmarking the performance of scGATE against other algorithms in terms of AUPRC, ACC, and Kappa-coefficient metrics for context specific network inference in scRNA-seq datasets from five mouse tissues.

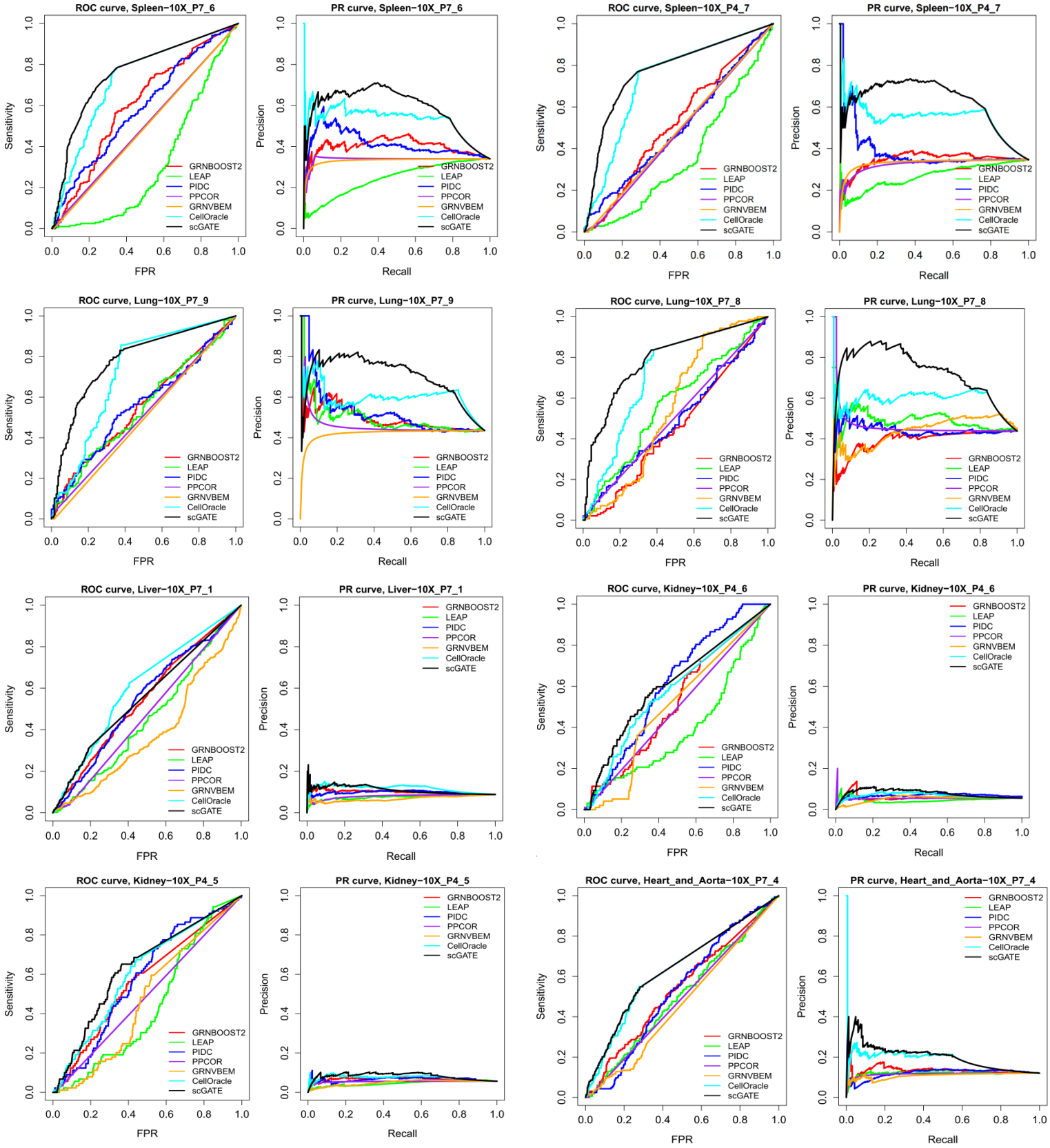


**Figure S9:** ROC and PR curves are plotted for context specific network inference in scRNA-seq datasets from five mouse tissues. Sample IDs are Spleen-10X_P7_6, Spleen-10X_P4_7, Lung-10X_P7_9, Lung-10X_P7_8, Liver-10X_P7_1, Kidney-10X_P4_6, Kidney-10X_P4_5, Heart_and_Aorta-10X_P7_4.

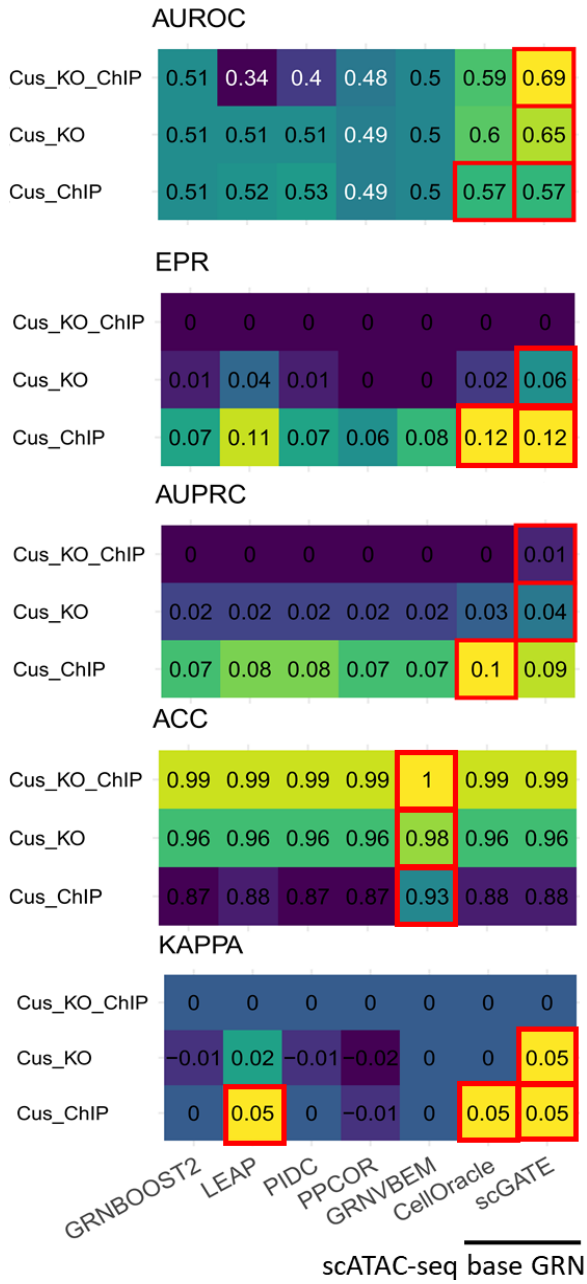


**Figure S10:** Benchmarking the performance of scGATE against other algorithms in terms of AUROC, EPR, AUPRC, ACC, and Kappa-coefficient metrics for context specific network inference in scRNA-seq dataset from human haematopoiesis cells. The predicted networks are compared to the ground-truth networks derived from TF perturbation experiments (Cus_KO) and ChIP-seq (Cus_ChIP) assays conducted in the GM12878 lymphoblastoid cell line [10], and also the intersection of the perturbation and ChIP-seq studies (Cus_KO_ChIP).

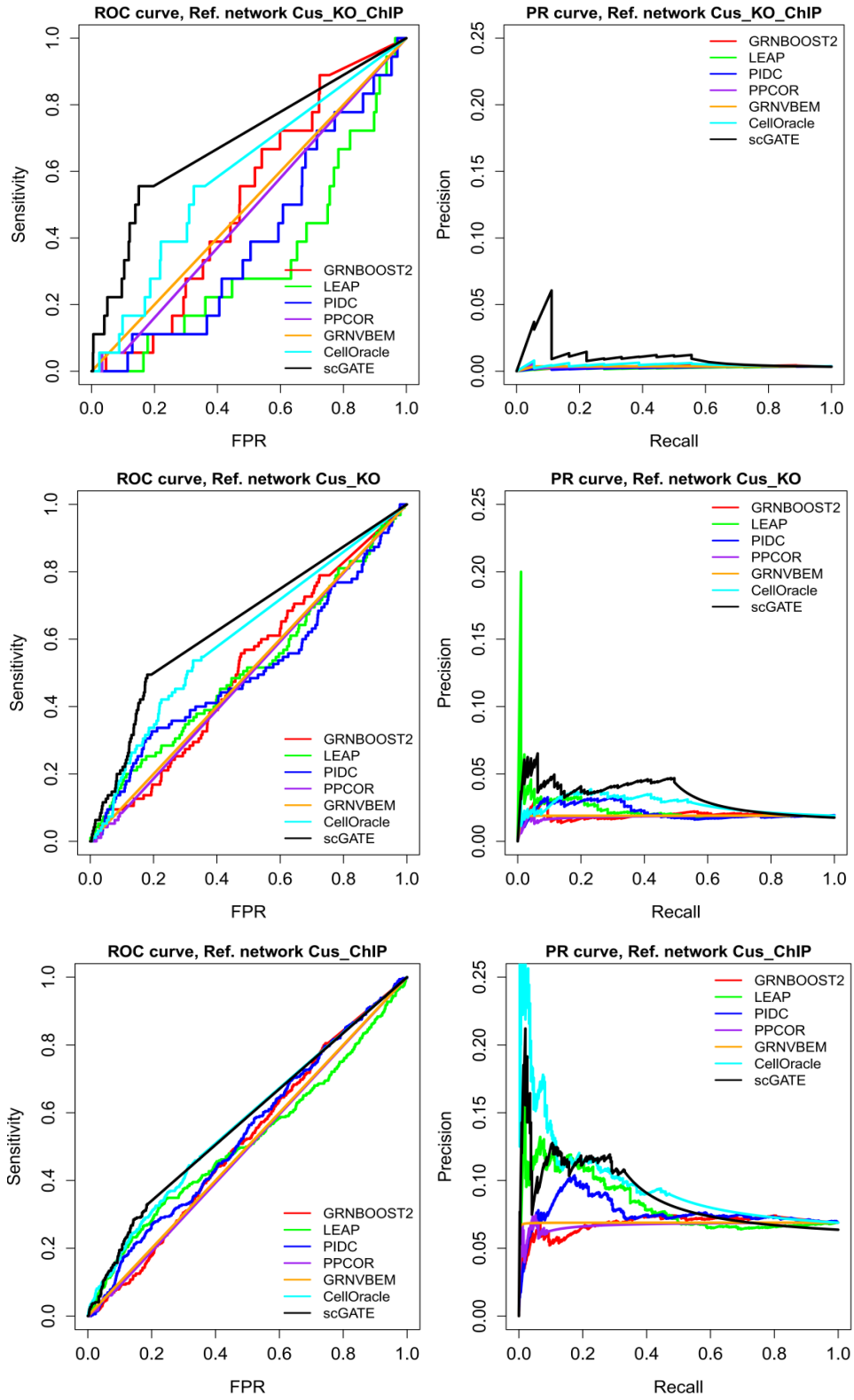


**Figure S11:** ROC and PR curves are plotted for context specific network inference in scRNA-seq dataset from human haematopoiesis cells. The predicted networks are compared to the ground-truth networks derived from TF perturbation experiments (Cus_KO) and ChIP-seq (Cus_ChIP) assays conducted in the GM12878 lymphoblastoid cell line [10], and also the intersection of the perturbation and ChIP-seq studies (Cus_KO_ChIP).

**Table S3**: The running time and memory usage are evaluated for GRNBOOST2, LEAP, PIDC, PPCOR, GRNVBEM, CellOracle, and scGATE on the Spleen-10X_P7_6 sample from the mouse tissue and the human haematopoiesis dataset. We reached comparable results for other samples from mouse tissues.

| | Mouse tissue Spleen-10X_P7_6 | | Human haematopoiesis | |
|---|---|---|---|---|
| | Run time | Memory | Run time | Memory |
| GRNBOOST2 | 00:03:23 | 767.8 | 00:05:18 | 1967.4 |
| LEAP | 00:49:45 | 182.2 | 02:09:08 | 336.3 |
| PIDC | 00:00:31 | 384.2 | 00:00:30 | 624.3 |
| PPCOR | 00:00:04 | 146.0 | 00:00:05 | 256.1 |
| GRNVBEM | 01:23:31 | 607.7 | 00:40:53 | 828.6 |
| CellOracle | 00:01:03 | 398.8 | 00:03:29 | 1400.7 |
| scGATE | 00:03:28 | 127.3 | 02:32:50 | 3566.7 |

Notes:

The running times are in hours:minutes:seconds orders. For example, 01:23:31 shows 1 hour, 23 minutes and 31 seconds. Memory usage is measured in Megabytes (MB).

In scGATE, the run time and memory usage are calculated using the 'peakRAM' package in R.

**scGATE guideline:**

## Step 1. scGATE installation

The scGATE codes are written in R version 4.1.3 and have been tested in both Windows and Linux environments.

Installation

1. Download the compiled package file scGATE_0.1.0.tar.gz from this GitHub page.
2. Install the scGATE package by running the following command in R:

```
install.packages("path/to/scGATE_0.1.0.tar.gz", repos = NULL, type = "source")
```

Dependencies

Please ensure that you have the following packages installed:

```
install.packages("VGAM")
install.packages("truncnorm")
install.packages("arrow")
```

These commands will install the VGAM, truncnorm, and arrow packages, which are required for running scGATE.

To load the packages, use the following commands:

```
library(scGATE)
library(VGAM)
library(truncnorm)
library(arrow)
```

## Step 2. Prepare input files

Preprocessing base GRN generated from external hints

To summarize information in the base GRN file in ".parquet" format, previously generated using external hints like scATAC-seq and TF binding motif analyses, you can use the read_base_GRN() function from the scGATE package.

```
# Read and summarize base GRN file
candidate_tf_target <- as.data.frame(read_parquet("Buenrostro2018_base_GRN_dataframe.parquet"))
candidate_tf_target <- read_base_GRN(candidate_tf_target)
```

Preprocessing scRNA-seq count data

To preprocess raw scRNA-seq data, including steps such as normalization and rescaling, you can use the scRNA_seq_preprocessing() function from the scGATE package.

```
# Preprocess scRNA-seq count data
```

```
normalized_counts <- scRNA_seq_preprocessing(data = data_scRNA_seq, library_size_normalization =
"True", tf_list = NA)

Parameter Descriptions
data: The scRNA-seq raw data matrix with cells in rows and genes in columns.
library_size_normalization: A flag indicating whether library size normalization should be
performed. The default value is "True". Set it to "False" if you don't want to perform library
size normalization.
tf_list: A list of transcription factors (TFs) to consider. The default value is NA, which means
all columns in the data matrix will be considered as TFs.
```

### Step 3. Run scGATE

```
scGATE provides two functions for TF-target network inference: scGATE_gate() and scGATE_edge().
These functions infer the TF-target network with and without predicted Boolean logic gates in the
output, respectively. The scGATE_gate() function in the scGATE package is more suitable for small
networks or when the base gene regulatory network (GRN) is available from external sources such as
scATAC-seq and TF motif data.
```

### TF-Target Network Inference (gate mode)

```
To infer the TF-target network with logic gates in the output, you can use
the scGATE_gate() function.
# Infer TF-target network without logic gates in the output
gates <- scGATE_logic(data = data, base_GRN = NA, h_set = NA, number_of_em_iterations = NA,
max_num_regulators = NA, abs_cor = NA, top_gates = NA, run_mode = NA)

Parameter Descriptions
data: A gene expression matrix with normalized counts within the (0,1) interval, where samples are
represented as rows and genes as columns. The gene expression matrix should have been preprocessed
using the scRNA_seq_preprocessing() function.
base_GRN: Base TF-gene interaction network derived from external hints (e.g., scATAC-seq data and
TF binding site motifs on DNA).
h_set: The range of possible values for the "h" parameter in the Hill climbing function.
number_of_em_iterations: The number of iterations in the expectation-maximization (EM) algorithm.
max_num_regulators: The Maximum number of TFs in a logic gate that can regulate the target gene
profile. In the main manuscript, a value of 3 is used.
abs_cor: This parameter varies in the (0, 1) interval and further removes edges with low absolute
Pearson correlations between TFs and their targets. A (default) value of 0 indicates no filtration
based on correlations.
top_gates: The number of top Boolean logic gates to be reported for each target gene, based on
Bayes Factor.
run_mode: Use "simple" for a faster algorithm run and "complex" for more precise results that take
more time. The argument is relevant to the possible complexities in the hill function parameter
space for regulatory TFs and target genes.
```

### TF-Target Network Inference (edge mode)

```
To infer the TF-target network without logic gates in the output, you can use
the scGATE_edge() function.
# Infer TF-target network without logic gates in the output
edges <- scGATE_edge(data = data, base_GRN = candidate_tf_target, h_act = NA,
number_of_em_iterations = NA, max_num_regulators = NA, abs_cor = NA)
```

Parameter Descriptions

data: A gene expression matrix with normalized counts within the (0,1) interval, where samples are represented as rows and genes as columns. The gene expression matrix should have been preprocessed using the scRNA_seq_preprocessing() function.

base_GRN: The TF-target gene network inferred from previous steps using external hints. Leave it empty if no base GRN is available.

h_act: Hill function parameter used in the inference process.

number_of_em_iterations: The number of iterations in the expectation-maximization (EM) algorithm.

max_num_regulators: The maximum number of TFs in a Boolean logic gate. In the main manuscript, a value of 3 is used.

abs_cor: This parameter varies in the (0, 1) interval and further removes edges with low absolute Pearson correlations between TFs and their targets. A (default) value of 0 indicates no filtration based on correlations.

**Example usage of scGATE**

**I. Context specific network and logic gate inference in synthetic toggle switch**

```
# 1. Please refer to the Jupyter notebook for instructions on how to perform Louvain clustering on the cells in the BoolODE simulated data.
# 2. Retrieve the data from Cluster I of cells, which was obtained in the previous step.
# Load scGATE package and data in example_data folder

rm(list = ls())
library(scGATE)

data <- as.matrix(read.csv("/example_data/ClusterI.csv")[ ,2:15])
print(head(data))

# 3. data preprocessing
# For scGATE simulated data, library size normalization is not performed.
# However, the simulated data is only re-scaled using the quantile normalization technique to fit the data within the (0,1) interval.
data <- scRNA_seq_preprocessing(data = data, library_size_normalization = "False")

# 4. Remove genes with low variability (scGATE operates on highly variable genes per context).
# This step is optional
data$n_counts <- data$n_counts[ , which(sqrt(apply(data$n_counts,2,var))> 0.20)]

# 5. Run scGATE_logic() function
# Please note that the likelihood values can be affected by the Louvain clustering results.
gates <- scGATE_logic(data = data, top_gates = 1, run_mode = "fast")

print(head(gates))
  gene_name -log10 L0 -log10 L1 log10 BF logic_gate
1        gE     173.9   -268.57   442.47        ~gF
2       gE1     51.85   -234.65   286.50    gE.~gE2
3       gE2     38.43   -235.48   273.91    gE.~gE1
4        gF    170.38   -278.57   448.95        ~gE
5       gF1     80.36   -215.32   295.68    gF.~gF2
6       gF2      67.6   -217.88   285.48    gF.~gF1
```

## II. Context specific network and logic gate inference in the mouse haematopoiesis scRNA-seq data

```
# 1. Please refer to the Jupyter notebook for instructions on how to perform Louvain clustering on
the cells in the mouse haematopoiesis scRNA-seq dataset.
# 2. Retrieve the data from Megakaryocyte cells (Cluster 11).
# Load scGATE package and data in example_data folder

rm(list = ls())
library(scGATE)
data <- as.data.frame(read.csv("/example_data/subset_counts_cluster_11.csv" , header = TRUE))

# select genes involved in the MegE differentiation
gene_list   <- c("Gata1", "Fli1", "Klf1", "Spi1", "Zfpm1", "Tal1", "Gata2")
data        <- data[  , gene_list]
data        <- na.omit(data)
print(head(data))
      Gata1      Fli1     Klf1      Spi1     Zfpm1      Tal1     Gata2
1 0.6931472 1.0986123 0.000000 0.6931472 0.0000000 0.6931472 0.0000000
2 0.0000000 1.3862944 0.000000 0.0000000 0.0000000 0.6931472 1.0986123
3 0.6931472 1.6094380 0.000000 0.0000000 0.0000000 0.0000000 0.6931472
4 0.0000000 0.0000000 1.098612 0.0000000 0.6931472 0.0000000 1.6094380
5 0.0000000 0.0000000 0.000000 0.0000000 0.6931472 0.6931472 1.3862944
6 0.0000000 0.6931472 0.000000 0.0000000 0.6931472 1.0986123 0.0000000

# Load base GRN
base_GRN <- read.csv("/example_data/base_grn_mouse_blood_cell_differentiation_toggle_switch.csv")

# 3. data preprocessing
# The dataset underwent library size normalization in Jupyter Notebook. To fit the scRNA-seq data
within the (0,1) interval, we applied quantile normalization as a technique to rescale the data.
data <- scRNA_seq_preprocessing(data = data, library_size_normalization = "False")

# 4. Run scGATE_logic() function
gates <- scGATE_logic(data = data, base_GRN = base_GRN, number_of_em_iterations = 10, top_gates =
1, run_mode = "slow")
print(head(gates))
```

## III. Context specific network inference in mouse tissue scRNA-seq datasets

```
# 1. Please refer to the Jupyter notebook for instructions on how to perform scATAC-seq analysis
to derive the candidate TF lists (base GRNs) in *.parquet file format.
# 2. Load scGATE package and data (base GRN and scRNA-seq data and TF list) in example_data folder

rm(list=ls())
library(scGATE)

# Load base GRN derived from external hints
candidate_tf_target <-
as.data.frame(read_parquet("/example_data/Cusanovich2018_Spleen_peak_base_GRN_dataframe.parquet"))
candidate_tf_target <- read_base_GRN(candidate_tf_target)

# Load scRNA-seq data
data <- as.data.frame(read.csv("/example_data/Tabula_Muris2018_Spleen-
10X_P4_7_ExpressionData.csv", header = TRUE))
gene_names     <- data[ ,1]
data           <- t(data[ ,2:ncol(data)])
colnames(data) <- gene_names
head(data[ , 1:10])
                  Batf Stat5b Ctcf H2-Eb1 AW112010 Ly6d Rplp0 Id2 Dok2 Gimap3
AAACCTGAGAAGGACA.1    0      0    0     18         0    0    10   0    0      0
AAACCTGAGCTAAGAT.1    0      0    1      0        19    0     5   1    1      1
AAACCTGCAACAACCT.1    0      0    0     22         0    5    12   0    0      2
AAACCTGCAGCCAATT.1    0      0    0     14         1    5    21   0    0      1
AAACCTGCAGCTCCGA.1    0      0    1     30         1    2    64   0    0      0
AAACCTGTCAGGTAAA.1    0      0    0     23         3    8    24   0    0      0

# Load TF list
# This step is optional
tf_names <- unlist(read.table("/example_data/Tabula_Muris2018_Spleen-10X_P4_7_tf_lists.txt"))
print(head(tf_names))
      V1       V2       V3
  "Batf" "Stat5b"   "Ctcf"

# 3. scRNA-seq data preprocessing (library size normalization, quantile normalization technique to
fit the scRNA-seq data within the (0,1) interval)
data <- scRNA_seq_preprocessing(data, library_size_normalization = "True", tf_list = tf_names)

# 4. Run scGATE_edge() function
ranked_edge_list <- scGATE_edge(data = data, base_GRN = candidate_tf_target, h_act = 7)
print(head(ranked_edge_list))
    from    to BF_score
1   Ctcf Rps19 2013.587
2   Batf Rps19 2012.551
3 Stat5b Rplp0 1850.334
4   Ctcf Rplp0 1849.896
5   Ctcf Rpl36 1649.263
6   Ctcf Eif5a 1559.044
```

## IV. Context specific network inference in human haematopoiesis scRNA-seq dataset

```
# 1. Please refer to the Jupyter notebook for instructions on how to perform scATAC-seq analysis
to derive the candidate TF lists (base GRNs) in *.parquet file format.
# 2. Load scGATE package and data (base GRN and scRNA-seq data and TF list) in example_data folder

rm(list=ls())
library(scGATE)

# Load base GRN derived from external hints
candidate_tf_target <-
as.data.frame(read_parquet("/example_data/Buenrostro2018_base_GRN_dataframe.parquet"))
candidate_tf_target <- read_base_GRN(candidate_tf_target)

# Load scRNA-seq data
data <- as.data.frame(read.csv("/example_data/Buenrostro2018_ExpressionData.csv", header = TRUE))
gene_names     <- data[ ,1]
data           <- t(data[ ,2:ncol(data)])
colnames(data) <- gene_names

head(data[ , 1:10])
      IRF8 FOS MAFF SPI1 JUNB SPIB IRF7 TFDP1 GATA1 RAD21
hsc_1    0   2    0    0    2    0    0     0     0     1
hsc_2    0   6    7    0    3    0    0     0     0     1
hsc_3    0   2    0    0    5    0    0     0     0     2
hsc_4    0   6    0    0    1    0    0     1     0     1
hsc_5    0   1    5    2    1    0    0     0     0     0
hsc_6    0   3    0    0    1    0    0     0     0     0

# Load TF list
# This step is optional
tf_names <- unlist(read.table("/example_data/Buenrostro2018_tf_lists.txt"))
print(head(tf_names))
    V1     V2     V3     V4     V5     V6
"IRF8"  "FOS" "MAFF" "SPI1" "JUNB" "SPIB"

# 3. scRNA-seq data preprocessing (library size normalization, quantile normalization technique to
fit the scRNA-seq data within the (0,1) interval)
data <- scRNA_seq_preprocessing(data, library_size_normalization = "True", tf_list = tf_names)

# 4. Run scGATE_edge() function
ranked_edge_list <- scGATE_edge(data = data, base_GRN = candidate_tf_target, h_act = 7)

print(head(ranked_edge_list))
     from     to BF_score
1    E2F1 MALAT1 13415.34
2 BHLHE40 MALAT1 13415.34
3   TFDP1 MALAT1 13415.32
4    NFE2 MALAT1 13414.98
5    IRF8 MALAT1 13414.26
6   RUNX2   PTMA 11592.68
```

References


1. Yoshida, H., et al., *The cis-Regulatory Atlas of the Mouse Immune System.* Cell, 2019. **176**(4): p. 897-912.e20.
2. Stuart, T., et al., *Comprehensive Integration of Single-Cell Data.* Cell, 2019. **177**(7): p. 1888-1902.e21.
3. Dahlin, J.S., et al., *A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice.* Blood, 2018. **131**(21): p. e1-e11.
4. Schaum, N., et al., *Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris.* Nature, 2018. **562**(7727): p. 367-372.
5. Cusanovich, D.A., et al., *A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility.* Cell, 2018. **174**(5): p. 1309-1324.e18.
6. Buenrostro, J.D., et al., *Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation.* Cell, 2018. **173**(6): p. 1535-1548.e16.
7. Krumsiek, J., et al., *Hierarchical Differentiation of Myeloid Progenitors Is Encoded in the Transcription Factor Network.* PLOS ONE, 2011. **6**(8): p. e22649.
8. Fujiwara, Y., et al., *Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1.* Proceedings of the National Academy of Sciences, 1996. **93**(22): p. 12355-12358.
9. Ohneda, K. and M. Yamamoto, *Roles of Hematopoietic Transcription Factors GATA-1 and GATA-2 in the Development of Red Blood Cell Lineage.* Acta Haematologica, 2002. **108**(4): p. 237-245.
10. Cusanovich, D.A., et al., *The Functional Consequences of Variation in Transcription Factor Binding.* PLOS Genetics, 2014. **10**(3): p. e1004226.