COMET: A tool for predicting multiple gene-marker panels from single-cell transcriptomic data

Conor Delaney[1*], **Alexandra Schnell**[2*], **Louis Cammarata**[3*],
Aaron Yao-Smith[4] , **Aviv Regev**[5,6] , **Vijay Kuchroo**[2,6], **and Meromit Singer**[1,6,7,†]
[1]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[2]Evergrande Center for Immunologic Diseases and Ann Romney Center for Neurologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, MA 02215, USA
[3]Department of Statistics, Harvard University, Cambridge, MA 02138, USA
[4]Department of Computer Science, Cornell University, Ithaca, NY 14850, USA
[5]Howard Hughes Medical Institute, Department of Biology and Koch Institute of Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02138, USA
[6]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[7]Department of Immunology, Harvard Medical School, Boston, MA 02115, USA
[†] Corresponding Author: msinger@jimmy.harvard.edu

Abstract

Single-cell transcriptomic studies are identifying novel cell populations with exciting functional roles in various in vivo contexts, but identification of succinct gene-marker panels for such populations remains a challenge. In this work we introduce COMET, a computational framework for the identification of candidate marker panels consisting of one or more genes for cell populations of interest identified with single-cell RNA-seq data. We show that COMET outperforms other methods for the identification of single-gene panels, and enables, for the first time, prediction of multi-gene marker panels ranked by relevance. Staining by flow-cytometry assay confirmed the accuracy of COMET's predictions in identifying marker-panels for cellular subtypes, at both the single- and multi-gene levels, validating COMET's applicability and accuracy in predicting favorable marker-panels from transcriptomic input. COMET is a general non-parametric statistical framework and can be used as-is on various high-throughput datasets in addition to single-cell RNA-sequencing data. COMET is available for use via a web interface or a standalone software package.

1 Motivation

Single-cell transcriptomic studies have enabled the exciting discovery of novel cell populations within various in vivo contexts [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Following the discovery of a new cell population based on full transcriptome analysis, follow-up studies require succinct gene-marker panels by which the cells of interest can be distinguished from the general cell population. Current techniques used in the literature for the identification of candidate marker-panels are substantially limited because they rely on statistical tests designed for other purposes (such as gene differential expression), do not consider gene combinations, and require extensive manual curation.

A broadly used technique for candidate marker-panel annotation from single-cell RNA-seq data consists in generating a ranked list of genes based on their upregulation in the cluster of choice and/or expression fold-change estimates [1, 4, 5, 13, 6, 7, 8, 11]. Extensive manual curation is then required to evaluate genes at the top of the list for their ability to provide good classifiers and for their ability to pair with each other to construct multiple-gene marker panels. Substantial limitations in the use of

such techniques is that they do not directly test for a gene's ability to isolate a given cell population from a background, and importantly, the genes constructing a successful multi-gene panel may not be favorable as single-gene markers.

Development of computational tools that provide useful guidance to researchers is difficult due to the scale and hardness of the algorithmic problem and limited availability of experimental reagents (e.g. antibodies for flow or in situ staining, probes for FISH). The latter requires that a marker panel prediction framework be broad by suggesting multiple (ranked) candidate marker-panels to the user, to be assessed for reagent availability and accuracy. Nonetheless, the need within the community to transition from observations at the single-cell level to functional studies calls for the development of a computational framework that can generate an informative ranking of candidate multi-gene marker panels.

2 COMET Tool

In this work we introduce COMET (COmbinatorial Marker dEtection from single-cell Transcriptomics), a computational framework that detects candidate marker panels consisting of one or more genes for cell populations of interest identified with high-throughput single-cell data. COMET takes as input a gene-by-cell expression matrix with cluster assignment for each cell and outputs a separate directory for each cluster that includes ranked lists of candidate marker panels along with informative statistics and visualizations.

COMET implements the XL-minimal Hypergeometric test (XL-mHG test) [14, 15] to binarize gene expression data in a gene-specific and cluster-specific manner. For each gene G and cluster K, all cells in the data set are sorted in decreasing order of the expression of G. The test selects a cutoff index that maximizes the hypergeometric enrichment of cells in K (with respect to cells in the complement of K, which we denote by C) at the top of the list. The chosen cutoff index translates into an expression threshold which is used to binarize gene expression data. X and L are parameters that can be used to control the minimum number of true-positives (X) and the maximum number of false-positives ($L - X$).

COMET outputs a ranked list of candidate single-gene markers (by integrating the XL-mHG p-values and the log-fold-change of gene expression) and provides the true-positive (TP) and true-negative (TN) rates for each marker candidate. Genes are also tested for their potential to be used as negative markers. COMET also leverages the binaried expression data to construct multiple-gene marker panels via logical operations. A ranking of candidate multi-gene panels is done based on enrichment of cells expressing the entire gene-panel in the cell cluster of choice (hypergeometric enrichment p-value) combined with a "Cluster-Clear Score" (CCS)

$$CCS = \sum_{C \in \mathcal{C} \setminus K} (TN_C^{after} - TN_C^{before})$$

where TN_C^{after} is the true negative percent in cluster C for the single gene in the panel with the lowest p-value when considered as a single-gene marker and TN_C^{before} is the true negative percent in cluster C for the panel (after addition of the remaining genes in the panel). COMET outputs a ranked list of candidate marker panels for each marker panel size (of 2-4 genes) along the true-positive (TP) and true-negative (TN) rates the given combination would achieve. TP and TN rates are efficiently computed using matrix multiplications on the binarized expression matrices.

3 Statistical Properties

The XL-mHG test enjoys desirable properties for the purpose of marker detection, as shown by Monte Carlo simulations using Gaussian synthetic expression data for one gene in many cells (Figure 1A,B). COMET was compared to several gene Differential Expression (DE) tests frequently used to identify single-gene marker panels [16, 17, 7]. Common gene DE tests included in the comparison are Welch's t-test, the Wilcoxon Rank-Sum test, the Kolmogorov-Smirnov test and the Likelihood Ratio test on a logistic regression model where cell cluster ($C_i = 1$ if cell i belongs to the cluster of interest, 0 otherwise) is regressed against an intercept only or both an intercept and the expression value of the gene (X_i) in that cell

$$C_i|X_i \sim Bernoulli(\sigma(\beta_0 + \beta_1 X_i)) \quad \text{vs.} \quad C_i|X_i \sim Bernoulli(\sigma(\beta_0))$$

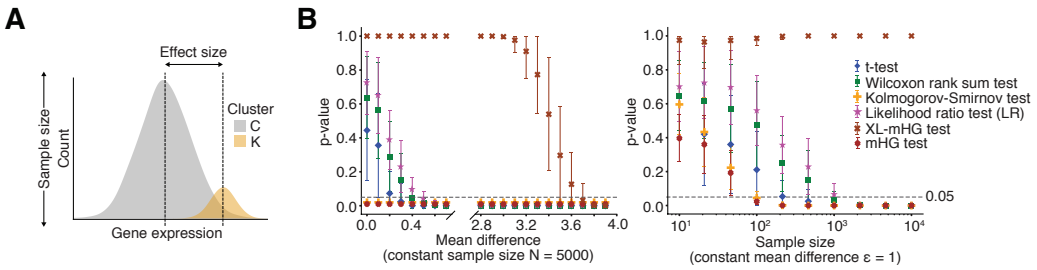

*Figure 1: COMET accurately computes single-gene markers for cell populations. The XL-mHG test outperforms various differential expression tests in identifying favorable marker genes from simulated datasets (**A**), with respect to both robustness to small effect-sizes (mean difference between the cluster of interest K and the cluster of all remaining cells C) (**B**, left) and sensitivity to sample size (**B**, right). Error bars indicate one standard deviation across 100 simulation runs.*

where $\sigma(\cdot)$ is the logistic function. Simulations showed that the COMET procedure detects good markers and discards poor markers regardless of sample size, contrary to other tests whose power increases rapidly with sample size (Figure 1B). The X and L parameters of the XL-mHG test play an important role in this favorable behavior.

The binarization of gene expression implemented in COMET can be related to a classification task. To assess COMET's performance compared to standard classifiers [17, 18], we performed simulations on cell-by-gene count matrices using a noisy Poisson-Gamma generative model for gene counts data (Figure 2A) which replicates both technical noise and efficiency noise in scRNA-seq [19, 20]. Synthetic expression data was generated for two cell clusters (the cluster of interest K and a 'background' cluster C) and many genes pertaining to three categories: good markers (s genes $G_1, ..., G_s$ which separate well the two clusters), poor markers (e.g. markers of cell sub-clusters, measurement outliers) and non-markers (genes with similar expression across both clusters).

We used each of XL-mHG test, Logistic Regression (LR), Random Forest (RF) and Extremely Randomized Trees (XT) to construct a ranking of potential markers, and compared the methods' rankings to the optimal ranking (known from the simulation engine) using the Scaled Sum of Ranks (SSR) metric. We defined SSR to determine the extent to which the good markers are ranked at the top of the list

$$SSR(M) = \frac{2}{s(s+1)} \sum_{j=1}^{s} rank(G_j|M)$$

where M refers to the method used to rank the genes (XL-mHG p-value, RF and XT Gini importance metric or LR p-value). An SSR score of 1 reflects a ranking in which all good markers are ranked at the top of the list, in higher places than any of the poor markers and the non-markers. Generally, for the SSR score lower is better. We compared the SSR scores across the LR, RF and XL-mHG classification methods and observed that poor markers had a detrimental effect on the identification of good single-gene markers by LR and RF, while the XL-mHG test was robust to the quantity and expression rates of poor markers in the data (Figure 2B,C). The X and L parameters play an important role in protecting COMET against the selection of genes which constitute poor markers for the cluster of interest yet enjoy a strong predictive power (such as sub-cluster markers).

4 Experimental Validation

To assess COMET's ability to identify novel surface single-gene markers from real data we evaluated COMET's prediction of cell surface markers for splenic B cell populations using data from the Mouse Cell Atlas [21]. We compared the rankings of known single-gene markers obtained by COMET to rankings obtained with other differential expression tests (Welch t-test, Wilcoxon Rank-Sum test, Likelihood Ratio test, MAST hurdled t-test). COMET performs well in identifying known single-gene markers for the different cell populations identified in the spleen, and performs slightly higher or comparable to other methods. Flow-cytometry based assay revealed that the additional

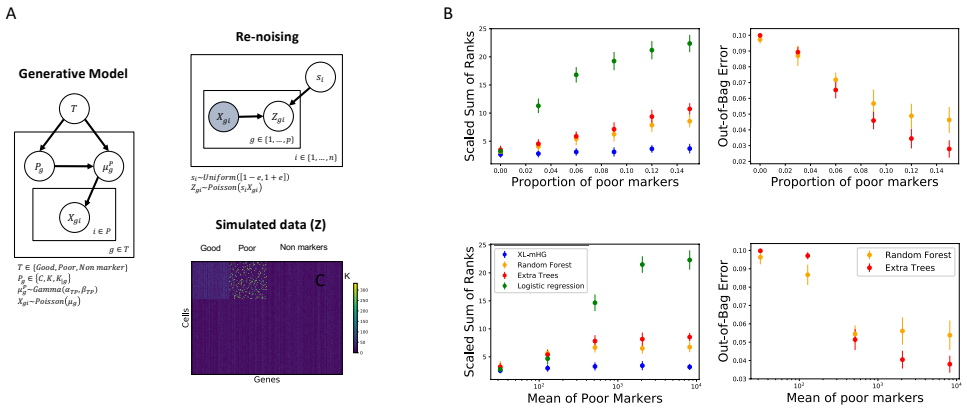

Figure 2: The XL-mHG test outperforms standard classifiers for single-gene marker recovery on simulated gene counts data. ***A****, a noisy Poisson-Gamma model (left and top right) is utilized to generate a cell-by-gene matrix of true counts. Technical and efficiency noises are introduced using an efficiency scaling factor followed by Poisson resampling (top right). This procedure produces gene count matrices of the type shown on the bottom right.* ***B****, SSR versus proportion of poor markers in the data set (top left). The XL-mHG picks up the correct good markers regardless of the proportion of poor markers, while this proportion affects both LR, RF and XT. Out-of-bag error (OOB error) is included for RF and XT (top right). We also display the SSR versus mean of poor markers in the data set (bottom left). The XL-mHG test picks up the correct good markers regardless of the mean of poor markers. Poor markers with very high expression are valuable for RF and XT, and contribute to increase the fold change between the cluster of interest K and the background C, resulting in suboptimal performances for LR. Out-of-bag error (OOB error) is included for RF and XT (bottom right). Error bars indicate one standard deviation across 20 simulation runs.*

top-ranking candidates Ly-6D and CD79b co-stain well with the well-known B cell marker CD19 [22] and showed limited co-staining with known T cell marker CD3 [23], showing their specificity as B cell markers. This confirms the accuracy of COMET's predictions for single-gene marker panels. Extended results and methods are available in the published COMET manuscript [24].

We envision a primary use for COMET in the identification of candidate marker-panels for subpopulations of a given cell type. Isolating cell subtypes requires identifying multiple-gene marker panels as single-gene markers may not be sufficient to accurately sort the cells. We therefore tested COMET's ability to detect marker combinations for the follicular B cell subpopulation using the Tabula Muris dataset [25]. COMET predicted the combination (CD62L+CD44-) for the isolation of follicular B cells. We observed that flow cytometry of CD62L+CD44- cells yields a significantly cleaner population of follicular B cells (defined as CD23-positive) than CD62L alone. We also sorted cells based on the highly ranked combination CD55+CD62L+ and observed an improvement compared to using either CD62L+ or CD55+ alone. Importantly, the combinations for validation were selected by their COMET ranking as well as by antibody availability. The combinations assessed ranked 22 (CD62L+CD44-) and 38 (CD62L+CD55+). More details and methods on experimental validation can be found in the published COMET manuscript [24].

5 Discussion

The fast-increasing number of single-cell RNA-seq datasets being generated and analyzed is revealing novel cell types and subtypes in a variety of systems. A main contribution of the COMET tool is the introduction of a principled framework for identifying multi-gene combinations that constitute favorable marker panels for cell clusters of interest. Along with its broad applicability to single-cell transcriptomic data, the COMET framework can be utilized for other instances by merely changing the input to the available software. We anticipate that the use of COMET will propel the transition from novel characterization-focused observations (made via methods such as single-cell RNA-sequencing) to targeted studies that focus on functional aspects of the identified cell populations.

References

[1] Maayan Baron, Adrian Veres, Samuel L. Wolock, Aubrey L. Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K. Wagner, Shai S. Shen-Orr, and Allon M. Klein. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4):346–360, 2016.

[2] Laurence Chapuy, Marwa Bsat, Siranush Sarkizova, Manuel Rubio, Amélie Therrien, Evelyne Wassef, Mickael Bouin, Katarzina Orlicka, Audrey Weber, Nir Hacohen, Alexandra-Chloé Villani, and Marika Sarfati. Two distinct colonic CD14 + subsets characterized by single-cell RNA profiling in Crohn's disease. *Mucosal Immunology*, 12(3):703, May 2019. ISSN 1935-3456. doi: 10.1038/s41385-018-0126-0. URL https://www.nature.com/articles/s41385-018-0126-0.

[3] Guangshuai Jia, Jens Preussner, Xi Chen, Stefan Guenther, Xuejun Yuan, Michail Yekelchyk, Carsten Kuenne, Mario Looso, Yonggang Zhou, Sarah Teichmann, and Thomas Braun. Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nature Communications*, 9(1):4877, November 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07307-6. URL https://www.nature.com/articles/s41467-018-07307-6.

[4] Eric M. Kernfeld, Ryan M. J. Genga, Kashfia Neherin, Margaret E. Magaletta, Ping Xu, and René Maehr. A Single-Cell Transcriptomic Atlas of Thymus Organogenesis Resolves Cell Types and Developmental Maturation. *Immunity*, 48(6):1258–1270.e6, June 2018. ISSN 1074-7613. doi: 10.1016/j.immuni.2018.04.015. URL https://www.cell.com/immunity/abstract/S1074-7613(18)30184-5.

[5] Sema Kurtulus, Asaf Madi, Giulia Escobar, Max Klapholz, Jackson Nyman, Elena Christian, Mathias Pawlak, Danielle Dionne, Junrong Xia, Orit Rozenblatt-Rosen, Vijay K. Kuchroo, Aviv Regev, and Ana C. Anderson. Checkpoint Blockade Immunotherapy Induces Dynamic Changes in PD1-CD8+ Tumor-Infiltrating T Cells. *Immunity*, 50(1):181–194.e6, January 2019. ISSN 1074-7613. doi: 10.1016/j.immuni.2018.11.014. URL https://www.cell.com/immunity/abstract/S1074-7613(18)30522-3.

[6] Franziska Paul, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, Eyal David, Nadav Cohen, Felicia Kathrine Bratt Lauridsen, Simon Haas, Andreas Schlitzer, Alexander Mildner, Florent Ginhoux, Steffen Jung, Andreas Trumpp, Bo Torben Porse, Amos Tanay, and Ido Amit. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell*, 163(7):1663–1677, December 2015. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2015.11.013. URL https://www.cell.com/cell/abstract/S0092-8674(15)01493-2.

[7] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, May 2015. ISSN 1546-1696. doi: 10.1038/nbt.3192. URL https://www.nature.com/articles/nbt.3192.

[8] Karthik Shekhar, Sylvain W. Lapan, Irene E. Whitney, Nicholas M. Tran, Evan Z. Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z. Levin, James Nemesh, Melissa Goldman, Steven A. McCarroll, Constance L. Cepko, Aviv Regev, and Joshua R. Sanes. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*, 166(5):1308–1323.e30, August 2016. ISSN 1097-4172. doi: 10.1016/j.cell.2016.07.054.

[9] Meromit Singer, Chao Wang, Le Cong, Nemanja D. Marjanovic, Monika S. Kowalczyk, Huiyuan Zhang, Jackson Nyman, Kaori Sakuishi, Sema Kurtulus, David Gennert, Junrong Xia, John Y. H. Kwon, James Nevin, Rebecca H. Herbst, Itai Yanai, Orit Rozenblatt-Rosen, Vijay K. Kuchroo, Aviv Regev, and Ana C. Anderson. A Distinct Gene Module for Dysfunction Uncoupled from Activation in Tumor-Infiltrating T Cells. *Cell*, 166(6):1500–1511.e9, September 2016. ISSN 1097-4172. doi: 10.1016/j.cell.2016.08.052.

[10] Raul German Spallanzani, David Zemmour, Tianli Xiao, Teshika Jayewickreme, Chaoran Li, Paul J. Bryce, Christophe Benoist, and Diane Mathis. Distinct immunocyte-promoting and adipocyte-generating stromal components coordinate adipose tissue immune and metabolic tenors. *Science Immunology*, 4(35):eaaw3658, May 2019. ISSN 2470-9468. doi: 10.1126/sciimmunol.aaw3658. URL https://immunology-sciencemag-org.ezp-prod1.hul.harvard.edu/content/4/35/eaaw3658.

[11] Roser Vento-Tormo, Mirjana Efremova, Rachel A. Botting, Margherita Y. Turco, Miquel Vento-Tormo, Kerstin B. Meyer, Jong-Eun Park, Emily Stephenson, Krzysztof Polański, Angela Goncalves, Lucy Gardner, Staffan Holmqvist, Johan Henriksson, Angela Zou, Andrew M. Sharkey, Ben Millar, Barbara Innes, Laura Wood, Anna Wilbrey-Clark, Rebecca P. Payne, Martin A. Ivarsson, Steve Lisgo, Andrew Filby, David H. Rowitch, Judith N. Bulmer, Gavin J. Wright, Michael J. T. Stubbington, Muzlifah Haniffa, Ashley Moffett, and Sarah A. Teichmann. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*, 563(7731): 347, November 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0698-6. URL https://www.nature.com/articles/s41586-018-0698-6.

[12] Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, Laura Jardine, David Dixon, Emily Stephenson, Emil Nilsson, Ida Grundberg, David McDonald, Andrew Filby, Weibo Li, Philip L. De Jager, Orit Rozenblatt-Rosen, Andrew A. Lane, Muzlifah Haniffa, Aviv Regev, and Nir Hacohen. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335):eaah4573, April 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aah4573. URL https://science.sciencemag.org/content/356/6335/eaah4573.

[13] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, June 2019. ISSN 1744-4292. doi: 10.15252/msb.20188746. URL https://www.embopress.org/doi/full/10.15252/msb.20188746.

[14] Eran Eden, Doron Lipson, Sivan Yogev, and Zohar Yakhini. Discovering Motifs in Ranked Lists of DNA Sequences. *PLOS Computational Biology*, 3(3):e39, March 2007. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0030039. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0030039.

[15] Florian Wagner. The XL-mHG Test For Enrichment: A Technical Report. *arXiv:1507.07905 [stat]*, July 2015. URL http://arxiv.org/abs/1507.07905. arXiv: 1507.07905.

[16] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16, 2015. ISSN 1474-7596. doi: 10.1186/s13059-015-0844-5. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4676162/.

[17] Vasilis Ntranos, Lynn Yi, Páll Melsted, and Lior Pachter. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nature Methods*, 16(2):163, February 2019. ISSN 1548-7105. doi: 10.1038/s41592-018-0303-9. URL https://www.nature.com/articles/s41592-018-0303-9.

[18] Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

[19] Florian Wagner, Dalia Barkley, and Itai Yanai. Accurate denoising of single-cell RNA-Seq data using unbiased principal component analysis. *bioRxiv*, page 655365, June 2019. doi: 10.1101/655365. URL https://www.biorxiv.org/content/10.1101/655365v2.

[20] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640, June 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2930. URL https://www.nature.com/articles/nmeth.2930.

[21] Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, Daosheng Huang, Yang Xu, Wentao Huang, Mengmeng Jiang, Xinyi Jiang, Jie Mao, Yao Chen, Chenyu Lu, Jin Xie, Qun Fang, Yibin Wang, Rui Yue, Tiefeng Li, He Huang, Stuart H. Orkin, Guo-Cheng Yuan, Ming Chen, and Guoji Guo. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, 172(5):1091–1107.e17, February 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.02.001. URL http://www.sciencedirect.com/science/article/pii/S0092867418301168.

[22] L. M. Nadler, K. C. Anderson, G. Marti, M. Bates, E. Park, J. F. Daley, and S. F. Schlossman. B4, a human B lymphocyte-associated antigen expressed on normal, mitogen-activated, and malignant B lymphocytes. *The Journal of Immunology*, 131(1):244–250, July 1983. ISSN 0022-1767, 1550-6606. URL http://www.jimmunol.org/content/131/1/244.

[23] STEFAN C. Meuer, Kathleen A. Fitzgerald, Rebecca E. Hussey, James C. Hodgdon, Stuart F. Schlossman, and Ellis L. Reinherz. Clonotypic structures involved in antigen-specific human T cell function. Relationship to the T3 molecular complex. *Journal of Experimental Medicine*, 157(2):705–719, 1983.

[24] Conor Delaney, Alexandra Schnell, Louis V Cammarata, Aaron Yao-Smith, Aviv Regev, Vijay K Kuchroo, and Meromit Singer. Combinatorial prediction of marker panels from single-cell transcriptomic data. *Molecular Systems Biology*, 15:e9005, 2019.

[25] Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727):367, October 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0590-4. URL https://www.nature.com/articles/s41586-018-0590-4.