Instruction of R scripts used in SCRIP for analysis

The datasets supporting the conclusions of this article are available in the Gene Expression Omnibus (GEO) accession, including GSE81608 (Xin), GSE65525 (Klein), GSE75140 (Camp), GSE77288 (Tung), GSE72056 (Tirosh), GSE139827 (Zhou) and GSE128890 (Seale). Read counts data were used for simulation.

Simulation

Gen simulation data using SCRIP.R

This file shows the R code for generating simulation data using methods from SCRIP (i.e., GP-commonBCV, GP-trendedBCV, BP, BGP-commonBCV, BGP-trendedBCV). To evaluate the simulation of a particular cell type, we simulated scRNA-seq counts from one cell type in each of 8 real datasets described above. Specifically, we simulated the cell type with the largest number of cells from each of Xin (human pancreatic islets), Tirosh (human melanoma tumor), Zhou (mouse artery wall Sca1+ cells), Seale (human adipose tissue) and Seale (mouse adipose tissue) datasets. We also simulated K562, induced pluripotent stem and whole brain organoid cells respectively from other three datasets (Klein, Tung, Camp), each of which measured a single type of cells. For each dataset, we simulated the same number of cells and genes as that in the corresponding real dataset. To know more details about how SCRIP was used and detailed description of SCRIP R functions, please check https://github.com/thecailab/SCRIP/blob/main/vignettes/SCRIPsimu.pdf.

Gen simulation data using other simulators.R

This file shows the R code for generating simulation data using other simulators (i.e., scDesign, scDD, SPARSim, powsimR, dyngen, SymSim).

Mean-variance-plots

Plot mean variance plot for simulation data using SCRIP.R

This file shows how we generate mean variance plots for simulation data from SCRIP (i.e., GP-commonBCV, GP-trendedBCV, BP, BGP-commonBCV, BGP-trendedBCV.).

Plot mean variance plot for other simulators.R

This file shows how we generate mean variance plots for simulation data from other simulators (i.e., scDesign, scDD, SPARSim, powsimR, dyngen, SymSim).

Heatmap-Boxplots

Heatmap and boxplots for simulation data using SCRIP.R

This file shows detailed analysis steps to evaluate simulation data from SCRIP using median absolute deviation (MAD) with five characteristics (i.e., gene-wise expression mean, variance, cell-wise zero proportion, gene-wise zero proportion, and library size) in eight real datasets (Xin, Klein, Tung, Camp, Tirosh, Zhou, Seale human, Seale mouse). $MAD = median(|log2(X_s/X_r)|)$, where X_s is the ordered data feature such as gene-wise data variance from each simulation data, and X_r is the ordered data feature from the corresponding real dataset. Smaller MAD indicates higher similarity between simulation and real data.

heatmap and Boxplots for other simulators.R

This file shows detailed analysis steps to evaluate simulation data from outstanding methods in SCRIP (GP-trendedBCV, BGP-trendedBCV) and other simulators using MAD with five characteristics (i.e., gene-wise expression mean, variance, cell-wise zero proportion, gene-wise zero proportion, and library size) in eight real datasets (Xin, Klein, Tung, Camp,Tirosh, Zhou, Seale human, Seale mouse).

DE

DE.R

This file shows how simulation data were generated for DE analysis methods. To simulate multi-group single-cell type data, we estimated hyperparameters from Tung's stem cell dataset and simulated two experimental groups for comparison. SCRIP is able to simulate different DE rates, fold change, dropout rate, library size, BCV.df and BCV.shrink. In this study, we generated 200 differentially expressed genes (DEGs) among 2000 genes of 100 cells in two samples for comparison. Multiple scenarios were simulated with fold changes (1.2, 1.4, 1.6, and 1.8) of base cell means of DE genes, library size (5k, 10k, 20k and 40k), dropout rate (0.1, 0.2, 0.3, and 0.4), BCV.df (2, 5, 10, and 20) and BCV.shrink (0.5, 1.0, 1.5, and 2.0) for evaluation in DE analysis. These 200 DE genes were simulated to be equally distributed that 100 genes were upregulated and the other 100 were downregulated.

Then DE analyses were conducted with methods including edgeR, DESeq2. Limma-voom, MAST, ZINB-Wave.edgeR, ZINB-Wave.DESeq2 and ZINB-Wave.limma_voom. We also provide scripts about how we evaluate the DE results using the areas under the ROC curve (AUC) and power with fixed alpha.

Trajectory

Gen simulation data for trajectory evaluation.R

This R file shows how we generate simulation data for each celltype using SCRIP with VEGs and other simulators. Then simulation data for each celltype were combined as the final simulation data for each dataset. Tirosh dataset was used as example. SCRIP preserves VEGs, which is not capable in existing methods. VEGs from each cell type were identified by the feature selection process from Seurat package (Satija et al., 2015) from real data. To simulate these VEGs, SCRIP used λ' captured from their expression means in real data and other simulation steps (including simulating over-dispersed expected count from Gamma or Beta distribution, and then simulating counts from Poisson distribution) to synthesize the final data. In this study, we simulated data based on 6 cell types from the Xin dataset, 6 cell types from Tirosh dataset, 10 cell types from Seale (mouse) and 12 cell types from Zhou dataset to evaluate the performance of SCRIP and other methods in recovering cell group information. 1000 VEGs were identified and simulated for each cell type.

Trajectory evaluation.R

This R file shows how we use simulation data to generate trajectory plots with Monocle R package and plot heatmap for cell distance evaluation. Pseudotime time of cells between simulation and real data were mapped from the first two dimensions. Therefore, each cell ordered by first two dimensions with both directions were compared between simulation data and real data separately. Then smallest rank of these four scenarios were used as the final order for comparison. Smaller orders with higher similarity of the cell trajectories to that of real data implies better performance for the simulation method to recover real data.