scientific reports

Check for updates

OPEN

Accurate identification of locally aneuploid cells by incorporating cytogenetic information in single cell data analysis

Ziyi Li[1]✉, Ruoxing Li[1,2], Irene Ganan-Gomez[3], Hussein A. Abbas[3,4], Guillermo Garcia-Manero[3] & Wei Sun[5,6,7]✉

Single-cell RNA sequencing is a powerful tool to investigate the cellular makeup of tumor samples. However, due to the sparse data and the complex tumor microenvironment, it can be challenging to identify neoplastic cells that play important roles in tumor growth and disease progression. This is especially relevant for blood cancers, where neoplastic cells may be highly similar to normal cells. To address this challenge, we have developed partCNV and partCNVH, two methods for rapid and accurate detection of aneuploid cells with local copy number deletion or amplification. PartCNV uses an expectation-maximization (EM) algorithm with mixtures of Poisson distributions and incorporates cytogenetic information to guide the classification. PartCNVH further improves partCNV by integrating a hidden Markov model for feature selection. We have thoroughly evaluated the performance of partCNV and partCNVH through simulation studies and real data analysis using three scRNA-seq datasets from blood cancer patients. Our results show that partCNV and partCNVH have favorable accuracy and provide more interpretable results compared to existing methods. In the real data analysis, we have identified multiple biological processes involved in the oncogenesis of myelodysplastic syndromes and acute myeloid leukemia.

Abbreviations

EM	Expectation-maximization
CNV	Copy number variation
HMM	Hidden markov model
PCA	Principal component analysis
TNBC	Triple negative breast cancer
AML	Acute myeloid leukemia
MDS	Myelodysplastic syndrome

Single-cell RNA sequencing (scRNA-seq) has greatly improved our ability to understand the cellular composition of the tissues and organs of interest, identify phenotype-associated cell groups, and elucidate the mechanisms behind many biological processes[1–3]. These advantages make scRNA-seq a powerful tool for studying a wide range of human diseases, including Alzheimer's disease[4], cardiovascular disease[5], and cancer[6]. In cancer research, a crucial step of scRNA-seq data analysis is to delineate tumor cells or neoplastic cells from other cell types[7,8]. The tumors of each patient have their unique tumoral and microenvironmental evolution, and thus the scRNA-seq data from cancer patients tend to be more heterogeneous. Such heterogeneity is an exciting opportunity for improving our understanding of cancer with scRNA-seq, but it also imposes computational challenges to dissect composing cell types[9].

Neoplastic cells are abnormal cells that are undergoing excessive and uncontrolled proliferation[10]. These cells, which may or may not be malignant, can be extracted experimentally through cell sorting, although this is not

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [2]Department of Biostatistics, The University of Texas Health Science Center, Houston, TX 78284, USA. [3]Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [4]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [5]Biostatistics Program, Public Health Science Division, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA. [6]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27516, USA. [7]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA. ✉email: zli16@mdanderson.org; wsun@fredhutch.org

always possible due to a lack of suitable markers or the high cost and labor requirements associated with these experiments[11]. In fact, it may be more useful to study all cells from a sequencing experiment simultaneously in order to understand the characteristics of neoplastic cells within their surrounding microenvironment. Neoplastic cells in certain types of cancer often have distinct features compared to non-neoplastic cells, such as high expression of certain cell type markers or genes belonging to some oncogenic pathways[12]. However, identifying neoplastic cells based on these markers or pathways can be difficult due to inter-individual heterogeneity, technical artifacts, and noise from the tumor microenvironment[13].

Recently, computational methods have been developed to identify large-scale copy number variations (CNVs) by comparing the smoothed scRNA-seq data against an internal or external normal reference, such as inferCNV, HoneyBADGER, and copyKAT[3,14,15]. For example, inferCNV is a popular visualization method for identifying large-scale CNVs. It uses smoothed averages over gene windows and compares the expression magnitude to the average over a set of reference 'normal' cells. CopyKat is a recently developed tool serving a similar purpose[15]. CopyKAT uses an integrative Bayesian segmentation approach combining CNV inference and hierarchical clustering, which has been shown to achieve high accuracy in distinguishing cancer cells from normal cells in multiple cancer types. Both inferCNV and CopyKAT generally work well with tumor cells that demonstrate extensive chromosomal alterations, but they do not work well for cancer types that have fewer and shorter CNVs. This is often the case in hematologic cancers such as myelodysplastic syndromes and acute myeloid leukemia[16]. Moreover, they couldn't incorporate additional clinical information for detecting specific CNVs. There are also methods that sought to integrate scRNA-seq data with bulk DNA sequencing (DNA-seq) or single-cell DNA-seq (scDNA-seq) data, such as CONGAS, clonealign, and CCNMF[17–19]. These methods serve a different purpose: to cluster cells based on CNV or mutation information.

In this paper, we propose to exploit cytogenetic information to improve the sensitivity and specificity for CNV identification. Cytogenetic data are routinely measured and recorded for patients with hematologic cancers[20] and they provide useful information to identify CNVs[21]. In the case of myelodysplastic syndromes, cytogenetic features, along with other factors such as morphology, immunophenotype, and clinical features, are included in the World Health Organization (WHO)-classification-based Prognostic Score System (WPSS) for myelodysplastic syndromes[22] and its revised version[23]. Similar risk scoring systems also exist for other types of hematologic malignancies[24,25]. For example, Leukemia patients with certain cytogenetic features, such as deletion of chromosome 7 or 7q, deletion of 3q, or amplification of chromosome 8, have been shown to have a poor prognosis[26]. Cytogenetic data provide location information of each CNV and the proportion of cells with specific CNVs based on the analysis of 20 metaphases. For example, if a patient has cytogenetic data as "46,XY,del(20)(q11.1q13.1)[5]/46,XY[15]," this means that approximately 25% percent of cells (5 out of 20) have a deletion of chromosome 20 in the region q11.1 to q13.1, while the rest of the cells have normal chromosomal features. Cytogenetic data are typically cheaper and more readily available in clinical settings compared to DNA-seq or scDNA-seq experiments. While the proportion in cytogenetic data is a crude estimate of the aberrant cells, they can still be useful in classifying cell status and identifying cells with chromosomal abnormalities, which may be markers for neoplastic cells[27]. None of the existing computational methods is able to incorporate such cytogenetic information in the analysis of scRNA-seq data.

Here, we introduce two methods, partCNV and partCNVH, for identifying cells with regional chromosomal abnormalities from scRNA-seq data by integrating cytogenetic information. Both methods are based on a statistical framework that models the count expression matrix of scRNA-seq data using a mixture of Poisson distributions while incorporating the cytogenetic information through prior specification. PartCNVH is built on partCNV and it further includes a hidden Markov model (HMM) to improve feature selection and clustering accuracy. It should be noted that our proposal is complementary to the existing methods such as copyKAT and inferCNV, as they focus on identifying large-scale CNVs while we detect smaller variations with the incorporation of external information. We implement our proposed methods in a computationally efficient expectation-maximization (EM) algorithm[28] and evaluate their performance through extensive simulation studies. We then apply them to three scRNA-seq data sets from patients with hematologic malignancies and show that they can identify cells with chromosomal deletions or amplifications in specific regions suggested by the cytogenetic data. We also perform additional analysis to understand the changes in the pathways and biologic processes in the identified aneuploid cells. Compared to existing methods, partCNV and partCNVH provide more interpretable results and additional findings. With the widespread use of single-cell technology in hematologic cancer research and clinical care of cancer patients, our methods offer a useful solution for fully leveraging cytogenetic data to identify cells with specific chromosomal abnormalities.

Method overview

PartCNV is a statistical framework that uses a hierarchical Poisson mixture model to differentiate two mixture components corresponding to normal and aberrant cells. PartCNVH is an extension of partCNV with the addition of HMM when there is a sufficient number of genes that allow feature selection. Figure 1 provides a schematic overview of the proposed methods. Our methods start with the normalized expression counts from the region with a known chromosomal deletion or amplification and explicitly incorporate the prior knowledge from cytogenetic data through imposing a Bernoulli prior on the cell status (i.e., normal or aberrant). We develop an EM algorithm that treats cell status as the missing variable and efficiently solves the mixture model. The inferred cell status from this step is the output of partCNV.

Taking the ouptut from partCNV, partCNVH further refines it by a HMM. Specifically, a group average is taken for the inferred two groups of cells and the rolling average of the ratios between the two groups is used to infer the hidden status of the regions by a HMM. There are two reasons that we adopt this combination of rolling average and HMM in partCNVH. First, as shown in our later results, the group mean and rolling average can effectively magnify the signal of the regional deletion or amplification on the expression level. This is especially

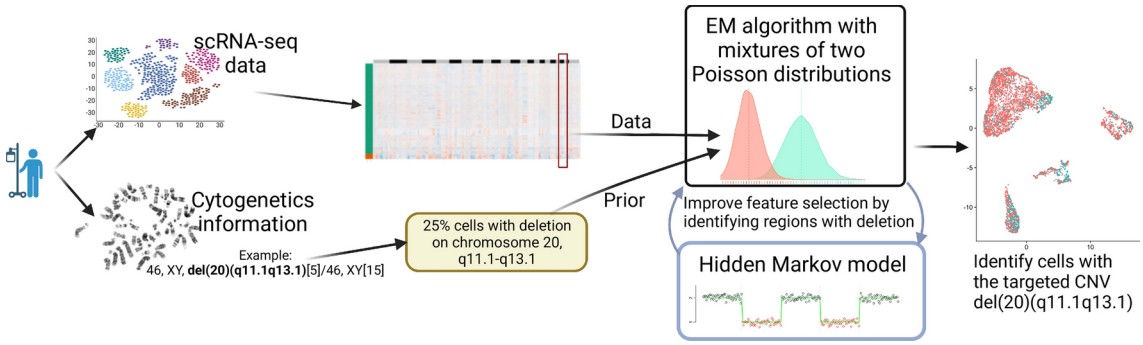


Fig. 1. Schematic of PartCNV and PartCNVH. With the input of normalized expression counts from scRNA-seq experiments and the cytogenetic information from the patient, we develop an EM algorithm with mixtures of two Poisson distributions to infer the aneuploid/diploid status for the regions of interest. We further include a hidden Markov model to improve feature selection and the classification accuracy.

important when the signal of copy number alternations is weak related to noise in gene expression measurement. Second, it is possible that only a subset of the regions of interest has copy number changes. HMM can identify regions that are more likely to contain the chromosomal changes, which in turn improves the performance of aneuploid/diploid cell classification. After this HMM-based feature selection step, partCNVH performs a second round of the EM algorithm using the Poisson mixture model and reports the inferred cell status.

Simulations

We design comprehensive simulation settings to evaluate the performance of partCNV and partCNVH. As comparisons, we also consider existing methods using dimension reduction by principal component analysis (PCA) followed by Louvain or Leiden clustering. Previous literature has reported that the Leiden algorithm can generate better connected communities through including an extra refinement step and run faster than Louvain[29]. Additionally, we include two widely used machine learning clustering algorithms, K-means clustering and hierarchical clustering. All the previous mentioned methods can be applied to detect *locally* aneuploid cells. Although our proposed method is not directly comparable to existing methods that classify cells based on whole-genome CNV inference, we still design a separate simulation study to compare the proposed methods versus the two large-scale CNV detection-based methods, inferCNV and copyKAT[14,15].

We consider two settings where the first one studies aneuploid cells with deletions and the second one studies amplifications: redSimulation data 1 and 2. The mean expression for these genes is generated by taking a ratio of the normal expression. This ratio (or log fold change) is randomly drawn from a Uniform distribution with different base levels (0.5/0.6 for Setting 1 and 1.5/1.4 in Setting 2) and different noise levels. A larger noise level makes the expression from the aneuploid cells similar to the normal cells, and thus creates harder scenarios for the methods. The evaluation criteria is the accuracy of the classification results of the proposed method evaluated by the true cell status (i.e., being aneuploid or normal). More details of the simulation settings are provided in the Methods section.First consider simulation data 1, where 500 out of 3000 cells have deletions. Our proposed methods partCNV and partCNVH have the highest accuracy among all the methods in all scenarios (Figure 2A-B). Using the same normalized gene expression counts input as our methods, K-means and hierarchical clustering have the lowest accuracy ranging from 0.5 to 0.7. PCA plus Louvain and Leiden have higher accuracy than K-means and hierarchical clustering. When the signal is strong (ratio = 0.5 in panel A) and noise is small, PCA plus Louvain/Leiden also have similar high accuracy as the proposed methods. But with the increase of the noise level, the accuracy of PCA plus Louvain or Leiden decreases. The advantage of the proposed methods becomes more obvious when the ratio is 0.6 and the noise level is high. For example, the mean accuracy of PCA plus Leiden is around 0.75 while partCNVH can achieve a high accuracy of 0.9. This is understandable since the proposed methods specifically model the data through two components for normal and aneuploid cells, and they allow mixtures of regions with and without deletions. Second, partCNV and partCNVH have similarly good performance with accuracy higher than 0.9, and partCNVH generally has higher accuracy than partCNV. To better understand the role of the HMM step of partCNVH and the result of feature selection, we use one simulation data set as an example and visualize the mean gene expression across the region (Figure 3A), ratios of the mean expressions of the two groups inferred by partCNV (Figure 3B), the rolling average of the mean expression ratios (Figure 3C), and the inferred status from HMM (Figure 3D). It can be seen that the rolling average of mean expression ratios between the two groups can effectively magnify the signal, and a majority of the HMM selected genes are located in the region with deletion. Figure 3E shows that our proposed method partCNVH has greater accuracy in classifying cells than the other methods.

Evaluation of PartCNV & PartCNVH for different prior information: amplification regions

Next we evaluate different methods in simulation data 2 with amplifications (Figure 4). Note that for cells with deletion, the expression change odds is 2 (from 1 to 0.5) while in cells with amplification the odds is 0.67 (from 1 to 1.5) or its inverse 1.5. Thus the signals from the amplified regions can be harder to detect than the first setting with deleted regions. In Figure 4(A), we observe the performance of all methods decrease, especially for PCA

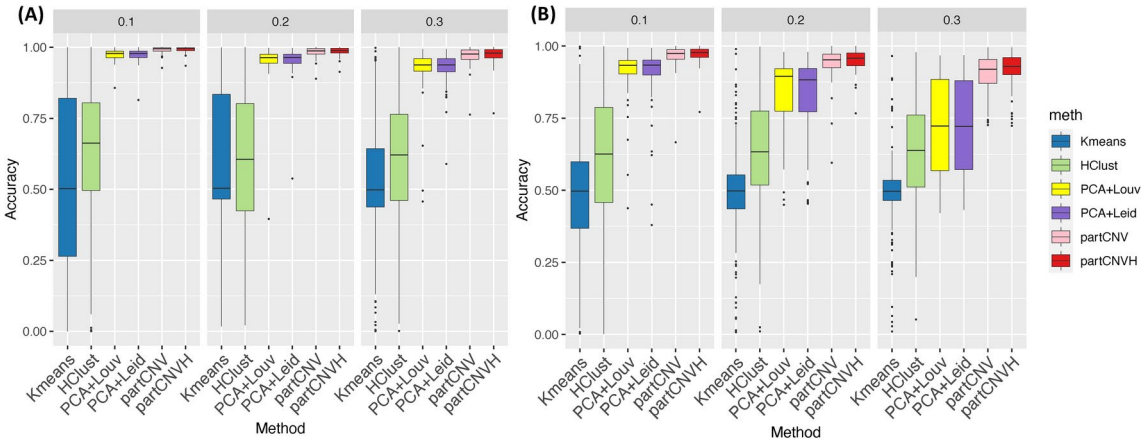


Fig. 2. Results of simulation Setting 1 with deletions. The methods that are compared include K-means clustering (Kmeans), hierarchical clustering (HClust), dimension reduction using PCA plus Louvain clustering (PCA+Louv), and dimension reduction using PCA and Leiden clustering (PCA+Leid). Each simulation dataset contains a total of 3000 cells: 2500 normal cells and 500 with deletions. (**A**) The accuracy of these methods when the ratio of gene expression in a deletion region versus normal expression is 0.5 at different noise levels (0.1: low, 0.2: medium, 0.3: high). (**B**) The results of the setting with ratio = 0.6 at different noise levels. All results are summarized over 100 Monte Carlo iterations.

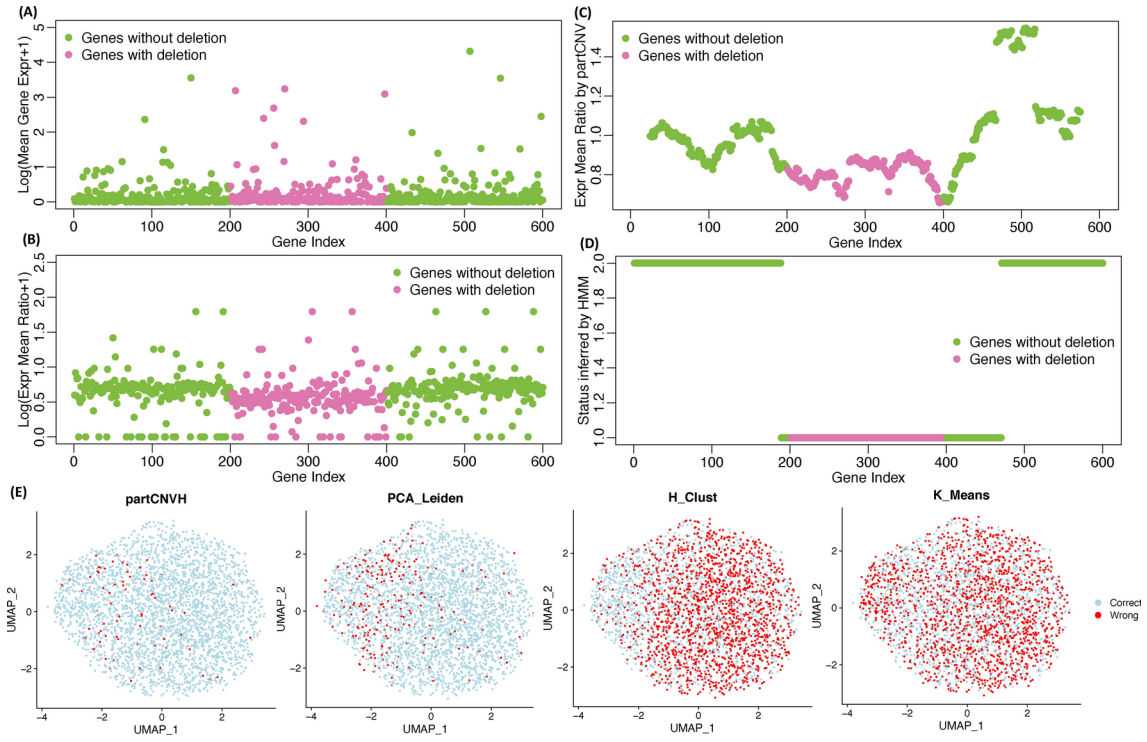


Fig. 3. Illustrating the procedure of the feature selection step using HMM in a simulation data set. (**A**) The log-transformed mean gene expression levels for the genes located in regions without and with deletion. Each dot is a gene. (**B**) The log-transformed ratio of the mean expression levels for cells without versus with deletion inferred by partCNV. (**C**) The expression mean ratio for the two groups of cells by partCNV after applying the rolling average with a bandwidth of 50. (**D**) The latent states inferred by HMM based on the rolling average from panel (**C**). (**E**) The results of classification accuracy of partCNVH, PCA plus Leiden, hierarchical clustering, and K-means clustering. The red dots are the cells incorrectly classified and the blue dots are the correct ones.

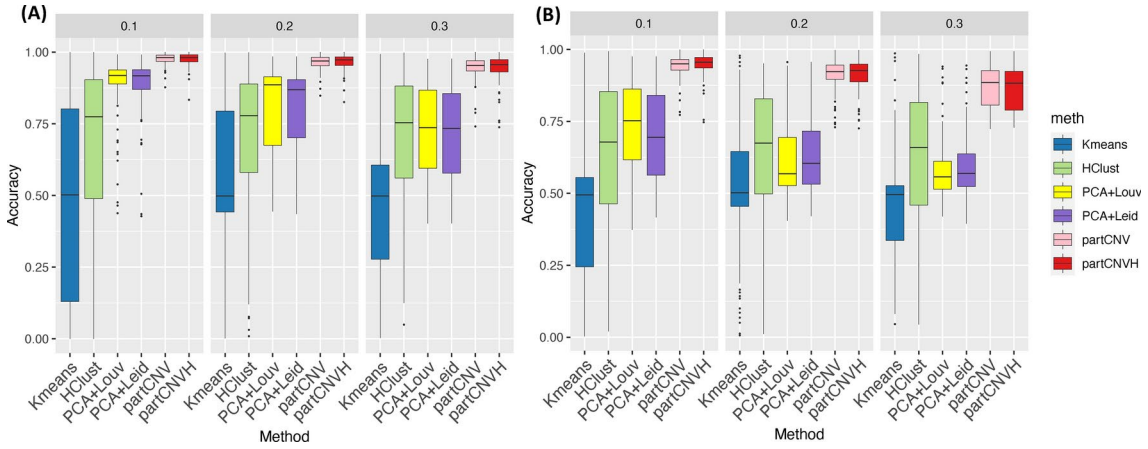


Fig. 4. Results of simulation Setting 2 with amplifications. Each simulation dataset contains a total of 3000 cells: 2500 normal cells and 500 with amplifications. (**A**) The accuracy of the methods compared when the ratio = 1.5 at different noise levels (0.1: low, 0.2: medium, 0.3: high). (**B**) The results when the ratio = 1.4 at different noise levels. All results are summarized over 100 Monte Carlo iterations.

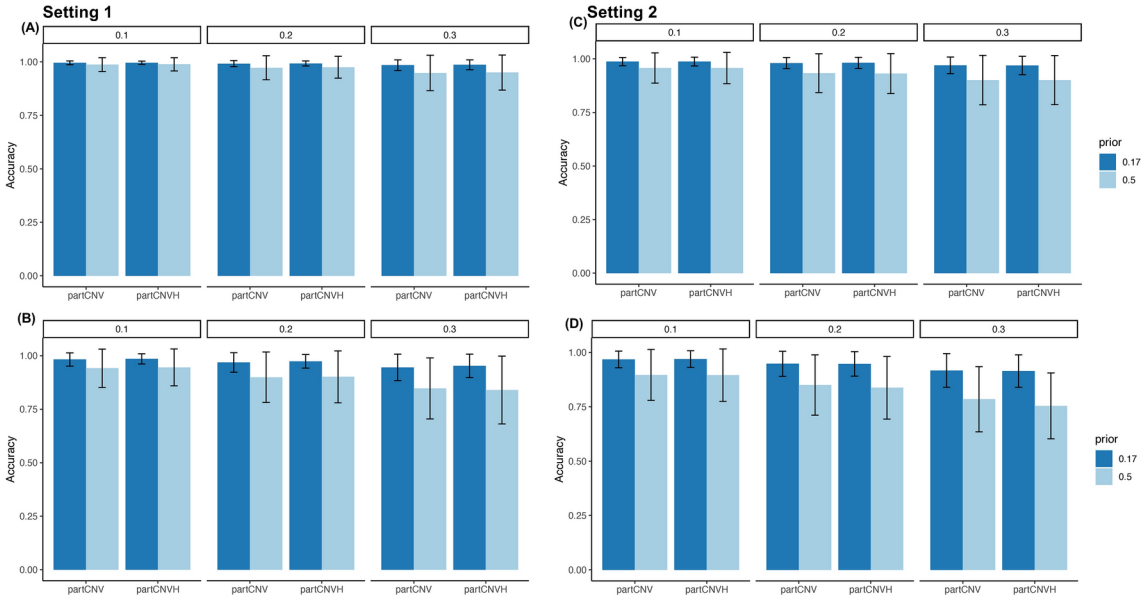


Fig. 5. Simulation results of evaluating the impact of different prior information on the classification accuracy of the proposed methods. The dark and light blue bars correspond to the results with and without correct specification of the prior information (true prior: 0.17). Panel (**A**) and (**B**) are the simulation results based on the first simulation setting with ratios = 0.5, 0.6, respectively. Panel (**C**) and (**D**) are based on the second simulation setting with ratios = 1.5, 1.4, respectively. All results are summarized over 100 Monte Carlo iterations.

plus Louvain and Leiden. When the noise level is high (0.3), PCA plus Leiden only reaches an accuracy of 0.74. In comparison, our proposed method still has a high accuracy of around 0.95. With the ratio level set as 1.4 in panel B, the signal level becomes lower and the existing methods have even lower accuracy ranging from 0.5 to 0.6, while the proposed methods still stay at a reasonable accuracy level around 0.9. These demonstrate the robustness of the proposed methods and highlight the importance of applying partCNV or partCNVH instead of existing methods when the region of interest has amplifications.

Evaluation of PartCNV & PartCNVH for different prior information: cell numbers

Our current simulation design considers a total of 3000 cells. When we study a region of interest suggested by cytogenetic data, more cells generally provide more information, and thus identifying signals from fewer cells can be more challenging. To evaluate the proposed methods under this scenario, we generate simulation data 3 by fixing the cell number as 1300, where 1000 cells are normal and 300 are aneuploid cells. Figure S1 shows

the simulation results with 1300 total cells and the region of interest (200 out of 600 genes) has a deletion in the aneuploid cells. Compared with the results in Figure 2, all the existing methods have worse performance for the same ratio and noise combinations. For example, both PCA plus Louvain and PCA plus Leiden have a high accuracy of around 0.90 when the ratio is 0.6 with a medium noise of 0.2 using 3000 total cells, while the accuracy decreases to around 0.8 using 1300 cells. The variation of the classification results also increases. Although our proposed methods have slightly decreased performance when the ratio is 0.6 and the noise is 0.3, their overall performance remains similar in other scenarios (ratio = 0.5 at all noise levels, ratio = 0.6 with low and medium noise levels).

Similar patterns can be observed in amplification settings by comparing the results in Figure S2 versus the results in Figure 4. Surprisingly, PCA plus Louvain or Leiden has even worse median accuracy than the hierarchical clustering, even though they all have quite low classification accuracy. These results suggest that our proposed methods tend to have more robust performance even with fewer cell numbers in the analyzed dataset, while the existing methods have decreased accuracy and more varied results, especially when the noise level is high.

Evaluation of PartCNV & PartCNVH for different prior information: proportions of aneuploid cells

As a methodology advantage, partCNV and partCNVH are able to incorporate the prior knowledge of an estimated proportion of aneuploid cells. If the prior is misspecified, we seek to understand the impact on the results. We generate simulation data 4 with the same data generation procedure but the prior information is specified as correct (0.17 for total cell number 3000 and 0.23 for total cell number 1300) and incorrect (0.5), and we examine the results of our proposed methods.

Figures 5 and S3 illustrate the accuracy for the total cell numbers 3000 and 1300. First, it is clear that the correct prior knowledge improves the classification accuracy than a non-informative prior of 0.5. This improvement is small when the ratio is 0.5 or 1.5, but it can be substantial when the signal is harder to detect (ratio = 0.6 or 1.4) and the noise level is high. For example, when the ratio is 1.4 and the noise is 0.3, the improvement of the accuracy for both partCNV and partCNVH using a correct prior can be about 10% compared to using the incorrect prior. Second, both Figures 5 and S3 demonstrate the robustness of the proposed methods against incorrect prior information, especially in panels A and C where the ratios are 0.5 and 1.5, respectively. In these experiments, we choose 0.5 as the incorrect prior knowledge, which is far from the true proportion 0.17 and illustrate a worst scenario that the prior is completely non-informative. In reality, when a closer prior such as 0.20 or 0.15 is used, the impact would be much smaller. Even in the worst scenario, with an amplification ratio 1.4 and a high noise level 0.3, our method with an incorrect prior still reaches a median accuracy above 0.75, which is better than the existing methods under the same scenario. These results highlight the advantage of the proposed methods in accurately identifying aneuploid cells.

Comparison with genome-wide CNV detection methods

Lastly, we compare the proposed methods with two widely used genome-wide CNV detection methods, inferCNV, copyKAT as well as the regional versions of inferCNV and copyKAT. For the regional versions, we applied hierarchical clustering on the normalized expression matrix by inferCNV/copyKAT for the regions of interest only. One example of simulation data 4 is visualized in Figure 6A. As inferCNV requires the input of normal cells, we use an additional 100 normal cells as the reference for inferCNV. Both inferCNV and copyKAT generally are much more computationally intensive. We summarize the accuracy of 20 Monte Carlo simulations. It can be observed that neither inferCNV/copyKat nor the regional versions of these methods is able to accurately infer aneuploid/diploid cell status (Figure 6B-D), since the regional aneuploid signal is too weak compared to genome-wide copy number alterations that inferCNV and copyKAT are designed to detect. It is also interesting that when we re-cluster the cells based on the normalized expression from the sub-regions of chromosome 20, the performance of copyKAT increased but not inferCNV. These findings highlight the need for region-specific detection tools with the considerations of cell type mixtures to distinguishing between aneuploid and diploid cells.

Real data application

We demonstrate the usage of the proposed methods on three real data applications. For each application, the scRNA-seq data from one patient were collected by the 10X genomics platform and the cytogenetic data were collected from patients' medical records. All three patients have a subset of the cells with regional copy number variations, and the rest are normal cells. The three applications have different complexity levels. The first subject (patient 1) is the most straightforward one, as a very long region (the whole chromosome Y) was reported as lost in a subset of the cells according to the patient's cytogenetic data. The second subject (patient 2) has a subset of cells with partial chromosome 20 lost. The third subject (patient 3) has a complicated situation as this patient has cells with partial deletions, as well as cells with partial amplifications.

Patient 1: MDS with loss of chromosome Y

We obtain the scRNA-seq data of a bone marrow sample from an MDS patient treated at MD Anderson Cancer Center. The data, after alignment and quality control, contains a total of 33,538 genes and 655 cells. For this patient, the bone marrow sample has been specifically sorted for CD34+ cells to enrich hematopoietic stem and progenitor cells (HSPCs). As a result, the cell number is smaller than regular scRNA-seq experiments. Based on the clinically obtained cytogenetic data, this patient has around 35% of cells with the loss of chromosome Y and our goal is to identify these aneuploid cells. We first apply copyKAT to the whole transcriptome data from this sample to infer copy number variations. As shown in Figure S4, copyKAT clusters the cells based on the inferred

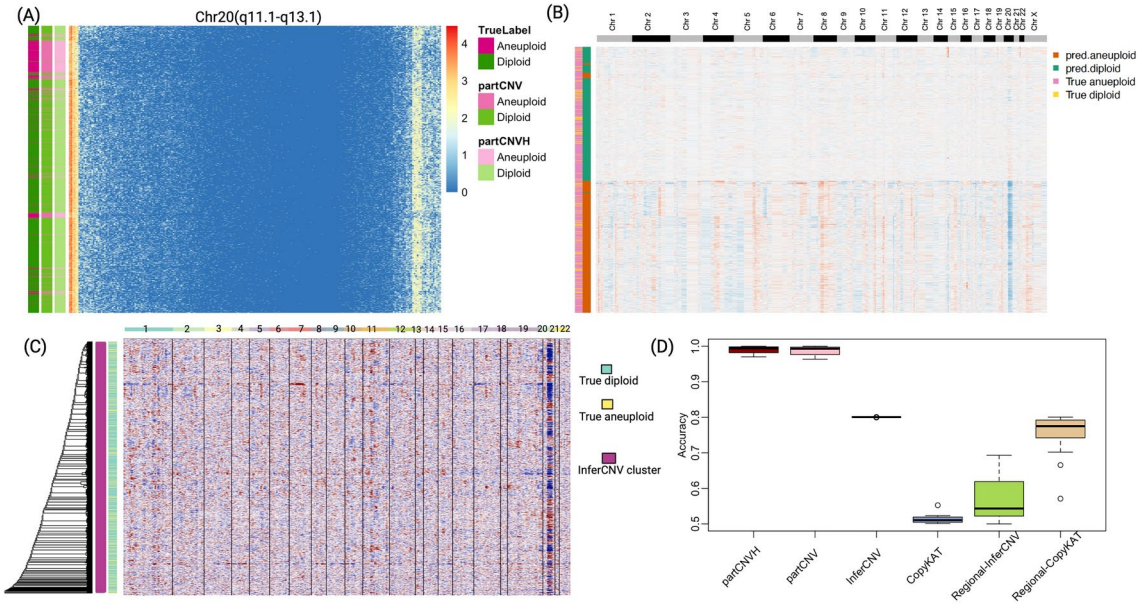


Fig. 6. Simulation results to compare the proposed method versus existing whole-genome based methods, InferCNV and CopyKAT. (**A**) Heatmap of the simulated expression values for genes in region Chr20(q11.1-q13.1), where the rows are the cells and columns are the genes. Logarithmic of the expression values plus one are used for visualization. Rows are labeled by the true aneuploid/diploid status and the inferred status by partCNV and partCNVH. (**B**) CopyKAT output of the aneuploid/diploid prediction versus the true cell status. The heatmap includes all the chromosomes. (**C**) Copy number results using InferCNV. (**D**) Boxplot of the aneuploid/diploid inference of the methods averaged over 20 Monte Carlo simulations. Regional InferCNV and Regional CopyKat are the hierarchical clustering results based on the normalized expression of the region of interest from InferCNV and CopyKAT.

CNV statuses across the whole genome, but it could not take regional data or the cytogenetic information into consideration when identifying the neoplastic cells.

To specifically identify cells with the chromosome Y loss, we apply the proposed methods and the five existing methods (PCA plus Louvain, PCA plus Leiden, K-means clustering, hierarchical clustering, and CopyKAT) on this dataset. The input data are the normalized counts from the genes located on chromosome Y. In this application, the HMM step from partCNVH selects the whole set of genes so we only present the results from partCNV. Since the CNV in this dataset encompasses the entire chromosome Y, we expect some of the existing methods also work well for this analysis. We find that partCNV, PCA plus Louvain, PCA plus Leiden, and K-means clustering all have proportions of aneuploid cells close to 35% (40.9%, 39.5%, and 39.5%, respectively) (Figure 7 A-E). Hierarchical clustering identifies a much smaller number of aneuploid cells (9.31%). From visualizing the pairwise ARI values of these results, we find that partCNV, PCA plus Louvain/Leiden have very similar results, while K-means clustering and hierarchical clustering have very different results (Figure 7 F). As the UMAP coordinates in Figure 7(A-E) are obtained using the whole transcriptome data, the fact that copyKAT-identified aneuploid cells cluster together suggests that copyKAT captures the whole transcriptome pattern instead of chromosome Y specific changes.

We further examine the average gene expression of chromosome Y genes among the aneuploid/normal cells identified by partCNV, copyKAT, K-means, and hierarchical clustering (Figure 7 G-I). It is apparent that partCNV-labeled aneuploid cells have much lower expression than the cells labeled as aneuploid by other methods but normal per partCNV, confirming that partCNV correctly identifies the cells with deletion on chromosome Y.

In summary, these results suggest that partCNV and the PCA plus Louvain or Leiden clustering have identified the cells with the chromosome Y loss. In contrast, K-means clustering, hierarchical clustering, and copyKAT failed to do so.

Patient 2: MDS with partial deletion of chromosome 20

We obtain the scRNA-seq data from the bone marrow sample for a different MDS patient. The data were also generated by the 10X genomics scRNA-seq technology. This sample was sequenced directly without the cell sorting step, and thus both HSPC and immune cells can be potentially identified. After alignment, preprocessing, and quality control, a total of 24,519 genes and 3,643 cells are kept for the analysis. Based on the cytogenetic data, about 20% of cells in the sample have deletions in chromosome 20 at regions q11.1 to q13.1, which is about 24.2 Mb long. We also apply CopyKAT to this data and present the heatmap result in Figure S5. Although cytogenetics reported deletions in chromosome 20, the log copy number ratio heatmap does not have an obvious

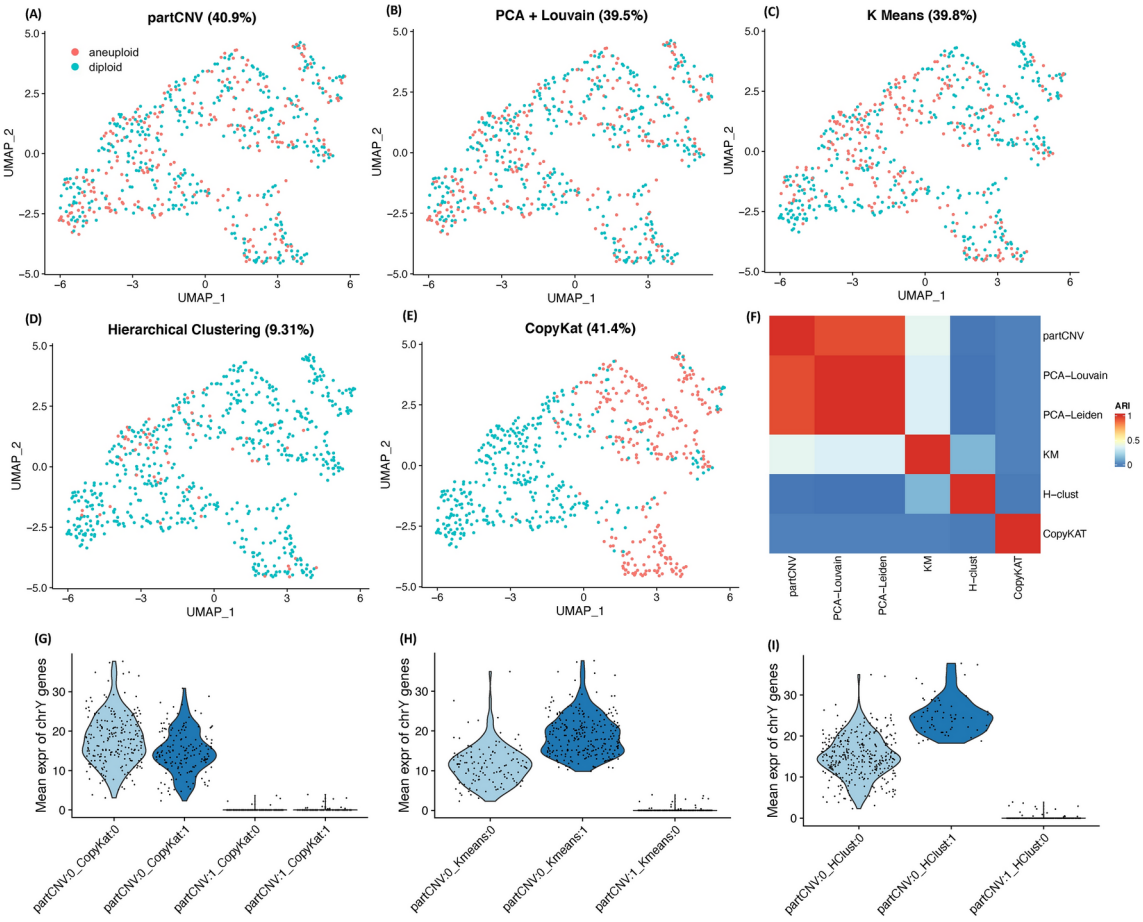
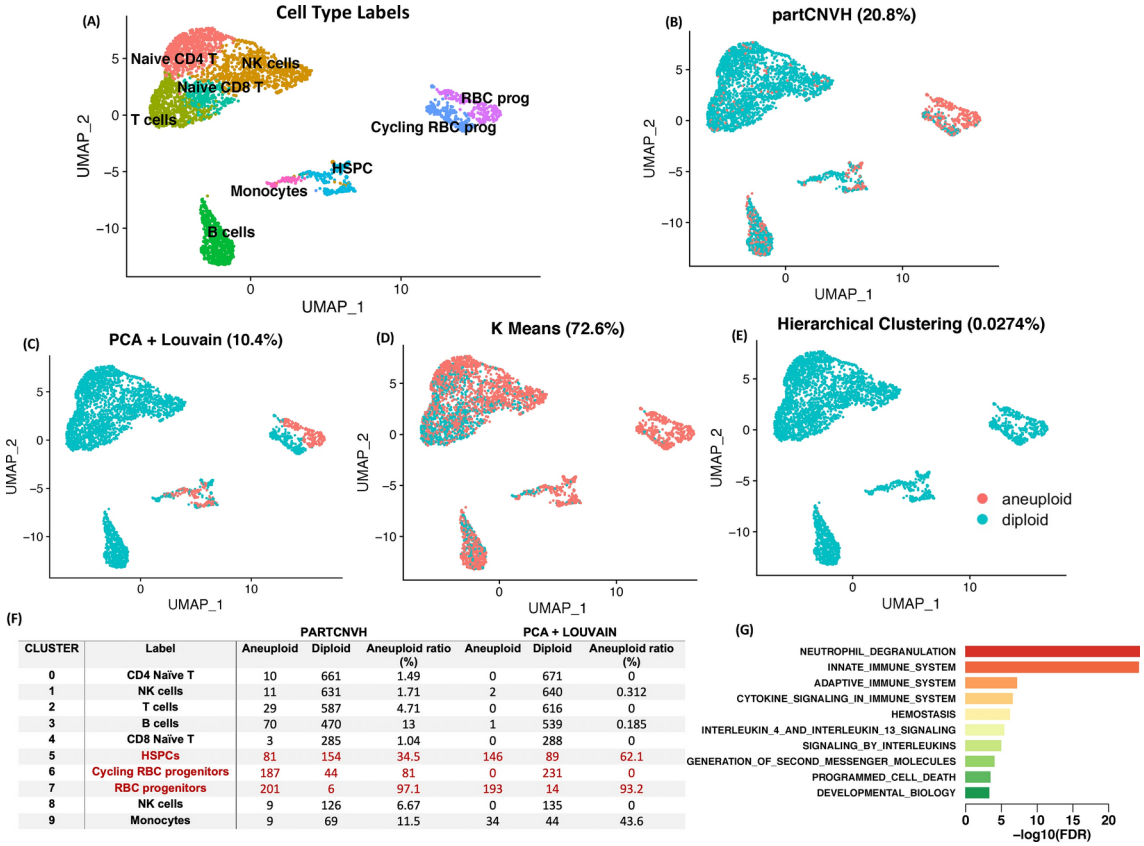


Fig. 7. Results of applying different methods to the scRNA-seq data using the bone marrow CD34 positive cells from patient 1. Cytogenetic information shows that $\sim 35\%$ of the cells from the sample have chromosome Y loss. (**A**-**E**) The cell classification based on the inference of chromosome Y loss using different methods. Panel F shows the heatmap of the ARI values from comparing the classification results using different methods. (**G**) The expression of chromosome Y genes for cells that are labeled as normal in partCNV (partCNV:0) and aneuploid in CopyKAT (CopyKat:1), and three similar groups. (**H**,**I**) The expression of chromosome Y genes comparing partCNV versus k-means clustering and comparing partCNV and Hierarchical clustering.

deletion pattern in the suggested regions. Based on the whole genome copy number inference, copyKAT only reports about 500 aneuploid cells (∼5.5%).

We analyze the whole-transcriptome data using Seurat[2] through identifying highly variable genes, extracting top principal components (PCs) based on these genes, and we perform UMAP and clustering analysis (Figure 8). UMAP is used for dimension reduction and unsupervised clustering is performed with the default Louvain clustering using the top PCs. A total of 10 clusters are identified, and the cluster specific markers are used for annotating the cell type labels based on biological knowledge by our MDS biologist. We also apply the proposed methods and existing methods targeting the region on chromosome 20 with known chromosomal deletions. Figure 8 A-E shows cell type labels and the aneuploid/diploid inference result using partCNVH, PCA plus Louvain, K-means clustering, and hierarchical clustering. The results for partCNV and PCA plus Leiden are presented in Figure S6. We find that the proposed methods have the closest proportion of aneuploid cells to the cytogenetics reported proportion; the other methods all have much lower or higher proportions. The major difference between our proposed method and PCA plus Louvain is in the cycling RBC progenitors (Figure 8 B, C, and F). Our method reports high proportions in all three MDS-related cell groups (i.e., HSPCs, RBC progenitors, and cycling RBC progenitors), while PCA plus Louvain only reports aneuploid cells in the former two clusters. Previous literature found impaired erythroid-proliferating capacities to be a prominent characteristic in patients with MDS[30,31]. Both RBC progenitors and cycling RBC progenitors are major cell types involved in the erythroid-proliferating function, and thus it makes sense to identify neoplastic cells in both cell types.

To understand the differences between the identified locally aneuploid cells and the normal cells, we conduct differential expression analysis for each cluster to compare the aneuploid versus diploid cells identified by our method[32]. A total of 177 cluster-specific differentially expressed genes were identified using a cutoff of 0.05 for adjusted p values. We also perform over-representation analysis using GO Biological Process database[33,34] and identify several functional categories over-represented by differential expression signals, including neutrophil



		PARTCNVH			PCA + LOUVAIN		
CLUSTER	Label	Aneuploid	Diploid	Aneuploid ratio (%)	Aneuploid	Diploid	Aneuploid ratio (%)
0	CD4 Naïve T	10	661	1.49	0	671	0
1	NK cells	11	631	1.71	2	640	0.312
2	T cells	29	587	4.71	0	616	0
3	B cells	70	470	13	1	539	0.185
4	CD8 Naïve T	3	285	1.04	0	288	0
5	HSPCs	81	154	34.5	146	89	62.1
6	Cycling RBC progenitors	187	44	81	0	231	0
7	RBC progenitors	201	6	97.1	193	14	93.2
8	NK cells	9	126	6.67	0	135	0
9	Monocytes	9	69	11.5	34	44	43.6

Fig. 8. Application results of the proposed method and existing methods using the scRNA-seq data from patient 2. (**A**) The cell type annotation based on marker genes' expression and biological knowledge. RBC progenitor: red blood cell progenitors. HSPC: hematopoietic stem and progenitor cells. NK: natural killer. (**B**-**E**) The inferred aneuploid cells with deletion chr20(q11.1q13.1) using different methods (B, partCNV; C, PCA+Louvian; D, K Means clustering; E, Hierarchical clustering). The proportions in the title bracket are the proportions of cells that are inferred as cells with this specific deletion. (**F**) The comparisons of results by partCNVH and PCA plus Louvain versus cell type labels. (**G**) Gene set enrichment analysis results for the genes that are differential expressed between the cells with and without deletion chr(20)(q11.1q13.1). Gene sets are defined using the Reactome Pathway Database.

degranulation, a few immune system related terms, and interleukin 4 and interleukin 13 signals (Figure 8 G). Many of these terms have been reported in previous literature to be related with MDS. For example, neutrophil degranulation and migration has been reported to be associated with MDS compared with normal controls[35]. The important role of the immune system and innate immune signaling in MDS has also been reported in multiple publications[36–38]. The over-representation results using the Reactome pathway and Hallmark database are presented in Figure S7[39,40]. In the Reactome pathway enrichment results, we also find several immune response-related and lymphocyte activation related terms. In Hallmark analysis, allograft rejection, complement, interferon gamma response, and TNFA signaling via NFKB are the top findings, which also have been associated with MDS pathogenesis[41–43]. There are also some terms that have not been reported in previous MDS studies, such as generation of second messenger molecules, hemostasis, defense response, and cell activation, which could be promising targets for future research.

Patient 3: AML with partial gain and partial deletion of chromosome 8

Lastly, we study the scRNA-seq data from an AML patient with complicated chromosomal variations[44]. Specifically, this patient has amplifications of the whole chromosome 8 in 25% of the cells, deletion of chromosome 8 at region q21.2 to q24.3 in 40% of the cells, and normal karyotype in the rest of the cells. We are interested in identifying which cells contain the chromosome 8 gain and which have del(8)(q21.2q24.3). The scRNA-seq data were generated using the 10X genomics technology platform. After preprocessing, the data contain a total of 20,521 genes from 4294 cells. Since there are overlaps between the deletion and amplification regions, we apply partCNVH through a two-step approach. We first focus on the region of chromosome 8 before q21.2 where about 25% of the cells have amplification; the rest of the cells are normal in the area. After the cells with chromosome 8 gain are detected, we apply partCNVH again to the rest of the cells, which contain del(8)(q21.2q24.3) in around 53.3% ($= 40\%/(1-25\%)$) of the cells. In the two steps, 25% and 53.3% are used as the prior knowledge of the aneuploid cell proportions in partCNVH.

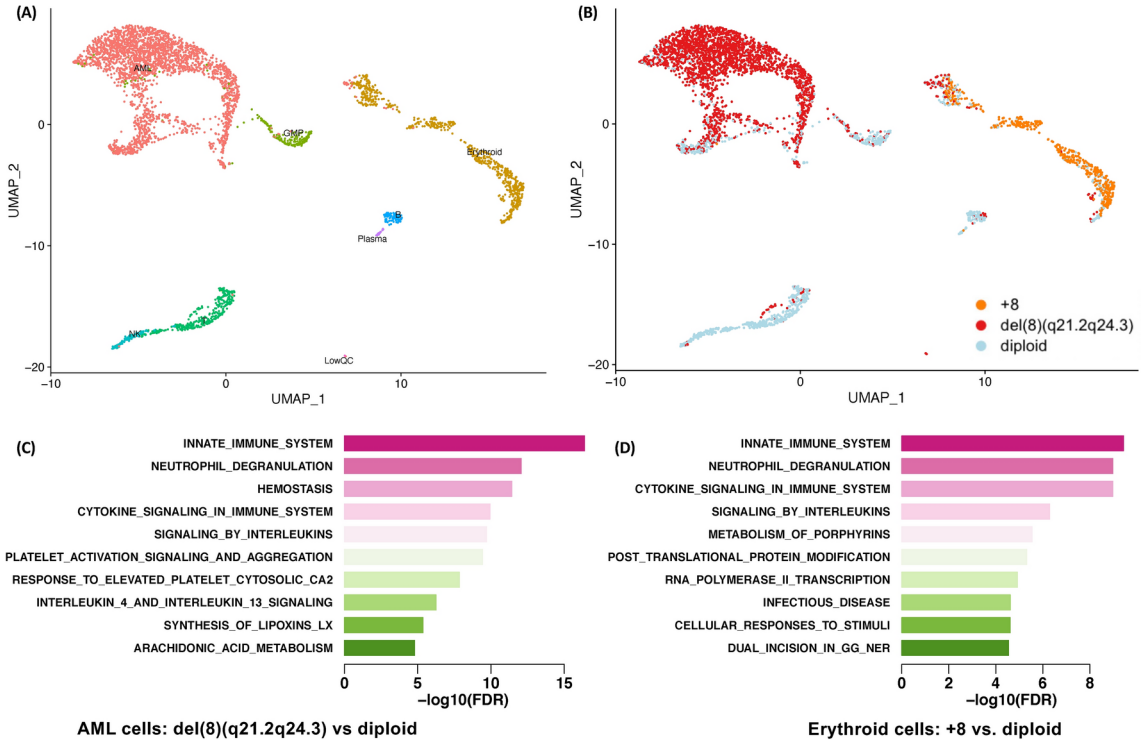


Fig. 9. Results of applying the proposed method partCNVH to the scRNA-seq data from an AML patient with $\sim 25\%$ cells having amplification of chromosome 8 and $\sim 40\%$ cells having chromosome 8 deletion at region q21.2-q24.3. (**A**) The true cell type labels annotated by a clinician-scientist. (**B**) The inferred cells with amplification of 8 ("+8") and the deletion of chromosome 8 at q21.2-q24.3 ("del(8)(q21.2q24.3)"). (**C**, **D**) The Reactome pathway enrichment analysis results using the DEGs by comparing the AML cells with del(8)(q21.2q24.3) versus diploid cells, and by comparing the erythroid cells with +8 versus diploid cells, respectively.

We find that the majority of the inferred cells with chromosome 8 amplification are from the erythroid cells, while the cells with del(8)(q21.1q24.3) are mostly AML cells (Figure 9 A-B). The proportions of cells with chromosome 8 amplification and del(8)(q21.1.q24.3) are about 13% and 60%, which are close to the prior knowledge from cytogenetics. We visualize the normalized expression for the region of interest on chromosome 8 in Figure S8A. We also present the results by inferCNV and copyKAT in Figure S8B and Figure S9, respectively. We observe that the patterns for chromosome 8 varies between the three cell groups in Figure S8A. However, whether the pattern shows gain or loss is not easy to identify due to data sparsity. Neither CopyKAT nor InferCNV can be applied to regional data, and thus they couldn't reproduce the patterns we observed in Figure S8A.

To understand the different molecular mechanisms related to chromosome changes, we perform differential analysis for comparing the AML cells with del(8)(q21.2q24.3) versus diploid AML cells and obtain 266 differentially expressed genes (DEGs). Similarly, we compare the erythroid cells with a gain of chromosome 8 versus diploid erythroid cells and obtain 426 DEGs. Gene set enrichment analysis identify a few shared significant terms between AML and erythroid cells (Figure 9C-D), including multiple immune system related terms, neutrophil degranulation, and cellular responses to stimuli. Some unique terms in AML are hemostasis, platelet activation signaling and aggregation, and arachidonic acid metabolism. Hemostatic and thrombotic complications are prevalent symptoms in AML patients and hemostasis has been studied before for AML pathogenesis-related mechanisms[45]. The term platelet activation signaling and aggregation is also consistent with the previous literature that the platelet defects and other hemorrhagic symptoms are widely observed in AML patients[46]. The arachidonic acid metabolism is a process highlighted in a few cancer research publications, but the evidence for their involvement in AML is still accumulating[47,48]. Overall, our results are consistent with literature and provide some novel disease-related biological processes for future research.

Discussion

We introduce partCNV/partCNVH, a statistical framework that distinguishes neoplastic cells with copy number alterations from normal diploid cells based on regional chromosomal deletions or amplifications. Unlike existing methods, our statistical framework can incorporate prior knowledge from cytogenetic data that includes both chromosomal locations of aberrations and the observed proportion. As demonstrated in our simulation study, this prior information can effectively improve the classification accuracy when the signal is weak. Our framework also includes a feature selection step using the hidden Markov model, which is able to filter the genes when part(s) of the region are diploid. This step further improves the signals of chromosomal changes and results in

higher accuracy to detect neoplastic cells. We have illustrated the benefits of partCNV and partCNVH through extensive simulation studies and in-depth analysis of three scRNA-seq datasets from MDS or AML patients.

Cytogenetic information is a key component of the scoring system of risk assessment, treatment selection, and outcome prediction for patients with hematologic malignancies[49,50]. In this work, we exemplify the use of the proposed methods in patients with MDS and AML. These methods can also be applied to other hematologic malignancies or even other cancer types if similar problems are encountered. As cytogenetic analysis is a mature technology and is widely used in clinical settings, the cytogenetic data should be fairly accessible from clinical collaborators who provide patient samples. Such cytogenetic data-guided analysis can be a useful tool for identifying subgroups of cells with the chromosomal changes of interest.

Implemented in an EM algorithm, our proposed methods have favorable computational performance. For a simulation dataset with 3000 cells and 600 genes, it takes 12 seconds and 21.5 seconds for partCNV and partCNVH to complete the analysis, respectively. This computation cost is similar to existing methods PCA plus Louvain (∼6.8 seconds) and PCA plus Leiden (∼30.7 seconds). In comparison, both inferCNV and CopyKAT take more than 10 minutes, sometimes more than an hour, to process a scRNA-seq dataset with a few thousand cells. Additionally, our proposed methods scale almost linearly to large datasets. If we increase the cell number from 3000 to 6000, partCNV takes about 26.5 seconds and partCNVH 41.5 seconds to complete the computation. Since cytogenetic data are generally unique for each subject, we expect that our methods are applied person by person in real applications. Thus, it is reasonable to assume a few thousand cells in the dataset, for which our methods can complete the computation within one minute.

For future work, the methods can be further extended to incorporate additional biological knowledge, such as the marker or mutation information mentioned in Fan et al[9]. Our method has the potential to be applied on multiple samples in parallel or even to borrow information across different samples from the same subject, such extension is not trivial and needs further evaluations. Moreover, due to the complexity of sequencing depth, gene expression variations, number of genes impacted by the aneuploid event, we have not evaluated the power of the proposed methods or the minimum required size for CNV events. These should be carefully evaluated in future works.

Previous work also developed machine learning models to predict the neoplastic/non-neoplastic status of the cells by splitting the annotated data and training a random forest model[16]. Such models usually rely on the training dataset and may not generalize well to other studies. With the accumulation of annotated single cell data, it is also possible to harness the power of deep learning algorithms to further improve existing models and achieve more accurate predictions. Our current methods assume a fixed prior based on the cytogenetic data. It is possible that different proportions of aneuploid cells have different confidence levels. The current method can be further extended to incorporate such confidence into the model to improve accuracy. When scDNA-seq data is available, it may be another prior knowledge to replace the cytogenetic information of clinical data as the input to PartCNV and PartCNHV. The method can be further extended to incorporate regions where the scDNA-seq data shows enriched hierarchical aneuploid CNVs with high resolution.

Methods

Details of partCNV

We aim to identify the cells with the known chromosomal deletion or amplification from the scRNA-seq data with incorporation of the prior cytogenetic knowledge. Assume a total of N cells were sequenced by scRNA-seq and G genes fall in the region with deletion or amplification. The count matrix is denoted by $\boldsymbol{Y} = (y_{gi})$, which is a $G \times N$ matrix with rows being the genes and columns being the samples. Without loss of generality, we assume the genes in $\boldsymbol{Y}$ are ordered by their locations on the chromosome. As the observed data contain the mixture of cells with and without chromosomal changes, denote the underlying status of the cell i by c_i. Assume the prior proportion of the aneuploid cells is q_0 for the region of interest. In our motivating problem, q_0 is calculated based on the number of metaphases observing the chromosomal changes divided by the total number of metaphases from an cytogenetics test. As shown in some recent literature regarding the distribution of the scRNA-seq count[51,52], the scRNA-seq data may not be zero-inflated and the excessive zeros are due to low expression level of each single cell. Additionally, since the region of interest generally contains a limited number of genes, the estimated dispersion parameters are not accurate enough if we use a negative binomial distribution. Thus, we assume the expression count of gene i follows a Poisson distribution with mean θ_{g1} if the cell is aneuploid or mean θ_{g0} if diploid, i.e.,

$$\Pr(y_{gi}|\theta_{g1}, c_i = 1) = \frac{\theta_{g1}^{y_{gi}} \exp(-\theta_{g1})}{y_{gi}!} \quad \text{and} \quad \Pr(y_{gi}|\theta_{g0}, c_i = 0) = \frac{\theta_{g0}^{y_{gi}} \exp(-\theta_{g0})}{y_{gi}!}.$$

We assume the cell status variable c_i follows a Bernoulli distribution $\Pr(c_i|q_i) = q_i(1 - q_i)$ where q_i denotes the probability of cell i having the chromosomal changes at the region of interest, i.e., $q_i = \Pr(c_i = 1)$. The prior knowledge of the aneuploid cell proportion is best described through a beta distribution, which we approximate through a normal distribution. Though cytogenetic information is obtained based on 20 metaphases, the involved cells can number in the hundreds of thousands, and thus a normal distribution can adequately approximate the underlying beta distribution. We assume q_i follows a prior Normal distribution with mean q_0 and variance λ^2:

$$\Pr(q_i|q_0, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} \exp\left(-\frac{(q_i - q_0)^2}{2\lambda^2}\right).$$

The variance λ^2 represents the confidence about the prior information. Smaller variance indicates stronger confidence in the prior knowledge. However, in our experiments, we found that the actual value of λ has minimal impact on the classification results as long as the value is within a decent range (e.g., λ between 0.01 and 1). Throughout our experiments, we use $\lambda = 0.1$. Together, the full likelihood of the problem can be written as

$$L(\boldsymbol{\theta_1}, \boldsymbol{\theta_0}, \boldsymbol{c}, \boldsymbol{q}|\boldsymbol{Y}, q_0, \lambda) = \prod_i \left[\Pr(y_{gi}|\theta_{g1})\right]^{\mathbb{1}(c_i=1)} \left[\Pr(y_{gi}|\theta_{g0})\right]^{\mathbb{1}(c_i=0)} \Pr(c_i|q_i) \Pr(q_i|q_0, \lambda) \tag{1}$$

and the detailed log-likelihood is

$$\begin{aligned} l(\boldsymbol{\theta_1}, \boldsymbol{\theta_0}, \boldsymbol{c}, \boldsymbol{q}|\boldsymbol{Y}, q_0, \lambda) = &\sum_i \mathbb{1}(c_i = 1)\{y_{gi}\log(\theta_{g1}) - \theta_{g1}\} + \sum_i \mathbb{1}(c_i = 0)\{y_{gi}\log(\theta_{g0}) - \theta_{g0}\} \\ &+ \sum_i c_i \log(q_i) + \sum_i (1 - c_i)\log(1 - q_i) - \sum_i \frac{(q_i - q_0)^2}{2\lambda^2}. \end{aligned}$$

Directly solving the likelihood (1) may not be feasible, and thus we use the EM algorithm by treating cell status $\{c_i\}$ as the missing variables. The objective function of the EM algorithm is

$$\begin{aligned} Q(\boldsymbol{\theta_1}, \boldsymbol{\theta_0}, \boldsymbol{c}, \boldsymbol{q}) = &\sum_i p_i\{y_{gi}\log(\theta_{g1}) - \theta_{g1}\} + \sum_i (1 - p_i)\{y_{gi}\log(\theta_{g0}) - \theta_{g0}\} + \sum_i c_i \log(q_i) \\ &+ \sum_i (1 - c_i)\log(1 - q_i) - \sum_i \frac{(q_i - q_0)^2}{2\lambda^2}. \end{aligned} \tag{2}$$

In the M step of the t-th iteration, we solve for $\theta_{g0}^{(t)}$, $\theta_{g1}^{(t)}$, and $q_i^{(t)}$ as follows:

$$\begin{aligned} \theta_{g1}^{(t)} &= \frac{\sum_i q_i^{(t-1)} y_{gi}}{\sum_i q_i^{(t-1)}}, \\ \theta_{g0}^{(t)} &= \frac{\sum_i \left(1 - q_i^{(t-1)}\right) y_{gi}}{N - \sum_i q_i^{(t-1)}}, \\ q_i^{(t)} &= \arg\max_{q_i} \left[\left(q_i^{(t-1)}\right)^{\sum_i c_i} \left(1 - q_i^{(t-1)}\right)^{N - \sum_i c_i} \exp\left(-\frac{n(q_i^{(t-1)} - q_0)^2}{2\lambda^2}\right) \right]. \end{aligned}$$

In the E step, we estimate the conditional expectation of the cell status by

$$\begin{aligned} p_i^{(t)} = \Pr(c_i = 1|y_{gi}) &= \frac{\Pr(y_{gi}|c_i = 1)\Pr(c_i = 1)}{\Pr(y_{gi}, c_i)} \\ &= \frac{\prod_g \Pr(y_{gi}|\theta_{g1}) \cdot q_i^{(t)}}{\prod_g \Pr(y_{gi}|\theta_{g0}) \cdot q_i^{(t)} + \prod_g \Pr(y_{gi}|\theta_{g0}) \cdot (1 - q_i^{(t)})}. \end{aligned}$$

The algorithm repeats the M step and E step until convergence. The convergence criteria is defined as the absolute difference between $\boldsymbol{p}^{(t)} = (p_1^{(t)}, \cdots, p_N^{(t)})$ and $\boldsymbol{p}^{(t-1)}$ smaller than 10^{-5}. The core of partCNV is the described EM algorithm and the output is the inferred cell status. With the estimated probability $\widehat{\boldsymbol{p}}$, the cells with $\widehat{p}_i \geq 0.5$ are assigned as aneuploid cells and the rest as diploid cells.

Details of partCNVH

Limited by the technology, cytogenetic data only provide crude information about the chromosomal deletion or amplification. When we include all the genes from the regions of interest, it is likely that not all of the genes have chromosomal changes in the aneuploid cells. It is helpful to select the genes that demonstrate chromosomal changing patterns. Motivated by this idea, we design partCNVH, the combination of partCNV and HMM for improving classification performance. Denote the underlying status for the observed G genes by $\boldsymbol{Z} = (z_1, \cdots, z_G)$. The first step of partCNVH is to apply partCNV on the scRNA-seq data from the region of interest. Denote the obtained cell status from partCNV by $\{\widehat{c}_i\}$. Then for each gene g, we compute the mean expression of this gene for the two groups, i.e.,

$$\bar{y}_g^{(1)} = \sum_{j \in \{i: c_i = 1\}} y_{gi} \quad \text{and} \quad \bar{y}_g^{(0)} = \sum_{j \in \{i: c_i = 0\}} y_{gi}.$$

Based on $\bar{y}_g^{(1)}$ and $\bar{y}_g^{(0)}$, we obtain the mean expression ratio for all the genes by $r_g = \bar{y}_g^{(0)}/\bar{y}_g^{(1)}$ if the region has deletion and $r_g = \bar{y}_g^{(1)}/\bar{y}_g^{(0)}$ if the region has amplification. As shown in Figure 3, the signals from the mean expression ratio are weak, and thus we apply a rolling average on the mean expression ratios to strengthen the signals. Denote the window size for the rolling average by K, and the rolling average at gene g becomes

$$\bar{r}_g^{rolling} = \frac{1}{K} \sum_{i=g-[K/2]}^{g+[K/2]} r_i.$$

In R, we implement the rolling average computation by function frollmean from the data.table package. We then build HMM using the rolling average of the genes. Denote the underlying state of gene *g* by z_g, where $z_g = 1$ is gene *g* with deletion or amplification in the aneuploid cells, and $z_g = 0$ otherwise. HMM aims to infer the hidden status $\boldsymbol{Z} = \{z_g\}$ through the observed sequence $\bar{\boldsymbol{R}}^{rolling} = (\bar{r}_g^{rolling})$ by solving the likelihood

$$\Pr(\bar{\boldsymbol{R}}^{rolling}, \boldsymbol{Z}) = \Pr(\bar{\boldsymbol{R}}^{rolling}|\boldsymbol{Z}) \times \Pr(\boldsymbol{Z}) = \prod_{g=1}^{G} \Pr(\bar{r}_g^{rolling}|z_g) \times \prod_{g=1}^{G} \Pr(z_g|z_{g-1}).$$

We solve HMM using the depmix function from the depmixS4 package in R by specifying the initial transition matrix between the two states ("genes in aneuploid region" or "A", "genes in diploid region" or "D") as

$$\begin{pmatrix} A \to A & A \to D \\ D \to A & D \to D \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}.$$

The initial states of the genes are decided as $z_g = 1$ if $\bar{r}_g^{rolling} \geq median\{\bar{r}_g^{rolling}, g = 1, \cdots, G\}$, and $z_g = 0$ otherwise. After the hidden states are inferred by HMM, we identify the states corresponding to the "genes in diploid region" by comparing the mean expression ratio of the two states, and the state with larger mean expression ratio is labeled as state "D". Denote the expression count matrix for the selected genes by $\widetilde{\boldsymbol{Y}}$. We apply the EM algorithm described in partCNV to $\widetilde{\boldsymbol{Y}}$ and infer the final cell states.

Simulation designs

To mimic the real data analysis, we generate the simulation datasets based on existing scRNA-seq data from patients with triple negative breast cancer (TNBC). The data were downloaded from the Gene Expression Omnibus (GEO) with accession number GSE148673. We compared the data characteristics between the TNBC dataset and the scRNA-seq data from our MDS patients. In Supplementary Figure S10, we presented the mean expression levels of genes and cells, as well as the dropout rate for the two datasets. We found that the data characteristics are similar between the two data sources. As a data processing step, genes with zero expression in all cells in the TNBC data are removed. We keep all of the normal cells based on the cell type annotation from the original study[15]. As our method focuses on the region with known deletion or amplification from the cytogenetic data and the region often covers tens or hundreds of genes, we generate the expression count for n_0 normal cells and n_1 aneuploid cells with a total of *G* genes. For each iteration, we randomly draw the expression of *G* genes from the normal cells of the TNBC dataset. We compute the mean expression of these *G* genes and denote it by $\widehat{\boldsymbol{\xi}} = (\xi_1, \cdots, \xi_G)$. For normal cells, the expression is generated from Poisson$(\widehat{\xi_g})$ for $g = 1, \cdots, G$. For aneuploid cells, assume G_1 genes are located at the deleted or amplified regions and $G_0 = G - G_1$ are from the normal regions.

We assume half of the G_0 genes are in the regions that are left-adjacent to a copy-number alteration and the other half are in regions right-adjacent that have normal expression in all cells. This partial chromosomal variation often happens in practice as cytogenetic data only provide a crude observation of the changed regions. Without loss of generality, we assume the beginning G_1 genes are from the aneuploid region. For a gene *g* from this region for an aneuploid cell *i*, we generate the expression from

$$y_{gi} \sim \text{Poisson}\{\lambda_g \cdot (r + \eta)\} \text{ and the noise } \eta \sim \text{Uniform}(0, \tau),$$

where *r* is the ratio controlling the impact level of chromosomal deletion or amplification on the expression and η is the noise, in the first setting with deletion in aneuploid cells. *r* takes value 0.5 or 0.6 in this setting (Simulation data 1). In the second setting, we consider amplification in the aneuploid cells. The expression is generated from

$$y_{gi} \sim \text{Poisson}\{\lambda_g \cdot (r - \eta)\} \text{ and the noise } \eta \sim \text{Uniform}(0, \tau),$$

and *r* takes value 1.5 or 1.4. (Simulation data 2) The noise parameter η controls the heterogeneity of the impacts among all the interested genes, mimicking the fact that gene expressions are heterogeneous when the genes are located in deleted or amplified regions. Through our experiments, we specify $\eta = 0.1, 0.2$, and 0.3 for low, medium, and high noise levels, respectively.

To understand the impact of different sample sizes, we consider the combination of $n_1 = 500$ and $n_0 = 2500$ in one scenario and $n_1 = 300$ and $n_0 = 1000$ in the other (Simulation data 3). We also evaluate the impact of prior specification in all the scenarios. In the first scenario, the true prior proportion of the aneuploid cell is 17% and we evaluate the method with both 17% and 50%. In the second scenario, the true proportion is 23% and we also compare the results with both 23% and 50% (Simulation data 4). To compare the proposed methods versus existing ones that only perform on whole genome data, we randomly sample 2000 normal cells from an existing scRNA-seq dataset (Simulation data 5). We then replace the expression of the genes located at the region chromosome 20 q11.1 to q13.1 using the simulated data as described above to mimic situations that 400 out of

2000 (20%) cells are locally aneuploid. The gene expressions in the region of interest are generated in the similar way as Simulation data 1. We then apply inferCNV and copyKAT using the suggested arguments as suggested by the original methods. All the simulation results are summarized over 100 Monte Carlo datasets.

Real data analysis

The data for the AML patient were downloaded from European Genome-phenome Archive (EGA) with accession EGAD00001007672[44]. The data from sample 7A were used for the analysis due to the available cytogenetic information. The raw data for the two MDS patients are currently not publicly available due to confidentiality regulations.

Data preprocessing

The raw sequencing data were preprocessed (demultiplexed cellular barcodes, read alignment, and generation of gene count matrix) using Cell Ranger Single Cell Software Suite (version 3.0, https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger) provided by 10X Genomics using Human Genome GRCh38. Genes were detected in $< 0.1\%$ of total sequenced cells and cells where < 200 genes had nonzero counts were filtered out and not included in the analysis. Low quality cells where $> 20\%$ of the counts were derived from the mitochondrial genome were also discarded. The doublet cell status was inferred using DoubletFinder[53] and only the singlet cells were kept for further analysis. Data were normalized using the total sequencing count per cell to adjust for differences in sequencing depth. The chromosome database associated with cytogenetic location is downloaded via the UCSC genome website.

Functional over-representation analysis

In the real data analysis section, after the aneuploid/diploid status has been inferred by the proposed or existing method, we perform cell type specific differential analysis using "MAST"[32] (available in Seurat package[2]) to compare the diploid versus aneuploid cells. The genes with an adjusted p value smaller than 0.05 are included as the DEGs. We then perform functional over-representation analysis using the MSigDB platform[33] (http://www.gsea-msigdb.org/gsea/msigdb/annotate.jsp) with the Reactome pathway[39], GO Biological Process[54], and Hallmark[40] databases. We present the top 10 pathways or biological process terms regardless of the significance level.

Implementation of existing methods

All of the existing methods take the normalized expression counts for the genes located in the region of interest as input. K-means and hierarchical clustering are implemented using the functions kmeans and hclust from the stat package in R, respectively. The PCA plus Louvain[55] and Leiden methods[29] are implemented using the Seurat package in function FindClusters with arguments algorithm$= 1$ and 4. Since the Seurat clustering function does not allow specification of cluster numbers, we design a loop with a precision parameter ranging from 0.001 to 0.5 with distance 0.005 until the algorithm identifies exactly two clusters. All the analyses and plotting are performed in R v4.0.3.

Data availability

The TNBC scRNA-seq data are downloaded from GEO with accession number GSE148673. The AML scRNA-seq data were donwloaded from European Genome-phenome Archive (EGA) with accession EGAD00001007672. The raw data for the two MDS patients are currently not publicly available due to confidentiality regulations. The processed data are available from the investigators upon reasonable request. Our software is publicly available at GitHub (https://github.com/ziyili20/partCNV) and the Bioconductor site (https://bioconductor.org/packages/devel/bioc/html/partCNV.html).

Received: 4 May 2023; Accepted: 3 October 2024
Published online: 15 October 2024

References

1. Saliba, A.-E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic acids research* **42**, 8845–8860 (2014).
2. Shalek, A. K. et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363–369 (2014).
3. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
4. Mathys, H. et al. Single-cell transcriptomic analysis of alzheimerâ€™s disease. *Nature* **570**, 332–337 (2019).
5. Lescroart, F. et al. Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. *Science* **359**, 1177–1181 (2018).
6. Suvà, M. L. & Tirosh, I. Single-cell RNA sequencing in cancer: lessons learned and emerging challenges. *Molecular cell* **75**, 7–12 (2019).
7. Chung, W. et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications* **8**, 1–12 (2017).
8. Shih, A. J. et al. Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq. *PLoS one* **13**, e0206785 (2018).
9. Fan, J., Slowikowski, K. & Zhang, F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Experimental & Molecular Medicine* **52**, 1452–1465 (2020).
10. Le, N. T. & Richardson, D. R. The role of iron in cell cycle progression and the proliferation of neoplastic cells. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer **1603**, 31–46 (2002).
11. Farahinia, A., Zhang, W. & Badea, I. Novel microfluidic approaches to circulating tumor cell separation and sorting of blood cells: A review. *Journal of Science: Advanced Materials and Devices* **6**, 303–320 (2021).

12. Cieślik, M. & Chinnaiyan, A. M. Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics* **19**, 93–109 (2018).
13. Peng, J. et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell research* **29**, 725–738 (2019).
14. Fan, J. et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome research* **28**, 1217–1227 (2018).
15. Gao, R. et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nature biotechnology* **39**, 599–608 (2021).
16. van Galen, P. et al. Single-cell RNA-seq reveals aml hierarchies relevant to disease progression and immunity. *Cell* **176**, 1265–1281 (2019).
17. Milite, S., Bergamin, R., Patruno, L., Calonaci, N. & Caravagna, G. A bayesian method to cluster single-cell rna sequencing data using copy number alterations. *Bioinformatics* **38**, 2512–2518 (2022).
18. Campbell, K. R. et al. clonealign: statistical integration of independent single-cell rna and dna sequencing data from human cancers. *Genome biology* **20**, 1–12 (2019).
19. Bai, X., Duren, Z., Wan, L. & Xia, L. C. Joint inference of clonal structure using single-cell genome and transcriptome sequencing data. bioRxiv 2020–02 (2020).
20. Mrozek, K., Heerema, N. A. & Bloomfield, C. D. Cytogenetics in acute leukemia. *Blood reviews* **18**, 115–136 (2004).
21. Gersen, S. L., Keagle, M. B., Gersen, S. & Keagle, M. The principles of clinical cytogenetics (Springer, 2013).
22. Malcovati, L. et al. Impact of the degree of anemia on the outcome of patients with myelodysplastic syndrome and its integration into the who classification-based prognostic scoring system (wpss). haematologica **96**, 1433 (2011).
23. Della Porta, M. G. et al. Validation of who classification-based prognostic scoring system (wpss) for myelodysplastic syndromes and comparison with the revised international prognostic scoring system (ipss-r). a study of the international working group for prognosis in myelodysplasia (iwg-pm). Leukemia **29**, 1502–1513 (2015).
24. Schanz, J. et al. New comprehensive cytogenetic scoring system for primary myelodysplastic syndromes (mds) and oligoblastic acute myeloid leukemia after mds derived from an international database merge. *Journal of Clinical Oncology* **30**, 820 (2012).
25. Löwenberg, B. Prognostic factors in acute myeloid leukaemia. *Best Practice & Research Clinical Haematology* **14**, 65–75 (2001).
26. Heerema, N. et al. Deletion of 7p or monosomy 7 in pediatric acute lymphoblastic leukemia is an adverse prognostic factor: a report from the children's cancer group. *Leukemia* **18**, 939–947 (2004).
27. Heim, S. & Mitelman, F. Cancer cytogenetics: chromosomal and molecular genetic aberrations of tumor cells (John Wiley & Sons, 2015).
28. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**, 1–22 (1977).
29. Traag, V. A., Waltman, L. & Van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports* **9**, 1–12 (2019).
30. Sibon, D. et al. Enhanced renewal of erythroid progenitors in myelodysplastic anemia by peripheral serotonin. *Cell Reports* **26**, 3246–3256 (2019).
31. Lanser, L. et al. Inflammation-induced tryptophan breakdown is related with anemia, fatigue, and depression in cancer. *Frontiers in immunology* **11**, 249 (2020).
32. Finak, G. et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome biology* **16**, 1–13 (2015).
33. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics **27**, 1739–1740 (2011).
34. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nature genetics* **25**, 25–29 (2000).
35. Cao, M. et al. Mechanisms of impaired neutrophil migration by microRNAs in myelodysplastic syndromes. The Journal of Immunology **198**, 1887–1899, https://doi.org/10.4049/jimmunol.1600622 (2017). https://www.jimmunol.org/content/198/5/1887.full.pdf.
36. Starczynowski, D. T. & Karsan, A. Innate immune signaling in the myelodysplastic syndromes. *Hematology/Oncology Clinics* **24**, 343–359 (2010).
37. Warlick, E. D. & Miller, J. S. Myelodysplastic syndromes: the role of the immune system in pathogenesis. *Leukemia & lymphoma* **52**, 2045–2049 (2011).
38. Wang, C. et al. Immune dysregulation in myelodysplastic syndrome: Clinical features, pathogenesis and therapeutic strategies. *Critical reviews in oncology/hematology* **122**, 123–132 (2018).
39. Fabregat, A. et al. The reactome pathway knowledgebase. *Nucleic acids research* **46**, D649–D655 (2018).
40. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell systems* **1**, 417–425 (2015).
41. Andreansky, M., Fiederlein, R. & Graham, M. Busulfan and melphalan as preparative therapy for stem cell transplantation in pediatric patients with acute myelogenous leukemia (aml) and myelodysplasia (mds). *Biology of Blood and Marrow Transplantation* **12**, 85 (2006).
42. Ostrand-Rosenberg, S. Cancer and complement. *Nature biotechnology* **26**, 1348–1349 (2008).
43. Yoyen-Ermis, D. et al. Myeloid maturation potentiates stat3-mediated atypical ifn-γ signaling and upregulation of pd-1 ligands in aml and mds. *Scientific reports* **9**, 1–11 (2019).
44. Abbas, H. A. et al. Single cell t cell landscape and t cell receptor repertoire profiling of aml in context of pd-1 blockade therapy. *Nature communications* **12**, 1–13 (2021).
45. Tallman, M. S. & Kwaan, H. C. Reassessing the hemostatic disorder associated with acute promyelocytic leukemia. *Blood* **79**, 543–553 (1992).
46. Bumbea, H. et al. Platelet defects in acute myeloid leukemiaâ€"potential for hemorrhagic events. *Journal of Clinical Medicine* **11**, 118 (2021).
47. Hyde, C. & Missailidis, S. Inhibition of arachidonic acid metabolism and its implication on cell proliferation and tumour-angiogenesis. *International immunopharmacology* **9**, 701–715 (2009).
48. Yang, P. et al. Arachidonic acid metabolism in human prostate cancer. *International journal of oncology* **41**, 1495–1503 (2012).
49. Harris, N. L. et al. The world health organization classification of neoplastic diseases of the hematopoietic and lymphoid tissues: report of the clinical advisory committee meeting, airlie house, virginia, november, 1997. *Annals of Oncology* **10**, 1419–1432 (1999).
50. Sole, F. et al. Identification of novel cytogenetic markers with prognostic significance in a series of 968 patients with primary myelodysplastic syndromes. haematologica **90**, 1168–1178 (2005).
51. Kim, T. H., Zhou, X. & Chen, M. Demystifying, "drop-outs" in single-cell UMI data. *Genome biology* **21**, 1–19 (2020).
52. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* **38**, 147–150 (2020).
53. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. Doubletfinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell systems* **8**, 329–337 (2019).
54. Botstein, D. et al. Gene ontology: tool for the unification of biology. *Nat genet* **25**, 25–29 (2000).
55. Que, X., Checconi, F., Petrini, F. & Gunnels, J. A. Scalable community detection with the louvain algorithm. In 2015 IEEE International Parallel and Distributed Processing Symposium, 28–37 (IEEE, 2015).

Acknowledgements
The authors acknowledge the language editing service by Ms. Jessica Swann at MD Anderson Cancer Center.

Author contributions
ZL and WS conceived the idea and designed the statistical methods. ZL implemented the methods, performed the simulation studies, and analyzed the real datasets with help from RL. RL and ZL prepared the package. IGG, SC, HA, and GGM provided the real datasets that motivates the problem and provided clinical insights for result interpretations. ZL, RL, and WS wrote the manuscript, with inputs from IGG, SC, HA, and GGM.

Funding
ZL and RL were partially funded by the National Cancer Institute grant R03CA270725. HAA was partially funded by Physician Scientist Award Grant from MDACC. WS was partially funded by the National Institute of Health grant R01GM105785.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-75226-2.

Correspondence and requests for materials should be addressed to Z.L. or W.S.