OPEN

Publisher Correction: Robustification of GWAS to explore effective SNPs addressing the challenges of hidden population stratification and polygenic effects

Zobaer Akond, Md. Asif Ahsan, Munirul Alam & Md. Nurul Haque Mollah

Correction to: *Scientific Reports* <https://doi.org/10.1038/s41598-021-90774-7>, published online 22 June 2021

The original version of this Article contained typographical errors in the Equations.

In Table 1, fourth column, the Equation “Polygenic effect variation” was incorrect:

$$\text{var} \left(\sum_{k=m_1+1}^{m_2} b_k Z_k \right)$$

now reads:

$$\text{var} \left(\sum_{k=m_1+1}^{m_2} b_k Z_k \right)$$

In the Materials and methods section, under the subheading “Robustification of LMM based GWAS by using the outlier modification rule (proposed)”,

“ $\mathbf{b}$ is the vector of random polygenic effects which follows $N(0, \sigma_g^2 K)$, where σ_g^2 is the polygenic variance component, and $K = (k_{ij})$ is the $m \times m$ genomic relationship matrix.”

now reads:

“ $\mathbf{b}$ is the vector of random polygenic effects which follows $N(0, \sigma_g^2 K)$, where σ_g^2 is the polygenic variance component, and $K = (k_{ij})$ is the $m \times m$ genomic relationship matrix.”

And,

“When the maximum likelihood (ML) or restricted maximum likelihood (REML) variance component $\hat{\mathbf{V}} = \hat{\sigma}_g^2 \mathbf{K} + \hat{\sigma}_e^2 \mathbf{I}$ is estimated, the classical F -statistic for testing the null hypothesis $Ma = 0$ for an arbitrary full-rank $p \times q$ matrix $M^{13,60}$.”

now reads:

“When the maximum likelihood (ML) or restricted maximum likelihood (REML) variance component $\hat{\mathbf{V}} = \hat{\sigma}_g^2 \mathbf{ZKZ}' + \hat{\sigma}_e^2 \mathbf{I}$ is estimated, the classical F -statistic for testing the null hypothesis $Ma = 0$ for an arbitrary full-rank $p \times q$ matrix $M^{13,60}$ ”

And,

“(ii) Divide the phenotypic data into m groups corresponding to the m genotypic labels of the selected most significant SNP. For example, let

$$\mathbf{y} = (y_1, y_2, \dots, y_n)' = (y_{11}, \dots, y_{1n_1}, \dots, y_{m1}, \dots, y_{mn_m})'$$

be the partition of phenotypic observations corresponding to the selected SNP, where, $n = n_1 + n_2 + \dots + n_k$.

(iii) Detect the outlying observations from the l^{th} ($l = 1, 2, \dots, k$) group using the β -weight function defined by

$$W_\beta(y_{li}|\hat{\theta}_l) = \exp\left\{-\frac{\beta}{2\sigma_{li}^2}(y_{li} - \hat{\mu}_l)^2\right\} \quad (4)$$

where $i = 1, 2, \dots, n$ ”

now reads:

“(ii) Divide the phenotypic data into m groups corresponding to the m genotypic labels of the selected most significant SNP. For example, let

$$\mathbf{y} = (y_1, y_2, \dots, y_n)' = (y_{11}, \dots, y_{1n_1}, \dots, y_{m1}, \dots, y_{mn_m})'$$

be the partition of phenotypic observations corresponding to the selected SNP, where, $n = n_1 + n_2 + \dots + n_m$.

(iii) Detect the outlying observations from the l^{th} ($l = 1, 2, \dots, k$) group using the β -weight function defined by

$$W_\beta(y_{li}|\hat{\theta}_l) = \exp\left\{-\frac{\beta}{2\sigma_{li}^2}(y_{li} - \hat{\mu}_l)^2\right\} \quad (4)$$

where $i = 1, 2, \dots, n_l$ ”

Additionally, Equation 7 was incorrect:

“An outlying phenotypic observation y_{lk} in the l^{th} group is defined based on the β -weight function mentioned below:

$$W_\beta(x_{jk}|\hat{\theta}_{j,\beta}) = \begin{cases} > \tau_j, & \text{if } x_{jk} \text{ is not an outlier} \\ \leq \tau_j, & \text{if } x_{jk} \text{ is an outlier} \end{cases} \quad (7)$$

where the threshold value τ_j is the p^{th} quantile value of the empirical distribution of $W_\beta(x_{li}|\hat{\theta}_{l,\beta})$.”

now reads:

“An outlying phenotypic observation y_{li} in the l^{th} group is defined based on the β -weight function mentioned below:

$$W_\beta(y_{li}|\hat{\theta}_{l,\beta}) = \begin{cases} > \tau_j, & \text{if } y_{li} \text{ is not an outlier} \\ \leq \tau_j, & \text{if } y_{li} \text{ is an outlier} \end{cases} \quad (7)$$

where the threshold value τ_j is the p^{th} quantile value of the empirical distribution of $W_\beta(y_{li}|\hat{\theta}_{l,\beta})$.”

In Table 3, fourth column, the Equation “Polygenic effect variation” was incorrect:

$$\text{var} \left(\sum_{j=1}^{m_2} b_k x_k \right)$$

now reads:

$$\text{var} \left(\sum_{k=m_1+1}^{m_2} b_k z_k \right)$$

Under the subheading “Genotype simulation”,

“For this purpose, $m = 2000$ SNPs were generated for $n = 1000$ individuals, and these individuals were taken from $k = 3$ distinct population by considering different minor allele frequencies (MAFs). To do this, first, a set of m latent vectors $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$ was generated from a multivariate normal distribution with mean zero and variance–covariance matrix $\text{Cov}(\mathbf{v}_j, \mathbf{v}_k) = \rho^{|j-k|^{64,65}}$. In our simulation, we considered $\rho = 0.5$ to avoid the linkage disequilibrium (LD) between the SNPs. Finally, two cutoff values s_1 and s_2 were used to convert the design matrix $V = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m] = [v_{ij}]$ of latent vectors to the genotypic score matrix $x_{ij} (i = 1, 2, \dots, n, j = 1, 2, \dots, m_1)$ and $z_{ij} (i = m_1 + 1, \dots, n, j = 1, 2, \dots, m_2)$ as follows:

$$x_{ij}, z_{ij} = \begin{cases} -1, & v_{ij} < s_1 \\ 0, & s_1 \leq v_{ij} \leq s_2 \\ 2, & v_{ij} > s_2 \end{cases}$$

where s_1 and s_2 determine the minor allele frequency.”

now reads:

“For this purpose, $m^* = 2000$ SNPs were generated for $n = 1000$ individuals, and these individuals were taken from $k = 3$ distinct population by considering different minor allele frequencies (MAFs). To do this, first, a set of latent vectors $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{m^*}\}$ was generated from a multivariate normal distribution with mean zero and variance–covariance matrix $\text{Cov}(\mathbf{v}_j, \mathbf{v}_k) = \rho^{|j-k|^{64,65}}$. In our simulation, we considered $\rho = 0.5$ to avoid the linkage disequilibrium (LD) between the SNPs. Finally, two cutoff values s_1 and s_2 were used to convert the design matrix $V = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{m^*}] = [v_{ij}]$ of latent vectors to the genotypic score matrix $x_{ij} (i = 1, 2, \dots, n, j = 1, 2, \dots, m_1)$ and $z_{ij} (i = 1, 2, \dots, n; j = 1, 2, \dots, m_2)$ as follows:

$$x_{ij}, z_{ij} = \begin{cases} 0, & v_{ij} < s_1 \\ 1, & s_1 \leq v_{ij} \leq s_2 \\ 2, & v_{ij} > s_2 \end{cases}$$

where s_1 and s_2 determine the minor allele frequency.”

Furthermore, in the section “Phenotype simulation”,

“In every situation, $m_1 = 4$ SNPs were considered as causal variants and the remaining $m_2 = m - m_1 = 1996$ SNPs were allocated as polygenic variants (effects).”

now reads:

“In every situation, $m_1 = 4$ SNPs were considered as causal variants and the remaining $m_2 = m^* - m_1 = 1996$ SNPs were allocated as polygenic variants (effects).”

Lastly, in the section “Consequence of phenotypic outliers on the partition of total variations”,

$$h = \frac{\text{var} \left(\sum_{k=1}^{m_1} a_k x_k \right)}{\text{var}(y)} > \frac{\text{var} \left(\sum_{k=1}^{m_1} a_k x_k \right)}{\text{var}(y^*)} = h^{2*}$$

now reads:

$$h^2 = \frac{\text{var}\left(\sum_{k=1}^{m_1} a_k x_k\right)}{\text{var}(y)} > \frac{\text{var}\left(\sum_{k=1}^{m_1} a_k x_k\right)}{\text{var}(y^*)} = h^{2*}$$

The original Article has been corrected.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021