# SCMarker: ab initio marker selection for single cell transcriptome profiling

Fang Wang[1]*; Shaoheng Liang[1]; Tapsi Kumar[2]; Nicholas Navin[1,2]; Ken Chen[1]*

[1] Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA
[2] Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA

* To whom correspondence should be addressed. Tel: +1 713-793-3109; Email: fwang9@mdanderson.org. Correspondence also may be addressed to Ken Chen. Tel: +1 713-794-1348; Email: kchen3@mdanderson.org.

## Abstract

Single-cell RNA-sequencing data generated by a variety of technologies, such as Drop-seq and SMART-seq, can reveal simultaneously the mRNA transcript levels of thousands of genes in thousands of cells. It is often important to identify informative genes or cell-type-discriminative markers to reduce dimensionality and achieve informative cell typing results. We present an *ab initio* method that performs unsupervised marker selection by identifying genes that have subpopulation-discriminative expression levels and are co- or mutually-exclusively expressed with other genes. Consistent improvements in cell-type classification and biologically meaningful marker selection are achieved by applying SCMarker on various datasets in multiple tissue types, followed by a variety of clustering algorithms. The source code of SCMarker is publicly available at https://github.com/KChen-lab/SCMarker.

## Author Summary

Single cell RNA-sequencing technology simultaneously provides the mRNA transcript levels of thousands of genes in thousands of cells. A frequent requirement of single cell expression analysis is the identification of markers which may explain complex cellular states or tissue composition. We propose a new marker selection strategy (SCMarker) to accurately delineate cell types in single cell RNA-sequencing data by identifying genes that have bi/multi-modally distributed expression levels and are co- or mutually-exclusively expressed with some other genes. Our method can determine the cell-type-discriminative markers without referencing to any known transcriptomic profiles or cell ontologies, and consistently achieves accurate cell-type-discriminative marker identification in a variety of scRNA-seq datasets.

## Introduction

Current single-cell RNA-sequencing (scRNA-seq) data generated by a variety of technologies such as Drop-seq and SMART-seq, can reveal simultaneously the mRNA transcript levels of thousands of genes in thousands of cells [1-3]. However, the increased dimensionality makes it challenging to delineate cell types, due to complex and often undefined associations between individual genes and cell-types [4,5]. It is well accepted that genes are not equally informative in delineating cell types [6,7]. Certain genes are only expressed in certain cell types, but not others [8]. Moreover, the expression levels of certain genes cannot be robustly measured (e.g., zero inflated), due to technological bias [9-11]. Thus, it has become a common practice to retain only highly expressed or highly variable genes for cell populational analysis [12-15]. Several scRNA-seq data clustering packages (**S1 Table**) perform marker selection through dimensionality reduction techniques such as principal component analysis and tSNE [16], which are equivalent to identifying the set of highly variable genes. Unfortunately, the biological implications and the technical optimality of these gene selection strategies retain unclear, despite their wide use in cell-type clustering.

2

Here, we propose an *ab initio* method, named SCMarker, which applies information-theoretic principles to determine the optimal gene subsets for cell-type identification, without referencing to any known transcriptomic profiles or cell ontologies. The central idea of our method is to select genes that are individually discriminative across underlying cell types, based on a mixture distribution model, and are co- or mutually exclusively expressed with some other genes, due to cell-type specific functional constraints. Although the techniques of applying a mixture distribution model for a set of continuous data points have been widely used in clustering analysis of gene expressions, it is unclear whether this approach can benefit this problem context [17,18]. In particular, because single-cell gene expression measurements have vast dimensions (>20,000 genes), are highly noisy (e.g., zero-inflated, drop-off errors), and are generated by technologies of varied properties [19,20]. For example, SMART-seq is aimed at sequencing the entire RNA transcript, while Drop-seq only the 3' end using unique molecular indices (UMI) to track individual transcript [2,21,22]. Part of our investigation here is to examine whether the previously applied data analytical techniques can be reapplied in the single-cell data-type that have different properties and population structures. Our main goal is to identify not only cell-types, but also biologically meaningful cell-type markers from scRNA-seq data at accuracies higher than results derived using canonical gene selection strategies.

## Materials and Methods

### Discriminativeness of gene expressions for subpopulation clustering

By definition, cell-type-discriminative markers (CTDMs) should have distinctive expression levels across cell subpopulations. Therefore, in a dataset with mixed cell subpopulations, the expression level of a CTDM should follow a bi- or multi-modal, instead of a unimodal distribution (**Fig. 1A and B**) [23-25]. Following this assumption, we quantify the degree of modality based on the probability density distribution ($f$) of each gene expression using a Gaussian kernel function, instead of a mixture model which requires knowing the number of mixture components:

$$\hat{f}_h(g) = \frac{1}{L}\sum_{j=1}^{L} K_h(g - g_j) = \frac{1}{L \cdot h}\sum_{j=1}^{L} K\left(\frac{g - g_j}{h}\right), \qquad (1)$$

where $g_j$ is the expression level of a gene in cell $j$; $h > 0$ is a smoothing parameter called bandwidth, which is estimated by the average level of gene expression in cell population and can be leveraged to alleviate biases introduced by sequencing depths; and $L$ is the number of cells. $K_h(x)$ is a scaled kernel function defined as $\frac{1}{h}K(x/h)$, where $K(x)$ is a standard Gaussian density function. For each gene, we count the number ($T$) of peaks in the estimated probability density function $\hat{f}_h(g)$. A peak is found at the density value $c$, if there exists a 2 times $h$ long interval $I$ centred at $c$ such that $\hat{f}_h(g) \leq \hat{f}_h(c)$ for all $g$ in $I$. A gene expression level follows a multi-modal distribution, if it has multiple ($T \geq 2$) local maximum probability density values. Only genes with multimodal probability density distributions are considered as markers.
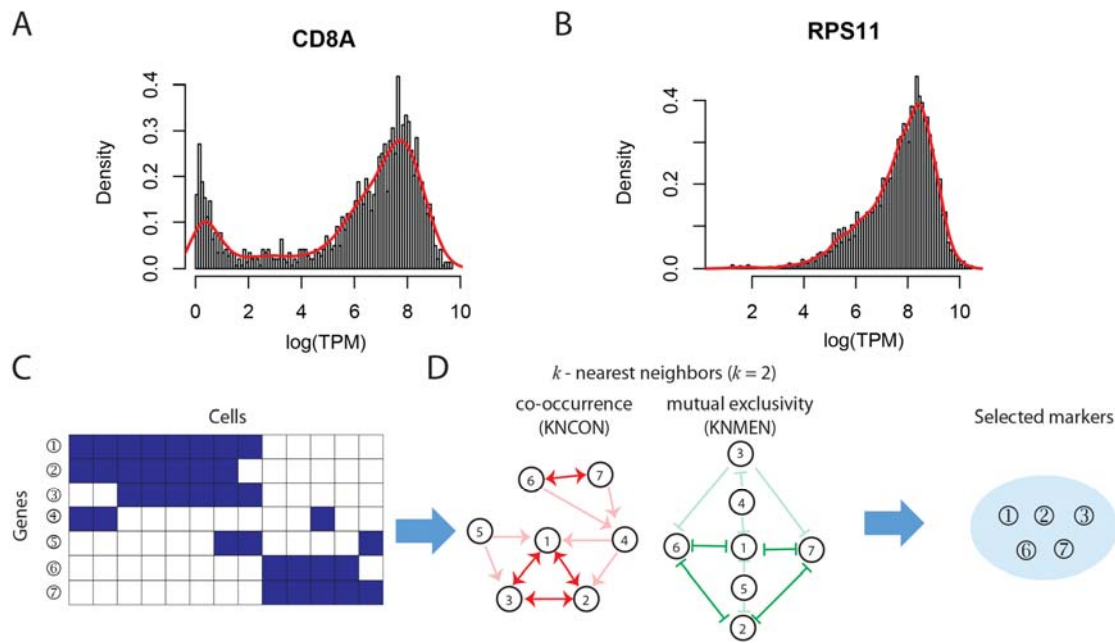
**Fig. 1**. Illustration of SCMarker. Plotted as examples are (**A**) a bimodally distributed gene expression and (**B**) a unimodally distributed gene expression. From a binarized gene-cell expression matrix (**C**), a *k*-nearest co-occurrence neighbour (KNCON) graph and a *k*-nearest mutually exclusive neighbour (KNMEN) graph are constructed (**D**), based on which co- or mutually exclusively expressing gene pairs (CMEGPs) in the KNCON (node 1, 2 and 3, node 6 and 7, connected by red double arrows) and in the KNMEM (node 1, 2, 6 and 7, connected by green double arrows) can be identified. Marker genes (node 1, 2, 3, 6 and 7) are subsequently selected based on the CMEGPs.

## Co- or mutually exclusively expressing gene pairs (CMEGPs)

CTDMs are often co- or mutually exclusively expressed, due to modularized regulatory interactions specific to cell types [26]. Consequently, identifying these CMEGPs, will help identify CTDMs. Because scRNA-seq data are often sparse with limited sequencing depth [27], binarization of the counts would help mitigate technical artifacts and improve robustness over different sequencing platforms (e.g., whole transcript vs 3' sequencing protocols). To identify CMEGPs, we only consider genes with multimodal distribution and discretize a gene-cell expression matrix with $N$ genes and $L$ cells into an $N \times L$ binary matrix $X \in \{0, 1\}^{N \times L}$, with $x_{ij} = 1$ designating an expressed gene $i$ in cell $j$, if the expression level is above the average and $x_{ij} = 0$ otherwise (**Fig. 1C**). For gene $i$, $x_{i.} = (x_{i1}, x_{i2}, \dots, x_{iL})$ is a binary string. We can calculate a co-occurrence matrix (S) that measures the pair-wise co-occurrence between all the gene pairs,

$$S = X \cdot X'. \tag{2}$$

S can also be represented as a directed graph (**Fig 1D**), in which a node denotes a gene and an edge from gene A to gene B represents that B co-occurs with gene A in at least $n$ cells. Among the connected nodes, the $k$ genes that co-occurred with A in $k$ largest sets of cells are termed the $k$-nearest co-occurrence neighbours (KNCONs).

4

In addition, we calculate a mutually exclusive matrix ($M$) that measures the pair-wise mutual exclusivity between all the gene pairs through equation (3),

$$M = (1 - X) \cdot X', \tag{3}$$

where $M$ represents a directed $k$-nearest mutually exclusive neighbour (KNMEN) graph (**Fig. 1D**). Similar to KNCON, but in opposite ways, the KNMENs of a gene A are the $k$ genes that occur mutually exclusively with A in $k$ largest sets of cells.

Under these definitions, an CMEGP is identified as two genes bi-directionally connected in the KNCEN or the KNMEN graph. We selected as markers the genes that belong to at least one CMEGP, because these genes are more likely associated with cell-type specific functions than those that do not have any CMEGP (likely due to random, non-function-related fluctuation). This concept has been previously examined in the RNA microarray data analysis, but has not been successfully applied in the context of single-cell RNA-seq data analysis, due to vastly different properties between the technologies [17,18].

## Results

### Comparison with other marker selection strategies

We applied SCMarker to the scRNA-seq data obtained from 1) 19 melanoma patients, which include 4,645 cells; and 2) 18 head and neck cancer patients, which include 5,902 cells sequenced by the SMART-seq2 platform [28,29]. In the original studies, each cell in the sets was labelled as a malignant or non-malignant cell through copy number analysis. The expression levels of known marker genes were used to further classify the non-malignant cells, such as T cells, B/plasma cells, macrophages, dendritic cells, mast cells, endothelial cells, fibroblasts, and myocytes. We found that most known marker genes (96%) demonstrated bi/multi-modal distributions across cells (**S1 Fig.**). Overall, around 6% of genes with bi/multi-modal distributions are identified as marker genes, among which half are the known marker genes.

For a fair comparison, we assessed SCMarker results with those obtained under two canonical strategies: selecting genes with A) the highest average expression levels and B) the highest variance across cells. In our experiments, the highest variable genes were determined using Seurat [12]. We used five clustering methods: k-means, Clara, hierarchical clustering, DBSCAN and Seurat to cluster single cells based on the selected markers [30,31]. Same numbers of clusters are specified for DBSCAN, k-means, Clara, and hierarchical clustering. The adjusted rand index (ARI), which measures the similarity of two sets of clustering results, was used to quantify the consistency between the clustering results and the known cell labels [32]. Compared to marker sets A and B, selected by the canonical strategies, the marker set selected by SCMarker (equal numbers of markers) resulted in a higher ARI with fairly evident margins (**Fig. 2**). The conclusion appeared to be robust over a range of $k$ and $n$ parameters and were unaffected by using different clustering methods (S2 **Fig.**).
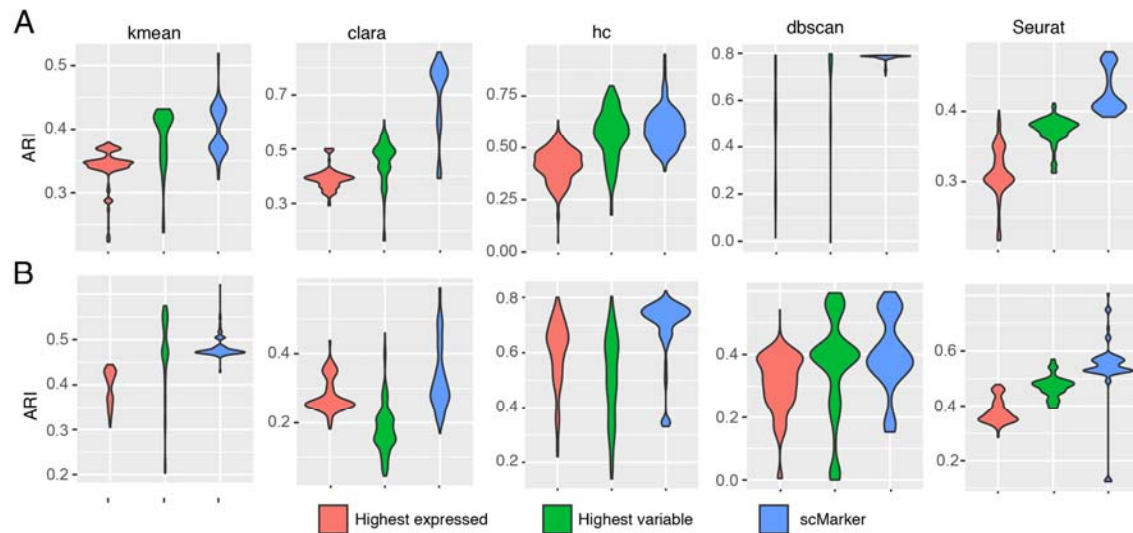
**Fig. 2** Comparison of 3 marker selection methods for cell-type identification. Accuracy of cell-type identification (in terms of adjusted rand index) are compared across 3 marker sets selected respectively by SCMarker, the highest expressed and the highest variable gene approaches, using two scRNA-seq datasets from (A) melanoma and (B) head-and-neck cancer samples by 5 clustering algorithms: k-means, Clara, hierarchical clustering (hc), DBSCAN, and Seurat.

These experiments indicated that setting $k$ between 100 and 300 resulted in the most accurate cell type identification results irrespectively to $n$ (**S3 Fig.**). Hence, we select $k = 300$ and $n = 30$ as the default parameters for applying SCMarker.

We obtained 902 markers from the melanoma data and more distinguishable cell types using SCMarker than using the canonical strategies (**Fig. 3A** to **C**). Better performance of SCMarker was also obtained in analysing the head and neck cancer data (**S4 Fig.**). Moreover, the genes selected by SCMarker had substantially higher degrees of overlap with the known cell-type markers reported in the original publications than the sets returned by other approaches (the same number of 902 top scoring genes were selected for fair comparison), including the "FindMarker" approach in Seurat (**Fig. 3D to G**). Notably, SCMarker selected significantly more immune cell surface markers specific to T cytotoxic, T helper, B lymphocyte, and macrophage cells that are likely present in the tumour microenvironment, as indicated by gene set enrichment analysis (**S5 Fig.**) [33].
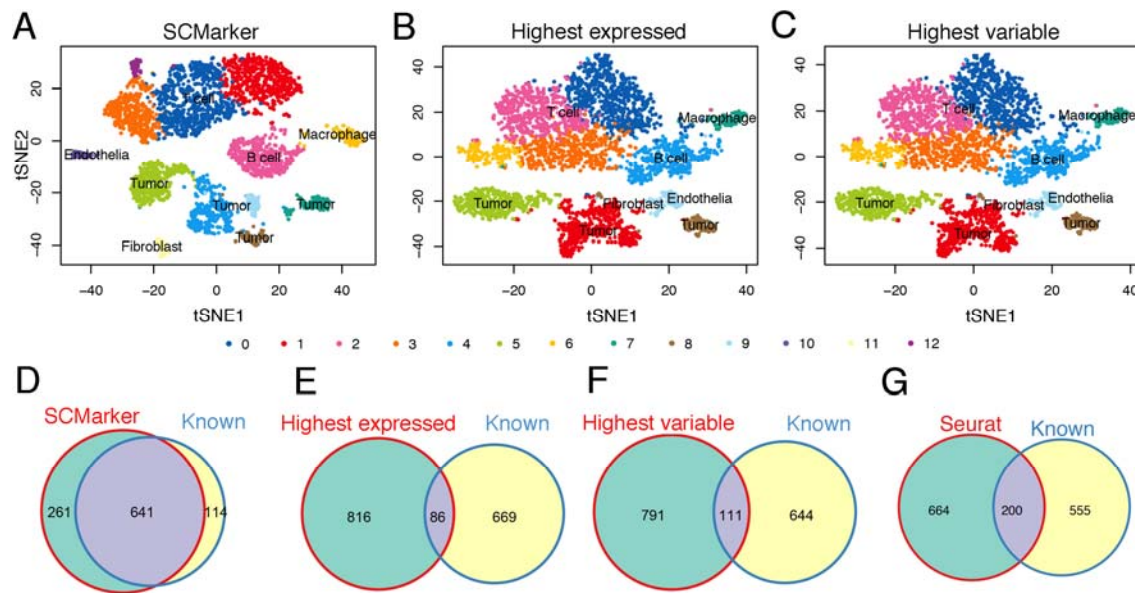
6

**Fig. 3.** Results on the melanoma data. Plotted in tSNE space are 4,645 melanoma cells with markers selected respectively by (**A**) SCMarker, (**B**) the highest expressed and (**C**) the highest variable genes. Also plotted are the Venn diagrams between the known cell-type markers and the marker sets determined respectively by (**D**) SCMarker, (**E**) the highest expressed, (**F**) the highest variable genes and (**G**) Seurat FindMarker in the melanoma data.

**Application of SCMarker to 3' UMI count data**

To avoid introducing biases due to potential overfitting and assess the utility of our approaches on other platforms, we further assessed the utility of SCMarker in analysing the 3' UMI count scRNA-seq data generated by the droplet platforms.

We first analysed a set of 5,602 cells from the cerebellar hemisphere of normal brain tissues generated by Drop-seq [34]. SCMarker selected 699 genes as CTDMs, which differentially expressed across cell subpopulations under the default parameters. Alternatively, the default mode of Seurat led to the selection of 6,111 highest variable genes (HVGs). For comparison, we selected 699 highest expressed genes (HEGs). Although SCMarker selected less markers than Seurat, the clustering result showed a clearer separation than that based on the Seurat HVGs and on the HEGs (**Fig. 4A** to **C**). In particular, SCMarker successfully delineated Purkinje neurons into purk1 (cluster4, **Fig. 4A**) and purk2 (cluster7, **Fig. 4A**) and recapitulated the differential levels of *SORCS3* between two clusters (**Fig. 4D**), which are consistent with the results in the original paper. In contrast, although the Purkinje neurons were clustered into four groups by Seurat (**Fig. 4B**), purk1 and purk2 were not well separated (**Fig. 4B**), and the expression levels of *SORCS3* showed mosaic patterns across the 4 groups (cluster4, 6, 11 and 12, **Fig. 4E**). As additional controls, we performed clustering using the top 500 and 1000 Seurat HVGs. That did not result in any improvement (**S6 Fig.**).
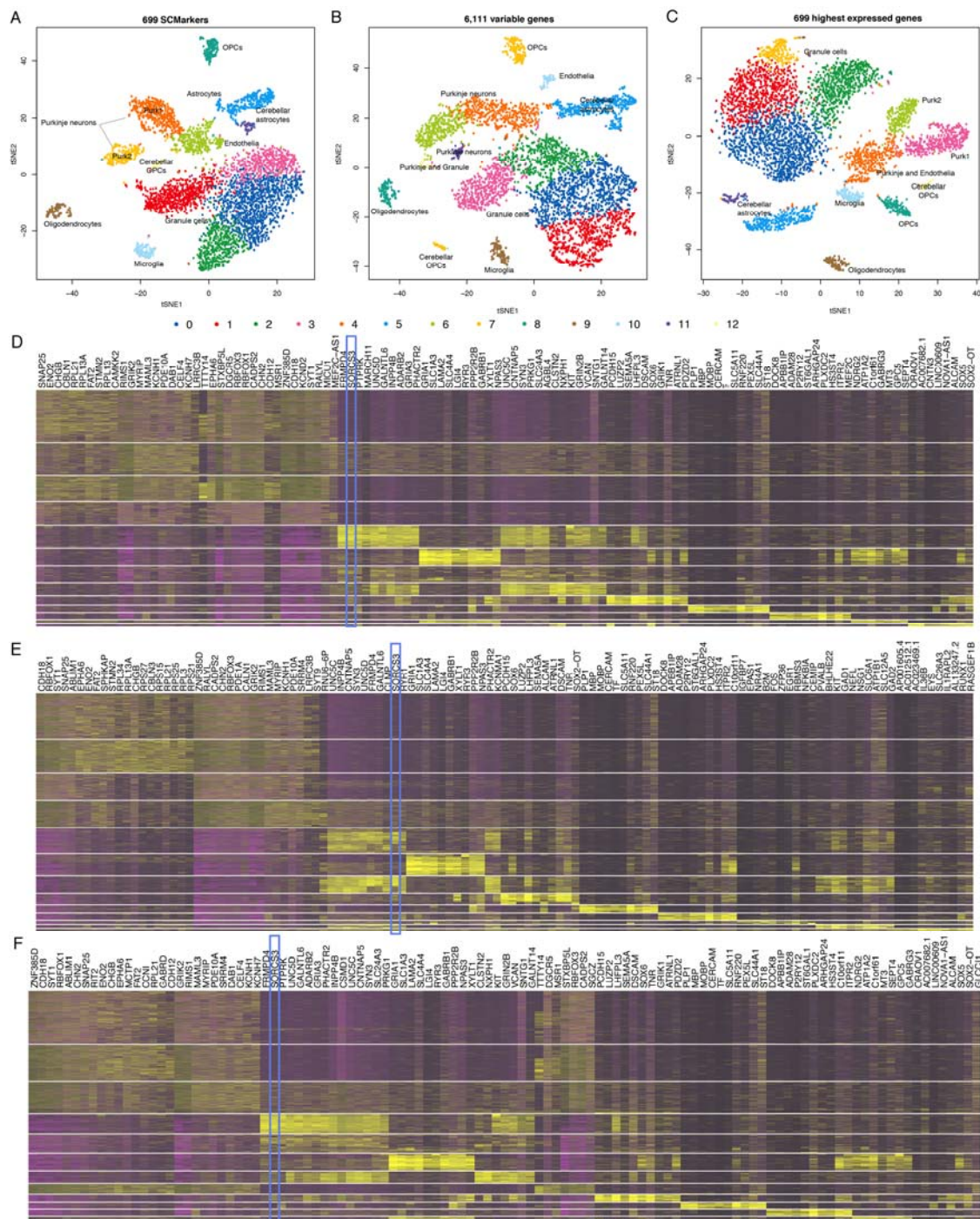
**Fig 4**. Results on the human brain data. Plotted in tSNE space are 5,602 cells in the cerebellar hemisphere of human brain tissue based on markers selected respectively by (**A**) SCMarker,(**B**) the highest variable genes and (**C**) the highest expressed genes, colored by performing clustering using Seurat. Cell types were labelled consistently as they were in the original paper. Also plotted are the heatmaps of the top 10 gene expression levels in each cluster derived respectively from (**D**) SCMarker,(**E**) the highest variable genes and (**F**) the highest expressed genes.

We then analysed the scRNA-seq data of 52,698 cells from 5 lung tumours generated by the 10X Chromium platform (10X Genomics) [35]. SCMarker identified 848 markers under the default

8

parameters, while Seurat identified 1,832 HVGs. We also selected 848 HEGs for comparison. SCMarker led to 22 clearly distinguishable clusters, while the Seurat HVGs led to 12 and the HEGs led to 18 (**Fig. 5A** to **C**). The 10 highest expressed markers per cluster derived by SCMarker showed a high degree of cluster-specificity in the heatmap (**Fig. 5D**). Among the selected markers were the 16 known markers reported by the original study (**Table 1**). SCMarker also discovered multiple putative subtypes for some cell types, such as the T, B, fibroblast, and myeloid cells (**Table 1, Fig. 5A**). For example, cluster 7, 10, 13, 19, and 22 are the B cells expressing known surface marker *CD79A* (**Fig. 5A**), yet cells in cluster 7, 13, 19, and 22 are evidently different from cells in cluster 10, due to differential *IGHG1* and *BANK1* expression levels (**Fig. 5D**). For comparison, we selected the 10 highest expressed genes from the clusters determined by the Seurat HVGs and by the HEGs, respectively (**Fig. 5E** and **F**). They appeared non-specifically distributed across clusters (**Fig. 5E** and **F**). These genes also contained fewer known markers (**Table 1**). For example, cluster 3 determined by the Seurat HVGs contained markers (*CLDN5*, *CAV1* and *IFITM3*) from 3 cell-types (endothelia, alveolar and B cell, respectively). Most clusters expressed *IFITM3*, except for clusters 1 and 6 (**Fig. 5E**). Only T cell and fibroblast markers appeared to be cluster-specific. As additional controls, we also performed analysis using fewer (i.e., 500 and 1000) Seurat HVGs. That resulted in worse results with fewer known markers and marker-specific clusters (**S7 Fig**.).
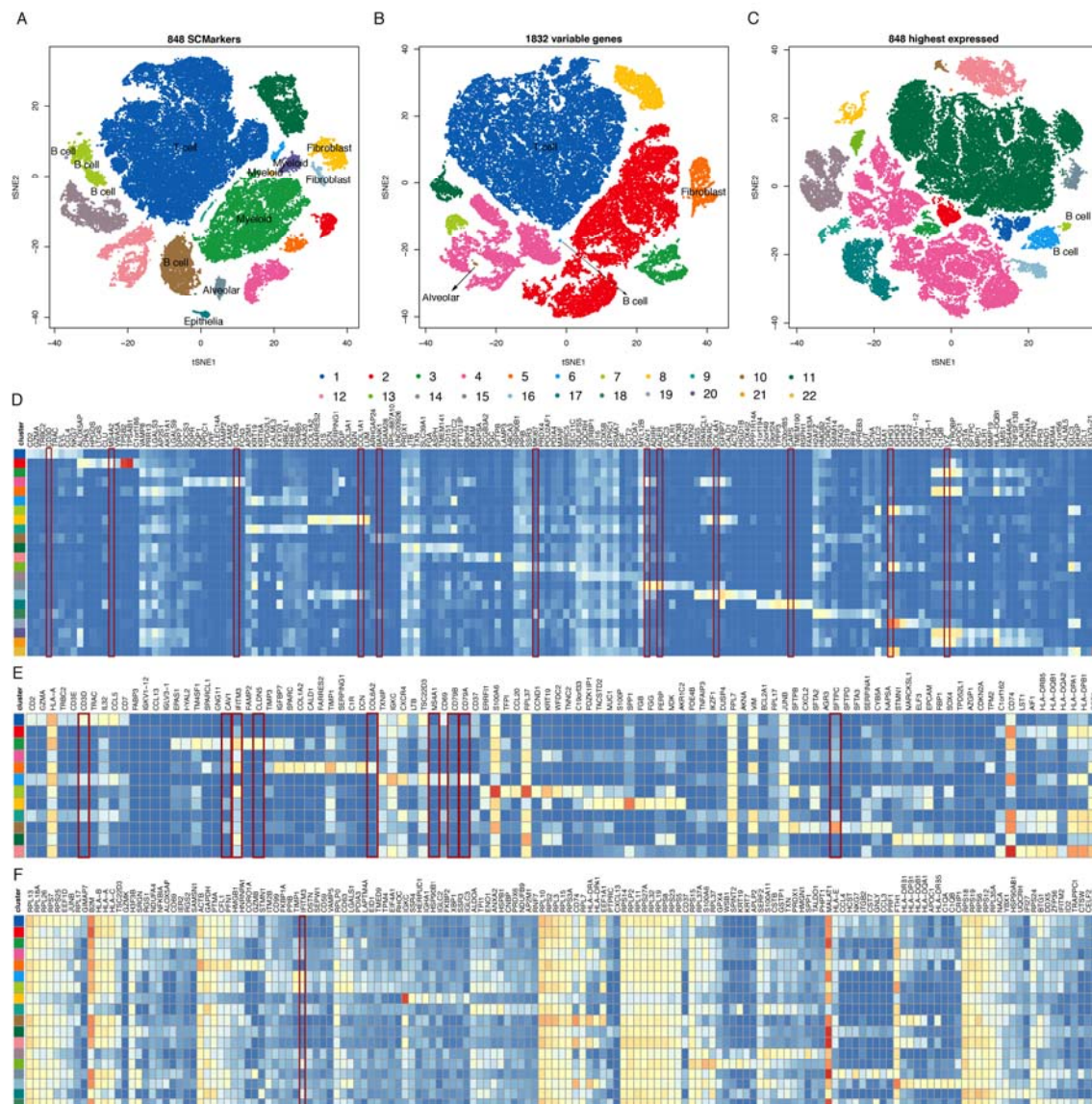
**Fig. 5**. Results on the lung cancer data generated by Dropseq. Plotted in tSNE space are 52,698 cells of 6 different lung cancer patients, clustered based on markers selected respectively by (**A**) SCMarker, (**B**) the highest variable genes and (**C**) the highest expressed genes. Colors correspond to clusters determined by DBSCAN. Heatmaps of the average expression levels of the 10 highest expressed genes per cluster identified respectively by SCMarker (**D**), the highest variable genes (**E**) and the highest expressed genes (**F**). Cell types in (**A**) to (**C**) are labelled based on the known cell-type specific markers, which are highlighted in red boxes in (**D**) to (**F**).

**Table 1. Known cell-type specific markers identified respectively by SCMarker and the highest variable gene approach.**

| Known markers | Cell type | Cluster ID (SCMarker) | Cluster ID (highest variable) | Cluster ID (highest expressed) |
|---|---|---|---|---|
| *CD3D* | T cell | 1 | 1 | |
| *MKI67* | T cell | 18 | | |
| *IGHG1* | B cell | 7, 13, 19, 22 | | |

10

| MS4A2 | Mast | 2 | | |
| LYZ | Myeloid | 3, 5, 20, 21 | | |
| CLDN5 | Endothelial | 4 | 3 | |
| COL1A1 | Fibroblast | 8 | | |
| COL4A1 | Fibroblast | 16 | | |
| CAPS | Epithelial | 17 | | |
| CAV1 | Alveolar | 15 | 3 | |
| AGER | Alveolar | 15 | | |
| IFIMT3 | Myeloid | | 3 | 6, 7 |
| COL6A2 | Fibroblast | | 5 | |
| MS4A1 | B cell | | 6 | |
| CD79A | B cell | | 6 | |
| SFTPC | Alveolar | | 10 | |

Only the top 10 highest expressed markers per cluster are included in the comparison.

Overall, SCMarker demonstrated higher sensitivity and specificity for cell type and cell-type specific marker identification than the alternative approaches. Moreover, markers selected by SCMarker were more significant among genes which were identified by Seurat to define clusters (**S8 Fig.**).

## Discussion

In this manuscript, we reported a new bioinformatics tool, SCMarker, which performs *ab initio* cell-type discriminative marker selection from scRNA-seq data. SCMarker operates based on two new information-theoretic metrics: 1) bi/multi-modal distribution of subpopulation-discriminative gene expression in mixed cell populations and 2) co- or mutually-exclusively expressing gene pairs, which quantifies populational structural properties intrinsic to single-cell RNA-seq data. We found that SCMarker can consistently significantly boost cell-type identification accuracy in datasets from a variety of tissues such as cancer and brain, generated by both SMART-seq and Drop-seq platforms. Because SCMarker does not depend on any prior knowledge, we anticipate that it will prove most useful in discovery settings for analysing cell populations of a high degree of plasticity and heterogeneity [36]. SCMarker can potentially be expanded to analyse other types of single-cell data, including mass cytometry and single cell ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) data [37,38]. It can be easily incorporated as a module into current scRNA-seq data analysis workflows to pre-process the cell-gene count/expression matrix before performing further downstream analysis.

## Acknowledgements

## Funding

# References

1. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA (2015) The technology and biology of single-cell RNA sequencing. Mol Cell 58: 610-620.
2. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, et al. (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161: 1202-1214.
3. Ramskold D, Luo S, Wang YC, Li R, Deng Q, et al. (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol 30: 777-782.
4. Campbell KR, Yau C (2018) A descriptive marker gene approach to single-cell pseudotime inference. Bioinformatics.
5. Bakken T, Cowell L, Aevermann BD, Novotny M, Hodge R, et al. (2017) Cell type discovery and representation in the era of high-content single cell phenotyping. BMC Bioinformatics 18: 559.
6. Trapnell C (2015) Defining cell types and states with single-cell genomics. Genome Research 25: 1491-1498.
7. Wagner A, Regev A, Yosef N (2016) Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol 34: 1145-1160.
8. Aevermann BD, Novotny M, Bakken T, Miller JA, Diehl AD, et al. (2018) Cell type discovery using single-cell transcriptomics: implications for ontological representation. Hum Mol Genet 27: R40-R47.
9. Pierson E, Yau C (2015) ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol 16: 241.
10. Yuan GC, Cai L, Elowitz M, Enver T, Fan G, et al. (2017) Challenges and emerging directions in single-cell analysis. Genome Biol 18: 84.
11. Chen L, Zheng S (2018) BCseq: accurate single cell RNA-seq quantification with bias correction. Nucleic Acids Res 46: e82.
12. Satija R, Farrell JA, Gennert D, Schier AF, Regev A (2015) Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 33: 495-502.
13. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, et al. (2017) SC3: consensus clustering of single-cell RNA-seq data. Nat Methods 14: 483-486.
14. Ho YJ, Anaparthy N, Molik D, Mathew G, Aicher T, et al. (2018) Single-cell RNA-seq analysis identifies markers of resistance to targeted BRAF inhibitors in melanoma cell populations. Genome Res.
15. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y (2015) SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. PLoS Comput Biol 11: e1004575.
16. van der Maaten L, Hinton G (2008) Visualizing Data using t-SNE. Journal of Machine Learning Research 9: 2579-2605.
17. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America 95: 14863-14868.
18. Quackenbush J (2001) Computational analysis of microarray data. Nature Reviews Genetics 2: 418-427.
19. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL (2001) Model-based clustering and data transformations for gene expression data. Bioinformatics 17: 977-987.
20. Fraley C, Raftery AE (2003) Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. Journal of Classification 20: 263-286.
21. Picelli S, Faridani OR, Bjorklund AK, Winberg G, Sagasser S, et al. (2014) Full-length RNA-seq from single cells using Smart-seq2. Nature Protocols 9: 171-181.

22. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, et al. (2017) Comparative Analysis of Single-Cell RNA Sequencing Methods. Molecular Cell 65: 631-+.
23. Montoro DT, Haber AL, Biton M, Vinarsky V, Lin B, et al. (2018) A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. Nature 560: 319-324.
24. Plasschaert LW, Zilionis R, Choo-Wing R, Savova V, Knehr J, et al. (2018) A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. Nature 560: 377-381.
25. Welch JD, Hartemink AJ, Prins JF (2017) MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol 18: 138.
26. Bartlett TE, Muller S, Diaz A (2017) Single-cell Co-expression Subnetwork Analysis. Sci Rep 7: 15066.
27. Hwang B, Lee JH, Bang D (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 50: 96.
28. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, et al. (2017) Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell 171: 1611-1624 e1624.
29. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, 2nd, Treacy D, et al. (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352: 189-196.
30. M. Ester H-PK, J. Sander, X. Xu (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96) AAAI Press, 1996: pp. 226–231.
31. Blashfield RK (1991) Finding Groups in Data - an Introduction to Cluster-Analysis - Kaufman,L, Rousseeuw,Pj. Journal of Classification 8: 277-279.
32. Santos J.M. EM (2009) On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. In: Alippi C, Polycarpou M, Panayiotou C, Ellinas G (eds) Artificial Neural Networks – ICANN 2009 ICANN 2009 Lecture Notes in Computer Science 5769.
33. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545-15550.
34. Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, et al. (2018) Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol 36: 70-80.
35. Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, et al. (2018) Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med.
36. Ye X, Weinberg RA (2015) Epithelial-Mesenchymal Plasticity: A Central Regulator of Cancer Progression. Trends Cell Biol 25: 675-686.
37. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, et al. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523: 486-490.
38. Qiu P, Simonds EF, Bendall SC, Gibbs KD, Jr., Bruggner RV, et al. (2011) Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol 29: 886-891.

**Supplementary Figure**

**S1 Fig.** Distributions of expression levels of known marker genes (CD3G, CD8A, IL7R, MS4A1, CD19, CD79A, CD79B and PMEL) from melanoma data.

**S2 Fig.** Comparison of 3 marker selection methods for cell-type identification. Tested were a range of parameters and 5 clustering algorithms: k-means, Clara, hierarchical clustering (hc), DBSCAN, and Seurat. Plotted in heatmaps are the ARI values calculated based on markers selected respectively by SCMarker, the highest expressed genes and the highest variable genes from (A) the melanoma and (B) the head and neck cancer data. X and Y axes in the SCMarker panel indicate the $n$ and $k$ parameters used by SCMarker and the corresponding (equal number of markers) results in the highest expressed or the highest variable gene panels.

**S3 Fig.** Determining the optimal parameters. Plotted in the heatmaps are the number of selected markers for (**A**) the melanoma and (**B**) the head and neck cancer data over a range of $n$ (X-axis) and $k$ (Y-axis) parameters. Bars on the side and the top are the mean values in the corresponding rows and columns. Also plotted are clustering accuracy measured by the adjusted rand index (ARI), a metric that measures the similarity of two clustering results, for (**C**) the melanoma and (**D**) the head and neck cancer data over various $n$ and $k$ parameters.

**S4 Fig.** Validation of genes selected by SCMarker. Plotted in tSNE space are 5,902 cells from the head and neck cancer data, based on genes selected respectively by (**A**) SCMarker, (**B**) the highest expressed and (**C**) the highest variable genes.

**S5 Fig.** Gene set enrichment analysis (GSEA) of markers selected by 3 methods: SCMaker, the highest expressed and the highest variable genes from the (**a**) melanoma; and (**b**) the head and neck cancer data, respectively. Only the top 15 terms are shown. The darkness of the colors corresponds to -log10 P values.

**S6 Fig.** Results on the human brain tissue data. Plotted in tSNE space are 5,602 cells in the cerebellar hemisphere of human brain tissue based on the highest 500 (**A**) and 1000 (**B**) variable genes, colored by cell types from the original paper.

**S7 Fig.** Results on the lung cancer data. Plotted in tSNE space are 52,698 cells of 6 different lung cancer patients, clustered based on the highest 500 (**A**) and 1000 (**B**) variable genes. Colors correspond to clusters determined by DBSCAN. Heatmaps of the average expression levels of the 10 highest expressed genes per cluster identified respectively by the highest 500 (**C**) and 1000 (**D**) variable genes. Cell types in (A) and (B) are labelled based on the known cell-type specific markers, which are highlighted in red box in (C) and (D).

**S8 Fig.** The distribution of significance of markers identified by Seurat and overlaps with SCMarker in melanoma, head and neck cancer (HNSCC), brain tissue and lung cancer.