Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort

Florian Privé,* Hugues Aschard, Shai Carmi, Lasse Folkersen, Clive Hoggart, Paul F. O'Reilly, and Bjarni J. Vilhjálmsson

(The American Journal of Human Genetics *109*, 12–23; January 6, 2022)

An unfortunate corruption of two equations on page 14 appeared in the version of this paper published on January 6. It has been corrected here and online. The publisher apologizes for this error.

New formula used in LDpred2

We also slightly modify the formula used in Privé et al.;[32] we have previously used

$$\mathrm{se}\left(\widehat{\gamma}_j\right)^2 = \frac{\left(\breve{\boldsymbol{y}} - \widehat{\gamma}_j \breve{\boldsymbol{G}}_{\boldsymbol{j}}\right)^T \left(\breve{\boldsymbol{y}} - \widehat{\gamma}_j \breve{\boldsymbol{G}}_{\boldsymbol{j}}\right)}{(n-K-1)\breve{\boldsymbol{G}}_{\boldsymbol{j}}^T \breve{\boldsymbol{G}}_{\boldsymbol{j}}} \approx \frac{\breve{\boldsymbol{y}}^T \breve{\boldsymbol{y}}}{n \breve{\boldsymbol{G}}_{\boldsymbol{j}}^T \breve{\boldsymbol{G}}_{\boldsymbol{j}}} \approx \frac{\mathrm{var}(\boldsymbol{y})}{n\ \mathrm{var}\left(\boldsymbol{G}_{\boldsymbol{j}}\right)},$$

where $\widehat{\gamma}_j$ is the marginal effect of variant *j*, and where $\breve{\boldsymbol{y}}$ and $\breve{\boldsymbol{G}}_{\boldsymbol{j}}$ are the vectors of phenotypes and genotypes for variant *j* residualized from *K* covariates, e.g., centering them. The first approximation expects $\widehat{\gamma}_j$ to be small, while the second approximation assumes the effects from covariates are small. However, we have found here that some variants can have very large effects, e.g., one variant explains about 30% of the variance in bilirubin log-concentration. Then, instead we compute

$$\left(\breve{\boldsymbol{y}} - \widehat{\gamma}_j \breve{\boldsymbol{G}}_{\boldsymbol{j}}\right)^T \left(\breve{\boldsymbol{y}} - \widehat{\gamma}_j \breve{\boldsymbol{G}}_{\boldsymbol{j}}\right) = \breve{\boldsymbol{y}}^T \breve{\boldsymbol{y}} - 2\widehat{\gamma}_j \breve{\boldsymbol{G}}_{\boldsymbol{j}}^T \breve{\boldsymbol{y}} + \widehat{\gamma}_j^2 \breve{\boldsymbol{G}}_{\boldsymbol{j}}^T \breve{\boldsymbol{G}}_{\boldsymbol{j}}$$
$$= \breve{\boldsymbol{y}}^T \breve{\boldsymbol{y}} - \widehat{\gamma}_j^2 \breve{\boldsymbol{G}}_{\boldsymbol{j}}^T \breve{\boldsymbol{G}}_{\boldsymbol{j}},$$

which now gives

$$(n-K-1)\mathrm{se}\left(\widehat{\gamma}_j\right)^2 = \frac{\breve{\boldsymbol{y}}^T \breve{\boldsymbol{y}} - \widehat{\gamma}_j^2 \breve{\boldsymbol{G}}_{\boldsymbol{j}}^T \breve{\boldsymbol{G}}_{\boldsymbol{j}}}{\breve{\boldsymbol{G}}_{\boldsymbol{j}}^T \breve{\boldsymbol{G}}_{\boldsymbol{j}}}$$
$$= \frac{\breve{\boldsymbol{y}}^T \breve{\boldsymbol{y}}}{\breve{\boldsymbol{G}}_{\boldsymbol{j}}^T \breve{\boldsymbol{G}}_{\boldsymbol{j}}} - \widehat{\gamma}_j^2 \approx \frac{\mathrm{var}(\breve{\boldsymbol{y}})}{\mathrm{var}\left(\boldsymbol{G}_{\boldsymbol{j}}\right)} - \widehat{\gamma}_j^2,$$

finally giving (note the added term $\widehat{\gamma}_j^2$)

$$\mathrm{sd}\left(\boldsymbol{G}_{\boldsymbol{j}}\right) \approx \frac{\mathrm{sd}(\breve{\boldsymbol{y}})}{\sqrt{n\ \mathrm{se}\left(\widehat{\gamma}_j\right)^2 + \widehat{\gamma}_j^2}}. \qquad \text{(Equation 1)}$$

Figure S23 shows that the updated formula Equation 1 is better; we now use it in the code of LDpred2, and also recommend using it for the QC procedure proposed in Privé et al.[32]

*Correspondence: florian.prive.21@gmail.com

https://doi.org/10.1016/j.ajhg.2022.01.007.