



Article

RDAClone: Deciphering Tumor Heterozygosity through Single-Cell Genomics Data Analysis with Robust Deep Autoencoder

Jie Xia ^{1,2,†}, Lequn Wang ^{2,3,†} , Guijun Zhang ^{1,*}, Chunman Zuo ^{4,*} and Luonan Chen ^{2,3,5} 

¹ College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China; xiajie@sibs.ac.cn

² Center for Excellence in Molecular Cell Science, State Key Laboratory of Cell Biology, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China; wanglequn2019@sibcb.ac.cn (L.W.); Inchen@sibcb.ac.cn (L.C.)

³ Key Laboratory of Systems Health Science of Zhejiang Province, Hangzhou Institute for Advanced Study, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Hangzhou 310024, China

⁴ Institute of Artificial Intelligence, Donghua University, Shanghai 201600, China

⁵ Guangdong Institute of Intelligence Science and Technology, Hengqin, Zhuhai, Guangdong 519031, China

* Correspondence: zgj@zjut.edu.cn (G.Z.); zuochunman@sibcb.ac.cn (C.Z.)

† These authors contributed equally.

Abstract: Rapid advances in single-cell genomics sequencing (SCGS) have allowed researchers to characterize tumor heterozygosity with unprecedented resolution and reveal the phylogenetic relationships between tumor cells or clones. However, high sequencing error rates of current SCGS data, i.e., false positives, false negatives, and missing bases, severely limit its application. Here, we present a deep learning framework, RDAClone, to recover genotype matrices from noisy data with an extended robust deep autoencoder, cluster cells into subclones by the Louvain-Jaccard method, and further infer evolutionary relationships between subclones by the minimum spanning tree. Studies on both simulated and real datasets demonstrate its robustness and superiority in data denoising, cell clustering, and evolutionary tree reconstruction, particularly for large datasets.

Keywords: single-cell genomics sequencing; deep learning; Louvain-Jaccard method; cell clustering; phylogenetic relationship



Citation: Xia, J.; Wang, L.; Zhang, G.; Zuo, C.; Chen, L. RDAClone: Deciphering Tumor Heterozygosity through Single-Cell Genomics Data Analysis with Robust Deep Autoencoder. *Genes* **2021**, *12*, 1847. <https://doi.org/10.3390/genes12121847>

Academic Editor: Anelia D. Horvath

Received: 22 October 2021

Accepted: 22 November 2021

Published: 23 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Understanding the evolutionary mechanisms related to cancer progression and characterizing the intra-heterogeneity are promising routes for predicting and further controlling cancer progression, metastasis, and treatment responses [1–5]. Tumor tissue is composed of different subpopulations of cells with different genotypes, each called a subclone [6], and the evolutionary relationship of these subclones can be constructed based on the order of the mutated genes in these cells [4,7]. Many methods have been developed to deconvolute the bulk DNA sequencing data to identify subclones, and then construct the evolutionary tree [8–13]. However, bulk DNA sequencing data obtained from a mixture of millions of cells masks the properties of intra-tumor heterozygosity [14].

The rapid development of single-cell genomics sequencing (SCGS) offers an unprecedented opportunity to profile the evolutionary relationship between subclones in cancer tissue [15–17]. However, the application of current SCGS data has been severely limited by high-level experimental noise from single cell isolation, whole genome amplification, genome interrogation, allelic dropout events inducing false negative (FN) and false positive (FP) mutations, missing bases resulting from the insufficient sequencing coverage, and doublets from the mistaken selection of more than one cell [14,15]. Many studies have reported that the FP rate ranges from 3×10^{-5} to 7×10^{-5} and the FN rate ranges from

0.1 to 0.43. In addition, the reported missing rate (MR) is often more than 50% and up to 58% [18,19].

Recently, many computational methods have been developed to analyze SCGS data. For instance, SCITE [20] was designed to identify the evolution of the subclones of a tumor by modeling the sequencing errors based on a stochastic search algorithm. In addition, oncoNEM [21], a probabilistic model, was developed to infer intra-tumor evolutionary lineage trees from SCGS data. SCG [22], a hierarchical Bayesian model, has been proposed to simultaneously cluster cells into subclusters and infer corresponding genotypes, but it cannot be used to infer the evolutionary relations between these subclones. However, these three methods perform analysis tasks under the infinite site assumption, without considering recurrent mutations. SiFit [23] was proposed to infer tumor phylogenies from noisy SCGS data under a finite-sites assumption. BEAM [24] was developed to improve the quality in the SCGS data by using evolutionary information in the SCGS data in a molecular phylogenetic framework. However, these methods are difficult to scale to large datasets, especially the probabilistic models with exponential time complexity. To solve this problem, RobustClone [25] was designed to analyze SCGS data with the following three steps: recovering genotype matrices from noisy SCGS data by robust principal component analysis (RPCA) [26] or extended RPCA [27], clustering cells into subclones based on recovered genotype matrix by Louvain-Jaccard clustering [28,29], and finally, reconstructing a subclone evolutionary tree by finding the minimum spanning tree of these subclones. However, the biological data is inherently nonlinear and too complex to be represented by the linear model used by RobustClone, and the time complexity for RPCA is relatively high.

Deep neural networks have been widely applied to learn complex nonlinear features in many bioinformatics domains [30–33], including single-cell data analysis [34–38]. In particular, the deep autoencoder has been shown to be a computationally effective approach to extract the low-dimensional nonlinear features to accurately represent the data. Inspired by RPCA, the robust deep autoencoder (RDA) [39] was proposed to simultaneously represent the nonlinear features, which are robust to the noise and outliers [26], by minimizing the rank of low-rank matrices and the number of nonzero entries in the sparse matrix. Its optimization can be implemented by the alternating direction method of multipliers (ADMM) [40]. Recently, RDA has been widely applied in image denoising and anomaly detection [39,41,42]. These advantages of RDA show great potential to solve the difficulties in the current analysis of SCGS data.

Here, we present a computational framework RDAClone to analyze the noised SCGS data based on our extended RDA. Specifically, RDAClone can (a) denoise and impute the experimental SCGS data; (b) identify subclones from cancer cells; and (c) construct the subclone evolution trajectory. Our method can handle very large datasets even with high error/noise rates. To demonstrate its effectiveness, we applied the RDAClone method and other widely used methods to both simulated and real datasets, which demonstrated the superior performance of RDAClone compared to current state-of-the-art methods.

2. Methods

2.1. RDAClone Model

RDAClone was proposed to analyze SCGS data through the following steps: (1) decomposing the observed matrix into the sum of a low-rank matrix (i.e., a recovered genotype matrix) and a sparse matrix through our extended RDA; (2) clustering cells into subclones based on the recovered genotype matrix using the Louvain-Jaccard clustering method [28,29]; and (3) reconstructing a subclone evolutionary tree using the minimum spanning tree of the genotypes among these subclones [25]. The whole computational framework is shown in Figure 1a. Each method is described in detail below.

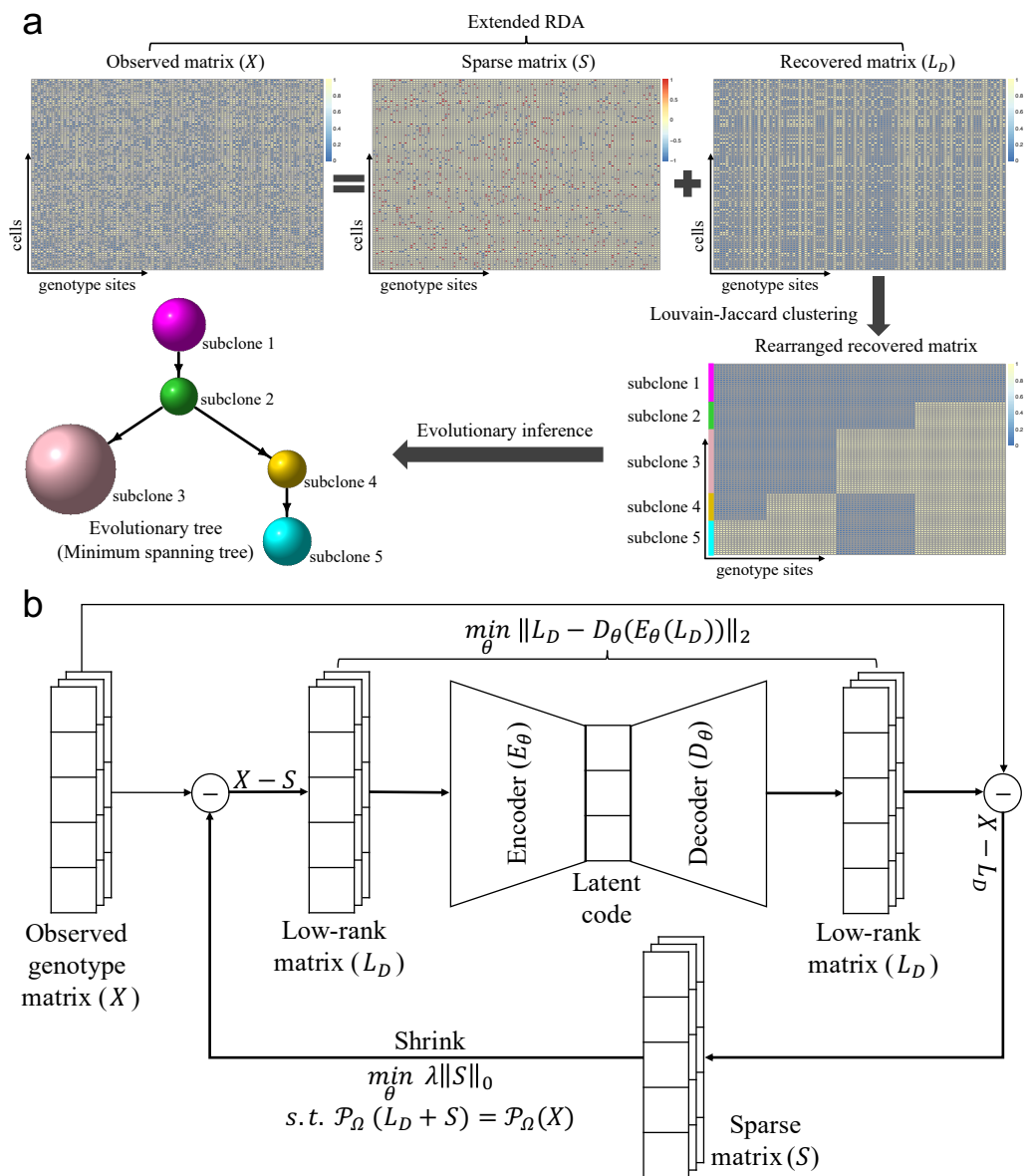


Figure 1. Overview of RDAClone framework and the computational graph for extended RDA model. (a) Given the SCGS data with row for cells and column for genotypes as input, RDAClone firstly decomposes the observed matrix (data) into a sparse matrix and a low rank recovered matrix by the extended RDA; clusters cells into subclones based on the recovered matrix by Louvain-Jaccard clustering method; and reconstructs subclone evolutionary tree by the minimum spanning tree of the subclone genotypes. (b) Recovering the genotype matrix by our extended RDA model as follows: first, initialize the low-rank matrix L_D and sparse matrix S as zero matrices. Then, given an observed matrix X , the low-rank matrix L_D calculated by $X - S$ as the input of an encoder to extract the low-dimensional features, reconstructed back to the original dimensional space L_D by a decoder, and the sparse matrix S calculated by $X - L_D$ is further optimized by a shrink operator to make S sparser. After the convergence of the computation, the recovered matrix L_D generated by the extended RDA is kept for the subsequent analysis.

2.1.1. RDA

Deep autoencoder learns the nonlinear low-dimensional features for the input data with an encoder and recovers the input data with a decoder, and the low-dimensional features can capture the information that accurately represents the input data. To process the experimental noisy data, RDA [41] was proposed to not only maintain a deep autoen-

coder's ability to discover high-quality nonlinear features but also to eliminate the outliers by splitting the input data (X) into a low-rank matrix (L_D) and sparsity matrix (S). This method has already been demonstrated to have powerful capability in image denoising and anomaly detection [39,41,42]. Hence, RDA is well adapted to process SCGS data due to its advantages over noisy data. The optimization objective of RDA can be summarized as two parts: minimizing the reconstruction loss of the deep autoencoder, and the number of nonzero entries in sparse matrix S , which can be formulated as

$$\begin{aligned} \min_{\theta} & \|L_D - D_{\theta}(E_{\theta}(L_D))\|_2 + \lambda \|S\|_0 \\ \text{s.t.} & L_D + S = X \end{aligned} \quad (1)$$

where E_{θ} and D_{θ} denote the encoder and decoder, respectively. S captures the noise information, θ is the learnable parameter in deep autoencoder, and λ is a hyperparameter to balance the two parts of the optimization objective, which can be fine-tuned manually. A larger λ makes the sparse matrix S sparser, while in contrast, a smaller λ will allow more noise to be isolated into S .

The above optimization is not computationally tractable. Following the idea in RPCA [27], the ℓ_0 norm of S can be approximately replaced by the ℓ_1 norm of S , and then the optimization problem takes the following form:

$$\begin{aligned} \min_{\theta} & \|L_D - D_{\theta}(E_{\theta}(L_D))\|_2 + \lambda \|S\|_1 \\ \text{s.t.} & L_D + S = X \end{aligned} \quad (2)$$

2.1.2. Extended RDA

To handle the amount of missing entries in the SCGS data, we proposed a novel algorithm named extended RDA (a nonlinear expansion of the extended RPCA [27]). We first define a projection as follows:

$$\mathcal{P}_{\Omega}(X) = \begin{cases} X_{ij}, & \text{if } (i, j) \in \Omega \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

The central idea of the extended RDA is to perform the decomposition only on the nonzero entries of the input SCGS data matrix. The missing entries are imputed to achieve a low rank in the recovered matrix, so that the optimization problem of the extended RDA can be formulated as follows:

$$\begin{aligned} \min_{\theta} & \|L_D - D_{\theta}(E_{\theta}(L_D))\|_2 + \lambda \|S\|_1 \\ \text{s.t.} & \mathcal{P}_{\Omega}(L_D + S) = \mathcal{P}_{\Omega}(X) \end{aligned} \quad (4)$$

However, the above optimization problem is computationally infeasible. The constraint on nonzero entries is equivalent to penalize the nonzero entries in the sparse matrix S , and, thus, the extended RDA problem can be transformed into the following form:

$$\begin{aligned} \min_{\theta} & \|L_D - D_{\theta}(E_{\theta}(L_D))\|_2 + \lambda \|\mathcal{P}_{\Omega}(S)\|_1 \\ \text{s.t.} & L_D + S = X \end{aligned} \quad (5)$$

After applying the ADMM method [40], the above optimization problem can be solved through multiple iterations, which optimize the two objective parts separately. The detailed optimization process is shown in Algorithm 1 and Figure 1b.

Algorithm 1 The extended RDA model

```

1: Input: Observed genotype matrix  $X \in R^{n \times m}$ ; Index set of observed entries  $\Omega$ 
2: Initialize  $L_D \in R^{n \times m}$ ,  $S \in R^{n \times m}$  to be zero matrices,  $LS \leftarrow X$ , and an autoencoder  $D(E(\cdot))$  with randomly initialized parameters.
3: while True do:
4:    $L_D = X - S$ 
5:    $\theta = \underset{\theta}{\operatorname{argmin}} \|L_D - D_\theta(E_\theta(L_D))\|_2$ 
6:    $L_D = D_\theta(E_\theta(L_D))$ 
7:    $S = X - L_D$ 
8:    $\mathcal{P}_\Omega(S) = \operatorname{prox}(\mathcal{P}_\Omega(S))$ 
9:    $c_1 = \|X - L_D - S\|_2 / \|X\|_2$  # check the convergence condition
10:   $c_2 = \|LS - L_D - S\|_2 / \|X\|_2$  # check the convergence condition
11:  if  $c_1 < \varepsilon$  or  $c_2 < \varepsilon$  break
12:   $LS = L_D + S$  # update  $LS$  for convergence checking in the next iteration
13: return  $L_D$  and  $S$ 

```

2.1.3. Identification of Subclones by Louvain-Jaccard Clustering

The Louvain–Jaccard clustering method [28,29], a hierarchical clustering algorithm, is used to detect communities from large networks and has been widely used for the clustering analysis of single-cell data [25,43]. The specific workflow for the Louvain–Jaccard method is as follows: constructing a k-nearest neighbor network for the cells based on their similarity by the Euclidean distance of their genotype sites; and then clustering cells into different subclones. The Louvain–Jaccard method is conducted very rapidly, which is scalable to large datasets. In addition, the Louvain–Jaccard method has been demonstrated to be robust to the choice of cluster K [25]; therefore, we do not need to specify the number of subclones in advance.

2.1.4. Construct Subclone Evolutionary Tree by Minimum Spanning Tree Method

Following the RDAClone framework [25], we adapted the following two steps to construct a subclone evolutionary tree: (1) the cells belonging to the same subclone should be homogeneous; hence, their genotypes should be almost identical. Specifically, for scSNV data, the genotype with the highest frequency among these cells was regarded as the consensus one, which was treated as subclone genotype; and (2) the Euclidean distance between any pair of subclones based on their subclone genotypes was calculated to find the minimum spanning tree.

2.2. Datasets and Preprocessing

Simulated datasets: we simulated five groups of data, with each containing five datasets, using the function “simulateData2” in the oncoNEM [21] R package. This is a two-step generation process that consists of constructing a clonal tree structure and simulating genotype observations based on the simulated subclone tree. Specifically, the data of 3000 cells for each of the five groups were simulated by changing one of the following five parameters while keeping the other four parameters fixed: the number of genotype sites (#GS), false-positive rate (FPR), false-negative rate (FNR), missing rate (MR), and the number of simulated subclones (#SC). The detailed parameters for each dataset are shown in Table 1.

For run-time comparison analysis, we simulated 10 datasets by changing the number of cells ranging from 500 to 5000 by 500, the number of genotypes ranging from 200 to 2000 by 200, and the number of subclones ranging from 10 to 100 by 10, with the following fixed parameters: 15% FPR, 15% FNR, and 30% MR.

Real Data: Two real scSNV datasets were used in this study: a high-grade serous ovarian cancer (HGSOC) dataset [22,44] and an essential thrombocythemia (ET) dataset from a sample of JAK2-negative myeloproliferative neoplasm [18]. To check whether our model was robust with the high missing rate data, we randomly changed non-missing

entries of both datasets to missing, with the proportion ranging from 20.7% to 60.7% by 10% for the HGSO dataset and 60.7% to 72.7% by 3% for the ET dataset. In this way, five evaluation datasets were generated for each real dataset.

Table 1. Simulation parameters for datasets of five groups.

	#CELL	#GS	FPR	FNR	MR	#SC
Group 1	3000	From 1200 to 2000 by 200	0.15	0.15	0.3	50
Group 2	3000	1000	From 0.2 to 0.4 by 0.05	0.15	0.3	50
Group 3	3000	1000	0.15	From 0.2 to 0.4 by 0.05	0.3	50
Group 4	3000	1000	0.15	0.15	From 0.4 to 0.8 by 0.1	50
Group 5	3000	1000	0.15	0.15	0.3	From 60 to 100 by 10

2.3. Evaluation Metrics

To compare RDA with other SCGS data analysis methods, RobustClone, BEAM, and SCG, we defined the following four metrics to evaluate the accuracy of recovered genotype data and clustering under the simulated datasets. Specifically, (1) the sum of the false positive rate and false negative rate (FPNR) of the recovered matrix obtained by each method divided by the FPNR of the input matrix was calculated by formula (6); (2) the proportion of missing entries correctly imputed by each method was calculated by formula (7); (3) the proportion of incorrectly recovered genotype matrix entries by each method was calculated by formula (8); and (4) the clustering accuracy based on the ARI [45] by comparing the predicted clustering result based on the recovered matrix and the subclone ground truth was obtained. Formulas (6) through (8) are defined as follows:

$$FPNR = FPR + FNR = \frac{FP}{FP + TN} + \frac{FN}{FN + TP} \quad (6)$$

$$accuracy_{\bar{\Omega}} = \frac{\|\mathcal{P}_{\bar{\Omega}}(\mathbb{I}(L_D = X_{true}))\|_0}{m \times n \times MR} \quad (7)$$

$$error = \frac{\|\mathbb{I}(L_D \neq X_{true})\|_0}{m \times n} \quad (8)$$

where FPR and FNR denote the false positive rate and false negative rate, respectively; FP , FN , TP , and TN represent false positive, false negative, true positive and true negative, respectively; \mathbb{I} is an indicator function, where when the condition is true \mathbb{I} returns 1, and is otherwise 0; L_D is the recovered genotype matrix; X_{true} is the corresponding simulated ground truth; MR is the abbreviation of the missing rate; and m and n denote the number of genotype sites and cells, respectively.

In addition, we defined the following evaluation metrics to determine whether these SCGS models could handle real SCGS data with a high missing rate using the 10 generated datasets based on real data. Specifically, (1) the proportion of the additional missing entries correctly recovered by each method was calculated by formula (7); and (2) the ARI between the predicted clustering result based on the genotype matrix recovered by each method and the clustering result from the previous publication were used.

3. Results

3.1. Model Evaluation and Comparison on the Simulated Datasets

We compared RDAClone with several state-of-the-art methods, including RobustClone, SCG (doublet-naïve variant), and BEAM, for the recovery of the genotype matrix and its clustering based on the defined four metrics using the simulated datasets. Because BEAM has difficulty handling large-scale datasets, we subsampled the input matrix (5 repeats) for the BEAM method. As summarized in Figure 2a, (1) for the five simulated datasets with different #GS, the accuracy of the genotype matrix recovered by RDAClone was higher than the other three methods, and the clustering accuracy of RDAClone were

higher than those of the other three methods. (2) RDAClone had a comparable performance to SCG for matrix recovery and clustering, and these two methods were better than RobustClone and BEAM under the five simulated datasets with varying FPRs. (3) RDAClone performed better than RobustClone and BEAM under the simulated datasets with changing FNR and was slightly weaker than SCG. Finally, (4) with the varying of MR or the number of subclones, the accuracy of matrix recovery and clustering for RDAClone was slightly higher than that of SCG and these two methods performed better than RobustClone and BEAM for both matrix recovery and clustering (Figure 2a). Additionally, we also evaluated the genotype matrix recovery accuracy using precision, recall and F1 score (Figure S1) and the clustering accuracy using normalized mutual information (NMI) and V measure (Figure S2). These results further supported the superiority and robustness of RDAClone. Furthermore, we performed sensitivity analysis on the resolution of Louvain-Jaccard clustering controlling the number of resulting clusters and found the clustering accuracy was less likely to be affected by resolution (Figure S3).

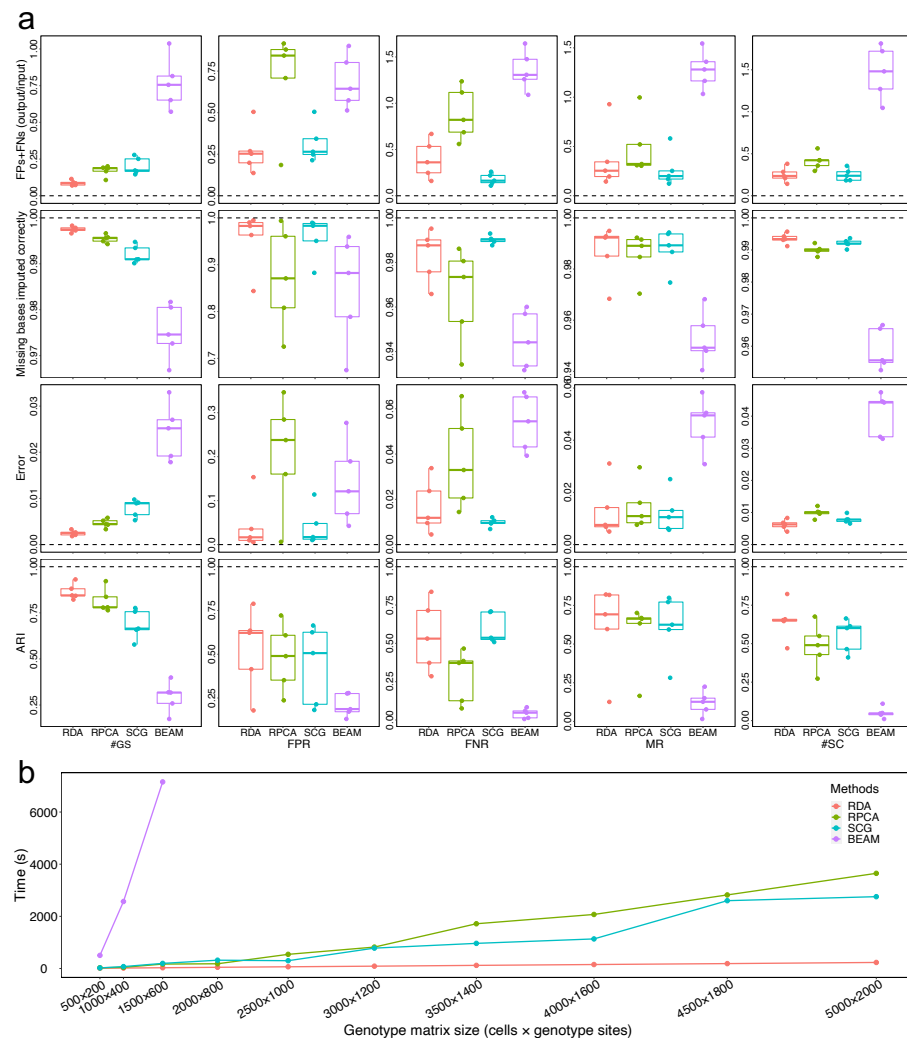


Figure 2. Genotype matrix recovery accuracy, clustering, and run-time comparison on the simulated datasets. **(a)** The matrix recovery accuracy (evaluated by FPs + FNs, missing bases imputed correctly, error), and clustering accuracy (evaluated by ARI between the known clustering and the predicted clustering based on recovered matrix) for each method. **(b)** Run-time comparison for fitting four models on the ten simulated datasets with different sizes of cells, genotypes and subclones. Clearly, the computational complexity of RDAClone is almost linear against the matrix size, which is much efficient comparing with other methods.

In addition, to assess the scalability of these methods, we simulated 10 datasets with different numbers of cells, genotypes, and subclones. Generally, RDAClone, RobustClone, and SCG could be scaled to large datasets, but RDAClone performed the fastest across all simulated datasets (Figure 2b) owing to its reliance on a fixed number of cells at each iteration of iterative stochastic optimization (Figure 2b). The computational complexity of RDAClone was almost linear against the matrix size, which was much more efficient compared with the other methods.

3.2. RDAClone Accurately Recovers Missing Entries and Identifies Subclones on a Real scSNV Dataset

We next evaluated the extent to which the genotype matrix recovered by the extended RDA could identify real subclones. We performed RDAClone on the HGSOc dataset (a real scSNV dataset) consisting of 420 cells \times 43 genotype sites with a 10.7% missing rate, with the other three methods, RobustClone, SCG, and BEAM, as a comparison [22,44]. The results showed that RDAClone could cluster 420 cells into five subclones, which contained 111, 99, 93, 82, and 35 cells, and 43 genotype sites into three blocks, which contained 18, 18, and seven genotype sites. The heatmaps of the recovered matrix and raw matrix for each method are shown in Figure 3a, and the ordering of cells and genotype sites is based on the predicted clustering result.

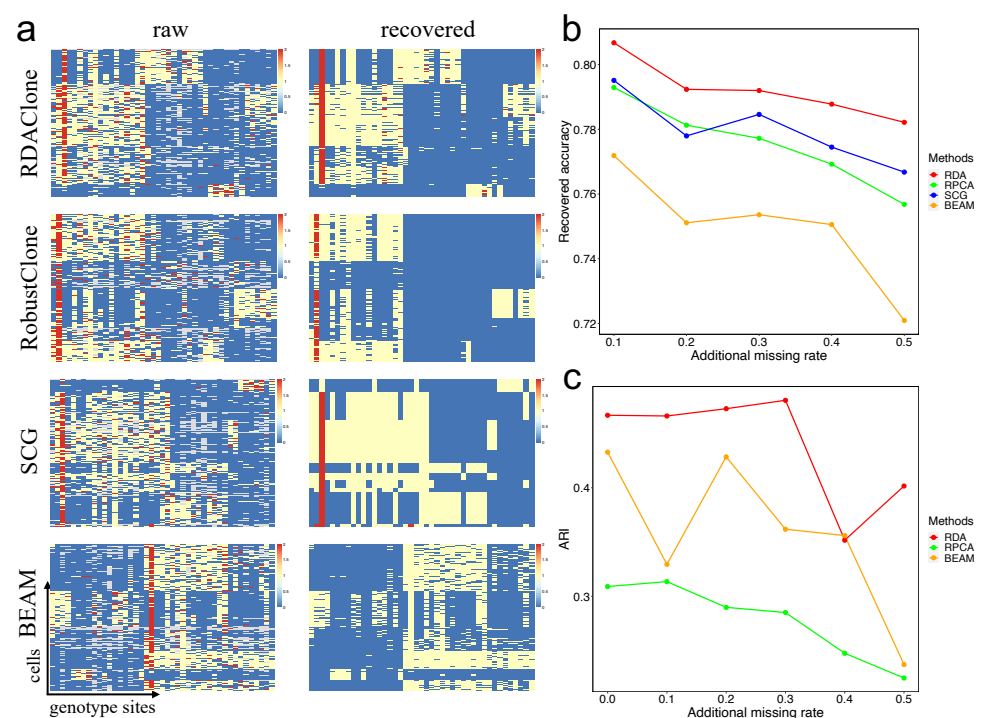


Figure 3. Genotype matrix recovery, and cell clustering on the HGSOc dataset. (a) The heatmap of raw and recovered genotypes matrix for each of methods: RDAClone, RobustClone, SCG and BEAM. The ordering of cells and genotype sites based on the clustering of recovered matrix. (b) The proportion of missing entries correctly recovered by RDAClone, RobustClone, SCG and BEAM under five sparsity data generated by additional missing rate. (c) The clustering accuracy was evaluated by ARI between the published SCG results and clustering results of recovered matrix by RDAClone, RobustClone and BEAM.

To test whether RDAClone was robust with sparse data, five datasets were generated by randomly converting non-missing entries of the HGSOc dataset to missing so that the missing data ranged from 20.7% to 60.7% in increments of 10%. By comparison, we found that RDAClone could recover missing entries with the highest accuracy and was robust relative to the other three methods under data with various levels of sparsity (Figure 3b).

Due to the lack of the real subclone structure for the HGSOc dataset as a benchmark, we used the published clustering result of the SCG model as the ground truth for the clustering assessment. ARI was applied to compare the clustering accuracy between RDAClone, RobustClone, and BEAM. We found that the clustering accuracy of RDAClone was significantly higher than RobustClone and BEAM, and RobustClone performed the worst (Figure 3c).

3.3. RDAClone Works Well on a Real scSNV Dataset with a High Missing Rate

We next demonstrated how RDAClone could recover a high missing rate genotype matrix to infer real subclone relationships. We applied RDAClone on the ET dataset (a real scSNV dataset), which consisted of 58 cells \times 712 genotype sites with a 57.7% missing rate, with the other three methods, RobustClone, SCG, and BEAM, used as a comparison [18]. The results showed that RDAClone could cluster 58 cells into three subclones, which contained 26, 19, and 13 cells, and 712 genotype sites into six blocks, which contained 168, 166, 133, 130, 71, and 44 genotype sites. The heatmaps of the recovered matrix and raw matrix for each method are shown in Figure 4a, and the ordering of cells and genotype sites are based on the predicted clustering results. In addition, the linear evolutionary relationship among three subclones could be briefly described as follows: subclone 3 was considered as a normal genotype (root of the evolutionary tree), and the mutation of some genotype sites converted some cells from subclone 3 to subclone 1, and then to subclone 2, as shown in Figure 4b. The RobustClone model had a similar linear evolutionary relationship between subclones to RDAClone, but some cells from subclones 1 and 2 were clustered into different subclones (Figure 4c).

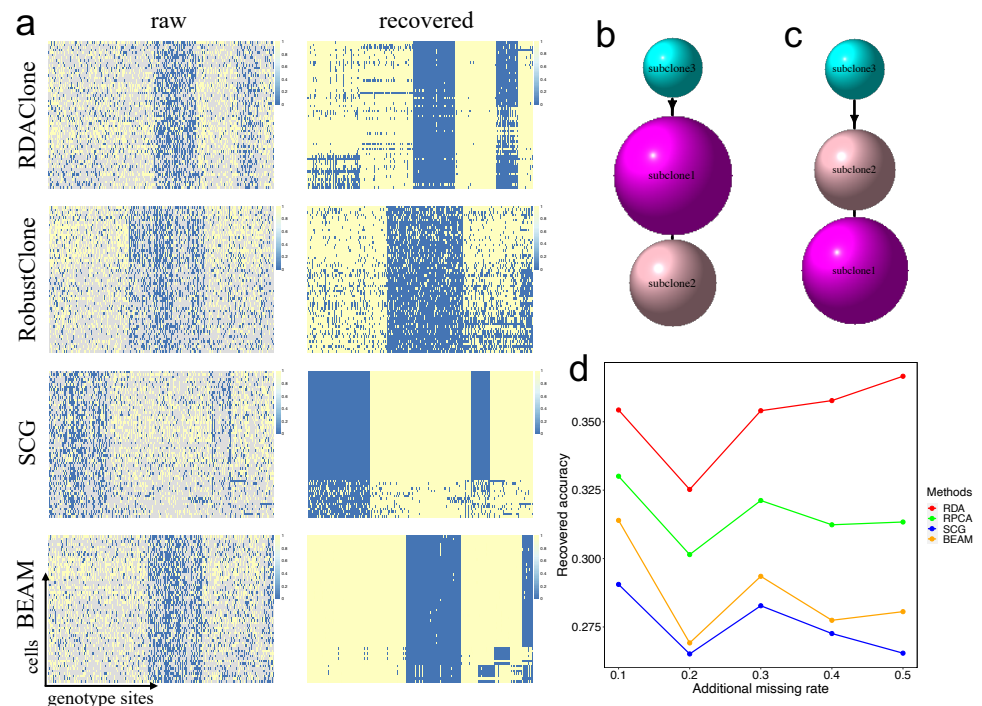


Figure 4. RDAClone works well on ET dataset with a high missing rate. (a) The heatmap of raw and recovered genotype matrix for each of methods: RDAClone, RobustClone, SCG and BEAM. The ordering of cells and genotype sites based on the clustering of recovered matrix. (b) The evolutionary tree inferred from the recovered matrix by RDAClone. (c) The evolutionary tree inferred from the recovered matrix by RobustClone. (d) The proportion of missing entries correctly recovered by RDAClone, RobustClone, SCG and BEAM under five sparsity data generated by additional missing rate.

We further evaluated the extent to which RDAClone could process the scSNV data with a higher missing rate. We randomly dropped out the non-missing entries of the ET dataset to missing so that the missing rate ranged from 60.7% to 72.7% with the increments of 3%, which generated five datasets. The results showed that RDAClone recovered missing entries with the highest accuracy and was robust relative to the other three methods under data with various levels of sparsity (Figure 4d).

4. Discussions

SCGS data provides unprecedented insights to characterize intra-tumor heterogeneity and infer tumor subclone evolutionary relationships. However, the sequencing errors resulting in FP, FN, and MB severely limit its application. In this study, we proposed a deep learning framework, RDAClone, for SCGS data analysis. RDAClone can be utilized to recover a genotype matrix from corrupted observations even with a high missing rate, cluster cells into subclones based on the recovered matrix, and then perform subclone relationship inference to construct a subclone evolutionary tree.

RDAClone is a nonlinear version of RobustClone, which has been shown to be a powerful and efficient tool for SCGS data analysis compared to most previous methods [22,24,25]. Based on our comparison, we found that the matrix recovery accuracy of RDAClone was significantly higher than that of RobustClone on the simulated and real datasets, as well as the subsequent clustering performance based on the recovered genotype matrix. In addition, RDAClone can be scaled to super large datasets of 10^7 entries with a run time of no more than 231 s, but RobustClone has a run time of up to 3649 s because RDAClone relies on a fixed number of cells at each iteration of iterative stochastic optimization. RDAClone also works well on high-missing rate scSNV data, i.e., with higher accuracy.

RDAClone has three components, i.e., extended RDA, the clustering method, and the evolution inference method. Each of these components can be adapted to other applications. Recent studies emerged to infer subclone evolution from single-cell RNA sequencing (scRNA-seq) data due to its high input amount [46–49]. The sparsity and noise are general for single-cell data (e.g., scRNA-seq and single-cell chromatin accessibility data) and, thus, our extended RDA can be also applied to other areas of single-cell data analysis. In addition, by combining with the cell specific network method [50], RDAClone can be adapted to analyze the cellular and gene networks for potential applications [51–55] in biology and medicine.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/genes12121847/s1>. Figure S1: Genotype matrix recovery accuracy comparison on the simulated datasets, Figure S2: Clustering accuracy comparison on the simulated datasets, Figure S3: Sensitivity analysis on the resolution (default to 1.0) of Louvain-Jaccard clustering.

Author Contributions: Conceptualization, L.C., G.Z. and C.Z.; methodology, J.X.; software, L.W.; validation, J.X. and L.W.; writing—original draft preparation, J.X., L.W. and C.Z.; writing—review and editing, J.X., L.W., L.C. and G.Z.; project administration, L.C.; funding acquisition, L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Basic Research Program of China (No. 2017YFA05 05500), the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB38040400), the National Natural Science Foundation of China (Nos. 31930022, 31771476, 12131020, 12026608), and JST Agency Moonshot R&D Project (No. JPMJMS2021).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nowell, P.C. The clonal evolution of tumor cell populations. *Science* **1976**, *194*, 23–28. [[CrossRef](#)]
2. Merlo, L.M.; Pepper, J.W.; Reid, B.J.; Maley, C.C. Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer* **2006**, *6*, 924–935. [[CrossRef](#)] [[PubMed](#)]
3. Pepper, J.W.; Scott Findlay, C.; Kassen, R.; Spencer, S.L.; Maley, C.C. Synthesis: Cancer research meets evolutionary biology. *Evol. Appl.* **2009**, *2*, 62–70. [[CrossRef](#)]
4. Greaves, M.; Maley, C.C. Clonal evolution in cancer. *Nature* **2012**, *481*, 306. [[CrossRef](#)] [[PubMed](#)]
5. Schwartz, R.; Schäffer, A.A. The evolution of tumour phylogenetics: Principles and practice. *Nat. Rev. Genet.* **2017**, *18*, 213. [[CrossRef](#)]
6. Lawson, D.A.; Kessenbrock, K.; Davis, R.T.; Pervolarakis, N.; Werb, Z. Tumour heterogeneity and metastasis at single-cell resolution. *Nat. Cell Biol.* **2018**, *20*, 1349–1360. [[CrossRef](#)]
7. Yates, L.R.; Campbell, P.J. Evolution of the cancer genome. *Nat. Rev. Genet.* **2012**, *13*, 795–806. [[CrossRef](#)]
8. Gerstung, M.; Beisel, C.; Rechsteiner, M.; Wild, P.; Schraml, P.; Moch, H.; Beerenwinkel, N. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **2012**, *3*, 811. [[CrossRef](#)] [[PubMed](#)]
9. Shah, S.P.; Roth, A.; Goya, R.; Oloumi, A.; Ha, G.; Zhao, Y.; Turashvili, G.; Ding, J.; Tse, K.; Haffari, G.; et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **2012**, *486*, 395–399. [[CrossRef](#)]
10. Oesper, L.; Mahmoody, A.; Raphael, B.J. THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.* **2013**, *14*, R80. [[CrossRef](#)]
11. Ha, G.; Roth, A.; Khattra, J.; Ho, J.; Yap, D.; Prentice, L.M.; Melnyk, N.; McPherson, A.; Bashashati, A.; Laks, E.; et al. TITAN: Inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **2014**, *24*, 1881–1893. [[CrossRef](#)]
12. Roth, A.; Khattra, J.; Yap, D.; Wan, A.; Laks, E.; Biele, J.; Ha, G.; Aparicio, S.; Bouchard-Côté, A.; Shah, S.P. PyClone: Statistical inference of clonal population structure in cancer. *Nat. Methods* **2014**, *11*, 396. [[CrossRef](#)]
13. Zare, H.; Wang, J.; Hu, A.; Weber, K.; Smith, J.; Nickerson, D.; Song, C.; Witten, D.; Blau, C.A.; Noble, W.S. Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.* **2014**, *10*, e1003703. [[CrossRef](#)] [[PubMed](#)]
14. Navin, N.E. Cancer genomics: One cell at a time. *Genome Biol.* **2014**, *15*, 452. [[CrossRef](#)]
15. Gawad, C.; Koh, W.; Quake, S.R. Single-cell genome sequencing: Current state of the science. *Nat. Rev. Genet.* **2016**, *17*, 175. [[CrossRef](#)]
16. Wang, Y.; Waters, J.; Leung, M.L.; Unruh, A.; Roh, W.; Shi, X.; Chen, K.; Scheet, P.; Vattathil, S.; Liang, H.; et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **2014**, *512*, 155–160. [[CrossRef](#)] [[PubMed](#)]
17. Zong, C.; Lu, S.; Chapman, A.R.; Xie, X.S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **2012**, *338*, 1622–1626. [[CrossRef](#)] [[PubMed](#)]
18. Hou, Y.; Song, L.; Zhu, P.; Zhang, B.; Tao, Y.; Xu, X.; Li, F.; Wu, K.; Liang, J.; Shao, D.; et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **2012**, *148*, 873–885. [[CrossRef](#)]
19. Gawad, C.; Koh, W.; Quake, S.R. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 17947–17952. [[CrossRef](#)]
20. Jahn, K.; Kuipers, J.; Beerenwinkel, N. Tree inference for single-cell data. *Genome Biol.* **2016**, *17*, 86. [[CrossRef](#)]
21. Ross, E.M.; Markowetz, F. OncoNEM: Inferring tumor evolution from single-cell sequencing data. *Genome Biol.* **2016**, *17*, 69. [[CrossRef](#)] [[PubMed](#)]
22. Roth, A.; McPherson, A.; Laks, E.; Biele, J.; Yap, D.; Wan, A.; Smith, M.A.; Nielsen, C.B.; McAlpine, J.N.; Aparicio, S.; et al. Clonal genotype and population structure inference from single-cell tumor sequencing. *Nat. Methods* **2016**, *13*, 573–576. [[CrossRef](#)] [[PubMed](#)]
23. Zafar, H.; Tzen, A.; Navin, N.; Chen, K.; Nakhleh, L. SiFit: Inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biol.* **2017**, *18*, 178. [[CrossRef](#)] [[PubMed](#)]
24. Miura, S.; Huuki, L.A.; Buturla, T.; Vu, T.; Gomez, K.; Kumar, S. Computational enhancement of single-cell sequences for inferring tumor evolution. *Bioinformatics* **2018**, *34*, i917–i926. [[CrossRef](#)] [[PubMed](#)]
25. Chen, Z.; Gong, F.; Wan, L.; Ma, L. RobustClone: A robust PCA method for tumor clone and evolution inference from single-cell sequencing data. *Bioinformatics* **2020**, *36*, 3299–3306. [[CrossRef](#)]
26. Candès, E.J.; Li, X.; Ma, Y.; Wright, J. Robust principal component analysis? *J. ACM (JACM)* **2011**, *58*, 1–37. [[CrossRef](#)]
27. Vidal, R.; Ma, Y.; Sastry, S. Generalized principal component analysis (GPCA). *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1945–1959. [[CrossRef](#)]
28. Blondel, V.D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [[CrossRef](#)]
29. Levine, J.H.; Simonds, E.F.; Bendall, S.C.; Davis, K.L.; Amir, E.-A.D.; Tadmor, M.D.; Litvin, O.; Fienberg, H.G.; Jager, A.; Zunder, E.R.; et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **2015**, *162*, 184–197. [[CrossRef](#)]
30. Min, S.; Lee, B.; Yoon, S. Deep learning in bioinformatics. *Brief. Bioinform.* **2017**, *18*, 851–869. [[CrossRef](#)]
31. Li, Y.; Huang, C.; Ding, L.; Li, Z.; Pan, Y.; Gao, X. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods* **2019**, *166*, 4–21. [[CrossRef](#)] [[PubMed](#)]

32. Lan, K.; Wang, D.-T.; Fong, S.; Liu, L.-S.; Wong, K.K.; Dey, N. A survey of data mining and deep learning in bioinformatics. *J. Med Syst.* **2018**, *42*, 139. [[CrossRef](#)]
33. Tang, B.; Pan, Z.; Yin, K.; Khateeb, A. Recent advances of deep learning in bioinformatics and computational biology. *Front. Genet.* **2019**, *10*, 214. [[CrossRef](#)]
34. Angermueller, C.; Lee, H.J.; Reik, W.; Stegle, O. DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **2017**, *18*, 67. [[CrossRef](#)]
35. Tian, T.; Wan, J.; Song, Q.; Wei, Z. Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nat. Mach. Intell.* **2019**, *1*, 191–198. [[CrossRef](#)]
36. Eraslan, G.; Simon, L.M.; Mircea, M.; Mueller, N.S.; Theis, F.J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **2019**, *10*, 390. [[CrossRef](#)]
37. Lopez, R.; Regier, J.; Cole, M.B.; Jordan, M.I.; Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **2018**, *15*, 1053. [[CrossRef](#)]
38. Ding, J.; Condon, A.; Shah, S.P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* **2018**, *9*, 2002. [[CrossRef](#)] [[PubMed](#)]
39. Sakurada, M.; Yairi, T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, Gold Coast, QLD, Australia, 2 December 2014; pp. 4–11.
40. Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J. *Distributed Optimization and statistical Learning via the alternating Direction Method of Multipliers*; Foundations and Trends® in Machine Learning: Hanover, MA, USA, 2011; Volume 3, pp. 1–122.
41. Zhou, C.; Paffenroth, R.C. Anomaly detection with robust deep autoencoders. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 665–674.
42. Chen, M.; Weinberger, K.; Sha, F.; Bengio, Y. Marginalized denoising auto-encoders for nonlinear representations. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1476–1484.
43. Shekhar, K.; Lapan, S.W.; Whitney, I.E.; Tran, N.M.; Macosko, E.Z.; Kowalczyk, M.; Adiconis, X.; Levin, J.Z.; Nemesh, J.; Goldman, M.; et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **2016**, *166*, 1308–1323.e30. [[CrossRef](#)] [[PubMed](#)]
44. McPherson, A.; Roth, A.; Laks, E.; Masud, T.; Bashashati, A.; Zhang, A.W.; Ha, G.; Biele, J.; Yap, D.; Wan, A.; et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.* **2016**, *48*, 758. [[CrossRef](#)] [[PubMed](#)]
45. Yeung, K.Y.; Ruzzo, W.L. Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. *Bioinformatics* **2001**, *17*, 763–774. [[CrossRef](#)]
46. Vu, T.N.; Nguyen, H.-N.; Calza, S.; Kalari, K.R.; Wang, L.; Pawitan, Y. Cell-level somatic mutation detection from single-cell RNA sequencing. *Bioinformatics* **2019**, *35*, 4679–4687. [[CrossRef](#)]
47. Zhou, Z.; Xu, B.; Minn, A.; Zhang, N.R. DENDRO: Genetic heterogeneity profiling and subclone detection by single-cell RNA sequencing. *Genome Biol.* **2020**, *21*, 10. [[CrossRef](#)] [[PubMed](#)]
48. Prashant, N.M.; Alomran, N.; Chen, Y.; Liu, H.; Bousounis, P.; Movassagh, M.; Edwards, N.; Horvath, A. SCReadCounts: Estimation of cell-level SNVs expression from scRNA-seq data. *BMC Genom.* **2021**, *22*, 689. [[CrossRef](#)] [[PubMed](#)]
49. Prashant, N.M.; Liu, H.; Dillard, C.; Ibeawuchi, H.; Alsaedy, T.; Chan, H.; Horvath, A.D. Improved SNV Discovery in Barcode-Stratified scRNA-seq Alignments. *Genes* **2021**, *12*, 1558. [[CrossRef](#)]
50. Dai, H.; Li, L.; Zeng, T.; Chen, L. Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Res.* **2019**, *47*, e62. [[CrossRef](#)]
51. Shi, Q.; Zhang, C.; Peng, M.; Yu, X.; Zeng, T.; Liu, J.; Chen, L. Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. *Bioinformatics* **2017**, *33*, 2706–2714. [[CrossRef](#)] [[PubMed](#)]
52. Chen, L.; Liu, R.; Liu, Z.-P.; Li, M.; Aihara, K. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci. Rep.* **2012**, *2*, 342. [[CrossRef](#)] [[PubMed](#)]
53. Shi, Q.; Zhang, C.; Guo, W.; Zeng, T.; Lu, L.; Jiang, Z.; Wang, Z.; Liu, J.; Chen, L. Local network component analysis for quantifying transcription factor activities. *Methods* **2017**, *124*, 25–35. [[CrossRef](#)] [[PubMed](#)]
54. Lu, Y.; Fang, Z.; Li, M.; Chen, Q.; Zeng, T.; Lu, L.; Chen, Q.; Zhang, H.; Zhou, Q.; Sun, Y.; et al. Dynamic edge-based biomarker non-invasively predicts hepatocellular carcinoma with hepatitis B virus infection for individual patients based on blood testing. *J. Mol. Cell Biol.* **2019**, *11*, 665–677. [[CrossRef](#)]
55. Shi, Q.; Li, X.; Peng, Q.; Zhang, C.; Chen, L. scDA: Single cell discriminant analysis for single-cell RNA sequencing data. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3234–3244. [[CrossRef](#)] [[PubMed](#)]