Estimation of Eosinophil Cells in Cord Blood with References Based on Blood in Adults via Bayesian Measurement Error Modeling

Yu Jiang1, Hongmei Zhang1*, Shan V. Andrews2, Hasan Arshad3, Susan Ewart4, John W. Holloway5, M. Daniele Fallin6, Kelly M. Bakulski7, Wilfried Karmaus1,

1Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN, USA, 2Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA, 3Clinical and Experimental Sciences, University of Southampton, Southampton, UK, 4Department of Large Animal Clinical Sciences, Michigan State University, East Lansing, MI, USA, 5Human Development and Health, University of Southampton, Southampton, UK, 6Department of Mental Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA, 7Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX
Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Eosinophils are phagocytic white blood cells with a variety of roles in the immune system. In situations where actual counts are not available, high quality approximations of their cell proportions using indirect markers are critical.

Results: We develop a Bayesian measurement error model to estimate proportions of eosinophils in cord blood, using the cord blood DNA methylation profiles, based on markers of eosinophil cell heterogeneity in blood of adults. The proposed method can be directly extended to other cells across different reference panels. We demonstrate the method’s estimation accuracy using B cells and show that the findings support the proposed approach. The method has been incorporated into the estimateCellCounts function in the minfi package to estimate eosinophil cells proportions in cord blood.

Availability: estimateCellCounts function is implemented and available in Bioconductor package minfi.

Contact: hzhang6@memphis.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Eosinophils are white blood cells that perform a variety of functions, e.g., helping combat multicellular parasites and certain infections in vertebrates. Eosinophils are implicated in various diseases including asthma and allergy. The proportion of eosinophils relative to other cells in umbilical cord blood has the ability to predict respiratory illnesses in high risk infants (Junge, Hornig et al. 2014, Berek 2016). It is important to correctly identify eosinophils and estimate their proportions in cord blood.

To estimate eosinophil counts in biological samples, cell counting is often needed. However, this is not feasible in many settings due to the low proportion of eosinophils in the blood, limitations of fresh cord blood samples, and difficulty in discriminating eosinophils from other granulocytes (Ethier, Lacy et al. 2014). An indirect method for estimating cell compositions, which uses DNA methylation (DNAm) profiles to infer cell proportions, has been developed (Houseman, Kim et al. 2015) and incorporated into an R package minfi. This approach estimates eosinophil proportions in peripheral blood in adults. However, estimation bias is noted when the method is applied to samples collected from cord blood or other tissues due to the lack of eosinophil-specific reference data (Aryee, Jaffe et al. 2014). We develop a Bayesian measurement error model aiming to correct the bias in the estimation of eosinophil cell proportions in cord blood when using a reference database based on blood in adults.

2 Methods

Let n denote the number of cord blood samples, and \( p_i \) be an estimated cell proportion of eosinophils using an incorrect reference profile, e.g., determined based on peripheral blood in adults, for the ith sample, \( i = 1, 2, \ldots, n \). For the purpose of model fitting, we apply logit transformation to \( p_i \), \( y_i = \log(p_i/(1-p_i)) \). Let \( \mu_i \) be the cell proportion of eosinophils in cord blood for sample i, inferred based on the correct reference data from cord blood samples. We define the following measurement error model:

\[
    y_i = \mu_i + \nu + \epsilon_i ,
\]

where \( \nu \) is the size of systematic error and \( \epsilon_i \sim N(0, \sigma^2) \). The parameter \( \sigma^2 \) determines the size of measurement error at the individual level. We assume that the prior distribution of \( \sigma^2 \) is a uniform distribution with lower and upper bounds \( a \) and \( b \), respectively. The values of \( a \) and \( b \) need to be pre-specified (Section 2 in the Supplementary file). For \( \mu_i \), we choose a

© The Author(s) (2019). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
flat prior distribution, $\mu \sim N(0, s^2)$, with variance $s^2$ large and known. In the current study, we set $s^2 = 10^4$. The prior distribution of systematic error $v$ is assumed to be normally distributed, $N(c, d^2)$, with $c$ and $d^2$ known and determined based on data from six existing studies (Sections 2 and 3 in the Supplementary file). We use the Gibbs sampler to estimate all the unknown parameters: 1) Sampling the conditional posterior distribution of $\mu_i$ conditional on data and other parameters (denoted as “rest”), $\mu_i | \text{rest} \sim N\left(\frac{\bar{y}_i - \hat{\mu}}{\hat{\sigma}^2}, \frac{1}{\hat{\sigma}^2}\right)$; 2) sampling the conditional posterior distribution of $v$, $v|\text{rest} \sim N\left(\frac{\sum_{i=1}^{n} (y_i - \hat{\mu} - v_i) \bar{v}}{\bar{v}^2}, \frac{\bar{v}^2}{\bar{v}^2}\right)$; and 3) sampling the conditional posterior distribution of $\sigma^2$, $\frac{1}{\sigma^2} | \text{rest} \sim \text{truncated Gamma}(\alpha, \beta)$, with $\alpha = \frac{n-1}{2}$ and $\beta = \sum_{i=1}^{n} (y_i - \hat{\mu} - v_i)^2/2$, with $a \leq \sigma^2 \leq b$.

As noted earlier, $\sigma^2$ and $\nu$ determine the size of measurement errors. Thus, the choice of prior parameters $a$, $b$, $c$, and $d$ is critical and should reasonably represent the range of errors. We select these parameters utilizing cell types that have both cord and adult blood references available. Here we briefly discuss the approach used to specify the four parameters and the details are in Section 2 of the Supplementary file. In the selection of parameters $a$ and $b$, using the existing method available in the minfi package (in the function estimateCellCounts), we are able to infer the cell type proportions for each sample using the reference profiles constructed based on cord blood (this gives $\mu_i$ for a cell) as well as the proportions via reference profiles based on blood in adults (this gives $y_i$ for that cell). Consequently, we are able to estimate the magnitude of measurement errors of each sample for each of the six cell types (by calculating the differences between $y_i$ and $\mu_i$). These measurement errors are then used to determine the values of $a$ and $b$ in the prior distribution of $\sigma^2$. To specify the prior parameters $c$ and $d$ objectively and informatively, we implement two strategies. When inferring $a$ and $b$, we use information on cells in the same or a similar category to eosinophils, which potentially increases the accuracy of inferred systematic error. Granulocytes are selected for this purpose since eosinophilic cells are a subset of this cell type. In addition, taking into account that systematic errors may vary between studies, we summarize measurement errors inferred from six independent studies.

To demonstrate the effectiveness of the proposed method, we use B cells and assess the accuracy of the estimated B cell proportions in six studies (Section 3 in the Supplementary material). We also examine the impact of different prior distributions of $v$ on the inference of cell proportions (Section 4 in the Supplementary file). The results show that using the proposed informative prior distributions improves the estimated cell proportions; it gives smaller mean squared errors and estimation bias compared to alternative priors, informative or none-informative.

3 Illustration of the R function for cell proportion estimation with correction

We have incorporated the Bayesian method for estimating cell proportions into the estimateCellCounts function in the minfi package. As shown in Fig. 1, the entire Bayesian estimation model and computation is in the background. The format of input data and output of the updated estimateCellCounts function are similar to the format of its previous version. It takes DNAm data in the format of a RGChannelSet as the input data and returns cell counts for all samples. To estimate the proportion of eosinophils, users should ensure that the cellTypes option includes "Eos", and the compositeCellType is defined as "CordBlood". The arguments in the estimateCellCounts function are specified as the following: estimateCellCounts (rgSet, compositeCellType = "CordBlood", cellTypes = c("CD8T", "CD4T", "NK", "Bcell", "Mono", "nRBC", "Gran", "Eos").

4 Summary

The proposed method has been built into the estimateCellCounts function in the minfi package. The use of the function is the same as that for the default cord blood estimateCellCounts function except that “Eos” needs to be specified in the cellTypes vector in order to infer its cell proportions. The added computational time is only 1.6 seconds (0.8% of total running time of estimateCellCounts) with the inclusion of the Bayesian measurement error modeling for an analysis of 70 samples. The current measurement error model is developed for the estimation of eosinophil cells. However, it is not restricted to this cell and can be directly applied to estimate proportions of other cell types as long as the cell has information in one reference database. In addition, the proposed method is not restricted to cells in cord blood. It can be extended to whole blood or other tissues when reference data are not available.

![Flowchart of cell proportion estimation in cord blood using estimateCellCounts function](image)

**Fig. 1.** Flowchart of eosinophil proportion estimation in cord blood using estimateCellCounts function in minfi package.

**Acknowledgements**

We are thankful to the computation support from the High Performance Computing (HPC) facility at the University of Memphis.

**Funding**

National Institutes of Health (NIH) R01AI121226 (PIs: Zhang, Holloway), R01HL132321 and R03HD092776 (PI: Karmaus), the start-up funds for Dr. Jiang from the school of Public Health at the University of Memphis. Dr. Bakulski is supported by NIH R01AG05546 (PIs: Bakulski, Ware), P30ES017885 (PI: Loehr-Caruso), R01ES025531 (PI: Fallin), and R01ES025574 (PI: Schmidt).

**Conflict of Interest:** none declared.

**References**


