


METHOD

Open Access



scCTS: identifying the cell type-specific marker genes from population-level single-cell RNA-seq

Luxiao Chen^{1†}, Zhenxing Guo^{2†}, Tao Deng^{2,3} and Hao Wu^{4,5*} 

[†]Luxiao Chen and Zhenxing Guo contributed equally to this work.

*Correspondence: wuhao@siat.ac.cn

¹ Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA

² School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-SZ), Shenzhen 518172, Guangdong, China

³ Shenzhen Research Institute of Big Data, Shenzhen 518172, China

⁴ Faculty of Computer Science and Control Engineering, Shenzhen University of Advanced Technology, Shenzhen 518055, Guangdong, China

⁵ Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, Guangdong, China

Abstract

Single-cell RNA-sequencing (scRNA-seq) provides gene expression profiles of individual cells from complex samples, facilitating the detection of cell type-specific marker genes. In scRNA-seq experiments with multiple donors, the population level variation brings an extra layer of complexity in cell type-specific gene detection, for example, they may not appear in all donors. Motivated by this observation, we develop a statistical model named scCTS to identify cell type-specific genes from population-level scRNA-seq data. Extensive data analyses demonstrate that the proposed method identifies more biologically meaningful cell type-specific genes compared to traditional methods.

Keywords: Single-cell RNA-seq, Cell type-specific genes, Differential expression, Hierarchical model

Background

Single-cell RNA sequencing (scRNA-seq) allows the quantification of gene expression levels in individual cells [1–3]. In recent years, the scRNA-seq technologies have been successfully applied to answer a variety of biological questions, for example, to discover new cell types [4], estimate cellular composition in tissue samples [5], uncover novel biological mechanisms in different biological systems [6–8], etc. Compared to traditional bulk RNA sequencing (RNA-seq), the major advantage of scRNA-seq is that the single-cell expression provides information for understanding the cellular heterogeneity of complex samples. A major source of cellular heterogeneity is the cell types, that is, a complex sample usually consists of many different types of cells which are functionally different. Traditionally, the cell types are defined by their morphological or phenotypical features. An often-used technology to define cell type is to use flow cytometry to sort cells according to certain cell surface markers. With the gene expression data, the cell types can be defined by the expression values of some cell type-specific (CTS) genes, which have distinct gene expression profiles in different cell types.



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

The CTS genes are defined as the genes with strong differential expression among cell types. These genes are often of great interest because they are closely related to the cellular identity and function, and potentially the pathologies of different diseases [9–11]. They are also useful in various data analysis tasks including cell type annotation and identification in scRNA-seq [12] and bulk data deconvolution [13]. For example, in scRNA-seq data analyses, a fundamental step is to identify the cell types for all cells. There are many cell type annotation methods, either unsupervised [14–16] or supervised [17–20]. Most of these methods contain a feature selection step where only the expression values for the CTS genes are used. The non-marker genes express uniformly in all cell types, thus do not contain information of cell types. Therefore, the feature selection step enhances the signal to noise ratio in the data and will lead to better results.

Studies on CTS genes have a long history. Before the wide application of high-throughput technologies, only limited number of CTS genes can be identified with low-throughput techniques such as western blot, northern blot, or RT-qPCR. Researchers have manually curated CTS genes to systematically study cell types under different conditions [21]. With the advances of high-throughput technologies (e.g., gene expression microarray or RNA-seq), CTS genes can be identified more efficiently. However, these methods require expressions from purified cell types to call CTS genes. The purification of cell types requires cell sorting which is expensive and laborious. In addition, cell sorting relies on specific cell surface markers, which are not always available. Compared to the traditional methods, scRNA-seq provides a much easier and efficient way to identify the CTS genes, i.e., they can be identified by differential expression analysis among cell types.

Various methods have been applied to identify CTS genes from scRNA-seq data. There are methods based on regular statistical tests for differential expression (DE) analysis, for example, Wilcoxon rank sum test and Student's *t* test are implemented in Seurat [22] and Scanpy [23]. There are also more sophisticated methods like Necessary and Sufficient Forest (NS-Forest), which leverages the non-linear attributes of random forest feature selection to identify markers that are highly expressed in one specific cell type only [24]. Moreover, CTS genes can also be identified by feature selection methods like FEAST [25], ELF [26], and scGeneFit [27], which are designed to select the most representative marker genes for cell clustering. ZINB-WaVe + DESeq2 selects markers from differentially expressed genes across different cell types [28].

All these methods were developed in early scRNA-seq days where the study only contains one or a few subjects. With the cost reduction, people start performing large-scale population level scRNA-seq studies [29–34]. To analyze scRNA-seq data from multiple subjects, all the existing methods ignore one important factor: the between subject heterogeneity. Their common approach is to pool cells from all subjects and perform differential expression test. Since the CTS genes are not guaranteed to appear in all subjects due to biological or technical reasons, such an approach could lead to undesirable results. For example, marker genes with low prevalence in the population will likely to be missed. Our major motivation in this work is to design a method for calling CTS marker genes in population-level scRNA-seq data, with consideration of subject heterogeneity in a rigorous way. The result from this type of analysis is both interesting and useful. Biologically, one wants to know the behavior of CTS genes, i.e., whether they would consistently show up in a population, or only appear in a proportion of the subjects.

Computationally, the CTS genes are used in several other tasks such as bulk data deconvolution and cell type identification, so their consistency is important. For example in supervised cell type identification, CTS genes are implicitly assumed to appear in both reference and target samples [18, 19, 35–37]. If this is violated, the result would suffer.

There are some previous works considering the consistency of CTS genes cross subjects. CellMarker is a manually curated resource that provides CTS genes either from scRNA-seq or from other experimental research in human and mouse [38]. In the CellMarker database, a CTS gene with more resources reported indicates greater consistency. GeneMarkeR is another database that provides manually curated CTS genes from published results [39]. It transforms marker gene statistics across publications to a “marker gene score” ranging from 0 to 1. A robust CTS gene should have marker gene score greater than 0.5 and be specific to at most two cell types. Fischer and Gillis [40] identified replicable CTS genes from Brain Initiative Cell Census Network (BICCN) [41, 42] based on two metrics: area under the receiver-operator curve (AUROC) and fold change, and demonstrated that they can improve bulk sample deconvolution and cell typing performance. Even though these works have provided valuable information about robust CTS genes, the methods are ad hoc and cannot be directly applied to analyzing new datasets.

In this work, we develop a novel statistical method named scCTS to identify CTS genes from population-level scRNA-seq data. We define a CTS gene as the one showing differential expression between one cell type and the others. For a gene, we consider both its frequency of being a CTS in a population and the strength of the differential expression in a Bayesian hierarchical model, and call CTS genes based on the derived posterior probability. Our method can identify different types of CTS genes, for example, the ones showing strong DE signal in only a small proportion of subjects, or the ones consistently showing weak DE signals across subjects. Real data analyses demonstrate that our method identifies CTS genes with more biological relevance, as well as provides more detailed characteristics for the CTS genes such as consistency and DE strength in a population.

Results

Method overview

scCTS incorporates between-subject heterogeneities in a hierarchical model to detect CTS markers from the population. Briefly speaking, for a specific cell type k , we first define a set of binary random variables to represent the underlying status for all genes being cell type specific. Then, for a particular gene g , if it is a marker of cell type k , we further define binary random variables to indicate its DE states in each subject, with a conditional prevalence prior. If gene g is not a marker, its conditional prevalence probability equals to zero and it will not show DE in any subject. Given a realization of all genes' DE states in all subjects for cell type k , we then model their within-subject log fold changes using normal distributions, with positive means for markers showing DE and zero mean otherwise. Complete data likelihood is established for each cell type by combining the conditional distribution of log fold changes and the prior of all latent variables. The posterior probability of a gene being a CTS marker in each cell type given its log fold change can be calculated and used to infer its CTS marker status in the population.

scCTS does not require CTS marker genes consistently show DE in all subjects. Instead, it assumes a probability for them to show DE in a randomly picked subject. Such an assumption and the modeling on log fold change allow scCTS to flexibly detect CTS markers with different characteristics: (1) consistently show strong DE signal in most subjects (high prevalence and large fold changes); (2) consistently show weak DE signals in most subjects (high prevalence but small fold changes); (3) show strong DE signals in only few subjects (low prevalence but large fold changes). In the following sections, we show the advantages of scCTS over Wilcoxon rank sum test using both simulations and real data analyses.

Data description

Two sets of scRNA-seq data are used as blueprint for simulation and for real data analyses. The first one is a PBMC Lupus data, which contains twenty-four samples from sixteen individuals. The samples come from two batches: in the first batch, there are eight control samples from eight individuals with systemic lupus erythematosus (SLE) disease; in the second batch, there are eight control samples and eight IFN-beta stimulated samples from another eight individuals with SLE disease. In each sample, there are seven cell types: B cells, CD14 + Monocytes, CD4 T cells, CD8 T cells, Dendritic cells, FCGR3A + Monocytes, and NK cells (Megakaryocytes were excluded due to its extremely small composition in samples). The second dataset is one COVID-19 dataset, which contains in total 284 samples from 196 individuals, including 171 COVID-19 patients (22 patients with mild/moderate symptoms, 54 hospitalized patients with severe symptoms, and 95 recovered convalescent persons), and 25 healthy controls. The data were obtained from various tissue types including human peripheral blood mononuclear cells (PBMCs), bronchoalveolar lavage fluid, and pleural effusion/sputum. In our analyses, we only select frozen PBMCs from healthy controls and severe patients without comorbidities, resulting in 13 samples in total. For each of the selected sample, eight major cell types are kept: B cells, CD14 + Monocytes, CD 4 T cells, CD8 T cells, dendritic cells, NK cells, Plasma cells, and Megakaryocytes cells.

Simulation results

Overview of simulation strategy

We conduct simulations to evaluate the performance of scCTS. We compare with traditional strategies where one pools cells from all subjects and perform Wilcoxon rank sum test, NS-Forest, and ZINB-WaVe + DESeq2. The simulations are constructed based on the PBMC Lupus data. Briefly, we simulate expression profiles of 6231 genes for 10,000 cells from 16 distinct subjects, with each subject having the same four different cell types. Both proportion of cells allocated to each subject, and the within-subject cell type proportions, as shown in Fig. 1a, align with real data observations. To pertain the heterogeneity in baseline expression across different subjects, mean expression of CD4 T cells from all 16 subjects in the PBMC lupus data are extracted and adopted as the baseline expression of cell type 1–4 (Additional file 1: Figure S1(a)). We assume 5% of all genes are CTS marker genes for each cell type, and their population prevalence positively correlates with their effect size estimated from real data (details in the “Methods” section). In our simulation, effect sizes of markers of cell type 1–4 are simulated based

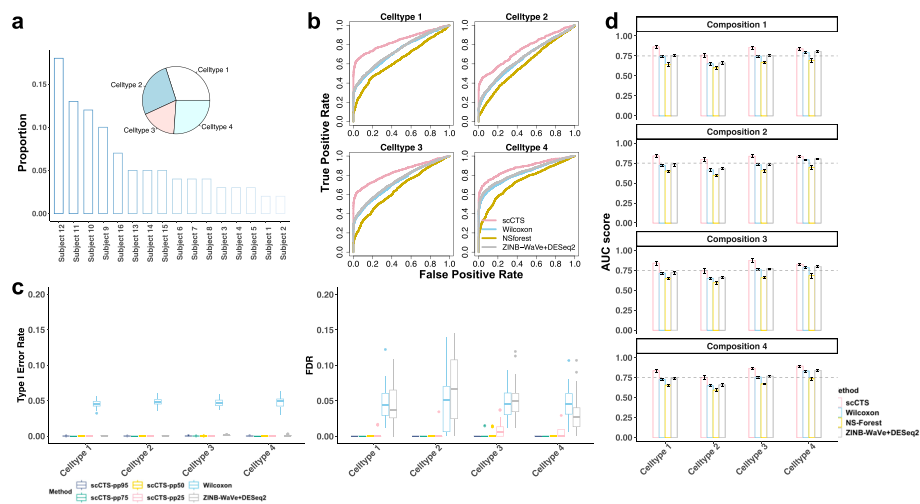


Fig. 1 Simulation results. **a** Barplot showing proportion of cells assigned to each subject in simulations, together with the pie chart displaying one set of proportions for the four simulated cell types. **b** ROC curves of detecting cell type-specific DE genes for all cell types, under proportions in **a**. The methods under comparison are scCTS, Wilcoxon rank sum test, NS-Forest and ZINB-WaVe + DESeq2. **c** Type I error rate and false discovery rate from scCTS, Wilcoxon rank sum test, and ZINB-WaVe + DESeq2. For scCTS, different thresholds (i.e., 0.25, 0.5, 0.75, 0.95) are applied onto posterior probabilities for the identification of CTS markers. **d** AUC scores of scCTS, Wilcoxon rank sum test, NS-Forest, and ZINB-WaVe + DESeq2 in detecting CTS DE genes, under four different cell type compositions suggested in Figure S2(e). Both results in **b**, **c**, and **d** are summarized from 100 simulations

on empirical positive estimates respectively for B cells, CD8 T cells, FCGR3A + Monocytes cells, and NK cells in the PBMC Lupus data. Because of the positive correlation between prevalence and effect size, some markers will have high prevalence and show DE effects in all subjects, while the others may occur in only a few of subjects due to low prevalence (Additional file 1: Figure S1(b) & Figure S1(c)). This characteristic resembles observations from the real data in Fig. 2. Under above simulation strategies, our synthetic data mimics real data well in the aspect of gene-wise mean expression (Additional file 1: Figure S1(d)). We then apply scCTS and its alternatives to detect CTS markers. The accuracy of prediction is evaluated by ROC curve and the area under the ROC curve, by averaging results of 100 simulations. Detailed simulation procedures are provided in the “Methods” section.

scCTS provides better CTS gene detection

As shown in Fig. 1b, the scCTS generates the highest ROC curves regarding the accuracy of detecting CTS markers compared to Wilcoxon rank sum test, NS-Forest, and ZINB-WaVe + DESeq2, and this performance gain is consistent among all four cell types with varied effect sizes and proportions. For cell type 2 whose markers generally show the weakest effect of DE compared to the other three cell types, all the four methods output the worst prediction results, but scCTS still maintains the highest accuracy than the other three alternatives, suggesting its stronger robustness against weak effect sizes. Results displayed in Fig. 1b are obtained when data were simulated using cell type proportions from Fig. 1a, which are in fact derived by normalizing the proportion of B cells, CD8 T cells, FCGR3A + Monocytes cells and NK cells in PBMC

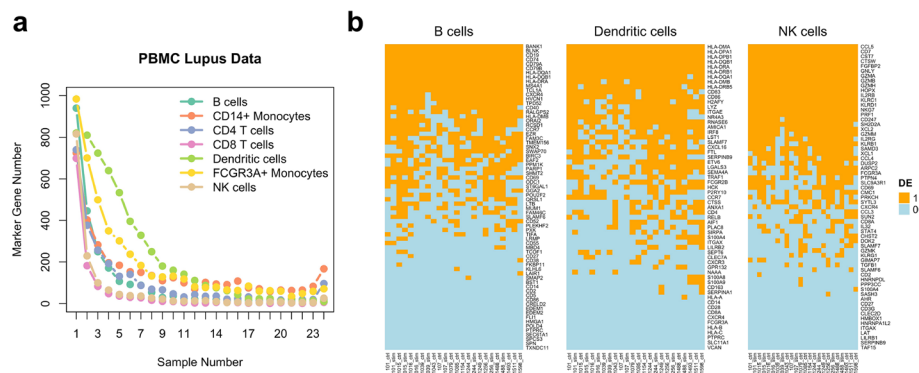


Fig. 2 CTS genes do not consistently appear in all samples. **a** Numbers of genes called as DE by Wilcoxon rank sum test in different numbers of samples for different PBMC cell types (B cells, CD14+ Monocytes, CD4 T cells, CD8 T cells, Dendritic cells, FCGR3A+ Monocytes, and NK cells). The y-axis represents the number of genes called DE by Wilcoxon rank sum test with $FDR < 0.05$ in different numbers of samples. The x-axis represents the number of samples (from 1 to 24). Different colors represent different cell types. **b** Heatmap represents DE state of CTS genes reported by GeneMarker or PanglaoDB in 24 samples of PBMC Lupus data for three cell types (B cells, Dendritic cells, and NK cells). Genes are sorted by the number of samples showing DE. The DE state represents whether the CTS genes can be called as DEG (one vs. others) by Wilcoxon rank sum test with $FDR < 0.05$ in one sample (1: yes; 0: no)

Lupus dataset. To explore the robustness against different cell type proportions, we exhaust 35 combinations by randomly sampling four cell types from the original seven cell types and display their proportion combinations in Additional file 1: Figure S1(e). As shown, four major groups are observed, and the mean of each group only supports one cell type to be the majority. With redefined cell type proportions, as summarized in Fig. 1c, still the proposed method consistently reports the highest AUC scores than the other three compared methods. Overall, the proposed method is more accurate in detecting cell type specific markers and is more robust to weak effect sizes and varied cell type compositions. This is not surprising, since scCTS captures more information in the data. For example, if a CTS gene has low prevalence but high conditional DE signal, the signal will be diluted when pooling data from all individuals. Thus, traditional methods will fail to detect this gene.

In addition to accuracies in overall ranking of detected CTS marker, we also examine the statistical inference from scCTS. We investigate the type I error rate and false discovery rate under nominal level of 0.05 from scCTS and compare it to the alternatives. For scCTS, instead of P -value, it relies on the posterior probability, where genes with posterior probability greater than certain threshold will be reported as CTS markers. Four monotonically increasing thresholds (i.e., 0.25, 0.5, 0.75, 0.95) are selected to explore their effects on final inference, the higher the stronger statistical evidence for being a CTS marker. For comparison, we also incorporate Wilcoxon rank sum test and ZINB-WaVe + DESeq2 into our evaluation. As shown in the left panel of Fig. 1c, scCTS and ZINB-WaVe + DESeq2 are more conservative than Wilcoxon test. Similar results were observed for FDR. We further explore the results and find the conservativeness is partially due to the first step filtering, where many genes with low prevalence and DE signal are filtered out in scCTS. These findings suggest that scCTS requires stronger evidence in detecting CTS markers. In DE analysis,

being conservative is usually better than liberal, and the ranking of DE genes is more important than the DE gene list itself. In real practice, investigators often take the top ranked genes for further analyses and validation. The ROC curves show that scCTS provides better DE genes ranking.

Real data analysis

We carefully analyze the two real datasets (PBMC Lupus and COVID-19) to compare scCTS with NS-Forest, ZINB-WaVe + DESeq2, scGeneFit, FEAST, and Wilcoxon rank sum test. The Wilcoxon rank sum test can be replaced by other method such as two-group t -test, but we found that these two approaches provide similar results. Thus, we do not include t -test in the comparison. There are different ways of applying the alternative methods. The common practice to apply alternative methods such as Wilcoxon rank sum test in Seurat is by pooling cells from all subjects. This requires a data integration step to remove technical artifacts such as batch effects. One can also apply traditional alternative methods on each subject separately and then combine the results from all subjects. Below we compare scCTS with both approaches.

CTS genes do not consistently appear across samples

We first explore the characteristics of the CTS genes detected from Wilcoxon rank sum test. For the PBMC Lupus dataset, we perform Wilcoxon rank sum test in each subject for each cell type (one vs. all others). The DE genes (CTS genes) were called by FDR < 0.05. From these results, we find that only a small proportion of genes are called as DEGs across all samples (Fig. 2a). For example, in CD4 T cells, there are totally 2529 genes called as DEGs in at least one sample among all 6231 genes. However, only 96 genes are called as DE in all 24 samples, while 740 genes are called as DE in only one sample. Same trend can be observed in other cell types (e.g., CD8 T cells, NK cells) with varying number of DEGs called across samples (Additional file 1: Table S1). We then collect CTS genes of PBMC cell types reported by GeneMarker [39] and PanglaoDB [43] and check whether these CTS genes can be identified by Wilcoxon rank sum test. For CTS markers reported by GeneMarker or PanglaoDB, we find that only part of them consistently show DE signal in all samples (Fig. 2b). For example, in B cells, CTS genes like *CD19*, *CD79A*, and *CD79B* are called as DE in all samples; but other CTS genes like *LTB*, *TMEM156* [39] are only called as DE in some samples (*LTB*: 10 out of 24 samples, *TMEM156*: 17 out of 24 samples). These results imply that CTS genes may not consistently appear in all samples (even under the same experimental condition). Thus, a thorough evaluation of CTS genes consistency across samples is needed for both biological understanding of different cell types and downstream analyses like cell typing or bulk sample deconvolution.

CTS genes called by scCTS show different characteristics

Next, we apply scCTS on the PBMC Lupus data to call CTS genes for different cell types. We set the threshold for LFC in estimation procedure loosely as 0 to ensure more CTS genes will be detected. The CTS genes are called by $P(D_{gk} = 1|Y_g) > 0.95$ for cell type $k = 1, \dots, K$ and gene $g = 1, \dots, G$. The genes called as CTS genes for one cell type have different characteristics: probability q_g measuring consistency of DE signal across

samples, mean and variance of LFC (m_g and τ_g^2) measuring strength of DE signals across samples (Fig. 3a, Additional file 1: Table S2).

The proposed method detects different types of CTS genes. First, some of the CTS genes have large LFC ($m_g > 1$) and high consistency ($q_g > 0.9$), for example, *CD14*, *FTL*, and *TYROBP* in CD14+ Monocytes. These three genes are well-known CTS genes for Monocytes. Our method identifies all of them but indicates that they have very different LFC variances across samples. Smaller LFC variance represents more stable cell type-specific gene expression signal in the population. Thus, with comparable mean LFC level, CTS genes with smaller LFC variance is more preferred for analysis like bulk sample

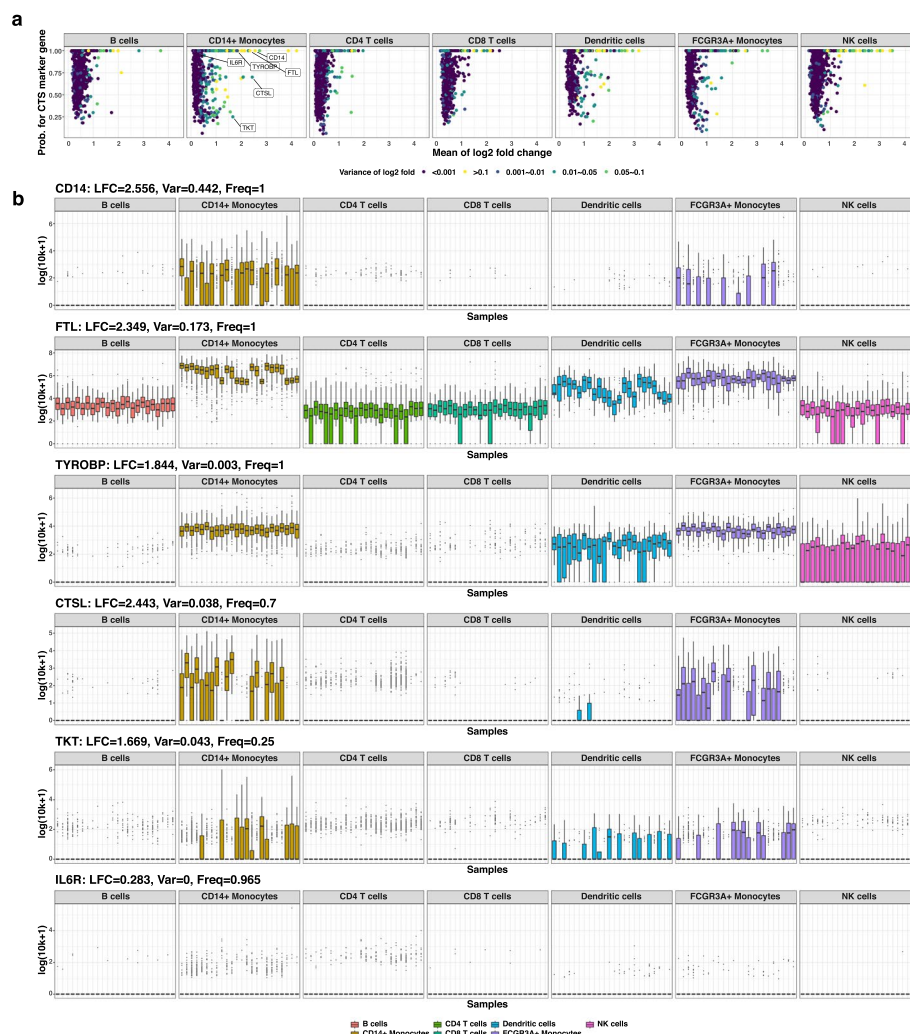


Fig. 3 Characteristics of CTS genes identified from samples. **a** Scatter plots showing different characteristics of identified CTS genes in PBMC cell types (B cells, CD14 + Monocytes, CD4 T cells, CD8 T cells, dendritic cells, FCGR3A + Monocytes, and NK cells). The y-axis represents estimated frequency of a CTS gene (q_g) showing DE signal across samples, which measures consistency. The x-axis represents the mean value of log2 fold change (m_g) of CTS genes in analyzed samples. The color of the points represents the variance of log2 fold change (τ_g^2) of CTS genes in analyzed samples (purple: small variance; yellow: large variance). **b** Boxplots of gene expression of all cells in different cell types for 24 samples. Six example CTS genes of CD14 + Monocytes (*CD14*, *FTL*, *TYROBP*, *CTSL*, *TKT*, and *IL6R*) are shown. They have different mean values of log2 fold change (LFC), variances of LFC (Var), and different probabilities to show DE signal (Freq) in samples. The y-axis is the log transformed 10 k counts. The x-axis represents samples

deconvolution, in which a fixed gene expression profile is used as reference for all samples. This variance difference (*CD14*: $\tau_g^2 = 0.442$; *FTL*: $\tau_g^2 = 0.173$; *TYROBP*: $\tau_g^2 = 0.003$) can be clearly observed in boxplot of gene expression in CD14+ Monocyte cells across 24 samples (Fig. 3b). Compared to gene *CD14* ($\tau_g^2 = 0.442$) and *FTL* ($\tau_g^2 = 0.173$), gene *TYROBP* has much smaller LFC variance ($\tau_g^2 = 0.003$) and its expression is more consistent across samples in both CD14+ Monocytes and other cell types (Fig. 3b). In contrast, gene *CD14* and *FTL* have greater expression variation in CD14+ Monocytes.

In some cell types (e.g., CD14+ Monocytes and Dendritic cells), scCTS also identifies some CTS genes have large LFC ($m_g > 1$), but lower consistency (e.g., $q_g < 0.9$) (Fig. 3a). For example, gene *CTSL* ($q_g = 0.7$) only shows high expression in CD14+ Monocytes in 16 out of 24 samples, while gene *TKT* ($q_g = 0.25$) in 8 out of 24 samples (Fig. 3b). Moreover, scCTS also identified some CTS genes with small LFC ($m_g < 0.3$) but very high consistency ($q_g > 0.9$) (Fig. 3a). One example gene is *IL6R* (estimated frequency is 0.96), which is called as DE in only 18 out of 24 samples by Wilcoxon rank sum test for CD14+ Monocytes. We can observe that *IL6R* has higher proportion of cells with non-zero counts in CD14+ Monocytes than in other cell types (Fig. 3b). Overall, these results show that scCTS detects CTS marker genes with different characteristics and provides more information for the marker gene properties including consistency and differential expression strength.

Comparison of the CTS genes detected by scCTS and Wilcoxon test

We carefully compare the CTS genes detected by scCTS (referred to as “s-markers”) and Wilcoxon rank sum test (“w-markers”). Note here we apply Wilcoxon test on each sample individually to call marker genes. Below we mainly present the comparison from the PBMC Lupus data (Fig. 4, Table 1). Among all genes called as CTS genes by scCTS, Table 1 shows the number of genes also called by Wilcoxon rank sum test in different number of samples. These numbers show that most scCTS marker genes only show in a proportion of samples. For example, in B-cell, 624 out of a total of 933 scCTS genes show up in only 1–8 sample. There are also some genes showing in 0 samples by Wilcoxon test, but called by scCTS. These are genes with low signals in all samples, so they are failed to be detected by Wilcoxon test.

There are also some w-markers not called as s-markers, as those are the ones with negative or very small positive LFC defined as Eq. (2) (blue and gold points in Fig. 4a). Since we are only interested in CTS genes with higher expression in cells from target cell type than from other cell types, these genes are not called as s-markers. One example is gene *CD74* in CD14+ Monocytes. In CD14+ Monocytes, *CD74* has much higher expression than in CD4 T cells, NK cells, and CD8 T cells, but much lower expression values than in B cells and Dendritic cells (Fig. 4b). It is more reasonable to define *CD74* as CTS gene for B cells and Dendritic cells instead of CD14+ Monocytes in these samples. The significant Wilcoxon rank sum test statistic of *CD74* is due to much higher proportion of CD14+ Monocytes than B cells and Dendritic cells (Additional file 1: Table S3, Figure S2), which leads to higher rank for expression in cells of CD14+ Monocytes. Besides, there are some w-markers with very small positive average LFC across samples (gold in Fig. 4a) that their DE signals are too weak to be called as s-markers.

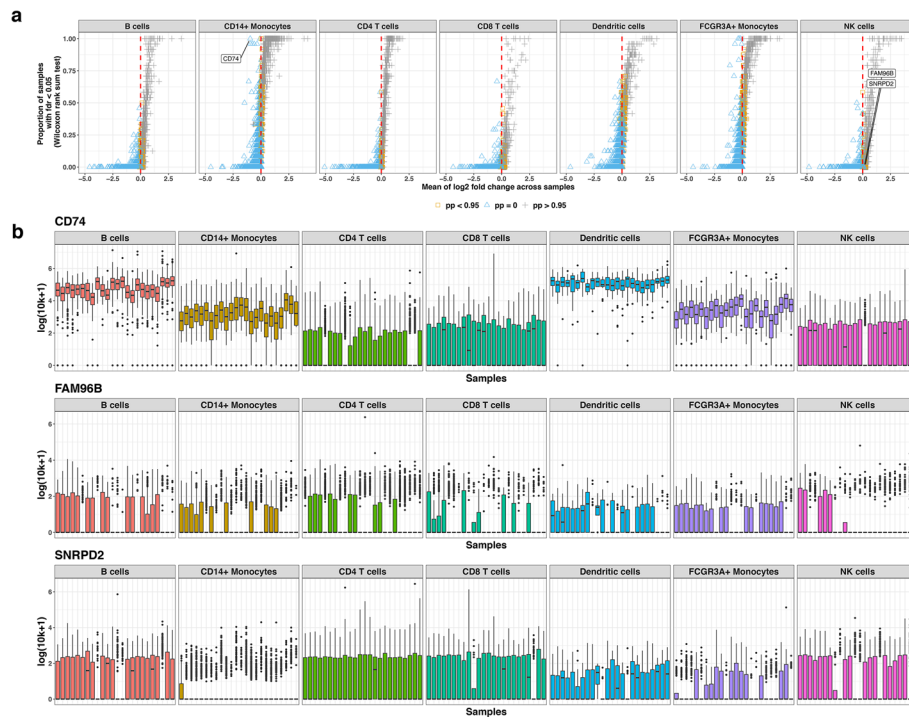


Fig. 4 Comparison between CTS genes called by Wilcoxon rank sum test (w-markers) and by scCTS (s-markers). **a** Scatter plot of DE state of genes in target cell type. The y-axis is proportion of samples in which a gene being called DE (w-marker) by Wilcoxon rank sum test with FDR < 0.05. The x-axis is the mean LFC defined in Eq. (2) across all twenty-four samples. Different colors represent posterior probability (pp) of genes that are s-markers (grey: pp > 0.95, is a s-marker; gold: pp < 0.95, not a s-marker and with positive LFC; blue: pp = 0, not a s-marker and with negative LFC). **b** Three example marker genes show difference between Wilcoxon rank sum test and scCTS. The y-axis is the log transformed 10 k counts. The x-axis represents samples. *CD74* (Monocytes) is w-marker in all samples, but not a s-marker. *FAM96B* (NK cell) is a s-marker with DE signal frequency 0.36, but not w-marker in any sample. *SNRPD2* (NK cell) is a s-marker, but not w-marker in any sample

Table 1 Number of genes by scCTS, categorized by number of samples showing DE from Wilcoxon test

Number of Samples showing DE by Wilcoxon test	B cells	CD14 + Monocytes	CD4 T cells	CD8 T cells	Dendritic cells	FCGR3A + Monocytes	NK cells
0	38	0	151	174	0	0	391
[1,8]	624	18	1135	408	11	13	660
[9,16]	178	347	300	45	257	272	90
[17,24]	93	622	233	25	137	453	65

The results are from the PBMC dataset without batch correction, where Wilcoxon test is applied on each of the 16 subjects

In addition to Wilcoxon, we also compare scCTS with other methods on the PBMC Lupus data. As shown in the figure S3, still most scCTS marker genes only show in a proportion of samples respectively by NS-Forest, FEAST, and scGeneFit. We further repeat above comparisons on the COVID-19 dataset. Still, among all CTS markers by

scCTS, most of them are only detected in a proportion of samples by alternative tools (Additional file 1: Table S4, Figure S4).

CTS markers detected by scCTS are more biologically meaningful

Since there are discrepancies between scCTS and alternative methods from each individual, we next explore which method provides more biologically meaningful CTS marker genes. Starting from the comparison with Wilcoxon rank sum test, for the genes identified as s-markers but not w-markers (referred to as “unique s-marker”), they are either with relatively strong DE signal in only a few samples, or with weak but consistent signal in most samples. *FAM96B* and *SNRPD2* are two example CTS genes in NK cells (Fig. 4b). We perform enrichment analysis with the unique s-marker genes for B cells (38 genes), CD4 T cells (151 genes), CD8 T cells (174 genes), and NK cells (394 genes). There are in total 87 terms corresponding to different cell types or tissues in the Human gene atlas database (<http://biogps.org/downloads/>) [44, 45] in package “enrichR” [46]. Table 2 contains all enriched terms with adjusted *p*-value smaller than 0.05. In Table 2, we can observe that in cell type B cells, CD4 T cells, and NK cells, the most significantly enriched terms are corresponding cell types. We repeat above analyses for the same four cell types using unique w-markers. Results in Additional file 1: Table S5 show that, although B cells, CD8 T cells, and NK cells are significantly enriched by respective unique w-markers, CD4 T cells are not enriched and the rank of enrichment of B cells is lower than that by s-markers. Furthermore, from the same analyses for the COVID-19 dataset, CD4 T cell, CD8T cells, and NK cells are significantly enriched by unique s-marker of corresponding cell types (Additional file 1: Table S6), while none of the 4 cell types are enriched by corresponding unique w-markers (Additional file 1: Table S7). We repeat the above analysis for the comparison between markers identified by scCTS with each of the other methods: NS-Forest, FEAST, scGeneFit, and ZINB-WaVe + DESeq2. Results of pathway enrichment from the PBMC Lupus dataset are summarized in Figure S5a, with detailed pathway information listed in Table S8-S15. As shown, s-markers for six cell types are significantly enriched in corresponding cell types, while at most three cell types (CD14 + Monocytes, CD8 T cells and NK cells) are significantly enriched in

Table 2 The significantly enriched terms in unique s-markers on the PBMC Lupus dataset without batch correction

Rank	B cells	CD4 T	CD8 T	NK cells
1	CD19 + B cells (neg. sel.) (p.adjust: 2.79e - 03)	CD4 + T cells (p.adjust: 7.57e - 05)	CD56 + NK cells (p.adjust: 1.42e - 03)	CD56 + NK cells (p.adjust: 8.33e - 12)
2	CD4 + T cells (p.adjust: 3.84e - 02)	CD8 + T cells (p.adjust: 7.57e - 05)	721 B lymphoblasts (p.adjust: 1.47e - 03)	721 B lymphoblasts (p.adjust: 1.29e - 06)
3	CD8 + T cells (p.adjust: 4.29e - 02)	721 B lymphoblasts (p.adjust: 1.40e - 03)	CD4 + T cells (p.adjust: 1.22e - 02)	CD4 + T cells (p.adjust: 1.43e - 04)
4			CD8 + T cells (p.adjust: 2.70e - 02)	CD8 + T cells (p.adjust: 9.06e - 04)
5			CD19 + B cells (neg. sel.) (p.adjust: 2.99e - 02)	Heart (p.adjust: 4.27e - 02)
6			Lymphoma burkitts (Raji) (p.adjust: 2.99e - 02)	

For each cell type, only terms with adjusted *p*-values < 0.05 are kept and the one corresponding to the same cell type is bolded. The unique s-markers are CTS genes called only by scCTS

markers uniquely detected by other methods. For the COVID-19 dataset, pathway enrichment results are summarized in Figure S5b, with detailed information listed in Table S16-S23. Still, s-markers for most cell types are significantly enriched in corresponding cell types, while at most one cell type (Dendritic cells) is significantly enriched in markers uniquely detected by other methods. All these results together indicate that the marker genes detected by scCTS are more biologically meaningful than those by the alternative methods. Discovery of such CTS genes are the result from a rigorous modeling of the data from many samples together by scCTS.

Prevalence estimates for common CTS marker genes

We further investigate the genes which are both s-markers and w-markers by comparing the estimated frequencies for DE from scCTS with the proportion of samples called DE by Wilcoxon rank sum test. In Fig. 5a, we can observe that the estimated frequency from scCTS is usually higher than the proportion of samples called DE by Wilcoxon rank sum test. One example gene is *NFATC1*. It is an s-marker for CD4 T cells with estimated frequency equals to one, while being called DE in only three out of twenty-four samples

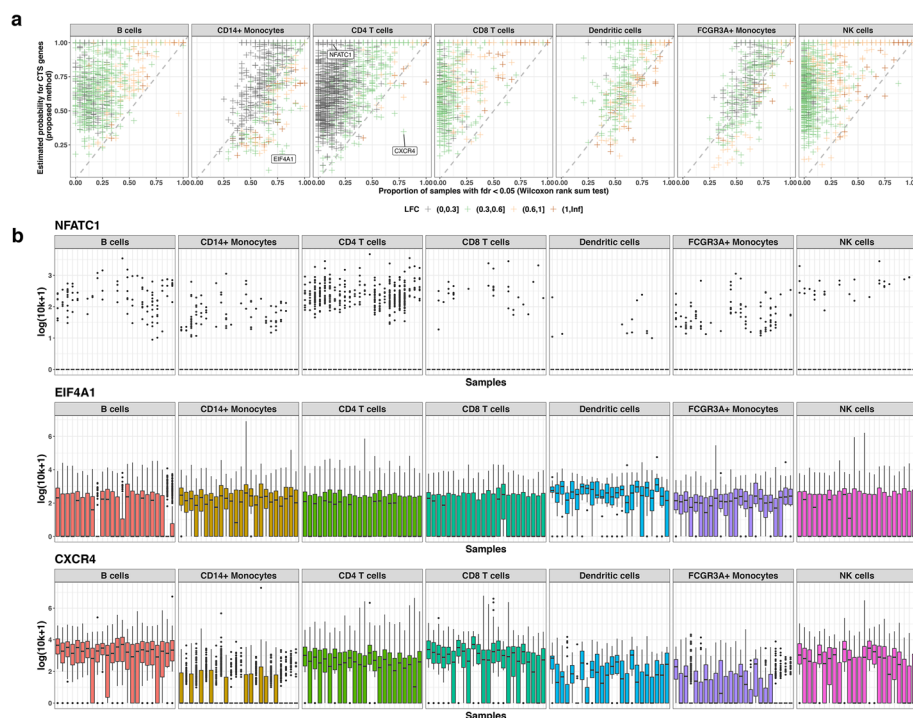


Fig. 5 Consistency comparison between scCTS and Wilcoxon test for genes that are both s-markers and w-markers. **a** Scatter plot of estimated frequency showing DE state by scCTS and Wilcoxon rank sum test of genes in target cell type. The y-axis is the estimated frequency showing DE state among samples by proposed method. The x-axis is proportion of samples in which a gene being called DE (w-marker) by Wilcoxon rank sum test with FDR < 0.05. Different colors represent estimated mean LFC among samples (grey: $0 < \text{LFC} \leq 0.30$; green: $0.30 < \text{LFC} \leq 0.6$; gold: $0.60 < \text{LFC} \leq 1.00$; brown: $\text{LFC} > 1.00$). **b** Three example genes show difference between Wilcoxon rank sum test and proposed method. The y-axis is the log transformed 10 k counts. The x-axis represents samples. *NFATC1* is a s-marker with weak but consistent DE signal in samples but called DE in only 3 out of 24 samples by Wilcoxon rank sum test for CD4 T cells. *EIF4A1* is a s-marker with DE signal frequency 0.19 but called DE in 19 out of 24 samples by Wilcoxon rank sum test for CD14 + Monocytes. *CXCR4* is a s-marker with DE signal frequency 0.35 but called DE in 19 out of 24 samples by Wilcoxon rank sum test for CD4 T cells

by Wilcoxon rank sum test. From Fig. 5b, we can observe that the expression pattern of *NEATC1* in cell types is similar across samples. However, its weak DE signal leads to small power to call DE in many samples. Benefitted from pooling samples in analysis, its estimated frequency showing DE is one with scCTS. At the same time, there are some genes with much lower estimated frequency showing DE with scCTS than the proportion from Wilcoxon rank sum test, such genes including *EIF4A1* in CD14+ Monocytes and *CXCR4* in CD4 T cells (Fig. 5b). In most samples, mean expression of gene *EIF4A1* in CD14+ Monocytes is lower than in Dendritic cells. With definition of Eq. (2), its estimated frequency showing DE is only 0.19. However, it is called DE in 19 out of 24 samples by Wilcoxon rank sum test. This is due to the high abundance of CD14+ Monocytes and low abundance of Dendritic cells (Additional file 1: Table S1). The same applies to gene *CXCR4* in CD4 T cells. These results show that the results from Wilcoxon test will be greatly affected by cell proportions, which is undesirable. On the other hand, scCTS, by modeling the log fold change, will avoid such problem.

Results from pooling subjects, batch-corrected data

Next, we compare scCTS with alternative methods on pooling cells from all subjects, which is a common approach in many tools such as Seurat. Due to subject level technical artifacts such as batch effect, we first run batch effect correction using scMerge [47], then call CTS genes on pooled cells using Wilcoxon test, FEAST, NS-Forest, scGeneFit, ZINB-WaVe + DESeq2, and scCTS. First discrepancy lies in the number of detected CTS markers. For example, the number of CTS marker genes detected by scCTS differs greatly from Wilcoxon rank sum test (Additional file 1: Figure S6-S7), and the proportion of detected genes shared by both methods is small, ranging from 5 to 14% in the PBMC Lupus dataset and 10 to 27% in the COVID-19 dataset. These results suggest a significant difference between CTS genes called by scCTS and Wilcoxon rank sum test. We repeat the pathway analyses using CTS markers from these batch-corrected data. Compared to the results in previous sections, results on pooled, batched-corrected data support a greater performance gain of scCTS over Wilcoxon rank sum test and the other methods, especially for the PBMC Lupus dataset. To be specific, in the comparison of scCTS with Wilcoxon rank sum test, six out of seven cell types in the PBMC Lupus dataset are strongly enriched with a high rank (top 1 or 2) by the corresponding unique s-markers (Table 3), and this specificity is clear for dendritic cell which is the only enriched term by its unique s-markers. In contrast, none of the seven cell types are enriched by w-markers for the corrected PBMC Lupus dataset (Additional file 1: Table S24). For the batch-corrected COVID-19 dataset, 6 out of 8 cell types are significantly enriched by the corresponding s-markers (Additional file 1: Table S25) while again none of them is enriched by w-markers (Additional file 1: Table S26). Pathway enrichment results by unique markers from comparisons between scCTS and each of the other methods on the batch corrected datasets are summarized in Figure S8, with all detailed information listed in table S27-42. For both datasets, still most cell types are significantly enriched in unique s-markers while at most two cell types are significantly enriched by unique markers from alternative methods.

Table 3 Top 5 significantly enriched terms in unique s-markers on the batch-corrected PBMC Lupus dataset

Rank	B cells	CD14 + Monocytes	CD4 T cells	CD8 T cells
1	721 B lymphoblasts (Padjust: 2.69e – 21)	CD33 + Myeloid (Padjust: 1.01e – 74)	CD4 + T cells (Padjust: 1.07e – 39)	CD56 + NK cells (Padjust: 6.31e – 32)
2	CD19 + B cells (neg. sel.) (Padjust: 1.53e – 20)	CD14 + Monocytes (Padjust: 4.15e – 58)	721 B lymphoblasts (Padjust: 1.35e – 39)	CD8 + T cells (Padjust: 3.23e – 15)
3	Lymphoma burkitts (Daudi) (Padjust: 1.63e – 06)	Whole Blood (Padjust: 2.64e – 44)	CD8 + T cells (Padjust: 3.69e – 34)	CD4 + T cells (Padjust: 2.03e – 14)
4	CD34 + (Padjust: 6.16e – 06)	BDCA4 + dendritic cells (Padjust: 1.95e – 07)	CD34 + (Padjust: 1.59e – 13)	Lymphoma burkitts (Raji) (Padjust: 2.40e – 02)
5	Lymphoma burkitts (Raji) (Padjust: 1.91e – 04)	Smooth Muscle (Padjust: 2.76e – 07)	CD105 + Endothelial (Padjust: 1.05e – 11)	Heart (Padjust: 2.40e – 02)
Rank	Dendritic cells	FCGR3A + Monocytes	NK cells	
1	BDCA4 + dendritic cells (Padjust: 3.26e – 07)	CD14 + Monocytes (Padjust: 1.15e – 53)	CD56 + NK cells (Padjust: 5.36e – 72)	
2		CD33 + Myeloid (Padjust: 2.82e – 41)	CD8 + T cells (Padjust: 3.55e – 08)	
3		Whole Blood (Padjust: 2.51e – 20)	721 B lymphoblasts (Padjust: 5.86e – 08)	
4		Smooth Muscle (Padjust: 7.77e – 04)	CD4 + T cells (Padjust: 7.26e – 08)	
5			Whole Blood (Padjust: 9.11e – 08)	

For each cell type, only terms with adjusted *p*-values < 0.05 are kept and the one corresponding to the same cell type is bolded. The unique s-markers are CTS genes called only by scCTS

In addition to unique markers, we also examine the scCTS markers with different characteristics: the ones showing consistently weak DE signals in most subjects (prevalence > 0.8 & LFC < 0.2) and showing strong DE signals in only few subjects (prevalence < 0.5 & LFC > 0.4). These genes are often failed to be detected by alternative tools such as Wilcoxon test in pooled analysis. As expected, both types of markers are biologically meaningful according to the pathway analyses. For example, in Additional file 1: Table S43, three cell types (CD14 + Monocytes, CD8 T cells and NK cells) of the PBMC Lupus dataset are significantly enriched in their markers that show strong DE signal but in only few subjects, and this situation is prominent for CD14 + Monocytes where the significance comes from only less than 20 marker genes. For consistently weak markers, two cell types (CD14 + Monocytes and CD4 T cells) are correspondingly significantly enriched (Additional file 1: Table S44). Repeated analyses for the COVID-19 dataset further support the biological importance of the two groups of markers, with stronger evidence (5 out of 8 cell types are enriched) coming from consistently weak markers (Additional file 1: Table S45 and S46).

Overall, results on batch effect corrected datasets further strengthen the conclusion that the marker genes detected by scCTS are more biologically meaningful, and the performance enhancement over alternative tools is even greater than in the uncorrected data.

Discussion

Cell type-specific (CTS) genes are of great interests in biological studies because they serve as cell type identities and provide insights for many biological and clinical mechanisms under various conditions. They can be also used in many scRNA-seq downstream analyses such as cell type annotation and bulk data deconvolution. Existing methods for identifying CTS genes from scRNA-seq data ignore the between-subject heterogeneity, thus generate results with low accuracy and robustness. Moreover, the characteristics of the CTS genes, such as their prevalence in a population, are not reported by traditional method such as Wilcoxon or t -test. This limits the biological interpretation and the application of CTS genes in downstream analyses. For example, if one wants to use CTS marker genes to annotation cell types, it is desirable to use markers that show up consistently in a population.

In this work, we explore the real data and discover that most of the CTS genes identified through Wilcoxon rank sum test or reported by public databases (PanglaoDB and GeneMarker) do not consistently appear in all subjects in a population. Inspired by this observation, we develop a novel statistical model scCTS to identify CTS genes from population level scRNA-seq data. Results show that our method not only identifies more biologically meaningful CTS genes, but also provide more information for these genes, including their population prevalence and conditional DE strength, which are important and interesting information. Overall, to identify CTS genes from population level scRNA-seq data, we recommend applying scCTS on the batch-corrected data since that will provide greater performance gain over traditional methods such as Wilcoxon test.

We want to note that the major goal in this work is to develop a method that can expand the existing CTS marker database to include the category 2 and 3 genes. In addition to including more marker genes, scCTS also reports the population level prevalence of marker genes, analogous to that the SNP annotation includes minor allele frequency. Also, CTS markers in category 3 are analogous to genes identified from bulk data by statistical methods such as Cancer Outlier Profile Analysis (COPA) [48, 49], Outlier Sums (OS) [50], Outlier Robust T-statistic (ORT) [51], Maximum-Ordered Subset T-statistics (MOST) [52], and Detection of Imbalanced Differential Signal (DIDS) [53], which aim at detecting DE genes activated only in a subset of data to decipher mechanisms that are present in a small subset of groups. For example, DIDS identified from a subgroup one confirmed gene ABCB1B showing resistance to docetaxel when comparing primary tumors resistant and sensitive to treatment with docetaxel [53, 54]. Besides, the new marker genes can help downstream analyses, such as bulk data deconvolution and cell type annotation. There are recent works on using individualized reference panel for cell type deconvolution [55] demonstrating the importance of considering the individual heterogeneity of marker genes. However, one cannot simply throw the category 2 and 3 marker genes into an existing deconvolution tool. Rigorous methods are needed to consider the results such as maker gene prevalence from scCTS, which is our research plan in the near future.

The reason why many marker genes do not show up in all subjects is from a combination of biological and technical reasons. The results do show that the DE strength has a positive correlation with the prevalence, i.e., stronger marker genes tend to show up in more subjects. However, there exist a non-trivial number of genes with

rather strong conditional DE signal but low prevalence. To carefully study the behavior of these genes will be interesting and our research plan in the near future. In addition, we plan to apply our proposed method on atlas-level scRNA-seq datasets and create an interactive database in various species or tissues. Moreover, the current method does not consider covariates. We will extend the model to incorporate subject-level covariates in analysis, which can help users to better identify the CTS genes under more complex study designs. Further note that DE analysis in single-cell data has more dimension than traditional bulk data. The DE can be performed among cell types in the same condition, or to compare each cell type between different conditions. We think both comparisons are important. Developing method for comparing varied biological conditions of the same cell type from population-level scRNA-seq data is our focus in near future.

Conclusions

In this work, we develop a novel statistical model called “scCTS” that incorporates between-subject heterogeneity into a hierarchical model to detect CTS markers from population level scRNA-seq data. The model is inspired by real data observation that CTS markers do not consistently show DE states in all studied subjects. By introducing a prevalence prior conditioning on the marker status, scCTS allows the CTS markers show up only in a fraction of the subjects. Under the hierarchical model, we derived EM procedures for parameter estimation and used the posterior probabilities of being CTS markers for inference. Benchmark simulation studies show that scCTS significantly improves the accuracy in identifying CTS markers compared to traditional method such as Wilcoxon rank sum test. Extensive real data analyses demonstrate that scCTS can identify CTS markers with different characteristics, including high consistency and strong/weak effect sizes, and low consistency but strong effect sizes, all of which could help to decipher complex mechanisms under different conditions. Also, supported by the enrichment evidence of corresponding cell types, unique genes identified by scCTS are more biologically meaningful compared to the ones from alternative methods.

Methods

Subject-level statistics representing gene’s cell type specificity

The input data of the model include scRNA-seq expression data from a population, with known cell types for all cells. Suppose there are N subjects from which we want to identify CTS genes. In each subject, there are G genes and K cell types. Let X_{gikc} be the normalized expression for g th gene ($g = 1, \dots, G$) of i th subject ($i = 1, \dots, N$) in c th cell ($c = 1, \dots, C_{ik}$) of k th cell type ($k = 1, \dots, K$). Here, C_{ik} represents the number of cells for subject i in cell type k . The normalization is done by computing the read counts per 10,000 reads. We assume the normalized expression X_{gikc} is independent between genes and cells for all subjects. Define $E\{X_{gikc}\} = \mu_{gik}$, and $Var\{X_{gikc}\} = \omega_{gik}^2$ as the mean and variance of the normalized expression value. Then the unbiased estimator for mean expression of gene g in cell type k of subject i is: $\bar{X}_{gik} = \frac{\sum_{c=1}^{C_{ik}} X_{gikc}}{C_{ik}}$. With Central Limit Theorem, when C_{ik} is large enough, \bar{X}_{gik} 's approximate distribution is:

$$\bar{X}_{gik} \sim AN(\mu_{gik}, \frac{\omega_{gik}^2}{C_{ik}}) \quad (1)$$

for $g = 1, \dots, G; i = 1, \dots, N; k = 1, \dots, K$.

For the following context, we treat k th cell type as the “target” cell type for which we want to identify its CTS genes. In this work, we focus on CTS genes with expression at a higher level in only one cell type (i.e., one vs. others).

Let Y_{gik} be the log2 fold change (LFC) of the expression for gene g in cell type k over the average of other cell types, in subject i . Y_{gik} is computed as shown in Eq. (2).

$$Y_{gik} = \log_2(\bar{X}_{gik} + 1) - \log_2\left(\frac{\sum_{k' \neq k} \bar{X}_{gik'}}{K - 1} + 1\right) \quad (2)$$

A large value of Y_{gik} indicates that gene g is a CTS gene of cell type k in subject i . Our computation of LFC is different from most existing methods for identifying CTS genes, which pool all cells and perform statistical test. In those methods, the results will be affected by the cell type proportions since larger cell types will have greater statistical power to obtain significant results. Our definition of mean expression in other cell types in Eq. (2) excludes the influence of cell type composition, thus will provide more stable results. For the procedures below, we will model Y_{gik} for CTS gene identification. Using Y_{gik} instead of the data from individual cells greatly improve the computational efficiency without losing much information.

A hierarchical model for CTS genes

We use the hierarchical model shown in equation (3) to combine the DE information from multiple subjects. We define D_{gk} as a binary random variable representing whether gene g is a CTS gene in cell type k (1 : yes; 0, no). If gene g is a CTS gene in cell type k ($D_{gk} = 1$), then it has a probability q_{gk} to be DE (higher expression than the average of other cell types) in a randomly picked subject i , which is represented by binary random variable $Z_{gik} = 1$. We further introduce a random variable Δ_{gik} to represent the expected value of the estimated LFC (Y_{gik}), and σ_{gik} is the corresponding standard deviation. If gene g is a CTS marker gene in cell type k and shows DE signal in subject i ($D_{gk} = Z_{gik} = 1$), then Δ_{gik} should be greater than 0; otherwise, it should have expected value 0 with a small variation. Putting all pieces together, we have following hierarchical model:

$$\begin{aligned} Y_{gik} | \Delta_{gik} &\sim N(\Delta_{gik}, \sigma_{gik}^2) \\ \Delta_{gik} | Z_{gik} = 1 &\sim N(m_{gk}, \tau_{gk}^2) \\ \Delta_{gik} | Z_{gik} = 0 &\sim N(0, \tau_{gk}^2) \\ Z_{gik} | D_{gk} &\sim \text{Bernoulli}(q_{gk} \times D_{gk}) \\ D_{gk} &\sim \text{Bernoulli}(\pi_k) \end{aligned} \quad (3)$$

Here, m_{gk} is the population level mean LFC of gene g in cell type k ; τ_{gk}^2 is the population level variance of LFC for gene g in cell type k . Specifically, we assume $m_{gk} \geq thres \geq 0$ and $Z_{gi} \perp Z_{gi} | D_{gk} = 1$. $thres$ is a threshold defined by users, since small LFC is less possible to be a marker and has less interest. In the estimation process, Y_{gik} and σ_{gik}^2 are estimated from

each individual subject. The detailed procedure is provided in the Additional file 1 Sect. "Background".

Identification of CTS marker genes

From the above model, we can obtain several interesting quantities from the model. First, the posterior probability of $D_{gk} = 1$ provides an overall assessment whether a gene is a CTS gene. At the highest level, a gene can be either CTS genes ($D_{gk} = 1$) or non-CTS genes ($D_{gk} = 0$). Next, the conditional probability q_{gk} represents the consistency for a CTS gene to show DE signals across subjects. The CTS genes are allowed to have different frequencies (q_{gk}) for showing DE in individual subjects and cell types. Finally, m_{gk} represents the conditional subject level DE strength once the gene is deemed CTS gene in a subject.

If we merely want to identify CTS marker genes, we only need to look at the posterior probability of $D_{gk} = 1$. However, a gene can have large posterior probability of $D_{gk} = 1$ if it has large q_{gk} or m_{gk} , or both. From our model, different types of CTS marker genes can be identified: (1) consistently show strong DE signal in most subjects (large q_{gk} and m_{gk}); (2) consistently show weak DE signals in most subjects (large q_{gk} , small m_{gk}); (3) show strong DE signals in only few subjects (small q_{gk} , large m_{gk}). Usually, the second type of CTS marker genes are difficult to detect from testing on individual subjects one by one, because tests for CTS markers with weak signals have very low statistical power, especially in minor cell types. The third type of markers are difficult to identify by testing on pooled data, because DE signal in partial subjects can be weakened after pooling with other subjects without DE signals. Our proposed method overcome these limitations and can identify all types of marker genes.

These different types of CTS marker genes could have distinct biological meanings and computational utilities. For example, CTS marker genes consistently showing strong DE signals in all subjects (have large q_{gk} and m_{gk}) are more preferred for downstream analyses such as cell typing or bulk sample deconvolution, since they can robustly provide clear signal to represent a cell type.

Parameters estimation with EM algorithm

The parameters to be estimated from the proposed model include the following: m_{gk} , population level mean LFC of gene g in cell type k ; τ_{gk}^2 , population level variance of LFC of gene g in cell type k ; q_{gk} , probability of CTS marker gene g in cell type k for a randomly picked subject; π_k , probability of a randomly picked gene to be a CTS marker gene for cell type k among the subjects. Since there are a number of latent variables in our model (Δ_{gik} , Z_{gik} , and D_{gk}), we develop an EM algorithm to separately estimate parameters in every cell type k .

Define $\phi(x; m, \tau^2)$ to be the probability density at a point x of a normal distribution with mean m and variance τ^2 . We further define following values: $\phi_{y_{gik}} = \phi\left(Y_{gik}; \Delta_{gik}, \sigma_{gik}^2\right)$; $\phi_{0_{gik}} = \phi\left(\Delta_{gik}; 0, \tau_{gk}^2\right)$; and $\phi_{1_{gik}} = \phi\left(\Delta_{gik}; m_{gk}, \tau_{gk}^2\right)$.

Denote $\Theta_k = \{\pi_k, \mathbf{q}_k, \mathbf{m}_k, \tau_k^2\}$, where $\mathbf{q}_k = \{q_{1k}, \dots, q_{Gk}\}$, $\mathbf{m}_k = \{m_{1k}, \dots, m_{Gk}\}$, $\tau_k^2 = \{\tau_{1k}^2, \dots, \tau_{Gk}^2\}$. Θ_t are the parameters derived at t -th iteration. We can derive the complete likelihood for cell type k as follows:

$$\begin{aligned}
L(\Theta_{\mathbf{k}}) &= \prod_g P(\mathbf{Y}_{gk}, \Delta_{gk}, \mathbf{Z}_{gk}, D_{gk} | m_{gk}, q_{gk}, \pi_k) \\
&= \prod_g \left\{ \left[(1 - \pi_k) \left\{ \prod_{i=1}^N [\phi_{\gamma_{gik}} \times \phi_{0_{gik}}]^{1-Z_{gik}} \right\} \right]^{1-D_{gk}} \right. \\
&\quad \left. \times \left[\pi_k \left\{ \prod_{i=1}^N \phi_{\gamma_{gik}} \times [(1 - q_{gk}) \phi_{0_{gik}}]^{1-Z_{gik}} \times [q_{gk} \phi_{1_{gik}}]^{Z_{gik}} \right\} \right]^{D_{gk}} \right\} \quad (4)
\end{aligned}$$

Then the log-likelihood is:

$$\begin{aligned}
l(\Theta_{\mathbf{k}}) &= \sum_{g=1}^G \left\{ (1 - D_{gk}) \log(1 - \pi_k) + D_{gk} \log \pi_k \right. \\
&\quad + \sum_{i=1}^N \left[(D_{gk} - D_{gk} Z_{gik}) \log(1 - q_{gk}) + D_{gk} Z_{gik} \log q_{gk} \right. \\
&\quad \left. \left. - \frac{\log \tau_{gk}^2}{2} - \frac{\Delta_{gik}^2}{2\tau_{gk}^2} - \frac{D_{gk} Z_{gik} m_{gk}^2}{2\tau_{gk}^2} + \frac{D_{gk} Z_{gik} m_{gk}^2 \Delta_{gik}}{\tau_{gk}^2} \right] \right\} + Constant \quad (5)
\end{aligned}$$

Theoretically, the estimation should be done by updating all four parameters jointly. For computation efficiency, we develop the following procedure to approximate the estimate of parameters. The general framework for the modified EM algorithm is as follows:

- S1. Assume all genes are CTS genes ($D_{gk} = 1$) and then estimate m_{gk} , τ_{gk}^2 and q_{gk} with EM algorithm (Z_{gik} is missing data) for each gene $g = 1, \dots, G$;
- S2. Based on estimated m_{gk} and given LFC threshold *thres* to arbitrarily assign $D_{gk} = 0$ for genes with $m_{gk} \leq \textit{thres}$;
- S3. Estimate π_k with EM algorithm, where m_{gk} , τ_{gk}^2 and q_{gk} are fixed as estimates derived in S1; D_{gk} is missing data.

The details of the steps S1 and S3 are shown in Additional file 1: Sect. 3.

Simulation

Data simulation

In our simulations, we assume there are in total 6231 genes and 10,000 cells from 16 subjects, and all subjects contain 4 cell types. The number of genes, subjects, and proportion of cells allocated to each subject are the same as those in the preprocessed PBMC Lupus dataset. Proportions of the four cell types are established by normalizing the proportion of B cells, CD8 T cells, FCGR3A + Monocytes, and NK cells from the PBMC Lupus dataset. In addition, we generate extra 4 groups of cell type proportions, by exploring the 35 sets of four-out-of-seven combinations from B cells, CD14 + Monocytes, CD4 T cells, CD8 T cells, Dendritic cells, FCGR3A + Monocytes, and NK cells (Additional file 1: Figure S1(e)). With the number of genes, cells, and the allocation of cells defined, the data are simulated according to the procedure as bellow.

First, we generate baseline expression matrix. To mimic the natural between-subject heterogeneity in baseline expression, we directly extract the profile of CD4 T cells from the 16 control subjects in the PBMC lupus dataset and utilize their subject-specific mean to construct baseline profile (Additional file 1: Figure S1(a)).

Second, for each cell type k , we simulate the marker status of gene g , by randomly sampling $D_{gk} \sim \text{Bernoulli}(\pi_k)$. We set $\pi_k = 0.05$ to allow only a small proportion of markers for each cell type. With all marker genes determined ($D_{gk} = 1$), we generate their mean log fold changes, m_{gk} , by sampling from the empirical distributions estimated from real data. The prevalence probability q_{gk} is then generated according to an observed positive correlation between q_{gk} and m_{gk} from the real data. Based on this strategy, some markers with large effect sizes will have high prevalence probabilities compared to the others with small effect sizes (Additional file 1: Figure S1(b)).

Then, we generate $Z_{gik} \sim \text{Bernoulli}(D_{gk} * q_{gk}), i = 1, 2, \dots, 16$, to determine whether maker g shows DE for cell type k in subject i .

Third, for each subject i and cell type k , suppose the number of cells is n_k^i . We simulate the expression count of gene g , $X_{gc}^{i,k}$, based on the following negative binomial model:

$$X_{gc}^{i,k} \sim NB\left(s_{gc}^{i,k} \mu_g^{i,k}, \phi_g^{i,k}\right), c = 1, 2, \dots, n_k^i,$$

$$\log_2 \mu_g^{i,k} = \alpha_g^i + \beta_g^{i,k},$$

$$\beta_g^{i,k} \sim \begin{cases} 0 & , \text{if } Z_{gik} = 0 \\ N\left(m_{gk}, \tau_{gk}^2\right) & , \text{if } Z_{gik} = 1 \end{cases}.$$

In above model, $\mu_g^{i,k}$ refers to the underlying mean and dispersion of gene g in cell type k from subject i . $s_{gc}^{i,k}$ is a size factor containing the technical noise cause by variations in sequencing depth. We simulate $s_{gc}^{i,k}$ from a uniform distribution of $U(0.5, 5)$. α_g^i reflects the baseline expression of gene g in subject i , which is obtained previously according to the expression of CD4 T cells.

$\beta_g^{i,k}$ quantifies the effect size of gene g when comparing cell type k to the remaining cell types in subject i . According to the above model, if gene g is a marker of cell type k and shows cell type DE in subject i , then $\beta_g^{i,k} \sim N(m_{gk}, \tau_{gk}^2)$ with τ_{gk}^2 randomly sampled from $U(0.1, 0.2)$, otherwise, $\beta_g^{i,k} = 0$. Under this setting, even if gene g is the marker of cell-type k , there would still be no difference between its expression and remaining cell types in subject i if its DE effect does not present in that subject, which is clearly shown by Additional file 1: Figure S1(c). Another notice about $\beta_g^{i,k}$ is that, our simulation allows one gene to be the marker of multiple cell types. In this case, $\beta_g^{i,k}$ will be further adjusted to avoid potential effect cancelation. For example, if gene g is the marker of both cell type k_1 and k_2 , then $\tilde{\beta}_g^{i,k_1} = \tilde{\beta}_g^{i,k_2} = \frac{\beta_g^{i,k_1} + \beta_g^{i,k_2}}{2}$.

Given α_g^i and $\beta_g^{i,k}$ and thus $\mu_g^{i,k}$, the dispersion parameter $\phi_g^{i,k}$ is simulated as a function of $\mu_g^{i,k} : \frac{0.01}{\mu_g^{i,k}} + 0.1$. With mean and dispersion specified, we simulate expression of all genes in cell type k from subject i according to the negative binomial distribution of $X_{gc}^{i,k} \sim NB\left(s_{gc}^{i,k} \mu_g^{i,k}, \phi_g^{i,k}\right), c = 1, 2, \dots, n_k^i$.

Evaluation

After deriving the simulated data, we compare scCTS method with Wilcoxon rank sum test. We use ROC and AUC score to evaluate the accuracy of proposed method, by averaging results of 100 simulations.

Real data analysis**Filtering of cells and genes**

For the lupus dataset, we first remove non-singlet cells which were already annotated in the dataset. We exclude megakaryocytes because they are very rare (average number of cells per patient < 20). Next, we filter out genes expressed in fewer than two cells for each subject and retain overlapped genes among all subjects. We normalize the gene expression by library sizes in each cell using a size factor of 1e4. For the COVID-19 dataset, we select frozen PBMCs from healthy controls and severe patients without comorbidities. In total we select 13 samples. We remove cells for two rare cell types (average number of cells per patient < 20), neutrophils and macrophages. Genes expressed in fewer than ten cells are filtered out for each subject, and the overlapped genes are preserved. The normalization step is performed in the same manner as the PBMC Lupus dataset.

Batch effect correction

To reduce batch effects introduced by different subjects while preserving cell-type heterogeneity, we compare multiple batch-effect correction methods: limma [56], FastMNN [57], ComBat-seq [58], scMerge [47] and scMerge2 [59]. We employ Uniform Manifold Approximation and Projection (UMAP) [60] visualizations and local inverse Simpson's index (LISI) [61] to evaluate the batch correction results. Specifically, following the methodology from a previous benchmark paper [62], we first calculate the iLISI and cLISI metrics for each cell on their gene-level principal components, and then obtain the median values of the two metrics. These two median values are respectively normalized by the maximum and minimum of iLISI and cLISI metrics, denoted as iLISI_{norm} and cLISI_{norm}. A higher iLISI_{norm}/cLISI_{norm} value indicates stronger homogeneity among subjects/cell types, and therefore we compute the F1 score as follows:

$$F1_{LISI} = \frac{2(1 - cLISI_{norm})(iLISI_{norm})}{1 - cLISI_{norm} + iLISI_{norm}}$$

Additional file 1: Figure S9-S11 illustrate that scMerge achieves the best overall performance on the two datasets. Thereby we utilize the expression matrix corrected by scMerge.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-024-03410-8>.

Additional file 1: Supplementary Section S1 (Standard error calculation for estimated log2 fold change in one sample); S2 (A more general framework for different types of marker identification); S3 (EM algorithm details); Table S1 – S46; Figure S1 – S11. (DOCX 5.9 MB).

Additional file 2. Review history.

Acknowledgements

Not applicable.

Review history

The review history is available as Additional File 2.

Peer review information

Veronique van den Berghe and Kevin Pang were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

H.W. conceived the project. H.W. and L.C. designed the method. L.C., Z.G., and T. D. performed real data analyses. L.C., Z.G., D.T., and H.W. drafted the manuscript. All authors read and approved the final manuscript.

Funding

H.W. and L.C. were partially supported by National Institutes of Health R01GM122083 and R01GM141392. H.W. was also partially supported by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDB38050100. Z.G. was partially supported by the intramural funding from The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), and the Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (2023B1212010001).

Availability of data and materials

The Gene expression data sets used and analyzed in this study are available in the Gene Expression Omnibus (GEO) repository under the following accession IDs: GSE96583 [63] and GSE158055 [64]. The proposed method is implemented in an R package scCST, which is freely available on GitHub at <https://github.com/ToryDeng/scCTS>, under the GPL-3.0 license [65]. The source code used in the manuscript can be accessed in Zenodo at <https://doi.org/10.5281/zenodo.13851702> [66]. All simulated and preprocessed real data can be accessed in Zenodo at <https://doi.org/10.5281/zenodo.13850742> [67].

Declarations**Ethics approval and consent to participate**

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 10 November 2023 Accepted: 30 September 2024

Published online: 14 October 2024

References

- Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. *Mol Cell*. 2015;58(4):610–20.
- Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8(1):1–12.
- Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*. 2015;161(5):1202–14.
- Li C, Liu B, Kang B, Liu Z, Liu Y, Chen C, et al. SciBet as a portable and fast single cell type identifier. *Nat Commun*. 2020;11(1):1–8.
- Deng Y, Bao F, Dai Q, Wu LF, Altschuler SJ. Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. *Nat Methods*. 2019;16(4):311–4.
- Jaitin DA, Weiner A, Yofe I, Lara-Astiaso D, Keren-Shaul H, David E, et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell*. 2016;167(7):1883–96 e15.
- Fan X, Dong J, Zhong S, Wei Y, Wu Q, Yan L, et al. Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis. *Cell Res*. 2018;28(7):730–45.
- Peng J, Sun B-F, Chen C-Y, Zhou J-Y, Chen Y-S, Chen H, et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res*. 2019;29(9):725–38.
- Saul D, Leite Barros L, Wixom AQ, Gellhaus B, Gibbons HR, Faubion WA, et al. Cell Type-Specific Induction of Inflammation-Associated Genes in Crohn's Disease and Colorectal Cancer. *Int J Mol Sci*. 2022;23(6): 3082.
- Velmeshev D, Schirmer L, Jung D, Haeussler M, Perez Y, Mayer S, et al. Single-cell genomics identifies cell type-specific molecular changes in autism. *Science*. 2019;364(6441):685–9.
- Park J, Shrestha R, Qiu C, Kondo A, Huang S, Werth M, et al. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science*. 2018;360(6390):758–63.
- Kim HJ, Wang K, Chen C, Lin Y, Tam PP, Lin DM, et al. Uncovering cell identity through differential stability with Cepo. *Nat Comput Sci*. 2021;1(12):784–90.
- Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun*. 2019;10(1):1–9.

14. Li D, Ding J, Bar-Joseph Z, editors. Unsupervised cell functional annotation for single-cell RNA-Seq. International Conference on Research in Computational Molecular Biology; 2022: Springer.
15. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods*. 2017;14(4):414–6.
16. Miao Z, Moreno P, Huang N, Papatheodorou I, Brazma A, Teichmann SA. Putative cell type discovery from single-cell gene expression data. *Nat Methods*. 2020;17(6):621–8.
17. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol*. 2019;20(2):163–72.
18. De Kanter JK, Lijnzaad P, Candelli T, Margaritis T, Holstege FC. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res*. 2019;47(16):e95–e.
19. Li Z, Wang Y, Ganan-Gomez I, Colla S, Do K-A. A machine learning-based method for automatically identifying novel cells in annotating single-cell RNA-seq data. *Bioinformatics*. 2022;38(21):4885–92.
20. Hu J, Li X, Hu G, Lyu Y, Susztak K, Li M. Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. *Nature machine intelligence*. 2020;2(10):607–18.
21. Kim JH, Ho SB, Montgomery CK, Kim YS. Cell lineage markers in human pancreatic cancer. *Cancer*. 1990;66(10):2134–43.
22. Hao Y, Hao S, Andersen-Nissen E, Mauck WM III, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573–87 e29.
23. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19(1):1–5.
24. Aevermann B, Zhang Y, Novotny M, Keshk M, Bakken T, Miller J, et al. A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell RNA sequencing. *Genome Res*. 2021;31(10):1767–80.
25. Su K, Yu T, Wu H. Accurate feature selection improves single-cell RNA-seq cell clustering. *Brief Bioinform*. 2021;22(5):bbab034.
26. Feng Z-Y, Wang Y. Elf: extract landmark features by optimizing topology maintenance, redundancy, and specificity. *IEEE/ACM Trans Comput Biol Bioinf*. 2018;17(2):411–21.
27. Dumitrascu B, Villar S, Mixon DG, Engelhardt BE. Optimal marker gene selection for cell type discrimination in single cell analyses. *Nat Commun*. 2021;12(1):1–8.
28. Van den Berge K, Perraudeau F, Soneson C, Love MI, Risso D, Vert JP, et al. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol*. 2018;19(1):24.
29. Azodi CB, Zappia L, Oshlack A, McCarthy DJ. splatPop: simulating population scale single-cell RNA sequencing data. *Genome Biol*. 2021;22(1):341.
30. De Donno C, Hedyeh-Zadeh S, Moinfar AA, Wagenstetter M, Zappia L, Lotfollahi M, et al. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nat Methods*. 2023;20(11):1683–92.
31. Jerber J, Seaton DD, Cuomo ASE, Kumasaka N, Haldane J, Steer J, et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat Genet*. 2021;53(3):304–12.
32. Lu S, Keles S. Debaised personalized gene coexpression networks for population-scale scRNA-seq data. *Genome Res*. 2023;33(6):932–47.
33. Ren X, Wen W, Fan X, Hou W, Su B, Cai P, et al. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell*. 2021;184(23):5838.
34. Stephenson E, Reynolds G, Botting RA, Calero-Nieto FJ, Morgan MD, Tuong ZK, et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nat Med*. 2021;27(5):904–16.
35. Andreatta M, Berenstein AJ, Carmona SJ. scGate: marker-based purification of cell types from heterogeneous single-cell RNA-seq datasets. *Bioinformatics*. 2022;38(9):2642–4.
36. Guo H, Li J. scSorter: assigning cells to known cell types according to marker genes. *Genome Biol*. 2021;22(1):1–18.
37. Zhang Z, Luo D, Zhong X, Choi JH, Ma Y, Wang S, et al. SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples. *Genes*. 2019;10(7): 531.
38. Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, et al. Cell Marker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res*. 2019;47(D1):D721–8.
39. Paisley BM, Liu Y. GeneMarker: a database and user interface for scRNA-seq marker genes. *Front Genet*. 2021;12:763431.
40. Fischer S, Gillis J. How many markers are needed to robustly determine a cell's type? *Iscience*. 2021;24(11):103292.
41. Yao Z, van Velthoven CT, Nguyen TN, Goldy J, Sedenó-Cortés AE, Baftizadeh F, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*. 2021;184(12):3222–41 e26.
42. Yao Z, Liu H, Xie F, Fischer S, Adkins RS, Aldridge AI, et al. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*. 2021;598(7879):103–10.
43. Franzén O, Gan L-M, Björkegren JL. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*. 2019;2019.
44. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol*. 2009;10(11):1–8.
45. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci*. 2004;101(16):6062–7.
46. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44(W1):W90–7.
47. Lin Y, Ghazanfar S, Wang KYX, Gagnon-Bartsch JA, Lo KK, Su X, et al. scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc Natl Acad Sci U S A*. 2019;116(20):9775–84.
48. MacDonald JW, Ghosh D. COPA—cancer outlier profile analysis. *Bioinformatics*. 2006;22(23):2950–1.
49. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*. 2005;310(5748):644–8.

50. Tibshirani R, Hastie T. Outlier sums for differential gene expression analysis. *Biostatistics*. 2007;8(1):2–8.
51. Wu B. Cancer outlier differential gene expression detection. *Biostatistics*. 2007;8(3):566–75.
52. Lian H. MOST: detecting cancer differential gene expression. *Biostatistics*. 2008;9(3):411–8.
53. de Ronde JJ, Rigauil G, Rottenberg S, Rodenhuis S, Wessels LF. Identifying subgroup markers in heterogeneous populations. *Nucleic Acids Res*. 2013;41(21): e200.
54. Rottenberg S, Vollebergh MA, de Hoon B, de Ronde J, Schouten PC, Kersbergen A, et al. Impact of intertumoral heterogeneity on predicting chemotherapy response of BRCA1-deficient mammary tumors. *Cancer Res*. 2012;72(9):2350–61.
55. Meng G, Pan Y, Tang W, Zhang L, Cui Y, Schumacher FR, et al. imply: improving cell-type deconvolution accuracy using personalized reference profiles. *Genome Med*. 2024;16(1):65.
56. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7): e47.
57. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol*. 2018;36(5):421–7.
58. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform*. 2020;2(3):lqaa078.
59. Lin Y, Cao Y, Willie E, Patrick E, Yang JYH. Atlas-scale single-cell multi-sample multi-condition data integration using scMerge2. *Nat Commun*. 2023;14(1):4272.
60. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. 2018; 37(1):38–44.
61. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. 2019;16(12):1289–96.
62. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol*. 2020;21(1):12.
63. Kang HM SM, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexing droplet-based single cell RNA-sequencing using genetic barcodes. GSE96583. *Gene Expression Omnibus*. 2017. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96583>.
64. Ren X WW, Fan X, Hou W et al. Large-scale single-cell analysis reveals critical immune characteristics of COVID-19 patients. GSE158055. *Gene Expression Omnibus*. 2020. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158055>.
65. Luxiao Chen, Zhenxing Guo, Tao Deng, Hao Wu. scCTS: v0.1.0. Github. 2024. <https://www.github.com/ToryDeng/scCTS>.
66. Luxiao Chen, Zhenxing Guo, Tao Deng, Hao Wu. scCTS: v0.1.0. Zenodo. 2024. <https://doi.org/10.5281/zenodo.13851702>.
67. Luxiao Chen, Zhenxing Guo, Tao Deng, Hao Wu. scCTS simulated and preprocessed real data. 2024. <https://www.doi.org/105281/zenodo13850742>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.