Simulated dataset:

In the simulation part, we generate different data for method validation.

Firstly, we consider two scenarios for cell types, i.e. one cell type and three cell types.

Secondly, we adjust parameters that decide dropout rate under each scenario to derive data with varying dropout rate, and each time we only change one parameter (each have 5 values) for the convenience of evaluating our method. In total, we have 40 different settings, and under each setting, we generate data with 100 iterations.

To be detailed, we set number of cells = 2000, number of genes = 10000, batch number = 2 with each batch containing 1000 cells .

Dropout rate parameters (GLMM, see paper): Alpha = (0,0.05,0.10,0.15,0.20), Beta = (0,0.05,0.10,0.15,0.20), Gamma = (-1.5,-1,-0.5,0,0.5), Theta = 1

Batch effect parameters (normal distribution, see paper): nu1 = (-0.2,-0.15,-0.10,-0.05,0), nu2 = (0.2,0.15,0.10,0.05,0), sigma1 = sigma2 = 1

When cell type = 1, all cells are indexed type 1. When cell type = 3, we separately allocate 250, 350, 400 cells in type 1, 2, 3 for batch 1, and 300, 350, 350 cells for batch 2.

Simulations:

In this section,we evaluate the performance of DABEA in datasets. We focus on ...

1 Settings:

1. we simulate 40 datasets, with G = 10000 genes and N = 2000 single cells in two different batches , each containing 1000 cells. gene length and total read count are generated from the exponential of Gaussian distribution, true expression matrix is generated from chi-square distribution. The parameters which are varied in the simulations are as follows.
2. Two scenarios are defined, i.e. K = 1 and 3 different cell types. When K = 3, batch 1 contains 250, 350, 400 cells in each cell type respectively, and batch 2 contains 300, 350, 350 cells respectively.
3. Recall the dropout rate model: ...... variables: Dropout rate parameters (GLMM, see paper): Alpha = {0,0.05,0.10,0.15,0.20}, Beta = {0,0.05,0.10,0.15,0.20}, Gamma = {-1.5,-1,-0.5,0,0.5}, Theta = 1

3) Batch effect parameters (normal distribution, see paper): cell specific batch effects bc ~(nui,sigmai), where mean of the normal distributions in all batch sum up to zero. To be specific, when generating simulated data, we set nu1 = {-0.2,-0.15,-0.10,-0.05,0}, nu2 = {0.2,0.15,0.10,0.05,0}, sigma1 = sigma2 = 1

Expression matrix Y is generated from zero-inflated Poisson distribution.

1. ...
2. ...
3. ...

Real data