Clustering Performance of Complete Single-cell RNA-seq Data

As announced in the paper, Table 1 contains the classification results for the individual transformations of the complete scRNA-seq data in comparison to SC3 and *adaSC*3:

Table 1: ARI of consensus clustering using a specific combination of methods PCA: Principal Component Analysis, LE: Laplacian Eigenmaps, DM: Diffusion Maps; IM: Isomaps and MDS: Multidimensional Scaling;
*: best clustering performance among all transformations;
bold: better performance comparing SC3 and *adaSC*3.

Data Set	PCA + LE (SC3)	LE + DM (adaSC3)	PCA	LE	DM	IM	LLE	MDS
Biase et al. [1]	**0.95**	**0.95**	0.95	0.95	0.95	0.95	1.00*	1.00*
Deng et al. [2]	0.67	**0.76**	0.67	0.95*	0.34	0.44	0.43	0.50
Goolam et al. [3]	0.69	0.68	0.69	0.68	0.94*	0.54	0.69	0.54
Treutlein et al. [5]	0.66	**0.77***	0.58	0.63	0.52	0.34	0.36	0.32
Yan et al. [6]	0.65	**0.75***	0.65	0.75*	0.64	0.60	0.91	0.65

Resampling on Single-cell RNA-seq Data

For classification we used the scRNA-seq data, described in Section 2 of the manuscript, and consider a resampling scheme to analyze how stable the results of the individual approaches are. For this purpose, each single cell was omitted once from each scRNA-seq data set, such that the classification was performed N times on $N-1$ data points. The sample that is drawn without replacement was compared to the corresponding underlying cell types of the sampled $N-1$ single cells. By these resampling experiments, we evaluated each iteration and underline the good performance of *adaSC*3 with the scRNA-seq data of Biase et al. [1], Deng et al. [2], Goolam et al. [3] and Treutlein et al. [5] in Figure 1 and of Yan et al. [6] in Figure 2.

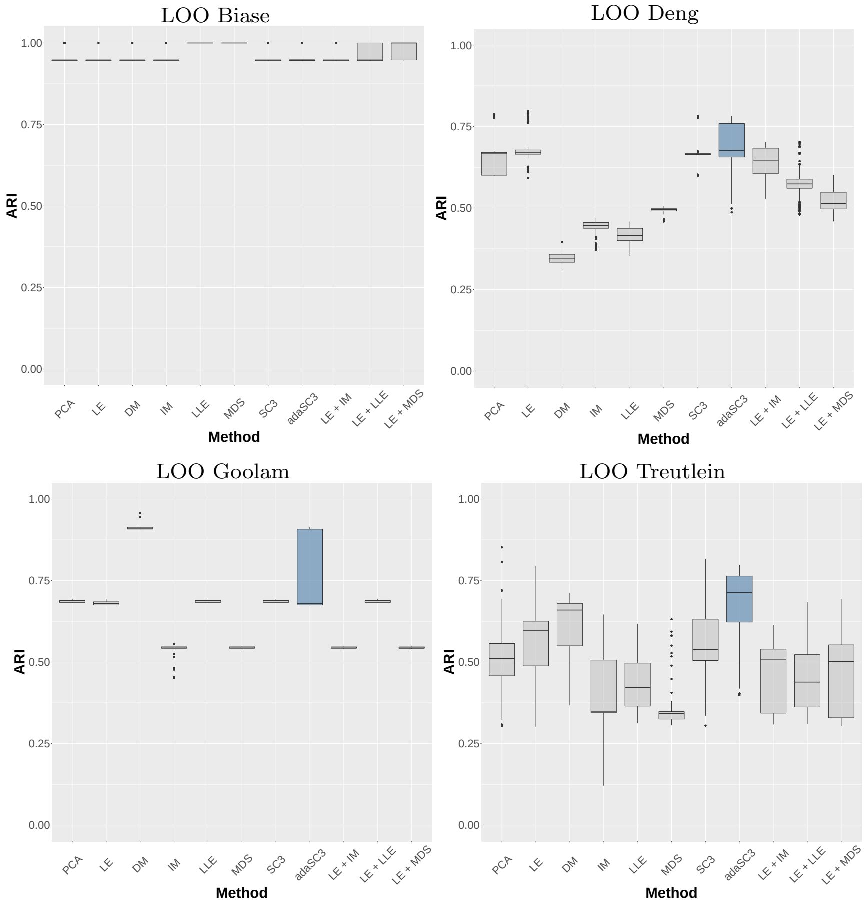


Fig. 1: Resampling results of the scRNA-seq data sets for *adaSC*3 and its competitors

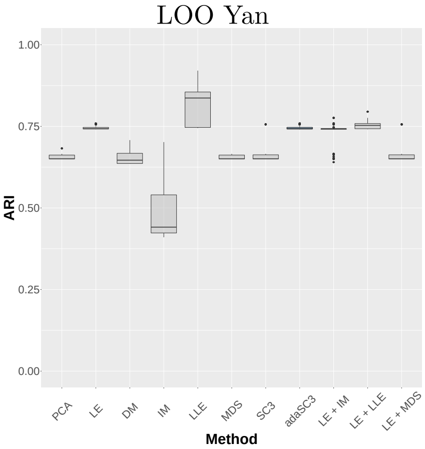


Fig. 2: Resampling results of the scRNA-seq data set Yan et al. [6] for $adaSC3$ and its competitors.

Construction of Simulation Data

To determine realistic simulation data, we mimic the data set of Kolodziejczyk et al. [4] that contains embryo stem cells of a mouse. This data set was selected for the construction of the simulation data because it contains a high number of single cells ($N = 704$) and a high number of genes ($G = 38.616$). Another advantage of this data is that it contains two quite balanced subgroups with 295 and 250 single cells representing the state "2i" and "serum", which are used for the construction of the simulation groups. The remaining 159 single cells of "a2i" are not considered for the construction of the simulation data.

We estimated the parameters of the zero-inflated Negative Binomial (ZINB) distribution of each gene for the respective cell population. From these estimates we determined ranges for the parameters of our simulation data with the 35% to 80% quantiles of the estimated parameters of the labelled group $2i$ for simulation group 1 and the 15% to 60% quantiles of the estimated parameters of the labeled group *serum* for simulation group 2. The lower quantiles of simulation group 2 should represent the healthy population with lower mean and lower dispersion of gene expression compared to simulation group 1. This leads to the following ranges of parameters μ and ϕ of the ZINB distribution for simulation group 1 and 2 shown in Table 2:

Table 2: Intervals of the estimated parameters of a ZINB distribution based on the cell types "2i" and "serum" used by steps of $^{(I)}$: 0.1; $^{(II)}$: 0.001; $^{(III)}$:0.0001.

Parameter Range	Constructed Ranges based on 2i	Constructed Ranges based on serum
$\boldsymbol{\mathcal{M}}^{\mathbf{(1)}}$	$\boldsymbol{\mathcal{M}}^{(1)} := [45, 293]^{(I)}$	$\boldsymbol{\mathcal{M}}^{(2)} := [12, 112]^{(I)}$
$\boldsymbol{\Phi}$	$\boldsymbol{\Phi}^{(1)} := [0.24, 0.94]^{(II)}$	$\boldsymbol{\Phi}^{(2)} := [0.12, 0.47]^{(II)}$
$\boldsymbol{\Pi}$	$\boldsymbol{\Pi} := [0.001, 0.01]^{(III)}$	$\boldsymbol{\Pi} := [0.001, 0.01]^{(III)}$

Results of Simulation data

Simulation Scenario (a) - Variation in Expectation Parameter μ

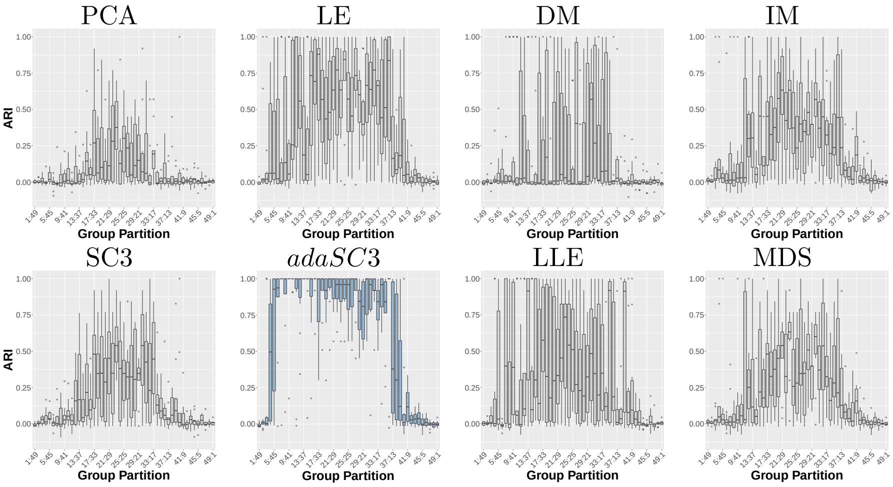


Fig. 3: Clustering accuracy (ARI) among all partitions of $N = 50$ with regard to the individual transformation techniques for simulation scenario (a) in comparison to SC and $adaSC3$.

Simulation Scenario (b) - Variation in Size Parameter ϕ

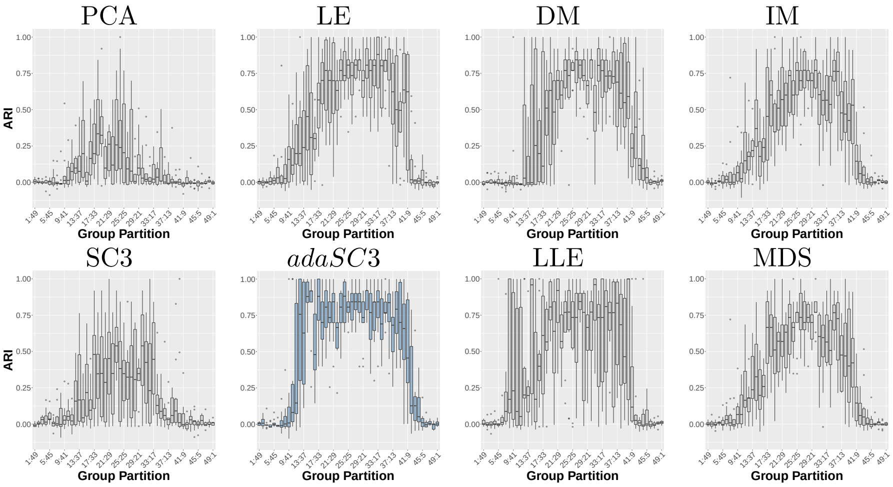


Fig. 4: Clustering accuracy (ARI) among all partitions of $N = 50$ with regard to the individual transformation techniques for simulation scenario (b) in comparison to SC and $adaSC3$.

Simulation Scenario (c) - Variation in both Parameter μ and Parameter ϕ

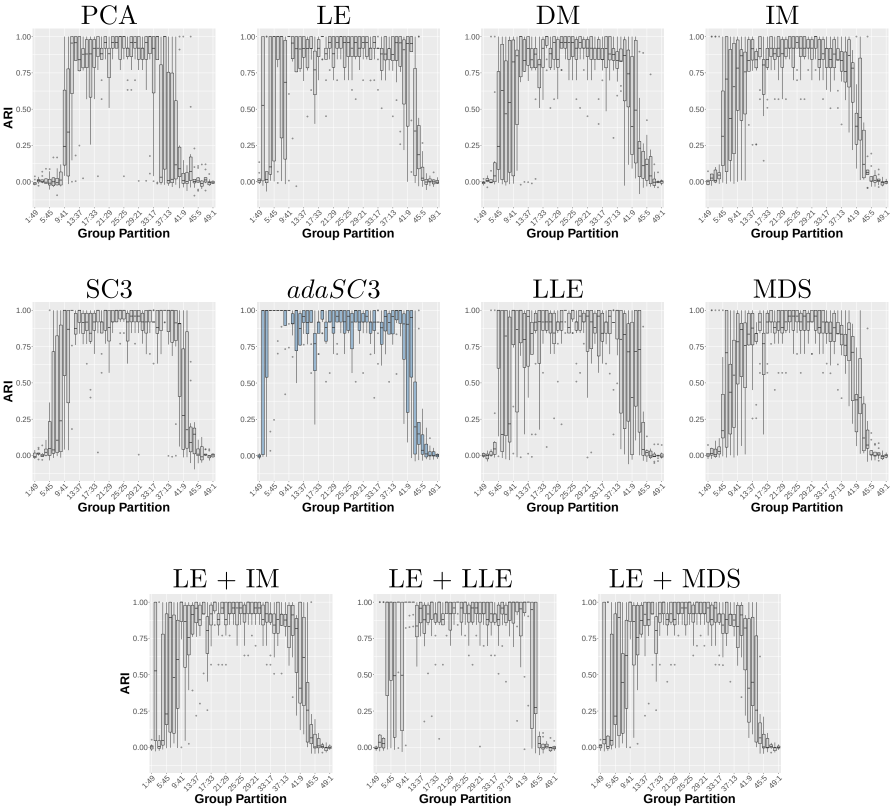
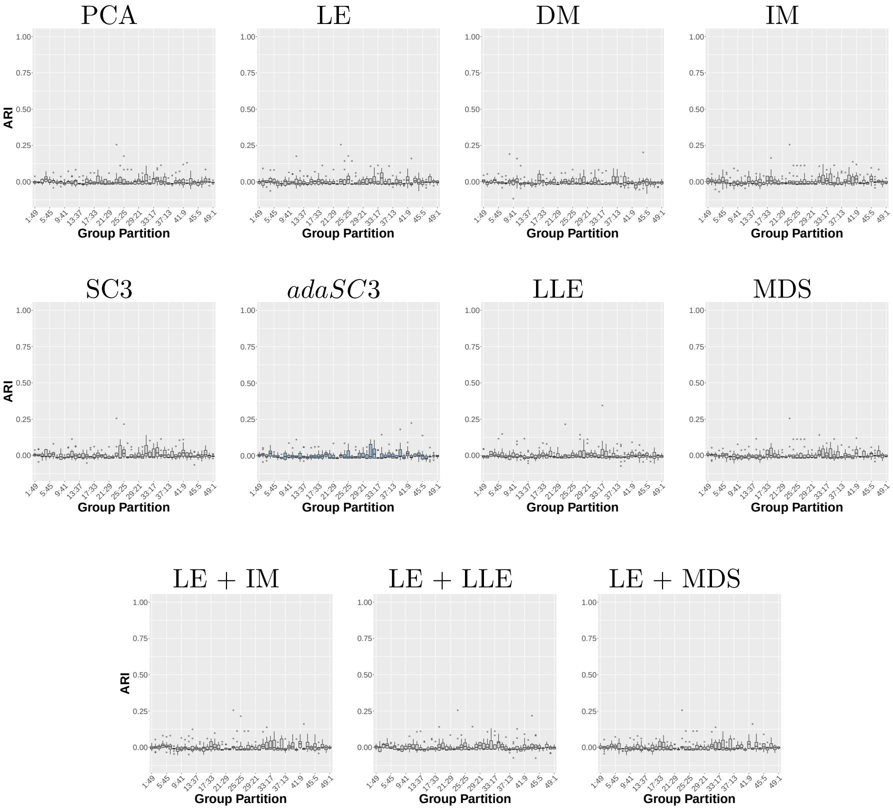


Fig. 5: Clustering accuracy (ARI) among all partitions of N= 50 with regard to the individual transformation techniques as well as in combination with LE for simulation scenario (c).

Simulation Scenario (d) - Variation in no Parameter μ and ϕ



PCA LE DM IM

SC3 $adaSC3$ LLE MDS

LE + IM LE + LLE LE + MDS

Fig. 6: Clustering accuracy (ARI) among all partitions of N= 50 with regard to the individual transformation techniques as well as in combination with LE for simulation scenario (d).

References

1. Biase FH, Cao X, Zhong S (2014) Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. Genome Research 24.11:1787-1796.
2. Deng Q, Ramsköld D, Reinius B, Sandberg R (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343.6167:193-196.
3. Goolam M, Scialdone A, Graham SJ, Macaulay IC, Jedrusik A, Hupalowska A et al. (2016) Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. Cell 165.1:61-74.
4. Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Bühler M, Pentao L, Marioni JC, Teichmann SA (2015) Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell 17.4:471-485.
5. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509.7500:371-375.
6. Yan L, Yang M, Guo H, Yang L, Wu J, Li R , Liu P, Lian Y, Zheng X, Yan, J. et al. (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nature structural and molecular biology 20.9:1131.