**SINCERITIES: User Manual**

**Version 2.0 (April 2017)**

Author: Nan Papili Gao

Institute for Chemical and Bioengineering

ETH Zurich

E-mail: [nanp@ethz.ch](mailto:nanp@ethz.ch)

***\*\*Overview\*\****

SINCERITIES: SINgle CEll Regularized Inference using TIme-stamped Expression profileS”. SINCERITIES is a computational tool for inferring gene regulatory network (GRN) from time-stamped cross-sectional single cell expression data.

For further detailed information about SINCERITIES, please refer to the following paper:

Papili Gao N., Ud-Dean S.M.M., Gandrillon O. and Gunawan R., **SINCERITIES: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles.***Bioinformatics* (2017).

***\*\* Systems Requirements\*\****

This SINCERITIES toolbox is written for R. The subroutines in SINCERITIES (version 2.0) have been successfully tested on R 3.3.0. SINCERITIES requires the following R packages:

1. glmnet
2. kSamples
3. ppcor
4. cvTools
5. pracma
6. R.matlab

***\*\*Usage\*\****

Please refer to the file ***MAIN.R*** for an example script on how to use SINCERITIES subroutines.

Users can prepare the single cell expression data in an Excel worksheet and upload the data to R after saving the Excel worksheet to **CSV** format. Please prepare the data in an EXCEL sheet using the format below:

Data: s-by-m+1 matrix, where s is the total number of observations/single cells and m is the number of genes. The first m columns contain the expression level of each m genes, and the last column contains the time-stamps.

Two data formats are accepted:

A) with row header

------------------------------------

Gene1 Gene2 Gene3 ... Genej Time

27 80 56 ... 69 0

73 20 90 ... 45 0

. . . ... . .

. . . ... . .

. . . ... . .

------------------------------------

B) without row header

------------------------------------

27 80 56 ... 69 0

73 20 90 ... 45 0

. . . ... . .

. . . ... . .

. . . ... . .

------------------------------------

Network inference can be done by calling out SINCERITIES subroutine using any of the following

command lines.

a. result = SINCERITIES(DATA)

DATA, a list containing the following information:

- DATA$singleCELLdata: list of length n, where n is the number of capture time points. DATA$singleCELLdata[[k]] is a s\_k by m matrix/data frame containing observed expression levels of m genes in s\_k single cells.

- DATA$totDATA: S by m matrix, where S is the total number of single cells (i.e., S=s\_1+s\_2+...+s\_n where n the number of capture time points) and m is the number of genes.

- DATA$time: vector of length n containing the cell capture time points or time-stamps).

b. result = SINCERITIES(DATA,distance)

distance: this parameter selects the distribution distance

1- for KS (Kolmogorov-Smirnov) (\* DEFAULT \*)

2- for CM (Cramer-von Mises)

3- for AD (Anderson-Darling)

4- for Mean expression difference

c. result = SINCERITIES(DATA,distance,method)

method: this parameter selects the regularization regression strategy

1- for RIDGE (\* DEFAULT \*)

2- for ELASTIC-NET with automatic detection of optimal alpha parameter

3- for LASSO

4- for ELASTIC-NET with manual selection of alpha parameter

d. result = SINCERITIES(DATA,distance,method,noDIAG)

noDIAG: this parameter selects whether the auto-regulatory edge is inferred

1. GRN contains no auto-regulatory edge (\* DEFAULT \*)
2. GRN contain auto-regulatory edge

e. result = SINCERITIES(DATA,distance,method,noDIAG,SIGN)

SIGN: this parameter selects whether the sign / mode of the gene regulations is inferred

0- for unsigned GRN

1- for signed GRN (\* DEFAULT \*)

SINCERITIES uses partial correlation analysis where a positive (negative) correlation is taken as an indication of activation (repression).

The output of SINCERITIES is a list including the following information:

-adj\_matrix: m by m matrix containing the weights of regulatory edges. The larger adj\_matrix[i,j] indicates higher confidence that the corresponding edge exists (i.e., gene i regulating gene j).

-DISTANCE\_matrix: n-1 by m matrix containing the (normalized) distribution distance (DD) computed during the network inference.

**NOTE**: the default SINCERITIES function could **NOT** accommodate datasets with **fewer than five** **time points** due to the limitation of LOOCV, in this case users may use SINCERITIES\_PLUS function. For further details please read the section below.

***\*\*Additional implementation\*\****

SINCERITIES toolbox includes an additional function called SINCERITIES\_PLUS.R. Users may choose this option to infer GRN from datasets with less than five time points (but at least three). SINCERITIES\_PLUS function adopts RIDGE regression as regularization strategy and KS (Kolmogorov-Smirnov) metric for the evaluation of gene distribution distances.

f. result = SINCERITIES\_PLUS(DATA,noDIAG)

noDIAG: this parameter selects whether the auto-regulatory edge is inferred

0- GRN contains no auto-regulatory edge (\* DEFAULT \*)

1- GRN contain auto-regulatory edge

g. result = SINCERITIES PLUS (DATA,noDIAG,SIGN)

SIGN: this parameter selects whether the sign / mode of the gene regulations is inferred

0- for unsigned GRN

1- for signed GRN (\* DEFAULT \*)

SINCERITIES uses partial correlation analysis where a positive (negative) correlation is taken as an indication of activation (repression).

h. result = SINCERITIES PLUS (DATA,noDIAG,SIGN,CV\_nfolds)

CV\_nfolds: this parameter defines a partition of the data into CV\_nfolds distinct subsets for cross validation

CV\_nfolds = 5 (\* DEFAULT\*)

The outputs of SINCERITIES\_PLUS is a list including the following information:

-adj\_matrix: m by m matrix containing the weights of regulatory edges. The larger adj\_matrix[i,j] indicates higher confidence that the corresponding edge exists (i.e., gene i regulating gene j).

-DISTANCE\_matrix: n-1 by m matrix containing the (normalized) distribution distance (DD) computed during the network inference.

***\*\*Examples\*\****

Three examples are provided following the SINCERITIES manuscript.

1. ***example1\_in\_silico\_data.R*** for the inference of *in silico* data contained in the folder “In silico single cell data”. This example was used as the gold standard dataset in an accompanying manuscript.

The folder “*In silico* single cell data” contains the following files:

* 20\_nets\_10genes\_6UNEVENtime\_sigma01B\_no\_initial\_points2.mat
* 20\_nets\_10genes\_7UNEVENtime\_sigma01B\_no\_initial\_points2.mat
* 20\_nets\_10genes\_8UNEVENtime\_sigma01B\_no\_initial\_points2.mat
* 20\_nets\_10genes\_9UNEVENtime\_sigma01B\_no\_initial\_points2.mat
* 20\_nets\_10genes\_10UNEVENtime\_sigma01B\_no\_initial\_points2.mat
* 20\_nets\_10genes\_8UNEVENtime\_sigma02B\_no\_initial\_points2.mat
* 20\_nets\_10genes\_8UNEVENtime\_sigma03B\_no\_initial\_points2.mat
* 20\_nets\_10genes\_8UNEVENtime\_sigma04B\_no\_initial\_points2.mat
* 20\_nets\_20genes\_8UNEVENtime\_sigma01B\_no\_initial\_points2.mat

In each \*file.mat, users can find the simulated time-stamped cross-sectional single cell data (and the relative gene-gene interactions) for 20 gene networks (10 from yeast and 10 from *Escherichia coli*) sampled at n uneven time points with different order of intrinsic noise (sigma).

1. ***example2\_THP1\_data.R*** for the inference of the gene regulatory network of monocytic THP-1 human myeloid leukemia cell differentiation [1]. The data can be found the folder “THP1 data”.

In particular the folder “THP1 data” contains the following files:

* single\_cell\_kouno\_data.csv: the expression profiles 45 transcription factors (TFs) from 960 THP-1 cells that were collected at 8 distinct time points (0, 1, 6, 12, 24, 48, 72, 96 hours) post induction [1]
* SUBNET2\_tomaru.csv and tomaru2.csv: information about RNAi knockdown experiments of 20 anti-/pro-differentiation TFs described in [2].

1. ***example3\_THP\_data\_with\_4\_time\_points.R*** for the inference of the GRN of monocytic THP-1 human myeloid leukemia cell differentiation [1] (see above) by using SINCERITIES\_PLUS function on only four time-stamped data (0, 1, 12, 48 hours).

***\*\*Questions and Comments\*\****

Please address any problem or comment to: [nanp@ethz.ch](mailto:nanp@ethz.ch) or [rudi.gunawan@chem.ethz.ch](mailto:rudi.gunawan@chem.ethz.ch).

***\*\*Change log\*\****

***\*\*References\*\****

1. Kouno T, de Hoon M, Mar JC, Tomaru Y, Kawano M, Carninci P, et al. Temporal dynamics and transcriptional control using single-cell gene expression analysis. Genome Biol. 2013;14:R118.

2. Tomaru Y, Simon C, Forrest AR, Miura H, Kubosaki A, Hayashizaki Y, et al. Regulatory interdependence of myeloid transcription factors revealed by Matrix RNAi analysis. Genome Biol. 2009;10:R121.