Briefings in Bioinformatics, 2022, pp. 1–10

doi: 10.1093/bib/bbad108
Advance Access Publication Date: Day Month Year
Problem Solving Protocol

PROBLEM SOLVING PROTOCOL

rcCAE: a convolutional autoencoder method for detecting intra-tumor heterogeneity and single-cell copy number alterations

Zhenhua Yu,[1,2,*] Furui Liu,[1] Fangyuan Shi[1,2] and Fang Du[1,2]

[1]School of Information Engineering, Ningxia University, 750021, Ningxia, China and [2]Collaborative Innovation Center for Ningxia Big Data and Artificial Intelligence Co-founded by Ningxia Municipality and Ministry of Education, Ningxia University, 750021, Ningxia, China
[*]Corresponding author. Email: zhyu@nxu.edu.cn

Abstract

Intra-tumor heterogeneity (ITH) is one of the major confounding factors that result in cancer relapse, and deciphering ITH is essential for personalized therapy. Single-cell DNA sequencing (scDNA-seq) now enables profiling of single-cell copy number alterations (CNAs) and thus aids in high-resolution inference of ITH. Here, we introduce an integrated framework called rcCAE, to accurately infer cell subpopulations and single-cell CNAs from scDNA-seq data. A convolutional autoencoder (CAE) is employed in rcCAE to learn latent representation of the cells as well as distill copy number information from noisy read counts data. This unsupervised representation learning via the CAE model makes it convenient to accurately cluster cells over the low-dimensional latent space, and detect single-cell CNAs from enhanced read counts data. Extensive performance evaluations on simulated datasets show rcCAE outperforms existing CNA calling methods, and is highly effective in inferring clonal architecture. Furthermore, evaluations of rcCAE on two real datasets demonstrate it is able to provide more refined clonal structure, of which some details are lost in clonal inference based on integer copy numbers.

Key words: Autoencoder, Hidden Markov model, Copy number alteration, Single-cell sequencing

Introduction

Intra-tumor heterogeneity and single-cell DNA sequencing

The evolutionary history of cancer is often depicted as a phylogenetic tree that elucidates the linage relationship between tumor clones [1, 2, 3, 4, 5, 6]. The emergence of each tumor clone is accompanied with a set of newly acquired genomic mutations such as single-nucleotide variation (SNV) and copy number alteration (CNA), and this genetic intra-tumor heterogeneity (ITH) has been one of the major confounding factors that contribute to therapy resistance and cancer relapse [1, 2, 6]. Without a full understanding of ITH, cancer treatment may often target only major clones whereas the low-prevalence clones may gain a growth advantage after therapy, resulting in cancer relapse [7]. Comprehensively charactering ITH is crucial to understand cancer progression and thus help designing personalized treatment [8, 9, 10]. This requires accurate detection of genomic mutations of each tumor clone as well as inference of the phylogenetic tree.

Single-cell DNA sequencing (scDNA-seq) makes it convenient to profile genomes of single cells and thus enables elucidation of ITH based on single-cell SNVs or CNAs [11]. In scDNA-seq, whole-genome amplification (WGA) is employed to amplify the DNA of isolated single cells to required amount of genetic material for sequencing. WGA tends to introduce amplification bias and errors into the produced sequences, which makes the sequenced reads error-prone and read count fluctuates extensively across genomic regions. The low sequencing coverage and amplification bias make the distribution of read counts over-dispersed, and this poses a critical challenge of how to accurately decipher CNAs from noisy read counts.

Existing methods for calling CNAs from scDNA-seq data

To date, a few methods have been specifically proposed to detect single-cell CNAs from scDNA-seq data [12, 13, 14, 15, 16, 17, 18, 19, 20]. These methods can be divided into two categories: detect single-cell CNAs with or without normal cells serving as negative controls. Ginkgo [12] is the first CNA calling method specifically designed for scDNA-seq data. It provides a web platform to detect single-cell CNAs using a circular binary segmentation (CBS) [21] algorithm, and also constructs a tumor phylogenetic tree based on the inferred single-cell copy-number profiles. SCNV [13] first identifies normal cells and pools them together as a composite control, then employs a bin-free segmentation approach to estimate copy numbers. It requires at least 20 normal cells to generate the control, and may be not applicable to datasets containing

© The Author 2022. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

few normal cells. Similarly, SCOPE [14] uses a generalized likelihood ratio test to jointly segment all cells, and leverages a modified Bayesian information criterion (BIC) [22] to determine the optimal number of segments. It automatically identifies normal cells based on calculation of the Gini coefficient and uses them to remove outlier cells. SCICoNE [15] is developed to simultaneously infer single-cell CNAs and a CNA mutation tree. One disadvantage of SCICoNE is that it treats breakpoint detection and CNA analysis as two separate procedures, which may disseminate errors made during breakpoint detection to downstream CNA analysis. SCYN [17] relies on SCOPE to prepare and preprocess read counts, thereafter employs a dynamic programming algorithm to find the optimal segmentation results by maximizing a simplified modified BIC. The authors also introduce SeCNV [20] to perform cross-sample segmentation on normalized read counts. Recently, a method called SCONCE [19] is proposed to call single-cell CNAs using a hidden Markov model (HMM), and it requires normal cells as negative controls to reduce false CNA calls.

Despite to the fact that these methods perform well on their respective test datasets, there are still two problems that are not well addressed in these methods: 1) how to accurately detect single-cell CNAs from noisy and extensively fluctuated read counts without normal cells as negative controls; 2) how to directly infer cell subpopulation composition from the raw read counts rather than the inferred integer copy numbers, thus avoid propagation of errors made in CNA analysis to clonal structure inference. These two problems are intertwined with each other, therefore integrated frameworks for simultaneously inferring single-cell CNAs and cell subpopulations from raw read counts are sorely required for scDNA-seq data.

Proposed method rcCAE

Deep generative models such as variational autoencoder (VAE) [23, 24] have shown great success in learning representations from single-cell sequencing data [25, 26, 27, 28, 29, 30]. The intuition behind these methods is that observed genetic heterogeneity among single cells actually results from some biological processes that are related to tumor evolution and could be modeled as latent variables in the autoencoders [28]. Inspired by this conception, we develop a novel computational method for analyzing single-cell read counts based on a Convolutional AutoEncoder framework (rcCAE) as shown in Figure 1. rcCAE uses a convolutional encoder network to project normalized single-cell read counts into a low-dimensional latent space where the cells are clustered into distinct subpopulations through a Gaussian mixture model (GMM), and leverages a convolutional decoder network to recover the read counts from learned latent representations. This encoding-decoding process could be treated as a distillation process that extracts copy number information from the noisy read counts by attenuating the effects of non-biological confounding factors, thus makes the recovered data being higher in signal-to-noise ratio. We then apply a novel hidden Markov model (HMM) on the recovered data to simultaneously segment the genome and infer absolute copy numbers for each cell. To the best of our knowledge, rcCAE is the first deep learning-based method for calling single-cell CNAs from scDNA-seq data. We comprehensively evaluate rcCAE on both simulated and real datasets to verify its high scalability on different-sized scDNA-seq datasets, and also compare rcCAE to the state-of-the-art (SOTA) CNA detection methods. The results showcase our method can accurately characterize cell subpopulation composition in latent space, and is more accurate than the competitors in calling absolute copy numbers of single cells.

Results

The workflow of rcCAE

The framework of rcCAE mainly consists of four components: 1) a module to fetch read counts of fixed-size bins from the BAM files of sequenced cells, and preprocess the read counts; 2) a convolutional autoencoder (CAE) to learn latent representations of single cells and enhance the quality of read counts; 3) a GMM based clustering model to find cell subpopulations; and 4) an HMM to predict copy number segments of each cell. Specifically, encoder of the CAE projects log2 transformed read counts (LRC) of each cell into a low-dimensional space and the decoder recovers the LRC data from the latent space. Given the learned latent features, the GMM model coupled with BIC for model selection is employed to cluster cells into distinct subpopulations. The recovered LRC data from the CAE are then analyzed using the HMM to detect CNAs of each cell. More methodological details of rcCAE are given in Methods section.

Evaluation on simulated datasets

Simulation of scDNA-seq datasets

As human cancers often exhibit aneuploidy, we generate datasets with different tumor ploidy (Simulation A: near-diploidy, Simulation B: near-triploidy and Simulation C: near-tetraploidy). To further investigate potential application of rcCAE to datasets with higher ploidy levels, we also simulate near-pentaploidy datasets (Simulation D). Following the simulation process adopted in [31], we set the minimum size of CNAs to 3Mb, and the maximum size to 20Mb. CNAs are randomly sampled and iteratively inserted into chromosomes 1-3 of human genome hg19 until there are no feasible loci to insert the simulated CNA. We generate 500 cells from 16 subpopulations including a normal subpopulation and 15 tumor clones for Simulations A-C, while generate 100 cells from 5 subpopulations including a normal subpopulation and 4 tumor clones for Simulation D. The simulation is repeated 5 times for each tumor ploidy, and the average numbers of CNAs for Simulations A-D are 18, 35, 36 and 40, respectively. More details about dataset simulation are provided in Supplementary Methods.

Performance evaluation

Estimating copy number of a bin can be treated as a classification problem where each copy number is a class. To measure copy number estimation accuracy, we make a comparison between the predicted copy number and the ground truth for each bin, and calculate three metrics for performance evaluation: 1) accuracy: proportion of bins whose copy numbers are accurately predicted; 2) balanced accuracy: mean of per-class accuracies [32]; and 3) macro-averaged F-measure: mean of per-class F1-scores. In addition, we also assess ploidy estimation performance by comparing the predicted average copy number (ACN) with the ground truth value. For inferring tumor clones, we use adjusted rand index (ARI) to measure the clustering accuracy.

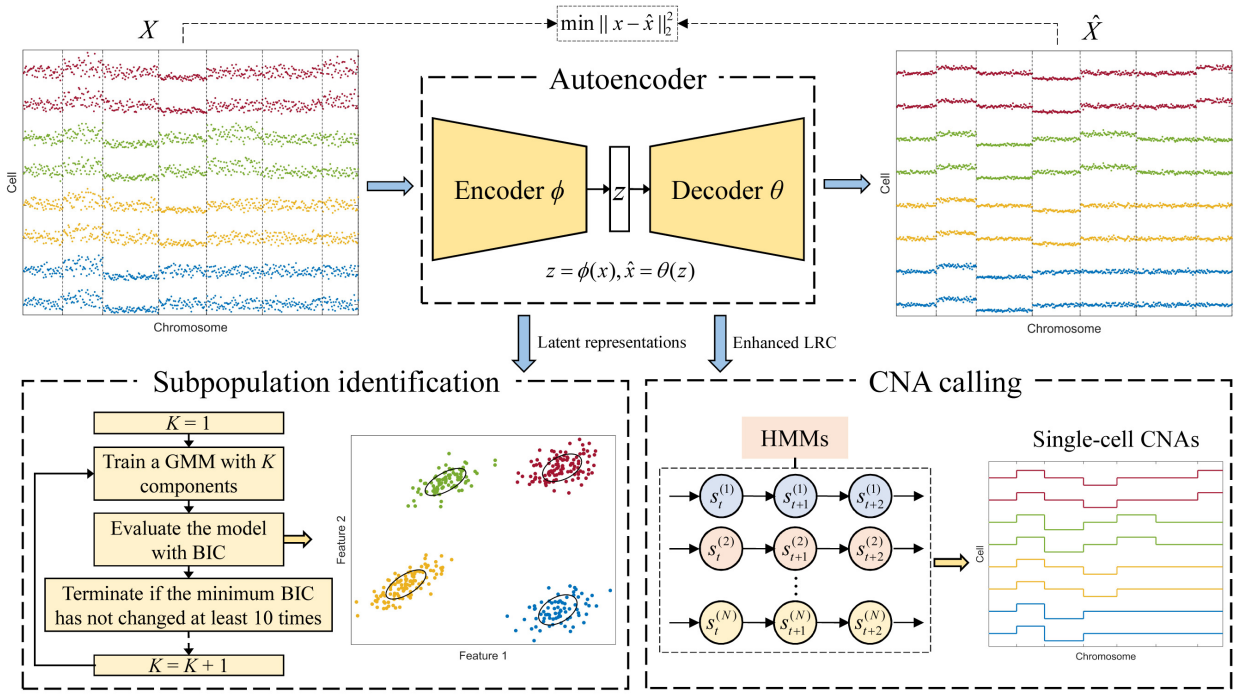

Fig. 1. The workflow of rcCAE. rcCAE takes the log2 transformed read counts (LRC) of all cells as inputs, and yields cell subpopulation composition as well as single-cell copy number segments as outputs. With a convolutional autoencoder for unsupervised representation learning, rcCAE projects the LRC data into a latent space where cells are efficiently clustered into distinct subpopulations, and effectively distills copy number information from noisy LRC data. A Gaussian mixture model is employed to identify cell subpopulations based on the latent representation of cells, and the number of subpopulations is determined by Bayesian information criterion. For calling CNAs for each cell, rcCAE employs a hidden Markov model to jointly segment the genome and infer absolute copy number for each segment.

When running rcCAE on the simulated datasets, we set bin size to 200kb and the number of epochs for training the CAE to 200 on all simulated datasets as training loss changes little after 200 epochs as shown in Figure S1. Example results of rcCAE on three simulated datasets from Simulations A-C are given in Figures S2-S4. It is observed that rcCAE significantly enhances the quality of LRC by decreasing the variance of LRC in each copy number state. This improvement benefits from the adopted encoding-decoding distillation procedure that makes it convenient to distinguish between different copy number states and between distinct tumor clones.

Comparison to other methods

We compare rcCAE to five state-of-the-art CNA calling methods including SCOPE [14], SCYN [17], SCICoNE [15], SeCNV [20] and SCONCE [19] on the simulated datasets from Simulations A-C. All these methods use read counts to call single-cell CNAs and perform well on their respective test datasets. We exclude CHISEL [18] from performance comparison as it uses both read counts and allele frequency data for estimating copy numbers. In addition, we introduce a null model (H0) as the baseline model that predicts the most common copy number (2 for diploid, 3 for triploid and 4 for tetraploid datasets) for all bins. The detailed description about the performance assessment and parameter configuration for the investigated methods are provided in the Supplementary Methods.

We first examine ploidy estimation accuracy of the investigated methods. Difference between the real and predicted ACNs (denoted as ΔACN) is calculated for each cell to indicate the ploidy estimation accuracy. The results in Figure 2 show rcCAE yields better predictions than the SOTAs across different simulations, especially on near-tetraploid datasets. SCICoNE overestimates the ploidy on near-diploid datasets (Simulation A) and predicts all cells to be near-triploid (median ΔACN is -1.02), while other methods perform well with median ΔACN close to 0. On near-triploid datasets (Simulation B), SCOPE, SCYN and SeCNV perform similarly and accurately estimate the ploidy for most of the samples. SCICoNE tends to overestimate the ploidy (median ΔACN is -0.21) and SCONCE erroneously identifies majority of the cells to be near-tetraploid (median ΔACN is 1). By comparison, our method produces accurate inferences of the ploidy for all cells, the mean of ΔACN is 0.0048. On near-tetraploid datasets (Simulation C), all of the existing methods tend to underestimate the ploidy, the median ΔACN of SCOPE, SCYN, SCICoNE, SeCNV and SCONCE are 1.87, 1.87, 0.69, 1.82 and 1.961, respectively, indicating they fail to distinguish between different copy number states on most of the tetraploid samples. Our method is able to assign correct ploidy to each cell, therefore achieves better performance than the competitors. Figure S5 gives a detailed view of the ploidy estimation results of all methods, and the results show rcCAE' predictions are in concordance with the ground truth.

We also assess how well rcCAE and other methods perform in estimating absolute copy numbers, and Figure 3 shows F-measures of the methods. The copy number accuracy and balanced accuracy are given in Figures S6-7. All performance metrics are calculated per-cell for each method. We use

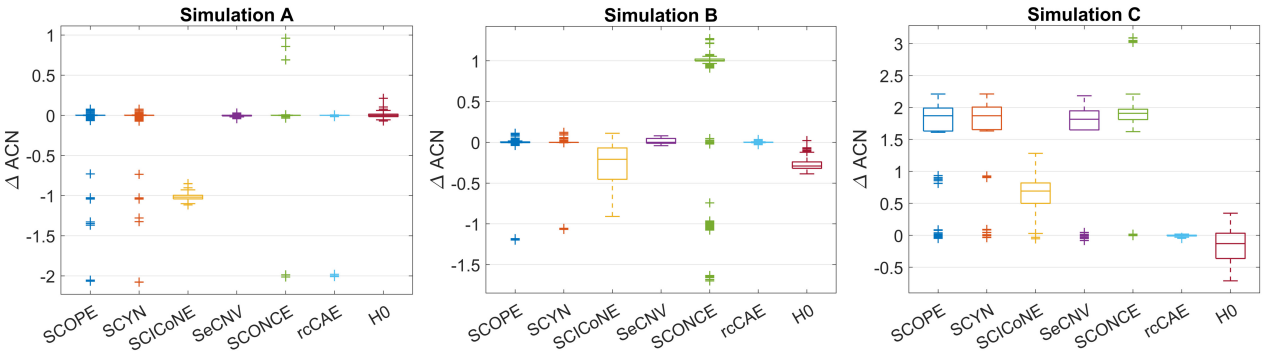

Fig. 2. Ploidy estimation accuracy of the investigated methods. Average copy number (ACN) is calculated for each cell, and the difference between real and predicted ACNs (denoted as ΔACN) is used to indicate ploidy estimation accuracy of each method. Results on simulated 5 near-diploid datasets (Simulation A), 5 near-triploid datasets (Simulation B) and 5 near-tetraploid datasets (Simulation C) are analyzed. Each boxplot is generated based on all cells included in the corresponding simulation. The average number of bins used for evaluation is 2509 for Simulation A, 2636 for Simulation B and 2166 for Simulation C. H0 denotes the baseline model that predicts the most common copy number (2 for diploid, 3 for triploid and 4 for tetraploid datasets) for all bins. rcCAE generally performs better than other methods, especially on the near-tetraploid datasets where all of the existing methods tend to underestimate the ploidy.

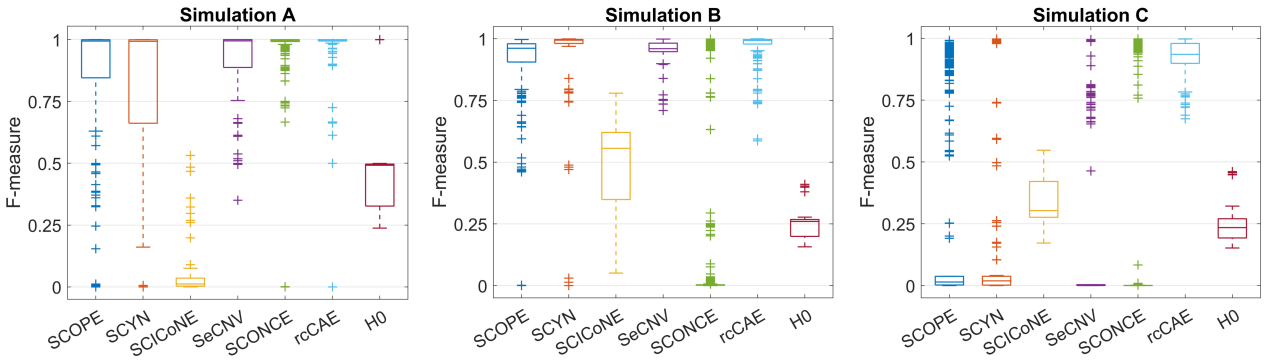

Fig. 3. F-measure of the investigated methods. For each cell, the macro-averaged F-measure is calculated as the mean of per-class F1-scores. Results on simulated 5 near-diploid datasets (Simulation A), 5 near-triploid datasets (Simulation B) and 5 near-tetraploid datasets (Simulation C) are analyzed. Each boxplot is generated based on all cells included in the corresponding simulation. The average number of bins used for evaluation is 2509 for Simulation A, 2636 for Simulation B and 2166 for Simulation C.

the common bins covered by the results of all methods for evaluation on each dataset. The average number of bins used for evaluation is 2509 for Simulation A, 2636 for Simulation B and 2166 for Simulation C. On near-diploid datasets, SCOPE, SCYN, SeCNV, SCONCE and rcCAE reaches ≥0.993 median F-measure, while SCICoNE fails to identify absolute copy numbers (median F-measure is 0.012) due to overestimation of the ploidy. Similar performance is observed for SCOPE, SCYN, SeCNV and rcCAE on near-triploid datasets. By comparison, SCICoNE yields median F-measure of 0.556 and SCONCE performs poorly as it overestimates the ploidy. On near-tetraploid datasets, the existing methods suffer from degraded performance due to underestimation of the ploidy, and rcCAE surpasses the competitors with median F-measure of 0.935. In addition, we measure the consistency of each method in identifying each copy number state, and the results are shown in Figure S8.

We proceed to evaluate the performance of rcCAE in identifying tumor clones. Two conventional dimensionality reduction methods including PCA [33] and t-SNE [34] are adopted for comparison by clustering the projected data using the same GMM. More details about how t-SNE is used are provided in the Supplementary Methods. An example of the clustering results is given in Figure S9 that shows rcCAE is more effective in disentangling distinct cell subpopulations. Figure S10 gives ARI scores of all methods, and suggests our method outperforms PCA and t-SNE across different tumor ploidy. rcCAE exactly deciphers clonal composition on 12 out of 15 datasets, and performs acceptedly well on the remaining 3 datasets (ARI > 0.94).

Finally, we assess rcCAE on the simulated near-pentaploidy datasets. An example of copy number estimation results (Figure S11) implies rcCAE significantly improves the quality of LRC data and accurately identifies the most abundant copy number of 5 for all tumor clones. The results in Figure S12 show rcCAE achieves median accuracy of 0.972, balanced accuracy of 0.984 and F-measure of 0.931, which suggests our method is still highly effective in dealing with high-ploidy datasets.

The effects of number of cells and hyper-parameters

We assess the effects of number of cells and hyper-parameters on rcCAE' performance. The results in Figure 4 and Figure S13 suggest our method is able to deliver accurate predictions of copy number (an example is given in Figure S14) and ploidy across different data sizes, and performs acceptedly well in

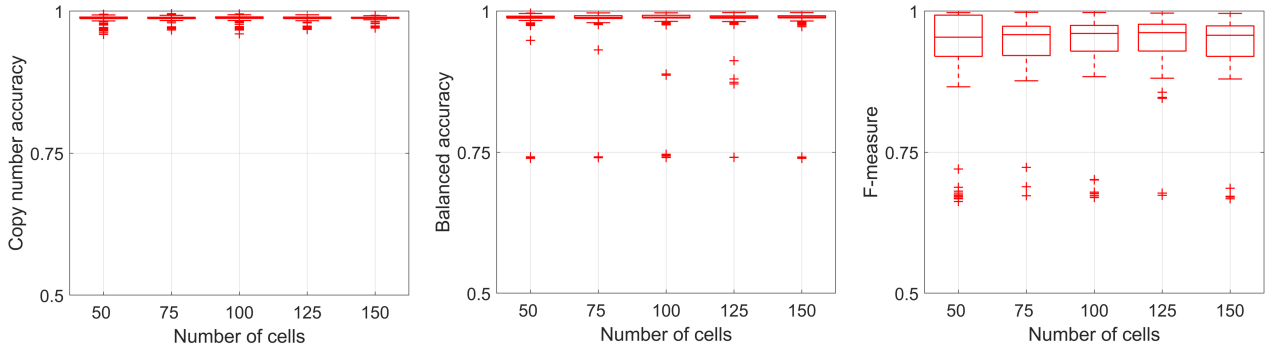

Fig. 4. Performance evaluation on different-sized datasets. Values in {50, 75, 100, 125, 150} are tested for the number of cells, and 50 datasets for performance evaluation are generated from a near-diploid dataset in the Simulation A. The cell clustering, copy number as well as ploidy estimation accuracy are measured with respected to the number of cells. Each boxplot is generated based on all cells included in the corresponding datasets. The number of bins used for evaluation is 2955 on all datasets.

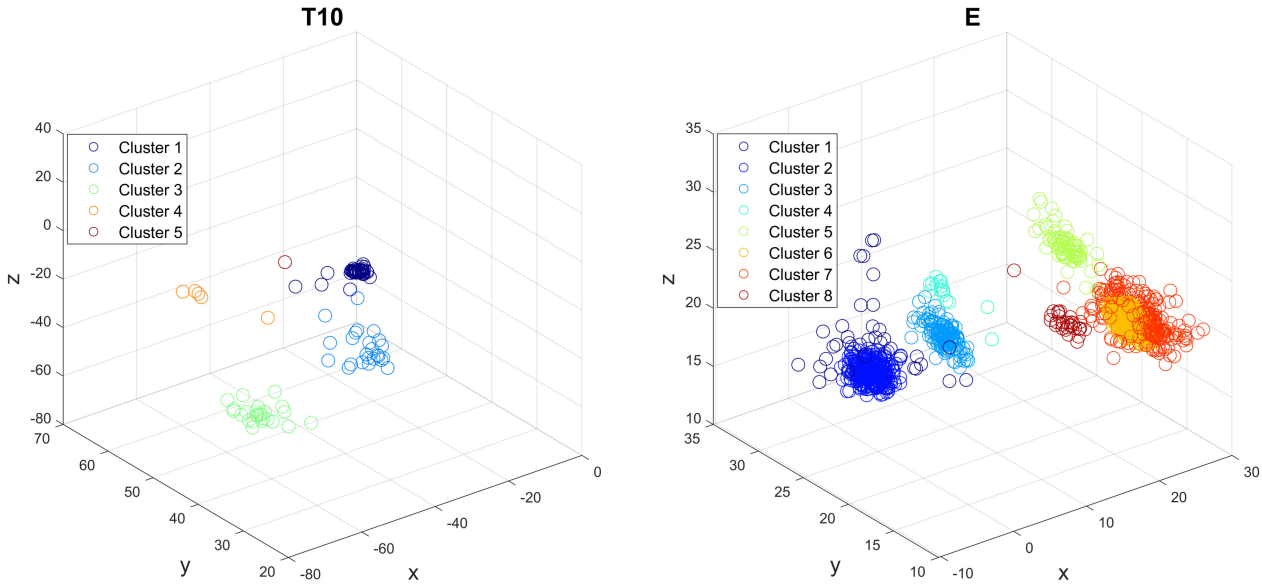

Fig. 5. Clustering results of rcCAE on real datasets. On the breast cancer dataset (T10) consisting of 100 cells, rcCAE clusters the cells into 5 subpopulations (labeled as 1-5) each with 45, 24, 25, 5 and 1 cells, respectively (left subfigure). On the 10X Genomics dataset, rcCAE clusters the cells into 8 subpopulations (labeled as 1-8) each with 37, 351, 152, 16, 75, 552, 234 and 29 cells, respectively (right subfigure).

inferring subpopulation composition when the number of cells is larger than 50.

We also examine the effects of hyper-parameters including kernel size, latent dimension and bin size on the performance of rcCAE. The results suggest rcCAE reaches consistently good performance across different values of the hyper-parameters (Figures S15-20), and detailed description is given in Supplementary Results.

Evaluation on a breast cancer dataset

We use rcCAE to infer tumor clonal structure from a breast ductal carcinoma dataset consisting of 100 single cells [35]. A previous study [35] has clustered the cells into 4 subpopulations (D+P, H, AA and AB) based on integer copy number profiles, and each subpopulation consists of 47, 24, 25 and 4 cells, respectively. Given the highly different copy number profiles between the subpopulations, we assess if rcCAE could decipher the same clonal structure from single-cell raw read counts data.

Read counts data are obtained for every 500kb bins, and we get a data matrix with size of 100×5120. The number of epochs for training the CAE is set to 200 (as suggested by the results in Figure S21). Our method clusters the cells into 5 subpopulations each with 45, 24, 25, 5 and 1 cells, respectively (Figure 5 and Figures S22-23). Cluster 1 consists of predominantly diploid cells and aligns well with the D+P subpopulation. Cells from cluster 2 are characterized by hemizygous deletions on multiple chromosomes, and the results are in high concordance with the H subpopulation, indicating our method precisely identifies this tumor clone. Cluster 3 is comprised of aneuploid cells that exhibit extensive copy number gains on all chromosomes and hemizygous deletions on chromosomes 4-7. The number of cells is consistent with that of the AA subpopulation, these results demonstrate our method exactly recovers the AA subpopulation. Cluster 4 contains 5 cells of which 4 cells show highly similar copy number profiles, and may be from the AB subpopulation. Our method finds another cluster 5 formed by only one cell that is featured by

hemizygous deletions on chromosomes 2, 3q, 8, 9 and 16. This cell is significantly different from the cells of cluster 2 in the copy number profiles (Figure S22), therefore classified into a separate cluster by our method. By directly analyzing the raw read counts data, rcCAE is able to provide more refined clonal structure of which some details are lost in clonal inference based on single-cell integer copy numbers.

Evaluation on a 10X Genomics dataset

To further verify the performance of rcCAE on large datasets containing thousands of cells, we apply rcCAE to a 10X Genomics dataset [18]. The cells are obtained from a frozen breast tumor tissue and each cell is sequenced with a coverage between 0.02X and 0.05X. A previous study has assigned 1446/2053 of the cells to 6 clusters including one diploid subpopulation (containing 388 cells) and 5 aneuploid tumor clones (each with 168, 58, 20, 782 and 30 cells, respectively) using CHISEL algorithm [18]. Tumor clones 2 and 3 show highly consistent copy number profiles across the whole genome but different allele-specific copy numbers on chromosome 2, therefore we merge the two clones into a same cluster as they are indistinguishable from each other solely based on the total copy numbers. We now label the new clusters with I-V (each with 388, 168, 78, 782 and 30 cells, respectively). We employ rcCAE to infer clonal structure from this dataset and check if our method could yield similar results as CHISEL by only using read counts data.

The bin size is set to 500kb, and the number of epochs for training the CAE is set to 50 (as shown in Figure S24). rcCAE assigns the cells into 8 clusters (labeled as 1-8) each with 37, 351, 152, 16, 75, 552, 234 and 29 cells, respectively (Figure 5 and Figure S23). A comparison between the original and enhanced LRC data is given in Figure 6 and the results show our method significantly improves the data quality. We find clusters 1 and 2 are diploid subpopulations having highly similar copy number profiles and the total number of cells from them is equal to that of CHISEL cluster I, which suggests rcCAE subdivides the cells of CHISEL cluster I into two sets that are in close proximity in the latent space (Figure 5). Similar results are observed for cluster 3 and 4 (labeled as clone A) exhibiting consistent copy numbers and aligning well with the CHISEL cluster II. Cluster 5 (labeled as clone B) has copy number of 4 on chromosomes 2-3 and copy number of 3 on chromosome 8, and presents very similar copy number profiles with the CHISEL cluster III. The results indicate rcCAE effectively deciphers this clone by identifying 75/78 of the cells. Clusters 6 and 7 (labeled as clone C) also show quite similar copy number profiles and have copy number of 3 on chromosome 2 when compared to clone B. In addition, 786 cells in clone C completely encompasses the 782 cells from the CHISEL cluster IV, which suggests clone C aligns well with the CHISEL cluster IV. Cluster 8 represents a minor clone (labeled as clone D) that contains 29 cells, and we found clone D is in concordance with the CHISEL cluster V. Taken together, rcCAE may be sensitive to LRC changes among the cells belonging to same cluster and tends to generate over-segmented clustering results, however it is still able to accurately recover the copy number profiles of single cells by exploiting strong reconstruction ability of the CAE model.

Discussion and conclusions

In this paper, we develop rcCAE, an autoencoder based method for jointly inferring ITH and single-cell copy numbers from scDNA-seq data. The approach leverages a convolutional autoencoder to learn latent representations of single cells as well as enhance the quality of read counts, thus provides an integrated framework that enables efficient characterization of underlying cell subpopulations in a low-dimensional latent space, and accurate detection of single-cell copy numbers from improved read counts. Compared to ITH inference based on integer single-cell copy numbers, rcCAE directly deciphers ITH from original read counts, which avoids potential error propagation from copy number analysis to ITH inference. For estimating copy number segments for each cell, rcCAE uses a novel hidden Markov model to explicitly represents tumor ploidy as a key parameter in the emission models to find the solution that best explains the observed read counts. rcCAE is the first method that borrows both the strong representation learning ability of autoencoder and good explanation capability of conventional statistical methods for analyzing single-cell read counts, thus opens a new paradigm for CNA-based ITH analysis.

Further improvements in single-cell CNA calling could be achieved by cross-cell segmentation of the read counts. First, cells in same clone have a common set of breakpoints, and incorporating this information into CNA detection could aid in better identifying breakpoints. Second, cells from different clones may share same breakpoints since tumor evolves by accumulating mutations [36], and cross-cell segmentation could attenuate the effects of read counts fluctuation and thus deliver better segmentation results. A more challenging task is to perform cross-cell segmentation along with identification of absolute copy numbers. Existing methods that are not based on HMMs often require a post-processing procedure to estimate absolute copy numbers after obtaining segmentation results. Despite to the fact that HMM-based methods can provide segmentation as well as absolute copy numbers for a single cell, how to apply HMM to multiple cells is still an open question, and a major challenge is that hidden space size may grow exponentially with the number of cells. Taken together, cross-cell segmentation with simultaneous identification of absolute copy numbers is a promising direction for future study of ITH inference from scDNA-seq data.

Methods

Obtain and preprocess read counts

We divide the genome into fix-sized bins and obtain read count of a bin by counting the reads mapped to the bin. BamTools [37] is used to fetch information of each read alignment. We also calculate GC percent and mappability score of each bin to correct read counts bias as well as exclude outlier bins. Specifically, bins that have mappability score of $\geq$0.9 or GC percent of $\leq$10% are excluded from downstream analysis [38]. To make read counts comparable among cells, we calculate mean of read counts across all bins for each cell, and divide read count of each bin with the mean. We further exclude bins with mean value of the normalized read counts being in the lower or upper 1% quantile. Considering cells may end up with low sequencing coverage or high amplification bias, we use the normalized read counts to calculate the Gini coefficient for each cell, and eliminate cells having $\geq$0.3 Gini coefficient.

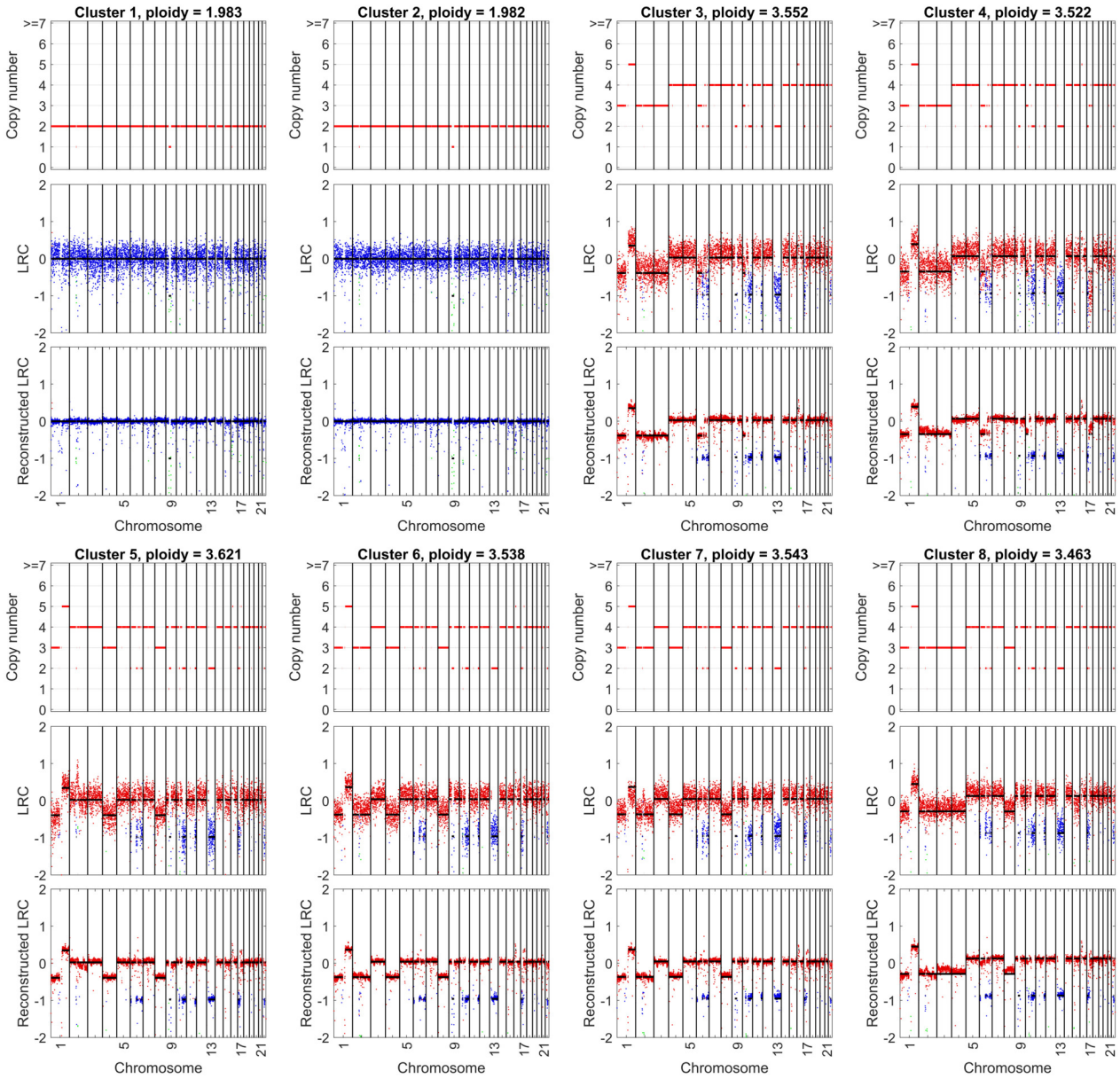

Fig. 6. Copy number estimation results of rcCAE on the 10X Genomics dataset. Clusters 1 and 2 are diploid subpopulations having highly similar copy number profiles. Clusters 3 and 4 showcase consistent copy numbers with average ploidy of 3.55 and may come from a same clone (labeled as clone A). Cluster 5 represents another clone (labeled as clone B) that has copy number of 4 on chromosomes 2-3 and copy number of 3 on chromosome 8. Clusters 6 and 7 show quite similar copy number profiles, and may derive from a same clone (labeled as clone C). Compared to clusters 1-7, cluster 8 represents a minor clone (labeled as clone D), and it has copy number of 3 on chromosomes 2-3. The quality of LCR data is significantly enhanced by rcCAE's encoding-decoding framework, thus single-cell CNAs could be accurately detected.

Finally, per-cell correction for GC-content bias using a median normalization approach [39] is conducted to improve the quality of read counts. For computational convenience, we calculate the log2 transformed read counts of all cells for downstream analysis, here denoted by a $N \times M$ matrix X (N denotes the number of cells and M is the number of bins).

Learn latent representations and enhance the quality of LRC

Inspired by recent success of deep generative models in learning representations of cells from single-cell RNA sequencing (scRNA-seq) data [26, 27], we introduce a convolutional autoencoder based approach to represent single cells in a latent space where cells could be efficiently clustered into distinct subpopulations. Formally, for each x (one row of the X), we use the encoder ϕ of the CAE model to get its latent representation $z = \phi(x)$, and recover x through the decoder network θ of the CAE, i.e. $\hat{x} = \theta(z)$. Here the latent z could be assumed to indicate underlying biological processes that result in the observed cell and are related to mutational events shared by other cells derived from the same tumor clone. The data reconstruction could be treated as a refinement process that distills copy number information from original

LRC data, therefore make it convenient to accurately estimate copy numbers from the recovered LRC.

We train the CAE model to minimize the reconstruction loss $\mathcal{L}(x, \hat{x}; \phi, \theta)$ which is measured using the ℓ_2-norm:

$$\mathcal{L}(x, \hat{x}; \phi, \theta) = \frac{1}{2M} \|x - \hat{x}\|_2^2 \tag{1}$$

During each training iteration, the mean loss is calculated for a batch of samples, and conventional gradient decent algorithm is employed to update the network parameters. After the model training is completed, the latent representations and recovered LRC data of all cells are inferred as Z and $\hat{X}$, respectively.

Infer cell subpopulations

Given the latent representations Z of all cells, we employ a GMM based clustering approach to find cell subpopulations by following a previous study [28] that uses a GMM to cluster cells based on scRNA-seq data. Specifically, the probability density function of z is defined as follows:

$$f(z|\lambda) = \sum_{k=1}^{K} \pi_k f(z|\mu_k, \Sigma_k) \tag{2}$$

where K is the number of clusters, π_k is the proportion of the k-th cluster, $f(z|\mu_k, \Sigma_k)$ is the density function associated with the k-th multivariate Gaussian distribution with mean μ_k and covariance matrix Σ_K, and $\lambda = (\pi_1, \mu_1, \Sigma_1, \ldots, \pi_K, \mu_K, \Sigma_K)$ are model parameters.

We adopt expectation maximization (EM) algorithm to update the model parameters, and assign each cell to the cluster having the maximum posterior probability. To determine the best number of clusters (the value of K), we calculate the BICs of the models associated with different values of K, and select the model having the minimum BIC as the best solution. For implementation, the value of K is initialized to 1 and then iteratively increased by one until the minimum BIC has remained unchanged at least 10 times. After the best model is found, the cells can be divided into distinct subpopulations, and cells belonging to the same subpopulation are considered to have highly similar copy number profiles across whole genome.

Detect single-cell CNAs

To jointly segment the genome and infer absolute copy numbers for each cell, we employ an HMM to analyze the enhanced LRC data. The HMM defines 11 hidden states each of which represents a copy number that ranges from 0 to 10. Here we use $\hat{X}_{ij}$ to denote the LRC of the i-th cell in the j-th bin. We assume $\hat{X}_{ij}$ follows a normal distribution under hidden state s:

$$f(\hat{X}_{ij}|s, \sigma_i, o_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(\hat{X}_{ij} - \mu_{is})^2}{2\sigma_i^2}\right) \tag{3}$$

with the mean μ_{is} defined as:

$$\mu_{is} = \log_2\left(\frac{c_s}{2}\right) + o_i \tag{4}$$

where o_i denotes baseline shift of the LRC due to change of ploidy, σ_i is a global parameter representing the standard variance of LRC, and c_s represents copy number in state s. The parameter o_i is introduced to recognize the most abundant copy number state. The value of o should be close to 0 for diploid samples, and lower than 0 for hyperploid samples. We use EM algorithm to estimate the model parameters. The expected partial log-likelihood is formulated as:

$$E(LL_i) = \sum_{j=1}^{M} \sum_{s=1}^{S} \gamma_{js} \log\left(f(\hat{X}_{ij}|s, \sigma_i, o_i)\right) \tag{5}$$

where γ_{js} is the posterior probability that the j-th bin is in state s, and S denotes the number of hidden states. During the n-th iteration of the EM algorithm, we update parameter o_i with the following formula:

$$o_i^{(n)} = \frac{\sum_{j=1}^{M} \sum_{s=1}^{S} \gamma_{js}^{(n)} \left(\hat{X}_{ij} - \log_2\left(\frac{c_s}{2}\right)\right)}{\sum_{j=1}^{M} \sum_{s=1}^{S} \gamma_{js}^{(n)}} \tag{6}$$

and update σ_i by:

$$\sigma_i^{(n)} = \sqrt{\frac{\sum_{j=1}^{M} \sum_{s=1}^{S} \gamma_{js}^{(n)} \left(\hat{X}_{ij} - \log_2\left(\frac{c_s}{2}\right) - o_i^{(n)}\right)^2}{\sum_{j=1}^{M} \sum_{s=1}^{S} \gamma_{js}^{(n)}}} \tag{7}$$

We test different initial values for o_i to check which ploidy should be the most abundant copy number, and initialize σ_i to the standard variance of LRC of the i-th cell. After the algorithm converges, the copy number of each bin is deduced from the state that has the maximum posterior probability. We also propose a post-processing step to recalibrate the copy number of highly amplified regions, and more details about the implementation are provided in the Supplementary Methods.

Key Points

- We develop a convolutional autoencoder based method rcCAE, which jointly detects cell subpopulations and single-cell CNAs from raw read counts.
- rcCAE exploits the encoding-decoding process to distill copy number information from extensively fluctuated read counts.
- rcCAE is able to directly infer cell subpopulations from read counts, and thus could give more robust characterization of ITH.

Competing interests

No competing interest is declared.

Authors' contributions

Z.Y. and F.D. conceived the study. Z.Y. developed the algorithm and drafted the manuscript. F.L. and F.S. analyzed the data. All authors read and approved the final manuscript for publication.

Ethics approval and consent to participate

Not applicable.

Funding

This work was supported in part by the National Natural Science Foundation of China (Grant Nos.61901238, 62062058) and West Light Foundation of The Chinese Academy of Sciences (XAB2019AW12).

Software availability

The source code and detailed documentation of rcCAE are available at https://github.com/zhyu-lab/rccae.

Data availability

The breast cancer dataset is available from NCBI SRA with accession number SRA018951, and the 10X Genomics dataset can be freely downloaded from https://support.10xgenomics.com/single-cell-dna/datasets/1.0.0/breast_tissue_E_2k.

References

1. Mel Greaves and Carlo C Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012.
2. Rebecca A Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345, 2013.
3. K. Yuan, T. Sakoparnig, F. Markowetz, and N. Beerenwinkel. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol*, 16(1):36, Feb 2015.
4. E. M. Ross and F. Markowetz. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol*, 17:69, Apr 2016.
5. M. El-Kebir. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34(17):i671–i679, 09 2018.
6. Samra Turajlic, Andrea Sottoriva, Trevor Graham, and Charles Swanton. Resolving genetic heterogeneity in cancer. *Nature Reviews Genetics*, 20(7):404–416, 2019.
7. Nicholas McGranahan and Charles Swanton. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, 168(4):613–628, 2017.
8. L. Oesper, A. Mahmoody, and B. J. Raphael. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol*, 14(7):R80, Jul 2013.
9. Ash A Alizadeh, Victoria Aranda, Alberto Bardelli, Cedric Blanpain, Christoph Bock, Christine Borowski, Carlos Caldas, Andrea Califano, Michael Doherty, Markus Elsner, et al. Toward understanding and exploiting tumor heterogeneity. *Nature medicine*, 21(8):846–853, 2015.
10. Z. Yu, A. Li, and M. Wang. CloneCNA: detecting subclonal somatic copy number alterations in heterogeneous tumor samples from whole-exome sequencing data. *BMC Bioinformatics*, 17:310, Aug 2016.
11. Charles Gawad, Winston Koh, and Stephen R Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175–188, 2016.
12. Tyler Garvin, Robert Aboukhalil, Jude Kendall, Timour Baslan, Gurinder S Atwal, James Hicks, Michael Wigler, and Michael C Schatz. Interactive analysis and assessment of single-cell copy-number variations. *Nature methods*, 12(11):1058–1060, 2015.
13. Xuefeng Wang, Hao Chen, and Nancy R Zhang. Dna copy number profiling using single-cell sequencing. *Briefings in bioinformatics*, 19(5):731–736, 2018.
14. Rujin Wang, Dan-Yu Lin, and Yuchao Jiang. Scope: a normalization and copy-number estimation method for single-cell dna sequencing. *Cell systems*, 10(5):445–452, 2020.
15. Jack Kuipers, Mustafa Anıl Tuncel, Pedro Ferreira, Katharina Jahn, and Niko Beerenwinkel. Single-cell copy number calling and event history reconstruction. *bioRxiv*, 2020.
16. Xian F Mallory, Mohammadamin Edrisi, Nicholas Navin, and Luay Nakhleh. Methods for copy number aberration detection from single-cell dna-sequencing data. *Genome biology*, 21(1):1–22, 2020.
17. Xikang Feng, Lingxi Chen, Yuhao Qing, Ruikang Li, Chaohui Li, and Shuai Cheng Li. Scyn: single cell cnv profiling method using dynamic programming. *BMC genomics*, 22(5):1–13, 2021.
18. Simone Zaccaria and Benjamin J Raphael. Characterizing allele-and haplotype-specific copy numbers in single cells with chisel. *Nature biotechnology*, 39(2):207–214, 2021.
19. Sandra Hui and Rasmus Nielsen. Sconce: a method for profiling copy number alterations in cancer evolution using single-cell whole genome sequencing. *Bioinformatics*, 38(7):1801–1808, 2022.
20. Wang Ruohan, Zhang Yuwei, Wang Mengbo, Feng Xikang, Wang Jianping, and Li Shuai Cheng. Resolving single-cell copy number profiling for large datasets. *Briefings in Bioinformatics*, 23(4):bbac264, 2022.
21. Adam B Olshen, E Seshan Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, 2004.
22. Nancy R Zhang and David O Siegmund. Model selection for high-dimensional, multi-sequence change-point problems. *Statistica Sinica*, pages 1507–1538, 2012.
23. Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.
24. Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.
25. Dongfang Wang and Jin Gu. Vasc: dimension reduction and visualization of single-cell rna-seq data by deep variational autoencoder. *Genomics, proteomics & bioinformatics*, 16(5):320–331, 2018.
26. Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune H Pers, and Ole Winther. scvae: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415–4422, 2020.
27. Valentine Svensson, Adam Gayoso, Nir Yosef, and Lior Pachter. Interpretable factor models of single-cell rna-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421, 2020.
28. Sabrina Rashid, Sohrab Shah, Ziv Bar-Joseph, and Ravi Pandya. Dhaka: variational autoencoder for unmasking tumor heterogeneity from single cell genomic data. *Bioinformatics*, 37(11):1535–1543, 2021.
29. Raktim Mitra and Adam L MacLean. Rvagene: generative modeling of gene expression time series data.

Bioinformatics, 37(19):3252–3262, 2021.

30. Gaoyang Li, Shaliu Fu, Shuguang Wang, Chenyu Zhu, Bin Duan, Chen Tang, Xiaohan Chen, Guohui Chuai, Ping Wang, and Qi Liu. A deep generative model for multi-view profiling of single-cell rna-seq and atac-seq data. *Genome biology*, 23(1):1–23, 2022.
31. Xian F Mallory, Mohammadamin Edrisi, Nicholas Navin, and Luay Nakhleh. Assessing the performance of methods for copy number aberration detection from single-cell dna sequencing data. *PLoS computational biology*, 16(7):e1008012, 2020.
32. Jeffrey P Mower. Prep-mt: predictive rna editor for plant mitochondrial genes. *BMC bioinformatics*, 6:1–15, 2005.
33. Ian T Joliffe and BJT Morgan. Principal component analysis and exploratory factor analysis. *Statistical methods in medical research*, 1(1):69–95, 1992.
34. Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
35. Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 2011.
36. Charles Swanton. Intratumor heterogeneity: evolution through space and time. *Cancer research*, 72(19):4875–4882, 2012.
37. Derek W Barnett, Erik K Garrison, Aaron R Quinlan, Michael P Strömberg, and Gabor T Marth. Bamtools: a c++ api and toolkit for analyzing and managing bam files. *Bioinformatics*, 27(12):1691–1692, 2011.
38. Gavin Ha, Andrew Roth, Daniel Lai, Ali Bashashati, Jiarui Ding, Rodrigo Goya, Ryan Giuliany, Jamie Rosner, Arusha Oloumi, Karey Shumansky, et al. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome research*, 22(10):1995–2007, 2012.
39. Zhenhua Yu, Yuanning Liu, Yi Shen, Minghui Wang, and Ao Li. Climat: accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data. *Bioinformatics*, 30(18):2576–2583, 2014.

Zhenhua Yu is an associate professor at School of Information Engineering, Ningxia University. His research interests include bioinformatics, computational biology and deep learning.

Furui Liu is a Master's student at School of Information Engineering, Ningxia University. His research interests include bioinformatics and cancer genomics.

Fangyuan Shi is an assistant professor at School of Information Engineering, Ningxia University. Her research interests include bioinformatics and computational biology.

Fang Du is a professor at School of Information Engineering, Ningxia University, Yinchuan, China. Her research interests include biomedical big data analysis and knowledge graph.