Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-020-00428-1

REGULAR ARTICLE

Adapted single-cell consensus clustering (adaSC3)

Cornelia Fuetterer[1] · **Thomas Augustin**[1] · **Christiane Fuchs**[2,3,4]

Received: 3 July 2019 / Revised: 11 August 2020 / Accepted: 8 November 2020

Abstract

The analysis of single-cell RNA sequencing data is of great importance in health research. It challenges data scientists, but has enormous potential in the context of personalized medicine. The clustering of single cells aims to detect different subgroups of cell populations within a patient in a data-driven manner. Some comparison studies denote single-cell consensus clustering (SC3), proposed by Kiselev et al. (Nat Methods 14(5):483–486, 2017), as the best method for classifying single-cell RNA sequencing data. SC3 includes Laplacian eigenmaps and a principal component analysis (PCA). Our proposal of unsupervised *adapted single-cell consensus clustering (adaSC3)* suggests to replace the linear PCA by diffusion maps, a non-linear method that takes the transition of single cells into account. We investigate the performance of $adaSC3$ in terms of accuracy on the data sets of the original source of SC3 as well as in a simulation study. A comparison of $adaSC3$ with SC3 as well as with related algorithms based on further alternative dimension reduction techniques shows a quite convincing behavior of $adaSC3$.

Keywords Diffusion maps · Non-linear embedding · Single-cell consensus clustering · Simulation data · Single-cell RNA sequencing data

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11634-020-00428-1.

✉ Cornelia Fuetterer
Cornelia.Fuetterer@stat.uni-muenchen.de

Thomas Augustin
Thomas.Augustin@stat.uni-muenchen.de

Christiane Fuchs
christiane.fuchs@helmholtz-muenchen.de

1 Department of Statistics, Ludwig-Maximilians-University Munich, 80539 Munich, Germany

2 Faculty of Business Administration and Economics, Bielefeld University, 33615 Bielefeld, Germany

3 Institute of Computational Biology, Helmholtz Zentrum Munich, 85764 Neuherberg, Germany

4 Department of Mathematics, Technical University of Munich, 85747 Garching, Germany

Published online: 15 December 2020

Mathematics Subject Classification 62H30 · 68U20

1 Introduction

Personalized medicine based on genomic data promises the precise and individualized treatment of diseases using information from a patient's genome (Cho et al. 2012). There is tremendous research interest in this field, especially with regard to cancer. Hereby it is of interest to determine the different stages of cancer as well as the understanding of the complex development of organs for instance, by analyzing the single cells that are obtained from the single-cell RNA sequencing. Data-driven approaches have led to projects such as The Human Cell Atlas (2020), which aims to establish an interpretable structure for the different cell types of single cells and serves as an orientation for the study of diseases. The mission of the Human Cell Atlas is "(t)o create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease." Based on the genetic profiles of these single-cell RNA sequencing data, an unsupervised classification allows a data-driven distinction of intra- and intertumoral heterogeneities as well as the determination of different pathways during the development (Duò et al. 2018). The approach of single-cell consensus clustering (SC3) by Kiselev et al. (2017) has gained much attention, not only due to its superior performance in the comparison study of Duò et al. (2018). SC3 is also explicitly tailored to single cell data. Nevertheless, the original incorporation of the linear dimension reduction of principal component analysis may offer a potential for improvement. Following Bendall et al. (2014) and Buettner and Theis (2012), for single cells the transition from one state to another is a non-linear continuous process. Therefore, we propose to replace the PCA of the SC3 method by diffusion maps (Haghverdi et al. 2015), resulting in a new unsupervised algorithm, which we call *adapted single-cell consensus clustering (adaSC3)*. The use of diffusion maps is not only motivated by the biological behavior of single cells but is also supported empirically: First, diffusion maps allow for a natural modeling of the transition of single cells by Markov processes. Secondly, according to Haghverdi et al. (2015), when applying diffusion maps to single cell data, they also seem to perform best compared to other non-linear transformation methods such as independent component analysis, Kernel PCA, Isomap, or Hessian Local Linear Embedding.

Our paper ist structured as follows: We first give an introduction into single-cell RNA sequencing data in Sect. 2. In Sect. 3, we present the methodological background, starting in Sect. 3.1 with the proposed *adapted single-cell consensus clustering (adaSC3)*, as well as some related competing methods. In Sect. 3.2, we focus on the special suitability of diffusion maps, included in our framework of $adaSC3$. Analyzing some characteristic single-cell RNA sequencing data sets, introduced already in Sect. 2, $adaSC3$ is compared to its competing methods in Sect. 4. The performance of $adaSC3$ and its competing methods is further evaluated with partition-wise simulation data in Sect. 5. Section 6 concludes with a brief discussion and outlook.

Table 1 Characteristics of the scRNA-seq data sets, with N single cells, G genes and k categories of cell types

Data set	N	G	k
Biase et al. (2014)	49	13,322	3
Deng et al. (2014)	269	10,333	10
Goolam et al. (2016)	124	11,154	5
Treutlein et al. (2014)	80	5757	5
Yan et al. (2013)	90	10,077	7

2 Single-cell RNA sequencing data (ScRNA-Seq data)

The health and development of living organisms can heavily be impacted by the kind and activity of their genes, referred to as gene expression. With the technique of single-cell RNA sequencing (scRNA-seq) introduced by Tang et al. (2009), it is possible to measure the gene expression for single cells. In scRNA-seq, genomic profiles are measured in terms of read counts, that is the number of small sequences ("reads") that result from a cell's RNA that can be identified as belonging to a particular gene. A data set comprising N cells and G genes will hence be a $N \times G$ matrix containing non-negative integers (including zero). The scRNA-seq data of Biase et al. (2014), Deng et al. (2014), Goolam et al. (2016) and Yan et al. (2013), provided by the Hemberg Group of the Sanger Institute (2020), contain read counts of single cells of a mouse or a human with different cell states that are passed during differential transcription for targeting the analysis of cell division in a pedigree. Gene expression is stochastic, and often the reads follow different distributions for different cell types. Another topic of interest is the examination of having reached varying pathway stages. For example, the experiments of Treutlein et al. (2014) were carried out to investigate cell transition during lung development. In detail, these experiments aim to analyze the development of the distal lung epithelium of the mouse based on the different transcriptional states.

The data sets shown in Table 1 had been used for the evaluation of SC3 by Kiselev et al. (2017). Fulfilling the reproducibility and the sample size criterion for the unsupervised classification leads to the data situation[1] described in Table 1. Except for the data set of Biase et al. (2014), the distributions of cell types are quite unbalanced encompassing between 80 and 269 single cells.

[1] Since in the SC3 framework, as described in the transformation step of *adaSC3* later, components in higher dimensions are chosen randomly, it is important to focus on data sets with a small amount of single cells, in order to keep the analysis replicable. Moreover, for providing the same data situation as Kiselev et al. (2017), we had to adapt the data set of Biase et al. (2014) such that the number of single cells corresponds to the data description of the original SC3 paper.

3 Methods

In this section, we first present the proposed *adapted single-cell consensus clustering (adaSC3)*, which allows, by replacing PCA with diffusion maps, to take the varying pathways and their different transcriptional states into account. Furthermore, we consider several SC3-like approaches that later serve as competitors including different transformation techniques.

In the second part of this section, we look more closely at diffusion maps, which allow an appropriate embedding of the complex data structure of single cells in the transformation step of $adaSC3$.

3.1 Unsupervised adapted single-cell consensus clustering (*adaSC3*) and its competitors

Single-cell consensus clustering aims at classifying the gene expression of single cells in an unsupervised way such that groups are determined in a data-driven manner for detecting new subgroups or confirming manually determined cell types. The classification process subdivides the cell population with regard to the homogeneity of the genetic profile into subgroups of single cells, which represent, for example, different stages of a disease or of a development process within a patient or within a mouse. The original SC3 is implemented in the software *R* (R Core Team 2020) and can be described as a pipeline consisting of several transformation steps including an automatic dimension reduction, resulting in a clustering respecting all combinations. For the construction of the adapted single-cell consensus clustering (*adaSC3*) and its competitors, we rely on the same principle framework as SC3 but consider different transformations.[2] The concrete procedure of $adaSC3$ consists of the following steps:

1. *Preprocessing* As a result of the scRNA-seq process, one obtains the gene expression matrix E containing the read counts of N single cells and G genes. As a preprocessing step, the original matrix E is reduced by a gene filter, as proposed in the original work, that aims to exclude rare and omnipresent genes.[3] This leads to the expression matrix E' of dimension $N \times G'$.
2. *Calculation of distance matrix D* Based on the expression matrix E', the Euclidean distance matrix is constructed for each pairwise single cell combination. Furthermore, two measures of dissimilarity are applied on the log-transformed data of E' using the Pearson and the Spearman correlation, respectively. For the sake of simplicity, the obtained distance and dissimilarity matrices will each be referred to as distance matrix D.
3. *Transformation technique T* For each of the obtained distance matrices D, we apply Laplacian eigenmaps[4], introduced by Belkin and Niyogi (2003) as proposed in the original SC3. In addition, we apply diffusion maps, described in more

[2] AdaSC3 and its competitors are also implemented in R and are available from the first author upon request.

[3] Genes with an expression value of > 2 in less than 6% of the cells as well as genes with a positive expression value in more than 94% of the single cell population are excluded.

[4] That are implemented as a spectral embedding in the Python software library `Scikit-learn` (Pedregosa et al. 2011) as the best size of the neighborhood for the aimed embedding.

detail below, instead of the originally proposed principal component analysis. To each of the six combinations, consisting of three different distance matrices D and two transformation techniques, an eigenvalue decomposition is applied. This leads in total to $3 \times 2 = 6$ different eigenvalue decompositions, resulting each in respectively $N - 1$ eigenvectors $\psi_1, \ldots, \psi_{N-1}$ with their ordered eigenvalues $1 > \lambda_1 \geq \cdots \geq \lambda_{N-1}$.

4. *Consensus clustering* In accordance to SC3, we adopt the automatic selection of the number of eigenvectors to be considered. For each eigenvalue decomposition, a k-means clustering, with k being deterministic, representing the number of categories of the underlying cell types as proposed in the original paper of Kiselev et al. (2017), is conducted. The automatic selection of eigenvectors of the described scenario starts incorporating the first eigenvector until the rounded integer of the 4% quantile of the set $\{1, \ldots, N\}$. The subsequent clusterings include each one further eigenvector until the 7% quantile is reached for including the maximal range of eigenvectors for the last clustering of each combination. The result of each k-means clustering m is summarized in a consensus matrix $\mathcal{C}$, indicating the relative frequency of how often a pair of single cells is grouped together over all n_m clusterings. Based on the obtained consensus matrix $\mathcal{C}$, a final complete-linkage clustering is performed. It aims to achieve higher performance and a more robust result for the classification of single cells, leading to the final grouping of k subgroups. The quality of clustering is evaluated ex-post by the Adjusted Rand Index (ARI) as proposed in the original work of Kiselev et al. (2017).

Apart from the original SC3 that includes a PCA and Laplacian eigenmaps as transformation techniques, we construct further competing algorithms following the same principle as of $adaSC3$. Instead of diffusion maps and Laplacian eigenmaps, we propose additional algorithms leading to two different types of constructions, differing in the number of incorporated transformations. The first construction only uses one single transformation technique T in Step 3 of $adaSC3$. We therefore analyze the influence of the non-linear manifolds of isomaps (IM), locally linear embedding (LLE) as well as the multidimensional scaling (MDS), in addition to the transformation of the principal component analysis (PCA), Laplacian eigenmaps (LE), and diffusion maps (DM), each on their own.[5] The second type of construction consists of the combination of each mentioned transformation T with Laplacian eigenmaps. This leads to the algorithms named by the incorporated transformations, resulting in LE + IM, LE + LLE and LE + MDS, in addition to the original SC3 (PCA + LE) and $adaSC3$ (DM+LE).

3.2 Diffusion maps

In the following, we describe the motivation of embedding the complex structure of single cells during transition into an appropriate global non-linear manifold, using

[5] For the construction of IM and LLE, Kayo (2006) proposes to use for IM and LLE the same estimate for the optimal neighborhood size, implemented by the R function `calc_k` of the R package `lle` (Diedrich et al. 2012). Isomaps are then constructed using the R package `vegan` (Oksanen et al. 2019); the locally linear embedding further relies on the R package `lle` (Diedrich et al. 2012) and MDS is based on the R package `stats` (R Core Team 2020).

diffusion maps. As stated in the introduction of Angerer et al. (2016), diffusion maps allow the reconstruction of the different states that are connected via different transitions. One possible transition is the mutation of one single cell into another. For the following construction of diffusion maps, we only consider the transition of one single cell into another within one step. Another decisive fact is the robustness of diffusion maps to noise. Furthermore, with the normalization, described in the following, diffusion maps are able to detect lowly represented cell types. Coifman and Lafon (2006) provide the general framework of diffusion maps that can be adapted to single cells following (Angerer et al. 2016) that is presented in the next steps. For the construction of diffusion maps, consider two states $x, y \in \Omega$, with Ω as the appropriate state space. x and y represent single cells; their gene expressions, measured by count data, lead to the pairwise distance $D(x, y)$.

1. For each choice of the distance measure D, each point (single cell) is considered as a node of a symmetric graph with weight function K_D

$$K_D(x, y) = \exp\left(-\frac{D(x, y)}{2\alpha^2}\right),$$

   indicating the affinity of a pair of single cells with scale parameter α, reflecting the best size of the included neighborhood.[6]
2. In the following, we construct the core of a transition kernel of a Markov chain

$$P_D(x, y) = \frac{K_D(x, y)}{Z(x)}, \quad \text{with } Z(x) = \sum_{y \in \Omega} K_D(x, y)\,.$$

3. With a density interpretation of the upper term, the following density normalized transition probability matrix

$$\tilde{P}_D(x, y) = \frac{1}{\tilde{Z}(x)} \frac{K_D(x, y)}{Z(x)Z(y)}, \quad \text{with } \tilde{Z}(x) = \sum_{y \in \Omega \setminus x} \frac{K_D(x, y)}{Z(x)Z(y)}$$

   can be obtained. As the research question consists of mapping the differentiation behavior of single cells, we are only interested in the transition between single cells. Thus, the diagonal of $\tilde{P}_D(x, y)$ is set to zero, and the normalization is adapted appropriately, summing up only the gene expression of differing pairs of single cells with $y \neq x$.
4. Based on the normalized matrix $\tilde{P}_D$, indicating the transition of one state to another by an ergodic Markovian diffusion process, the aimed transformation is obtained.

4 Results

In this section, we evaluate the clustering performance of *adaSC3* and its competitors. The accuracy of combining each of the mentioned transformations in combination with Laplacian eigenmaps is illustrated in Table 2.

[6] The estimation of α relies on the methods implemented in the R package `destiny` (Angerer et al. 2016).

Table 2 ARI of all algorithms including two transformation techniques with
*: overall best clustering performance of the combination Laplacian eigenmaps (LE) with isomaps (IM), locally linear embedding (LLE), and multidimensional scaling (MDS), as well as SC3 and $adaSC3$; bold: best performance comparing SC3 and $adaSC3$

Data Set	LE+ IM	LE + LLE	LE + MDS	SC3	$adaSC3$
Biase et al. (2014)	0.95	1.00*	1.00*	**0.95**	**0.95**
Deng et al. (2014)	0.68	0.70	0.56	0.67	**0.76***
Goolam et al. (2016)	0.54	0.69*	0.54	**0.69***	0.68
Treutlein et al. (2014)	0.53	0.42	0.56	0.66	**0.77***
Yan et al. (2013)	0.75	0.75	0.65	0.65	**0.75***

$AdaSC3$ leads in three out of five cases (Deng et al., Treutlein et al. and Yan et al.) to better clustering results, compared to the original SC3, and it is identical in the case of the data set of Biase et al.. Concerning the competing algorithms, the combinations of LE with IM and MDS tend to deliver worse results compared to $adaSC3$. However, LE + LLE achieves two times the best performing classification but fails extremely in the case of the Treutlein et al. data set. The slightly worse performance of $adaSC$ compared to SC3 concerning the complete Goolam et al. data set of Table 2 should not be over-interpreted as the resampling results based on leaving out each single cell once, we reach considerably higher performance compared to SC3. Furthermore, we can state that $adaSC3$ delivers the highest overall performance concerning both the resampling study as well as using only one transformation technique as illustrated in the Supplementary Material. We therefore consider our proposal as generally the best approach among its competitors, of SC3 and its related approaches, based on the benchmarking data sets. This result is especially surprising as the scRNA-seq data sets were originally used for determining the proposed default settings of SC3, such as e.g. the automatic choice for the lower dimension.

5 Simulation data

The classification accuracy of the simulation data is investigated in the same way as the scRNA-seq data. We are interested in the consensus clustering accuracy of two simulation groups, which are constructed with different ranges of distribution parameters describing the read counts. With shifted parameter ranges, one can consider the simulation groups as representing a healthy and a diseased population. According to the literature, the use of a zero-inflated negative binomial (ZINB) distribution is recommended as the most adequate approximate distribution for modeling the read counts of single cells. It allows larger variability of read counts compared to the former used Poisson distribution (see e.g. Wagner et al. 2013). Based on the constructed simulation data following a (generalized version of a) ZINB distribution for the expression of each gene, we will investigate the influence of various parameters describing each gene for all possible group partitions for a fixed total number N of single cells.

5.1 Construction of simulation data

The ZINB distribution (Kleiber and Zeileis 2016) is a mixture between a negative binomial probability mass function and a point mass at zero. For generating ZINB distributed simulation data we use the R package `emdbook` (Bolker B, Bolker Maintainer Ben and Imports, MASS 2020), which is based on a generalization of the negative binomial (NB) distribution with parameters μ and ϕ for the non-zero inflated part.[7] The parameter μ is a continuous positive real value, describing the mean. The dispersion parameter ϕ represents the shape parameter of the gamma distribution underlying the generalization of the NB. The fraction of zero-inflation is taken into account by the parameter π.

In the following scenarios, we investigate the influence of different parameters of the distribution family. In order to mimic a realistic situation, the scRNA-seq data of Kolodziejczyk et al. (2015) is taken for estimating the parameters of a ZINB distribution[8] and allow the construction of ranges for each parameter looking at the shifted quantiles of the estimates of the parameters μ and ϕ. This leads to the parameter ranges $\boldsymbol{\mathcal{M}}^{(1)}$ and $\boldsymbol{\Phi}^{(1)}$ for cell population 1 and $\boldsymbol{\mathcal{M}}^{(2)}$ and $\boldsymbol{\Phi}^{(2)}$ for cell population 2. The parameter range $\boldsymbol{\Pi}$ for π is set to be the same for both populations.[9] Thus, the simulated read counts of each gene follow a ZINB(μ_1, ϕ_1, π_1) distribution for simulation group 1 and a ZINB(μ_2, ϕ_2, π_2) for simulation group 2, according to the following scenarios:

- Simulation scenario (a) for different ranges of μ:
  $\mu_1 \in \boldsymbol{\mathcal{M}}^{(1)}$ and $\mu_2 \in \boldsymbol{\mathcal{M}}^{(2)}$, $\phi_1 = \phi_2 \in \boldsymbol{\Phi}^{(2)}$, $\pi_1 = \pi_2 \in \boldsymbol{\Pi}$
- Simulation scenario (b) for different ranges of ϕ:
  $\mu_1 = \mu_2 \in \boldsymbol{\mathcal{M}}^{(2)}$, $\phi_1 \in \boldsymbol{\Phi}^{(1)}$ and $\phi_2 \in \boldsymbol{\Phi}^{(2)}$, $\pi_1 = \pi_2 \in \boldsymbol{\Pi}$
- Simulation scenario (c) for different ranges of μ and ϕ:
  $\mu_1 \in \boldsymbol{\mathcal{M}}^{(1)}$ and $\mu_2 \in \boldsymbol{\mathcal{M}}^{(2)}$, $\phi_1 \in \boldsymbol{\Phi}^{(1)}$ and $\phi_2 \in \boldsymbol{\Phi}^{(2)}$, $\pi_1 = \pi_2 \in \boldsymbol{\Pi}$
- Simulation scenario (d) for the same range of μ and ϕ:
  $\mu_1 = \mu_2 \in \boldsymbol{\mathcal{M}}^{(2)}$, $\phi_1 = \phi_2 \in \boldsymbol{\Phi}^{(2)}$, $\pi_1 = \pi_2 \in \boldsymbol{\Pi}$

For each of the simulation scenarios (a) to (d), we sample N_1 times out of ZINB$(\mu_1$ $\phi_1, \pi_1)$ and N_2 times out of ZINB(μ_2, ϕ_2, π_2) such that, for comparison purposes, the gene-specific parameters remain the same when generating all possible partitions of $N_1 : N_2$ (with $N_1 + N_2 = N$), starting with $1 : (N-1)$ until $(N-1) : 1$, with $N = 50$ for the respective scenario. In order to obtain simulation data with the dimension $N \times G$ for each partition, we repeat this procedure 200 times. Thus, read counts of $G = 200$ genes are generated with the new parameter values drawn uniformly from the respective intervals.

[7] Details explaining the generalization of the negative binomial distribution function based on a mixture of Poisson distributions with gamma distributed Poisson rates can be found e.g. in Fuetterer et al. (2019). They investigate the influence of different heterogeneity degrees of count data using simulation data as well as up- and downwardly distorted measurements via the ZINB distribution describing the case of measurements tending to lower read counts and upper read counts.

[8] The manual construction of two cell populations rely on the differentially cultured murine embryonic stem cell populations "2i" and "serum" for each of the 38.616 genes.

[9] The constructed parameter ranges are part of the Supplementary Material.

5.2 Clustering performance based on simulation data

The following plots show how the group partition (x-axis) of a combination of two transformation methods influences the clustering accuracy, measured by the Adjusted Rand Index (ARI) (y-axis). In the ideal case, the accuracy is 1 for each of the partitions, which would indicate that the classification perfectly corresponds to the underlying group allocation. This criterion is best met for $adaSC3$, not only in the case of using only one transformation technique (see Supplementary Material), but also in combination of those with Laplacian eigenmaps (LE) for simulation scenarios (a) to (c). Simulation (d) serves as a reference where no difference in the gene-specific parameters was simulated and no accurate grouping should be detected. Each partition of each scenario is repeated 10 times and the accuracy of the respective clustering results is visualized by boxplots. Results of simulation scenario (c) and (d) can be found in the Supplementary Material.

5.2.1 Simulation scenario (a): variation in expectation parameter μ

In the case of differing parameter μ represented in scenario (a), $adaSC3$ seems to perform best among the combined methods (see Fig. 1) as well as compared to each method on its own. It can also be seen that the inter quantile range of boxplots have the tendency to be shorter for $adaSC3$ and reach higher ARI values compared to its competitors. Therefore, we conclude in general that our approach generates

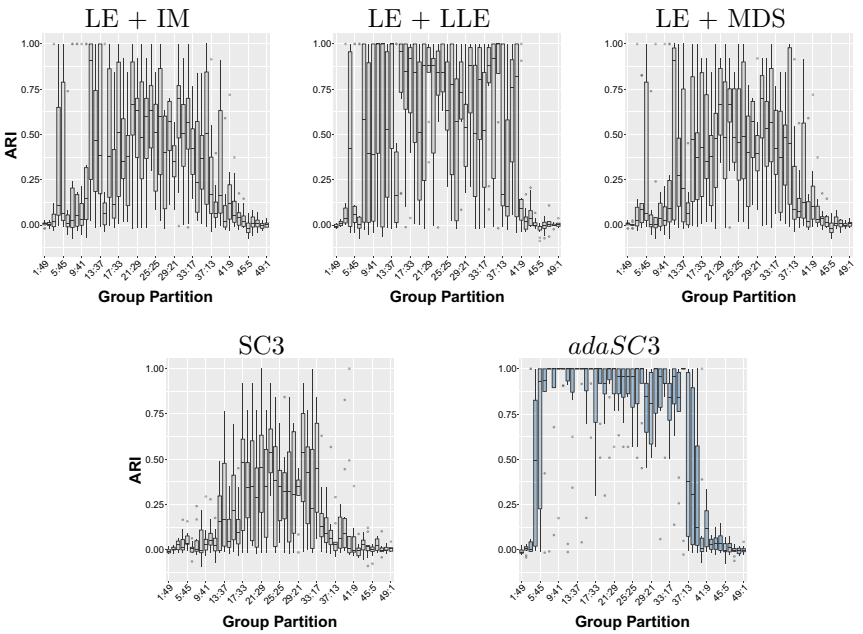

Fig. 1 ARI among all partitions of $N = 50$ with regard to the combination of different transformation techniques for simulation scenario (a)

a more efficient clustering allocation. Furthermore, it should also be noted that the performance of the method $adaSC3$ seems to be better when more single cells are part of simulation group 1 compared to simulation group 2. This might allow the interpretation that the detection of true positives achieves higher accuracy compared to the detection of false negatives, given the diseased population has on average higher gene expression.

5.2.2 Simulation scenario (b): variation in size parameter ϕ

For partition-wise created simulation data differing in the parameter ϕ one can state, referring to Fig. 2, that apart from $adaSC3$ the combination of LE + LLE performs quite well, too. However, this method needs more partitions before it starts detecting the difference in the simulation groups and fails drastically earlier compared to $adaSC3$. For approximately balanced data, LE + LLE often leads to worse results. The tendency that the clustering performance depends on the partitions can be confirmed over all methods for simulation scenarios (a) to (c), in which $adaSC3$ is affected the less.

With regard to scenario (c), the simulated differences of both parameters μ and ϕ lead to a quite accurate classification for most methods with an overall superiority of $adaSC3$, representing the scenario being the closest to the reality. For the simulation design with no difference in the simulation groups, the allocation of single cells is as expected and represents random allocation.

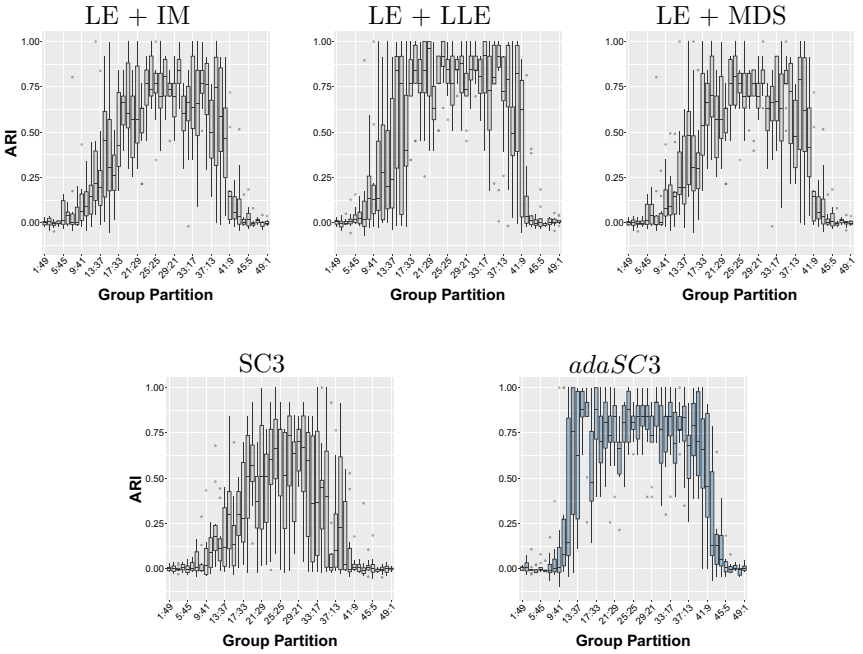

Fig. 2 ARI among all partitions of $N = 50$ with regard to the combination of different transformation techniques for simulation scenario (b)

6 Conclusions

The approach of *adapted single-cell consensus clustering (adaSC3)* is tailored to the clustering of single cells. It reflects the biological structure of single cells by including diffusion maps with the aim to respect the transition process of the underlying data. Indeed, the inclusion of diffusion maps instead of the originally proposed PCA led to a better clustering performance. We consider $adaSC3$ to be the best method compared to all investigated competitors, both in the analyzed scRNA-seq data as well as in the simulation study. This motivates further research that takes into account the biological basis of the data before constructing or combining some methods, as this could be rewarded both in terms of interpretation and accuracy, as shown in this paper.

Based on the discovery that balanced data seems to be detected correctly with higher quality, the distribution of classified classes could be taken into account for an unsupervised evaluation. Furthermore, studies of additional scRNA-seq data and further simulations are needed to reinforce the results of this paper. This is especially due to the fact that $adaSC3$ was evaluated on the same scRNA-seq data used for the development of the original SC3 method. This makes the overall superiority of $adaSC3$ over SC3 even more surprising, while on the other hand some bias of these data sets favoring SC3-like methods cannot be excluded.

Acknowledgements We are very grateful to the two anonymous referees and the editors for their stimulating and constructive comments. We also want to thank Florian Pfisterer for discussions about diffusion maps. Furthermore, the public data sharing of the scRNA-sequencing data by the Hemberg Group of the Sanger Institute is gratefully acknowledged. In addition, the first author is very thankful to the LMUMentoring program, connecting young researchers with experienced researchers and providing financial support.

Funding Open Access funding enabled and organized by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Angerer P, Haghverdi L, Büttner M, Theis FJ, Marr C, Buettner F (2016) Destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics 32(8):1241–1243

Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15(6):1373–1396

Bendall SC, Davis KL, el Amir AD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Pe'er D (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell 157(3):714–725

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 Res 24(11):1787–1796
Bolker B, Bolker Maintainer Ben and Imports, MASS (2020) Package 'emdbook' R package version 1.3.11
Buettner F, Theis FJ (2012) A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. Bioinformatics 28(18):i626–i632
Cho SH, Jongsu J, Seung IK (2012) Personalized medicine in breast cancer: a systematic review. J Breast Cancer 15(3):265–272
Coifman RR, Lafon S (2006) Diffusion maps. Appl Comput Harmonic Anal 21(1):5–30
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
Diedrich H, Abel M., Diedrich Maintainer Holger (2012) Package "lle" R package version 1.1
Duò A, Robinson MD, Soneson C (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F10000Research 7.2
Fuetterer C, Schollmeyer G, Augustin T (2019) Constructing simulation data with dependency structure for unreliable single-cell RNA-sequencing data using copulas. ISIPTA '19. Proc Mach Learn Res 103:216–224
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
Haghverdi L, Buettner F, Theis FJ (2015) Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics 31(18):2989–2998
Hemberg Group at Sanger Institute (2020) scRNA-Seq Datasets. https://hemberg-lab.github.io/scRNA.seq.datasets/. Accessed 11 Aug 2020
Kayo O (2006) Locally linear embedding algorithm–Extensions and applications. Technical Report, Faculty of Technology, Department of Electrical and Information Engineering, University of Oulo
Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Rei W, Barahona M, Green AR et al (2017) SC3: consensus clustering of single-cell RNA-Seq data. Nat Methods 14(5):483–486
Kleiber C, Zeileis A (2016) Visualizing count data regressions using rootograms. Am Stat 70(3):296–303
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
Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H (2019) vegan: Community Ecology Package. Package 'vegan' R package version 2.5-6
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg R, Vanderplas J, Passos A, Cournapeau D, Perrot Brucher M, Duchesnay ME (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/. Accessed 11 Aug 2020
Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA (2009) MRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6:377–382
The Human Cell Atlas. https://www.humancellatlas.org. Accessed 11 Aug 2020
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
Wagner GP, Kin K, Lynch VJ (2013) A model based criterion for gene expression calls using RNA-seq data. Theory Biosci 132(3):159–164
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. Nat Struct Mol Biol 20(9):1131