OXFORD

Briefings in Bioinformatics, 2024, **25(4)**, bbae277

https://doi.org/10.1093/bib/bbae277

Problem Solving Protocol

LoRA-TV: read depth profile-based clustering of tumor cells in single-cell sequencing

Junbo Duan*, Xinrui Zhao, Xiaoming Wu

*Corresponding author. Key Laboratory of Biomedical Information Engineering of Ministry of Education and Department of Biomedical Engineering, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China. E-mail: junbo.duan@mail.xjtu.edu.cn

Abstract

Single-cell sequencing has revolutionized our ability to dissect the heterogeneity within tumor populations. In this study, we present LoRA-TV (Low Rank Approximation with Total Variation), a novel method for clustering tumor cells based on the read depth profiles derived from single-cell sequencing data. Traditional analysis pipelines process read depth profiles of each cell individually. By aggregating shared genomic signatures distributed among individual cells using low-rank optimization and robust smoothing, the proposed method enhances clustering performance. Results from analyses of both simulated and real data demonstrate its effectiveness compared with state-of-the-art alternatives, as supported by improvements in the adjusted Rand index and computational efficiency.

Keywords: tumor cells; single-cell sequencing; clustering; read depth profile; robust smoothing; low-rank approximation

Introduction

It is well known that copy number variation (CNV) is a subtype of unbalanced structural variation (SV), and involves gain or loss of genetic segments of size more than 1 kbp [1],[2], resulting in an abnormal number of copies of specific genes or noncoding genomic regions. CNV was reported to be discovered frequently in both human and other mammal genomes, and has been associated with complex diseases such as cancer [3], schizophrenia [4], Alzheimer disease [5], *etc.* CNVs are frequently observed in tumor cells, and play a crucial role in the progression of cancer by promoting genomic instability and disrupting key genes involved in cell growth regulation [6, 7].

Read depth profile is a traditional signature for CNV detection [8, 9]. By splitting whole reference genome into fix-sized nonoverlapping bins [10, 11], variable-sized bins [12] or a sliding window [8], and counting the number of whole-genome sequencing reads mapped within each bin, the read depth profile can be obtained, and consecutive bins with significant high/low read depth are identified as CNV gain/loss, respectively. By counting the number of whole-exome sequencing reads mapped within each exonic bins, CNV can also be inferred. Furthermore, read depth profile, with the terminology expression level [13], was also used in RNA-seq as a signature to describe gene expression quantitatively.

Single-cell sequencing (SCS) has emerged as a powerful technique to study the heterogeneity of cellular populations with unprecedented resolution, and hence an important tool for investigations in cancer, developmental biology [14–16] *etc.* Compared with the previous bulk sequencing, SCS enables the detection of CNVs in individual tumor cells, and hence provides valuable insights into the underlying molecular mechanisms of cancer. Navin *et al.* [17] explored the use of SCS techniques to study the evolution of tumors. The authors highlight the heterogeneity and complexity of tumors, and the limitations of bulk sequencing methods in capturing the full spectrum of genetic changes. By analyzing individual cancer cells, they reveal the clonal dynamics and genetic diversity within tumors, providing insights into tumor evolution, metastasis and potential therapeutic targets.

The processing and analysis of SCS data present significant challenges due to the complexity of the data. (i) The huge volume of data generated from SCS presents a computational burden. The high-throughput nature of the SCS technology results in millions of reads per cell, leading to massive datasets that require substantial computational resources for data processing; (ii) the presence of technical noise, such as amplification bias that related to GC-content inhomogeneity, demands careful bias correction; (iii) the heterogeneity of single cell populations adds another factor of

Junbo Duan received the BS degree in information engineering and the MS degree in communication and information system from the Xi'an Jiaotong University, Xi'an, China, in 2004 and 2007, respectively, and the PhD degree in signal processing from the Université Henry Poincaré, Nancy, France, in 2010. After graduation, he was a postdoctoral fellow at the Department of Biomedical Engineering and Bio- statistics and Bioinformatics, Tulane University, USA, until 2013. He iscurrently an associate professor at the Department of Biomedical Engineering, Xi'an Jiaotong University.

Xinrui Zhao received the BS degree in biological science from the Inner Mongolia University in 2018. She is currently pursuing the doctor's degree from the School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, Shannxi, China. Her research interests include bioinformatics and molecular biology.

Xiaoming Wu is an associate professor at the School of Life Science and Technology at Xi'an Jiaotong University since 2013. With a PhD in biomedical engineering, his research interests include biomedical engineering and bioinformatics.

Received: April 1, 2024. **Revised:** May 17, 2024. **Accepted:** May 29, 2024

complexity to data processing. To tackle these challenges, the field of SCS data analysis has witnessed remarkable advancements. Novel open-source packages and pipelines have been developed to facilitate data preprocessing and downstream analysis. [12] presents a protocol for performing genome-wide copy number analysis at the single-cell level. The authors proposed a step-by-step procedure for isolating single cells, amplifying their DNA and using next-generation sequencing to assess copy number profiles. The protocol offers a comprehensive approach to study CNVs in heterogeneous cell populations and has implications in understanding genomic instability, tumor heterogeneity and genetic diseases at the single-cell resolution. [18] focuses on the development of an interactive analysis framework, and proposed Ginkgo, a user-friendly, open-source web platform for the analysis of single-cell CNVs (http://qb.cshl.edu/ginkgo). *Seurat* package [19] and the *Monocle* package [20, 21] are two popular bioinformatics toolkits for SCS data analysis. Both packages provide a range of functions and methods for tasks such as quality control, normalization, dimensionality reduction, clustering, differential expression analysis and visualization of single-cell data.

Given that tumor cells from the same subclone share common signatures, such as identical CNVs on tumor suppressor genes and aneuploidy, it is crucial to consider these shared features in the analysis. However, traditional SCS tools typically analyze the read depth profiles of each cell individually before clustering cells into subclones, overlooking these common signatures. The work by Navin *et al.* [17] addressed this issue by detecting common breakpoints before clustering, although the approach was somewhat empirical. In this paper, we propose LoRA-TV (Low Rank Approximation with Total Variation), a model to process the read-depth profiles of tumor cells from SCS jointly to aggregate common signature dispersed among individual cells. In the proposed model, the read-depth profile of a single cell is stored in a column vector, and vectors of all cells are cascaded horizontally to form a matrix. Since profiles suffer from high fluctuation caused by sequencing, in order to smooth profile and capture common signature, robust smoothing and low-rank approximation are introduced. In our proposed model, robust smoothing is established with modern techniques involving total variation [22] and L-1 norm optimization [23] to reduce the impact of outliers and extreme values. Low-rank approximation [24] is a mathematical technique that can capture essential features while reducing dimensionality, and is widely used in data compression, noise reduction, machine learning, *etc.* Since optimization involving matrix rank is combinatorially complicated and intractable to solve in general, the rank of a matrix is frequently relaxed to its nuclear norm, which is the sum of its singular values [25]. As a result, singular value decomposition (SVD) is employed. The detailed model is presented in the sequel.

Methods

Model

The LoRA-TV model is as follows:

$$f(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\|_F^2 + \lambda\|\boldsymbol{DX}\|_1 + \mu\|\boldsymbol{X}\|_*, \tag{1}$$

where $\boldsymbol{Y} \in \mathbb{R}^{N\times M}$ is a matrix containing the read-depth profiles of M cells; each profile is of length N. Detailed descriptions of how $\boldsymbol{Y}$ was organized are presented in Section 3.1, while the process of extracting $\boldsymbol{Y}$ from real SCS data is outlined in Section 4; $\boldsymbol{X} \in \mathbb{R}^{N\times M}$ is the denoised version of $\boldsymbol{Y}$; $\boldsymbol{D} \in \mathbb{R}^{(N-1)\times N}$ is the first-order difference operator, which has a Toeplitz structure with main diagonal elements $D_{i,i} = -1$ and upper diagonal elements $D_{i,i+1} = 1$ [26]; $\|\|_F, \|\|_1, \|\|_*$, are the Frobenius norm, ℓ_1 norm and nuclear norm [25], respectively; the first term is the data fitting fidelity term, the second term is total variation used to smooth the profile of each cell while preserve CNVs [22] and the last term is the convex relaxation of matrix rank [27], which is used to concentrate cluster of cells; hyperparameters λ and μ are used to balance the tradeoff of smoothness and low-rank, respectively. By minimizing $f(\boldsymbol{X})$ for given $\boldsymbol{Y}, \boldsymbol{D}, \lambda$ and μ, the minimizer $\boldsymbol{X}$ is the refined read-depth profiles of cells that promise better clustering.

Optimization

Following the standard alternating direction method of multipliers (ADMM) [28], above optimization problem (1) is decomposed into following three iterations over k:

$$\boldsymbol{X}^{k+1} = \arg\min_{\boldsymbol{X}} \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\|_F^2 + \mu\|\boldsymbol{X}\|_* + \frac{\beta}{2}\|\boldsymbol{DX} - \boldsymbol{Z}^k - \boldsymbol{W}^k\|_F^2 \tag{2}$$

$$\mathbf{Z}^{k+1} = \arg\min_{\mathbf{Z}} \frac{\beta}{2}\|\boldsymbol{DX}^{k+1} - \boldsymbol{W}^k - \boldsymbol{Z}\|_F^2 + \lambda\|\boldsymbol{Z}\|_1 \tag{3}$$

$$\boldsymbol{W}^{k+1} = \boldsymbol{W}^k - (\boldsymbol{DX}^{k+1} - \boldsymbol{Z}^{k+1}) \tag{4}$$

Variable $\boldsymbol{W}$ is the Lagrange multiplier associated with the consensus constraint $\boldsymbol{DX} = \boldsymbol{Z}$, which is introduced to encourage the primal variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ to satisfy the equality.

β is a positive penalty parameter that controls the trade-off between fitting the data and enforcing the equality constraint.

Sub-optimization problem (3) is trivial, which has closed-form solution

$$\boldsymbol{Z}^{k+1} = TH_{\frac{\lambda}{\beta}}^{soft}(\boldsymbol{DX}^{k+1} - \boldsymbol{W}^k), \tag{5}$$

where $TH_{\lambda/\beta}^{soft}$ is the soft thresholding function with cutoff value λ/β applying on each element of the matrix [29].

Sub-optimization problem (2) further reads

$$\boldsymbol{X}^{k+1} = \arg\min_{\boldsymbol{X}} \frac{1}{2}\left\|\begin{bmatrix}\boldsymbol{Y}\\ \beta(\boldsymbol{Z}^k + \boldsymbol{W}^k)\end{bmatrix} - \begin{bmatrix}\boldsymbol{I}\\ \beta\boldsymbol{D}\end{bmatrix}\boldsymbol{X}\right\|_F^2 + \mu\|\boldsymbol{X}\|_* \tag{6}$$

$$= \arg\min_{\boldsymbol{X}} \frac{1}{2}\|\boldsymbol{B}^k - \boldsymbol{AX}\|_F^2 + \mu\|\boldsymbol{X}\|_*, \tag{7}$$

which can be tackled with singular value thresholding [27] and iterative thresholding [29] over l:

$$\boldsymbol{F}^l = (\boldsymbol{I} - \boldsymbol{A}^T\boldsymbol{A})\boldsymbol{X}^l + \boldsymbol{A}^T\boldsymbol{B}^k \tag{8}$$

$$\boldsymbol{V}_1^l\boldsymbol{\Sigma}^l(\boldsymbol{V}_2^l)^T = SVD(\boldsymbol{F}^l) \tag{9}$$

$$\boldsymbol{\Sigma}^{l+1} = TH_\mu^{soft}(\boldsymbol{\Sigma}^l) \tag{10}$$

$$\boldsymbol{X}^{l+1} = \boldsymbol{V}_1^l\boldsymbol{\Sigma}^{l+1}\boldsymbol{V}_2^{l^T} \tag{11}$$

Table 1 shows the pseudo code, and the Matlab implementation is available online with URL in the conclusion.

Simulation study

Date simulation

To test the performance of the proposed model, we first generated an SCS dataset *in silico* including the read depth profiles of $M = 64$ cells to mimic tumor evolution.

Table 1. Algorithm pseudo code.

Input: Data $\boldsymbol{Y}$ and parameters $\lambda, \mu, \beta, s, t_1, t_2$.

Initialize $\boldsymbol{Z}, \boldsymbol{W}$ as full zero matrices.
while until meet relative tolerance t_1

$$\bar{\boldsymbol{B}} = \begin{bmatrix} \frac{1}{s}\boldsymbol{Y} \\ \frac{\beta}{s}(\boldsymbol{Z}+\boldsymbol{W}) \end{bmatrix},$$
$$\bar{\boldsymbol{A}} = \begin{bmatrix} \frac{1}{s}\boldsymbol{I} \\ \frac{\beta}{s}\boldsymbol{D} \end{bmatrix},$$
$$\boldsymbol{E} = \boldsymbol{I} - \bar{\boldsymbol{A}}^T\bar{\boldsymbol{A}},$$
$$\boldsymbol{C} = \bar{\boldsymbol{A}}^T\bar{\boldsymbol{B}},$$
$$\boldsymbol{X} = \bar{\boldsymbol{A}}^+\bar{\boldsymbol{B}},$$

while until meet relative tolerance t_2

$$\boldsymbol{F} = \boldsymbol{E}\boldsymbol{X} + \boldsymbol{C},$$
$$\boldsymbol{V}_1\boldsymbol{\Sigma}\boldsymbol{V}_2^T = SVD(\boldsymbol{F}),$$
$$\boldsymbol{\Sigma} = TH^{soft}_{\mu/s^2}(\boldsymbol{\Sigma}),$$
$$\boldsymbol{X} = \boldsymbol{V}_1\boldsymbol{\Sigma}{\boldsymbol{V}_2}^T,$$

end while

$$\boldsymbol{Z} = TH^{soft}_{\frac{\lambda}{\beta}}(\boldsymbol{D}\boldsymbol{X} - \boldsymbol{W}),$$
$$\boldsymbol{W} = \boldsymbol{W} - (\boldsymbol{D}\boldsymbol{X} - \boldsymbol{Z}),$$

end while
Output: $\boldsymbol{X}$.

- CNV profiles evolution: Half cells were labeled as normal (with label A1 to A32), *i.e.* the copy number was assigned as 2 to represent diploid. The left half cells were assigned as abnormal (B1 to B16, C1 to C8, D1 to D4 and E1 to E4). Cells in cluster B share a CNV with random location, length following a uniform distribution with boundary of minimal 3 and maximal 10 bins and a random CNV status (0 homozygous deletion, 1 heterozygous deletion, 3 heterozygous duplication and 4 homozygous duplication). Cells in cluster C share the same CNV in cluster B, and one more random CNV. Cells in clusters D and E follow the same procedure.
- Read depth profiles generation: Since SCS read depth profiles are commonly modeled as random variables obeying Poisson or Negative Binomial distributions [10, 30, 31], in this work we sampled $N = 300$ i.i.d. random numbers following Poisson distribution with density parameter (or mean) 10 for each normal cell. For abnormal cells incorporating CNVs, read depth profiles were simulated according to its CNV profile, and Poisson density parameter was set as 1, 5, 15 and 20 for CNV status 0, 1, 3 and 4, respectively.

Figure 1 showcases example read depth profiles for each cell cluster. In normal cells A1 and A2, the average read depth is 10. Cells B1 and B2 exhibit a CNV deletion of approximately 30 bins near coordinate 150. Cells C1 and C2 show a CNV duplication of approximately 10 bins near coordinate 50. Cells D1 and D2 display a CNV deletion of approximately 30 bins near coordinate 270. Cells E1 and E2 present a CNV duplication of approximately 20 bins near coordinate 230.

Numerical optimization

The simulated dataset is a matrix of $N = 300$ rows by $M = 64$ columns, and each column represents the read depth profile of a cell. This matrix was then input as $\boldsymbol{Y}$ in Eq. (1), and optimization procedure was executed.

As the proposed algorithm is an iterative optimization method, four issues needs attention:

- β: Since ADMM was employed for optimization, the penalty parameter β plays a crucial role in the algorithm. β not only controls the trade-off between data fitting and equality constraint enforcing, *i.e.* a larger β places more emphasis on satisfying the constraint, leading to a more accurate fulfillment of the equality, but also impacts the convergence behavior and stability of the algorithm; an appropriate value can improve the numerical properties and convergence speed of the algorithm. By performing a grid search over 1e-2 to 1e2, Fig. 2 shows that the value 2 works best for the simulated dataset.
- Initialization: For the ADMM iterates, $\boldsymbol{Z}$ and $\boldsymbol{W}$ were initialized as zero matrices; for the singular value thresholding iterates, $\boldsymbol{X}$ was initialized with $\boldsymbol{A}^+\boldsymbol{B}^k$, where $^+$ is the Moore–Penrose inverse.
- Stopping criterion: Relative tolerance was employed in the experiments. Iteration terminates if the relative change in the objective function $\frac{f^k - f^{k+1}}{f^k} < 1e-3$. Figure 3 shows a typical plot of objective versus iterations, indicating that 20 iterations are sufficient for convergence, taking approximately 8 s on a Windows desktop.
- Scale of $\boldsymbol{A}$: It is known that for iterative thresholding methods, convergence to a local minimum is guaranteed if the operator norm (or ℓ_2 norm $\|\boldsymbol{A}\|_2$ [25]) of the matrix $\boldsymbol{A}$ in Eq. (7) is less than one [25]. For our optimization problem, the operator norm is larger than one, therefore we have to scale the objective to meet this requirement. To be more specific, suppose $s > \|\boldsymbol{A}\|_2$ is a scale factor opted by user, then the optimization problem (7) is equivalent to its modification by replacing $\boldsymbol{B}^k$, $\boldsymbol{A}$ and μ with $\bar{\boldsymbol{B}}^k = \boldsymbol{B}^k/s$, $\bar{\boldsymbol{A}} = \boldsymbol{A}/s$ and $\bar{\mu} = \mu/s^2$, respectively. For the simulated dataset, $\|\boldsymbol{A}\|_2 = 4.1$, so we chose $s = 10$.

Phylogenetic tree

After the above optimization procedure, optimizer matrix $\boldsymbol{X}$ was output, which is a denoised and low-rank approximated version of original data $\boldsymbol{Y}$. Then we calculated the pairwise Euclidean distances between the columns of $\boldsymbol{X}$, and performed hierarchical clustering (neighbor-joining algorithm) [32] based on the distance matrix. The result was visualized as a phylogenetic tree. An example phylogenetic tree is shown in Fig. 4(A), showing that cell lines A to E are clustered successfully. As a comparison, panel (B) also shows the phylogenetic tree of matrix $\boldsymbol{Y}$. The phylogenetic structure remains unclear without the application of LoRA-TV processing.

Since the simulated dataset has a ground true label, we can evaluate the clustering result quantitatively. The Rand index is a measure used to assess the similarity between two data clusters [33]. It is often employed when evaluating the performance of a clustering algorithm. For two clusters of a set of elements, the Rand index measures the percentage of pairs of elements that are either in the same cluster or in different clusters. The Rand index is expressed as a value between 0 and 1, where 0 indicates no similarity, and 1 indicates identical clusterings. We employed the adjusted Rand index (ARI) [34], which is a variation of the original one that corrects for the expected value of the index under random clustering.

The parameters λ and μ significantly impact clustering performance, so we fine-tuned these two parameters using ARI as the criterion. A two-dimensional grid search on a logarithm scale was employed, with 1000 Monte-Carlo replications for each grid point. The mean ARI results are demonstrated in Fig. 5. It is shown that grid with $\lambda = 3.2$ *and* $\mu = 32$ achieves the highest ARI for the simulated dataset. The red curves in Fig. 1 are the denoised read depth signals with this parameter configuration.

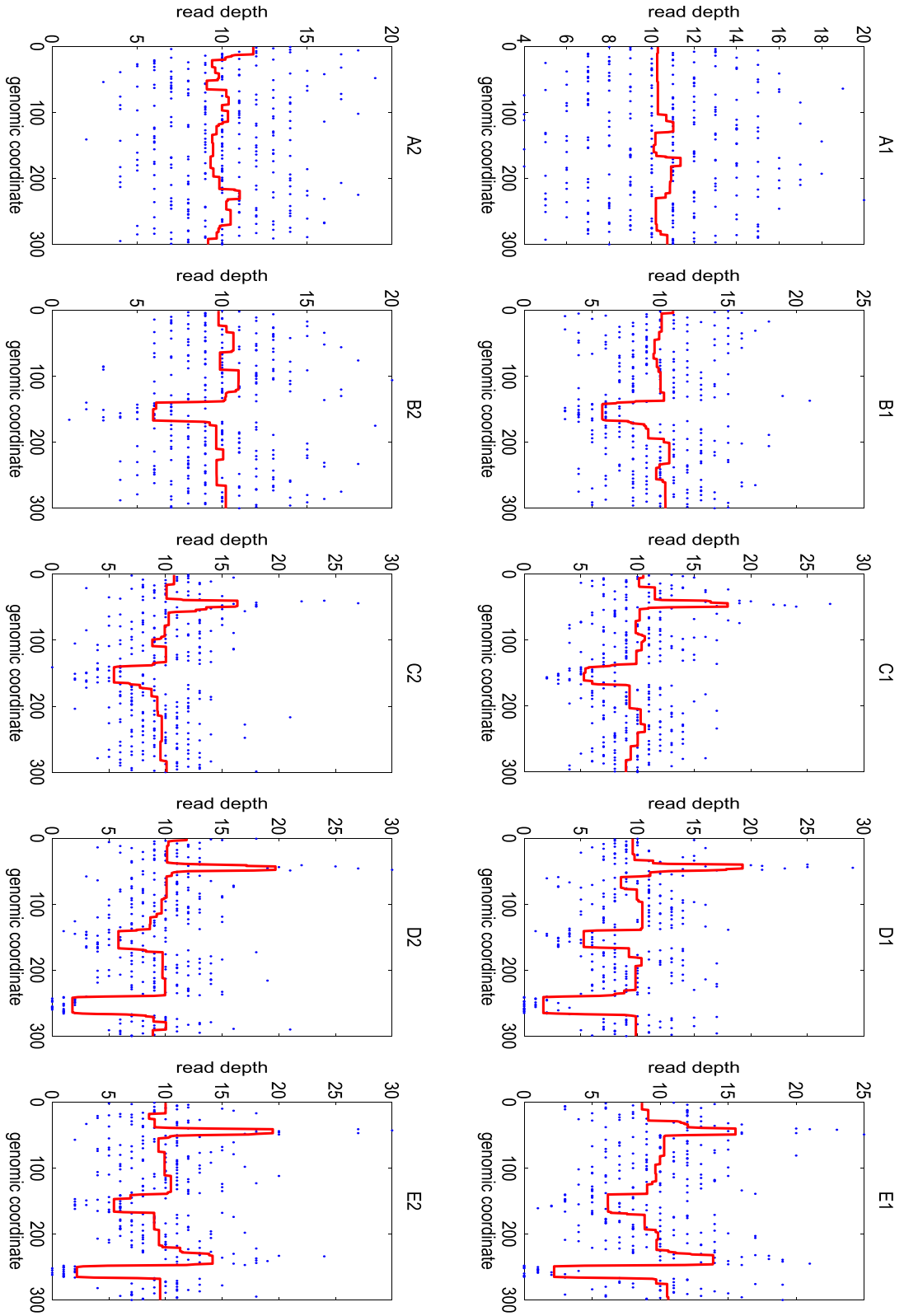

Figure 1. Read depth of five (A to E) simulated cell lines. Each row displays one cell and each column display one cell lines.

Comparisons

The proposed LoRA-TV method was compared with both existing bioinformatics and optimization toolkits.

Bioinformatics toolkits

There are several methods/toolkits/packages that are developed for CNV detection and cell clustering based on read depth profile

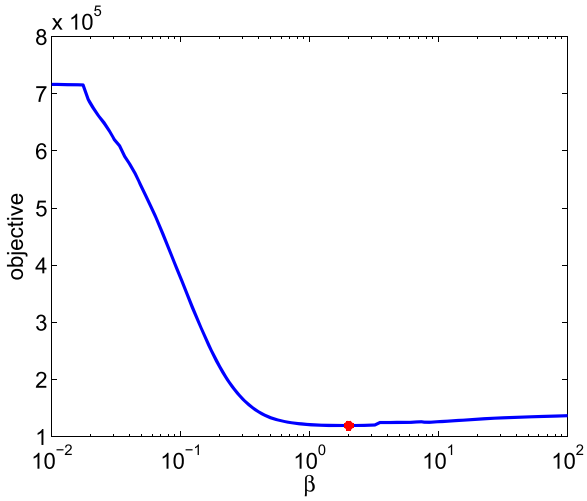

Figure 2. The minimal objective value with respect to ADMM parameter β. The red dot represents the lowest value, which is used in the sequel.

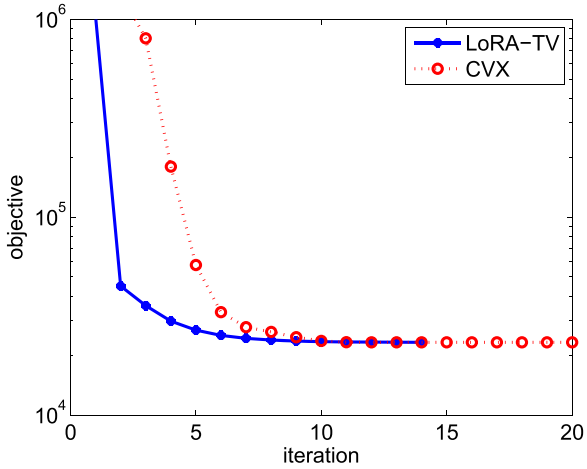

Figure 3. An example objective value versus iteration curve.

from SCS, and we list representative ones in Table 2. We also compared LoRA-TV with state-of-the-art alternatives, namely, the *Seurat* package [19] and the *Monocle* package [20, 21], which are popular bioinformatics toolkits for SCS data analysis. These packages provide a range of functions and methods for tasks such as quality control, normalization, dimensionality reduction, clustering, differential expression analysis and visualization of single-cell data.

To call the Seurat package, firstly a seurat object was created with $\boldsymbol{Y}$ as the count matrix; then functions *NormalizeData, FindVariableFeatures, ScaleData, RunPCA, FindNeighbors and FindClusters* were called consecutively to cluster cells; finally the ARI was calculated. Default parameters were used, except *resolution* in *FindClusters*, which is crucial for the number of clusters, and was set as 1.1 empirically; otherwise, the clustering performance of Seurat is notably poor. For the Monocle package, firstly, a CellDataSet object was created (using *new_cell_data_set* function) with $\boldsymbol{Y}$ as the expression matrix; then functions *preprocess_cds, reduce:dimension and cluster_cells* were called to cluster cells and the ARI was calculated finally. Principle component analysis and uniform manifold approximation and projection were used in preprocessing and dimension reduction processes, respectively. Both Louvain and Leiden algorithms were tested [37].

Table 2. Representative methods for CNV detection and cell clustering based on read depth profile from SCS

Tools	Reference	Feature & website
Seurat	[19]	A R toolkit for single cell genomics. (https://satijalab.org/seurat/)
Monocle3	[20, 21]	An analysis toolkit for SCS. (https://cole-trapnell-lab.github.io/monocle3/)
Ginkgo	[12, 18]	A cloud-based single-cell CNV analysis tool. (http://qb.cshl.edu/ginkgo/)
SCONCE	[35]	A package for profiling CNV in cancer evolution using SCS. (https://github.com/NielsenBerkeleyLab/sconce)
CaSpER	[36]	A toolkit for identifying CNV events by integrative analysis of SCS. (https://github.com/akdess/CaSpER)

A total of 1000 Monte-Carlo replicates were conducted, resulting in the proposed method achieving a mean score of 0.77 with a standard deviation of 0.24 (0.77±0.24) on the aforementioned five-subclone dataset. As a comparison, Seurat default (Seurat–Louvain) got an average ARI 0.63, with standard deviation 0.21. By switching Louvain to Leiden algorithm, Seurat (Seurat–Leiden) got an average ARI 0.61, with standard deviation 0.22. The results of Monocle3 with Louvain and Leiden algorithms (Monocle3-Louvain and Monocle3-Leiden) are also shown in Table 3. To have a detailed distribution of ARI, Fig. 6 further shows the histograms of ARIs of methods. It shows in panel (a) that most ARIs of the proposed methods locate near to 1, while in (b) the ARIs of Seurat–Louvian concentrate between 0.6 and 0.7, which is the case subclones D and E are mixed; in (c) the ARIs of Monocle3-Leiden locates almost uniformly.

Above performance evaluation was based on five-subclone dataset, and in order to have a more comprehensive evaluation, one- to four-subclone dataset were analyzed. For example, four-subclone dataset was generated by eliminating subclone E in the subsection 3.1, *etc.* ARIs are listed in Table 3. It is shown that the proposed method outperformed Seurat and Monocle3 packages for all numbers of subclones, indicating the superior performance for a broad spectrum. Note that for one subclone clustering task, both Seurat and Monocle3 packages failed, which is aligned with the conclusion of previous publication [38].

Optimization toolkit

LoRA-TV was also compared with existing optimization toolkit *CVX* [39], which is a versatile toolkit for disciplined convex programs, offering solutions for linear/quadratic programs, second-order cone programs and semidefinite programs. Implemented in Matlab, CVX simplifies optimization problem formulation and solution, and hence a popular toolkit for numerical optimization.

We compared the proposed ADMM algorithm in Table 1 with CVX in solving the objective in Eq. (1). Both codes were run on a Windows desktop with an Intel i7-3770 CPU and 32 gigabytes of memory. Figure 3 shows a typical objective versus iterate plot. It shows that LoRA-TV iterated 14 loops and met stopping criterion. The CVX toolkit was also called to solve the same problem (same data and hyperparameters) on the same computer, and the objective iterate is displayed as the red curve, which converged to an approximate objective value as LoRA-TV, and exited with 20 iterates. In comparison for the runtime computation resources consumption, LoRA-TV took 8 s and used a megabyte memory, while the CVX package processed 124 031 variables and 38 400

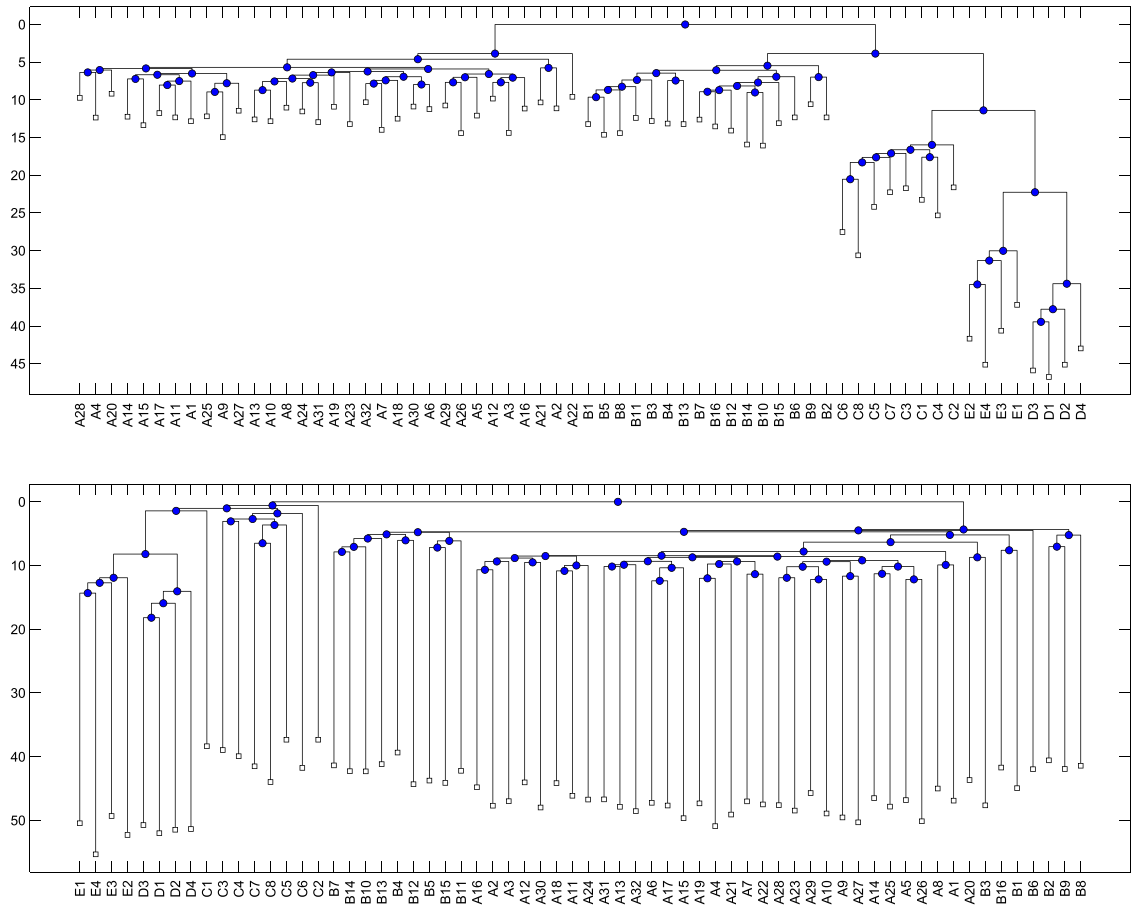

Figure 4. Typical phylogenetic trees of simulated dataset. The upper and lower trees are the results with and without LoRA-TV processing. It is shown in the upper panel that with LoRA-TV five cell lines (A to E) are well clustered.

Table 3. ARI performance comparison on simulated data

Method\subclone	1	2	3	4	5
Proposed	1.00±0.00	0.76±0.42	0.78±0.27	0.79±0.22	0.77±0.24
Seurat-Louvain	0.31±0.46	0.77±0.40	0.74±0.21	0.61±0.21	0.63±0.21
Seurat-Leiden	0.19±0.39	0.80±0.36	0.72±0.22	0.64±0.18	0.61±0.22
Monocle3-Louvain	0.00±0.00	0.14±0.05	0.24±0.07	0.21±0.06	0.23±0.08
Monocle3-Leiden	0.00±0.03	0.30±0.15	0.50±0.28	0.34±0.16	0.49±0.27
None	1.00±0.00	0.32±0.47	0.38±0.46	0.43±0.31	0.46±0.30

equality constraints, and hence took an hour and used 29 gigabytes memory at peak.

To further demonstrate the computational performance of LoRA-TV, Fig. 7 shows the resources consumption with respect to varying scales of read depth length N and cell number M; each point is the average of 10 Monte-Carlo experiments. It is shown that with the increase of N, memory usage increases steadily, and CPU time increases significantly, while with the increase of M, memory usage increases significantly, and CPU time increases slowly. For large-scale problem (10 000 of cells and read depths), LoRA-TV cost approximately an hour and a gigabytes, demonstrating its effectiveness and generalization ability.

Real data study

Two real SCS dataset were processed to test the performance of the proposed method. The datasets were sampled from

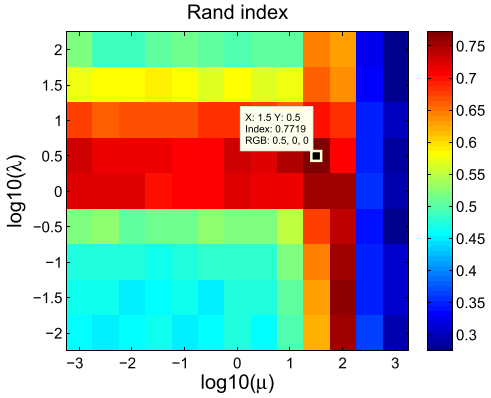

Figure 5. Tuning of parameters λ and μ. 1000 Monte-Carlo replications were tested, and the best configuration is shown.

Table 4. Cell labels and SRR ID in T10, T16P and T16M dataset

T10				T16P		T16M	
label	SRR	label	SRR	label	SRR	label	SRR
D1	052047	H9	054566	P1	089523	D2	089706
D2	052148	H10	054567	P2	089526	D5	089709
D4	053437	H11	054568	P3	089529	D6	089710
D5	053600	H12	054569	P4	089533	D7	089711
D6	053602	H13	054570	P5	089542	D8	089712
D3	053604	H14	054571	P6	089550	D9	089713
D7	053605	H15	054572	P7	089561	D10	089714
D8	053606	H16	054573	P8	089564	D11	089715
D9	053607	H17	054574	P9	089568	D12	089716
D10	053608	H18	054575	P10	089573	D13	089717
D11	053609	H19	054576	P11	089577	D14	089719
D12	053610	H20	054577	P12	089578	D15	089722
D13	053611	H21	054578	D1	089580	D16	089730
D14	053615	AA1	054592	D2	089583	D17	089731
D15	053616	AA2	054594	D3	089586	D3	089733
D16	053617	AA3	054596	D4	089589	D21	089750
D17	053618	AA4	054597	D5	089591	D22	089751
D18	053619	AA5	054598	D6	089592	D23	089752
D19	053620	AA6	054599	D7	089593	D24	089753
D20	053623	AA7	054600	D8	089594	D25	089754
D21	053624	AA8	054601	D9	089595	D26	089755
D22	053629	AA9	054602	D10	089596	A1	089756
D23	053630	AA10	054603	D11	089597	A2	089757
D24	053631	AA11	054604	D12	089598	A3	089817
D25	053632	AA12	054605	D13	089599	A4	090126
D26	053633	AA13	054606	D14	089600	A5	090129
D27	053634	AA14	054608	D15	089601	A6	090130
D28	053635	AA15	054609	D16	089602	A7	090131
D29	053636	AA16	054610	D17	089603	A8	090133
D30	053637	AA17	054611	D18	089604	A10	090142
D31	053638	AA18	054612	D19	089605	A11	090144
D32	053639	AA19	054613	D20	089606	A12	090155
D33	053666	AA20	054614	D21	089607	A13	090156
D34	053667	AA21	054615	D22	089608	A14	090158
P1	053668	AA22	054616	D23	089609	A15	090159
P2	053669	AA23	054618	D24	089610	A16	090198
P3	053670	AA24	054620	A1	089646	A17	090206
P4	053671	AA25	054622	A2	089659	A18	090209
P5	053672	AB1	054626	A3	089662	A19	090210
P6	053673	AB2	054632	A4	089663	A20	090211
P7	053674	AB3	054633	A5	089664	A21	090212
P8	053675	AB4	054634	A6	089665	A22	090213
H1	053676	D37	089377	A7	089666		
H2	053677	D38	089378	A8	089694		
H3	053678	D39	089379	A9	089695		
H4	053679	D36	089397	A10	089696		
H5	053680	H22	089401	A11	089697		
H6	053681	H23	089402	A12	089698		
H7	054213	H24	089403	A13	089699		
H8	054565			A14	089700		
				A15	089701		
				A16	089702		

polygenomic breast tumors, particularly two high-grade, triple-negative ductal carcinomas (T10 and T16P) and a paired metastatic liver carcinoma (T16M) [17].

The T10 dataset contains 63 normal and 37 tumor cells, with infiltrating leukocytes. Five major ploidy distributions are obtained by fluorescence-activated cell sorting, and these cells are labeled with Diploid (D), Pseudodiploid (P), Hypodiploid (H), Aneuploid A (AA) and Aneuploid B (AB). The T16P and T16M dataset contains 52 and 48 cells, respectively, and we combined them into T16PM dataset.

The SCS data for each cell were downloaded from National Center for Biotechnology Information Sequence Read Archive server (https://www.ncbi.nlm.nih.gov/sra). The SRR IDs and labels are listed in Table 4. Note that one cell in T10 and six cells in T16M have no labels, so the data of those cells were not included in our analysis.

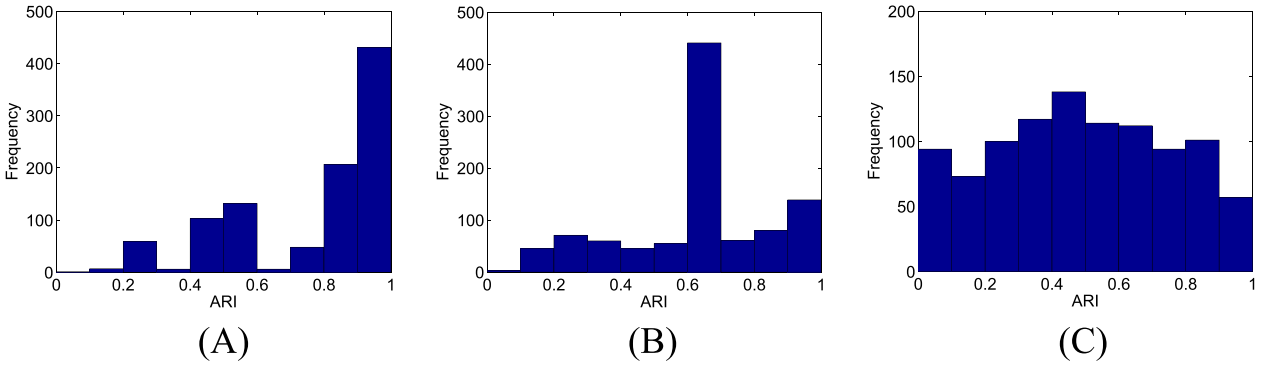

Figure 6. Distributions of ARIs of 1000 Monte-Carlo experiments. (A) proposed method, (B) Seurat-Louvain, (C) Monocle3-Leiden.

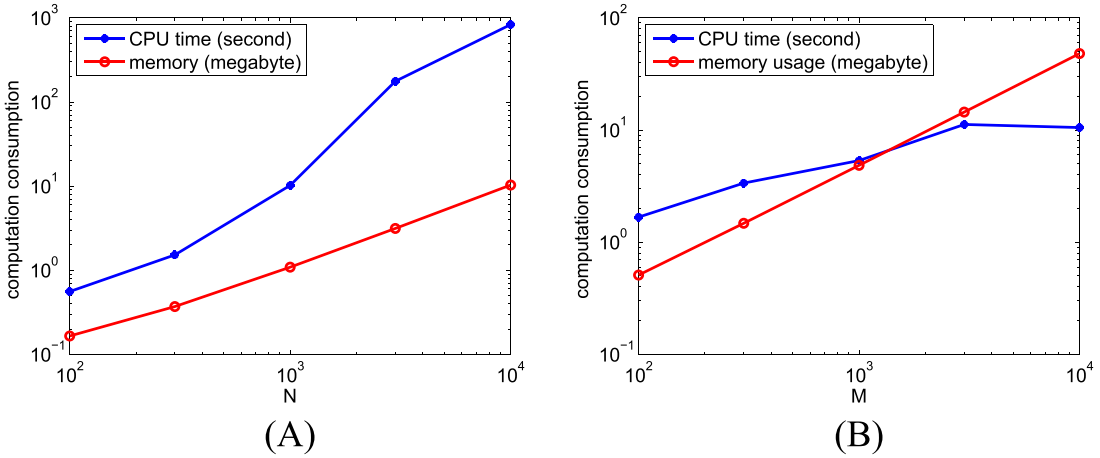

Figure 7. Computation consumptions of CPU time and memory (A) with respect to read depth length N while M is fixed to 64; (B) with respect to cell number M while N is fixed to 300.

The processing pipeline is as follows, and preprocessing described in [12] is referred:

1. Map each fastq file to human reference genome hg19 with *Bowtie2* [40], yielding an SAM file;
2. Convert and sort each SAM file to a BAM file with *samtools* [41];
3. Calculate a read depth profile of each cell from BAM file with *bedtools* [42]; chromosomal bin regions are defined in *hg19.varbins.bed* [12];
4. Normalize each read depth profile by dividing the median value, yielding the median of normalized profile to 1;
5. Smooth read depth *versus* GC content in logarithm scale curve with Lowess [43];
6. Correct GC content bias in logarithm scale;
7. Process the corrected read depth profiles in logarithm scale with LoRA-TV;
8. Calculate pairwise Euclidean distances of profiles;
9. Draw phylogenetic tree using neighbor-joining method [32];
10. Evaluate ARI [34].

Figure 8 shows the phylogenetic tree of T10 dataset. It is demonstrated that four subpopulations have been identified and clustered as follows: D and P, H, AA and AB. Within the first cluster, pseudodiploid cells gradually diverge from diploid cells but remain within the same cluster. As evolution progresses, hypodiploid cells form the second cluster. Ultimately, aneuploid A and B cells constitute the last two clusters. The clustering results are nearly flawless, with the exception of SRR054604 labeled as AA11, which was clustered into AB, resulting in an ARI of 0.987. Further investigation reveals that this outcome aligns with the findings on the Ginkgo website [18](http://qb.cshl.edu/ginkgo/). Bootstrap was employed to estimate the statistical significance of the clustering result, 1000 sampling with replacement was carried out. The reported P-value of 0.013 indicates that the clustering result is statistically significant, and is unlikely to have occurred by random chance alone, and suggests that there is a meaningful structure in the data that leads to the observed clustering pattern.

Figure 9 shows the phylogenetic tree of T16PM dataset. It is evident that three subpopulations have been clustered: A-P, A-M and D and P. The first cluster comprises aneuploid cells from the T16P dataset, the second cluster consists of aneuploid cells from the T16M dataset and the third cluster comprises other normal cells. The statistical significance of the clustering result is 1e-3.

Conclusion and discussion

The processing of SCS data is a complex task that requires specialized computational methods and careful consideration of various sources of variability. In this paper, we proposed LoRA-TV, a model that can cluster tumor cells based on read depth profiles. Demonstrations on both simulated and real data show that LoRA-TV achieved better ARI over its alternatives, and support that the proposed method can illustrate the evolutionary branches and relationships between subpopulations.

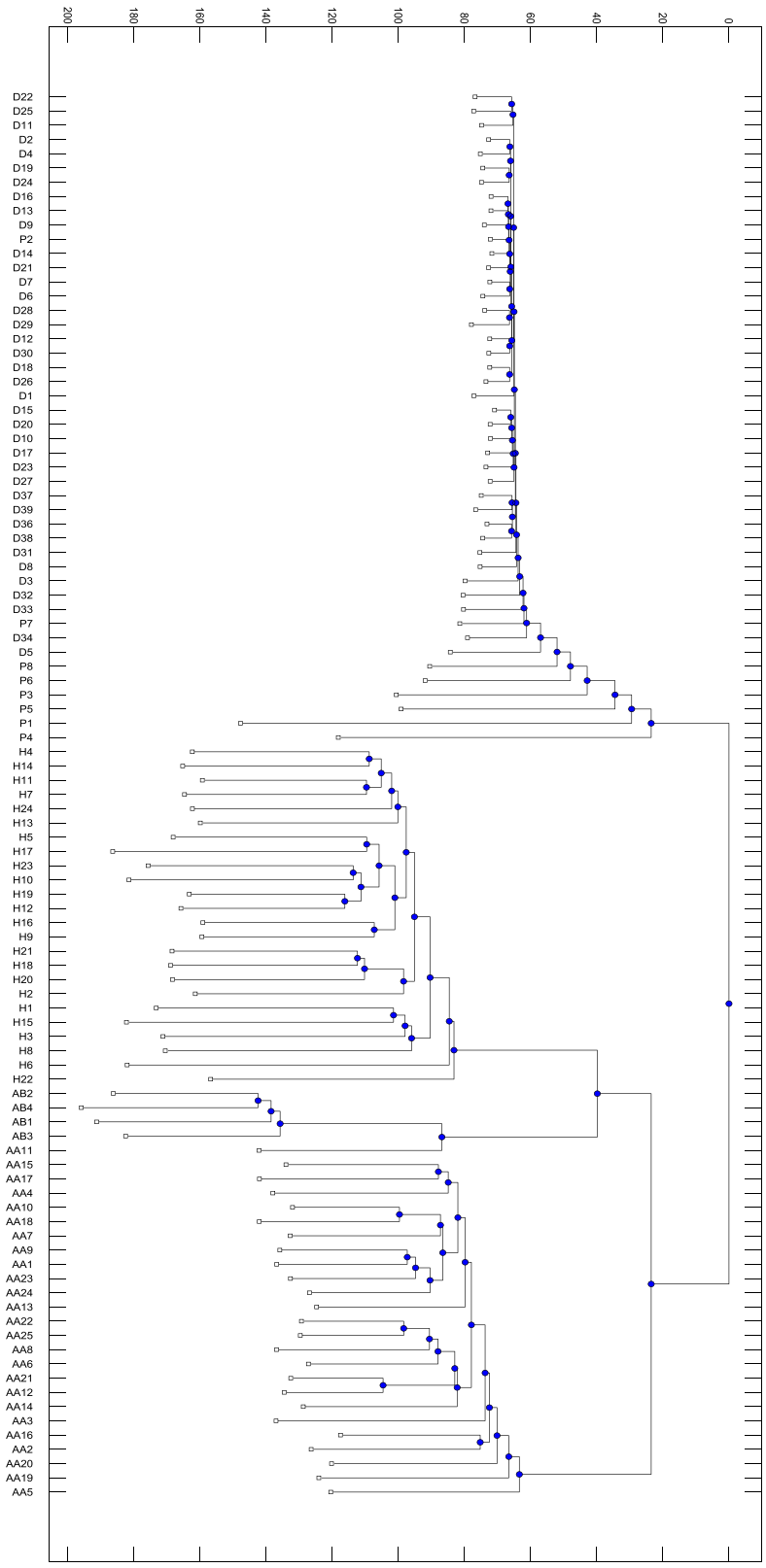

Figure 8. The clustering result of T10 dataset. Prefix: Diploid (D), Pseudodiploid (P), Hypodiploid (H), Aneuploid A (AA), Aneuploid B (AB).

The highlights of LoRA-TV are as follows: (i) traditional methods smooth and -reduce dimension of read depth profile from SCS data individually, while the proposed model processes the read depth profiles jointly before tumor cell clustering, and hence can aggregate common feature that disperses within cells in a subclone. Therefore, improved clustering results were expected. (ii) Since the proposed model consists of three terms, namely, the data fitting term (the Frobenius norm), the total variation smoothing term (the L-1 norm) and the low-rank approximation term (the nuclear norm), and there is no specific solver for this model, we

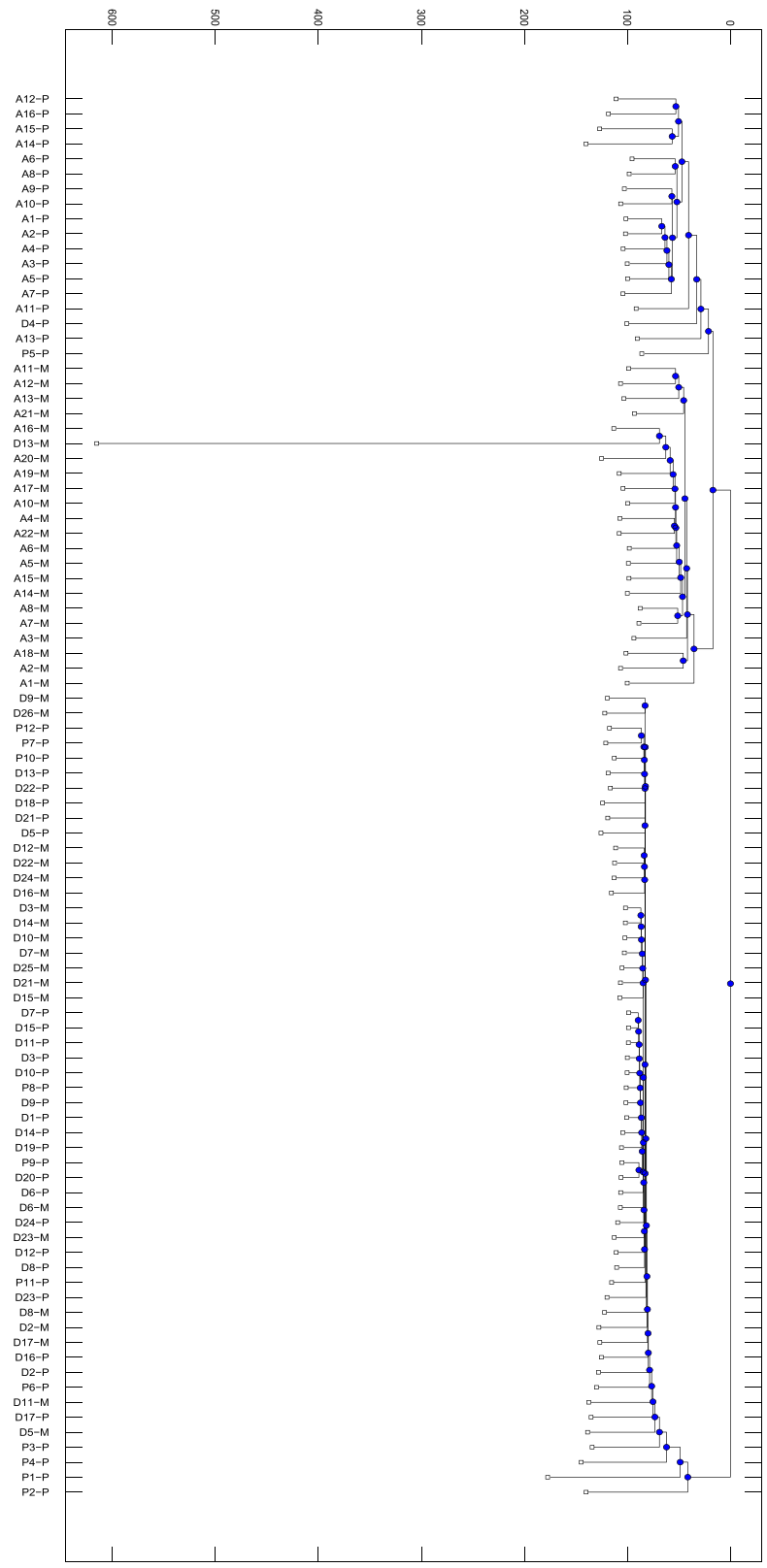

Figure 9. The clustering result of T16PM dataset. Prefix: Diploid (D), Pseudodiploid (P), Aneuploid (A); suffix: -P (T16P), -M (T16M).

employed ADMM to design a numerical optimization algorithm. Compared with the general convex optimization toolbox *CVX*, in the simulation data analysis section, the proposed algorithm showed its computational efficiency in terms of CPU and memory usage. (iii) Current popular learning-based methods have the generalization issue: these models tend to perform well only on data they have seen before, but their predictive capabilities are limited when it comes to new data. The proposed model is based on

traditional method, and does not suffer from the generalization issue, and hence have practical applications.

Note that there are two identified issues resulting in imperfect clustering (ARI 0.83): (i) SRR089717 labeled as D13 from the T16M dataset is clustered as a singleton, and (ii) SRR089589 labeled as D4 and SRR089542 labeled as P5 from the T16P dataset are clustered into A. Notably, these two issues are also observed in the Ginkgo website.

Last, there is still room for improvement of LoRA-TV. Since SVD is called densely in ADMM iteration (Eq. (9)), computation is slow for long read depth profile. In our real data analysis section, read depth profiles are of length 50 000 for a whole-genome sequencing, and were processed chromosome by chromosome. This took about tens of minutes. If we can tailor the incremental SVD [44] into ADMM, computation efficiency maybe increased.

Code of LoRA-TV is available at Matlab file exchange https://www.mathworks.com/matlabcentral/fileexchange/158481-lora-tv-low-rank-approximation-with-total-variation.

Key Points

- LoRA-TV clusters tumor cells based on the read depth profile from single-cell sequencing;
- The proposed method aggregates common genomic signatures through low-rank optimization and robust smoothing;
- ADMM was employed to solve the introduced optimization problem;
- Both clustering effectiveness and computational efficiency were supported by simulated and real data.

Funding

This work was supported by the National Natural Science Foundation of China (61771381).

Data availability

The data are available upon request.

References

1. Freeman JL, Perry GH, Feuk L. *et al*. Copy number variation: new insights in genome diversity. *Genome Res* 2006;**16**:949–61.
2. Duan J, Fu X, Zhang J-G. *et al*. The next generation sequencing and applications in clinical research. In: Wang X, Baumgartner C, Shields D. *et al*. (eds). Translational Bioinformatics, chapter 4 *Application of Clinical Bioinformatics*. Springer, 2016, 83–113.
3. Campbell PJ, Stephens PJ, Pleasance ED. *et al*. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* 2008;**40**:722–9.
4. Stefansson H. *et al*. Large recurrent microdeletions associated with schizophrenia. *Nature* 2008;**455**:232–6.
5. Rovelet-Lecrux A, Hannequin D, Raux G. *et al*. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat Genet* 2006;**38**:24–6.
6. Shlien A, Malkin D. Copy number variations and cancer. *Genome Med* 2009;**1**:62.
7. Tonini GP. Growth, progression and chromosome instability of neuroblastoma: a new scenario of tumorigenesis? *BMC Cancer* 2017;**17**.
8. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* 2009;**10**:80.
9. Duan J, Zhang J-G, Deng H-W. *et al*. Comparative studies of copy number variation detection methods for next generation sequencing technologies. *PloS One* 2013;**8**:e59128.
10. Yoon S, Xuan Z, Makarov V. *et al*. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* 2009;**19**:1586–92.
11. Chiang DY, Getz G, Jaffe DB. *et al*. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* 2009;**6**:99–103.
12. Baslan T, Kendall J, Rodgers L. *et al*. Genome-wide copy number analysis of single cells. *Nat Protoc* 2012;**7**:1024–41.
13. Talevich E, Shain H, Botton T. *et al*. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput Biol* 2016;**12**:e1004873.
14. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* 2016;**17**:175–88.
15. Lei Y, Tang R, Xu J. *et al*. Applications of single-cell sequencing in cancer research: progress and perspectives. *J Hematol Oncol* 2021;**14**:91.
16. Menon S, Lui VCH, Tam PKH. Bioinformatics tools and methods to analyze single-cell rna sequencing data. *Bioinformatics* 2021;**6**.
17. Navin N, Kendall J, Troge J. *et al*. Tumour evolution inferred by single-cell sequencing. *Nature* 2011;**472**:90–4.
18. Garvin T, Aboukhalil R, Kendall J. *et al*. Interactive analysis and assessment of single-cell copy-number variations. *Nat Methods* 2015;**12**:1058–60.
19. Satija R, Farrell JA, Gennert D. *et al*. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**:495–502.
20. Trapnell C, Cacchiarelli D, Grimsby J. *et al*. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014;**32**:381–6.
21. Cao J, Spielmann M, Qiu X. *et al*. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 2019;**566**:496–502.
22. Chambolle A, Lions P-L. Image recovery via total variation minimization and related problems. *Numer Math* 1997;**76**:167–88.
23. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Statist Soc B* 1996;**58**:267–88.
24. Chandrasekaran V, Sanghavi S, Parrilo P, Willsky A, " Sparse and low-rank matrix decompositions", in *Forty-Seventh Annual Allerton Conference, Allerton House*, UIUC, Illinois, USA, 2009, vol. **42**, pp. 1493–8.
25. Boyd S. *Convex Optimization*. Cambridge University Press, 2009.
26. Duan J, Zhang J-G, Deng H-W. *et al*. CNV-TV: a robust method to discover copy number variation from short sequencing reads. *BMC Bioinformatics* 2013;**14**:1–12.
27. Cai J-F, CandèS EJ, Shen Z. A singular value thresholding algorithm for matrix completion. *SIAM J Optimization* 2010;**20**:1956–82.
28. Boyd S, Parikh N, Chu E. *et al*. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* 2011;**3**:1–122.
29. Xu F, Duan J, Liu W. Comparative study of non-convex penalties and related algorithms in compressed sensing. *Digit Signal Process* 2023;**135**:103937.
30. Abyzov A, Urban AE, Snyder M. *et al*. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs

from family and population genome sequencing. *Genome Res* 2011;**21**:974–84.
31. Zhang Y, Liu W, Duan J. On the core segmentation algorithms of copy number variations detection tools. *Brief Bioinform* 2024;**25**: 1–10.
32. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;**4**:406–25.
33. Rand WM. Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc* 1971;**66**:846–50.
34. McComb C. Adjusted rand index. *GitHub* 2023;
35. Hui S, Nielsen R. Sconce: a method for profiling copy number alterations in cancer evolution using single-cell whole genome sequencing. *Bioinformatics* 2022;**38**:1801–8.
36. Harmanci AS, Harmanci AO, Zhou X. Casper identifies and visualizes cnv events by integrative analysis of single-cell or bulk rna-sequencing data. *Nat Commun* 2020;**11**:1–16.
37. Waltman L, Eck N. "A smart local moving algorithm for large-scale modularity-based community detection", *the European physical journal. B Condensed matter physics* 2013;**86**:1–14.
38. Grabski IN, Street K, Irizarry RA. Significance analysis for clustering with single-cell RNA-sequencing data. *Nat Methods* 2023;**20**:1196–202.
39. CVX Research Inc., "*CVX: Matlab software for disciplined convex programming, version* 2.0", http://cvxr.com/cvx, 2012.
40. Langmead B, Wilks C, Antonescu V. *et al*. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* 2019;**35**:421–32.
41. Danecek P, Bonfield JK, Liddle J. *et al*. Twelve years of samtools and bcftools. *GigaScience* 2021;**10**. https://doi.org/10.1093/gigascience/giab008.
42. Quinlan AR, Hall IM. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2.
43. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* 1979;**74**:829.
44. Brand M, "Incremental singular value decomposition of uncertain data with missing values", *Tech. Rep. TR-2002-24, Mitsubishi Electric Information Technology Center America, 201 Broadway, Cambridge, Massachusetts 02139*, 2002.