frontiers | Frontiers in Genetics

TYPE Opinion
PUBLISHED 01 June 2023
DOI 10.3389/fgene.2023.1205499

Check for updates

OPEN ACCESS

EDITED BY
Blanka Rogina,
University of Connecticut Health Center,
United States

REVIEWED BY
Lei Li,
St. Jude Children's Research Hospital,
United States
Xinjun Wang,
Memorial Sloan Kettering Cancer Center,
United States

*CORRESPONDENCE
Tanya T. Karagiannis,
✉ Tanya.Karagiannis@tuftsmedicine.org

RECEIVED 13 April 2023
ACCEPTED 17 May 2023
PUBLISHED 01 June 2023

CITATION
Karagiannis TT, Monti S and Sebastiani P (2023), Bayesian differential analysis of cell type proportions: opinion.
Front. Genet. 14:1205499.
doi: 10.3389/fgene.2023.1205499

Bayesian differential analysis of cell type proportions: opinion

Tanya T. Karagiannis[1]*, Stefano Monti[2,3,4] and Paola Sebastiani[1,5]

[1]Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, United States, [2]Division of Computational Biomedicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States, [3]Department of Biostatistics, Boston University School of Public Health, Boston, MA, United States, [4]Bioinformatics Program, Boston University, Boston, MA, United States, [5]Department of Medicine, Tufts University, Boston, MA, United States

KEYWORDS

single cell transcriptomics, cell type composition, cell type dependence, Bayesian multinomial regression, estimated cell type probabilities

1 Introduction

The paper by Ahlers et al., 2022 uses single cell transcriptomics data to analyze dermal sheath cells in younger and older individuals, including a focus on cell type compositional changes between the two age groups. We noticed that in Figure 1 (Ahlers et al., 2022), the investigators compared the proportions of each cell type between age groups one at a time. This analysis approach is consistent with most methods used to analyze single cell distribution data (Luecken and Theis, 2019). However, by comparing the proportions of one cell type at a time, one does not account for the constraint that the proportions must add up to 1. Indeed, the proportions of the cell types for each group (old and young) in Figure 1 of (Ahlers et al., 2022) do not seem to add up to 1. This challenges the interpretation of the results.

Methods have been created to account for this constraint on the cell type proportions, including scCODA (Büttner et al., 2021). This method uses a multinomial distribution to describe the vector of probabilities (proportions) of all cell types in a sample, and a logit-type parameterization that relies on a reference cell type to avoid issues of convergence of the Bayesian estimation algorithm. This approach is ideal when one can identify a reference cell type whose proportion is unaffected by the condition under study and/or is stable in relative abundance across samples. However, there are situations where no such reference cell type can be determined. For example, in our recent study of the distribution of peripheral blood mononuclear cells (PBMCs) with age, we could not identify a cell type with stable proportion in various age groups (Karagiannis et al., 2023).

We recently introduced a simple and robust method for the analysis of single cell distribution data using a Bayesian multinomial regression model (Karagiannis et al., 2023). The approach estimates the cell type proportions without the need to provide a reference cell type and guarantees that the cell type proportions add up to 1. Here, we take the opportunity to describe the advantages of this method in the analysis of single cell distribution data and provide an example analysis script in the R software.

2 Methods

2.1 Modeling approach

We have configured the Bayesian multinomial regression using the R package *rjags* (Plummer, 2008) to model the cell type abundance distribution as a function of covariates of interest. We model the vector of cell type counts in a sample using the following parameterization:

$$Y_{i,1:J} \sim Multinomial\left(p_{i,1:J}, N.total_i\right)$$
$$log\ (q_{i,j}) = \alpha_j + \beta_{1,j} X_{1\,i,j} + \ldots + \beta_{c,j}\, X_{c\,i,j}$$
$$p_{i,j} = \frac{q_{i,j}}{\sum_{k=1}^{J} q_{i,k}}$$
$$\alpha_j \sim Normal\,(0, 0.001)$$
$$\beta_{1,j} \sim Normal\,(0, 0.001)$$
$$\ldots$$
$$\beta_{c,j} \sim Normal\,(0, 0.001)$$

where $Y_{i,1:J}$ represents the vector of numbers of cell types 1: J in sample i , and is modeled using a multinomial distribution with probabilities $p_{i,1:J}$ such that $\sum_{j=1}^{J} Y_{i,j} = N.total_i$ and $\sum_{j=1}^{J} p_{i,j} = 1$, for all sample i. The probabilities $p_{i,1:J}$ can depend on covariates $X_1 \ldots X_c$ through the function $log\ (q_{i,j})$. The regression parameters $\alpha_j, \beta_{1,j}, \ldots, \beta_{c,j}$, $j = 1: J$ can be estimated using Markov Chain Monte Carlo (MCMC) sampling as implemented in rjags (Plummer, 2008), and used to estimate the probabilities of cell types.

The advantage of this Bayesian and unconditional approach is that one can use many tools to monitor the goodness of fit of the model and the convergence of the parameter estimates, including graphical diagnostics for Bayesian modelling (Plummer et al., 2006). Commonly used graphical diagnostics include trace plots to display posterior samples of the model parameters over the MCMC iterations to visualize convergence, density plots to plot the posterior distribution of the model parameters that can provide insight into the uncertainty of the estimates, autocorrelation plots to identify any correlation between samples in the Markov chain to identify any issues in mixing or slow convergence, and the Geweke diagnostic to examine the difference in the means between the early and late portions of the MCMC chain to assess convergence.

In addition, one can estimate the absolute proportion of each cell type and provide measures of the uncertainty of the estimates (Supplementary Figure S1). To assess the effect of covariates in each cell type, we can use the MCMC estimates of the regression coefficients and their standard errors to calculate approximate two-sided p-values and use Benjamin-Hochberg correction for multiple testing. In addition, the implementation of the multinomial regression in rjags does not require a logistic parameterization (Lunn et al., 2013) and thus the analysis produces estimates of the absolute proportions of cell types per sample that are easier to interpret compared to odds ratios.

2.2 Analysis script

We developed an example analysis script that uses this approach in the R packages rjags (Plummer, 2008) and coda (Plummer et al., 2006). The script can be easily adjusted based on the study design and covariates of interest (Supplementary Figure S1). To run the analysis scripts, the program JAGS (https://mcmc-jags.sourceforge.io/) is required for download and installation. JAGS is a program for statistical analysis in the Bayesian framework using MCMC simulations. To run the analysis scripts for model configuration, initialization and parameter inference, the R packages rjags (Plummer, 2008) and coda (Plummer et al., 2006) are required for installation. Additional R packages required for data initialization, manipulation, and visualization include packages in tidyverse, and the hablar and patchwork packages.

2.3 Data

We demonstrate this model and approach using cell type distribution data of 66 subjects from single cell transcriptomics datasets of aging and longevity. The data is described in Karagiannis et al. (Karagiannis et al., 2023).

3 Example

As an example, we show how we used this approach to characterize the distribution of PBMCs at different ages, based on previously published work (Karagiannis et al., 2023). We used single cell transcriptomics data of PBMCs from 66 male and female subjects across four age groups with ages 20–119 years to identify 13 immune cell types based on specific gene signatures (Karagiannis et al., 2023). We applied the Bayesian multinomial regression model to the distributions of the 13 immune cell types and used 1,000 MCMC iterations with 500 iterations for burn-in to estimate cell type proportions and 95 percent credible intervals for males and female subjects for each age group across all immune cell types (Karagiannis et al., 2023).

Figure 1 displays the estimated proportions (predicted probabilities) of the 13 cell types using this approach and the observed cell type proportions calculated in the 66 subjects grouped by age and sex. The plots show a very good agreement between observed and estimated proportions, particularly for not small probability values. This analysis identified significant age-related changes of cell type composition in EL including a significant reduction of lymphocyte subtypes nCD4TC and mCD4TC (females: 6.00%–9.34%; males: 5.38%–7.78%) compared to younger age (females: 21.62%–32.10%; males: 18.18%–29.08%) and a significant decrease of mDC and pDC in EL (females: 0.31%–0.70%; males: 0.32%–0.88%) compared to younger age (females: 0.80%–1.05%; males: 0.82%–1.33%). Comparing the estimated cell type proportions to the relative proportions across subjects, we found similar results across cell types. Full results are described in Karagiannis et al. (Karagiannis et al., 2023).

For comparison, we applied scCODA to the distribution of the 13 immune cell types. When no obvious reference cell type is available, the recommended use of scCODA is to run the analysis using each cell type as a reference, for a total of 13 tests for comparison in our case, and to then call as significant those changes observed with a credible effect in more than 50% of the runs. Supplementary Table S1 displays the credible compositional changes between EL and younger age for each cell type identified by the Bayesian multinomial regression and by scCODA. We found that scCODA identified compositional

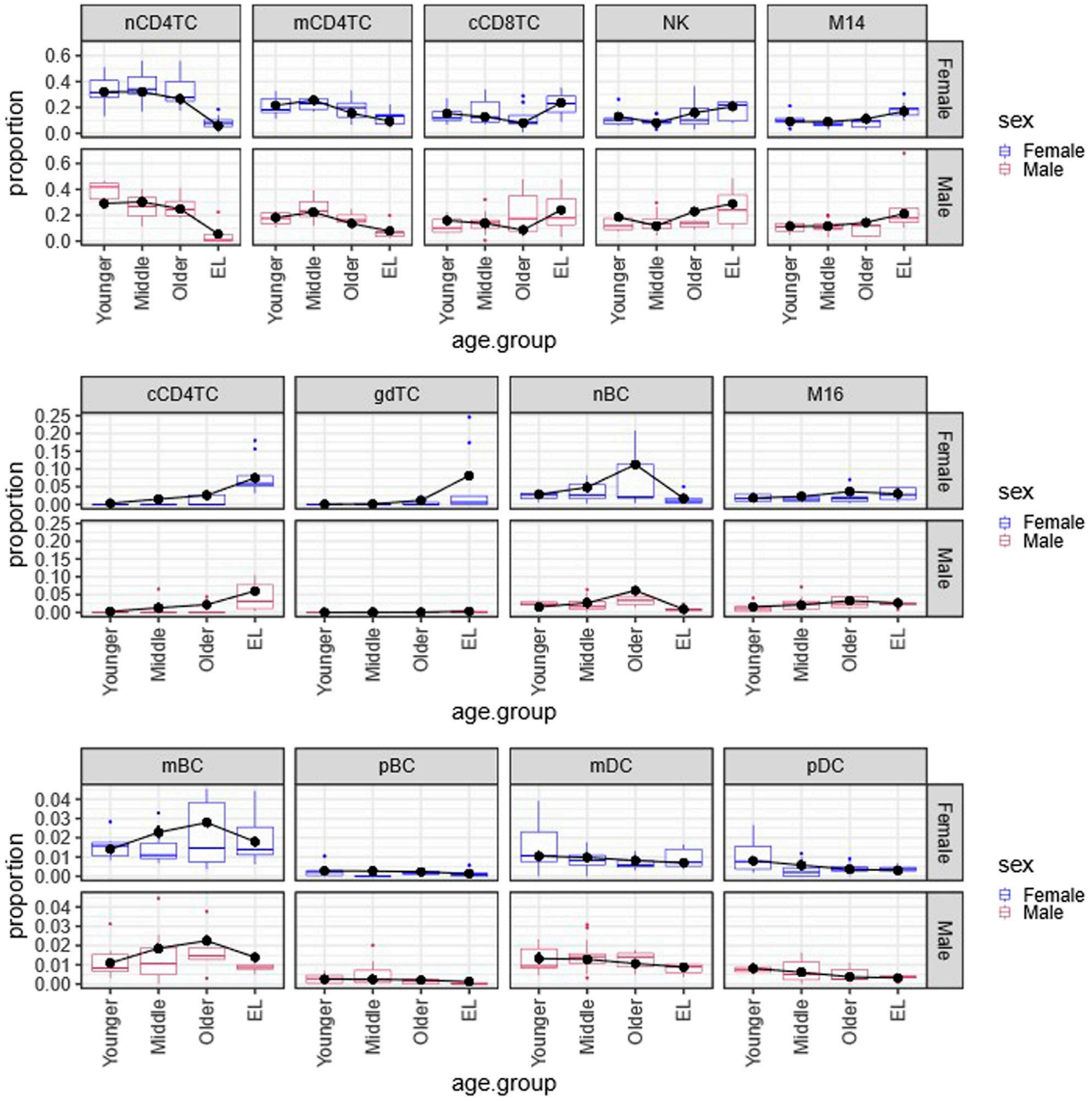

FIGURE 1
Multinomial regression cell type composition estimates across age and sex in PBMCs. Plot of the Bayesian estimates and observed relative proportions of the 13 immune cell types in PBMCs in each age group (Younger, Middle, Older, EL), for males and females. We applied the Bayesian multinomial regression to a matrix of the13 cell counts across the 66 subjects. The 13 cell types are: noncytotoxic naive and memory CD4$^+$ T cells (nCD4TC, mCD4TC), cytotoxic CD4$^+$ T cells (cCD4TC), cytotoxic CD8$^+$ T cells (cCD8TC), gamma-delta T cells (gdTC), naive, memory and plasma B cells (nBC, mBC, and pBC), Natural Killer cells (NK), CD14$^+$ and CD16$^+$ monocytes (M14 and M16), and myeloid and plasmacytoid dendritic cells (mDC and pDC). The relative proportions per subject are represented as boxplots for Females (blue) and Males (maroon). For each cell type, the estimates are overlayed with points (black) and connected by a line (black) to highlight trends across age groups.

changes in 4 of the 9 cell types identified as significantly changed by the multinomial regression model. Of note, although scCODA found nBC to have a credible change in EL compared to younger age, it only identified this change in 10 out of the 13 tests run. The decrease in composition of nBC in centenarians has been previously reported (Hashimoto et al., 2019) and we were able to confirm this credible decrease using the multinomial regression model. We also identified a significant increase in M14 composition using the multinomial regression that supports previous reports of increased composition of M14 with age (Zheng et al., 2020). However, scCODA only identified the credible change in M14 in 2 out of the 13 tests run.

In summary, we show the advantages of using the Bayesian multinomial regression model to identify and provide simple interpretations of changes in the distribution of cell types without a reference cell type. As shown through the example analysis from Karagiannis et al., 2023, we leveraged this method and identified multiple age-related changes including those previously reported and showed how this methodology can be applied to single cell distribution.

4 Discussion

We have presented a new perspective in the analysis of single cell distribution data using a Bayesian multinomial regression that accounts for cell type proportion compositional constraints within each sample and does not require the choice of a reference cell type. The analysis script we developed uses the rjags package in the R software and can be easily generalized to different sets of covariates and study design. We provide detailed documentation for model and parameter configuration and initialization as well as for application to single cell distribution data to obtain posterior distributions of sample proportions across conditions. As shown in the example application to distributions of PBMCs with age from Karagiannis et al., 2023, the Bayesian multinomial regression allows for the investigation of cell type specific compositional changes applicable to studying disease and other conditions. An important feature of our unconditional approach is to estimate the absolute proportions of cell type per sample that are easier to interpret compared to odds ratios or other relative metrics.

Author contributions

TK, PS, and SM conceived of the presented opinion on this method for single cell distribution data. TK wrote the paper with feedback from all authors. All authors contributed to the article and approved the submitted version.

Code availability

The model and application scripts are available on github (https://github.com/Integrative-Longevity-Omics/Bayesian-Multinomial-Regression).

Funding

TK, SM, PS are supported by NIH-NIA UH2AG064704.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1205499/full#supplementary-material

References

Ahlers, J. M. D., Falckenhayn, C., Holzscheck, N., Solé-Boldo, L., Schütz, S., Wenck, H., et al. (2022). Single-cell RNA profiling of human skin reveals age-related loss of dermal sheath cells and their contribution to a juvenile phenotype. *Front. Genet.* 12, 797747. doi:10.3389/fgene.2021.797747

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

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. *PNAS* 116, 24242–24251. doi:10.1073/pnas.1907883116

Karagiannis, T. T., Dowrey, T. W., Villacorta-Martin, C., Montano, M., Reed, E., Belkina, A. C., et al. (2023). Multi-modal profiling of peripheral blood cells across the human lifespan reveals distinct immune cell signatures of aging and longevity. *eBioMedicine* 90, 104514. doi:10.1016/j.ebiom.2023.104514

Luecken, 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

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013). *The bugs book: A practical introduction to bayesian analysis*. London: Chapman Hall.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for MCMC. *R. News* 6, 7–11.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics* 9, 523–539. doi:10.1093/biostatistics/kxm049

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