**Supplementary Information for “Phenome-wide Heritability Analysis of the UK Biobank”**

Tian Ge1,2,3, Chia-Yen Chen2,3,4, Benjamin M. Neale2,3,4, Mert R. Sabuncu1,5,\*, and Jordan W. Smoller2,3,\*

1Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital / Harvard Medical School, Charlestown, MA 02129, USA;

2Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA;

3Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02138, USA;

4Analytic and Translational Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA;

5School of Electrical and Computer Engineering and School of Biomedical Engineering, Cornell University, Ithaca, NY 14853, USA.

\*MRS and JWS contributed equally to this work.

Correspondence to: tge1@mgh.harvard.edu (TG) or jsmoller@hms.harvard.edu (JWS)

**The Moment-matching Method for SNP Heritability Estimation**

We consider the linear random effect model $y=g+e$, where an $N$-dimensional trait $y$is partitioned into the sum of additive genetic effects $g$ and unique (subject-specific) environmental effects $e$. The covariance structure of $y$ is $cov\left[y\right]=σ\_{g}^{2}K+σ\_{e}^{2}I\_{N}$, where $K$ is the empirical genetic similarity matrix for each pair of individuals estimated from genome-wide SNP data, $I\_{N}$is an $N×N$ identity matrix, $σ\_{g}^{2}$ and $σ\_{e}^{2}$ are the total additive genetic variance captured by genotyped common SNPs and the variance of unique environmental factors across individuals, respectively.

To obtain unbiased estimates of $σ\_{g}^{2}$ and $σ\_{e}^{2}$, we regress the empirical estimate of the phenotypic covariance onto the matrices $K$and $I$: $vec\left[yy^{T}\right]=σ\_{g}^{2}vec\left[K\right]+σ\_{e}^{2}vec\left[I\right]+ϵ$, where $vec[∙]$ is the matrix vectorization operator that converts a matrix into a vector by stacking its columns, and $ϵ$is the residual of the regression. The ordinary least squares (OLS) estimator of this multiple regression problem can be obtained by solving the linear system:

$$\left[\begin{matrix}tr\left[K^{2}\right]&tr\left[K\right]\\tr\left[K\right]&N\end{matrix}\right]\left[\begin{matrix}σ\_{g}^{2}\\σ\_{e}^{2}\end{matrix}\right]=\left[\begin{matrix}y^{T}Ky\\y^{T}y\end{matrix}\right].$$

In the presence of covariates, i.e., $y=Xβ+g+e$, where $X$ is an $N×q$ covariate matrix and $β$ is a vector of fixed effects, an $N×(N-q)$ matrix $U$ always exists, which satisfies $U^{T}U=I$, $UU^{T}=P\_{0}$, $U^{T}X=0$, and $P\_{0}=I-X(X^{T}X)^{-1}X^{T}$. Applying $U^{T}$to both sides of the model removes the covariate matrix and gives $U^{T}y=U^{T}g+U^{T}e$. The covariance structure of the transformed trait is $cov\left[U^{T}y\right]=σ\_{g}^{2}U^{T}KU+σ\_{e}^{2}I\_{N-q}$, and the linear system becomes

$$\left[\begin{matrix}tr\left[P\_{0}KP\_{0}K\right]&tr\left[P\_{0}K\right]\\tr\left[P\_{0}K\right]&N-q\end{matrix}\right]\left[\begin{matrix}σ\_{g}^{2}\\σ\_{e}^{2}\end{matrix}\right]=\left[\begin{matrix}y^{T}P\_{0}KP\_{0}y\\y^{T}P\_{0}y\end{matrix}\right].$$

We note that for large sample size $N$, the $N×N$ genetic similarity matrix $K$and the $N×N$ residual forming matrix $P\_{0}$can be very large, making the computation of $y^{T}P\_{0}y$,$y^{T}P\_{0}KP\_{0}y$,$tr\left[P\_{0}K\right]$, and $tr\left[P\_{0}KP\_{0}K\right]$ memory intensive. To reduce the memory demand, we note that (1) the number of covariates $q$ is typically orders of magnitude smaller than $N$; (2) the quantities $y^{T}K$,$X^{T}K$, $tr\left[K\right]=\sum\_{i}^{}k\_{ii}$, and $tr\left[K^{2}\right]=\sum\_{ij}^{}k\_{ij}^{2}$ can be computed by iteratively reading columns (or block columns) of $K$into the memory; and (3) we have the following derivations:

$$y^{T}P\_{0}y=y^{T}y-y^{T}X\left(X^{T}X\right)^{-1}X^{T}y,$$

$$tr\left[P\_{0}K\right]= tr\left[K\right]-tr\left[X^{T}KX\left(X^{T}X\right)^{-1}\right],$$

$$y^{T}P\_{0}KP\_{0}y=y^{T}Ky-2y^{T}X\left(X^{T}X\right)^{-1}X^{T}Ky+y^{T}X\left(X^{T}X\right)^{-1}X^{T}KX\left(X^{T}X\right)^{-1}X^{T}y,$$

$$tr\left[P\_{0}KP\_{0}K\right]=tr\left[K^{2}\right]-2tr\left[\left(X^{T}X\right)^{-1}X^{T}KKX\right]+tr\left[X\left(X^{T}X\right)^{-1}X^{T}KX\left(X^{T}X\right)^{-1}X^{T}K\right].$$

Therefore, it can be seen that once $y^{T}K$,$X^{T}K$, $tr\left[K\right]=\sum\_{i}^{}k\_{ii}$, and $tr\left[K^{2}\right]=\sum\_{ij}^{}k\_{ij}^{2}$ have been computed by iteratively loading columns (or block columns) of $K$into the memory, all quantities in the linear system can be computed without manipulating any $N×N$ matrix. In particular, the residual forming matrix $P\_{0}$does not need to be explicitly computed. This makes the moment-matching algorithm computationally and memory efficient even if the sample size is very large.