PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

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, United States of America, **2** Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, United States of America

* fwang9@mdanderson.org (FW); kchen3@mdanderson.org (KC)

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

OPEN ACCESS

Citation: Wang F, Liang S, Kumar T, Navin N, Chen K (2019) SCMarker: Ab initio marker selection for single cell transcriptome profiling. PLoS Comput Biol 15(10): e1007445. https://doi.org/10.1371/journal.pcbi.1007445

Editor: Rachel Karchin, Johns Hopkins University, UNITED STATES

Received: April 10, 2019

Accepted: September 28, 2019

Published: October 28, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The work has been made possible in part by grant number 2018-182735 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. The work was supported by a Chan-Zuckerberg Initiative Human Breast Cell Atlas Seed Network Grant (HCA3-0000000147) to NN and KC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

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 population 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.

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 should have distinctive expression levels across cell subpopulations. Therefore, in a dataset with mixed cell subpopulations, the expression level of a marker should follow a bi- or multi-modal, instead of a unimodal distribution (**Fig 1A and 1B**) [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:

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

where g_j is the expression level of gene g in cell j, L is the number of cells and $h>0$ is a smoothing parameter called bandwidth, which can be leveraged to alleviate biases introduced by uneven sequencing depths. $K_h(x) = \frac{1}{h}K(x/h)$ is a scaled kernel function, where $K(x)$ is a standard Gaussian density function. The optimal value for h can be calculate from

$$h = \left({}^{4}\sigma^{5}/_{3}L\right)^{\frac{1}{5}}, \tag{2}$$

where the standard deviation $\sigma = \sqrt{\frac{1}{L-1}\sum_{i=1}^{L}\left(g_i - \bar{g}\right)^2}$, g_i is the expression level of gene g in

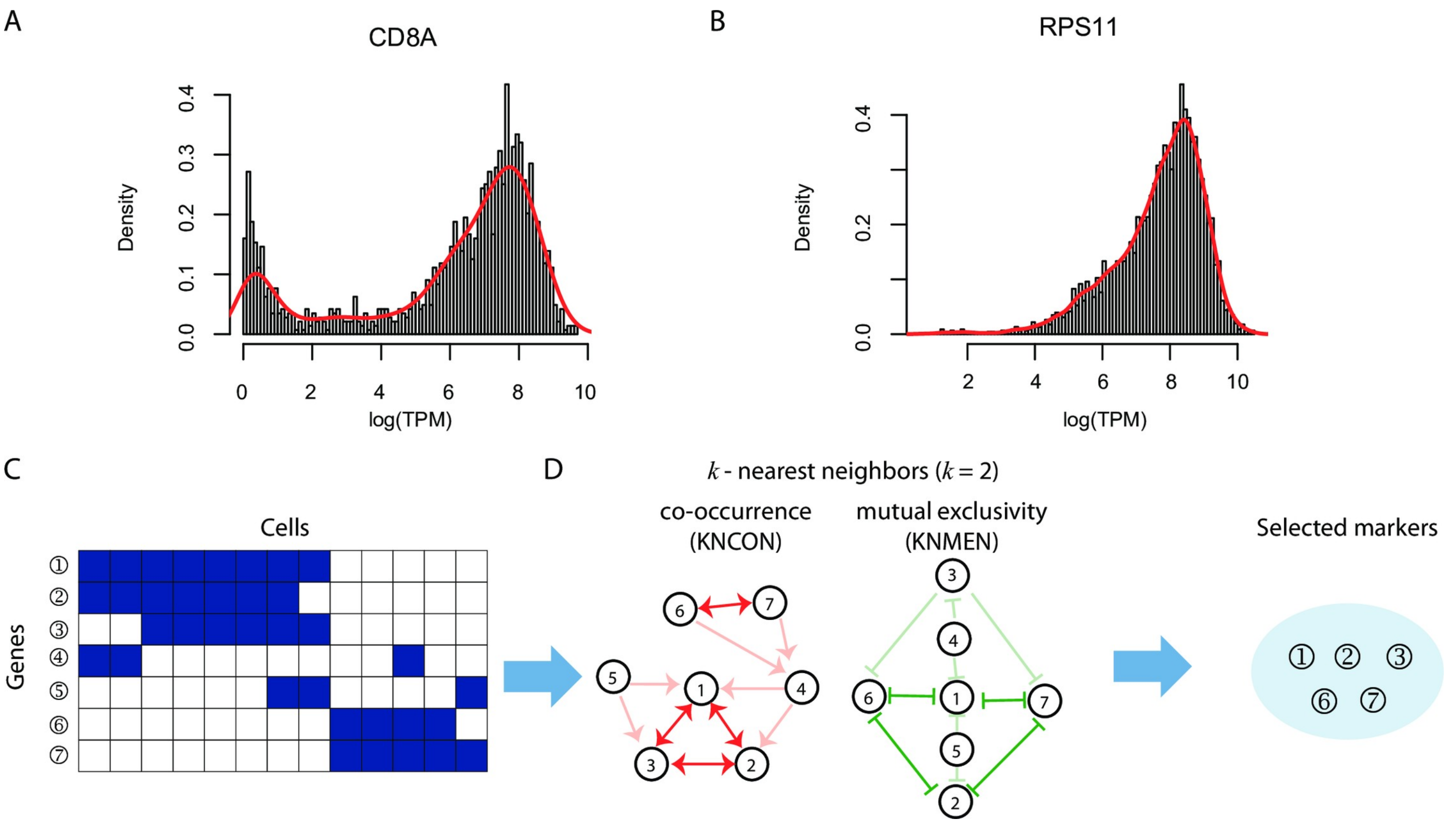

Fig 1. Illustration of SCMarker. Plotted as examples are (A) a bimodally distributed gene expression and (B) a unimodally distributed gene expression. From a binarized gene-cell expression matrix (C), a k-nearest co-occurrence neighbour (KNCON) graph and a k-nearest mutually exclusive neighbour (KNMEN) 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.

https://doi.org/10.1371/journal.pcbi.1007445.g001

cell i and $\bar{g}$ is the average expression level of g across all the cells [26]. We estimated h from the datasets used in this study [27–30] and obtained a mean $h = 0.3$ on the datasets generated by the SMART-seq platform and 0.05 on the datasets generated by the Drop-seq platform. We set these as the default values for subsequent analyses. 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.

Co- or mutually exclusively expressing gene pairs (CMEGPs)

Cell-type-discriminative markers are often co- or mutually exclusively expressed, due to modularized regulatory interactions specific to cell types [31]. Consequently, identifying these CMEGPs, will help identify markers which delineate cell type. Because scRNA-seq data are often sparse with limited sequencing depth [32], 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, $x_{i\cdot} = (x_{i1}, x_{i2}, \ldots, x_{iL})$ is a binary string. We can calculate a co-occurrence matrix (S) that measures the pair-wise co-occurrence between all the gene pairs,

$$\mathrm{S} = X \cdot X'. \tag{3}$$

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).

In addition, we calculate a mutually exclusive matrix (M) that measures the pair-wise mutual exclusivity between all the gene pairs through Eq (4),

$$M = (1 - X) \cdot X', \tag{4}$$

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. The detailed algorithm for constructing these graphs is shown in S1 Fig.

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 [27,28]. 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 (S2 Fig). Overall, around 6% of genes with bi/multi-modal distributions are identified as marker genes, among which half are the known marker genes.

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 [33,34]. 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 [35]. 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

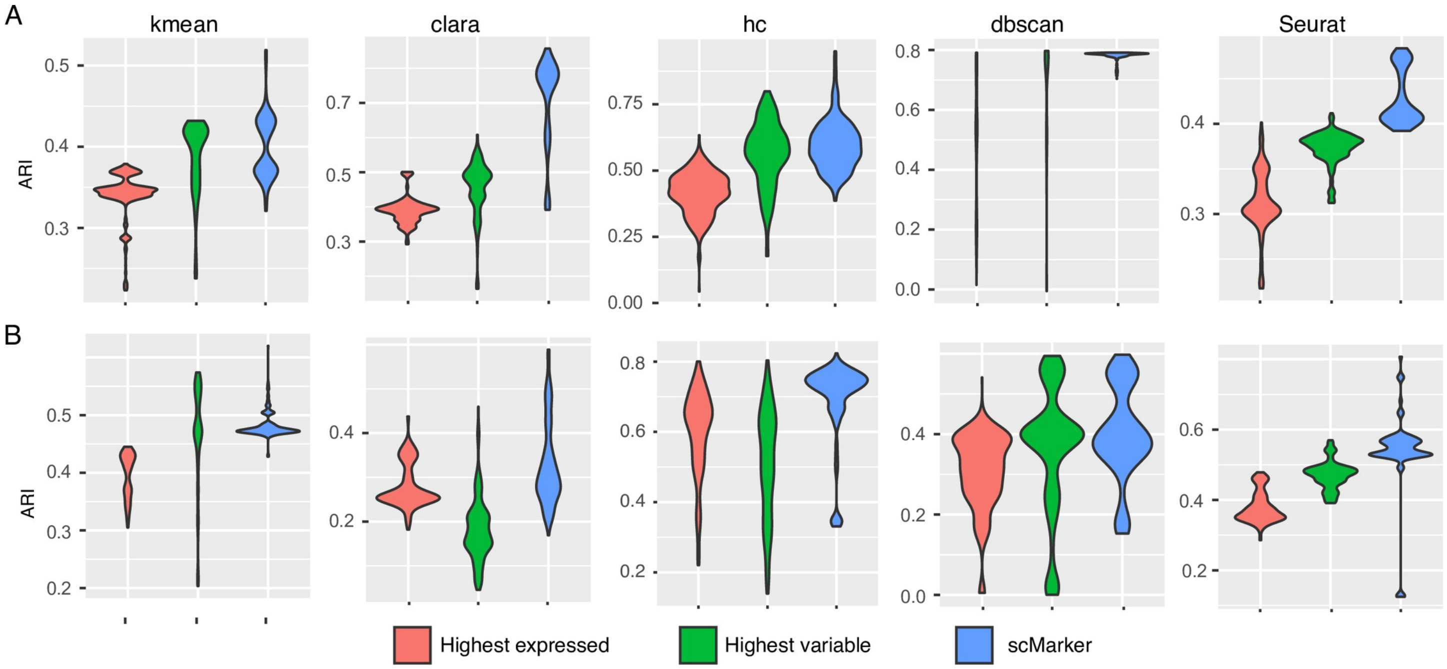

Fig 2. Comparison of 3 marker selection methods for cell-type identification over a range of k (from 50 to 1000 at a step-size of 50) and n (from 5 to 100 at a step-size of 5). Accuracy of cell-type identification (in terms of adjusted rand index) are compared across 3 marker sets selected respectively 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.

https://doi.org/10.1371/journal.pcbi.1007445.g002

clustering methods (S3 Fig).These experiments indicated that setting *k* between 100 and 300 resulted in the most accurate cell type identification results irrespectively to *n* (S4 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 3C, S1 Data). Better performance of SCMarker was also obtained in analysing the head and neck cancer data (S5 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 3G). 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 (S6 Fig) [36].

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 determined the optimal values of *k* and *n* under various sample sizes via resampling the melanoma data (S7 Fig). When the sample size was 500, the optimal values were *k* = 100 and *n* = 20 (or 30). When the sample size was 2,500 or 25,000, the optimal values were *k* = 300, *n* = 30, the same as what we obtained from the original dataset (4,645 cells). When sample size increased to 50,000, the optimal values increased to *k* = 400, *n* = 100. Overall, although the sample size did affect the optimal choices of *k* and *n*, their influences were relatively modest.

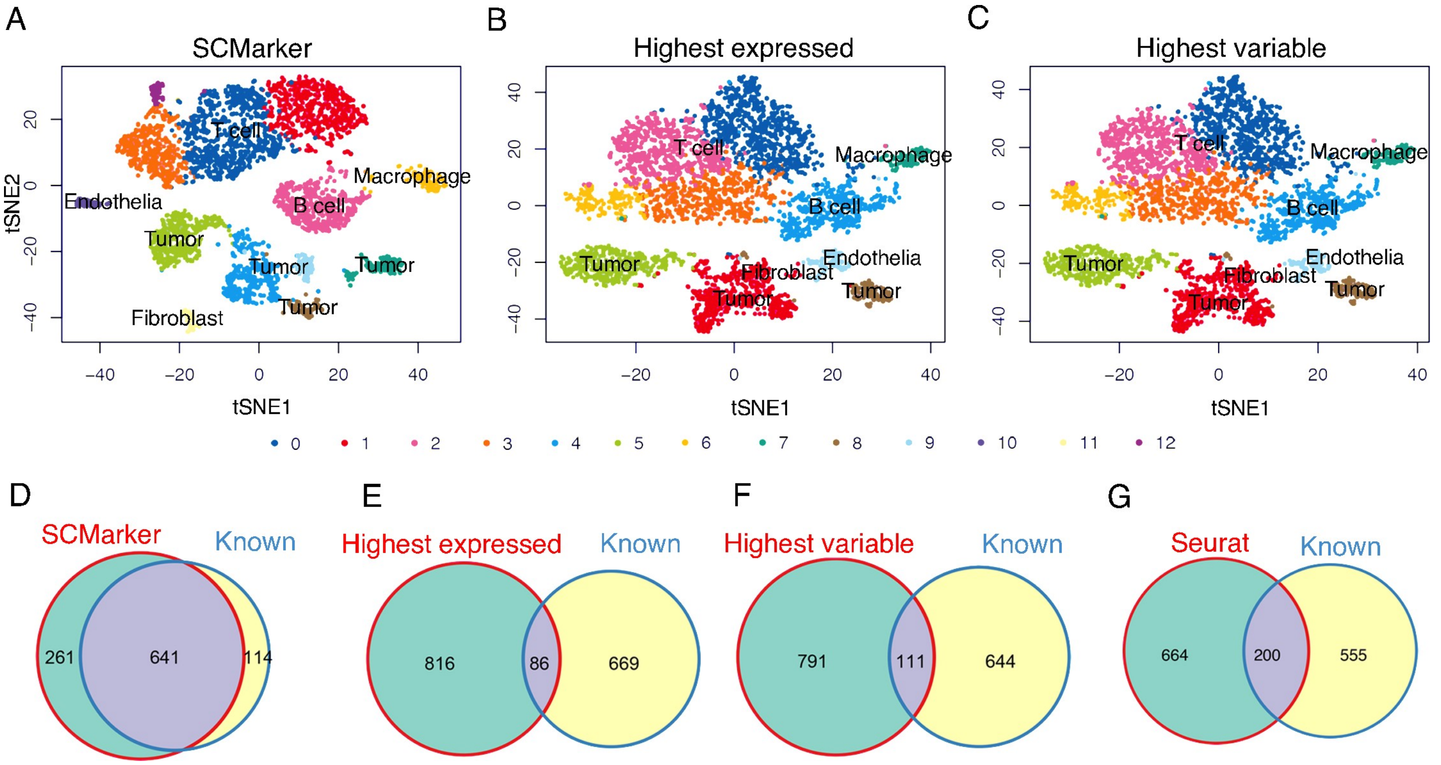

Fig 3. Results on the melanoma data. Plotted in tSNE space are 4,645 melanoma cells with markers selected respectively by (A) SCMarker, (B) the highest expressed and (C) the highest variable genes. Also plotted are the Venn diagrams between the known cell-type markers and the marker sets determined respectively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

https://doi.org/10.1371/journal.pcbi.1007445.g003

We first analysed a set of 5,602 cells from the cerebellar hemisphere of normal brain tissues generated by Drop-seq under k = 300 and n = 30 [29]. SCMarker selected 699 genes as markers, which differentially expressed across cell subpopulations under the default parameters (**S1 Data**). 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 **4C**). 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 (**S8 Fig**).We then analysed the scRNA-seq data of 52,698 cells from 5 lung tumours generated by the 10X Chromium platform (10X Genomics) under k = 400 and n = 100 [30]. SCMarker identified 950 markers under the parameters (**S1 Data**), while Seurat identified 1,832 HVGs. We also selected 950 HEGs for comparison. SCMarker led to 23 clearly distinguishable clusters, while the Seurat HVGs led to 12 and the HEGs led to 19 (**Fig 5A** to **5C**). 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 17 known markers reported by the original study (**Table 1**). SCMarker also discovered multiple putative subtypes for some cell types, such as the

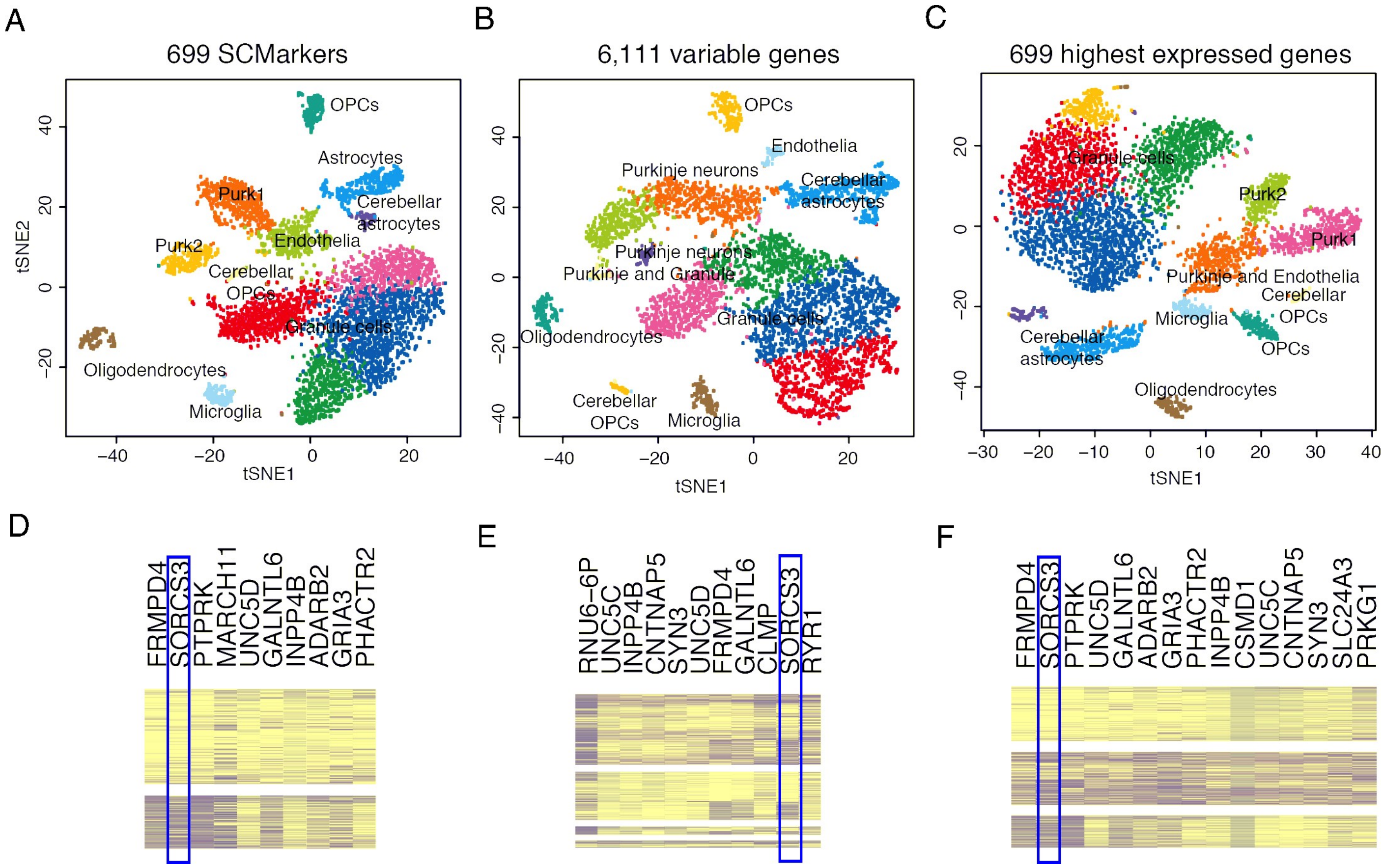

Fig 4. Results on the human brain data. Plotted in tSNE space are 5,602 cells in the cerebellar hemisphere of human brain tissue based on markers selected respectively by (A) SCMarker, (B) the highest variable genes and (C) the highest expressed genes, colored by performing clustering using Seurat. Cell types were labelled consistently as they were in the original paper. Also plotted are the heatmaps of the top 10 gene expression levels derived respectively from (D) SCMarker in cluster 4 and 7, (E) the highest variable genes in cluster 4, 6 11 and 12 and (F) the highest expressed genes in cluster 3, 4 and 6.

https://doi.org/10.1371/journal.pcbi.1007445.g004

T, B and myeloid cells (Table 1, Fig 5A). For example, cluster 4 and 23 are the B cells expressing known surface marker *CD79A* (Fig 5A), yet cells in cluster 4 are evidently different from cells in cluster 23, 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 5F). They appeared non-specifically distributed across clusters (Fig 5E and 5F). 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 (S9 Fig).

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 (S10 Fig).

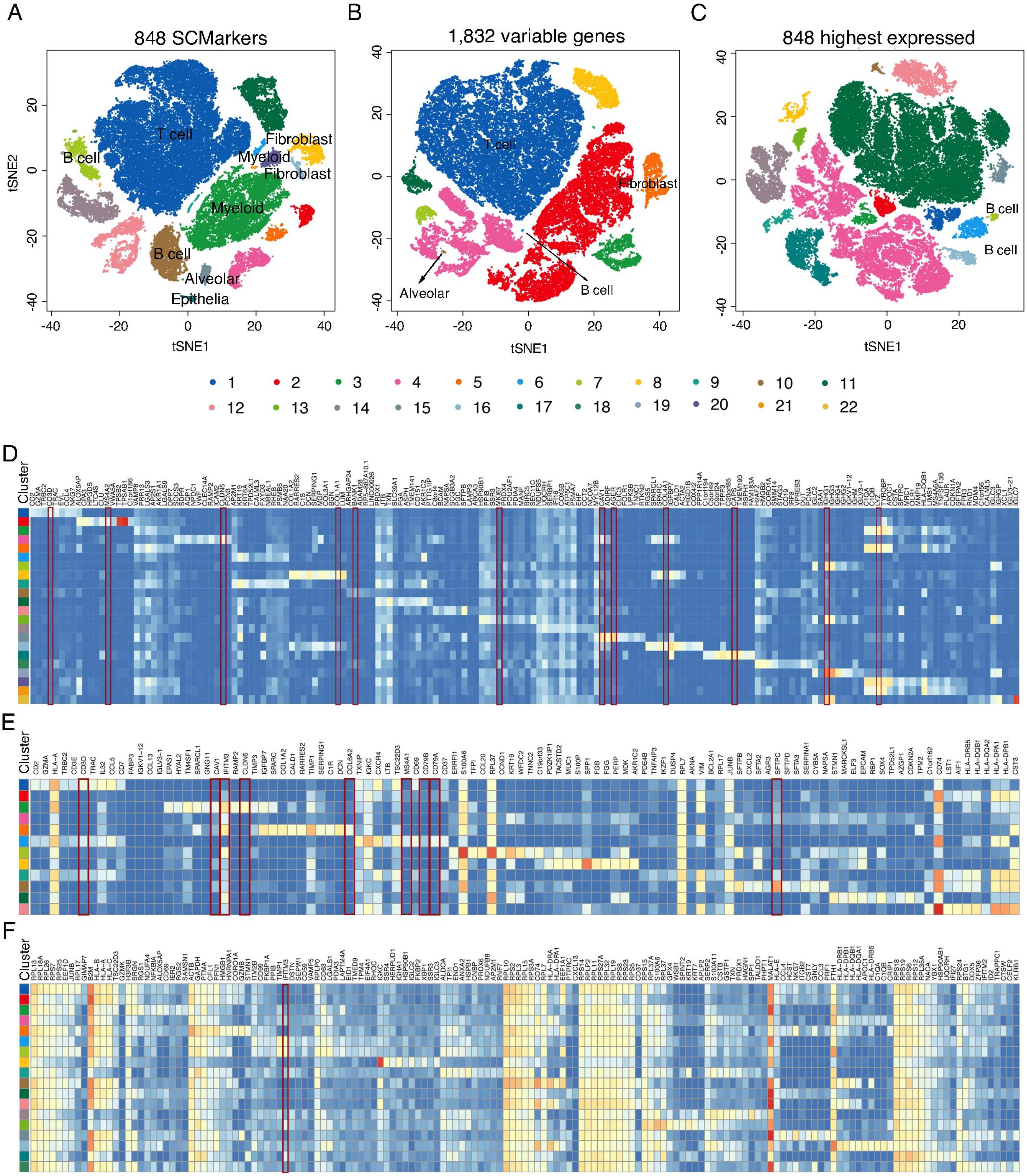
A
848 SCMarkers
T cell
Fibroblast
B cell
Myeloid
Fibroblast
Myeloid
B cell
Alveolar
Epithelia
tSNE2
tSNE1
20
0
-20
-40
-40
-20
0
20
40
B
1,832 variable genes
T cell
Fibroblast
Alveolar
B cell
tSNE1
40
20
0
-20
-40
-40
-20
0
20
40
C
848 highest expressed
B cell
B cell
tSNE1
40
20
0
-20
-40
-40
-20
0
20
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
D
Cluster
CD2 GZMA TRBC2 CD3D TRAC EVL CCL4 NKG7 ALOX5AP CPA3 HPGDS LTC4S CLU MS4A2 VWA5A TPSB2 TPSAB1 C1orf186 VAMP8 PRR13 LGALS3 AP2S1 AKR1A1 LGALS9 DPP7 SOCS3 SDPR AQP1 NPDC1 VWF CLEC14A RAMP2 ICAM2 CLDN5 FCN3 AP2M1 KRT19 KRT6A TPD52L1 CALML3 FXYD3 NBEAL1 RHEB PSMB5 NAA20 COL1A2 RARRES2 C1S DCN SERPING1 MGP COL3A1 BGN COL1A1 LUM ARHGAP24 BANK1 ADAM28 RP5−887A10.1 LINC00926 PRDX1 JTB TXN SLC39A1 FGA ASPH TMEM141 CD151 AKR1C2 PTTG1IP C8orf4 BCAM NAPSA SCGB3A2 PGC SFTPB LAMP3 ABCA3 HSP90B1 PPIB SSR3 MKI67 PRDX4 POU2AF1 PDIA4 MANF BIRC5 SEC11C NDUFS5 UQCRH SERBP1 IFI16 COX5B ATP5C1 PSMA7 EHF CCT2 NCOA7 MYL12B CAV1 ADIRF AGER CLIC3 FOLR1 UPK3B TNNC1 RTKN2 RGS5 SPARCL1 SPARC COL4A1 IGFBP7 CALD1 ACTA2 HIGD1B COX4I2 PPP1R14A C1orf194 C5orf49 C9orf24 TPPP3 C20orf85 CAPS TMEM190 RSPH1 FAM183A H2AFZ HMGB2 CORO1A SMIM14 STAG3 CD19 IRF8 VPREB3 DUT PCNA IGLC2 SAA1 IGHG1 IGHG3 IGHG4 IGHG2 IGKV1−12 IGHM IGLV3−1 C1QA C1QB LYZ TYROBP APOC1 CSTA SFTPC MRC1 OLR1 MMP19 HLA−DQB1 LIMS1 MS4A6A TNFSF13B PLAUR CDKN1A SFTPA2 FPR3 RND1 MDM4 C1orf56 CALML5 IGLC3 IGHGP XCL1 IGLV3−21 IGLC7
E
Cluster
CD2 GZMA HLA−A TRBC2 CD3E CD3D TRAC IL32 CCL5 CD7 FABP3 IGKV1−12 CCL13 IGLV3−1 EPAS1 HYAL2 TM4SF1 SPARCL1 GNG11 CAV1 IFITM3 RAMP2 CLDN5 TIMP3 IGFBP7 SPARC COL1A2 CALD1 RARRES2 TIMP1 SERPING1 C1R DCN COL6A2 TXNIP IGKC CXCR4 LTB TSC22D3 MS4A1 CD69 CD79B CD79A CD37 ERRFI1 S100A6 TFPI CCL20 RPL37 CCND1 KRT19 WFDC2 TNNC2 C19orf33 PDZK1IP1 TACSTD2 MUC1 S100P SPP1 FGB FGG PERP MDK AKR1C2 PDE4B TNFAIP3 IKZF1 DUSP4 RPL7 AKNA VIM BCL2A1 RPL17 JUNB SFTPB CXCL2 SFTA2 AGR3 SFTPC SFTPD SFTA3 SERPINA1 CYB5A NAPSA STMN1 MARCKSL1 ELF3 EPCAM RBP1 SOX4 TPD52L1 AZGP1 CDKN2A TPM2 C1orf162 CD74 LST1 AIF1 HLA−DRB5 HLA−DQB1 HLA−DQA2 HLA−DPA1 HLA−DPB1 CST3
F
Cluster
RPL13 RPL18A RPL26 RPS7 RPS25 EEF1D JUNB RPL17 GIMAP7 B2M HLA−B HLA−A HLA−C TSC22D3 GZMK H3F3B SRGN RGS1 NDUFA4 NFKBIA ALOX5AP CD69 IER2 RGS2 SAMSN1 ACTB GAPDH PTMA CFL1 PFN1 HMGB1 HNRNPA1 CORO1A GZMB STMN1 ITM2B CD99 FKBP1A PPIB TIMP1 IFITM3 DSTN SEPW1 CD59 VAMP5 RPLP0 CD63 LGALS1 PDIA3 LAPTM4A EID1 TMED9 TPM4 EIF4A1 RHOC IGKC SSR4 HERPUD1 IGHA1 HSP90B1 IGLC2 FKBP2 XBP1 SSR3 IGLC3 ALDOA TPI1 ENO1 ANXA2 HSPB1 CNBP PRDX6 NDUFB9 AP2M1 RNF7 RPL10 RPS2 RPL3 RPL15 RPS3A CD74 RPL7 HLA−DRA HLA−DPA1 EEF1A1 PTPRC CXCL13 RPS14 RPLP2 RPL11 RPS27A RPL32 RPL19 RPS8 RPS23 RPS5 CD37 RPS15 RPL37A S100A6 RPL37 GPX4 WSB1 SPINT2 KRT19 KRT7 APLP2 SERF2 S100A11 CSTB GSTP1 TXN PRDX1 HMGN1 SPP1 TALDO1 PHPT1 MALAT1 HLA−E CCL4 HCST NKG7 ITGB2 CST7 GNLY CCL3 PRF1 FTH1 HLA−DRB1 HLA−DPB1 HLA−DQB1 HLA−DQA1 APOC1 HLA−DRB5 C1QA C1QB CRIP1 RPS18 RPS19 RPS6 RPS12 RPL35A NACA YBX1 HSP90AB1 UQCRH IFI27 RPS24 BTG1 DDX5 ZFP36 IFITM2 ID2 TRAPPC1 CTSW CELF2 KLRB1

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 (Table 1), which are highlighted in red boxes in (D) to (F).

https://doi.org/10.1371/journal.pcbi.1007445.g005

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	1
S100A2	Myeloid	8, 9		
IGHG1	B cell	23		
MS4A2	Mast	2		
LYZ	Myeloid	2, 15, 17, 19		3, 4
CLDN5	Endothelial	3	3	
COL1A1	Fibroblast	5		
CD27	T cell	6		
CAPS	Epithelial	14		
CAV1	Alveolar	11	3	
GZMH	T cell	1		
GZMA	T cell	1		11
IFITM3	Myeloid		3	
COL6A2	Fibroblast		5	
MS4A	B cell		6	
CD79A	B cell		6	
SFTPC	Alveolar		10	

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

https://doi.org/10.1371/journal.pcbi.1007445.t001

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 [37]. 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 [38,39]. 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.

Supporting information

S1 Table. Clustering methods.
(DOCX)

S1 Data. Marker gene list.
(XLSX)

S1 Fig. Algorithm for constructing the k-nearest mutually co/anti-occurrence neighbour graph.
(JPG)

S2 Fig. The expression levels of most known marker genes (*CD3G*, *CD8A*, *IL7R*, *MS4A1*, *CD19*, *CD79A*, *CD79B* and *PMEL*) follow a bi/multi-modal distribution in the melanoma data.
(JPG)

S3 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.
(JPG)

S4 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.
(JPG)

S5 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.
(JPG)

S6 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.
(JPG)

S7 Fig. Estimation of *k* and *n* through resampling of the melanoma data. Plotted in heatmap are clustering accuracy measured by the adjusted rand index (ARI). The sample sizes of each dataset were labelled above each of the figures.
(JPG)

S8 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.
(JPG)

S9 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).
(JPG)

S10 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.
(JPG)

Acknowledgments

We thank Miao Qi and Yu-yu Lin for their important suggestions and feedback.

Author Contributions

Methodology: Fang Wang, Shaoheng Liang.

Software: Fang Wang.

Validation: Tapsi Kumar, Nicholas Navin.

Writing – original draft: Fang Wang.

Writing – review & editing: Ken Chen.

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. https://doi.org/10.1016/j.molcel.2015.04.005 PMID: 26000846
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. https://doi.org/10.1016/j.cell.2015.05.002 PMID: 26000488
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. https://doi.org/10.1038/nbt.2282 PMID: 22820318
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. https://doi.org/10.1186/s12859-017-1977-1 PMID: 29322913
6. Trapnell C (2015) Defining cell types and states with single-cell genomics. Genome Research 25: 1491–1498. https://doi.org/10.1101/gr.190595.115 PMID: 26430159
7. Wagner A, Regev A, Yosef N (2016) Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol 34: 1145–1160. https://doi.org/10.1038/nbt.3711 PMID: 27824854
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. https://doi.org/10.1093/hmg/ddy100 PMID: 29590361
9. Pierson E, Yau C (2015) ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol 16: 241. https://doi.org/10.1186/s13059-015-0805-z PMID: 26527291
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. https://doi.org/10.1186/s13059-017-1218-y PMID: 28482897
11. Chen L, Zheng S (2018) BCseq: accurate single cell RNA-seq quantification with bias correction. Nucleic Acids Res 46: e82. https://doi.org/10.1093/nar/gky308 PMID: 29718338
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. https://doi.org/10.1038/nbt.3192 PMID: 25867923

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. https://doi.org/10.1038/nmeth.4236 PMID: 28346451

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. https://doi.org/10.1371/journal.pcbi.1004575 PMID: 26600239

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. https://doi.org/10.1073/pnas.95.25.14863 PMID: 9843981

18. Quackenbush J (2001) Computational analysis of microarray data. Nature Reviews Genetics 2: 418–427. https://doi.org/10.1038/35076576 PMID: 11389458

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. https://doi.org/10.1093/bioinformatics/17.10.977 PMID: 11673243

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. https://doi.org/10.1038/nprot.2014.006 PMID: 24385147

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–+. https://doi.org/10.1016/j.molcel.2017.01.023 PMID: 28212749

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. https://doi.org/10.1038/s41586-018-0393-7 PMID: 30069044

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. https://doi.org/10.1038/s41586-018-0394-6 PMID: 30069046

25. Welch JD, Hartemink AJ, Prins JF (2017) MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol 18: 138. https://doi.org/10.1186/s13059-017-1269-0 PMID: 28738873

26. Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability. London: Chapman and Hall.

27. 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. https://doi.org/10.1016/j.cell.2017.10.044 PMID: 29198524

28. 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. https://doi.org/10.1126/science.aad0501 PMID: 27124452

29. 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. https://doi.org/10.1038/nbt.4038 PMID: 29227469

30. 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 24: 1277–1289. https://doi.org/10.1038/s41591-018-0096-5 PMID: 29988129

31. Bartlett TE, Muller S, Diaz A (2017) Single-cell Co-expression Subnetwork Analysis. Sci Rep 7: 15066. https://doi.org/10.1038/s41598-017-15525-z PMID: 29118406

32. Hwang B, Lee JH, Bang D (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 50: 96. https://doi.org/10.1038/s12276-018-0071-8 PMID: 30089861

33. 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.

34. Blashfield RK (1991) Finding Groups in Data—an Introduction to Cluster-Analysis—Kaufman,L, Rousseeuw,Pj. Journal of Classification 8: 277–279.

35. 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.

36. 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. https://doi.org/10.1073/pnas.0506580102 PMID: 16199517

37. Ye X, Weinberg RA (2015) Epithelial-Mesenchymal Plasticity: A Central Regulator of Cancer Progression. Trends Cell Biol 25: 675–686. https://doi.org/10.1016/j.tcb.2015.07.012 PMID: 26437589

38. 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. https://doi.org/10.1038/nature14590 PMID: 26083756

39. 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. https://doi.org/10.1038/nbt.1991 PMID: 21964415