bioRxiv preprint doi: https://doi.org/10.1101/356634; this version posted July 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SCMarker: ab initio marker selection for single cell transcriptome profiling

Fang Wang[1,3], Tapsi Seth[2], Shaoheng Liang[1], Nicholas Navin[1,2], Ken Chen[1,3]

1. Department of Bioinformatics and Computational Biology

2. Department of Genetics

3. Corresponding: Ken Chen (kchen3@mdanderson.org)

The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Abstract

Current single-cell RNA-sequencing (scRNA-seq) data generated by a variety of technologies such as DropSeq and SMART-seq can reveal simultaneously the mRNA transcript levels of thousands of genes in thousands of cells. Cell subpopulations (e.g., cell-types) that have similar transcriptomes can be further delineated in the high dimensional gene expression space. However, genes are not equally informative in delineating cell subpopulations. Therefore, it is often important to select informative genes or subpopulation-informative markers (SIMs) to reduce dimensionality and achieve informative clustering. Here, we present an ab initio method that performs unsupervised marker selection, based on two novel metrics 1) discriminative power of individual gene expressions and 2) mutually coexpressed gene pairs (MCGPs). Consistent improvement in cell-type classification and biologically meaningful marker selection are achieved when applying SCMarker on data generated by scRNA-seq datasets, including UMI data by the 10X Chrimium and TPM data by SMART-seq2, from various tissue types (melanoma, brain, etc.), followed by a variety of clustering algorithms such as k-means, shared nearest neighbor (SNN), etc. The R package of SCMarker is publicly available at https://github.com/KChen-lab/SCMarker.

Introduction

Current single-cell RNA-sequencing (scRNA-seq) data generated by a variety of technologies such as DropSeq and SMART-seq can reveal simultaneously the mRNA transcript levels of thousands of genes in thousands of cells[1-3]. Cell subpopulations that have similar transcriptomes can be further delineated in the resulting high dimensional gene-expression space. However, genes are not equality informative to delineate the cell subpopulations[4,5]. For example, genes whose expression levels cannot be robustly measured (e.g., zero inflated[6]) or have low variability across cells are often eliminated before any further analysis. It has become a common practice to keep only highly expressed or highly variable genes for further analysis [7]. Here, we propose an ab initio method (SCMarker) that further determines subsets of

bioRxiv preprint doi: https://doi.org/10.1101/356634; this version posted July 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

genes informative to subpopulation clustering, without referencing to any known transcriptomic profiles. Our approach revealed that by applying information-theoretic principles, we can simultaneously reveal cell-types and cell-type specific biology from existing scRNA-seq data, with higher accuracy and resolution than existing techniques.

Methods

SCMarker consists of two steps.

1. Filtering based on discriminative power of individual gene expressions

Subpopulation informative markers (SIMs) should have distinctive expression levels across cell subpopulations. Therefore, in a dataset with mixed cell subpopulations, the expression level of a SIM should follow a bi- or multi-modal, instead of a unimodal distribution (**Figure 1a and b**). Based on this assumption, we quantified the degree of modality based on the probability density distribution (*f*) of each gene expression using a Gaussian kernel function through formula (1).

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

Where g_{ij} is the expression level of gene i in cell j. $h > 0$ is a smoothing parameter called bandwidth, which was set at 2 here. $K_h(x)$ is a scaled kernel defined as $\frac{1}{h}K(x/h)$, where $K(x)$ is a normal kernel $K(x) = \varphi(x)$ and $\varphi(\cdot)$ is the standard normal density function. For each gene, we counted the number of local maximum probability density based on estimated probability density function. A gene expression level follows a multi-modal distribution if it has multiple local maximum probability density values. We filtered out genes whose levels follow unimodal probability density distribution with only one local maximum probability density value.

2. Co-expression of genes in a cell subpopulation

SIMs should co-express with other SIMs in cells of identical types. To identify co-expressed gene pairs, we discretize a gene-cell expression matrix into a n $\times$ m binary matrix $X \in \{0,1\}^{n \times m}$, with a 1 in each entry x_{ij} designating an expressed gene i in cell j and a 0 otherwise (**Figure 1c**). For gene i, $x_{i\cdot} = (x_{i1}, x_{i2}, \ldots, x_{im})$ is a binary string. We can calculate a co-occurrence matrix (S) that measures the pair-wise cooccurrence between all the gene pairs,

$$S = X \cdot X'. \quad (2)$$

We can further represent S as a directed k-nearest co-expressed neighbor (KNCEN) graph (**Figure 1d**), in which a node represents a gene and an edge from gene A to gene B representing that B is A's KNCEN.

bioRxiv preprint doi: https://doi.org/10.1101/356634; this version posted July 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The KNCENs of a gene A are the *k* genes that co-occurred with A in the *k* highest number of cells. We further required that a KNCEN must cooccur with the host gene in at least *n* cells. This is necessary to remove spurious co-occurrence.

Under these definitions, a mutually co-expressed gene pair (MCGP) refers to two genes bidirectionally connected in the KNCEN graph and are each other's KNCEN. Only genes included in at least one MCGP are selected as Markers.

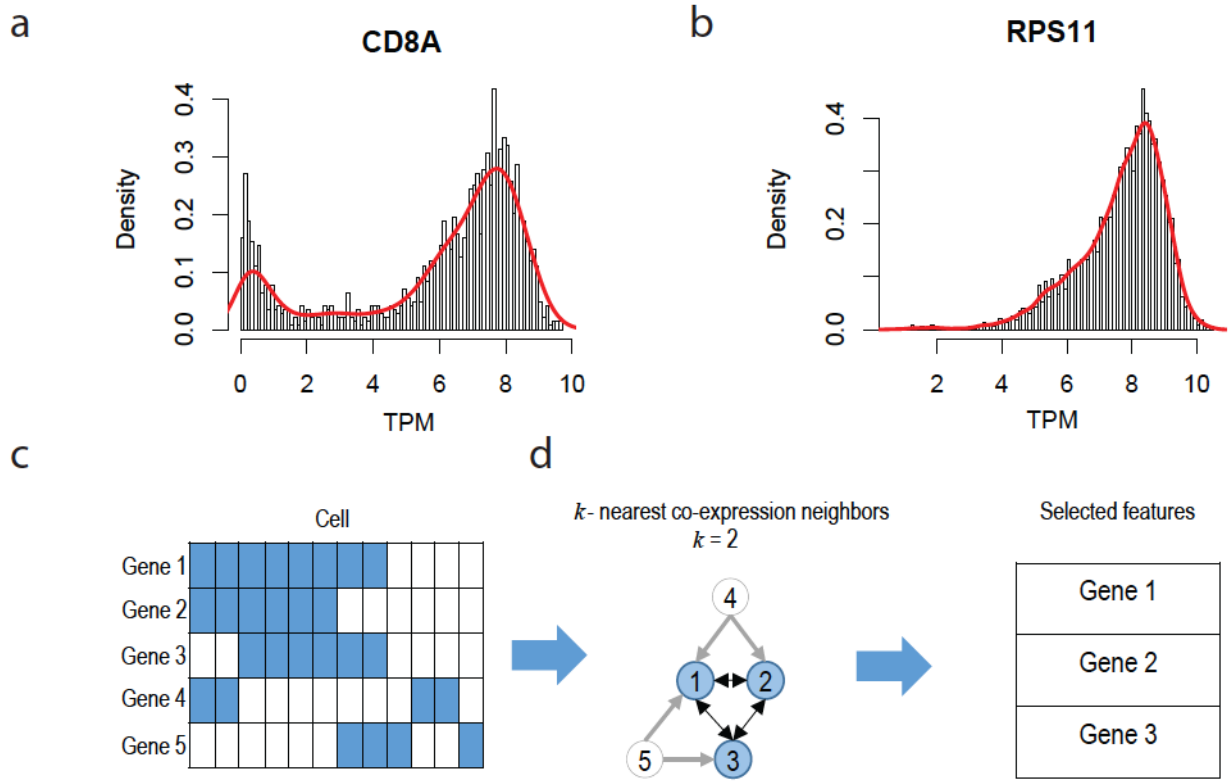


Figure 1. Illustration of SCMarker. Plotted are a bimodally distributed gene (**a**) and unmodally distributed gene (**b**). From a binarized gene-cell matrix (**c**), a *k*-nearest co-expression neighbor graph can be constructed and mutually coexpressed gene pairs (MCGP) (node 1, 2 and 3) can be identified (**d**). Marker genes are subsequently selected based on the above metrics.

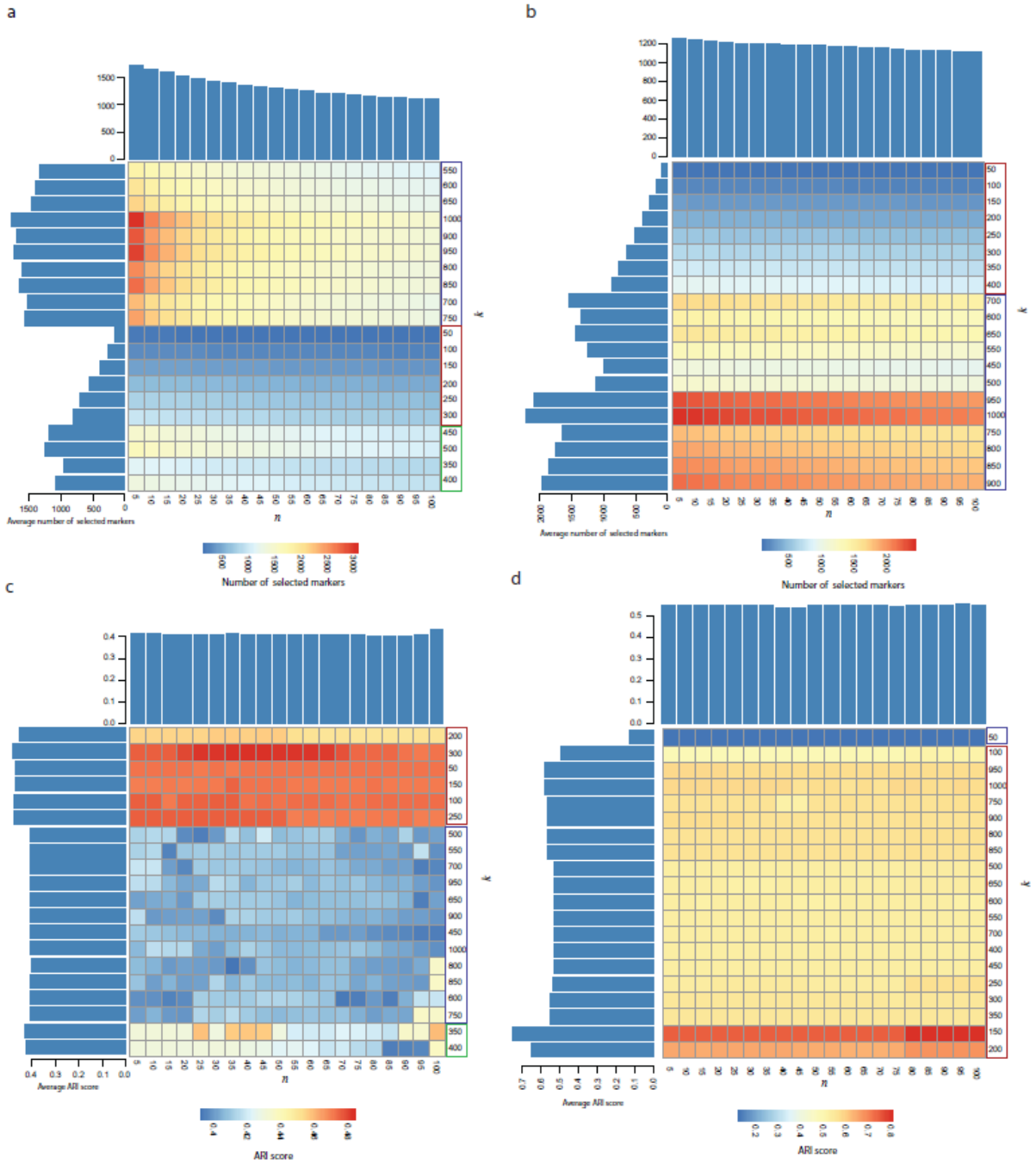
Results

1. Effects of parameterization

We applied SCMarker (https://github.com/KChen-lab/SCMarker) to the scRNA-seq data obtained from 1) a melanoma patient sample, which includes 4,605 cells[8] and 2) a head and neck cancer patient sample, which includes 5,902 cells[9] sequenced by the SMART-seq platform. Each cell in the sets is labeled with a unique cell-type determined by orthogonal technologies.

bioRxiv preprint doi: https://doi.org/10.1101/356634; this version posted July 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We first tested the two parameters required by SCMarker: number of the nearest neighbors (*k*) and the minimal number of co-expressed cells (*n*). We found that choice of *k* appeared to have a large effect on the number of selected markers, whereas choice of *n* did not (**Figure 2a and b**). We further evaluated the accuracy of clustering results by comparing cell types identified from the data with the known cell-type labels determined by the original studies. We found that setting *k* between 100 and 300 resulted in the most accurate cell type identification results, indicated by high adjusted rand index (ARI) scores [10] (**Figure 2c and d**) and stable numbers of markers (**Figure 2a and b**).



bioRxiv preprint doi: https://doi.org/10.1101/356634; this version posted July 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2. Determining the optimal parameters. Number of selected markers given a range of *n* and *k* parameters for the melanoma (**a**) and the head and neck cancer (**b**) data. Clustering accuracy measured by ARI scores, given various *n* and *k* for the melanoma (**c**) and the head and neck cancer (**d**) data.

2. Comparison with other marker selection strategies

Based on the analysis above, we selected $k = 300$ and $n = 30$ for SCMarker analysis and obtained 891 and 643 genes as markers for the melanoma and the head and neck cancer datasets, respectively. For comparison, we also selected the same number of A) most highly expressed and B) most variable genes. Four clustering methods including k-means, Clara[11], hierarchical clustering and Seurat[7] were used to cluster single cells based on selected markers. The adjusted rand index (ARI) was used to measure the consistence between clustering results and cell labels[10]. Compared to marker sets A and B, the marker set selected by SCMarker have substantially higher degrees of overlap with the genes used to generate the cell-type labels in the original publications (**Figure 3**) and resulted in the highest ARI with the fairly evident differences (**Figure 4**).

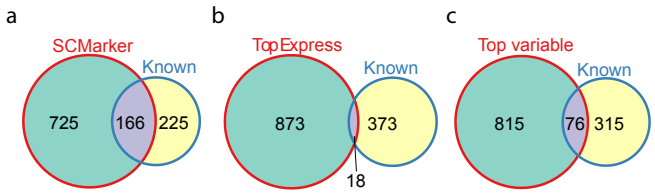
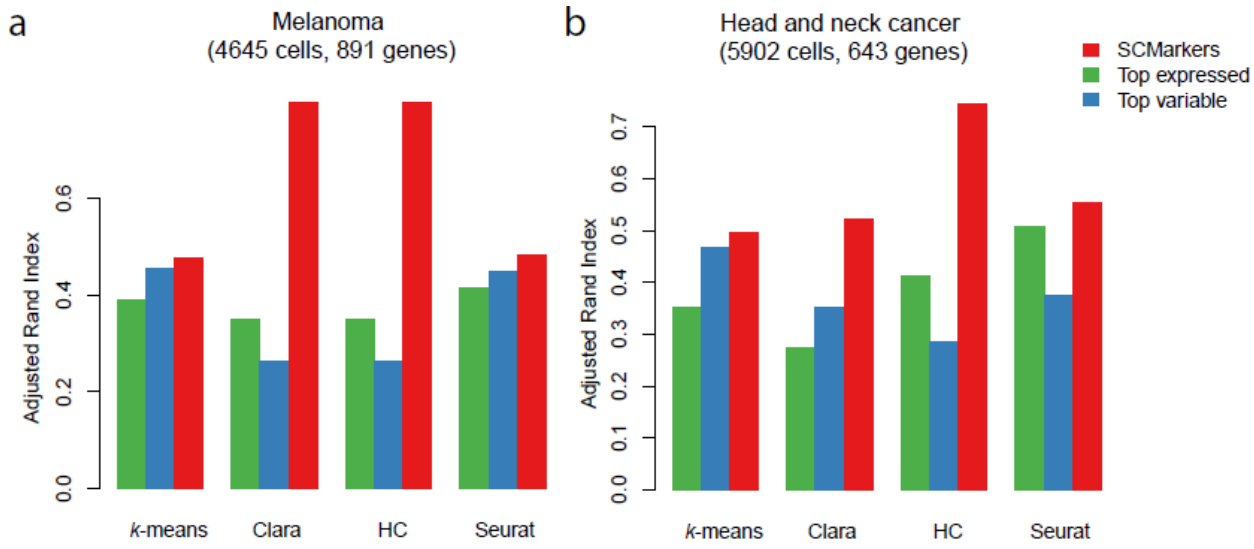


Figure 3. Venn diagrams between marker sets determined by SCMarker (**a**), top expressed (**b**) and top variable genes (**c**) with the known cell-type markers in the melanoma data[8].



bioRxiv preprint doi: https://doi.org/10.1101/356634; this version posted July 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 4. 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 by SCMarker, top expressed and top variable genes, using two scRNA-seq datasets from a melanoma (a) and a head-and-neck (b) cancer samples by 4 clustering algorithms: k-means, Clara, hierarchical clustering (HC) and Seurat.

To obtain unbiased assessment, we further compare the 3 marker-selection methods using a range of *k* and *n* parameters and Seurat as the classifier (**Figure 5**). The results clearly indicated superior, robust classification accuracy achieved by SCMarker, compared with the top expressed and top variable approaches, when identical numbers of markers were selected.

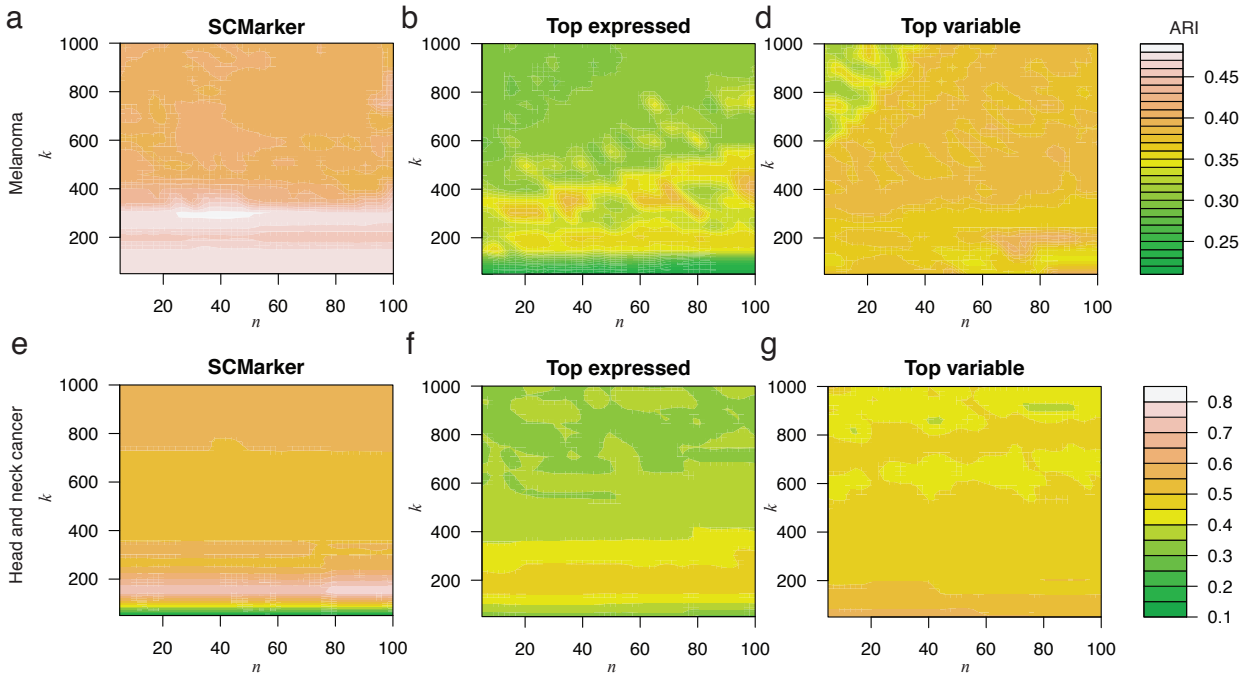


Figure 5. Comparison of 3 marker selection methods for cell-type identification across ranges of parameters. Plotted as heatmaps are ARI values using markers selected by SCMarker (**a** and **e**), top expressed genes (**b** and **f**) and top variable genes (**d** and **g**) and Seurat as the classifier from the melanoma (top panel) and the head and neck cancer data (bottom panel). X and Y axes in (**a** and **e**) indicate the *n* and *k* parameters used by SCMarker and in (**b,d,f,g**) indicate identical numbers of markers were selected by the other marker-selection methods.

3. Functional annotations of selected Markers

Compared with other methods, SCMarker selected significantly more immune cell surface markers specific to T cytotoxic cells, T helper cells, B lymphocyte cell and macrophage that are likely present in the tumor microenvironment through gene set enrichment analysis[12]. In addition, cancer pathways VEGF,

bioRxiv preprint doi: https://doi.org/10.1101/356634; this version posted July 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

angiogenesis, tumor suppressor signaling, cell cycle (particularly G1/S check point) were only enriched in markers selected by SCMarker but not in those selected by the other methods (**Figure 6**).

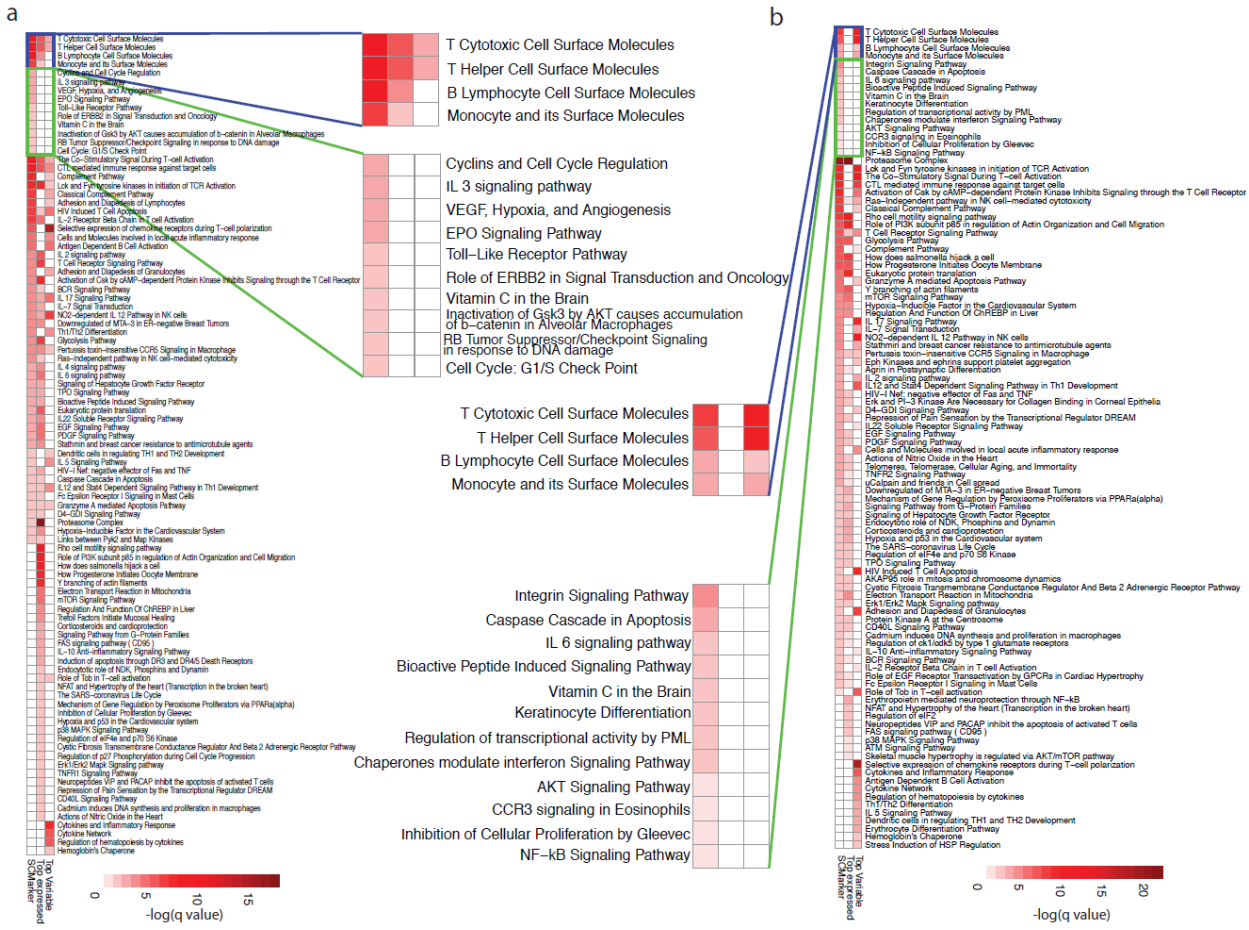


Figure 6. Gene set enrichment analysis (GSEA) of markers selected based on 3 methods: SCMaker, top expressed and top variable genes, from the melanoma (a) and the head and neck cancer (b) data. Darkness of the colors correspond to -log10 P values.

4. Application of SCMarker to brain data

We applied SCMarker to an independent dataset that includes 5,204 cells from the cerebellar hemisphere of 6 different postmortem adult human brains based on Drop-seq platform[13]. We clustered cells using Seurat[14] with default parameters and markers selected by SCMarker. The tSNE[15] plots generated based on SCMarker genes (**Figure 7**) showed more clear separation than those based on top variable genes selected by Seurat in default mode. Moreover, purkinje neurons (Purk1) and non-neurons (Purk2) cells were successfully clustered into two sub-groups based on markers selected by SCMarker. In addition, cerebellar-specific astrocytes (Ast_Cer) were successfully separated from astrocytes (Ast).

bioRxiv preprint doi: https://doi.org/10.1101/356634; this version posted July 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

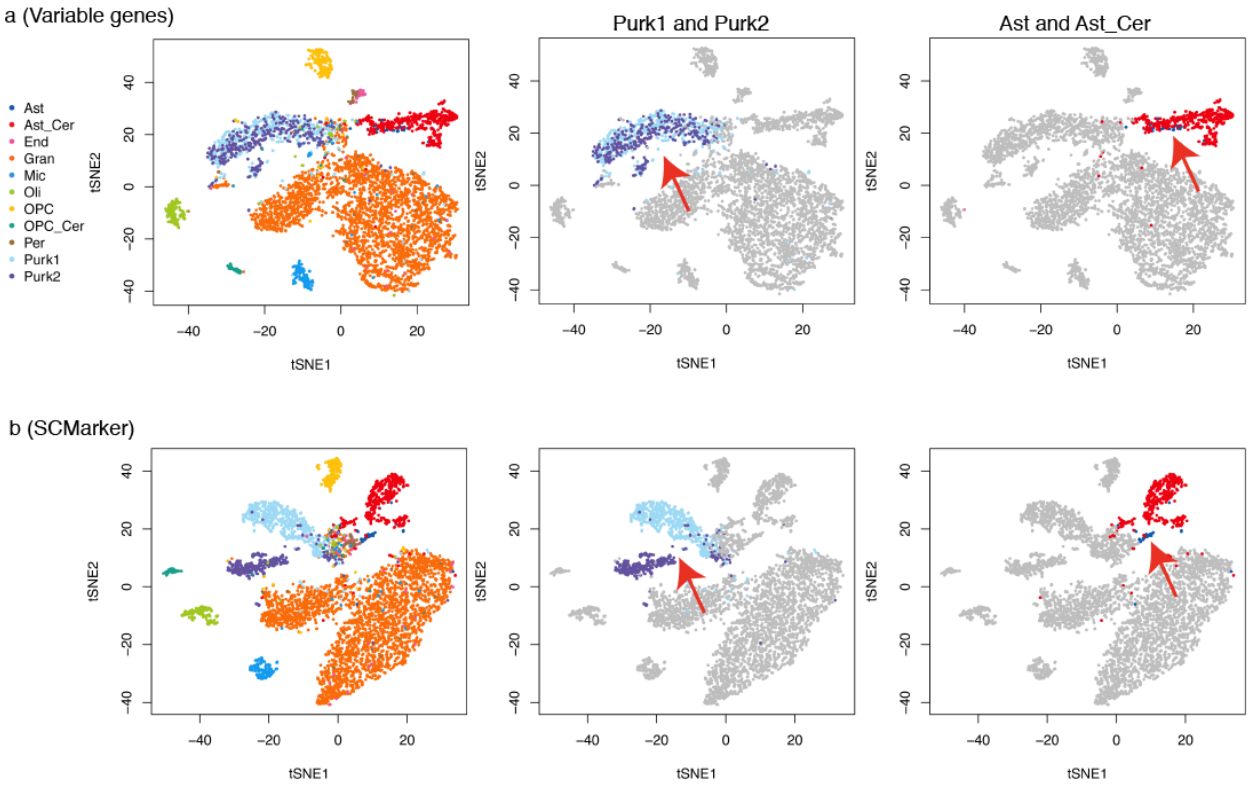


Figure 7. tSNE plots of 5,204 cells in the cerebellar hemisphere of 6 different postmortem adult human brains based on markers selected by **a**) top variably expressed genes and **b**) SCMarker into 11 cell types (left panel). Notice that Purk1 and Purk2 cells (middle panel) are mixed in **a**) but clearly separated in **b**), highlighted by red arrows. So are cerebellar-specific astrocytes (Ast_Cer) and astrocytes (Ast) (right panel).

Conclusions

In this manuscript, we reported a new marker selection strategy SCMarker that performs ab initio marker selection from scRNA-seq data. Using information-theoretic approaches without any biological priors, SCMarker selects markers by scrutinizing two subpopulation discriminative features: 1) bi/mul-modal distribution of subpopulation-informative gene expression in mixed cell population and 2) level of co-expression among subpopulation-specific gene pairs. We found that SCMarker can consistently significantly boost cell-type identification accuracy in several cancer and brain scRNA-seq datasets. Because SCMarker does not depend on any prior knowledge, we anticipate that it will prove most useful in analyzing cancer cells of a high degree of plasticity and heterogeneity in transcriptomic profiles[16].

SCMarker can be easily incorporated as a preprocessing module into current scRNA-seq data analysis workflow to preprocess cell-gene count/expression matrix before performing further downstream analysis. The source code of SCMarker is publicly available at https://github.com/KChen-lab/SCMarker.

bioRxiv preprint doi: https://doi.org/10.1101/356634; this version posted July 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

This work was supported by a Chan-Zuckerberg Initiative award to Ken Chen.

References

1. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. & Teichmann, S.A. The technology and biology of single-cell RNA sequencing. *Mol Cell* **58**, 610-20 (2015).
2. Macosko, E.Z. *et al*. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
3. Ramskold, D. *et al*. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**, 777-82 (2012).
4. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Research* **25**, 1491-1498 (2015).
5. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol* **34**, 1145-1160 (2016).
6. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol* **16**, 241 (2015).
7. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).
8. Tirosh, I. *et al*. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-96 (2016).
9. Puram, S.V. *et al*. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624 e24 (2017).
10. Santos J.M., E.M. 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**(2009).
11. Blashfield, R.K. Finding Groups in Data - an Introduction to Cluster-Analysis - Kaufman,L, Rousseeuw,Pj. *Journal of Classification* **8**, 277-279 (1991).
12. Subramanian, A. *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
13. Lake, B.B. *et al*. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80 (2018).
14. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
15. van der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research* **15**, 3221-3245 (2014).
16. Ye, X. & Weinberg, R.A. Epithelial-Mesenchymal Plasticity: A Central Regulator of Cancer Progression. *Trends Cell Biol* **25**, 675-86 (2015).