TRANSCOMPP user guide

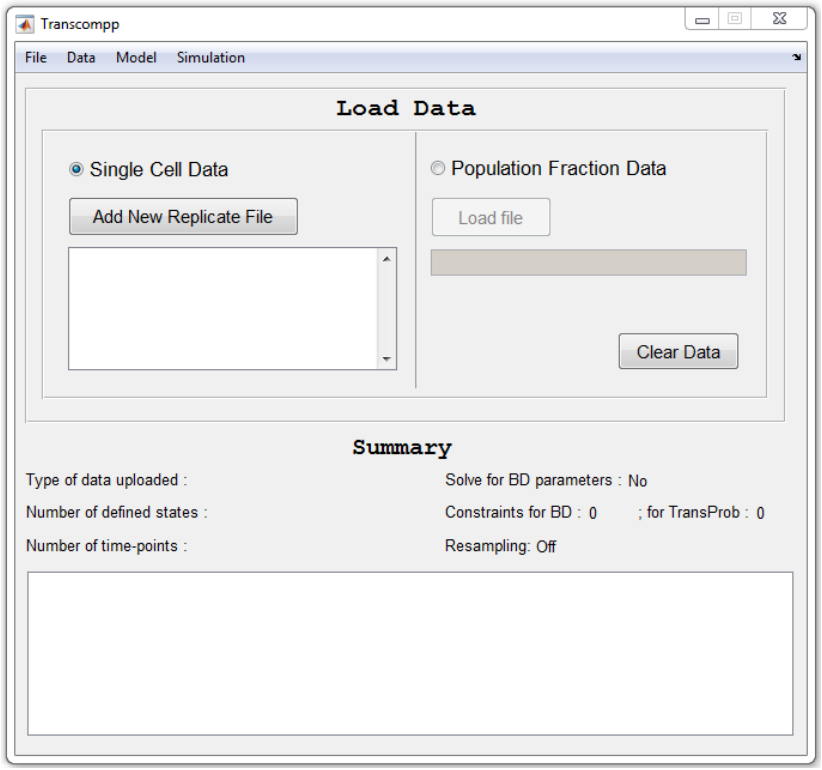
Author: N. Suhas Jagannathan (*suhas@duke-nus.edu.sg*)

TRANSCOMPP is a tool that allows users to quantify phenotypic plasticity starting from single-cell or bulk measurements of cellular phenotype. TRANSCOMPP uses Markov modeling to characterize distinct phenotypes as *phenotypic states*, and optimization to estimate best-fit values for stochastic transition rates, transition rate intervals, and phenotype-specific proliferation parameters.

Input data to TRANSCOMPP is either in the form of single-cell measurements of phenotype-related attributes, or in the form of summary statistics from bulk data indicating the relative abundance of each phenotypic state in the population. More details about the types and formatting requirements for input data can be found in additional files in the SampleData folder. The folder also contains example data files (for each data type) that can be used to verify if TRANSCOMPP is installed and functional.

The following guide introduces the user to the different windows of the TRANSCOMPP software and the options and parameters available for the user.

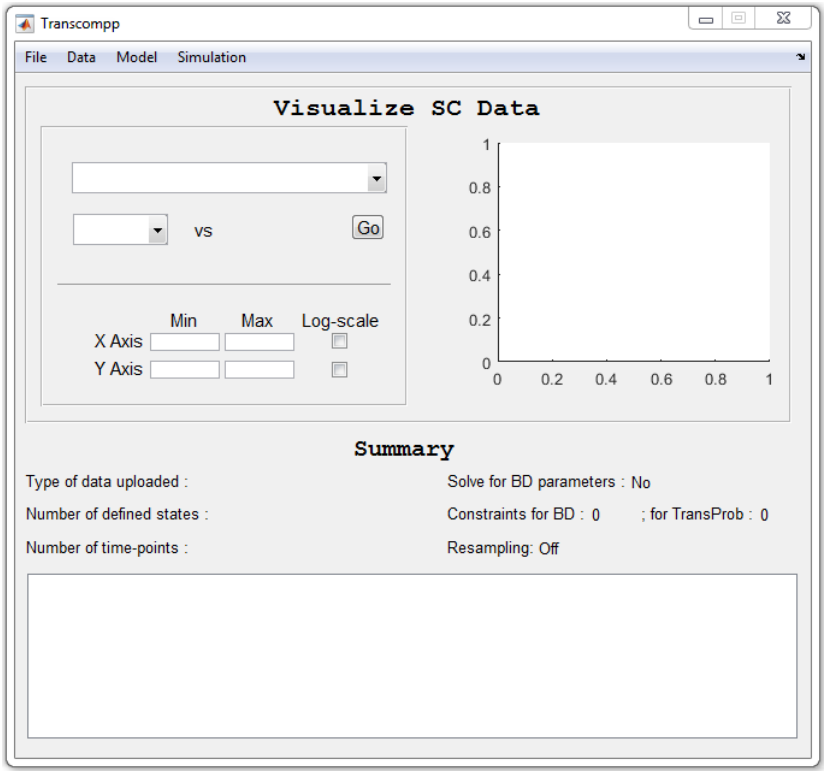
Load Data Window



The ***Load Data*** window is where the user can input the data required by TRANSCOMPP. Input data is expected to be in the form of spreadsheets (.xls, or .xlsx). The data in the spreadsheets can be either of two types, with radio buttons allowing the user to toggle between the two choices -

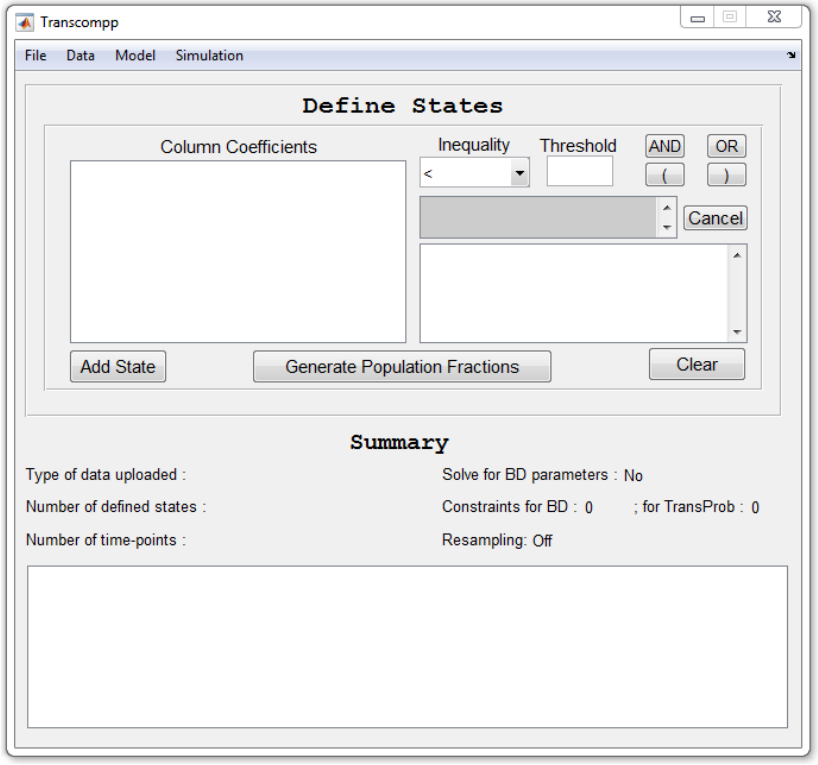
1) **Single Cell Data**: The input data in this case, are phenotypic measurements (biomarkers/gene expression/fluorescence etc.) of *single cells*. Each input file corresponds to a replicate of the biological experiment, and has multiple sheets inside, each of which corresponds to the time-point at which the experimental measurements were made. These time-points are in arbitrary units (minutes/hours/days) but are scaled such that they are all positive integers. It is *mandatory* that the sheets are named by the corresponding time-point in numeral form (e.g., "2", "4" etc.). Typically the earliest measurement made is considered to be time-point "0" and all subsequent time-points are shifted accordingly. Within each sheet, the columns represent the measured markers/genes etc., while each row represents a single cell. The first row contains the column names for each column (*mandatory*). In each sheet, a minimum of 100 cells are required for the analysis to proceed. Each time-point/replicate can have a different number of cells, but should have the same columns in the same order. As each replicate is added, the text box below the "**Add New Replicate File**" button is auto-populated with the name of the input file.
2) **Population Fraction Data**: The input data in this case, are fractional compositions (fraction of each phenotypic state in a population) for each replicate and time-point. As before, each sheet in the spreadsheet represents an experimental time-point (positive integers) and are named as such. Within each sheet, the rows represent individual replicates, and the columns represent each phenotypic state. No column/row headers are required. All sheets are expected to have the same column order for states and the same row order for replicates. In case of experiments that involve enrichment/purification of each phenotypic state, an empty row can be used to differentiate groups of replicates that correspond to trajectories following the fate of enriched populations of each state. Successful loading of the file is indicated by the auto-population of the text box beneath the "**Load file**" button.

Visualize SC Data window



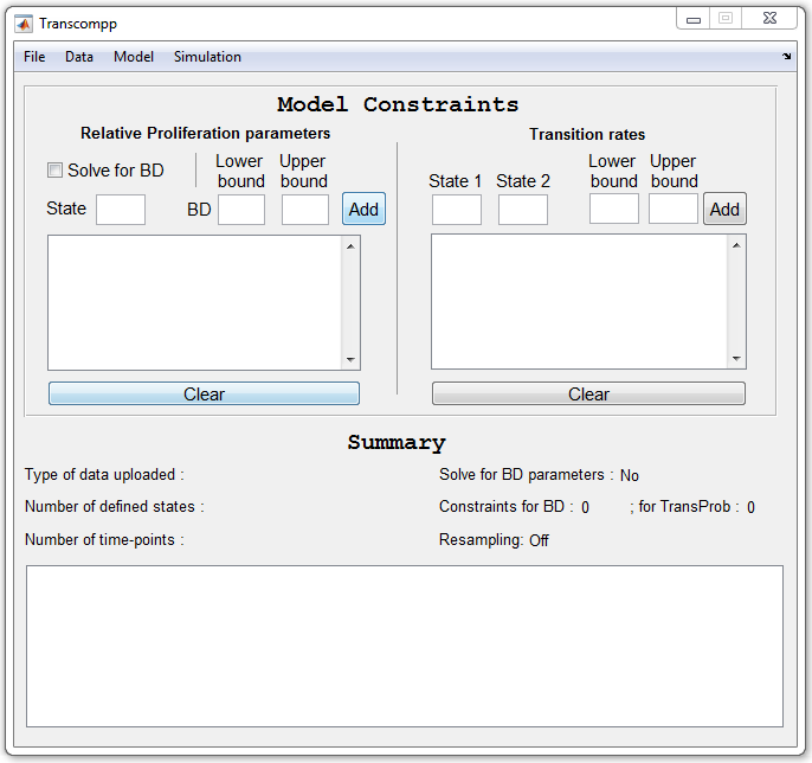
The ***Visualize SC Data*** window is intended for use only when Single cell (SC) data has been input in the "***Load Data***" window. It is used for users to visualize patterns in the data by plotting any two columns from any replicate against each other. Options allow users to change each axis to log-scale, and to restrict axes bounds. By default, TRANSCOMPP performs PCA analysis on the input data to identify the principal components for each replicate. These components are then represented as "PC1", "PC2"… in the choice of X-axis or Y-axis parameters (dropdown menu), and are available for visualization.

Define States window



The ***Define States*** window is where the user can input rules about how to use the Single cell data to define individual phenotypic states. Once SC data is loaded in the ***Load Data*** window, the first column of the **Column Coefficients** table in the ***Define States*** window is auto-populated with the marker names (column names in the input spreadsheets), plus the principal components (PC1, PC2 etc). The second column of this table is editable and is expected to contain the "weights" associated with each marker for each state definition. For example, if two example marker names are CD24 and CD44 and states are classified by the value of (CD24 marker + CD44 marker) being greater than or lesser than a threshold, the second column would be edited such that the rows corresponding to CD24 and CD44 have 1, while all other rows are 0. The inequality dropdown and the appropriate threshold would also have to be populated by the user. By clicking "**Add State**", the user adds this defined state to the list and the white text box is populated to reflect the new state. The additional buttons "AND", "OR", "(", and ")" can be used to chain rules to define an individual state (e.g., CD24 > 100 AND (CD44 < 50 OR EpCAM = 10)) . When chaining, the current state definition is displayed in the gray text box and can be canceled at any point before adding the state definition. Once all states are added, "**Generate Population Fractions**" should be clicked to get fractional composition data. It is the responsibility of the user to define states such that the same cells are not classified into multiple states.

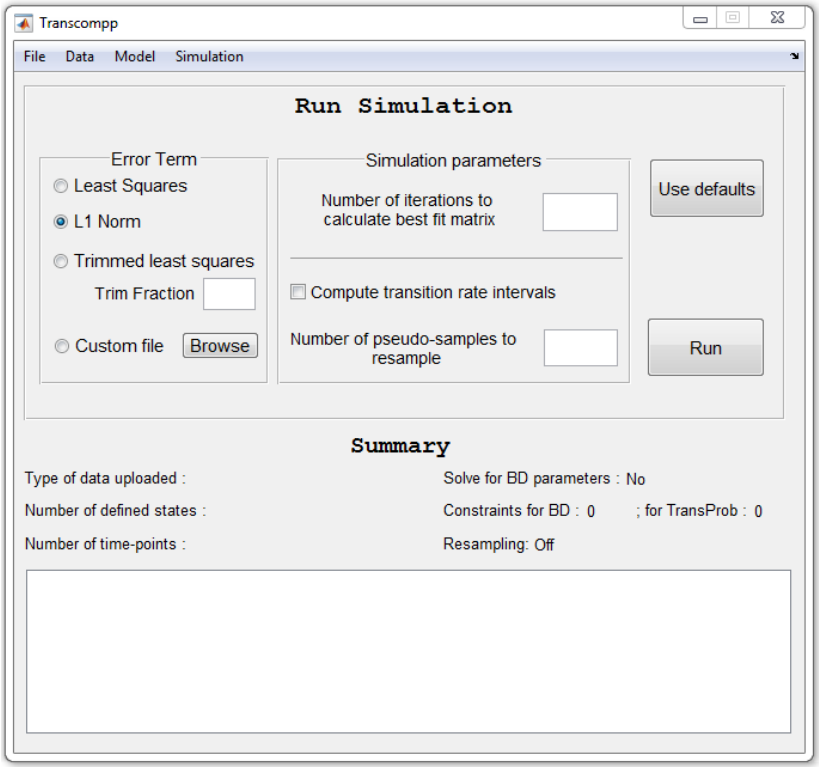
Model Constraints window



The ***Model Constraints*** window is where the user can add constraints for the optimization problem. The constraints can be of two types – Constraints in proliferation (BD) parameters, and constraints in the transition rates between states. When using SC data, it is *mandatory* to define states before adding constraints.

1) **Relative proliferation parameters**: A checkbox toggles the option to solve for the best-fit proliferation parameters (BD vector, see main text) in addition to the transition rates. When this is checked, TRANSCOMPP will compute these parameters such that the BD scalar for each state falls within any range given by the user. By default this scalar for the first state is taken to be 1 (can be overridden) and those of all other states are allowed to vary between 0.1 and 10 (10-fold difference). This can be overridden. For any constraint to be added, the user inputs the state number, followed by the presumed lower and upper bound for the relative BD scalar for that state. When these relative values are known and are not to be solved for, the "**Solve for BD params**" checkbox is unchecked and the constraints are entered by using the state number and *the same known value* for the lower and upper bound.
2) **Transition rates**: Transition rates (probabilities) are *always* in the range 0-1. If any particular transition rate is known or expected to fall in a narrower range (e.g 0.9 – 0.95), those constraints can be added here. By default, self-transitions (transitions of a state to itself) are expected to be greater than 0.5. This can however be overridden. The user enters constraints by inputting details about the the source state (State 1), the destination state (State 2) and the expected range within which these rates lie (lower bound and upper bound).

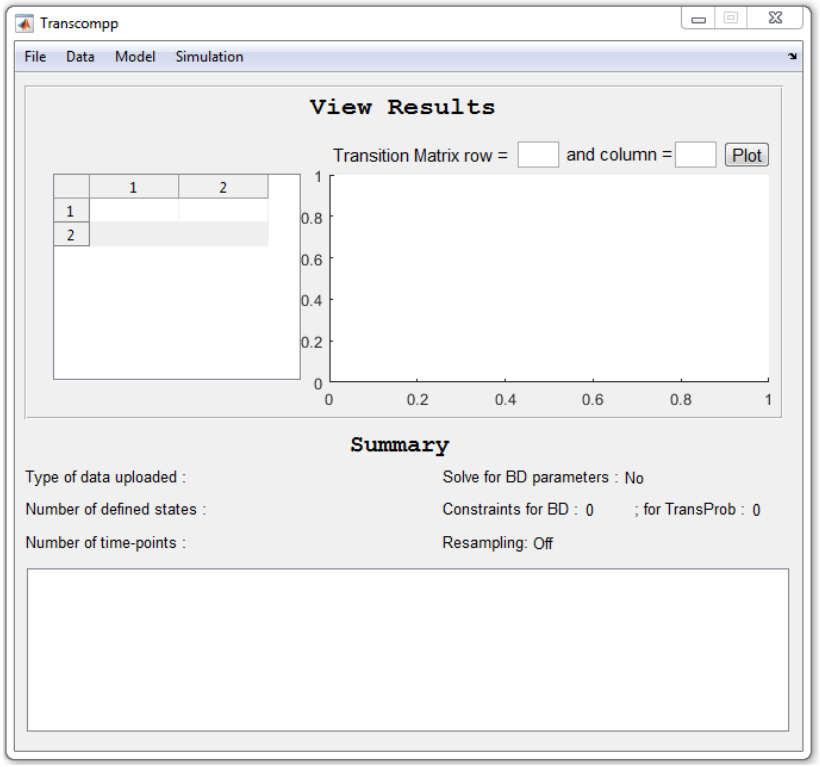
Run Simulation window



The ***Run Simulation*** window is where the final parameters of the simulation are set and the simulation is executed. First, the user has a choice of error term that defines the objective function to be optimized. TRANSCOMPP has inbuilt support for three error terms – Least Squares (Sum of squared residuals), L1 norm (Manhattan distance), and Trimmed Least squares. In addition, TRANSCOMPP also allows users to enter their own error term function. Such a custom function takes two arguments – the predicted and observed population fractional compositions *at a single time-point*. The function then outputs a single value that is the error of fit for that replicate and time-point.

Next, the user enters simulation parameters such as the "**Number of iterations to calculate best fit matrix**" – which is the number of times each optimization process is repeated, starting from a different random seed (searching for global minima). The user can also check or uncheck the box that controls whether or not resampling is performed. When unchecked, only the best-fit matrix is computed. When this box is checked (Simulation takes longer), a vector of values for each transition rate is returned (See resampling method in the main text). The size of the vector is given by the user-customizable "**Number of pseudo samples to generate**".

View Results window

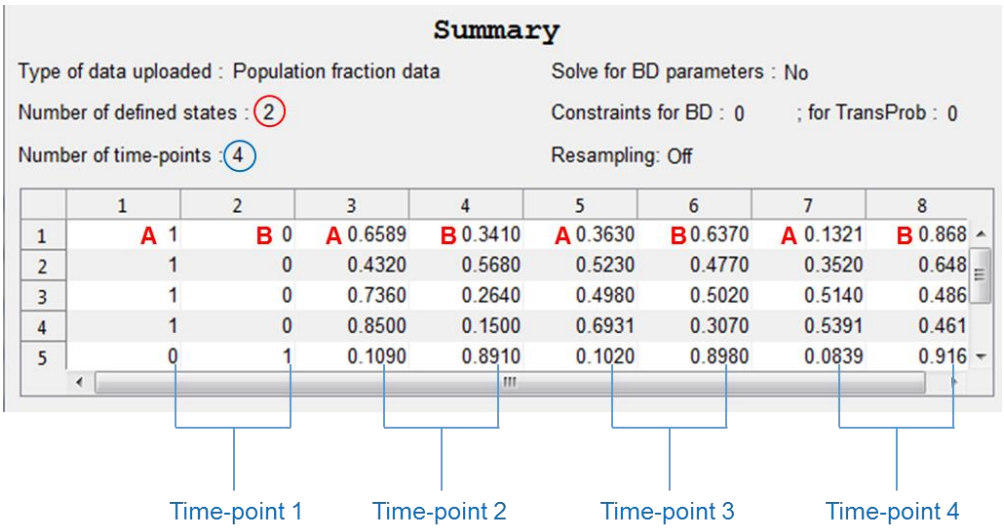


The ***View Results*** window is where users can view the output of the simulation. When no resampling is done, the user only sees a populated table on the left that contains the best-fit transition matrix. When resampling is performed, each element in the table is of the form: *mean transition rate* ± *standard deviation*. Resampled simulations also allow users to plot distributions of any particular transition rate, by entering the row and column numbers of the particular rate, in the transition matrix.

Summary Panel

At the bottom part of each of these above windows is a panel called summary, which is updated as the user makes their data/parameter choices. It indicates what kind of data is uploaded (Single-cell or population fractions), how many phenotypic states have been defined. How many time-points do the experiments have, whether the user opts to solve for BD parameters, How many BD or Transition probability constraints have been added by the user, and whether the user intends to do resampling to obtain transition rate intervals and distributions.

This panel also displays a table containing the population fraction data (either loaded directly, or computed from Single-cell data) that will be used to obtain the best-fit transition matrices. Each row of this table represents an individual experimental replicate. Within each row are the fractions of each phenotypic state, concatenated across time-points. For example consider the example below.



	1	2	3	4	5	6	7	8
1	A 1	B 0	A 0.6589	B 0.3410	A 0.3630	B 0.6370	A 0.1321	B 0.868
2	1	0	0.4320	0.5680	0.5230	0.4770	0.3520	0.648
3	1	0	0.7360	0.2640	0.4980	0.5020	0.5140	0.486
4	1	0	0.8500	0.1500	0.6931	0.3070	0.5391	0.461
5	0	1	0.1090	0.8910	0.1020	0.8980	0.0839	0.916

This example displays a data set with two phenotypic states (States A and B) and four measured time-points. So the first row has the fractions of states A and B at time-point 1, followed by corresponding fractions of A and B at time-points 2, 3 and 4. Note that the order of states within each time-point is the same. Also note that within a time-point, the sum of all (both in this case) phenotypic fractions equals 1.

Menu

Each of these above windows can be accessed from the menu bar at the top of the program. In addition, the menu bar option "File" allows the user to do the following.

New Analysis

The user can delete all entered fields in all windows, and start a fresh simulation workspace. All entered data will be deleted in this case.

Load Analysis Workspace File

The user can load in a previously saved analysis workspace file.

Save Analysis Workspace File

The user can save current analysis workspace as a workspace file. The file includes all entered data, constraints, state definitions and simulation parameters/checkbox options. Note that in the case of single-cell data, although state definitions are included, the user will have to manually click "**Generate Population Fractions**" before proceeding with the simulation.

Export Results

The results of the simulations (namely the computed best-fit or resampled transition rates, and best-fit BD parameters if required) can be saved as a .mat file

Exit

Exit the program