Supplementary Note: Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution

Maarten van Iterson[1], Erik van Zwet[2], the BIOS Consortium[3], P. Eline Slagboom[1], and Bastiaan T. Heijmans[1]

[1]Leiden University Medical Center, Department of Molecular Epidemiology, Leiden, The Netherlands.
[2]Leiden University Medical Center, Department of Medical Statistics and Bioinformatics, Leiden, The Netherlands.
[3]the BIOS consortium

May 25, 2016

Contents

1 The BIOS consortium

The mission of the BIOS Consortium is to create a large-scale data infrastructure and to bring together BBMRI researchers focusing on integrative omics studies in Dutch Biobanks (https://www.bbmri.nl/?p=259).

Management Team: Bastiaan T. Heijmans (chair)[1], Peter A.C. 't Hoen[2], Joyce van Meurs[3], Rick Jansen[5], Lude Franke[6].

Cohort collection: Dorret I. Boomsma[7], René Pool[7], Jenny van Dongen[7], Jouke J. Hottenga[7] (Netherlands Twin Register); Marleen MJ van Greevenbroek[8], Coen D.A. Stehouwer[8], Carla J.H. van der Kallen[8], Casper G. Schalkwijk[8] (Cohort study on Diabetes and Atherosclerosis Maastricht); Cisca Wijmenga[6], Lude Franke[6], Sasha Zhernakova[6], Ettje F. Tigchelaar[6] (LifeLines Deep); P. Eline Slagboom[1], Marian Beekman[1], Joris Deelen[1], Diana van Heemst[9] (Leiden Longevity Study); Jan H. Veldink[10], Leonard H. van den Berg[10] (Prospective ALS

Study Netherlands); Cornelia M. van Duijn[4], Bert A. Hofman[11], Aaron Isaacs[4], André G. Uitterlinden[3] (Rotterdam Study).

Data Generation: Joyce van Meurs (Chair)[3], P. Mila Jhamai[3], Michael Verbiest[3], H. Eka D. Suchiman[1], Marijn Verkerk[3], Ruud van der Breggen[1], Jeroen van Rooij[3], Nico Lakenberg[1].

Data management and computational infrastructure: Hailiang Mei (Chair)[12], Maarten van Iterson[1], Michiel van Galen[2], Jan Bot[13], Dasha V. Zhernakova[5], Rick Jansen[4], Peter van 't Hof[12], Patrick Deelen[5], Irene Nooren[13], Peter A.C. 't Hoen[2], Bastiaan T. Heijmans[1], Matthijs Moed[1].

Data Analysis Group: Lude Franke (Co-Chair)[6], Martijn Vermaat[2], Dasha V. Zhernakova[6], René Luijk[1], Marc Jan Bonder[6], Maarten van Iterson[1], Patrick Deelen[6], Freerk van Dijk[14], Michiel van Galen[2], Wibowo Arindrarto[12], Szymon M. Kielbasa[15], Morris A. Swertz[14], Erik. W van Zwet[15], Rick Jansen[5], Peter-Bram 't Hoen (Co-Chair)[2], Bastiaan T. Heijmans (Co-Chair)[1].

1. Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
2. Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
3. Department of Internal Medicine, ErasmusMC, Rotterdam, The Netherlands
4. Department of Genetic Epidemiology, ErasmusMC, Rotterdam, The Netherlands
5. Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands
6. Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands
7. Department of Biological Psychology, VU University Amsterdam, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands
8. Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, The Netherlands
9. Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands
10. Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands
11. Department of Epidemiology, ErasmusMC, Rotterdam, The Netherlands
12. Sequence Analysis Support Core, Leiden University Medical Center, Leiden, The Netherlands
13. SURFsara, Amsterdam, the Netherlands
14. Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
15. Medical Statistics Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

2 The genomic inflation factor is affect by true associations

Given a set of test-statistics $z_2, \cdots, z_p$ the (squared) genomic inflation factor is given by[1]:

$$\lambda^2 = \frac{\text{med}\{z_2^2, \cdots, z_p^2\}}{0.457}. \quad (1)$$

The median of the squared test-statistics will be the ordered test-statistic at position $p/2$ or $(p+2)/2$, if p is odd or even, respectively. Since, the set of test-statistics represents p_1 test-statistics following the null distribution and $p - p_1$ the alternative. The set ordered test-statistics will be given by $\{z_2^2, \cdots, z_{p_1}^2, z_{p_1+2}^2 \cdots, z_p^2\}$. Furthermore, it is known in advance that $p_1 > p/2$ or $p_1 > p+1/2$ it follows that $\text{med}\{z_2^2, \cdots, z_p^2\} > \text{med}\{z_2^2, \cdots, z_{p_1}^2\} > 0.457$ and thus $\lambda^2 > 1$.

3 Applying genomic control is the same as using an inflated or overdispersed empirical null

Consider a two-sided test for normally distributed test-statistics, T_i for $i = 1, \cdots, p$. Genomic control divides test-statistics by the inflation factor, λ, before the calculation of P values, U_i.

$$\begin{aligned} U_i &= 2\left[1 - \Phi\left(\left|\frac{T_i}{\lambda}\right|\right)\right] \\ &= 2\left[1 - \Pr\left\{Z \le \left|\frac{T_i}{\lambda}\right|\right\}\right] \\ &= 2\left[1 - \Pr\{\lambda Z \le |T_i|\}\right] \\ &= 2\left[1 - \Pr\{X \le |T_i|\}\right] \end{aligned} \tag{2}$$

Here, $Z \sim \mathcal{N}(0,1)$ with CDF given by Φ. Furthermore, $X \sim \mathcal{N}(0,\lambda^2)$, represents the overdispersed or inflation normal distribution.

And in general if both inflation and bias are present.

$$\begin{aligned} U_i &= 2\left[1 - \Phi\left(\left|\frac{T_i - \mu}{\sigma}\right|\right)\right] \\ &= 2\left[1 - \Pr\{\sigma Z + \mu \le |T_i|\}\right] \\ &= 2\left[1 - \Pr\{X \le |T_i|\}\right] \end{aligned} \tag{3}$$

Here, μ, represents the bias and σ the inflation. Furthermore, now $X \sim \mathcal{N}(\mu, \sigma^2)$.

4 Unobserved covariates introduce inflation and bias

In a note from P. Rao[2] it was shown that the omission of a variable introduces bias and decreases the variance of all least squares estimates, i.e. introduces both bias and inflation. Here is a sketch of the proof for the introduction of bias, within the framework of the main text.

Considered the omission of the unobserved technical or biological covariates, $\mathbf{W}$. For the sake of simplicity, we assume there are no known covariates, $\mathbf{Z}$, without loss of generality.

$$\begin{aligned} \mathbf{y}_j &= \mathbf{x}\tilde{\beta}_j + \tilde{\varepsilon}_{.j} \\ \mathbf{y}_j &= \mathbf{x}\beta_j + \mathbf{W}\gamma_j + \varepsilon_{.j} \end{aligned} \tag{4}$$

The latter model is true but we are unaware of this and continue estimating the regression coefficient of interest, $\tilde{\beta}_j$ of the former, misspecified, model.

$$\begin{aligned} \tilde{b_j} &= \frac{\mathbf{x}^T \mathbf{y}_j}{\mathbf{x}^T \mathbf{x}} \\ &= \frac{\mathbf{x}^T (\mathbf{x}\beta_j + \mathbf{W}\gamma_j + \varepsilon_{.j})}{\mathbf{x}^T \mathbf{x}} \\ &= \beta_j + \frac{\mathbf{x}^T \mathbf{W}\gamma_j + x^T \varepsilon_{.j}}{\mathbf{x}^T \mathbf{x}} \end{aligned} \tag{5}$$

where, we substituted the true model for $\mathbf{y}_j$. Now, since $E[\varepsilon_{.j}] = 0$, the expected regression coefficient is given by:

$$E[\tilde{b}_j] = \beta_j + \frac{\mathbf{x}^T \mathbf{W}}{\mathbf{x}^T \mathbf{x}} \gamma_j \tag{6}$$

and the bias is given by the last fraction, which can be interpreted as a weighted sum of correlations between the covariate of interest, $\mathbf{x}$ and q omitted variables.

$$\sum_{k}^{q} \gamma_{jk} \mathrm{cor}(\mathbf{x}, \mathbf{w}_k) \tag{7}$$

If all weights are zero or all correlations than the bias will be zero to. The bias is not equal to zero if an omitted variable is confounded with the outcome, i.e. both γ_{jk} and $\mathbf{x}^T \mathbf{W}$ are not zero.

5 Partial least squares for imputation of white blood cell composition

White blood cell counts (WBC), i.e., neutrophils, lymphocytes, monocytes, eosinophils and basophils, were measures by the standard WBC differential as part of the CBC (Complete Blood Count). However, a minority of samples lack CBC measurements. Since DNA methylation levels are informative of the white blood cell composition[3] we build a linear predictor to infer the white blood cell composition of those samples lacking WBC measurements.

Obviously, this model can not be fitted using ordinary least-squares, since $p >> n$, we need some kind of regularization. Furthermore, the multivariate responses, white blood counts on five cell types, represents compositional data, i.e., the data are percentages that sum up to 100%.

We have chosen to use partial least-squares for fitting a model with cell counts as a multivariate response and the > 400.000 CpGs age and sex as covariates. It is known that the WBCC is dependent on age and gender. The advantage of partial least-squares is that it both can handle multivariate responses and high-dimensional ($p >> n$) covariates. We used the R-package pls[4] to fit the model and optimize the number of pls-components using five-fold cross-validation. The fitted model was used to predict WBCC using the 450K data, age and sex of those samples lacking WBCC.

The pls-approach has been validated by splitting the data with WBCC available in a train and test set. Fit the pls-model on the train set and predict WBCC on the test set. Correlation (Pearson) between predicted and measured WBCC range from 0.86 – 0.37 for lymphocytes and basophils respectively, the intraclass correction was 0.84 – 0.25.

This approach has been implemented in a R package https://github.com/mvaniterson/wbccPredictor.

6 Different ways to calculate genomic inflation

Devlin and Roeder [1] propose the genomic inflation factor is as the ratio of the median of χ^2_1-distributed test-statistics divided by the median of a χ^2_1 of 0.456.

$$\lambda = \frac{\mathrm{median}(t_1, t_2, \cdots, t_p)}{0.456}, \tag{8}$$

where $t_i \sim \chi^2_2$ at least under the null hypothesis.

Since, in EWAS/TWAS usually test-statistics, $z_1, z_2, \cdots, z_p$ are derived that follow (approximately) a normal distribution the following formula can be used to estimate the amount of inflation:

$$\lambda^2 = \frac{\mathrm{median}(z_1^2, z_2^2, \cdots, z_p^2)}{0.456}, \tag{9}$$

since, z_i^2 will approximately follow a χ^2_1-distribution. Furthermore, the inflation of the z-scale is the square-root of the χ^2_2-scale to indicate this we introduced the λ^2.

Another way to estimate the amount of inflation is using the absolute value of z-score and divide them by 0.456^2.

Occasionally, P values are used to estimate the amount of inflation by comparing median of the minus $\log_{10}$-transformed P values with the median of minus $\log_{10}$-transformed random uniformly distributed statistics. However, a simple calculation shows these follow an exponential distribution with rate parameter $\log_e 10$ and thus the theoretical median is $\log_{10} 2$.

References

[1] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4), 1999.

[2] P. Rao. Some notes on misspecification in multiple regression. *The American Statistican*, 25(5), 1971.

[3] E.A. Houseman et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 8(13), 2012.

[4] Wehrens R. Mevik, B.-H. The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, 18(2), 2007.