Supplementary Information

JOINT for Large-scale Single-cell RNA-Sequencing Analysis via Soft-clustering and Parallel Computing

Fig. S1: Convergence of the JOINT algorithm with iterations.

Convergence of $q_{g,k,l}$ (**a**), $\alpha_{g,k,l}$ (**b**), $\beta_{g,k,l}$ (**c**), and π_k (**d**) for different genes and cell clusters to true values with iterations.

Fig. S2: Convergence of the JOINT algorithm with number of samples.

Convergence of $q_{g,k,l}$ (**a**), $\alpha_{g,k,l}$ (**b**), $\beta_{g,k,l}$ (**c**), π_k (**d**), m (**e**, $\left|m_{g,k} - \hat{m}_{g,k}\right|/m_{g,k}$, the mean of absolute difference between the theoretical mean from zero-inflated negative binomial model and the mean from model using estimated parameters over the theoretical mean),

p^0 (**f**, $\left|p^0_{g,k} - \hat{p}^0_{g,k}\right|/p^0_{g,k}$, the mean of absolute difference between the theoretical zero-count probability from zero-inflated negative binomial model and the zero-count probability from model using estimated parameters over the theoretical probability), *var* (**g**, $\left|var_{g,k} - \widehat{var}_{g,k}\right|/ var_{g,k}$, the mean of absolute difference between the theoretical variance from zero-inflated negative binomial model and variance from model using estimated parameters over the theoretical variance) to true values with the number of samples. Error bars in (**a**) - (**d**) indicate the full range of data variation.

Fig. S3: Convergence of the JOINT algorithm with dropout probabilities.

Convergence of $q_{g,k,l}$ (**a**), $\alpha_{g,k,l}$ (**b**), $\beta_{g,k,l}$ (**c**), π_k (**d**), and m (**e,** $\left|m_{g,k} - \hat{m}_{g,k}\right|/m_{g,k}$, i.e. the mean of absolute difference between the theoretical mean from zero-inflated negative binomial model and the mean from model using estimated parameters over the theoretical mean) to true values with dropout probabilities. Error bars in (**a**) - (**d**) indicate the full range of data variation.

Fig. S4: The ratio of mean gene expression between pyramidal CA1 neurons and oligodendrocytes in the Zeisel dataset.

(**a**) - (**b**) Histogram of α (**a**) and β (**b**) values for each gene when pyramidal CA1 neuron expression counts were used in model training. (**c**) Histogram of the ratio of mean gene expression between pyramidal CA1 neurons and oligodendrocytes. Note the median of the gene expression ratio between cells with "CA1 Pyramidal" and "Oligodendrocytes" labels in the Zeisel dataset is 1.5.

Fig. S5: Simulated data at different dropout probabilities and DEG numbers.

(**a**) Simulated datasets with three clusters when there is no dropout and DEG number set to 150, 100, and 50. (**b**) Simulated dataset with three clusters when dropout probability is set to 0.1, and DEG number set to 150, 100, and 50. (**c**) Simulated dataset with three clusters when dropout probability is set to 0.2, and DEG number set to 150, 100, and 50. (**d**) Simulated dataset with three clusters when dropout probability is set to 0.3, and DEG number set to 150, 100, and 50. (**e**) Simulated dataset with three clusters when dropout probability is set to 0.4, and DEG number set to 150, 100, and 50. For datasets with dropout, we applied the PCA from the original dataset without dropout to obtain the 2-dimensional plot. These simulated data show the impact of dropout probability and DEG number on the destruction of single-cell data.

Fig. S6: Comparison of clustering performance of different algorithms at various dropout probabilities and DEG numbers.

(**a**) Cell clustering by Saver, scImpute, and JOINT on a simulated dataset with three clusters (dropout probability set to 0.1 and DEG number set to 50). Original data without dropout is shown on the left. K-means clustering method is used for published imputation algorithms. Adjusted Rand Index for each algorithm is shown. Imputation algorithm in JOINT is used for data visualization. (**b**) Cell clustering by Saver, scImpute, and JOINT on a simulated dataset with three clusters (dropout probability set to 0.1 and DEG number set to 100). (**c**) Cell clustering by Saver, scImpute, and JOINT on a simulated dataset with three clusters (dropout probability set to 0.1 and DEG number set to 150). (**d**) Cell clustering scores are compared for Saver, scImpute, and JOINT algorithms at different dropout probabilities on a dataset with 100 DEG. (**e**) Correlation of cell clustering results from Saver, scImpute, and JOINT to original "true labels" averaged across all genes (Gene Correlation) or cells (Cell Correlation) at different dropout probabilities. Correlation coefficients generated from a dataset with 100 DEG are shown. (**f**) - (**g**) The JOINT algorithm determines cell cluster numbers automatically by likelihood (**f**) and AIC (**g**) tests. For each dataset, we applied the PCA from the original dataset without dropout to obtain the 2-dimensional plot.

Fig. S7: Cell clustering data visualization by the JOINT imputation algorithm at different dropout probabilities and DEG numbers.

(**a**) - (**d**) Cell clustering by JOINT on a simulated dataset with three clusters when dropout probability is set to 0.1 (**a**), 0.2 (**b**), 0.3 (**c**), and 0.4 (**d**), and DEG number set to 150, 100, and 50. For each dataset, we applied the PCA from the original dataset without dropout to obtain the 2-dimensional plot.

Fig. S8: EM algorithm in JOINT improves the performance of cell clustering.

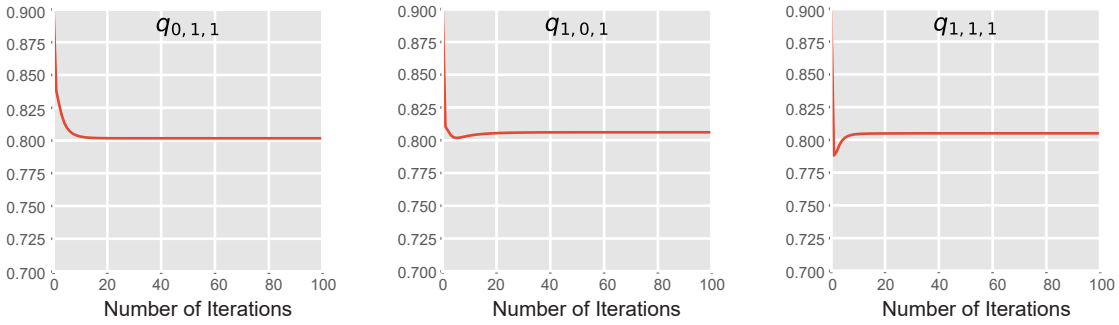
(**a**) Clustering scores that JOINT obtained on the Zeisel dataset when the initial points were selected by the K-means method, with and without application of the EM algorithm. (**b**) Clustering scores that JOINT obtained on the Zeisel dataset when the initial points were randomly selected, with and without application of the EM algorithm.

Table S1: Comparison of clustering performance for JOINT and published imputation algorithms on a simulated dataset.

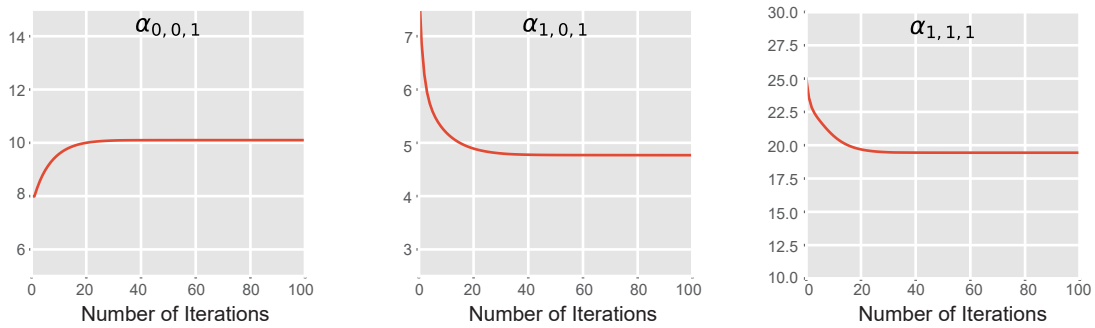
Table S2: Comparison of computing time when JOINT is run on GPU *vs.* CPU.

Fig. S1

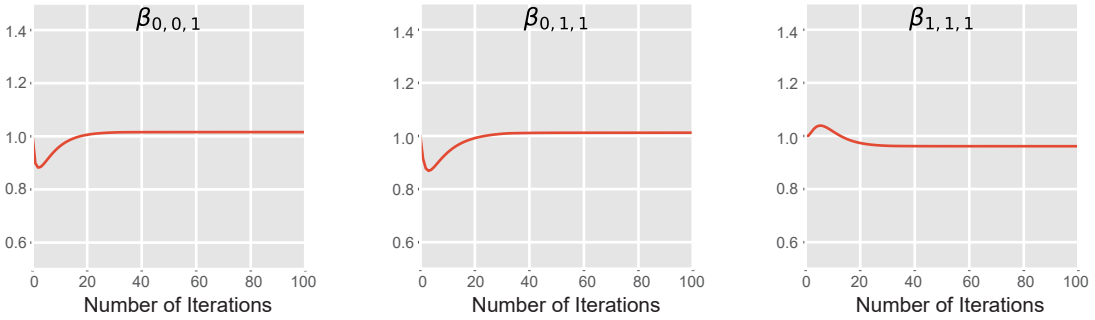
a


$q_{0,1,1}$
$q_{1,0,1}$
$q_{1,1,1}$
Number of Iterations

b


$\alpha_{0,0,1}$
$\alpha_{1,0,1}$
$\alpha_{1,1,1}$
Number of Iterations

c


$\beta_{0,0,1}$
$\beta_{0,1,1}$
$\beta_{1,1,1}$
Number of Iterations

d

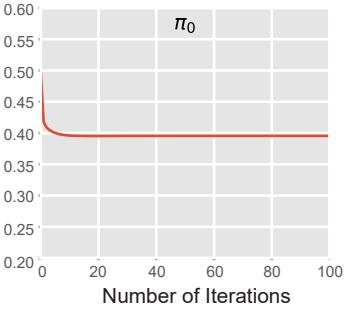

π_0
Number of Iterations

Fig. S2

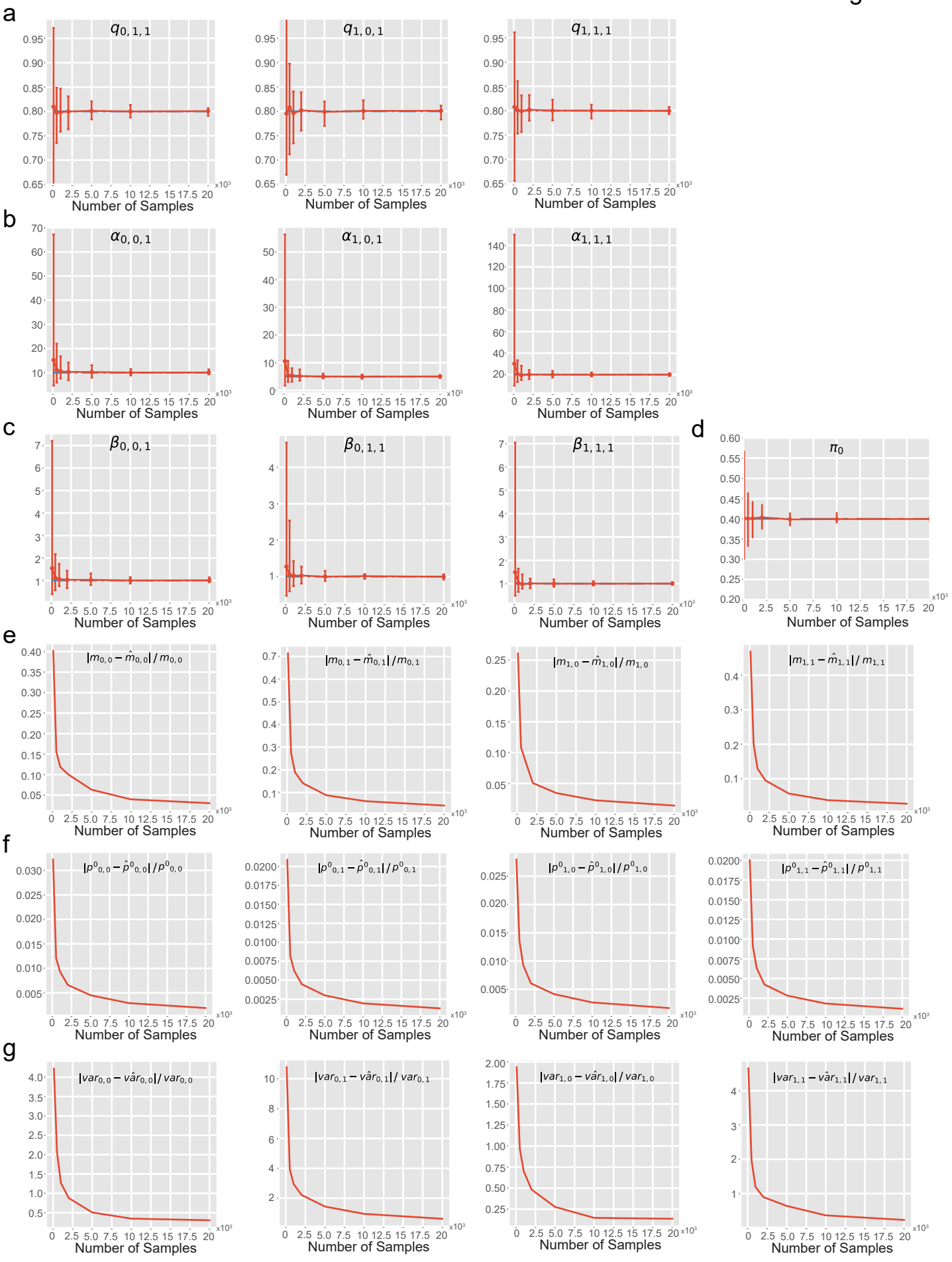

a
$q_{0,1,1}$
$q_{1,0,1}$
$q_{1,1,1}$
b
$\alpha_{0,0,1}$
$\alpha_{1,0,1}$
$\alpha_{1,1,1}$
c
$\beta_{0,0,1}$
$\beta_{0,1,1}$
$\beta_{1,1,1}$
d
π_0
e
$|m_{0,0} - \hat{m}_{0,0}| / m_{0,0}$
$|m_{0,1} - \hat{m}_{0,1}| / m_{0,1}$
$|m_{1,0} - \hat{m}_{1,0}| / m_{1,0}$
$|m_{1,1} - \hat{m}_{1,1}| / m_{1,1}$
f
$|p^0_{0,0} - \hat{p}^0_{0,0}| / p^0_{0,0}$
$|p^0_{0,1} - \hat{p}^0_{0,1}| / p^0_{0,1}$
$|p^0_{1,0} - \hat{p}^0_{1,0}| / p^0_{1,0}$
$|p^0_{1,1} - \hat{p}^0_{1,1}| / p^0_{1,1}$
g
$|var_{0,0} - \hat{var}_{0,0}| / var_{0,0}$
$|var_{0,1} - \hat{var}_{0,1}| / var_{0,1}$
$|var_{1,0} - \hat{var}_{1,0}| / var_{1,0}$
$|var_{1,1} - \hat{var}_{1,1}| / var_{1,1}$
Number of Samples
x10³

Fig. S3

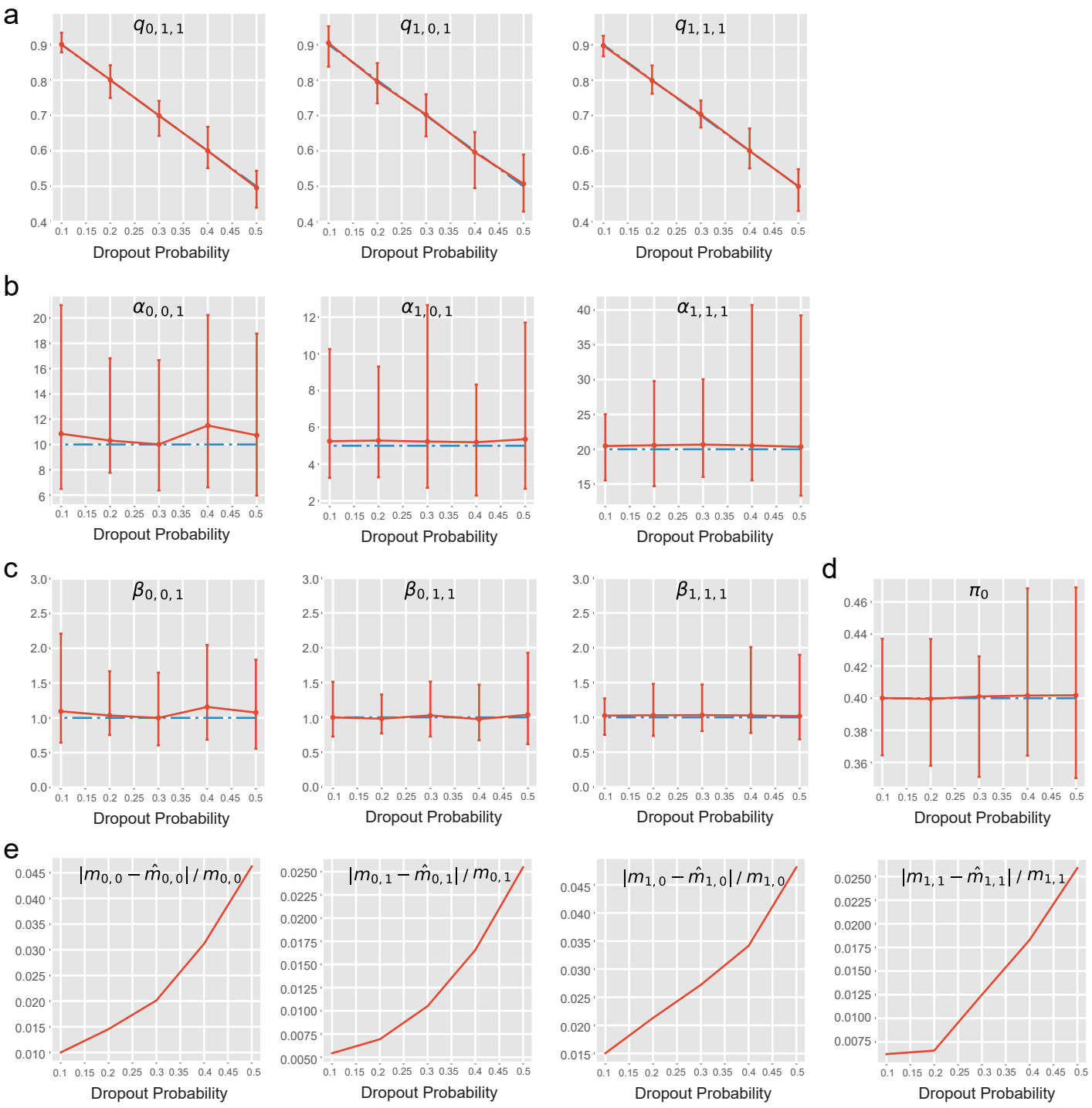

a
$q_{0,1,1}$
$q_{1,0,1}$
$q_{1,1,1}$
b
$\alpha_{0,0,1}$
$\alpha_{1,0,1}$
$\alpha_{1,1,1}$
c
$\beta_{0,0,1}$
$\beta_{0,1,1}$
$\beta_{1,1,1}$
d
π_0
e
$|m_{0,0} - \hat{m}_{0,0}| / m_{0,0}$
$|m_{0,1} - \hat{m}_{0,1}| / m_{0,1}$
$|m_{1,0} - \hat{m}_{1,0}| / m_{1,0}$
$|m_{1,1} - \hat{m}_{1,1}| / m_{1,1}$
Dropout Probability

Fig. S4

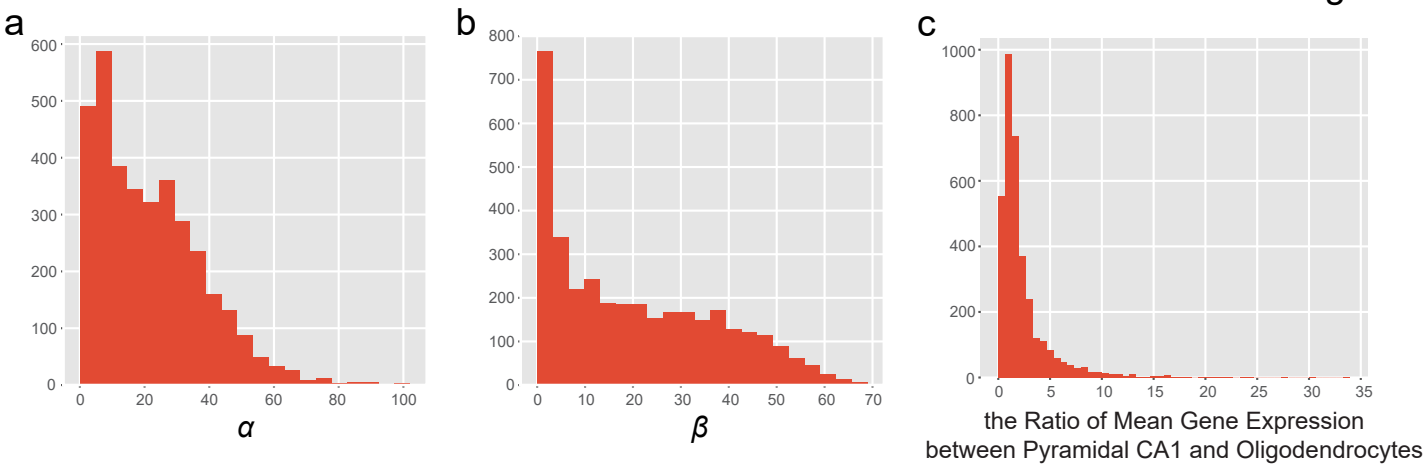

a
600
500
400
300
200
100
0
0
20
40
60
80
100
α
b
800
700
600
500
400
300
200
100
0
0
10
20
30
40
50
60
70
β
c
1000
800
600
400
200
0
0
5
10
15
20
25
30
35
the Ratio of Mean Gene Expression
between Pyramidal CA1 and Oligodendrocytes

Fig. S5

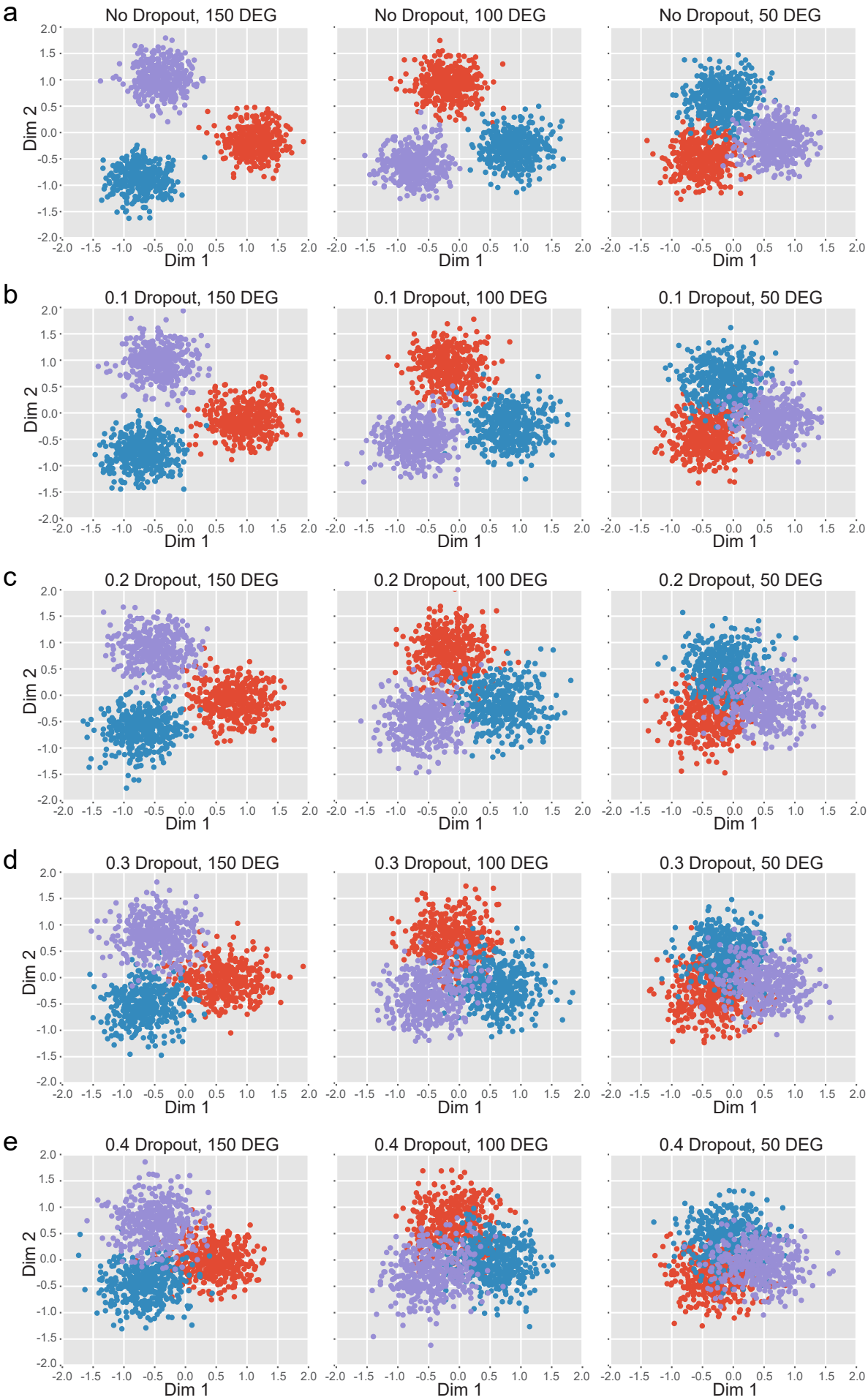

a
No Dropout, 150 DEG
No Dropout, 100 DEG
No Dropout, 50 DEG
b
0.1 Dropout, 150 DEG
0.1 Dropout, 100 DEG
0.1 Dropout, 50 DEG
c
0.2 Dropout, 150 DEG
0.2 Dropout, 100 DEG
0.2 Dropout, 50 DEG
d
0.3 Dropout, 150 DEG
0.3 Dropout, 100 DEG
0.3 Dropout, 50 DEG
e
0.4 Dropout, 150 DEG
0.4 Dropout, 100 DEG
0.4 Dropout, 50 DEG
Dim 1
Dim 2
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0

Fig. S6

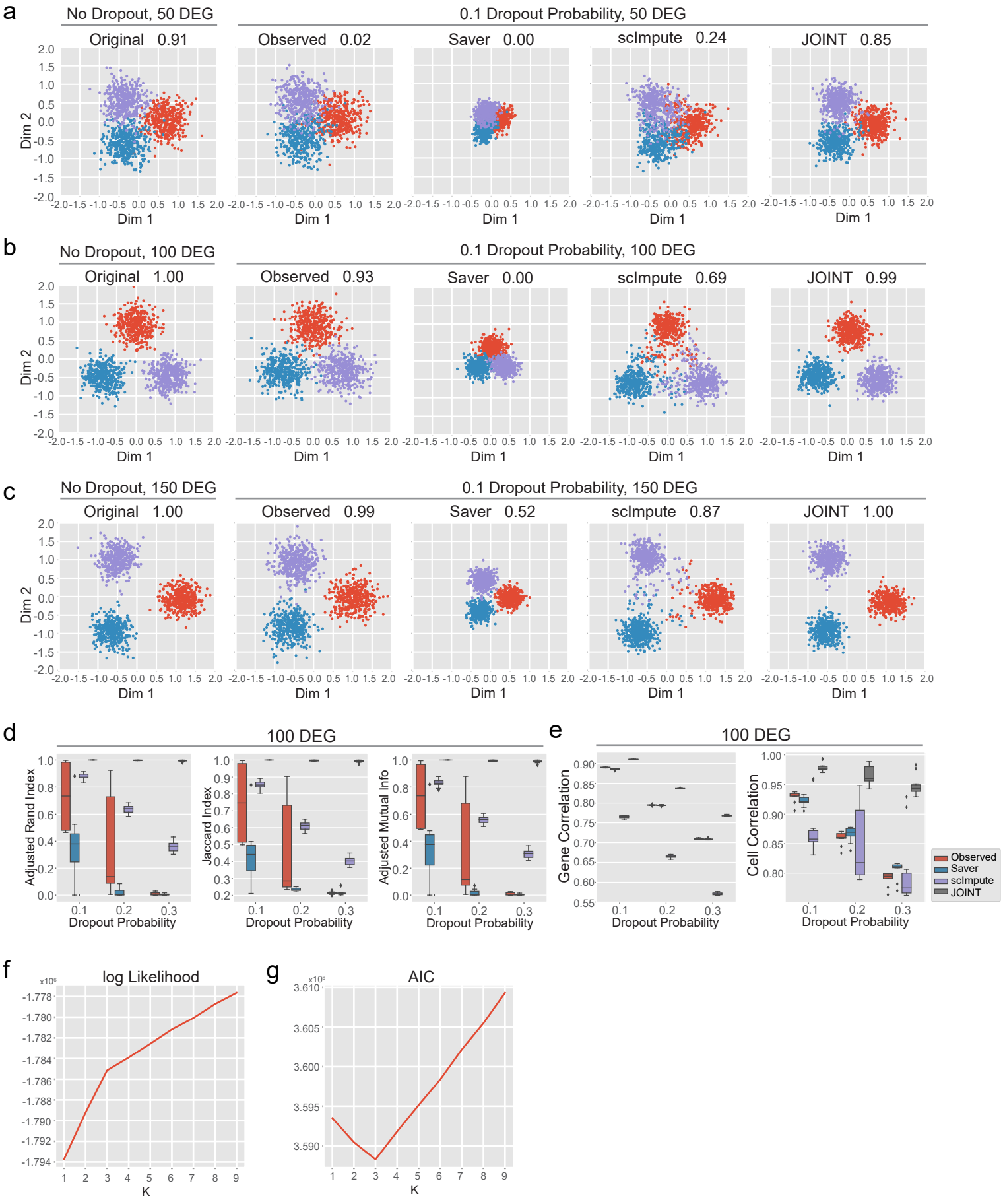

a
No Dropout, 50 DEG
0.1 Dropout Probability, 50 DEG
Original 0.91
Observed 0.02
Saver 0.00
scImpute 0.24
JOINT 0.85
b
No Dropout, 100 DEG
0.1 Dropout Probability, 100 DEG
Original 1.00
Observed 0.93
Saver 0.00
scImpute 0.69
JOINT 0.99
c
No Dropout, 150 DEG
0.1 Dropout Probability, 150 DEG
Original 1.00
Observed 0.99
Saver 0.52
scImpute 0.87
JOINT 1.00
Dim 1
Dim 2
d
100 DEG
Adjusted Rand Index
Jaccard Index
Adjusted Mutual Info
Dropout Probability
e
100 DEG
Gene Correlation
Cell Correlation
Dropout Probability
Observed
Saver
scImpute
JOINT
f
log Likelihood
K
g
AIC
K

Fig. S7

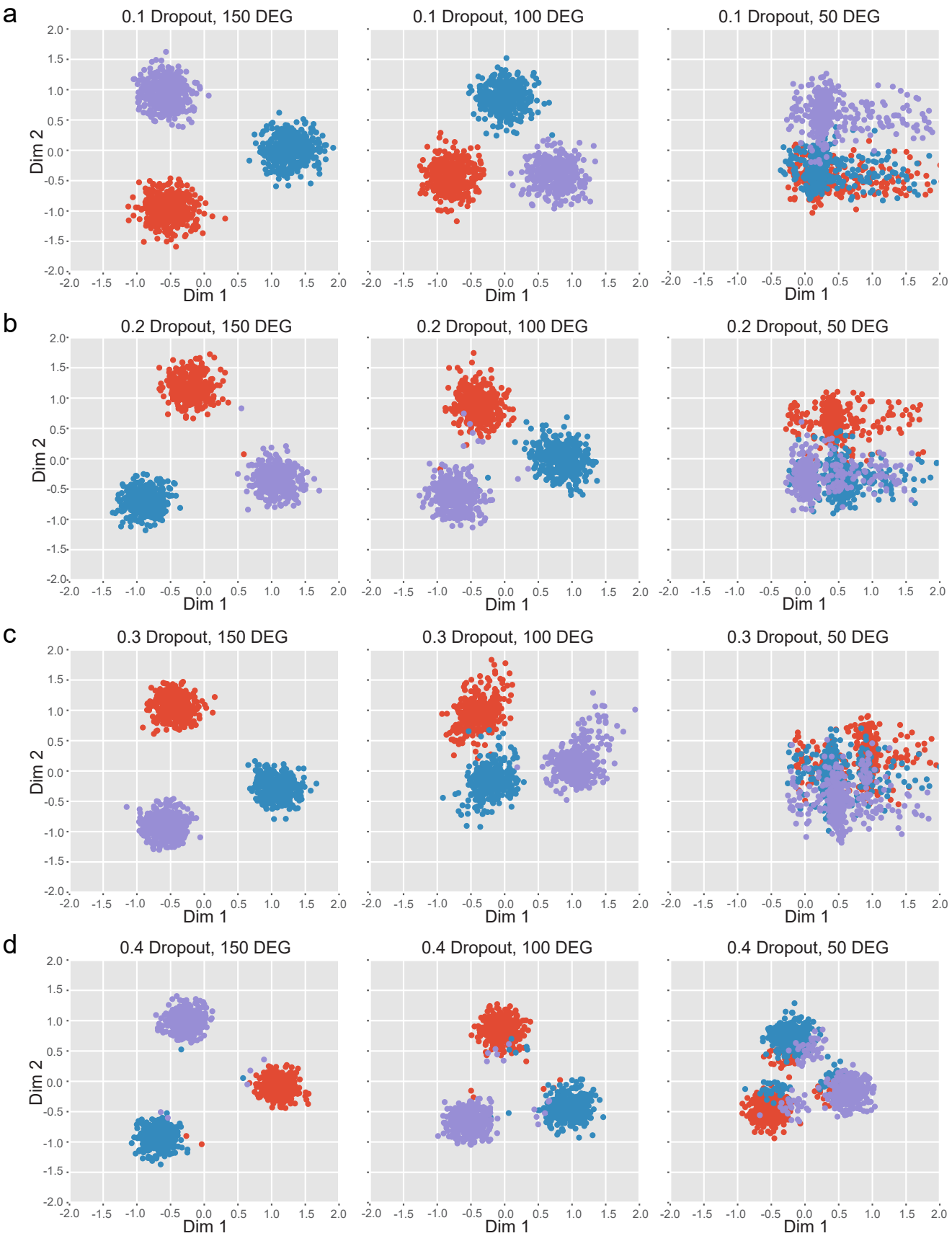
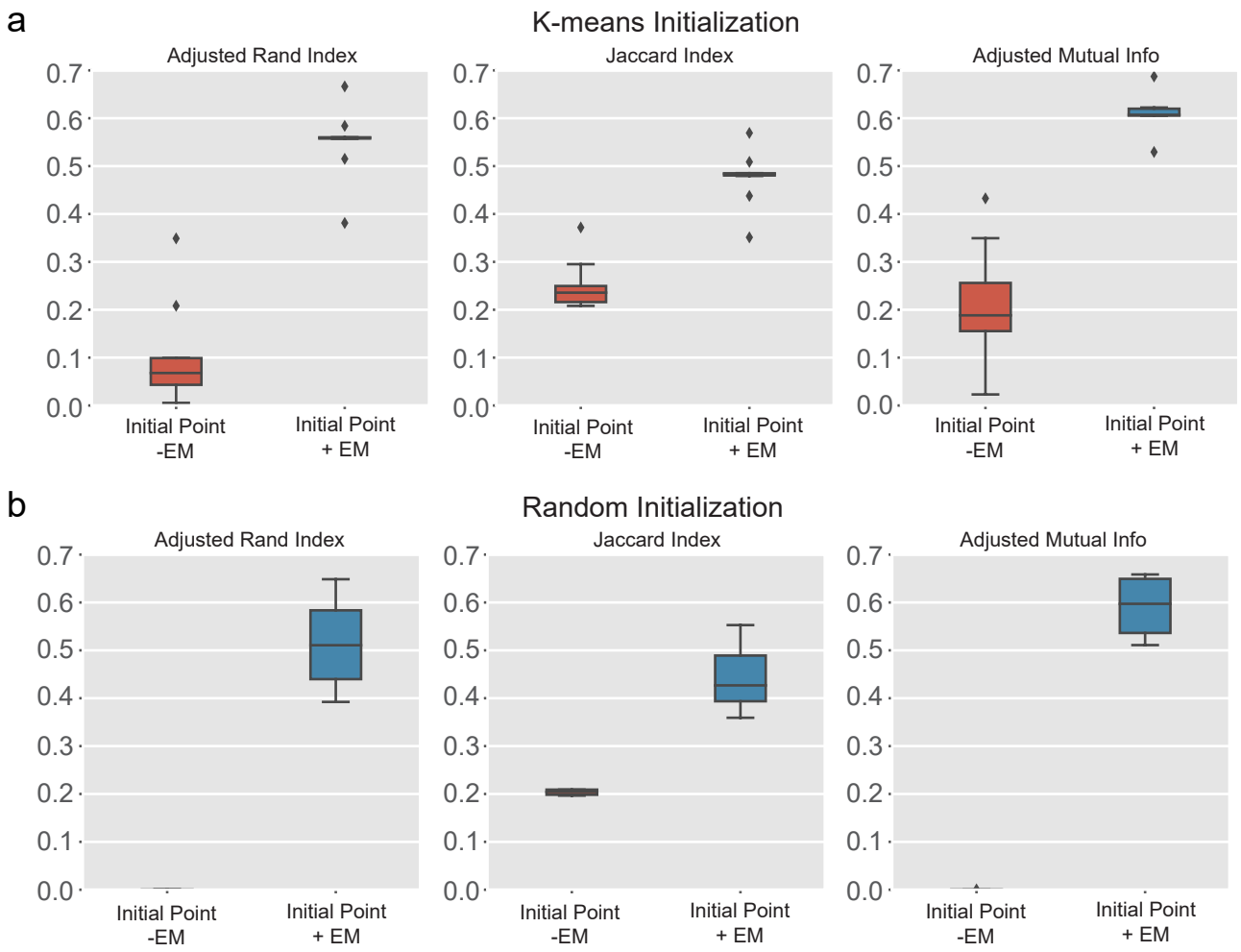

a
0.1 Dropout, 150 DEG
0.1 Dropout, 100 DEG
0.1 Dropout, 50 DEG
b
0.2 Dropout, 150 DEG
0.2 Dropout, 100 DEG
0.2 Dropout, 50 DEG
c
0.3 Dropout, 150 DEG
0.3 Dropout, 100 DEG
0.3 Dropout, 50 DEG
d
0.4 Dropout, 150 DEG
0.4 Dropout, 100 DEG
0.4 Dropout, 50 DEG
Dim 1
Dim 2

Fig. S8


a
K-means Initialization
Adjusted Rand Index
Jaccard Index
Adjusted Mutual Info
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Initial Point -EM
Initial Point + EM
b
Random Initialization
Adjusted Rand Index
Jaccard Index
Adjusted Mutual Info
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Initial Point -EM
Initial Point + EM

Supplementary Table 1

Performance Scores	Original (K-means)	Observed (K-means Non-log)	Observed (K-means log)	Saver	JOINT
Adjusted Rand Index	0.90	0.54	0.01	0.54	0.92
Jaccard Index	0.91	0.63	0.43	0.63	0.93
Adjusted Mutual Info	0.85	0.52	0.00	0.52	0.85

Supplementary Table 2

Gene = 2,000	TensorFlow GPU	TensorFlow CPU	NumPy CPU
Computing Time (s)	0.167	5.950	589.950