



Cell Type Diversity Statistic: An Entropy-Based Metric to Compare Overall Cell Type Composition Across Samples

Tanya T Karagiannis^{1,2*}, Stefano Monti^{1,3,4} and Paola Sebastiani¹

¹Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, United States, ²Bioinformatics Program, Boston University, Boston, MA, United States, ³Division of Computational Biomedicine, Boston University School of Medicine, Boston, MA, United States, ⁴Department of Biostatistics, Boston University School of Public Health, Boston, MA, United States

OPEN ACCESS

Edited by:

Tao Huang,
Shanghai Institute of Nutrition and
Health, (CAS), China

Reviewed by:

Cong Liang,
Tianjin University, China
Alsu Missarova,
European Bioinformatics Institute
(EMBL-EBI), United Kingdom

*Correspondence:

Tanya T Karagiannis
tkaragiannis@
tuftsmedicalcenter.org

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 14 January 2022

Accepted: 18 March 2022

Published: 08 April 2022

Citation:

Karagiannis TT, Monti S and
Sebastiani P (2022) Cell Type Diversity
Statistic: An Entropy-Based Metric to
Compare Overall Cell Type
Composition Across Samples.
Front. Genet. 13:855076.
doi: 10.3389/fgene.2022.855076

Changes of cell type composition across samples can carry biological significance and provide insight into disease and other conditions. Single cell transcriptomics has made it possible to study cell type composition at a fine resolution. Most single cell studies investigate compositional changes between samples for each cell type independently, not accounting for the fixed number of cells per sample in sequencing data. Here, we provide a metric of the distribution of cell type proportions in a sample that can be used to compare the overall distribution of cell types across multiple samples and biological conditions. This is the first method to measure overall cell type composition at the single cell level. We use the method to assess compositional changes in peripheral blood mononuclear cells (PBMCs) related to aging and extreme old age using multiple single cell datasets from individuals of four age groups across the human lifespan.

Keywords: single cell transcriptomic analysis, cell type composition, sample level analysis, sample-to-sample comparison, diversity statistics

INTRODUCTION

Tissues are composed of heterogeneous cell types that demonstrate differences in biological function (Raj and van Oudenaarden, 2008; Choi and Kim, 2019). Gene expression profiling methods such as single cell RNA-sequencing (scRNA-seq) have made it possible to profile the genome-wide gene expression levels for each single cell of a sample, to account for cell-to-cell variability (Chen et al., 2019; Tanay and Regev, 2017; Choi and Kim, 2019), and to identify and characterize cell types in a given tissue (Jaitin et al., 2014; Macosko et al., 2015; Zheng et al., 2017). ScRNA-seq has been extensively applied in multiple research areas to study cell types and states, as well as cell types compositional changes, across diseases and conditions (Shalek et al., 2014; Baron et al., 2016; Muraro et al., 2016; Villani et al., 2017; Butler et al., 2018; Schaum et al., 2018; Mathys et al., 2019; Velmeshev et al., 2019).

Most methods to analyze cell type composition at a single cell level model each cell type independently from other cell types (Haber et al., 2017; Luecken and Theis, 2019; Hashimoto et al., 2019; Wilk et al., 2020; Zheng et al., 2020; Zhu et al., 2020). For example, changes of peripheral blood mononuclear cells (PBMCs) composition observed between supercentenarians and younger age controls in Hashimoto et al., 2019 were assessed for each cell type *independently* using a Wilcoxon rank sum test. Other studies have taken a similar approach when assessing compositional changes

between groups of samples at the single cell level (Haber et al., 2017; Luecken and Theis, 2019; Hashimoto et al., 2019; Wilk et al., 2020; Zheng et al., 2020; Zhu et al., 2020). However, high throughput sequencing data are in fact compositional (Gloor et al., 2016, 2017; Lin and Peddada, 2020). The approach we propose rests on the observation that a sample in scRNA-seq data is composed of cell abundances across cell types that are in constrained proportions, given the total number of cells in the sample (Gloor et al., 2016; Gloor et al., 2017; Lin and Peddada, 2020). In other words, the proportion of cell types within a sample are in fact dependent on each other: if the proportion of one type increases, then others need to decrease (Luecken and Theis, 2019). It is thus necessary to account for this dependency when assessing overall cell type compositional changes across samples. In addition, there is no method that provides a numerical summary of a sample overall cell type composition that can be used to compare samples in different conditions (Luecken and Theis, 2019).

Here, we introduce a statistic to summarize the distribution of the proportions of cell types in a sample. Using three single cell transcriptomic datasets of PBMCs comprising four age groups, we show the utility of this statistic to describe changes in PBMCs composition in aging and extreme old age.

MATERIALS AND METHODS

Cell type diversity statistic. The statistic makes three assumptions: 1) To make different samples of cells comparable, cell abundances must be normalized based on the total number of cells in a sample; 2) After conditioning on the total number of cells in a sample (Gloor et al., 2017), the cell type composition data is a simplex (Aitchison, 1982), and when the proportion of one cell type changes, the proportion of the other cell types must change as well to maintain the total fixed; and 3) To make the statistic comparable across different cell type resolutions, the statistic must be normalized. Formally, we denote by $p_i = \frac{n_i}{n_s}$ the proportion of cell type i , for $i = 1, \dots, k$ in a sample s with n_s cells, so that $\sum_{i=1}^k p_i = 1$.

The statistic is adapted from alpha diversity measures applied in ecology and microbiome studies (Whittaker, 1972; Olde Loohuis et al., 2018; Calle, 2019). We measure the overall cell type composition of a sample by the adjusted entropy

$$E_s = \frac{-\sum_{i=1}^k p_i \log(p_i) - \log(k)}{\log(k)} = \frac{-\sum_{i=1}^k p_i \log(p_i)}{\log(k)} - 1$$

In the formula, $\log(k)$ is the maximum value of $-\sum_{i=1}^k p_i \log(p_i)$ that is reached when $p_i = \frac{1}{k}$ for all indexes i , so that the distribution is uniform. The minimum value of $-\sum_{i=1}^k p_i \log(p_i)$ is 0, which corresponds to a mass-point distribution with $p_i = 0$ for all indexes i but one. The adjusted entropy E_s therefore ranges between $[-1, 0]$. A sample with more uniformity in cell type proportions, and hence more variability, will result in a greater cell type diversity statistic and $E_s = 0$ in a sample with equal proportions of all cell types. A sample with cell type

proportions that are skewed towards specific cell types, and less variability, will have a lower statistic and $E_s = -1$ when all cells are of one type.

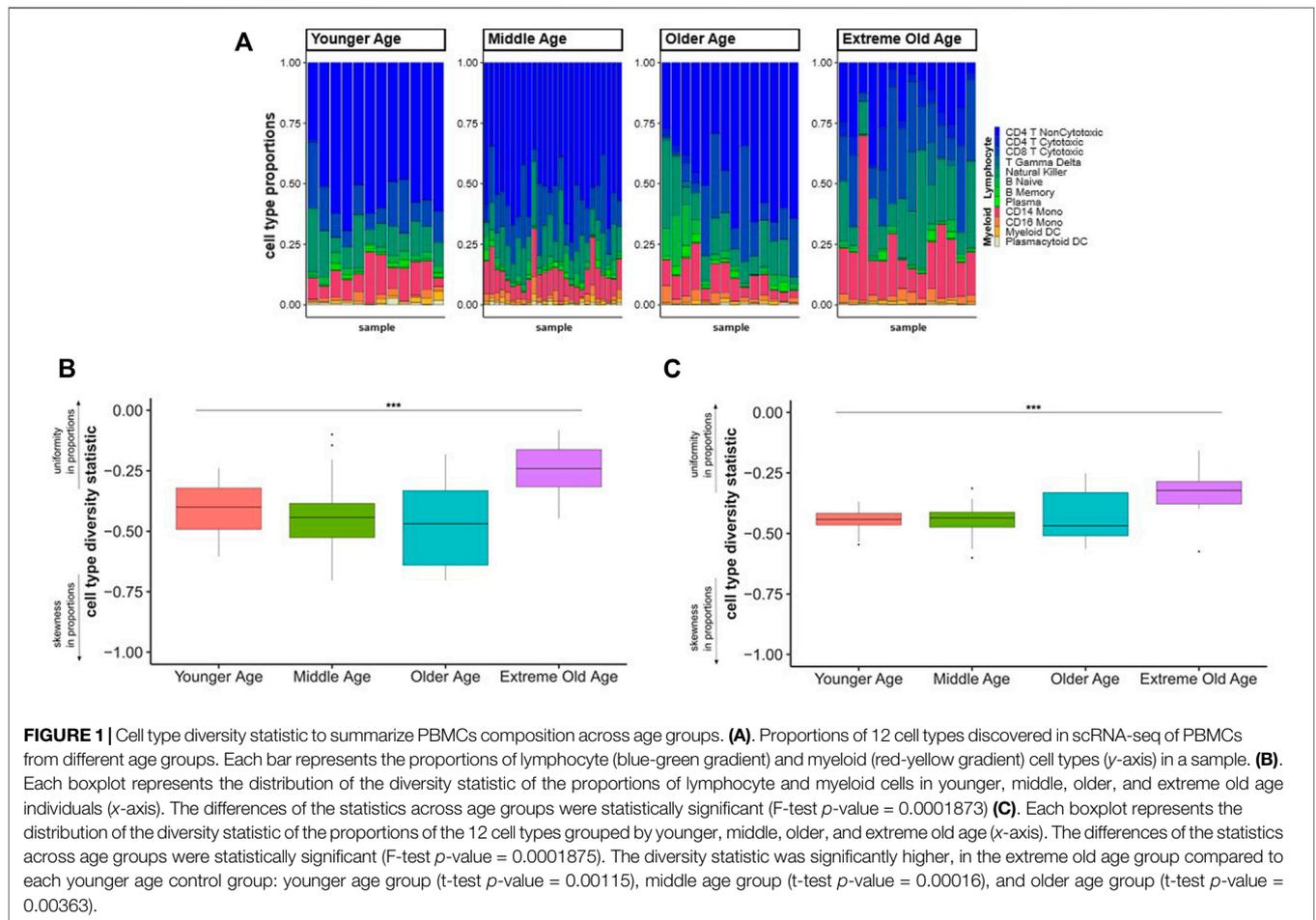
Data. To demonstrate the utility of the cell type diversity statistic, we analyzed three single cell transcriptomic datasets of PBMCs representing regular aging and extreme old age. One dataset comprised samples of 7 centenarians from the New England Centenarian Study (NECS) (Sebastiani and Perls, 2012) and 2 younger age controls. We downloaded a publicly available scRNA-seq dataset of PBMCs from 45 younger age controls (van der Wijst et al., 2018), which we will refer to as NATGEN, and a publicly available scRNA-seq dataset of PBMCs from 5 younger age controls and 7 supercentenarians, which we will refer to as PNAS (Hashimoto et al., 2019). We integrated these datasets and stratified the samples into four age groups of the human lifespan: 12 subjects of younger age (20–39), 26 subjects of middle age (40–59), 14 subjects of older age (60–89), and 14 subjects of extreme longevity (100–119). Data processing steps and identification of the 12 cell types are described in the Supplement.

Application of cell type diversity statistic. We integrated the datasets to generate a matrix of cell type abundances across samples from all three datasets. We calculated the cell type proportions for each sample such that the sum of the cell type proportions for a particular sample equals to 1. We applied the cell type diversity statistic to different cell type resolutions: 1) based on the proportions of lymphocytes and myeloid cells; and 2) based on the proportions of the 12 lymphocyte and myeloid subpopulations that were detected in the data. For both resolutions, we measured the cell type diversity statistic per sample and compared the differences of the statistics between the four age groups using ANOVA and pairwise T-tests with significance level 0.05.

RESULTS AND DISCUSSION

We applied the cell type diversity statistic to the cell type proportions from the three scRNA-seq datasets of younger age individuals and centenarians to assess overall compositional changes across four age groups: younger age (20–39), middle age (40–59), older age (60–89), and extreme old age (100–119 years of age). We first calculated the cell type proportions for each sample across the four age groups (Figure 1A, Supplementary Table S1) and we observed a shift in the distribution of cell proportions from lymphocyte and myeloid cell types from younger ages to centenarians (Figure 1A).

We then calculated the cell type diversity statistic to measure the variability of the proportion of lymphocyte and myeloid cells in each sample (Supplementary Table S2). Comparing the cell type diversity statistics across the four age groups, we found a significant difference in the distribution of the statistics across the four age groups (F-test p -value = 0.0001873) (Figure 1B). The increased value of the cell type diversity statistic in the extreme old age group is consistent with the shift in abundances from lymphocytes to myeloid cells, which is an expected change in the



immune system with aging (Geiger et al., 2013). We also applied the cell type diversity statistic to measure the variability of the proportions of 12 lymphocyte and myeloid subpopulations in each sample (Supplementary Table S3). We again found a significant difference in the distribution of the statistic in the four age groups (F-test p -value = 0.0001875) (Figure 1C). Specifically, centenarians had significantly increased cell type diversity statistics compared to each younger age control group: younger age group (t-test p -value = 0.00115), middle age group (t-test p -value = 0.00016), and older age group (t-test p -value = 0.00363) (Figure 1C). The pattern of the cell type diversity with age groups suggests that centenarians have a more uniform distribution of cell types compared to individuals of younger ages even at a finer resolution of cell types.

The analyses illustrate how the cell type diversity statistic can be used in combination with visualizations of cell type proportions to provide a numerical summary of the distribution of cell types in different conditions. We showed an application of this metric in the context of aging to summarize changes of the distribution of cell types across different age groups, at different resolutions. The metric showed a significant change of the distribution of 12 cell types in extreme old age compared to younger age groups, as well as a significant change of the proportion of lymphocytes and myeloid

cells that are biologically relevant to aging (Geiger et al., 2013). Although in our analysis the distribution of the cell type diversity statistics did not change with different cell type resolutions, in other applications the statistic could change since the distribution of the proportions of subpopulations of cells can be very different.

One major challenge in the analysis of single cell transcriptomics data is in the identification and annotation of cell types. There are varying methods to identify cell types (Andrews et al., 2021; Adil et al., 2021; Shekhar and Menon, 2019; Luecken and Theis, 2019) and the resolution of cell type for analysis should be selected based on the biological question of interest (Luecken and Theis, 2019). Another challenge of this type of analyses is accounting for cell types that are not detectable under specific conditions. Other metrics are needed to account for cell types that are not detected in all conditions.

The cell type diversity statistic is applied as a global summary of cell type composition, and additional analyses are required to quantify individual cell type changes and to adjust this analysis for additional covariates. The recent method scCoda uses a Bayesian Dirichlet regression model to examine individuals cell type changes and accounts for the constrained proportions in single cell composition data is particularly promising (Büttner et al., 2021).

Entropy as a metric to study composition level data has been applied in many fields including analyses of microbiome data (Whittaker, 1972; Olde Loohuis et al., 2018; Calle, 2019). The importance in applying this metric to single cell transcriptomics is that it accounts for the constrained proportions of cell types in each sample, and ignoring these constraints can result in inconsistencies when assessing compositional changes (Gloor et al., 2016; Gloor et al., 2017; Calle, 2019; Luecken and Theis, 2019).

CONCLUSION

We present the cell type diversity statistic, an entropy-based measure to assess and summarize the overall cell type composition of samples in single cell gene expression data. The diversity statistic allows for the investigation of global cell type compositional changes applicable to studying disease and other conditions at the single cell level. We demonstrate the utility of this method by its application to single cell datasets of aging and extreme old age, and show that it can reveal novel changes in composition in aging at different resolutions.

CODE AVAILABILITY STATEMENT

The cell type diversity statistic is available as a function in R at <https://github.com/tanya-karagiannis/Cell-Type-Diversity-Statistic>. The function can be applied to a matrix of cell type proportions per sample, a Seurat object, and a Single Cell Experiment object.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The data that support these findings are

REFERENCES

- Adil, A., Kumar, V., Jan, A. T., and Asger, M. (2021). Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis. *Front. Neurosci.* 15. doi:10.3389/fnins.2021.591122
- Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B (Methodological)* 44, 139–160. doi:10.1111/j.2517-6161.1982.tb01195.x
- Andrews, T. S., Kiselev, V. Y., McCarthy, D., and Hemberg, M. (2021). Tutorial: Guidelines for the Computational Analysis of Single-Cell RNA Sequencing Data. *Nat. Protoc.* 16, 1–9. doi:10.1038/s41596-020-00409-w
- Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., et al. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cel Syst.* 3, 346–360. e4. doi:10.1016/j.cels.2016.08.011
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species. *Nat. Biotechnol.* 36, 411–420. doi:10.1038/nbt.4096
- Büttner, M., Ostner, J., Müller, C. L., Theis, F. J., and Schubert, B. (2021). scCODA Is a Bayesian Model for Compositional Single-Cell Data Analysis. *Nat. Commun.* 12, 6876. doi:10.1038/s41467-021-27150-6

publicly available and were accessed from several repositories. NATGEN single cell expression data and subject level data were publicly available as referenced in (van der Wijst et al., 2018): <https://molgenis58.target.rug.nl/scrna-seq/>. PNAS single cell expression data and subject level data was available as referenced in (Hashimoto et al., 2019): <http://gerg.gsc.riken.jp/SC2018/>. NECS will be available from Synapse (URL <https://adknowledgeportal.synapse.org/Explore/Projects/DetailsPage?Grant%20Number=UH2AG064704>).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Boston University IRB. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TK, PS, and SM conceived of the presented method for single cell transcriptomics data. TK implemented the method and wrote the paper with feedback from all authors. All authors contributed to the final version of the manuscript.

FUNDING

This work was supported by NIH-NIA UH2AG064704.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.855076/full#supplementary-material>

- Calle, M. L. (2019). Statistical Analysis of Metagenomics Data. *Genomics Inform.* 17, e6. doi:10.5808/GI.2019.17.1.e6
- Chen, G., Ning, B., and Shi, T. (2019). Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front. Genet.* 10, 317. doi:10.3389/fgene.2019.00317
- Choi, Y. H., and Kim, J. K. (2019). Dissecting Cellular Heterogeneity Using Single-Cell RNA Sequencing. *Mol. Cell* 42, 189–199. doi:10.14348/molcells.2019.2446
- Geiger, H., de Haan, G., and Florian, M. C. (2013). The Ageing Haematopoietic Stem Cell Compartment. *Nat. Rev. Immunol.* 13, 376–389. doi:10.1038/nri3433
- Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* 8, 2224. doi:10.3389/fmicb.2017.02224
- Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., and Egozcue, J. J. (2016). It's All Relative: Analyzing Microbiome Data as Compositions. *Ann. Epidemiol.* 26, 322–329. doi:10.1016/j.annepidem.2016.03.003
- Haber, A. L., Biton, M., Rogel, N., Herbst, R. H., Shekhar, K., Smillie, C., et al. (2017). A Single-Cell Survey of the Small Intestinal Epithelium. *Nature* 551, 333–339. doi:10.1038/nature24489
- Hashimoto, K., Kouno, T., Ikawa, T., Hayatsu, N., Miyajima, Y., Yabukami, H., et al. (2019). Single-cell Transcriptomics Reveals Expansion of Cytotoxic CD4 T Cells in Supercentenarians. *Proc. Natl. Acad. Sci. U.S.A.* 116, 24242–24251. doi:10.1073/pnas.1907883116

- Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., et al. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-free Decomposition of Tissues into Cell Types. *Science* 343, 776–779. doi:10.1126/science.1247651
- Lin, H., and Peddada, S. D. (2020). Analysis of Microbial Compositions: a Review of Normalization and Differential Abundance Analysis. *npj Biofilms Microbiomes* 6, 60–13. doi:10.1038/s41522-020-00160-w
- Lueckel, M. D., and Theis, F. J. (2019). Current Best Practices in Single-Cell RNA-Seq Analysis: a Tutorial. *Mol. Syst. Biol.* 15, e8746. doi:10.15252/msb.20188746
- Macosko, E. Z., Basu, A., Satija, R., Nemes, J., Shekhar, K., Goldman, M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161, 1202–1214. doi:10.1016/j.cell.2015.05.002
- Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., et al. (2019). Single-cell Transcriptomic Analysis of Alzheimer's Disease. *Nature* 570, 332–337. doi:10.1038/s41586-019-1195-2
- Muraro, M. J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., et al. (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cel Syst.* 3, 385–394. e3. doi:10.1016/j.cels.2016.09.002
- Olde Loohuis, L. M., Mangul, S., Ori, A. P. S., Jospin, G., Koslicki, D., Yang, H. T., et al. (2018). Transcriptome Analysis in Whole Blood Reveals Increased Microbial Diversity in Schizophrenia. *Transl Psychiatry* 8, 96–99. doi:10.1038/s41398-018-0107-9
- Raj, A., and van Oudenaarden, A. (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and its Consequences. *Cell* 135, 216–226. doi:10.1016/j.cell.2008.09.050
- Schaum, N., Karkanas, J., Neff, N. F., May, A. P., Quake, S. R., Wyss-Coray, T., et al. (2018). Single-cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris. *Nature* 562, 367–372. doi:10.1038/s41586-018-0590-4
- Sebastiani, P., and Perls, T. (2012). The Genetics of Extreme Longevity: Lessons from the New England Centenarian Study. *Front. Genet.* 3. doi:10.3389/fgene.2012.00277
- Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., et al. (2014). Single-cell RNA-Seq Reveals Dynamic Paracrine Control of Cellular Variation. *Nature* 510, 363–369. doi:10.1038/nature13437
- Shekhar, K., and Menon, V. (2019). "Identification of Cell Types from Single-Cell Transcriptomic Data," in *Computational Methods for Single-Cell Data Analysis, Methods in Molecular Biology*. Editor G.-C. Yuan (New York, NY: Springer), 45–77. doi:10.1007/978-1-4939-9057-3_4
- Tanay, A., and Regev, A. (2017). Scaling Single-Cell Genomics from Phenomenology to Mechanism. *Nature* 541, 331–338. doi:10.1038/nature21350
- van der Wijst, M. G. P., Brugge, H., de Vries, D. H., Deelen, P., Swertz, M. A., and Franke, L. (2018). Single-cell RNA Sequencing Identifies Celltype-specific Cis-eQTLs and Co-expression QTLs. *Nat. Genet.* 50, 493–497. doi:10.1038/s41588-018-0089-9
- Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., et al. (2019). Single-cell Genomics Identifies Cell Type-specific Molecular Changes in Autism. *Science* 364, 685–689. doi:10.1126/science.aav8130
- Villani, A. C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., et al. (2017). Single-cell RNA-Seq Reveals New Types of Human Blood Dendritic Cells, Monocytes, and Progenitors. *Science* 356. doi:10.1126/science.aah4573
- Whittaker, R. H. (1972). Evolution and Measurement of Species Diversity. *Taxon* 21, 213–251. doi:10.2307/1218190
- Wilk, A. J., Rustagi, A., Zhao, N. Q., Roque, J., Martínez-Colón, G. J., McKechnie, J. L., et al. (2020). A Single-Cell Atlas of the Peripheral Immune Response in Patients with Severe COVID-19. *Nat. Med.* 26, 1070–1076. doi:10.1038/s41591-020-0944-y
- Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively Parallel Digital Transcriptional Profiling of Single Cells. *Nat. Commun.* 8, 14049. doi:10.1038/ncomms14049
- Zheng, Y., Liu, X., Le, W., Xie, L., Li, H., Wen, W., et al. (2020). A Human Circulating Immune Cell Landscape in Aging and COVID-19. *Protein Cell* 11, 740–770. doi:10.1007/s13238-020-00762-2
- Zhu, L., Yang, P., Zhao, Y., Zhuang, Z., Wang, Z., Song, R., et al. (2020). Single-Cell Sequencing of Peripheral Mononuclear Cells Reveals Distinct Immune Response Landscapes of COVID-19 and Influenza Patients. *Immunity* 53, 685–696. e3. doi:10.1016/j.immuni.2020.07.009

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Karagiannis, Monti and Sebastiani. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.