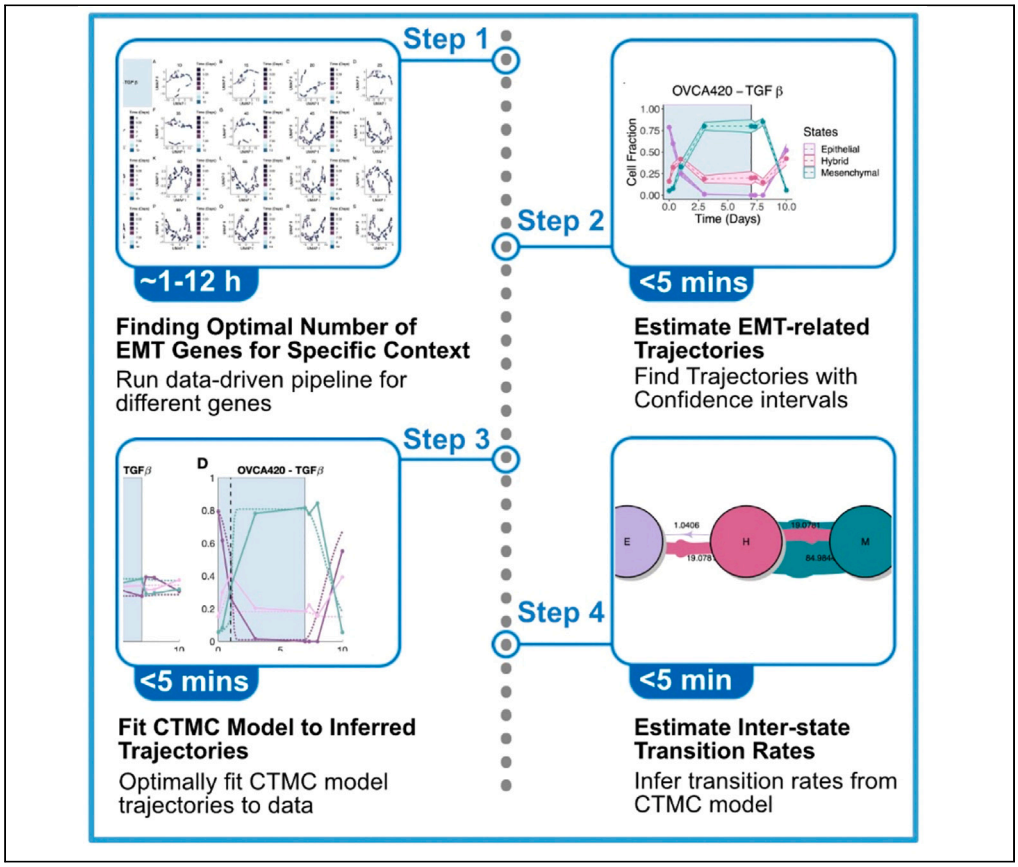
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for inferring epithelial-to-mesenchymal transition trajectories from single-cell RNA sequencing data using R



The epithelial-to-mesenchymal transition (EMT) provides crucial insights into the metastatic process and possesses prognostic value within the cancer context. Here, we present COMET, an R package for inferring EMT trajectories and inter-state transition rates from single-cell RNA sequencing data. We describe steps for finding the optimal number of EMT genes for a specific context, estimating EMT-related trajectories, optimal fitting of continuous-time Markov chain to inferred trajectories, and estimating inter-state transition rates.

Publisher's note: Undertaking any experimental protocol requires adherence to local institutional guidelines for laboratory safety and ethics.

Annice Najafi, Mohit Kumar Jolly, Jason T. George

jason.george@tamu.edu (J.T.G.)
mkjolly@iisc.ac.in (M.K.J.)

Highlights

Accurate inference of EMT-related trajectories from time-course scRNA-seq data

Identifying critical genes driving the EMT process in each context

Optimally fitting CTMC models of EMT phenotypic transitions to data

Estimating inter-state transition rates through the CTMC models

Najafi et al., STAR Protocols 5, 102819
March 15, 2024
https://doi.org/10.1016/j.xpro.2023.102819

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for inferring epithelial-to-mesenchymal transition trajectories from single-cell RNA sequencing data using R

Annice Najafi,[1,5] Mohit Kumar Jolly,[2,*] and Jason T. George[1,3,4,6,*]

[1]Department of Biomedical Engineering, Texas A&M University, College Station, TX 77843, USA

[2]Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore 560012, India

[3]Intercollegiate School of Engineering Medicine, Texas A&M University, Houston, TX 77030, USA

[4]Center for Theoretical Biological Physics, Rice University, Houston, TX 77030, USA

[5]Technical contact

[6]Lead contact

*Correspondence: jason.george@tamu.edu (J.T.G.), mkjolly@iisc.ac.in (M.K.J.)
https://doi.org/10.1016/j.xpro.2023.102819

SUMMARY

The epithelial-to-mesenchymal transition (EMT) provides crucial insights into the metastatic process and possesses prognostic value within the cancer context. Here, we present COMET, an R package for inferring EMT trajectories and inter-state transition rates from single-cell RNA sequencing data. We describe steps for finding the optimal number of EMT genes for a specific context, estimating EMT-related trajectories, optimal fitting of continuous-time Markov chain to inferred trajectories, and estimating inter-state transition rates.

BEFORE YOU BEGIN

Overview

EMT is a dynamic cellular process that contributes to cancer metastasis by enabling tumor cells to dissociate from their epithelial state while also acquiring migratory and mesenchymal characteristics.[1–3] Previously viewed as a binary process, stable hybrid intermediate EMT states are now well-established and associated with enhanced phenotypic plasticity, thus prolonging tumor survival and worsening patient outcomes.[4] Understanding the timing of EMT and the distribution of cells in each state in detail would enable more complete characterizations of intertemporal patterns of phenotypic intra-tumoral heterogeneity in a context-specific manner.

Here, we introduce an R package that provides a computational framework for inferring EMT-related states and trajectories from time-course scRNA-seq data. We describe this protocol for a time-course scRNA-seq dataset which was acquired from a cancer cell line treated with an EMT-related transcription factor.[5] We show that our methodology allows context-specific inference of EMT-related trajectories and transition rates and is of direct utility for the scientific community for estimating dynamic EMT status.

Install tools and packages

1. To run the scripts, please install R and optionally RStudio (MATLAB code is also provided on our GitHub for the CTMC model).
2. Install 'devtools' and load it.

Note: Currently, the COMET package is available on GitHub. To proceed with using COMET, you should install "devtools" which allows installation of GitHub package.

```
>Install.package("devtools")

>library(devtools)
```

3. As of now, the functions required for running this pipeline are available as an R package on GitHub. To install the package, please run:

```
>devtools::install_github("TAMUGeorgeGroup/COMET")
```

4. Please load all relevant R packages that are listed in the key resources table (The dependencies are imported by default upon installing COMET).

```
#Load relevant libraries

>library(COMET)

>library(dplyr)

>library(ggplot2) #For plots

>library(tidyverse)

>library(data.table)

>library(tidyr)

>library(reshape2)

>library(Rmagic)

>library(umap)

>library(readxl)

>library(Seurat)

>library(dtw)

>library(pracma)

>Library(Rmagic)
```

Note: You should install the 'phateR' and 'Rmagic' packages separately. 'Rmagic' is removed from CRAN, consequently you would have to install the archived package (https://cran.r-project.org/src/contrib/Archive/Rmagic/Rmagic_2.0.3.tar.gz) and place it on a directory of your choice. In the example below the package is downloaded in the 'Downloads' directory and hence we load the package from there.

```
>devtools::install_github("KrishnaswamyLab/phateR")

>install.packages("~/Downloads/Rmagic_2.0.3.tar.gz")
```

Either from a terminal window of RStudio or your computer, run the command below.

```
>pip install --user magic-impute
```

Load the Rmagic library.

```
>library(Rmagic)
```

Download or prepare datasets

Note: The dataset used in this protocol and provided on Zenodo (https://zenodo.org/records/10050380) and GitHub is from Cook et al. 2020.[5] The metadata for this dataset is modified for COMET. More specifically, the time points in the original dataset were modified to a numeric form (0, 0.33, 1, 3, 7, 7.33, 8, and 10). Please refer to Cook et al. 2020[5] for more information regarding the experimental data.

Note: If using your own dataset, please ensure that the count matrix and metadata are stored in separate csv.gz files. Both the data and metadata should be objects of the 'data.frame' class in R. When running

Note: COMET requires that the count matrix is quality controlled prior to processing. Common quality controlling steps for single-cell RNA sequencing data include checking for mitochondrial percentage and filtering for number of expressed genes in each cell. We leave this step to the user as the quality controlling steps are variable depending on the dataset and outside the scope of our pipeline.

Note: If your input data is stored in a Seurat object (named 'Seur' in the following code), you can utilize the following code to save your count matrix and metadata in csv.gz files.

```
>write_csv(as.data.frame(Seur@assays$RNA@counts),"data.csv.gz")

>write_csv(as.data.frame(Seur@meta.data),"metadata.csv.gz")
```

Note: Please ensure that the metadata file has a column named 'Time' that specifies the corresponding time point (or dose if using dose-dependent data) for each cell in a numeric format. You can check for the existence of the Time column and ensure it is in a numeric format by running the following code.

```
>check_1 <- "Time" %in% colnames(meta.data)

>check_2 <- is.numeric(meta.data$Time)

>checks <- sum(check_1+check_2)

> if (checks==2) print("metadata file is in correct format") else "metadata is in incorrect format"
```

Note: If you have data from different batches, please ensure that your data is batch adjusted and the gene expression values from the corresponding cells are all combined into one dataframe. We leave the batch adjustment step to the user as this step is dependent on the input data and the experimental setting and outside the scope of our pipeline.

Note: The input count matrix and metadata should be in the following format (Figures 1 and 2).

Note: By installing the COMET package, the dependencies mentioned in Key Resource Table are installed by default other than Rmagic and phateR. If facing issues regarding the installation of dependencies, please refer to the corresponding sources mentioned in the Key Resource Table for each dependency or kindly request the technical contact author for additional help.

Filter | Cols: « < 1 – 50 > »

	V1	Mix1_AAAGCAACACTTCGAA	Mix1_AACTCTTCACAGCCCA	Mix1_AAGCCGCAGGAATCGC	Mix
1	FO538757.2	0	0	0	
2	AP006222.2	0	0	0	
3	RP11-206L10.9	1	0	0	
4	LINC00115	0	0	0	
5	FAM41C	0	0	0	
6	RP11-54O7.3	0	0	0	
7	SAMD11	1	0	0	
8	NOC2L	0	2	2	
9	KLHL17	0	0	0	
10	PLEKHN1	0	0	0	
11	HES4	2	0	0	
12	ISG15	0	2	2	
13	AGRN	1	0	1	
14	C1orf159	0	0	0	
15	SDF4	3	2	2	
16	B3GALT6	0	0	0	
17	FAM132A	1	0	0	

Figure 1. Format of the count matrix

This figure shows how the count matrix is formatted. The count matrix is stored in a data.frame object with the gene names stored in a column named "V1" and cell names as the other column names.

Set up project directory

5. Create a main directory named "COMET" and create two subdirectories named "Input_Data", and "Tables":

Filter

	V1	orig.ident	nCount_RNA	nFeature_RNA	percent.mito	Sample	C
1	Mix1_AAAGCAACACTTCGAA	Mix1	16979	3679	0.04505352	A549_TGFB1_3d	A
2	Mix1_AACTCTTCACAGCCCA	Mix1	14496	3380	0.02384398	A549_TGFB1_3d	A
3	Mix1_AAGCCGCAGGAATCGC	Mix1	30190	4554	0.03075295	A549_TGFB1_3d	A
4	Mix1_ACACTGAGTAACGCGA	Mix1	21772	3886	0.05087156	A549_TGFB1_3d	A
5	Mix1_ACACTGATCAACGCTA	Mix1	18431	3724	0.03013713	A549_TGFB1_3d	A
6	Mix1_ACAGCTAAGTAGTGCG	Mix1	15720	3439	0.04594268	A549_TGFB1_7d	A
7	Mix1_ACCTTTACAGACACTT	Mix1	5533	2120	0.07410746	A549_TGFB1_3d	A
8	Mix1_ACGCCAGGTTCTGAAC	Mix1	15984	3257	0.03450647	A549_TGFB1_1d	A
9	Mix1_ACGGCCATCTCCCTGA	Mix1	17217	3488	0.03451677	A549_TGFB1_7d	A
10	Mix1_ACGGGTCAGATATGGT	Mix1	14157	3338	0.08985283	A549_TGFB1_3d	A
11	Mix1_ACTTACTAGTGCGATG	Mix1	21624	3982	0.04480163	A549_TGFB1_1d_rm	A
12	Mix1_AGAATAGAGTATCGAA	Mix1	19787	3896	0.03159382	A549_TGFB1_3d	A
13	Mix1_AGCGGTCTCACCACCT	Mix1	11543	2844	0.02812879	A549_TGFB1_3d	A
14	Mix1_AGCTTGAGTCAGATAA	Mix1	15723	3264	0.04669335	A549_TGFB1_3d	A
15	Mix1_AGCTTGAGTCATTAGC	Mix1	21577	3895	0.04272357	A549_TGFB1_3d	A
16	Mix1_AGGGATGAGATGGCGT	Mix1	17007	3248	0.03278111	A549_TGFB1_3d	A
17	Mix1_AGGGATGTCAGCACAT	Mix1	11025	2717	0.03443277	A549_TGFB1_3d	A

Figure 2. Format of the metadata

This figure shows how the metadata is formatted. The metadata is stored in a data.frame object with the cell names stored in a column named "V1" and (not shown here) a column named "Time" which stores the time points (or dose for dose-dependent data) in a numeric column.

a. Within the "Tables" directory, you must store a file that holds the name of the data and metadata files
b. The input data and metadata files should be stored within the "Input_Data" subdirectory.

Institutional permissions

The data used in this paper for analysis have been previously published and are publicly available. Institutional permissions do not apply.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Analyzed dataset	Cook and Vanderhyden, 2020[5]	GSE147405 (data uploaded on Zenodo: https://zenodo.org/records/10050380)
Software and algorithms		
R (v4.2.1)	R CRAN	https://cran.r-project.org/
RStudio 2023.6.1.524	RStudio	https://www.rstudio.com/
COMET 0.1.0	Najafi et al., 2023[6]	https://github.com/TAMUGeorgeGroup/COMET
MAGIC 2.0.3	Van Dijk et al., 2018[7]	https://github.com/KrishnaswamyLab/MAGIC
Seurat 4.3.0.1	Satija et al., 2015[8]	https://satijalab.org/seurat/
phateR 1.0.7	Moon et al. 2019[9]	https://github.com/KrishnaswamyLab/PHATE#r
umap 0.2.10.0	McInnes et al. 2018[10]	https://github.com/tkonopka/umap
dtw 1.23.1	Toni Giorgino[11]	https://dynamictimewarping.github.io/
pracma 2.4.2	R Foundation	https://cran.r-project.org/web/packages/pracma/index.html
dplyr	R Foundation	https://cran.r-project.org/package=dplyr
ggplot2	R Foundation	https://cran.r-project.org/package=ggplot2
ggpubr	R Foundation	https://cran.r-project.org/package=ggpubr
readxl	R Foundation	https://cran.r-project.org/package=readxl
devtools	R Foundation	https://cran.r-project.org/package=devtools
tidyr	R Foundation	https://cran.r-project.org/web/packages/tidyr/index.html
tidyverse	R Foundation	https://cran.r-project.org/web/packages/tidyverse/index.html
reshape2	R Foundation	https://cran.r-project.org/web/packages/reshape2/index.html
Grid	R Foundation	https://cran.r-project.org/web/packages/grid/index.html
gridExtra	R Foundation	https://cran.r-project.org/web/packages/gridExtra/index.html
diagram	R Foundation	https://cran.r-project.org/web/packages/diagram/index.html
plotly	R Foundation	https://cran.r-project.org/web/packages/plotly/readme/README.html
data.table	R Foundation	https://cran.r-project.org/package=data.table
Kolmogorov-Smirnov test function	Chakraborty et al., 2020[2]	https://github.com/priyanka8993/EMT_score_calculation
Other		
Terra (running on operating system CentOS 7) with 28 cores and 56 GB of memory allocated for each job	Texas A&M University's High Performance Research Computing cluster	https://hprc.tamu.edu/kb/

MATERIALS AND EQUIPMENT

All the analyses performed here were evaluated on Texas A&M University's High Performing Research Computing cluster, Terra (running on operating system CentOS 7) with specifications described within the key resources table. This was also repeated on a personal computer with an Apple M1 Pro chip with 8 cores and 16 GB of memory (running on macOS Monterey).

STEP-BY-STEP METHOD DETAILS

Here, we provide step-by-step instructions for running the COMET pipeline on the time-course scRNA-seq data of Cook et al. 2020.[5] Following the steps described below would allow the user

to infer EMT trajectories and inter-state transition rates. The example data utilized in this section is publicly available on Zenodo (Please refer to the key resources table for more information).

Start the pipeline

Timing: ~5–10 min

COMET requires that the user creates a directory that contains the data they would like to utilize and would be generated by COMET. Prior to running the pipeline, the user should store the data in a directory named "Input_Data" with information about the data and metadata files in a csv file in the "Tables" directory. The path to the "Input_data" and "Tables" directory should be given as input arguments to the "start_pipeline" function of COMET.

1. Load the COMET package in R.

```
#Load relevant libraries

>library(COMET)
```

2. Set the working directory to the "COMET" directory you have created which contains two subdirectories "Input_Data", and "Tables".

   Note: For demonstration purposes you can set the working directory to the example directory mentioned earlier (https://zenodo.org/records/10050380) in the Data Preparation section.

```
>setwd("~/Desktop/COMET_STAR_protocol/")

>input.data.dir <- "Input_Data/"

>tables.dir <- "Tables/"
```

3. Read the input file.

```
>data.inputs <- read.csv(paste0(tables.dir, "DataTableTest.csv"), sep=",")
```

4. Start COMET and indicate the path to your input files.

```
> COMET::start_pipeline(tables.dir, input.data.dir)
```

   Note: You must have a comma separated file stored within the Tables directory with three columns "Sample", "DataPath", "MetaData" which store the sample name (please do not include any spaces), name of the count matrix and metadata respectively.

Find the optimal cutoff of EMT genes

Timing: ~1–12 h

To find the optimal number of EMT genes that would most likely represent the EMT process, we infer trajectories for different number of highly variable EMT genes and assess their similarity to a flow

cytometry data by Jia et al. 2019[12] provided by the package through Dynamic Time Warping (DTW) alignment. The cutoff that minimizes the DTW distance will be chosen as the optimal cutoff.

5. Use COMET to run the pipeline for different number of EMT genes from data and infer EMT trajectories.

   Note: The resulting inferred EMT trajectories will be stored within a subdirectory named "COMET_populated_files". In the code below, the biological sample is going through a reverse mesenchymal to epithelial transition (MET) when coming from the 7.33, 8, and 10 timepoints. The pipeline will store the generated DTW distances for every cutoff in a directory named "DTW_Matrix", and the confidence intervals in the "Confidence_Interval_calculations" directory.

```
>COMET::generate_pipeline_files(data.inputs, >tables.dir, input.data.dir)

>COMET::calculate_conf_intervals(data.inputs)

>COMET::DTW_calculate(data.inputs, c(7.33, 8, 10))
```

Find optimal stochastic model trajectories

Timing: <5 min

A stochastic model is optimally fitted to the data by minimizing the mean squared error between empirical and theoretical trajectories. The mean squared error is calculated for all three trajectories of epithelial, hybrid, and mesenchymal states and summed up.

6. Fit Continuous Time Markov Chain Models to the data.

```
>COMET::fit.all.data(data.inputs,c(7.33, 8, 10))-> final.result
```

Visualize the pipeline results

Timing: <5 min

The code for this section is provided within the vignette of the R package for the purpose of visualizing the results obtained in the previous steps. Please only run the code within this section once you have generated the required files through running the previous steps of the protocol.

7. For a particular dataset, you can visualize the confidence intervals for every cutoff of highly variable EMT genes using the code below.

```
>populated.files.dir <- "COMET_populated_files/"

>h<-1

>for(cutoff in seq(5, 100, 5)){

>setNames(data.frame(matrix(ncol = 3, nrow = 0)),c("time","variable", "value"))->binded

>for(k in 1:10){
```

```
>file.path <- paste0(main.dir, "/", populated.files.dir, >data.input$Sample,"_",k, "_", cutoff,".Rds")
>readRDS(file.path)->data
>rbind(binded, data)->binded
>}
>ggplot(binded, aes(x=time, y=value, group=variable, color=variable, >stroke=1.5),
>fill=c("#24A19C","#D96098", "#BE79DF"))+
>ggtitle(cutoff)+
>geom_rect(aes(xmin = -Inf,xmax = 7,ymin = -Inf, ymax = Inf),
>fill="#DAEAF1",
>alpha = .2)+
>geom_vline(xintercept=7)+
>stat_summary(geom="ribbon", fun.data=mean_cl_normal, width=0.1, >conf.int=0.95, fill =
>c(rep(emt.color.scheme[3], 8), rep(emt.color.scheme[2], 8),
>rep(emt.color.scheme[1], 8)))+
>stat_summary(geom="line", fun.y=mean, linetype="dashed", >fill=c("#24A19C","#D96098",
>"#BE79DF"))+
>stat_summary(geom="point", fun.y=mean, color=c(rep("#24A19C", 8), >rep("#D96098", 8),
>rep("#BE79DF", 8)), shape=8, size=1)+
>scale_color_manual(values=(c("#24A19C","#D96098", "#BE79DF")),
>labels = c("Epithelial", "Hybrid", "Mesenchymal"))+labs(x="Time >(Days)",
>y="Cell Fraction", tag = LETTERS[h], color="States", shape=8)+
>theme(
># Remove panel border
>panel.border=element_blank(),
>#plot.border = element_blank(),
># Remove panel grid lines
>panel.background = element_blank(),
>panel.grid.major = element_blank(),
>panel.grid.minor = element_blank(),
># Add axis line
>axis.line = element_line(colour = "black"),
>#legend.position = "none",
>plot.title = element_text(hjust = 0.5, size=20),
>axis.text = element_text(size = 15),
>text = element_text(size=18)
>) +
>plt <->guides(color=guide_legend(override.aes=list(fill=NA)))+scale_x_continuous(limits=c(0,10))
```

```
>nam <- paste("plt.", h, sep = "")
>assign(nam, plt)
>
>h<-h+1}
>
>ggplot() +            # Draw ggplot2 plot with text only
>annotate("text",
>      x = 1,
>      y = 1,
>      size = 8,
>      label = data.input$Sample) +
>  theme_void()+ theme(panel.background = element_rect(fill = >'#DAEAF1', colour = 'black'),
>text = element_text(size=28))->plt
>grid.arrange(plt, plt.1, plt.2, plt.3, plt.4,
>      plt.5,plt.6, plt.7, plt.8, plt.9, plt.10,
>      plt.11, plt.12, plt.13, plt.14, plt.15,
>      plt.16, plt.17, plt.18, plt.19,
>      top =data.input$Sample, nrow = 4)
```

8. Visualize the heatmap for the DTW distances for every cutoff of highly variable EMT genes.

```
>DTW.dir <- "DTW_Matrix"
>dtw_mat <- readRDS(paste0(main.dir, "/", DTW.dir, "/", >data.input$Sample,
>"_DTW_Matrix.Rds"))
>colnames(dtw_mat)<-c("E", "H", "M", "Total")
>plot_ly(z=dtw_mat, type="heatmap",colors = c(color.scheme[1], >color.scheme[5],
>color.scheme[7], color.scheme[8]))%>% layout(title = list(text = >paste0(data.input
$CellLine,"
>- ", data.input$Factor), y=0.99),
> xaxis = list(title = '  E
>H M Total',zeroline = TRUE, showticklabels = >FALSE), yaxis = list(title = 'Cutoff',
>showticklabels = FALSE, nticks=29))
```

9. We used heatmaps to display the DTW alignment distances acquired through COMET. You can visualize the heatmap for the DTW distances for every cutoff of highly variable EMT genes using the code below.

```
>#To find out about the optimal cutoff, please run the commented >#command outside notebook

>#optimal.cutoff <- find.optimal.cutoff(data.input)

>optimal.cutoff <- 45

>#Get the final inferred trajectories from data

>conf.dat <- readRDS(paste0(main.dir, >"/Confidence_Interval_Calculations/", data.input$-
Sample,"_", >optimal.cutoff, ".Rds"))

>#Get the final fit

>final.result <- readRDS(paste0(main.dir, "/Results/final_df.Rds"))

>final.result[[1]]->final.df

>reshape(

>    conf.dat,

>    idvar = "time",

>    timevar = "variable",

>    direction = "wide"

>)->reshaped_data

>ggplot()+

> geom_line(data=reshaped_data, aes(x=time, y=value.Epithelial), >color= emt.color.
scheme[3], size=1.5)+

> geom_point(data=reshaped_data, aes(x=time, y=value.Epithelial), >color= emt.color.
scheme[3], stroke=3, shape=8)+

> geom_line(data=final.df,  aes(x=time,  y=E_final),  >color=  emt.color.scheme.bold
[3],size=1.5, linetype="dashed")+

> geom_line(data=reshaped_data, aes(x=time, y=value.Hybrid), >color= emt.color.scheme
[2], size=1.5)+

> geom_point(data=reshaped_data, aes(x=time, y=value.Hybrid), >color= emt.color.scheme
[2], stroke=3, shape=8)+

> geom_line(data=final.df, aes(x=time, y=H_final), >color= >emt.color.scheme.bold[2],
size=1.5, linetype="dashed")+

> geom_line(data=reshaped_data, aes(x=time, y=value.Mesenchymal), >color= emt.color.-
scheme[1], size=1.5)+

> geom_point(data=reshaped_data, aes(x=time, y=value.Mesenchymal), >color= emt.color.-
scheme[1], stroke=3, shape=8)+

> geom_line(data=final.df,  aes(x=time,  y=M_final),  >color=  emt.color.scheme.bold[1],
size=1.5, linetype="dashed")+

> ggtitle(data.input$Sample)+

>   geom_rect(aes(xmin = -Inf,xmax = 7,ymin = -Inf, ymax = Inf),

>        fill="#ADC4CE",

>        alpha = .2)+

>   geom_vline(xintercept=7)+

> theme(

>   # Remove panel border

>   panel.border=element_blank(),
```

```
>    #plot.border = element_blank(),
>    # Remove panel grid lines
>    panel.background = element_blank(),
>    panel.grid.major = element_blank(),
>    panel.grid.minor = element_blank(),
>    # Add axis line
>    axis.line = element_line(colour = "black"),
>    #legend.position = "none",
>    plot.title = element_text(hjust = 0.5, size=20),
>    axis.text = element_text(size = 15),
>    text = element_text(size=18))
>    +labs(y="Cell Fraction", x="Time (Days)", tag="A")
```

10. The inter-state transition rates from the CTMC model can be found and visualized by utilizing the code below.

```
>#To find out about the optimal cutoff, please run the commented >command outside notebook
>#optimal.cutoff <- find.optimal.cutoff(data.input)
>optimal.cutoff <- 45
>#Get the final inferred trajectories from data
>conf.dat <- readRDS(paste0(main.dir, >"/Confidence_Interval_Calculations/", data.input$-
Sample, "_", >optimal.cutoff, ".Rds"))
>#Get the final fit
library(diagram)
>M <- matrix(nrow = 3, ncol =3, byrow = TRUE, data = 0)
>A<-M
>M[1,2]<-paste0(final.result[[3]])
>M[2,1]<-paste0(final.result[[2]])
>M[2,3]<-paste0(final.result[[2]])
>M[3,2]<-paste0(final.result[[4]])
>A[1,2]<-paste0(final.result[[2]])
>A[2,1]<-paste0(final.result[[3]])
>A[2,3]<-paste0(final.result[[3]])
>A[3,2]<-paste0(final.result[[4]])
>col <- M
>col[] <- "red"
>col[1, 2] <- "#BEAEE2"
>col[2, 1] <- col[2, 3] <- "#D96098"
>col[3, 2] <- "#24A19C"
```

```
>plotmat(M, pos = c(3), name = c("E", "H", "M"),box.col=rev(emt.color.scheme.bold),
    #box.size=c(0.05,0.03,0.03,0.05), box.prop = 1,
    arr.lwd=A,
    lwd = 1, box.lwd = 2, box.cex = 1, cex.txt = 0.8,
    arr.lcol = col, arr.col = col, box.type = "ellipse",
    lend=3)
```

EXPECTED OUTCOMES

After running the pipeline for different numbers of highly variable EMT genes, the "COMET_populated_files" directory will be filled with the inferred trajectories for different cutoffs of highly variable genes (Figure 3). The results from this directory can be visualized using the code in step 7. An example of the results of this code is shown for the A549 cell line treated with TGF β from Cook et al. 2020[5] in Figure 3 below. COMET relies on a DTW alignment score to a reference flow cytometry data for the identification of the optimal cutoff of highly variable EMT genes (Please refer to Najafi et al. 2023[6] for further information). To visualize the DTW alignment scores to each of the three EMT trajectories and the total score, the code in step 8 can be utilized to generate heatmaps as shown in Figure 4A. Lastly, the code in steps 9 to 10 can be used to visualize the stochastic trajectories fitted to the time-course data and the inter-state transition rates (the results of this code were run on the A549 sample treated with TGF β from the Cook et al. 2020[5] dataset and are shown in Figures 4B and 4C).

QUANTIFICATION AND STATISTICAL ANALYSIS

In step 5, we note that the program run time has a linear dependence on the number of cells in the count matrix (Figure 5B) and is weakly correlated with the number of EMT genes included as shown in Figure 5 below (Figure 5A). The Spearman correlation coefficient (R = 0.99) indicates a linear time complexity (O(n), n is the number of cells) and occurs with fair computational performance. The program run time can be significantly improved upon the allowance of multithreading and parallelization over cores (Figure 5C). The run time relative to the number of tasks is included in the figure below.

LIMITATIONS

Here, we presented an R package for the reliable inference of EMT trajectories from time-course scRNA-seq data. We note that this pipeline has only been tested on cell line data and in the absence of the elements of the tumor microenvironment such as immune cells. Further investigation is required to be able to successfully expand our framework to tumor data in the presence of stroma.

TROUBLESHOOTING

Problem 1

Error in is.null (x = genes) || is.na(x = genes) : 'length = 3' in coercion to 'logical(1)' in Step I when finding the optimal cutoff of highly variable EMT genes.

Potential solution

- This issue is related to the incompatibility of the Rmagic package which is one of the dependencies of RCOMET with the new R version 4.3.1+, please use an alternative R version to fix this issue.

Problem 2

I am getting the following error when running the pipeline with my own data in Step I when finding the optimal cutoff of highly variable EMT genes.

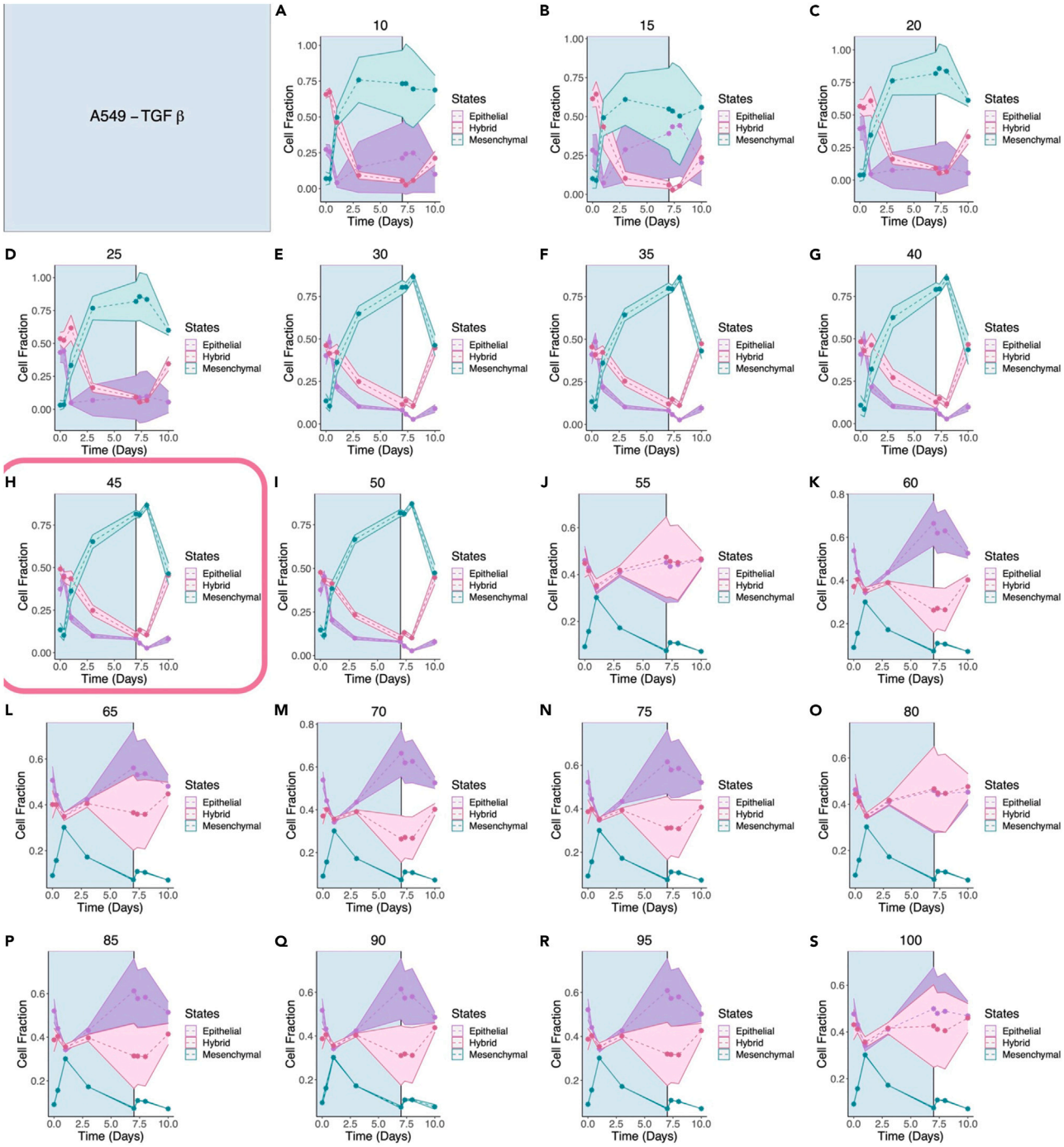


Figure 3. Pipeline predictions for different cutoffs of highly variable EMT genes included

Panels A-S show the inferred EMT trajectories from single-cell RNA sequencing data over a range of cutoffs of highly variable EMT genes from 5 to 100 in increments of 5.

```
Error in '$<-.data.frame'(x, name, value) :

replacement has 0 rows, data has 3568

In addition: Warning message:
```

```
In is.null(x = genes) || is.na(x = genes) :
'length(x) = 5 > 1' in coercion to 'logical(1)'
```

Potential solution

- Please check that you have the time points stored in a numeric column named "Time" within the metadata file. Failure to do so would result in this error. We require that the user changes the format of the metadata to adhere to guidelines (Figure 6).

Problem 3

I would like to apply this methodology to another biological context and thus would need to change the loaded parameters such as imported EMT genes.

Potential solution

- After installing the R package from GitHub, you can change the contents of the data stored within the "extdata" directory to your parameters of interest. These parameters are loaded by default when you start the pipeline.

Problem 4

I have dose-dependent data of cell lines treated with various doses of an EMT transcription factor that I would like to use for the inference of EMT trajectories using COMET.

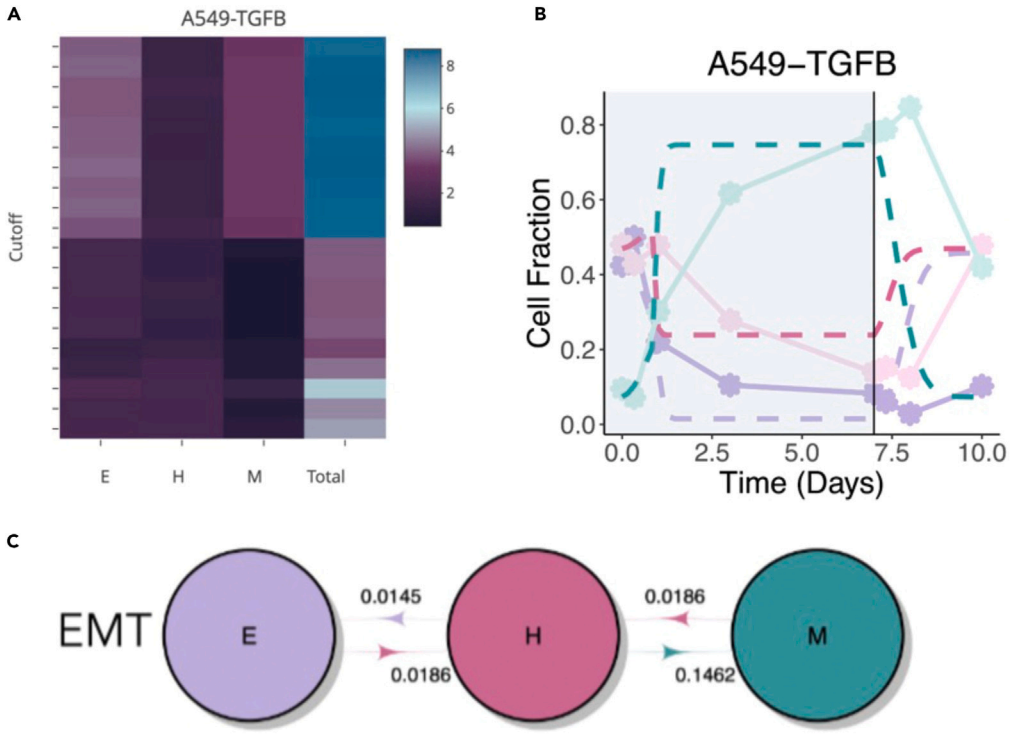


Figure 4. Visualization of the results of COMET

(A) depicts the DTW distance calculated over a range of highly variable EMT genes from 5 to 100 in increments of 5. (B) shows the CTMC model trajectories fitted to the inferred EMT trajectories from data. (C) illustrates the inter-state transition rates calculated from the optimal fit of the CTMC model.

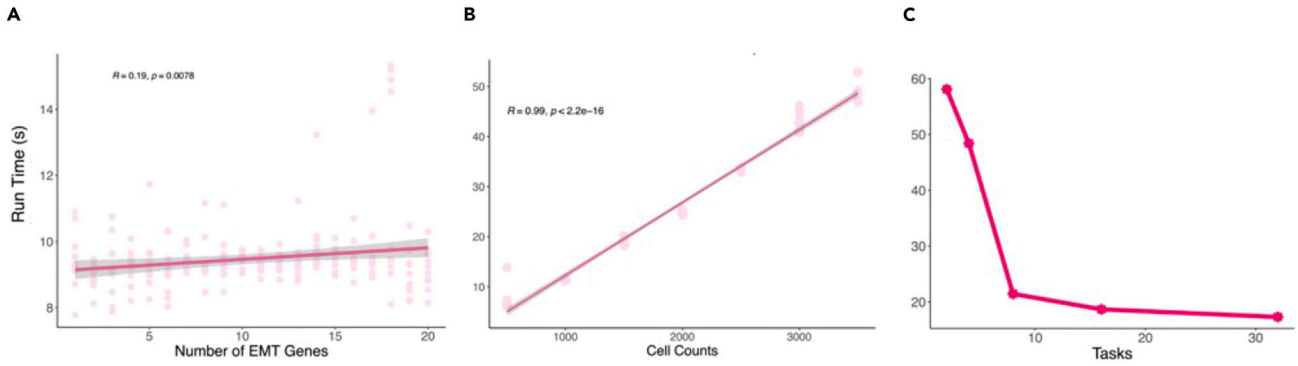


Figure 5. Run time as a function of cell counts

(A) This figure shows the run time of the algorithm over different numbers of EMT genes which indicates a weak correlation between the number of EMT genes and the run time. A regression line was fitted to the data which was obtained from 10 runs of the algorithm.

(B) Shows a regression model fitted to the run time of the algorithm (10 cases per cell count) which indicates a Pearson correlation coefficient of 0.99.

(C) Shows the run time of the algorithm over the number of tasks.

Potential solution

- Please change the metadata of the file such that the "Time" column stores the corresponding dose of the EMT transcription factor. Please note that the column should still be in a numeric format.

Problem 5

I have multiple scRNA-seq datasets coming from the same experiment but different timepoints and different batches, how do I run COMET on my dataset?

Potential solution

- COMET requires that your scRNA-seq datasets are quality controlled, merged (if multiple single-cell RNA sequencing files related to different time points such as data from 24 h, 36 h, etc.), and batch corrected prior to EMT trajectory inference (batch effects resulting from experiments

```
> meta.data
      V1                     V1 orig.ident nCount_RNA nFeature_RNA percent.mito          Sample CellLine
  1:   1  Mix1_AAAGCAACACTTCGAA       Mix1      16979         3679   0.04505352    A549_TGFB1_3d     A549
  2:   2  Mix1_AACTCTTCACAGCCCA       Mix1      14496         3380   0.02384398    A549_TGFB1_3d     A549
  3:   3  Mix1_AAGCCGCAGGAATCGC       Mix1      30190         4554   0.03075295    A549_TGFB1_3d     A549
  4:   4  Mix1_ACACTGAGTAACGCGA       Mix1      21772         3886   0.05087156    A549_TGFB1_3d     A549
  5:   5  Mix1_ACACTGATCAACGCTA       Mix1      18431         3724   0.03013713    A549_TGFB1_3d     A549
 ---
496: 496 Mix3b_CTACATTAGTCACGCC      Mix3b      15845         3629   0.02200227    A549_TGFB1_7d     A549
497: 497 Mix3b_CTACCCATCCGGGTGT      Mix3b      20366         3895   0.02202276    A549_TGFB1_3d     A549
498: 498 Mix3b_CTACGTCTCAACACGT      Mix3b      16630         3700   0.02583048    A549_TGFB1_3d     A549
499: 499 Mix3b_CTCATTAAGTGGTAAT      Mix3b      20910         4303   0.04563265    A549_TGFB1_3d     A549
500: 500 Mix3b_CTCGAAAGTCCAGTTA      Mix3b      27729         4755   0.04199986 A549_TGFB1_1d_rm     A549
     Treatment Time Doublet     S.Score  G2M.Score Phase   Mix SCT_snn_res.0.8 SCT_snn_res.0.1
  1:     TGFB1    3 Singlet  0.44107571 -1.3621855     S  Mix1               9               0
  2:     TGFB1    3 Singlet  0.01627270  0.6234093   G2M  Mix1               9               0
  3:     TGFB1    3 Singlet  1.42214800 -1.6454059     S  Mix1               9               0
  4:     TGFB1    3 Singlet -0.44278863 -1.6124782    G1  Mix1               9               0
  5:     TGFB1    3 Singlet  0.42702980 -1.0752030     S  Mix1               9               0
 ---
496:     TGFB1    7 Singlet -0.01670092 -1.2418760    G1 Mix3b               9               0
497:     TGFB1    3 Singlet  0.28682768 -0.5169677     S Mix3b               9               0
498:     TGFB1    3 Singlet -0.09018499 -1.8303710    G1 Mix3b               9               0
499:     TGFB1    3 Singlet -0.15493320  0.9431755   G2M Mix3b               9               0
500:     TGFB1    8 Singlet -0.31928743  2.2497361   G2M Mix3b               9               0
```

Figure 6. An overview of an acceptable metadata file for COMET

COMET requires the metadata file to include a column named "Time" with numeric values.

performed in different experimental settings). Please ensure that typical quality controlling steps such as checking for mitochondrial percentage and batch adjustment is performed on data prior to running the COMET pipeline.

RESOURCE AVAILABILITY

Lead contact

For further information and requests for resources and code availability, please contact Dr. Jason T. George (Jason.george@tamu.edu).

Technical contact

Technical questions on executing this protocol should be directed to and will be answered by the technical contact, Annice Najafi (annicenajafi@tamu.edu).

Materials availability

This study did not generate any new reagents.

Data and code availability

- The dataset used by this paper has been published by Cook et al. 2020.[5]
- The example processed dataset is archived and available on Zenodo: https://doi.org/10.5281/zenodo.10050380
- The specific version of the package (v0.0.1) utilized within this manuscript is archived on Zenodo: https://doi.org/10.5281/zenodo.10327903
- All of the analyses performed within this manuscript has been conducted in R and is published on GitHub: https://github.com/TAMUGeorgeGroup/COMET.
- Additional information and MATLAB code is provided on the GitHub page, https://github.com/TAMUGeorgeGroup/Stochastic_EMT_2023.

ACKNOWLEDGMENTS

We utilized the Terra cluster from Texas A&M University's High Performance Research Computing clusters to test the run time of our pipeline. We would like to thank Abhijeet Deshmukh and Sendurai A. Mani from UT MD Anderson Cancer Center for making the flow cytometry data used within this protocol available to us. J.T.G. was supported by the Cancer Prevention and Research Institute of Texas (RR210080) and is a CPRIT Scholar in Cancer Research. M.K.J. was supported by the Ramanujan Fellowship (SB/S2/RJN-049/2018) awarded by the Science and Engineering Research Board (SERB), Department of Science and Technology, Government of India.

AUTHOR CONTRIBUTIONS

Research design and methodology, A.N., J.T.G., and M.K.J.; data analysis and interpretation, A.N., J.T.G., and M.K.J.; writing – review and editing, A.N., J.T.G., and M.K.J.; conceptualization, J.T.G. and A.N.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Brabletz, T., Kalluri, R., Nieto, M.A., and Weinberg, R.A. (2018). EMT in cancer. Nat. Rev. Cancer *18*, 128–134.

2. Chakraborty, P., George, J.T., Tripathi, S., Levine, H., and Jolly, M.K. (2020). Comparative study of transcriptomics-based scoring metrics for the epithelial-hybrid-mesenchymal spectrum. Front. Bioeng. Biotechnol. *8*, 220.

3. George, J.T., Jolly, M.K., Xu, S., Somarelli, J.A., and Levine, H. (2017). Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric. Cancer Res. *77*, 6415–6428.

4. Jolly, M.K., Boareto, M., Huang, B., Jia, D., Lu, M., Ben-Jacob, E., Onuchic, J.N., Levine, H., and Levine, H. (2015). Implications of the hybrid epithelial/mesenchymal

phenotype in metastasis. Front. Oncol. *5*, 155.

5. Cook, D.P., and Vanderhyden, B.C. (2020). Context specificity of the EMT transcriptional response. Nat. Commun. *11*, 2142.

6. Najafi, A., Jolly, M.K., and George, J.T. (2023). Population Dynamics of EMT Elucidates the Timing and Distribution of Phenotypic Intra-tumoral Heterogeneity. iScience *26*, 106964.

7. Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. Cell *174*, 716–729.e27.

8. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

9. Moon, K.R., Stanley, J.S., III, Burkhardt, D., van Dijk, D., Wolf, G., and Krishnaswamy, S. (2018). Manifold learning-based methods for analyzing single-cell RNA-sequencing data. Curr. Opin. Syst. Biol. *7*, 36–46.

10. McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at arXiv.

11. Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. J. Stat. Softw. *31*, 1–24.

12. Jia, W., Deshmukh, A., Mani, S.A., Jolly, M.K., and Levine, H. (2019). A possible role for epigenetic feedback regulation in the dynamics of the epithelial–mesenchymal transition (EMT). Phys. Biol. *16*, 066004.