Bayesian Differential Analysis of Cell Type Proportions

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

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

*To whom correspondence should be addressed
Email: tkaragiannis@tuftsmedicalcenter.org

SUMMARY

The analysis of cell type proportions in a biological sample should account for the compositional nature of the data but most analyses ignore this characteristic with the risk of producing misleading conclusions. The recent method scCODA appropriately incorporates these constraints by using a Bayesian Multinomial-Dirichlet model that requires a reference cell type to normalize the distribution of all cell types. However, a reference cell type that is stable across biological conditions may not always be available. Here, we present an approach that uses a Bayesian multinomial regression for the analysis of single cell distribution data without the need for a reference cell type. We show an implementation example using the rjags package within the R software.

Keywords
Single cell transcriptomic, cell type composition, Bayesian multinomial regression

1. INTRODUCTION

Single cell transcriptomics allows us to explore changes of cell type composition across conditions (Luecken and Theis 2019). Most methods analyze the changes of the proportion of each cell type across biological conditions independently of the other cell types, when in fact the cell type proportions within a sample are dependent on each other and constrained to sum to 1 (Haber *and others* 2017; Luecken and Theis 2019; Hashimoto *and others* 2019; Wilk *and others* 2020; Zhu *and others* 2020; Zheng *and others* 2020). The recent method scCODA, appropriately accounts for these constraints by using a Bayesian Multinomial-Dirichlet model (Büttner *and others* 2021). scCODA 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 ongoing 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 *and others* 2022).

To circumvent this problem, we present an alternative Bayesian multinomial regression analysis of cell type composition. The approach estimates the cell type proportions without the need to provide a reference cell type. We describe the approach 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:

$$Y_{i,1:J} \sim Multinomial(p_{i,1:J}, N.total_i)$$

$$log(q_{i,j}) = \alpha_j + \beta_{1j} X_{1\,i,j} + \cdots + \beta_{cj} X_{c\,i,j}$$

$$p_{i,j} = \frac{q_{i,j}}{\sum_{k=1}^{N.ct} q_{i,k}}$$

$$\alpha_j \sim Normal(0, 0.001)$$

$$\beta_{1,j} \sim Normal(0, 0.001)$$

$$\dots$$

$$\beta_{c,j} \sim Normal(0, 0.001)$$

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 $X_1 \dots X_c$ through the function $log(q_{i,j})$. The regression parameters $\alpha_j, \beta_{1j}, \dots, \beta_{cj}, j = 1{:}J$ 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. In addition, one can estimate the absolute proportion of each cell type and provide measures of the uncertainty of the estimates. 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 analysis produces estimates of the absolute proportions of cell type 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 *and others* 2006). The script can be easily adjusted based on the study design and covariates of interest (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 *and others* 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 *and others* 2022).

3. APPLICATION

As an example, we used this approach to characterize the distribution of PBMCs at different ages. 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 *and others* 2022). We applied the proposed 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.

Figure 1 displays the estimates 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 *and others* 2022).

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. Table 1 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 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 *and others* 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 *and others* 2020). However, scCODA only identified the credible change in M14 in 2 out of the 13 tests run.

In summary, application of our Bayesian multinomial regression model identified multiple age-related changes including those previously reported and showed that this method can be used to identify and provide simple interpretations of changes in the distribution of cell types without a reference cell type.

4. DISCUSSION

We have presented an implementation of Bayesian multinomial regression to analyze single cell distribution data 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 application to distributions of PBMCs with age, 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.

Software

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

Supplementary Material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgements

TK, SM, PS are supported by NIH-NIA UH2AG064704.
Conflict of Interest: None declared.

References

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. *Nature Communications* **12**, 6876.

HABER, A. L., BITON, M., ROGEL, N., HERBST, R. H., SHEKHAR, K., SMILLIE, C., BURGIN, G., DELOREY, T. M., HOWITT, M. R., KATZ, Y., ET AL. (2017). A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339.

HASHIMOTO, K., KOUNO, T., IKAWA, T., HAYATSU, N., MIYAJIMA, Y., YABUKAMI, H., TEROOATEA, T., SASAKI, T., SUZUKI, T., VALENTINE, M., ET AL. (2019). Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. *Proceedings of the National Academy of Sciences* **116**, 24242–24251.

KARAGIANNIS, T. T., DOWREY, T. W., VILLACORTA-MARTIN, C., MONTANO, M., REED, E., ANDERSEN, S. L., PERLS, T. T., MONTI, S., MURPHY, G. J. AND SEBASTIANI, P. (2022). Multi-modal profiling of peripheral blood cells across the human lifespan reveals distinct immune cell signatures of aging and longevity. Immunology. doi:10.1101/2022.07.06.498968.

LUECKEN, M. D. AND THEIS, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* **15**, e8746.

PLUMMER, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics* **9**, 523–539.

PLUMMER, M., BEST, N., COWLES, K. AND VINES, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**, 7–11.

WILK, A. J., RUSTAGI, A., ZHAO, N. Q., ROQUE, J., MARTÍNEZ-COLÓN, G. J., MCKECHNIE, J. L., IVISON, G. T., RANGANATH, T., VERGARA, R., HOLLIS, T., ET AL. (2020). A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nature Medicine* **26**, 1070–1076.

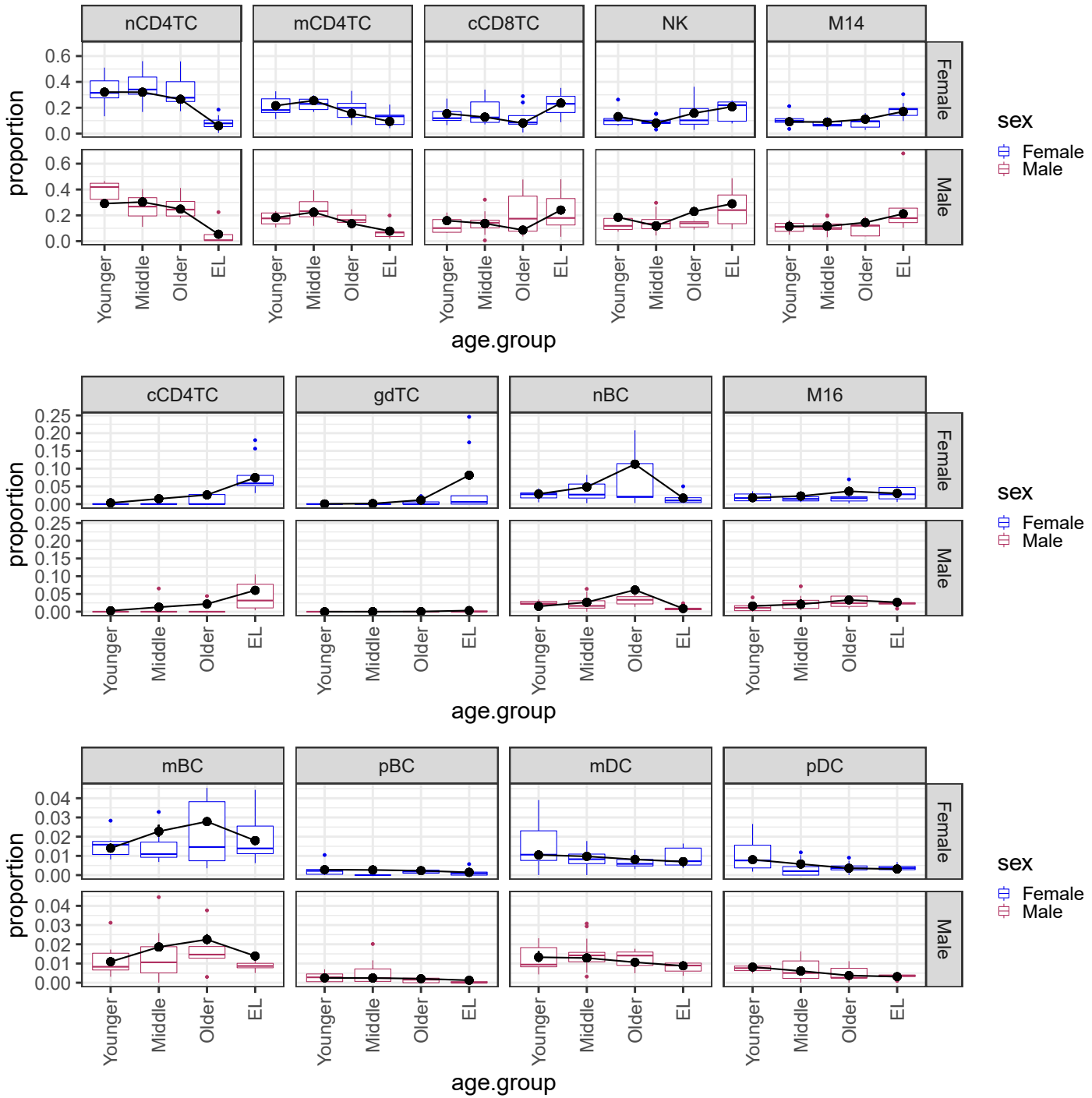
ZHENG, Y., LIU, X., LE, W., XIE, L., LI, H., WEN, W., WANG, S., MA, S., HUANG, Z., YE, J., ET AL. (2020). A human circulating immune cell landscape in aging and COVID-19. *Protein & Cell* **11**, 740–770.

ZHU, L., YANG, P., ZHAO, Y., ZHUANG, Z., WANG, Z., SONG, R., ZHANG, J., LIU, C., GAO, Q., XU, Q., 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.

Table 1. Table of significant cell type specific credible changes in composition between EL and younger age based on the Bayesian multinomial regression and scCODA.

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.

cell types	scCODA: number of credible changes	scCODA: percent of credible changes	scCODA: credible effect	multinomial: credible effect
nCD4TC	12	0.923077	TRUE	TRUE
mCD4TC	11	0.846154	TRUE	TRUE
cCD4TC	12	0.923077	TRUE	TRUE
cCD8TC	2	0.153846	FALSE	FALSE
gdTC	0	0	FALSE	TRUE
nBC	10	0.769231	TRUE	TRUE
mBC	0	0	FALSE	FALSE
pBC	0	0	FALSE	TRUE
NK	2	0.153846	FALSE	FALSE
M14	2	0.153846	FALSE	TRUE
M16	0	0	FALSE	FALSE
mDC	0	0	FALSE	TRUE
pDC	NaN	NaN	FALSE	TRUE


nCD4TC
mCD4TC
cCD8TC
NK
M14
cCD4TC
gdTC
nBC
M16
mBC
pBC
mDC
pDC
Female
Male
proportion
age.group
Younger
Middle
Older
EL
sex
Female
Male