CALISTA: User Manual

Version 1.0.1 (6 March 2018)

Authors: Nan Papili Gao and Rudiyanto Gunawan
Institute for Chemical and Bioengineering
ETH Zurich
Contact e-mail: nanp@ethz.ch

Table of contents

1 Overview

This user manual is for the MATLAB distribution of CALISTA (Clustering And Lineage Inference in Single Cell Transcriptional Analysis).

CALISTA provides a user-friendly toolbox for the analysis of single cell expression data. CALISTA accomplishes three major tasks:

(1) Identification of cell clusters in a cell population based on single-cell gene expression data;
(2) Reconstruction of lineage progression and produce transition genes;
(3) Pseudotemporal ordering of cells along any given developmental paths in the lineage progression.

For detailed information about CALISTA, please refer to the following manuscript.
Papili Gao N., Hartmann T. and Gunawan R., **CALISTA: Clustering and lineage inference in single-cell transcriptional analysis**, bioRxiv, 2018. https://doi.org/10.1101/257550

DEVELOPER NOTE: CALISTA is originally implemented in MATLAB and all results described in the related manuscript are obtained by the corresponding MATLAB version of the software. Although the R version of CALISTA represents an accurate translation of CALISTA functions written in MATLAB, the results from the two programming languages might slightly diverge due to the use of different packages and subroutines.

2 System requirements

This distribution of CALISTA is written for and developed in R[1]. CALISTA (version 1.0) has been successfully tested on R version 3.4.3 and RStudio Version 1.1.423

2.1 Required R packages

(to quickly install all required packages please run the commands in **install_packs.R**)

```
library(R.matlab)
library(doParallel)
library(foreach)
library(Matrix)
library(cluster)
library(ClusterR)
library(tsne)
library(destiny)
library(expm)
library(Rcpp)
library(openxlsx)
library(data.table)
library(tools)
library(scatterplot3d)
library(plot3D)
library(rgl)
library(ggplot2)
library(igraph)
library(tcltk2)
library(gridExtra)
library(primes)
library(numbers)
library(matrixStats)
library(TTR)
library(ppcor)
library(Hmisc)
library(fields)
library(gplots)
```

[1] https://www.r-project.org/

3 CALISTA package

CALISTA package contains the following files:

1. This `CALISTA_USER_MANUAL.doc` file
2. `License.txt` modified BSD license for CALISTA
3. `MAIN.R` example script on how to use CALISTA subroutines
4. The folder `subfunctions` containing the following main subroutines (and other subroutines):
    a. `import_data.R` : upload single-cell expression data and perform preprocessing.
    b. `CALISTA_clustering_main.R:` single-cell clustering in CALISTA.
    c. `CALISTA_transition_main.R:` infer lineage progression among cell clusters.
    d. `CALISTA_transition_genes_main.R:` identify the key genes in lineage progression.
    e. `CALISTA_ordering_main.R:` perform pseudotemporal ordering of cells.
    f. `CALISTA_path_main.R:` perform post-analysis along developmental path(s).
    g. The folder `Two-state model parameters` containing:
        i. `Parameters.mat` steady-state distribution functions of mRNA level
5. The folder `EXAMPLES` containing single-cell expression datasets used in the examples below.

For further information on running the main subroutines in CALISTA, please check comments contained in the previous subroutines.

4 Examples

In the following, we describe the main steps of CALISTA applied to publicly available single-cell gene expression data. For each dataset, ONLY the most important results are reported. Please refer to the file `MAIN.R` for an example R script of CALISTA implementation.

4.1 Example 1. iPSC differentiation into mesodermal and endodermal cells

Analysis of RT-qPCR data of Bargaje et al. (Bargaje, et al, Cell population structure prior to bifurcation predicts efficiency of directed differentiation in human induced pluripotent cells. *Proc. Natl. Acad. Sci. U. S. A*. 114, 2271–2276 (2017)).

4.1.1 Data Import and Preprocessing

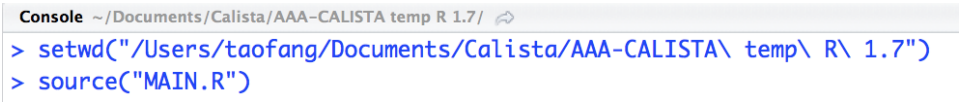
We begin by editing the `MAIN.R` script in the main folder of CALISTA and set the fields of `INPUTS` as follows:

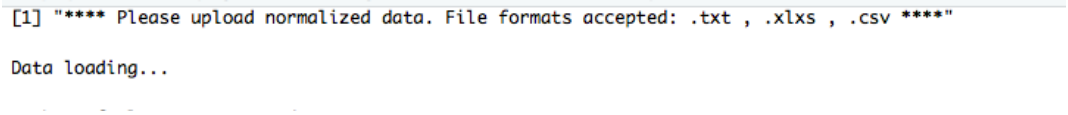
```
% Specify data types and settings for pre-processing
INPUTS$data_type=1; % Single-cell RT-qPCR CT data
INPUTS$format_data=1; % Rows= cells and Columns= genes with time/stage info in the
last column
INPUTS$data_selection=integer(); % Include data from all time points
INPUTS$perczeros_genes=100; % Remove genes with > 100% of zeros
INPUTS$perczeros_cells=100; % Remove cells with 100% of zeros
INPUTS$cells_2_cut=0; % No manual removal of cells
INPUTS$perc_top_genes=100; % Retain only top X the most variable genes with X=min(200,
INPUTS$perc_top_genes * num of cells/100, num of genes)
% Specify single-cell clustering settings
INPUTS$optimize=1; % Set the number of clusters based on Eigengap Heuristics
INPUTS$parallel=1; % Use parallelization option
INPUTS$runs=50; % Perform 50 independent runs of greedy algorithm
INPUTS$max_iter=100; % Limit the number of iterations in greedy algorithm to 100
INPUTS$cluster='kmedoids'; % Use k-medoids in consensus clustering
% Specify transition genes settings
INPUTS$thr_transition_genes=50; % Set threshold for transition genes determination to
50%
% Specify path analysis settings
INPUTS$plot_fig=1; % Plot figure of smoothened gene expression along path
INPUTS$hclustering=1; % Perform hierarchical clustering of gene expression for each
path
INPUTS$method=2; % Use pairwise correlation for the gene co-expression network
(value_cutoff=0.8, pvalue_cutoff=0.01)
```

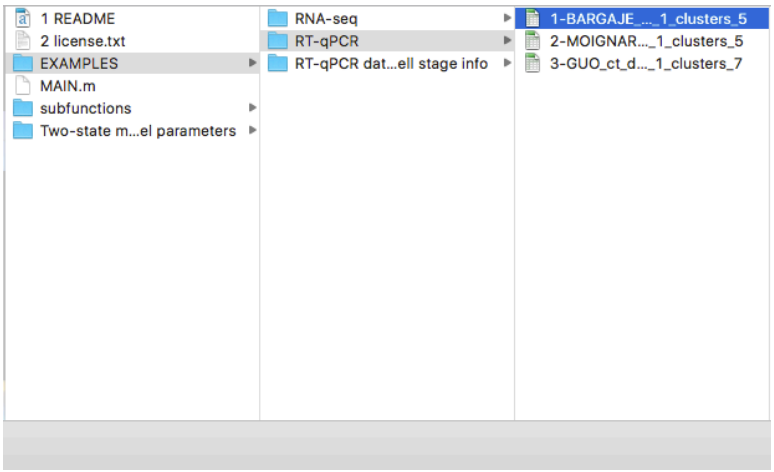
```
INPUTS$moving_average_window=10; % Set the size of window (percent of cells in each path) used for the moving averaging
```

Then, we change the current directory in R to the CALISTA folder (with "`setwd`" command). Subsequently, we run `MAIN.R` and import Bargaje dataset (available in the subfolder EXAMPLES/RT-qPCR).

The following are screenshots from running CALISTA on R.

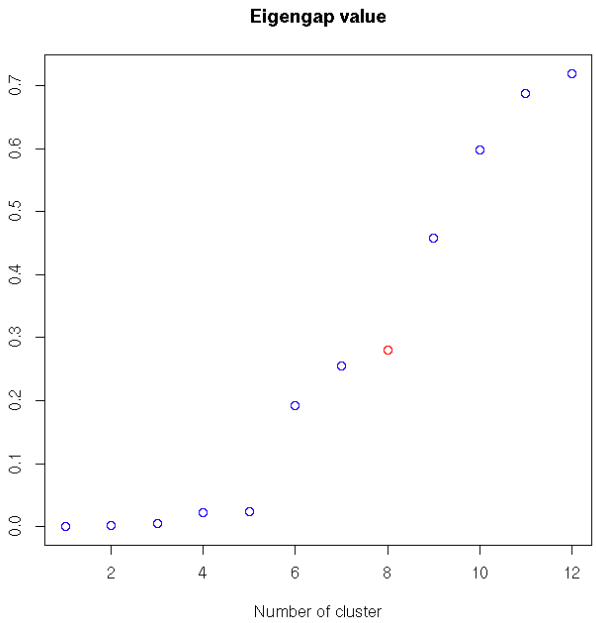




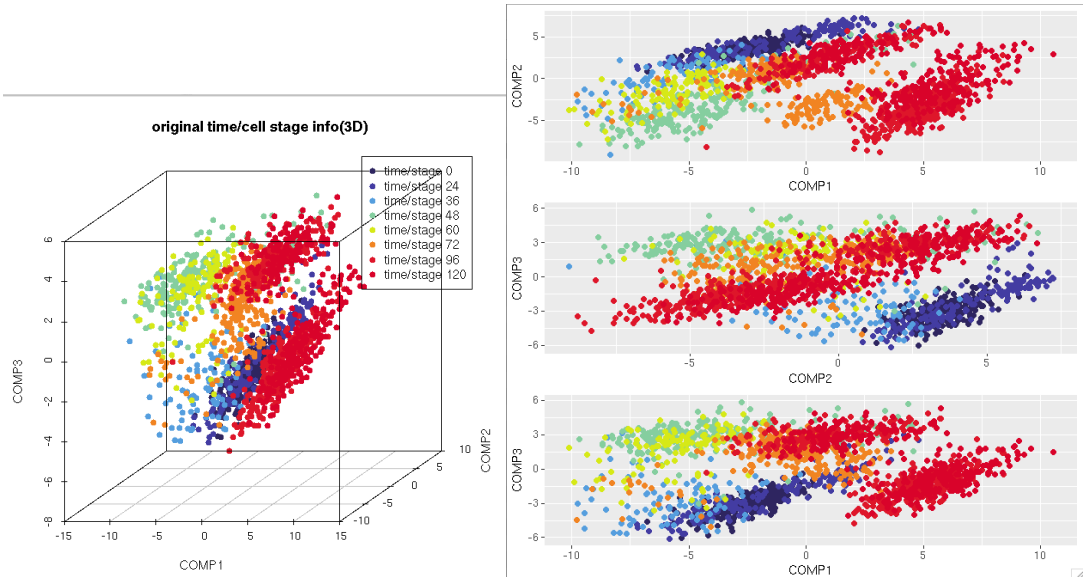


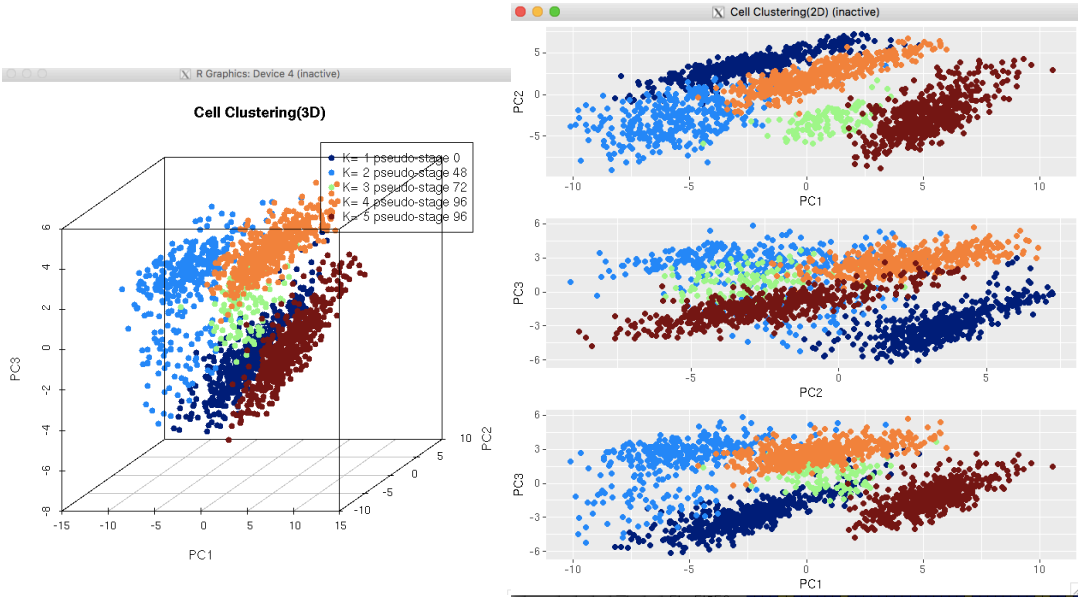
4.1.2 Single-cell clustering

In this case, the number of clusters is determined using the eigengap plot. According to the eigengap plot below, we set the number of clusters to **5**.



NOTE: CALISTA automatically returns the optimal number of clusters based on the MAXIMUM eigengap value. However, the user might choose the number of clusters to adopt based on the FIRST eigengap.





If desired, users can remove cells from specific clusters from further analysis. In this example, we do not want to remove any clusters. Hence, we enter **0** (no cluster removal) and then **1** to proceed with lineage inference.

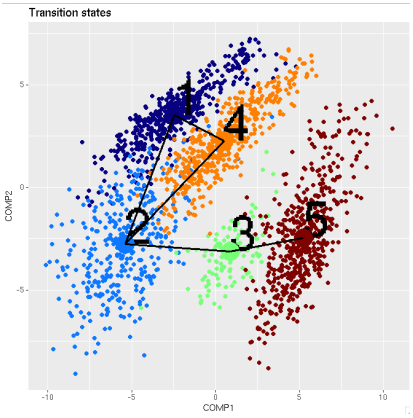
```
Plotting...

Press 1 if you want to remove one cell cluster, 0 otherwise:
0
 Press 1 if you want to perform additional analysis (e.g. lineage inference, cell ordering) , 0 otherwise:
1

CALISTA_transtion is running...
```

4.1.3 Reconstruction of lineage progression

During the lineage inference step, CALISTA automatically generates and displays a lineage graph, obtained by adding an edge between two clusters in increasing cluster distances, until all clusters are connected to at least one other cluster. Subsequently, users can manually add or remove one edge at time based on the cluster distances.

```
CALISTA_transtion is running...


5 edge(s) have been added and the graph is connected.
If you want to add another edge press "1" (then Enter)
If you want to remove an edge press "2" (then Enter)
If you want to continue with the next step press "0" (then Enter)

0
```

ATTENTION: to add an edge (press "**1**"), remove an edge (press "**2**") or finalize the lineage progression graph (press "**0**"), the MATLAB figure of the graph must appear in foreground without any modification (e.g., zooming, rotation). Note that the addition/removal of the edges are performed according to increasing/decreasing order of cluster distance.

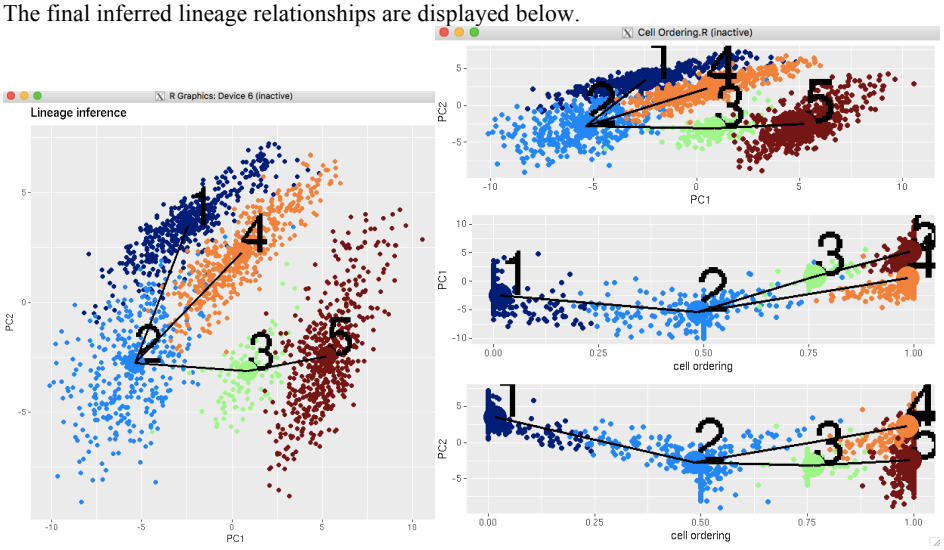


ATTENTION: the final graph must be connected (i.e. there exists a path from any node/cluster to any other node/cluster in the graph), otherwise a warning will be returned.

Since the transition from cluster 1 to cluster 4 is inconsistent with the capture time info (i.e. cluster pseudotime values for cluster 1 and 4 are 0 and 1 respectively) we remove the spurious edge between cluster 1 and 4, by entering **1** and entering **[1 4]**, upon the following query.

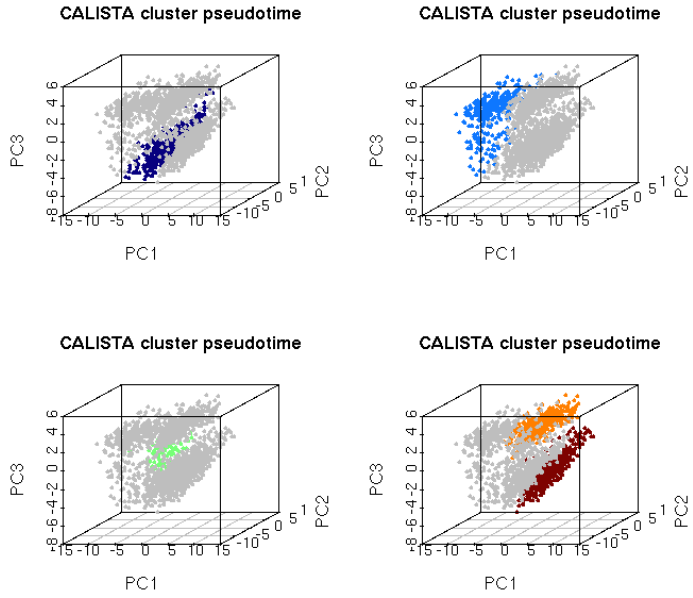
```
Press 1 if you want to remove edges, 0 otherwise:
1

 ********************************************************

Specify the node pairs (e.g. 4 5):
1 4
 Press 1 to remove another edge, 0 otherwise:
0
1 edges to remove
```

The final inferred lineage relationships are displayed below.

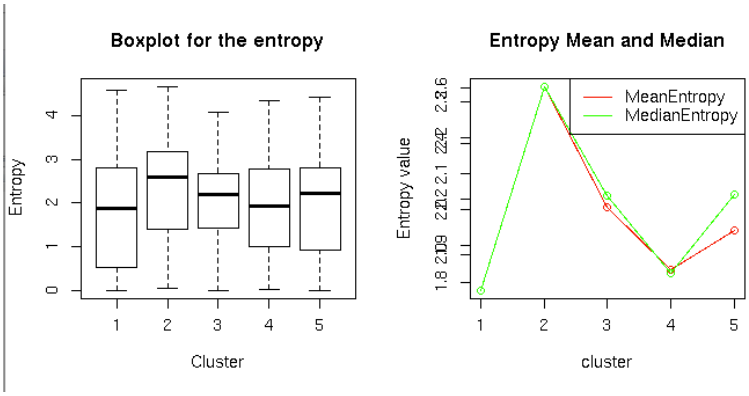


In addition, CALISTA provides the following.

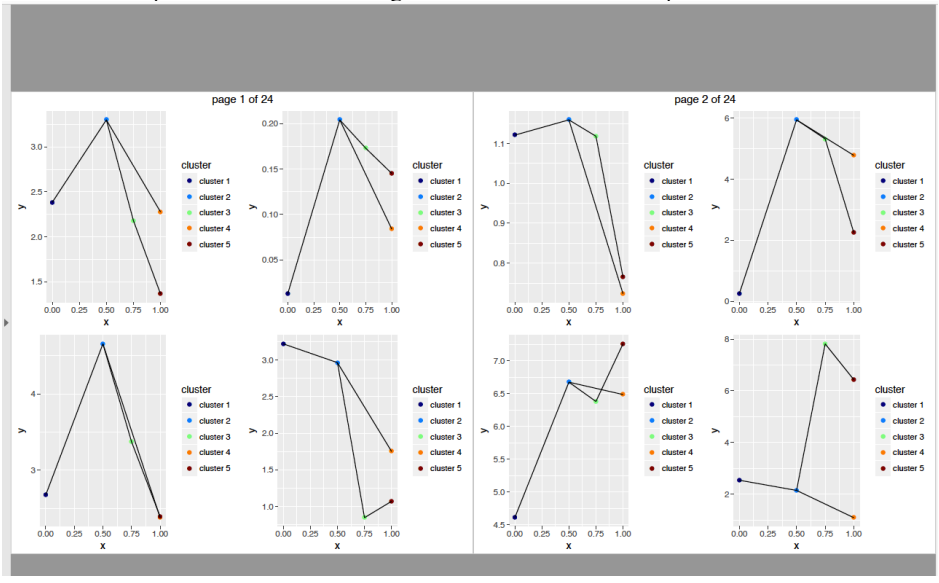
- Cell clustering plot based on the cluster pseudotime



- Boxplot, mean, median entropy values calculated for each cluster

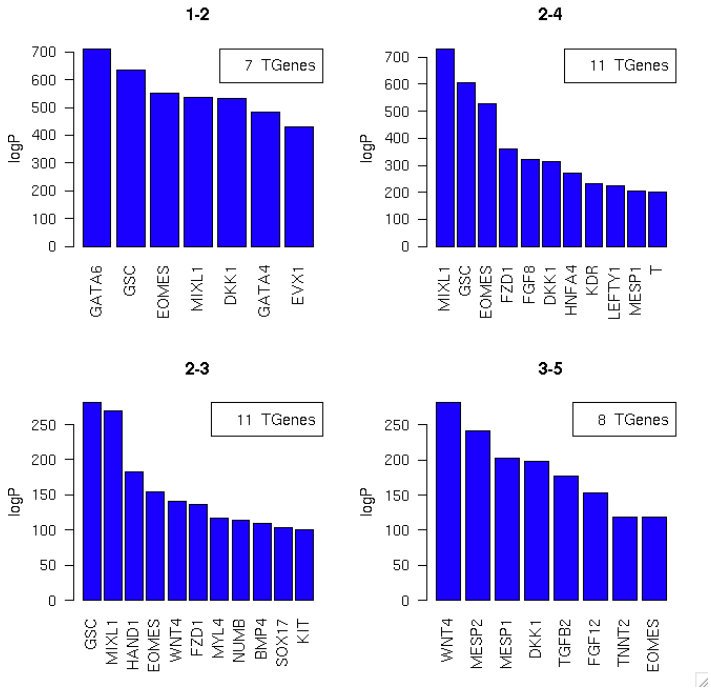


- Plot of mean expression values for each gene based on cell cluster expression level



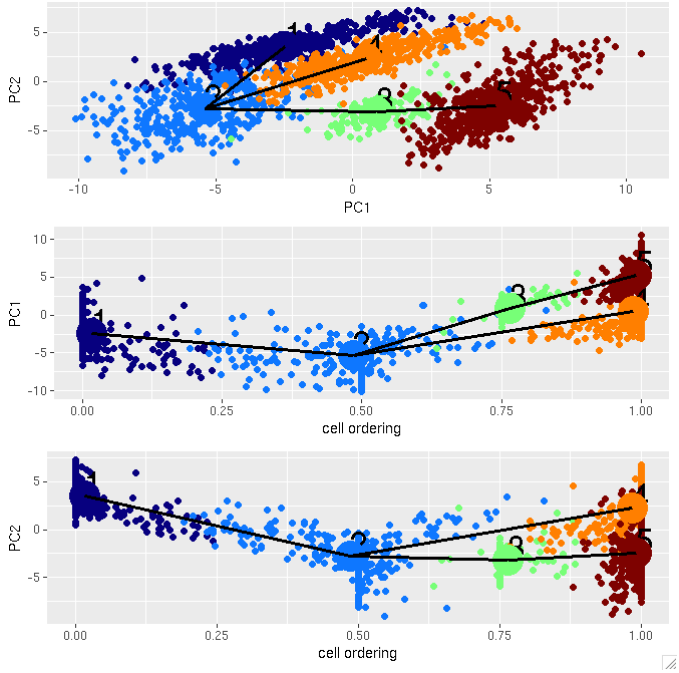
4.1.4 Determination of transition genes

After reconstructing the lineage progression, we identify the key transition genes for any two connected clusters in the graph, based on the gene-wise likelihood difference between having the cells separately as two clusters and together as a single cluster. Larger differences in the gene-wise likelihood point to more informative genes. The transition genes are selected as those whose gene-wise likelihood differences make up to more than a certain percentage of the cumulative sum of the likelihood differences of all genes – set by `INPUTS$thr_transition_genes`.



4.1.5 Pseudotemporal ordering of cells

For pseudotemporal ordering of cells, CALISTA performs maximum likelihood optimization for each cell using a linear interpolation of the cell likelihoods between any two connected clusters. The pseudotimes of the cells are computed by linear interpolation of the cluster pseudotimes, and correspond to the maximum point of the likelihood optimization above. Cells are subsequently assigned to the edges in the lineage progression graph. The following screenshot gives the results of this cell-to-edge assignment.

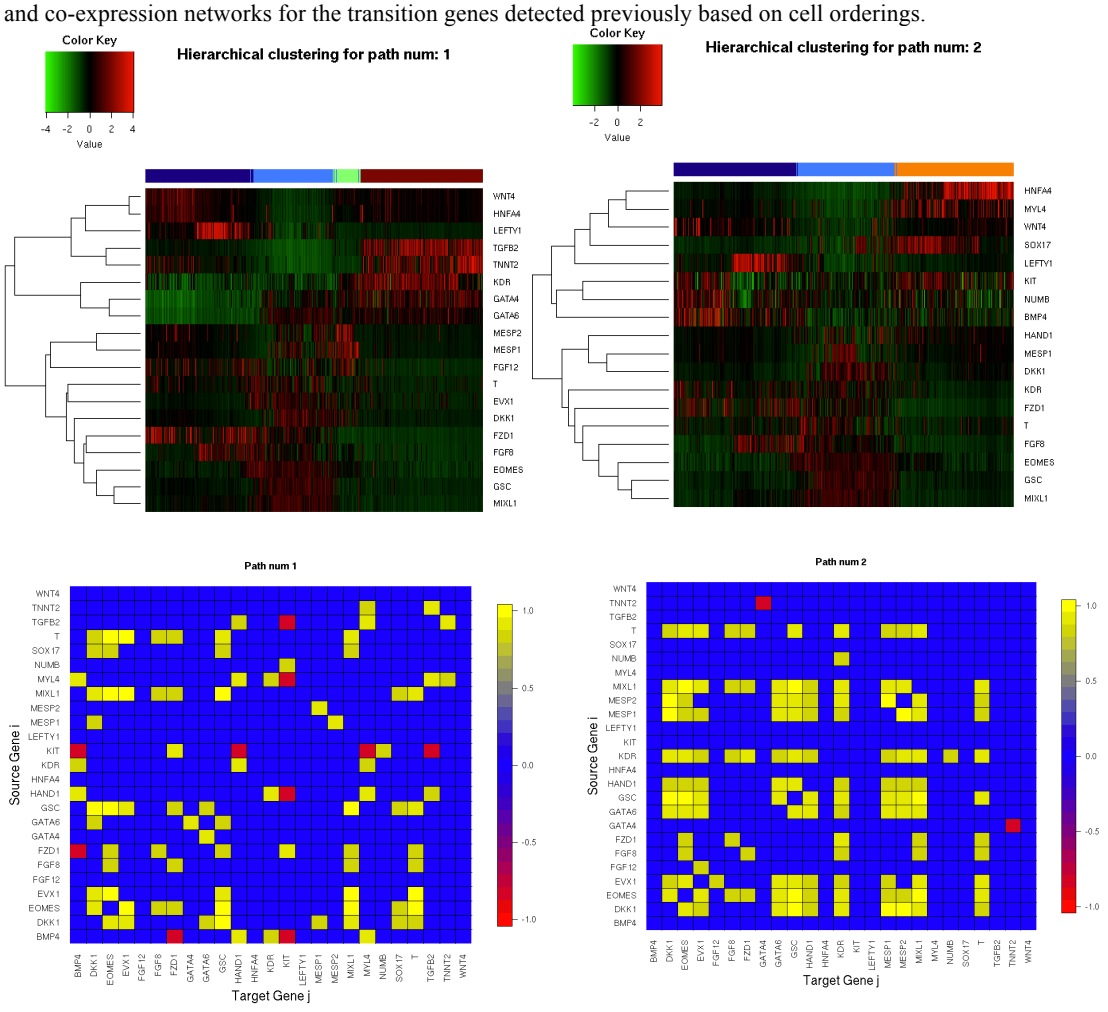


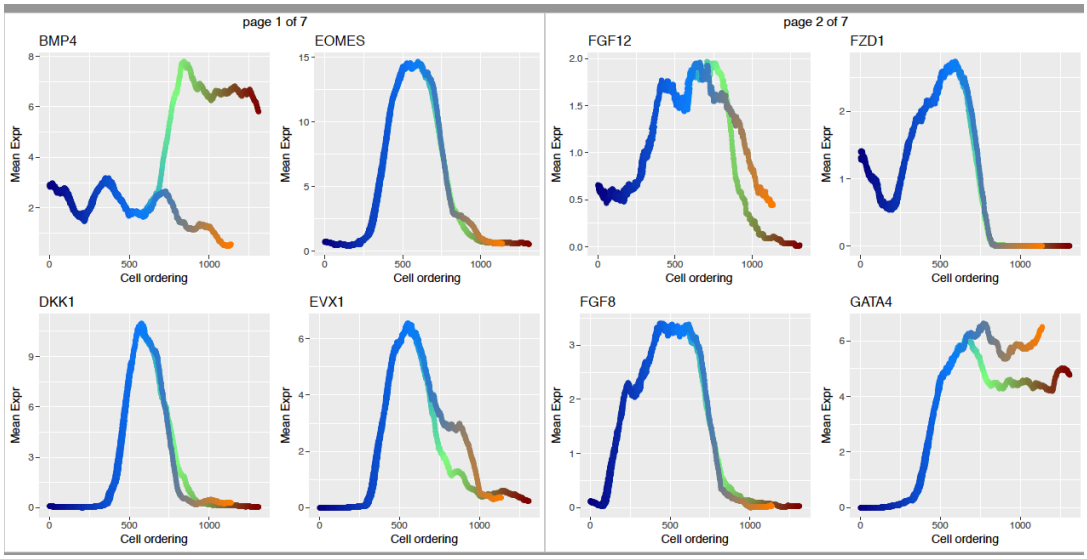
4.1.6 Path analysis

To perform post-analysis, users can enter **1** upon queried. In the following, we input two developmental paths of interest: **[1 2 3 5]** (mesodermal fate) and **[1 2 4]** (endodermal fate).

```
Press 1 if you want to select transition paths and perform additional analysis, 0 otherwise:
1
Path num:  1

**********************************

Please check the figure 'Cell Ordering.R' for reference and
Key the clusters in the path (e.g. 1 2 3 4):
1 2 3 5
Please 1 to add another path, 0 otherwise:
1
Path num:  2

**********************************

Please check the figure 'Cell Ordering.R' for reference and
Key the clusters in the path (e.g. 1 2 3 4):
1 2 4
Please 1 to add another path, 0 otherwise:
0
```

For each path, the post-analysis in CALISTA generates Clustergrams, moving-averaged gene expression profiles and co-expression networks for the transition genes detected previously based on cell orderings.





4.2 Example 2. Hematopoietic stem cell differentiation

Analysis of RT-qPCR data in Moignard et al., Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis, *Nat. Cell Biol*. **15**, 363–72 (2013).

4.2.1 Data Import and Preprocessing

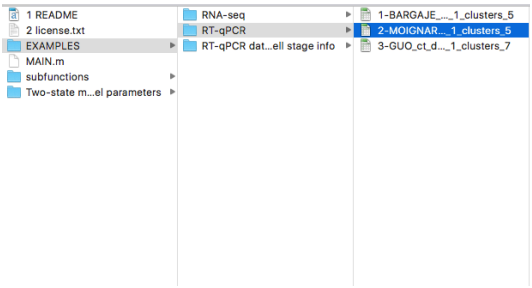
We begin by editing the `MAIN.R` script in the main folder of CALISTA and set the fields of `INPUTS` as follows:

```
% Specify data types and settings for pre-processing
INPUTS$data_type=1; % Single-cell RT-qPCR CT data
INPUTS$format_data=1; % Rows= cells and Columns= genes with time/stage info in the
last column
INPUTS$data_selection= integer();; % Include data from all time points
INPUTS$perczeros_genes=100; % Remove genes with > 100% of zeros
INPUTS$perczeros_cells=100; % Remove cells with 100% of zeros
INPUTS$cells_2_cut=0; % No manual removal of cells
INPUTS$perc_top_genes=100; % Retain only top X the most variable genes with X=min(200,
INPUTS$perc_top_genes * num of cells/100, num of genes)
% Specify single-cell clustering settings
INPUTS$optimize=0; % The number of cluster is known a priori
INPUTS$parallel=1; % Use parallelization option
INPUTS$runs=50; % Perform 50 independent runs of greedy algorithm
INPUTS$max_iter=100; % Limit the number of iterations in greedy algorithm to 100
INPUTS$cluster='kmedoids'; % Use k-medoids in consensus clustering
% Specify transition genes settings
INPUTS$thr_transition_genes=50; % Set threshold for transition genes determination to
50%
% Specify path analysis settings
INPUTS$plot_fig=1; % Plot figure of smoothened gene expression along path
INPUTS$hclustering=1; % Perform hierarchical clustering of gene expression for each
path
INPUTS$method=2; % Use pairwise correlation for the gene co-expression network
(value_cutoff=0.8, pvalue_cutoff=0.01)
INPUTS$moving_average_window=10; % Set the size of window (percent of cells in each
path) used for the moving averaging
```

The we change the current directory in R to the CALISTA folder (use "`setwd`" command).
Subsequently we run `MAIN.R` and import Bargaje dataset (available in the subfolder EXAMPLES/RT-qPCR).

The following are screenshots from running CALISTA on R.

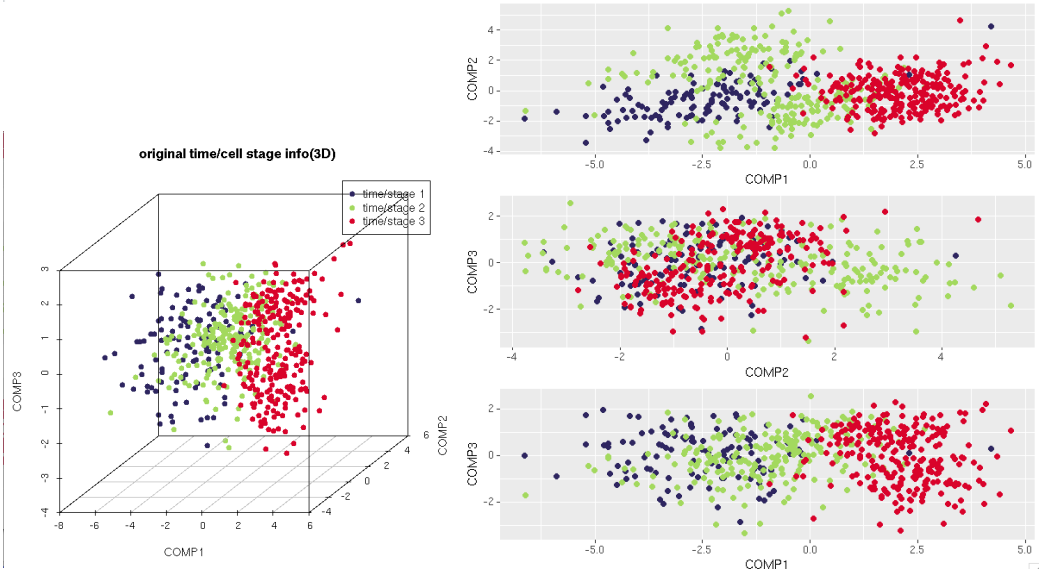
```
Console ~/Documents/Calista/AAA-CALISTA temp R 1.7/
> setwd("/Users/taofang/Documents/Calista/AAA-CALISTA\ temp\ R\ 1.7")
> source("MAIN.R")
```

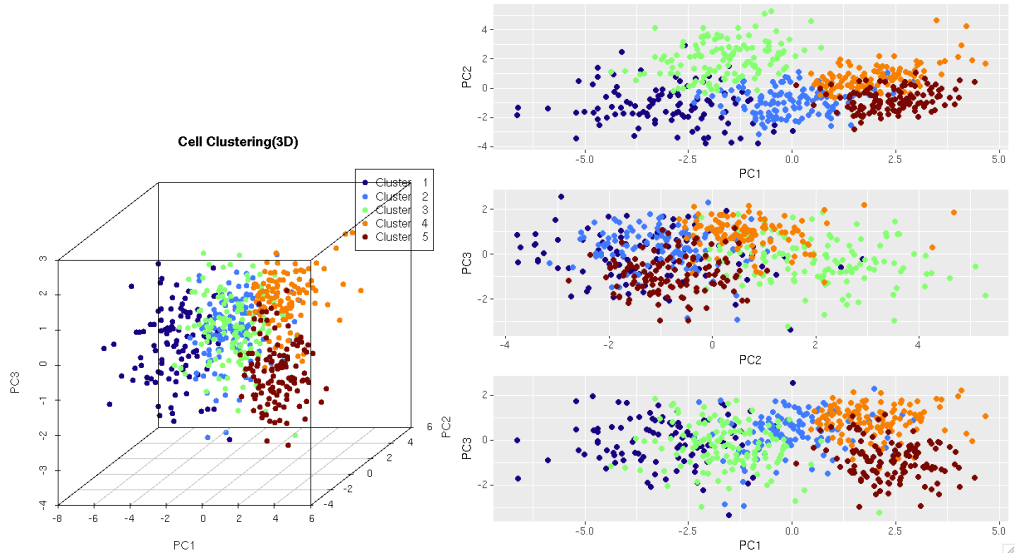


4.2.2 Single-cell clustering

Following the original publication, we set the number of clusters equals to **5**.

```
Number of clusters expected:
5

CALISTA_clustering is running...
```





In this case, we do not need to remove any clusters (by pressing **0** upon queried). Then we proceed with further analysis (by pressing **1** upon queried).

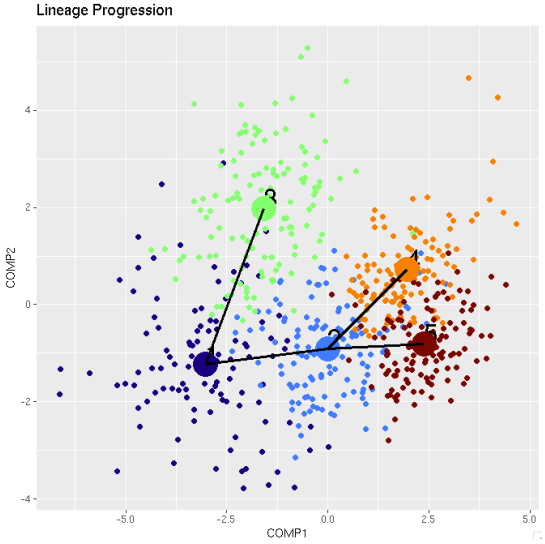
```
Plotting...

Press 1 if you want to remove one cell cluster, 0 otherwise:
0
 Press 1 if you want to perform additional analysis (e.g. lineage inference, cell ordering) , 0 otherwise:
1

CALISTA_transtion is running...
```

4.2.3 Reconstruction of lineage progression

During the lineage inference step, CALISTA automatically generates and displays a lineage graph, obtained by adding an edge between two clusters in increasing cluster distances, until all clusters are connected to at least one other cluster. Subsequently, users can manually add or remove one edge at time based on the cluster distances.

```
CALISTA_transtion is running...


4 edge(s) have been added and the graph is connected.
If you want to add another edge press "1" (then Enter)
If you want to remove an edge press "2" (then Enter)
If you want to continue with the next step press "0" (then Enter)

0
```

ATTENTION: to add an edge (press "**1**"), remove an edge (press "**2**") or finalize the lineage progression graph (press "**0**"), the MATLAB figure of the graph must appear in foreground without any modification (e.g., zooming, rotation). Note that the addition/removal of the edges are performed according to increasing/decreasing order of cluster distance.



ATTENTION: The final lineage progression graph must be connected (i.e. there is a path from any node/cluster to any other node/cluster in the graph) otherwise a warning will be returned.

Here, we do not need to remove any spurious edges, and hence we enter **0** upon queried.

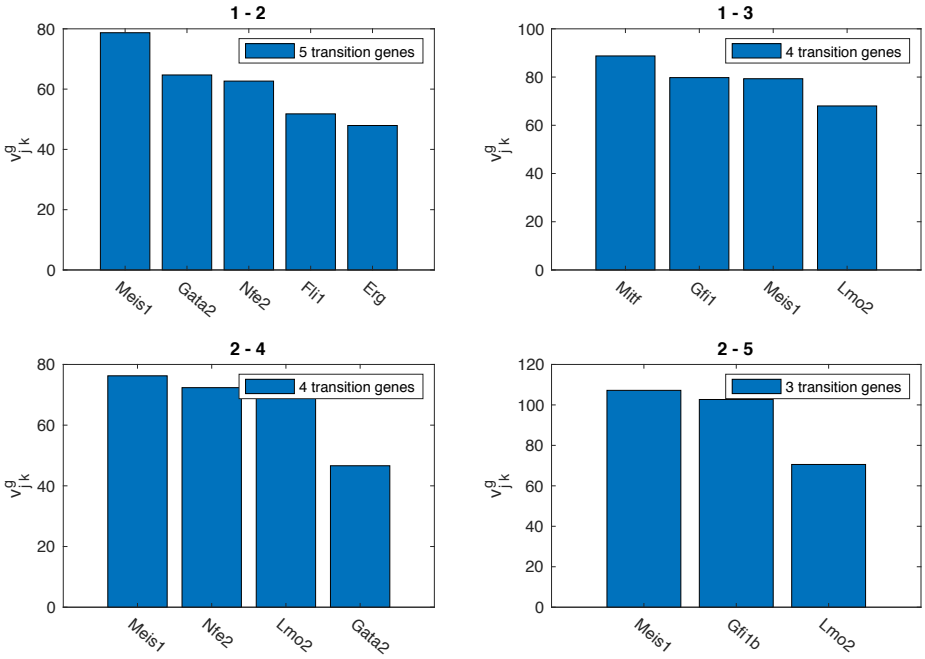
```
Press 1 if you want to remove edges, 0 otherwise:
0
```

In addition, CALISTA returns (not shown):

- Cell clustering plot based on the cluster pseudotime
- Boxplot, mean, median entropy values calculated for each cluster
- Plot of mean expression values for each gene based on cell cluster expression level

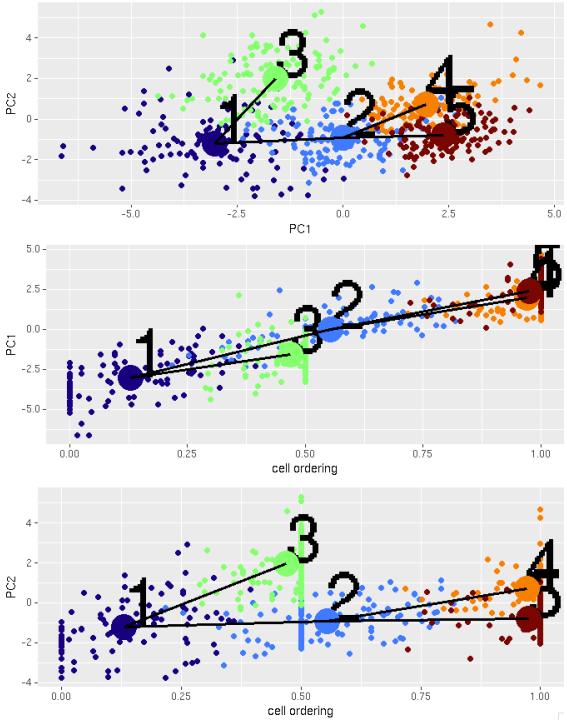
4.2.4 Determination of transition genes

After reconstructing the lineage progression, we identify the key transition genes for any two connected clusters in the graph, based on the gene-wise likelihood difference between having the cells separately as two clusters and together as a single cluster. Larger differences in the gene-wise likelihood point to more informative genes. The transition genes are selected as those whose gene-wise likelihood differences make up to more than a certain percentage of the cumulative sum of the likelihood differences of all genes – set by `INPUTS$thr_transition_genes`.



4.2.5 Pseudotemporal ordering of cells

For pseudotemporal ordering of cells, CALISTA performs maximum likelihood optimization for each cell using a linear interpolation of the cell likelihoods between any two connected clusters. The pseudotimes of the cells are computed by linear interpolation of the cluster pseudotimes, and correspond to the maximum point of the likelihood optimization above. Cells are subsequently assigned to the edges in the lineage progression graph. The following screenshot gives the results of this cell-to-edge assignment.



4.2.6 Path analysis

Finally, we perform post-analysis by entering **1** upon queried. Here, we input three developmental paths: **[1 3]**, **[1 2 5]**, and **[1 2 4]**.

```
Press 1 if you want to select transition paths and perform additional analysis, 0 otherwise:
1
Path num:  1

*********************************

Pleaseheck the figure 'cell ordering for reference '  and
Key the clusters in the path (e.g. 1 2 3 4):
1 3
Please 1 to add another path, 0 otherwise:
1
Path num:  2

*********************************

Pleaseheck the figure 'cell ordering for reference '  and
Key the clusters in the path (e.g. 1 2 3 4):
1 2 4
Please 1 to add another path, 0 otherwise:
1
Path num:  3

*********************************

Pleaseheck the figure 'cell ordering for reference '  and
Key the clusters in the path (e.g. 1 2 3 4):
1 2 5
Please 1 to add another path, 0 otherwise:
0
```

For each path, the post-analysis in CALISTA generates Clustergrams, moving-averaged gene expression profiles and co-expression networks for the transition genes detected previously based on cell orderings (not shown).

4.3 Example 3. Mouse embryonic fibroblast differentiation into neurons (Manual data import)

Analysis of RNA-seq data in Treutlein et al., Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq, *Nature* **534,** 391–395 (2016).

Please unzip the file "3-TREUTLEIN_data_type_3_format_data_5_clusters_4.txt.zip**" in **EXAMPLES/RNA-seq/** before running CALISTA**

4.3.1 Data Import and Preprocessing

Here, we edit the `MAIN.R` script in the main folder of CALISTA and set the fields of `INPUTS` as follows:

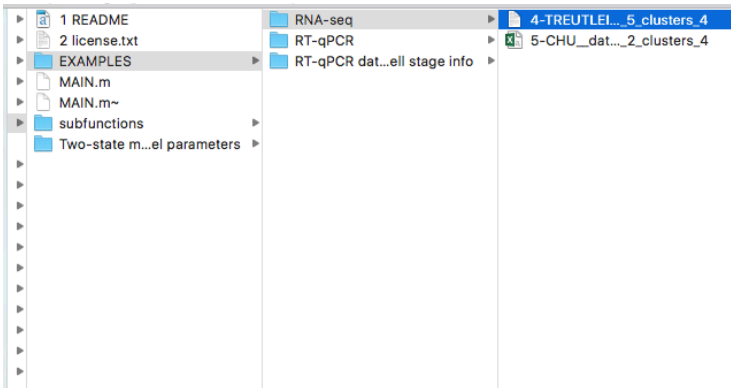
```
% Specify data types and settings for pre-processing
INPUTS$data_type=3; % Single-cell RNA-seq data
INPUTS$format_data=5; % Manual selection from data table
INPUTS$data_selection= integer(); % Include data from all time points
INPUTS$perczeros_genes=90; % Remove genes with > 90% of zeros
INPUTS$perczeros_cells=100; % Remove cells with 100% of zeros
INPUTS$cells_2_cut=0; % No manual removal of cells
INPUTS$perc_top_genes=10; % Retain only top X the most variable genes with X=min(200,
INPUTS$perc_top_genes * num of cells/100, num of genes)
% Specify single-cell clustering settings
INPUTS$optimize=1; % Set the number of clusters based on Eigengap Heuristics
INPUTS$parallel=1; % Use parallelization option
INPUTS$runs=50; % Perform 50 independent runs of greedy algorithm
INPUTS$max_iter=100; % Limit the number of iterations in greedy algorithm to 100
INPUTS$cluster='kmedoids'; % Use k-medoids in consensus clustering
% Specify transition genes settings
INPUTS$thr_transition_genes=75; % Set threshold for transition genes determination to
75%
```

Then we change the current directory in R to the CALISTA folder (use "`setwd`" command).
Subsequently, we run `MAIN.R` and import Bargaje dataset (available in the subfolder EXAMPLES/RT-qPCR).

The following are screenshots from running CALISTA on R.

```
Console ~/Documents/Calista/AAA-CALISTA temp R 1.7/
> setwd("/Users/taofang/Documents/Calista/AAA-CALISTA\ temp\ R\ 1.7")
> source("MAIN.R")
```

```
[1] "**** Please upload normalized data. File formats accepted: .txt , .xlxs , .csv ****"

Data loading...
```



The file.txt containing the original dataset can be summarized as follows (preview with the first 25 rows and 12 column):

cell_name	assignment	log_tauGFP_int	experiment	time_point	0610005C13Rik	0610007C21Rik	0610007L01Rik	0610007N19Rik	0610007P08Rik	0610007P14Rik	0610007P22Rik
1_iN1_C01	d2_induced	0	iN1	2	0	0	0	4.75176408	0	0	5.122584584
1_iN1_C02	d2_induced	0	iN1	2	0	7.37704663	0	0	0	0	0
1_iN1_C03	d2_induced	0	iN1	2	0	0	3.544937108	2.770697871	2.563243746	0	5.501432671
1_iN1_C04	d2_intermediat	0	iN1	2	0	3.926875324	0	9.082312119	0	0.975336905	0
1_iN1_C05	d2_intermediat	0	iN1	2	0	6.399809926	6.946602177	5.269901346	4.251042833	5.389056375	1.916400044
1_iN1_C07	d2_induced	0	iN1	2	0	7.995805052	2.724768024	4.74824605	0	0	0.780314952
1_iN1_C08	d2_intermediat	0	iN1	2	0	4.541419746	1.147394416	2.297672823	0	5.217150702	0
1_iN1_C09	d2_induced	0	iN1	2	0	7.936363922	1.116471648	2.52097415	0	0	0
1_iN1_C10	d2_induced	0	iN1	2	0	6.391090605	2.26164537	0	3.671925406	4.388118622	1.599668841
1_iN1_C11	d2_intermediat	0	iN1	2	0	5.120672139	4.497759092	3.694826533	0	0	0
1_iN1_C12	d2_induced	0	iN1	2	0	5.667198479	3.21066394	4.242001325	0	3.339223359	4.329961698
1_iN1_C13	d2_induced	0	iN1	2	0	2.995850255	0	1.077606307	0	4.363507545	0
1_iN1_C14	d2_induced	0	iN1	2	0	7.55398606	0.805955813	4.632348754	3.066893067	5.993690988	0
1_iN1_C15	d2_induced	0	iN1	2	0	0	0.338310627	0	0.03966647	0	0
1_iN1_C16	d2_intermediat	0	iN1	2	0	5.693022646	7.262878865	2.17209037	0.36848359	5.139190271	0
1_iN1_C17	d2_induced	0	iN1	2	0	7.777611627	5.066452204	0.204849805	0	0	0
1_iN1_C19	d2_intermediat	0	iN1	2	0	8.245371124	5.54280365	5.541281792	0	0	2.500978591
1_iN1_C20	d2_induced	0	iN1	2	0	1.229217679	0	0	0	0	0
1_iN1_C21	d2_intermediat	0	iN1	2	0	7.701893565	4.541058261	2.975951694	0	3.092083605	1.789176911
1_iN1_C22	d2_induced	0	iN1	2	0	6.206550133	0	7.272218899	0	0	0
1_iN1_C23	d2_induced	0	iN1	2	0	5.254699661	0	3.644292022	5.359889928	0.262936771	0
1_iN1_C25	MEF	0	iN1	2	0	0	1.888426352	0	0	2.673734565	2.207363104
1_iN1_C26	d2_induced	0	iN1	2	0	4.711219374	0	1.757716009	0	0	0
1_iN1_C27	MEF	0	iN1	2	0	6.205006575	0	4.487598366	5.959469132	5.763890139	3.095969378

CALISTA provides a preview and the dimensions of the IMPORTED DATA.

```
IMPORTED DATA preview::


  cell_name      assignment log_tauGFP_intensity experiment time_point 0610005C13Rik
1 1_iN1_C01      d2_induced                    0        iN1          2             0
2 1_iN1_C02      d2_induced                    0        iN1          2             0
3 1_iN1_C03      d2_induced                    0        iN1          2             0
4 1_iN1_C04 d2_intermediate                    0        iN1          2             0
5 1_iN1_C05 d2_intermediate                    0        iN1          2             0
  0610007C21Rik 0610007L01Rik 0610007N19Rik 0610007P08Rik
1      0.000000      0.000000      4.751764      0.000000
2      7.377047      0.000000      0.000000      0.000000
3      0.000000      3.544937      2.770698      2.563244
4      3.926875      0.000000      9.082312      0.000000
5      6.399810      6.946602      5.269901      4.251043
Dimension of IMPORTED DATA::
 Rows= 405
 Columns= 22529
```

We press **1** since columns refer genes and rows refer cells.

```
Press 1 if columns=genes, 0 otherwise
1
```

Then select starting column of expression data by excluding the capture time info.

```
Key the starting coloumn for expression data
6
```

CALISTA also provides a preview and the dimensions of the EXPRESSION DATA.

```
Selected EXPRESSION DATA preview::


  0610005C13Rik 0610007C21Rik 0610007L01Rik 0610007N19Rik 0610007P08Rik 0610007P14Rik
1             0      0.000000      0.000000      4.751764      0.000000     0.0000000
2             0      7.377047      0.000000      0.000000      0.000000     0.0000000
3             0      0.000000      3.544937      2.770698      2.563244     0.0000000
4             0      3.926875      0.000000      9.082312      0.000000     0.9753369
5             0      6.399810      6.946602      5.269901      4.251043     5.3890564
  0610007P22Rik 0610008F07Rik 0610009B14Rik 0610009B22Rik
1      5.122585             0             0      8.077869
2      0.000000             0             0      6.489290
3      5.501433             0             0      0.000000
4      0.000000             0             0      0.000000
5      1.916400             0             0      2.948811
Dimension of selected EXPRESSION DATA::
 Rows= 405
 Columns= 22524
```

Based on the EXPRESSION DATA preview, we set the starting and ending rows and columns for the expression values: as **[1 405]** and **[2 22524]**, respectively, when queried.

```
* Key starting and ending rows for the expression values (e.g. 1 405):*
1 405
* Key starting and ending columns for the expression values (e.g. 1 22524):*
1 22524
```

We define the gene's names using the IMPORTED DATA preview **[6 22529]** (starting and ending columns).

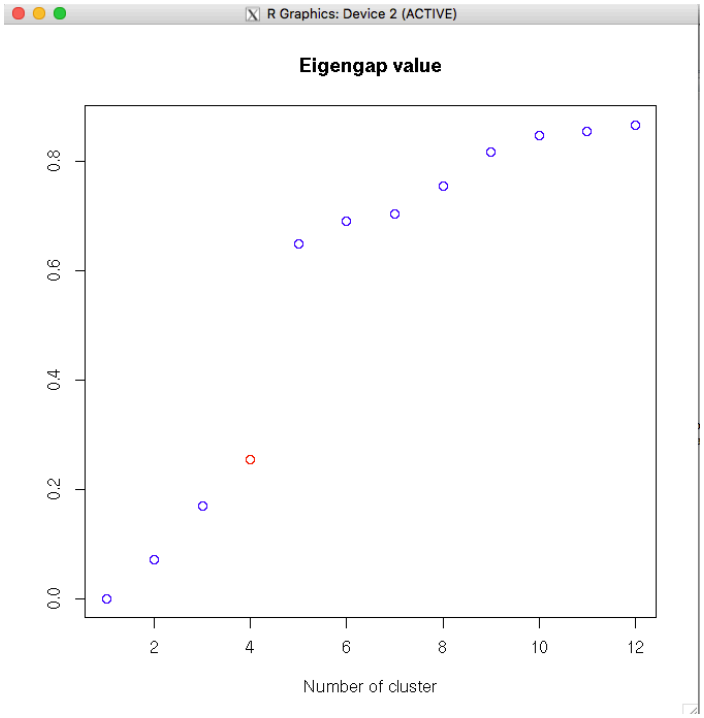
```
*Key the starting and ending columns for gene names(e.g. 6 22529)*
6 22529
```

We load the capture time/cell stage by pressing **1** (i.e. time/cell stage information is in IMPORTED DATA) and selecting column **5** in the data matrix.

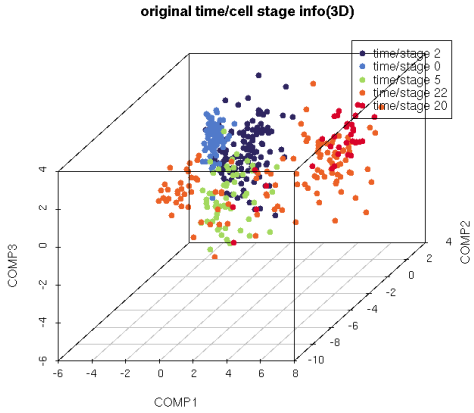
```
*Add time info(1-Yes,0-No):*
1
*Key the column or row(e.g 1)vector in the raw data set with time/cell stage info:*
5
```

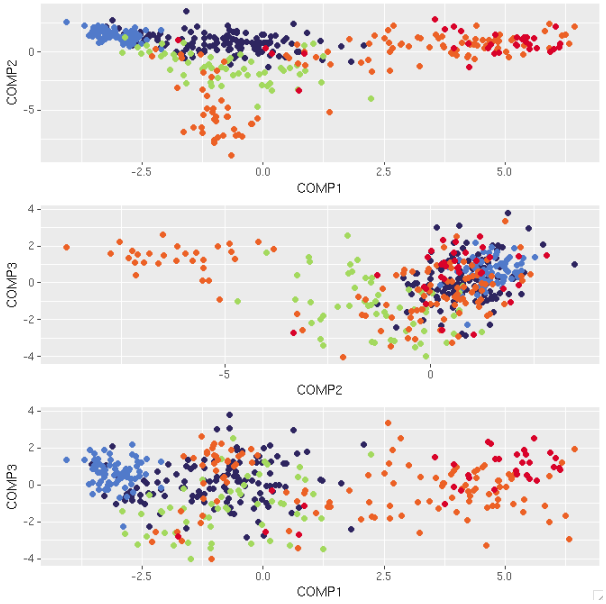
4.3.2 Single-cell clustering

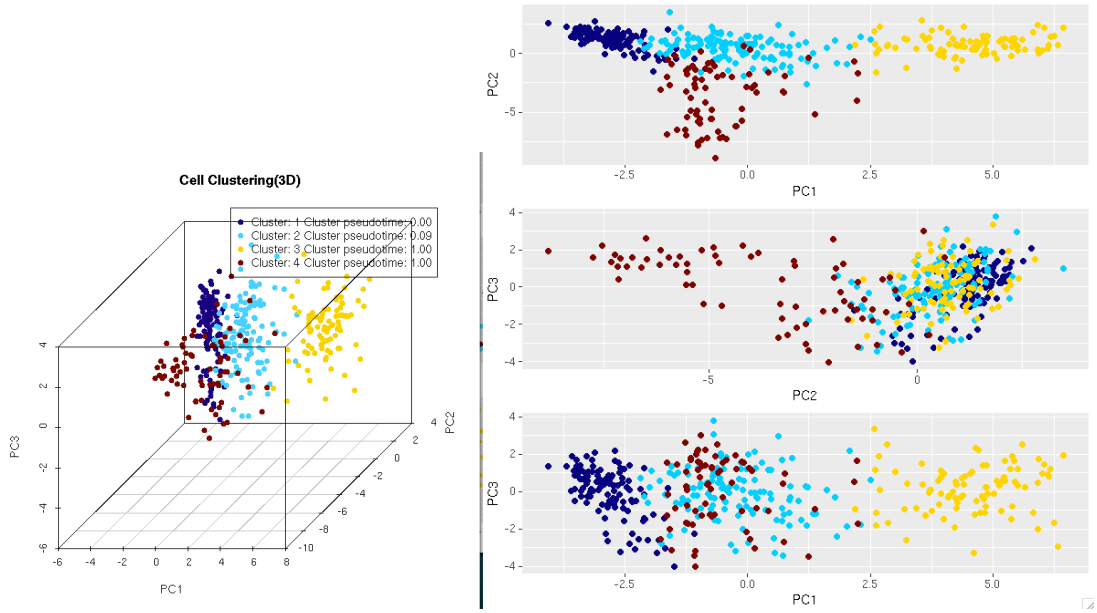
In this case, the number of clusters is determined using the eigengap plot. According to the eigengap plot below, we set the number of clusters to **4**. The following are screenshots from CALISTA single-cell clustering analysis.




```
Number of clusters expected:
4
```







We do not need to remove any cluster (by entering **0** upon queried), and continue with further analysis (by entering **1** upon queried):

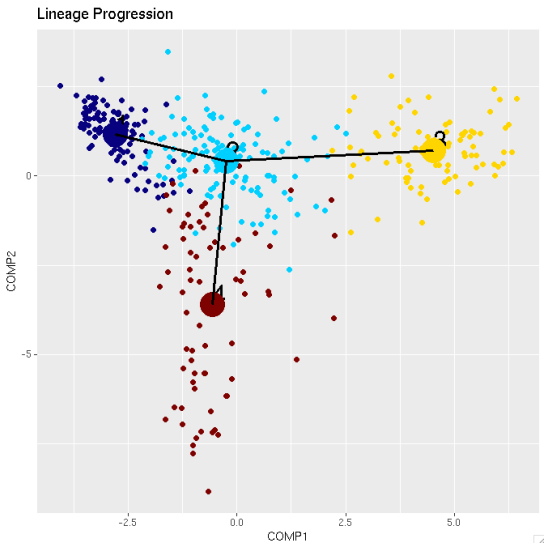
```
Plotting...

Press 1 if you want to remove one cell cluster, 0 otherwise:
0
 Press 1 if you want to perform additional analysis (e.g. lineage inference, cell o
rdering) , 0 otherwise:
1
```

4.3.3 Reconstruction of lineage progression

During the lineage inference step, CALISTA automatically generates and displays a lineage graph, obtained by adding an edge between two clusters in increasing cluster distances, until all clusters are connected to at least one other cluster. Subsequently, users can manually add or remove one edge at time based on the cluster distances.

```
CALISTA_transtion is running...


4 edge(s) have been added and the graph is connected.
If you want to add another edge press "1" (then Enter)
If you want to remove an edge press "2" (then Enter)
If you want to continue with the next step press "0" (then Enter)

0
```

ATTENTION: to add an edge (press "**1**"), remove an edge (press "**2**") or finalize the lineage progression graph (press "**0**"), the MATLAB figure of the graph must appear in foreground without any modification (e.g., zooming, rotation). Note that the addition/removal of the edges are performed according to increasing/decreasing order of cluster distance.



ATTENTION: the final graph must be connected (i.e. there is a path from any node/cluster to any other node/cluster in the graph) otherwise a warning will be returned.

Here, we do not need to remove any spurious edges, and hence we enter **0** upon queried.

```
Press 1 if you want to remove edges, 0 otherwise:
0
```

In addition, CALISTA returns (not shown):

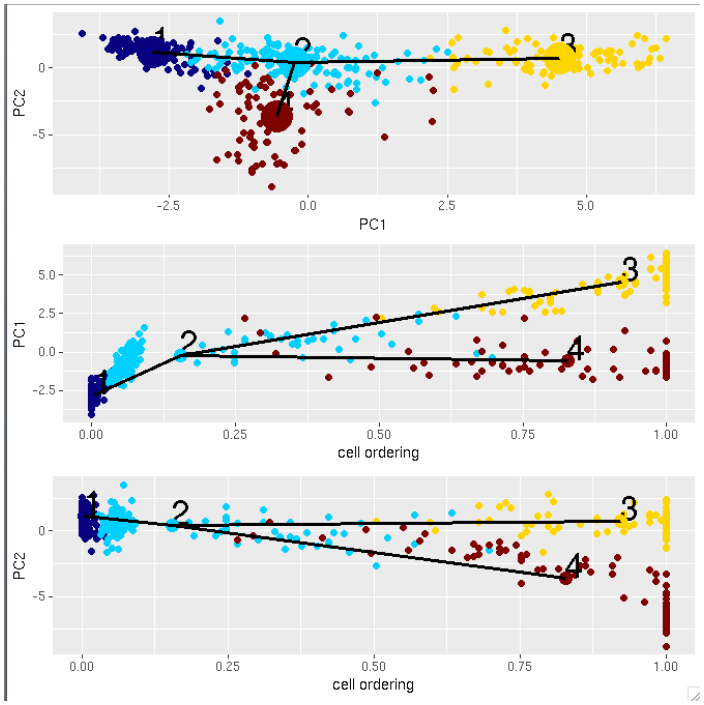
- Cell clustering plot based on the cluster pseudotime
- Boxplot, mean, median entropy values calculated for each cluster
- Plot of mean expression values for each gene based on cell cluster expression level

4.3.4 Determination of transition genes

After reconstructing the lineage progression, we identify the key transition genes for any two connected clusters in the graph (results not shown here), based on the gene-wise likelihood difference between having the cells separately as two clusters and together as a single cluster. Larger differences in the gene-wise likelihood point to more informative genes. The transition genes are selected as those whose gene-wise likelihood differences make up to more than a certain percentage of the cumulative sum of the likelihood differences of all genes – set by `INPUTS$thr_transition_genes`.

4.3.5 Pseudotemporal ordering of cells

For pseudotemporal ordering of cells, CALISTA performs maximum likelihood optimization for each cell using a linear interpolation of the cell likelihoods between any two connected clusters. The pseudotimes of the cells are computed by linear interpolation of the cluster pseudotimes, and correspond to the maximum point of the likelihood optimization above. Cells are subsequently assigned to the edges in the lineage progression graph. The following screenshot gives the results of this cell-to-edge assignment.



4.4 Example 4. Human embryonic stem cell differentiation into endodermal cells

Analysis of RNA-seq data in Chu et al., Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm, *Genome Biol.* **17,** 173 (2016).

Please unzip the file "4-CHU__data_type_4_format_data_2_clusters_4.csv.zip**" in **EXAMPLES/RNA-seq/** before running CALISTA**

4.4.1 Data Import and Preprocessing

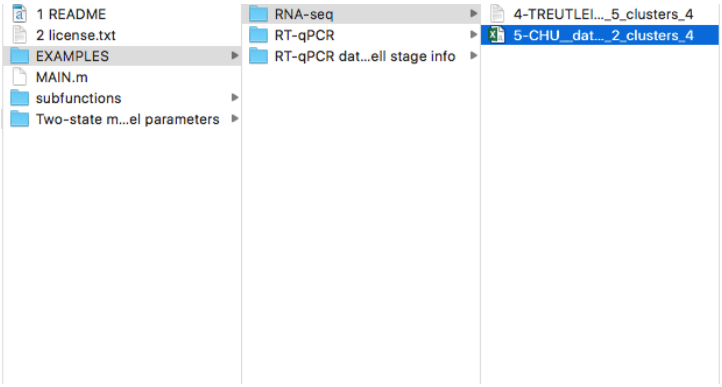
We first edit the `MAIN.R` script in the main folder of CALISTA and set the fields of `INPUTS` as follows:

```
% Specify data types and settings for pre-processing
INPUTS$data_type=4; % Single-cell RNA-seq data
INPUTS$format_data=2; % Rows= genes and Columns= cells with time/stage info in the
first row
INPUTS$data_selection= integer();; % Include data from all time points
INPUTS$perczeros_genes=90; % Remove genes with > 90% of zeros
INPUTS$perczeros_cells=100; % Remove cells with 100% of zeros
INPUTS$cells_2_cut=0; % No manual removal of cells
INPUTS$perc_top_genes=10; % Retain only top X the most variable genes with X=min(200,
INPUTS$perc_top_genes * num of cells/100, num of genes)
% Specify single-cell clustering settings
INPUTS$optimize=1; % Set the number of clusters based on Eigengap Heuristics
INPUTS$parallel=1; % Use parallelization option
INPUTS$runs=50; % Perform 50 independent runs of greedy algorithm
INPUTS$max_iter=100; % Limit the number of iterations in greedy algorithm to 100
INPUTS$cluster='kmedoids'; % Use k-medoids in consensus clustering
% Specify transition genes settings
INPUTS$thr_transition_genes=75; % Set threshold for transition genes determination to
75%
```

Then we change the current directory in R to the CALISTA folder (use "`setwd`" command).
Subsequently, we run `MAIN.R` and import Bargaje dataset (available in the subfolder EXAMPLES/RT-qPCR).

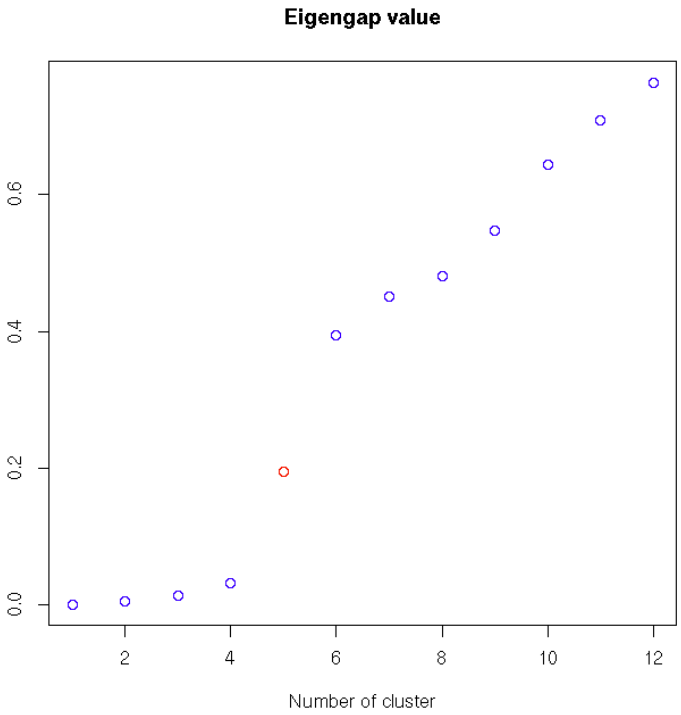
The following are screenshots from running CALISTA on R.

```
Console ~/Documents/Calista/AAA-CALISTA temp R 1.7/
> setwd("/Users/taofang/Documents/Calista/AAA-CALISTA\ temp\ R\ 1.7")
> source("MAIN.R")
[1] "**** Please upload normalized data. File formats accepted: .txt , .xlxs , .csv ****"

Data loading...
```



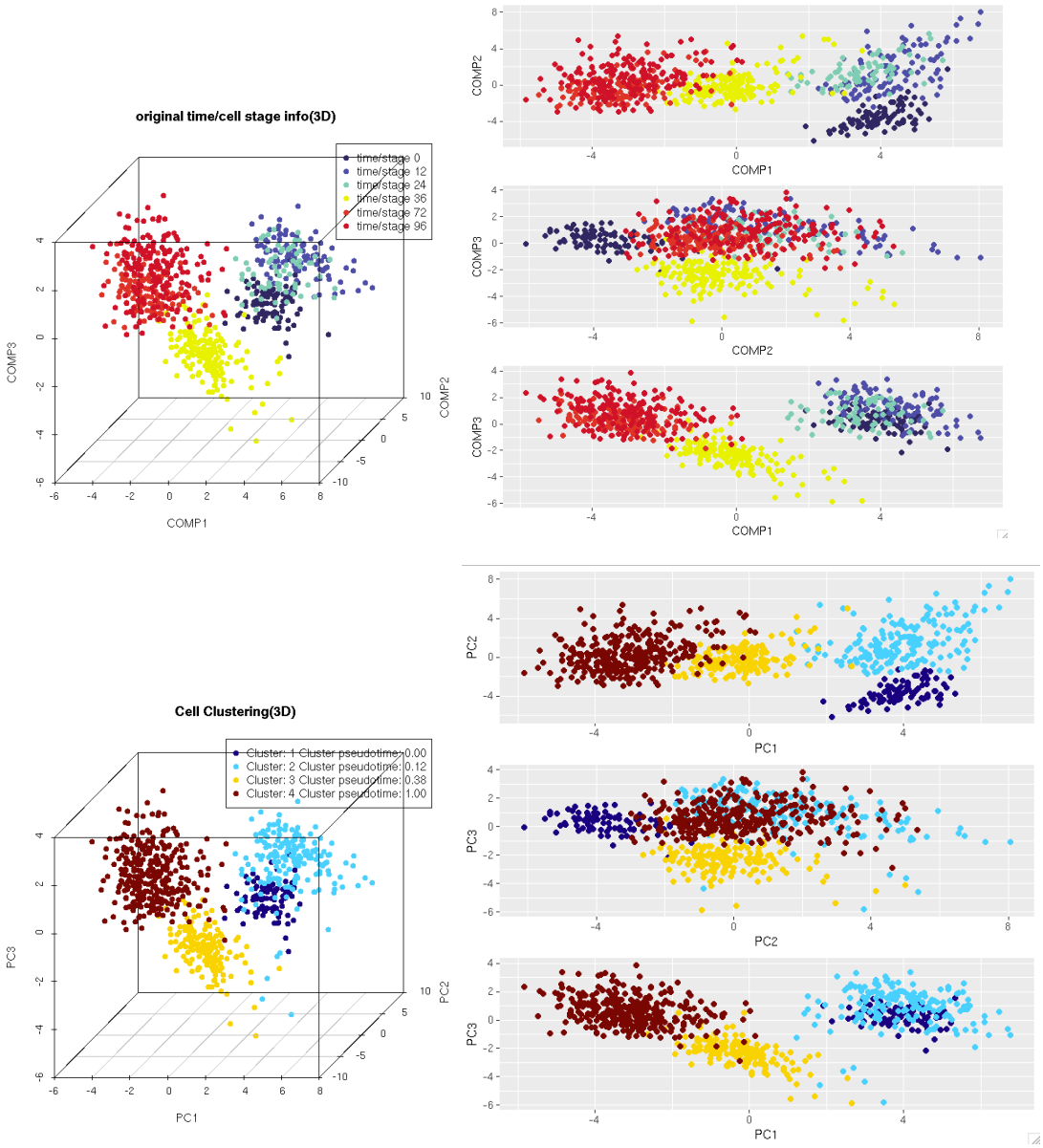
4.4.2 Single-cell clustering

In this case, the number of clusters is determined using the eigengap plot. According to the eigengap plot below, we set the number of clusters to **4**. The following are screenshots from CALISTA single-cell clustering analysis.



NOTE: CALISTA automatically returns the optimal number of clusters based on the MAXIMUM eigengap value. However, the user might choose the number of clusters to adopt based on the FIRST eigengap.

```
Number of clusters expected:
4
```



We do not need to remove any clusters (by entering **0** upon queried), and continue with further analysis (by entering **1** upon queried).

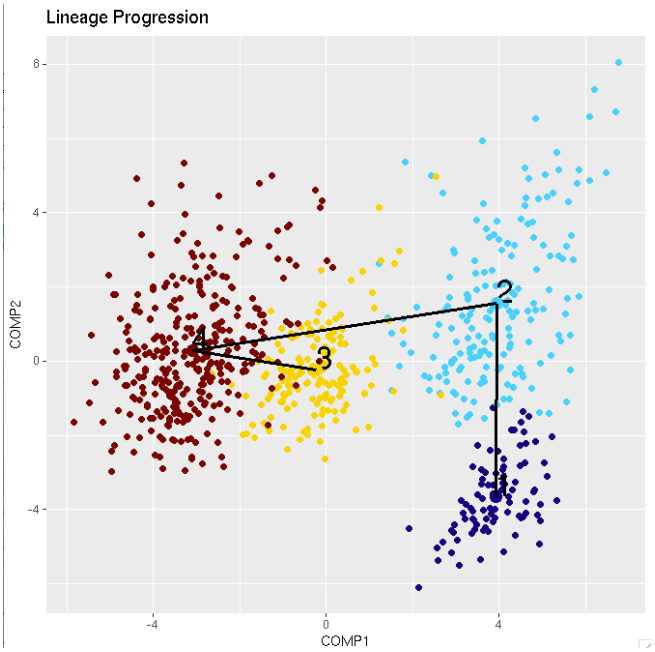
```
Press 1 if you want to remove one cell cluster, 0 otherwise:
0
 Press 1 if you want to perform additional analysis (e.g. lineage inference, cell
ordering) , 0 otherwise:
1
```

4.4.3 Reconstruction of lineage progression

During the lineage inference step, CALISTA automatically generates and displays a lineage graph, obtained by adding an edge between two clusters in increasing cluster distances, until all clusters are connected to at least one other cluster. Subsequently, users can manually add or remove one edge at time based on the cluster distances.

```
CALISTA_transtion is running...


3 edge(s) have been added and the graph is connected.
If you want to add another edge press "1" (then Enter)
If you want to remove an edge press "2" (then Enter)
If you want to continue with the next step press "0" (then Enter)
```

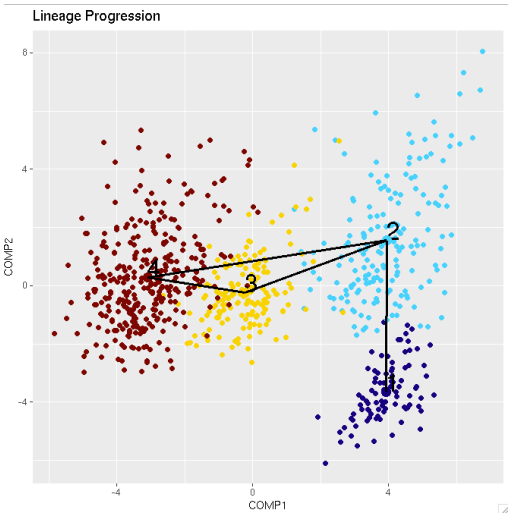
ATTENTION: to add an edge (press "**1**"), remove an edge (press "**2**") or finalize the lineage progression graph (press "**0**"), the MATLAB figure of the graph must appear in foreground without any modification (e.g., zooming, rotation). Note that the addition/removal of the edges are performed according to increasing/decreasing order of cluster distance.



ATTENTION: the final graph must be connected (i.e. there is a path from any node/cluster to any other node/cluster in the graph), otherwise a warning will be returned.

Since the transition from cluster 4 to cluster 3 is inconsistent with the capture time info (i.e. cluster pseudotime values for cluster 4 and 3 are 1 and 0.38 respectively), we add one further edge to produce the following lineage progression graph by pressing "**1"** and then "**0**":

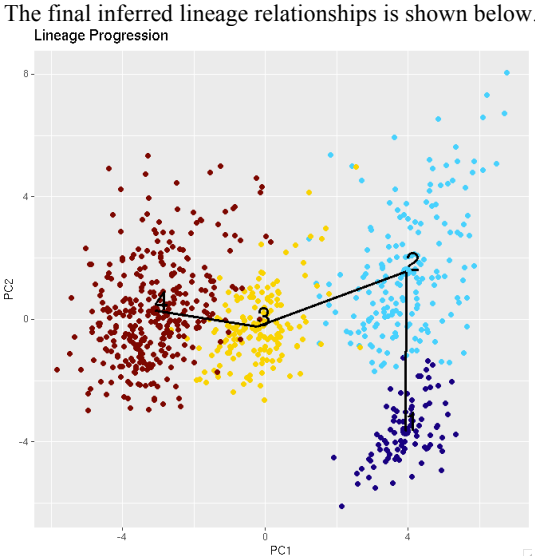
```
CALISTA_transtion is running...


3 edge(s) have been added and the graph is connected.
If you want to add another edge press "1" (then Enter)
If you want to remove an edge press "2" (then Enter)
If you want to continue with the next step press "0" (then Enter)

1
4 edge(s) have been added and the graph is connected.
If you want to add another edge press "1" (then Enter)
If you want to remove an edge press "2" (then Enter)
If you want to continue with the next step press "0" (then Enter)

0
```



Based on our definition of branching point, we consider the previous inferred lineage graph still linear, since there is only one final cell cluster (cluster 4). Moreover, since the transition from cluster 2 to cluster 4 is inconsistent with the capture time info (i.e. cluster pseudotime values for cluster 2 and 4 are 0.12 and 1 respectively), we remove the spurious edge between cluster 2 and 4, by entering **1** and entering **[2 4]**, upon the following query:

```
Press 1 if you want to remove edges, 0 otherwise:
1

 ********************************************************

Specify the node pairs (e.g. 4 5):
2 4
 Press 1 to remove another edge, 0 otherwise:
0
1 edges to remove
```

The final inferred lineage relationships is shown below.



In addition, CALISTA returns (not shown):

- Cell clustering plot based on the cluster pseudotime
- Boxplot, mean, median entropy values calculated for each cluster
- Plot of mean expression values for each gene based on cell cluster expression level

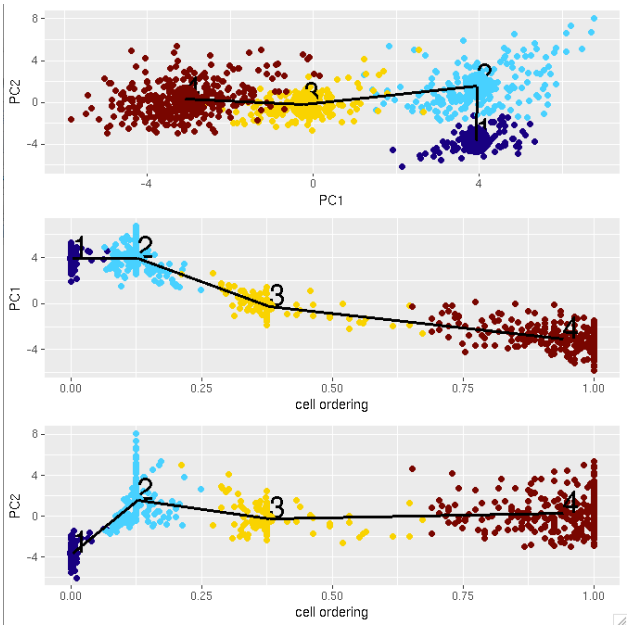
4.4.4 Determination of transition genes

After reconstructing the lineage progression, we identify the key transition genes for any two connected clusters in the graph (results not shown here), based on the gene-wise likelihood difference between having the cells separately as two clusters and together as a single cluster. Larger differences in the gene-wise likelihood point to more informative genes. The transition genes are selected as those whose gene-wise likelihood differences make

up to more than a certain percentage of the cumulative sum of the likelihood differences of all genes – set by `INPUTS$thr_transition_genes`.

4.4.5 Pseudotemporal ordering of cells

For pseudotemporal ordering of cells, CALISTA performs maximum likelihood optimization for each cell using a linear interpolation of the cell likelihoods between any two connected clusters. The pseudotimes of the cells are computed by linear interpolation of the cluster pseudotimes, and correspond to the maximum point of the likelihood optimization above. Cells are subsequently assigned to the edges in the lineage progression graph. The following screenshot gives the results of this cell-to-edge assignment.



4.5 Example 5. Running CALISTA without time or cell stage information

*Analysis of RT-qPCR data in **Moignard et al.** "Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis". Nat. Cell Biol. 15, 363–72 (2013).*

Here, we report only the main steps of the analysis. For the complete analysis please check **Example 2**.

4.5.1 Data Import and Preprocessing

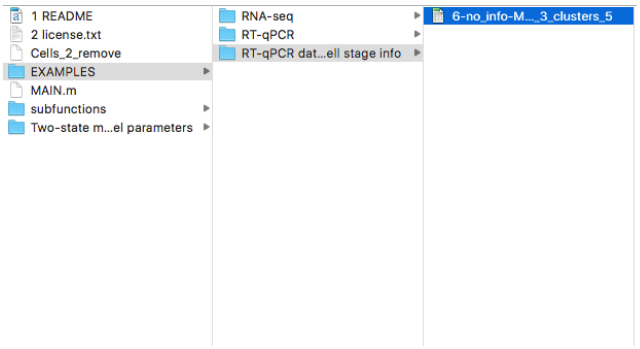
We edit the `MAIN.R` script in the main folder of CALISTA and set the fields of `INPUTS` as follows:

```
% Specify data types and settings for pre-processing
INPUTS$data_type=1; % Single-cell RT-qPCR CT data
INPUTS$format_data=3; % Rows= cells and Columns= genes (no time/stage info)
INPUTS$data_selection= integer();; % Include data from all time points
INPUTS$perczeros_genes=100; % Remove genes with > 100% of zeros
INPUTS$perczeros_cells=100; % Remove cells with 100% of zeros
INPUTS$cells_2_cut=0; % No manual removal of cells
INPUTS$perc_top_genes=100; % Retain only top X the most variable genes with X=min(200,
INPUTS$perc_top_genes * num of cells/100, num of genes)
% Specify single-cell clustering settings
INPUTS$optimize=0; % The number of cluster is known a priori
INPUTS$parallel=1; % Use parallelization option
INPUTS$runs=50; % Perform 50 independent runs of greedy algorithm
INPUTS$max_iter=100; % Limit the number of iterations in greedy algorithm to 100
INPUTS$cluster='kmedoids'; % Use k-medoids in consensus clustering
% Specify transition genes settings
INPUTS$thr_transition_genes=50; % Set threshold for transition genes determination to
50%
```

Then we change the current directory in R to the CALISTA folder (use "`setwd`" command).
Subsequently, we run `MAIN.R` and import Bargaje dataset (available in the subfolder EXAMPLES/RT-qPCR).

The following are screenshots from running CALISTA on R.

```
Console ~/Documents/Calista/AAA-CALISTA temp R 1.7/
> setwd("/Users/taofang/Documents/Calista/AAA-CALISTA\ temp\ R\ 1.7")
> source("MAIN.R")
```



4.5.2 Single-cell clustering

Following the original publication, we set the number of clusters equals to **5**.

```
Data loading...

Number of clusters expected:
5
```

4.5.3 Reconstruction of lineage progression and pseudotemporal ordering of cells

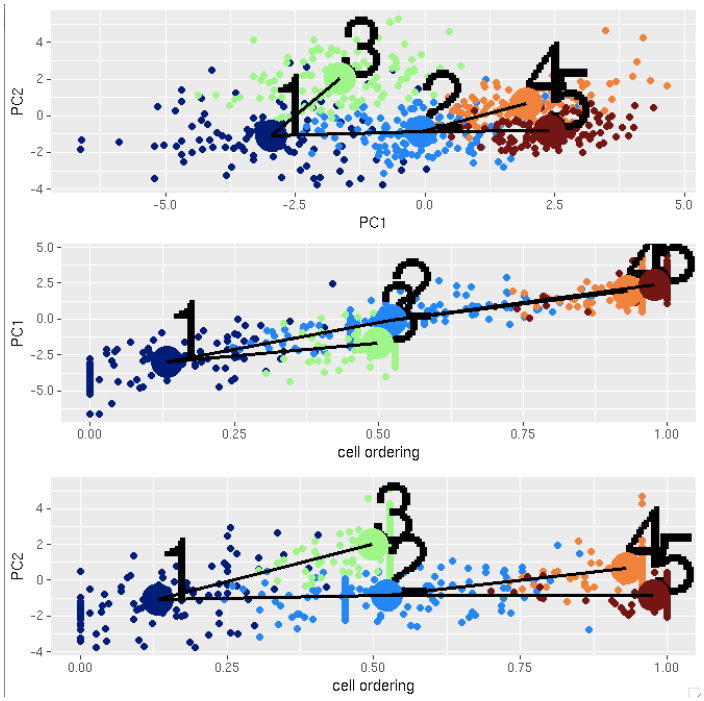
We follow the steps as outlined in the other examples above to infer the lineage progression and carry out pseudotemporal ordering of single cells.
Without the time or cell stage info, CALISTA is still able to recover the cluster progression based on:

a. *The specification of the starting cell* (e.g. cell **1**):

```
No time info found. Please enter the starting cell or the marker gene whenever available

Press 1 to enter the starting cell, 2 to enter the marker gene, 3 otherwise:
1

 Key the index of the starting cell(e.g. 1):
1
```

The final inferred lineage relationships are as follow.

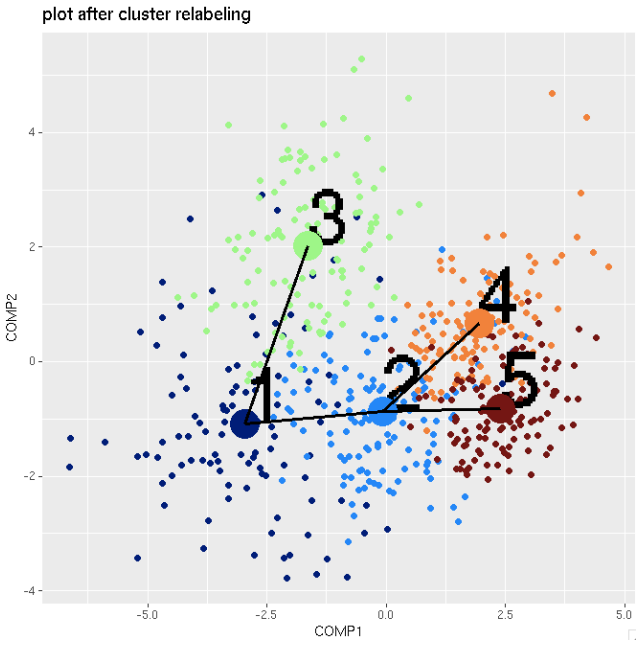
CALISTA pseudotemporal ordering gives the following outcome.



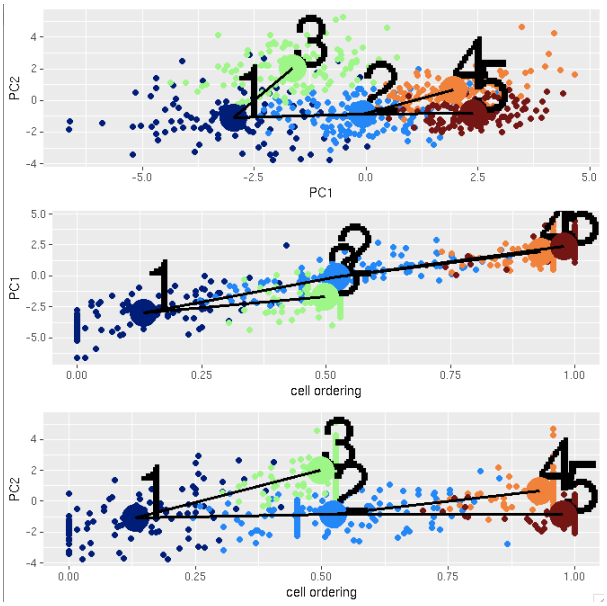
b. *The specification of a marker gene* (e.g. **'Erg'**) which is downregulated (press **2**):

```
Press 1 to enter the starting cell, 2 to enter the marker gene, 3 otherwise:
2

 Enter the marker gene chosen(e.g. Erg, case insensitive):
erg

press 1 if it is upregulated(minEXp in cluster 1),2 if it is downregulated(maxExp in
cluster 1)(e.g.2):
2
```

The final inferred lineage relationships:



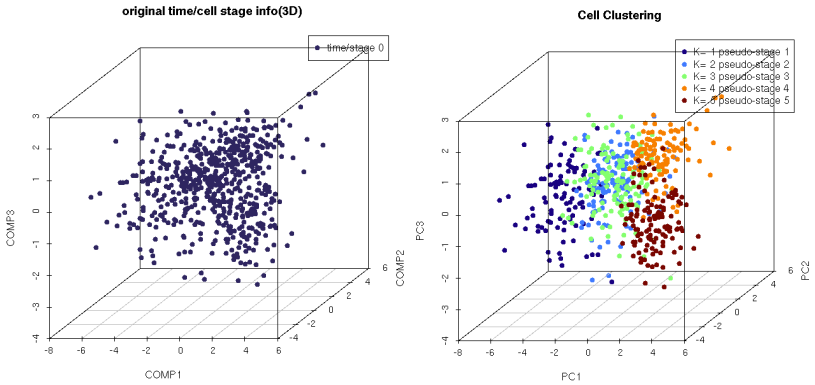
CALISTA pseudotemporal ordering of cells gives the following result.



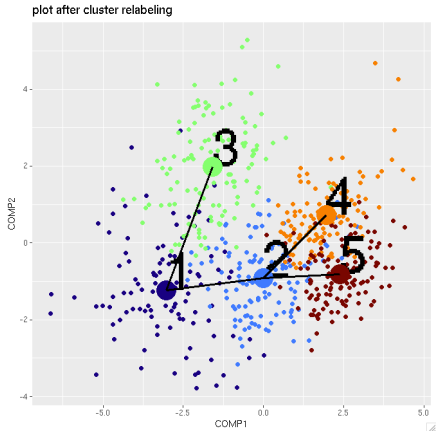
Without any information of the time information, cell stage, starting cell and marker genes, CALISTA is still able to find the topology of the lineage graph, but the edges are undirected.

```
No time info found. Please enter the starting cell or the marker gene whenever avaib
le

Press 1 to enter the starting cell, 2 to enter the marker gene, 3 otherwise:
3

Skip relabeling step
```

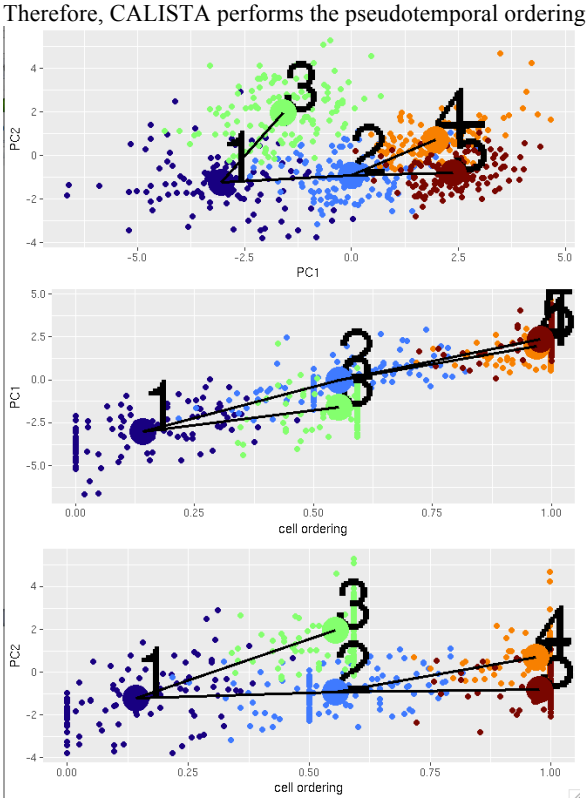
CALISTA single-cell clustering result is as follows.



The final inferred lineage relationships are shown below.



Therefore, CALISTA performs the pseudotemporal ordering of cells as follows:



4.6 Example 6. Removing undesired clusters

Analysis of RT-qPCR data in Moignard et al., Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis, *Nat. Cell Biol.* **15**, 363–72 (2013).

4.6.1 Data Import and Preprocessing

We edit the `MAIN.R` script in the main folder of CALISTA and set the fields of `INPUTS` as follows:

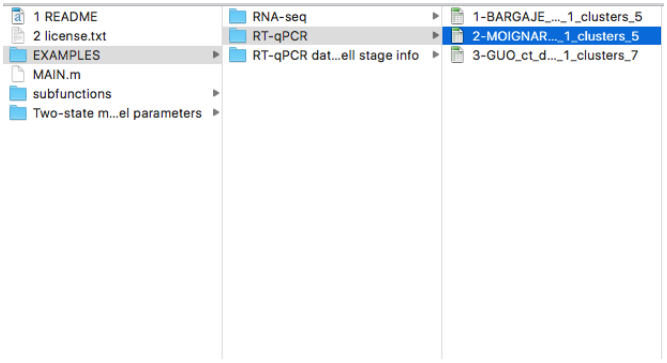
```
% Specify data types and settings for pre-processing
INPUTS$data_type=1; % Single-cell RT-qPCR CT data
INPUTS$format_data=1; % Rows= cells and Columns= genes with time/stage info in the
last column
INPUTS$data_selection= integer();; % Include data from all time points
INPUTS$perczeros_genes=100; % Remove genes with > 100% of zeros
INPUTS$perczeros_cells=100; % Remove cells with 100% of zeros
INPUTS$cells_2_cut=0; % No manual removal of cells
INPUTS$perc_top_genes=100; % Retain only top X the most variable genes with X=min(200,
INPUTS$perc_top_genes * num of cells/100, num of genes)
% Specify single-cell clustering settings
INPUTS$optimize=0; % The number of cluster is known a priori
INPUTS$parallel=1; % Use parallelization option
INPUTS$runs=50; % Perform 50 independent runs of greedy algorithm
INPUTS$max_iter=100; % Limit the number of iterations in greedy algorithm to 100
INPUTS$cluster='kmedoids'; % Use k-medoids in consensus clustering
% Specify transition genes settings
INPUTS$thr_transition_genes=50; % Set threshold for transition genes determination to
50%
```

Then we change the current directory in R to the CALISTA folder (use "`setwd`" command).
Subsequently, we run `MAIN.R` and import Bargaje dataset (available in the subfolder EXAMPLES/RT-qPCR).

The following are screenshots from running CALISTA on R.

```
Console ~/Documents/Calista/AAA-CALISTA temp R 1.7/
> setwd("/Users/taofang/Documents/Calista/AAA-CALISTA\ temp\ R\ 1.7")
> source("MAIN.R")
```

```
[1] "**** Please upload normalized data. File formats accepted: .txt , .xlxs , .csv ****"

Data loading...
```

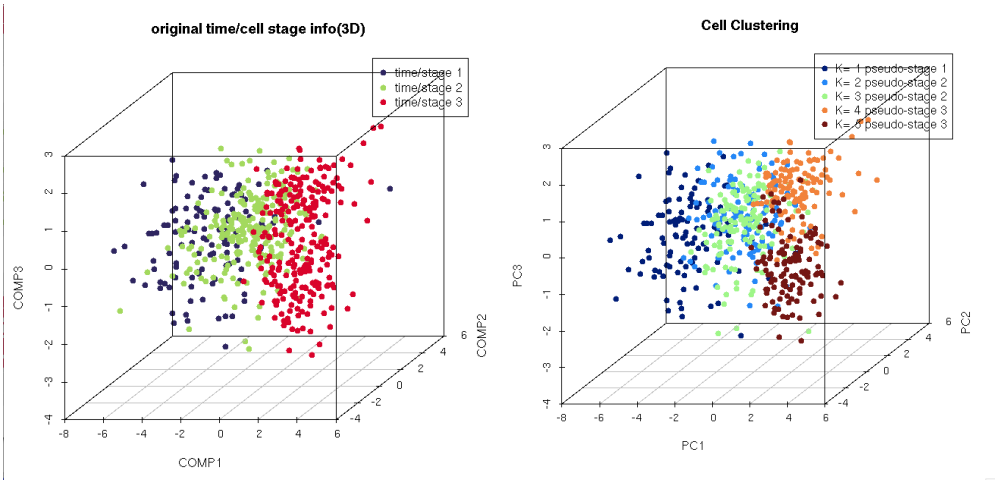


4.6.2 Single-cell clustering

We set the number of clusters equals to **5** following the original publication.

```
Data loading...

Number of clusters expected:
5
```

CALISTA single-cell clustering result is shown below.



Let us proceed with removing cluster 3 and 5, by entering **1** and type **[5 3]** upon queried.

```
Press 1 if you want to remove one cell cluster, 0 otherwise:
1
key cluster number(e.g 1 or 5 3)
5 3
cell's indices to remove are saved in 'cells 2 remove.csv',

                please rerun MAIN script again
```

The indices of cells to remove are saved in a csv file.

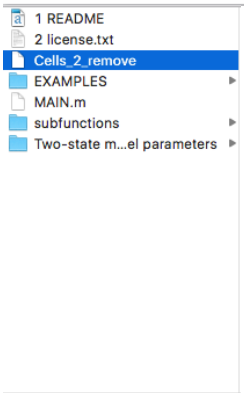
```
cell's indices to remove are saved in 'cells 2 remove.csv',

            please rerun MAIN script again
```

We then edit `MAIN.m` script again, but we set the `INPUTS` as described previously except:

```
INPUTS$cells_2_cut=0; % Manual removal of cells
```

We run `MAIN.R` once more from the workspace and import Moignard dataset (in subfolder EXAMPLES/RT-qPCR). We also upload the csv file containing cell's indices to remove.

```
> source("MAIN.R")
```

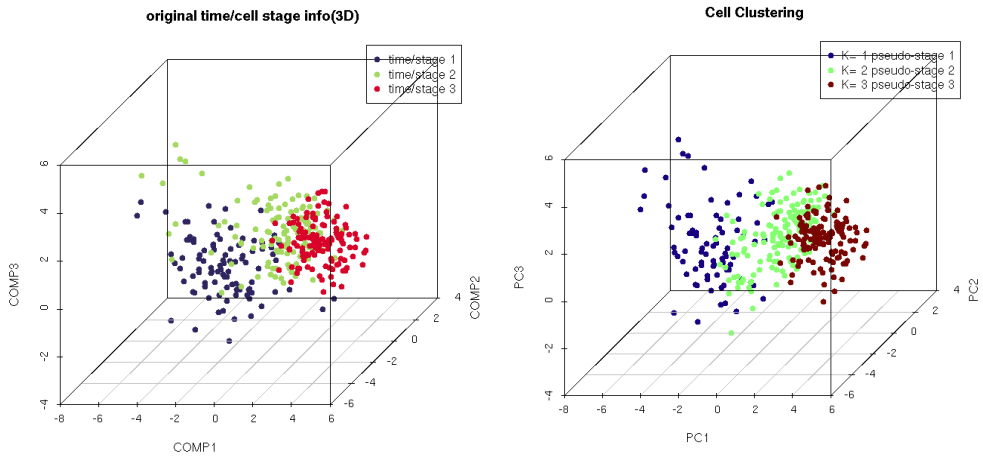


4.6.3 Single-cell clustering after removing undesired clusters

We now set the number of clusters equals to **3**.

```
Optimal number of cluster according to max. eigenvalue: 3
if you want to use this value press 0, else provide desired number of cluster:
3
```

We obtain the following clustering results:



4.6.4 Reconstruction of lineage progression

We continue with lineage inference step. During the lineage inference step, CALISTA provides the minimal connected graph (with nodes= cell clusters and edges= state transitions) as starting prediction for the developmental hierarchy. In addition, the user can also manually add or remove one edge at time based on the cluster distance values:

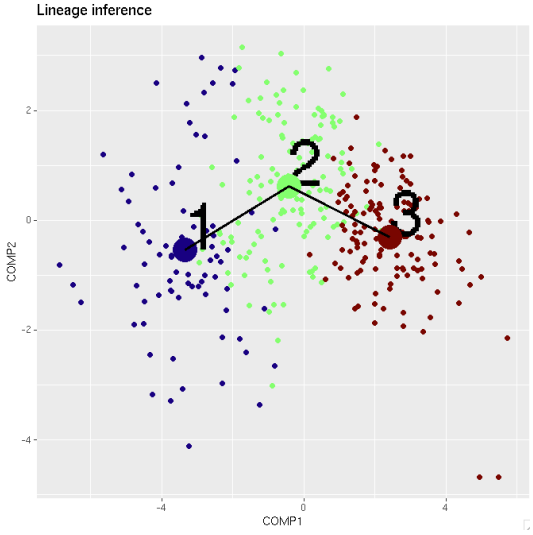
ATTENTION: to add an edge (press "**1**"), remove an edge (press "**2**") or finalize the lineage progression graph (press "**0**"), the MATLAB figure of the graph must appear in foreground without any modification (e.g., zooming, rotation). Note that the addition/removal of the edges are performed according to increasing/decreasing order of cluster distance.

ATTENTION: the final graph must be connected (i.e. there exists a path from any node/cluster to any other node/cluster in the graph), otherwise a warning will be returned.

We do not need to remove spurious edges (entering **0** upon queried)

```
Press 1 if you want to remove edges, 0 otherwise:
0
```

The final inferred lineage relationship is shown below



In addition, CALISTA returns (not shown):

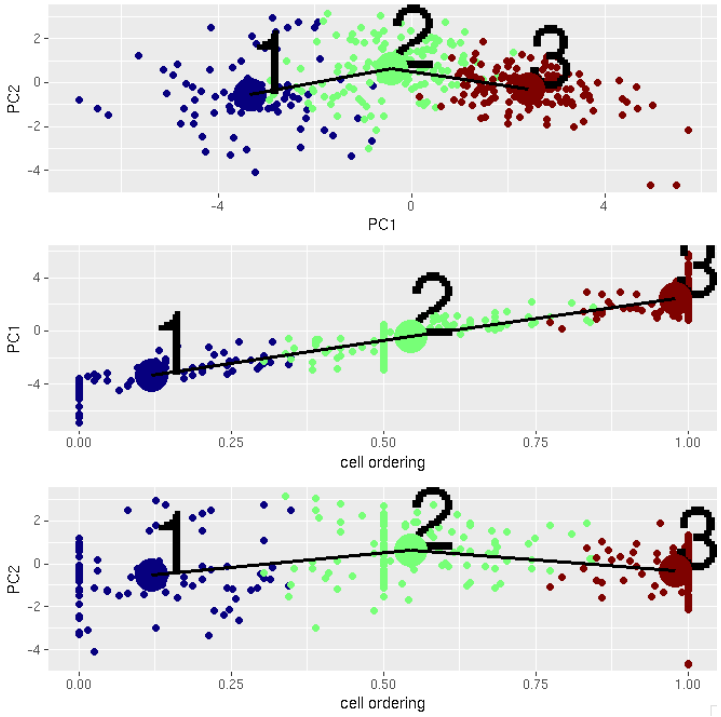
- Cell clustering plot based on the cluster pseudotime
- Boxplot, mean, median entropy values calculated for each cluster
- Plot of mean expression values for each gene based on cell cluster expression level

4.6.5 Determination of transition genes

After reconstructing the lineage progression, we identify the key transition genes for any two connected clusters in the graph (results not shown here), based on the gene-wise likelihood difference between having the cells separately as two clusters and together as a single cluster. Larger differences in the gene-wise likelihood point to more informative genes. The transition genes are selected as those whose gene-wise likelihood differences make up to more than a certain percentage of the cumulative sum of the likelihood differences of all genes – set by `INPUTS$thr_transition_genes`.

4.6.6 Pseudotemporal ordering of cells

For pseudotemporal ordering of cells, CALISTA performs maximum likelihood optimization for each cell using a linear interpolation of the cell likelihoods between any two connected clusters. The pseudotimes of the cells are computed by linear interpolation of the cluster pseudotimes, and correspond to the maximum point of the likelihood optimization above. Cells are subsequently assigned to the edges in the lineage progression graph. The following screenshot gives the results of this cell-to-edge assignment.



5 Questions and comments

Please address any problem or comment to: nanp@ethz.ch or rudi.gunawan@chem.ethz.ch.

6 Change log