Heritability and Genetic Correlations Explained by Common SNPs for Metabolic Syndrome Traits: Revised Methods for multivariate model, May, 2012.

Citation

Please cite the manuscript describing the methods. The paper is publicly available at: `http://www.ncbi.nlm.nih.gov/pubmed/22479213`.

Vattikuti S, Guo J, Chow CC. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. PLoS Genet. 2012;8(3):e1002637.

Bivariate (multivariate) linear mixed-effects linear model

We considered the following multivariate linear mixed-effects model for m individuals, n loci, and t traits [1–6]:

$$\mathbf{y_i} = \mathbf{X_i v_i} + \mathbf{Z u_i} + \mathbf{e_i} \tag{1}$$

where i indexes one of t traits, $\mathbf{y_i}$ is the m dimensional phenotype vector of trait i, $\mathbf{X_i}$ is an $m \times s$ fixed effects matrix, $\mathbf{Z}$ is a trait invariant $m \times n$ matrix of standardized genotypes, $\mathbf{v_i}$ is an $s \times 1$ vector of fixed effects parameters, $\mathbf{u_i}$ is an $n \times 1$ vector of random genetic effects, and $\mathbf{e_i}$ is an $m \times 1$ vector of residual effects. The genetic effects vectors $\mathbf{u_i}$ obey the law $\mathbf{u_i} \sim \mathcal{N}(0,\mathbf{G})$, where $\mathbf{G}$ is the covariance matrix across all traits with matrix blocks $\mathbf{G_{ij}} = \mathrm{cov}_g(i,j)\mathbf{I_n}$ for traits i and j; the residual effects obey $\mathbf{e_i} \sim \mathcal{N}(0,\mathbf{R})$, where $\mathbf{R}$ is the covariance matrix across all traits with matrix blocks $\mathbf{R_{ij}} = \mathrm{cov}_e(i,j)\mathbf{I_m}$ for traits i and j, and $\mathbf{I_p}$ is the $p \times p$ identity matrix. This model can be used for single or multiple traits. For two traits, it is called a bivariate model. The model is identical to that used in [1, 2, 4].

We considered only biallelic SNPs in Hardy-Weinberg equilibrium. Denote the minor allele by q and the major allele by Q. Let the minor allele frequency at locus l (of n loci) have frequency p_l. We assign a value of 2 for genotype qq, 1 for genotype qQ, and 0 for genotype QQ. The Hardy-Weinberg mean frequency for the genotype at locus l is $2p_l$ and the variance is $2p_l(1-p_l)$. The standardized genotype entries have values of $\dfrac{2-2p_l}{\sqrt{2p_l(1-p_l)}}$ for qq, $\dfrac{1-2p_l}{\sqrt{2p_l(1-p_l)}}$ for qQ, and $\dfrac{-2p_l}{\sqrt{2p_l(1-p_l)}}$ for the QQ genotype.

The log of the likelihood function is given by

$$l = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{V}| - \sum_{i=1,j=1}^{t} (\mathbf{y_i} - \mathbf{X_i v_i})' {\mathbf{V_{ij}}}^{-1} (\mathbf{y_j} - \mathbf{X_j v_j}) \tag{2}$$

where the covariance matrix is $\mathbf{V} = \mathbf{G} \otimes \mathbf{A} + \mathbf{R} \otimes \mathbf{I}$, its inverse is written in terms of $m \times m$ blocks ${\mathbf{V_{ij}}}^{-1}$, and $\mathbf{A}$ is the genetic relationship matrix given by $\mathbf{A} = \dfrac{\mathbf{Z}\mathbf{Z}'}{n}$. Following Yang et al. [1] we used a modified covariance matrix for $\mathbf{A}$, where the diagonals of $\mathbf{A}$ are computed using the formula

$$\mathbf{A_{kk}} = \frac{1}{n}\sum_{l=1}^{n} \mathbf{Z_{kl}} \left(\mathbf{Z_{kl}} + \frac{2p_l - 1}{\sqrt{2p_l(1-p_l)}} \right) \tag{3}$$

where k is indexed over m subjects and l over n loci.

We use the restricted maximum likelihood (REML) approach [3] where the gradients of the log likelihood for traits i, j are given by

$$\frac{\delta l}{\delta \mathrm{cov}_g(i,j)} = \frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{A_{ij}}\mathbf{P}\mathbf{y} - \frac{1}{2}\mathrm{Tr}(\mathbf{P}\mathbf{A_{ij}}) \tag{4}$$

$$\frac{\delta l}{\delta \mathrm{cov}_e(i,j)} = \frac{1}{2}\mathbf{y'PI_{ij}Py} - \frac{1}{2}\,\mathrm{Tr}(\mathbf{PI_{ij}}) \tag{5}$$

where $\mathbf{I_{ij}}$ is a $t \cdot m \times t \cdot m$ dimensional matrix with zero entries except for an $m \times m$ identity matrix at block location ij, $\mathbf{A_{ij}} = \mathbf{A} \otimes \mathbf{I_{ij}}$, and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{W}(\mathbf{W'V}^{-1}\mathbf{W})^{-1}\mathbf{W'V}^{-1}$ where $\mathbf{W} = \mathbf{I_t} \otimes \mathbf{X_i}$ with $\mathbf{I_t}$ the identity matrix of $t \times t$ dimensions and $\mathbf{W}$ is a tiled transformation of $\mathbf{X_i}$.

We solved the REML equations using an EM algorithm [3] given by

$$(\mathrm{cov}_g(i,j))^{k+1} = (\mathrm{cov}_g(i,j))^k \left[1 + \frac{(\mathrm{cov}_g(i,j))^k}{m}\left(\mathbf{y'P}^k\mathbf{A_{ij}P}^k\mathbf{y} - \mathrm{Tr}(\mathbf{P}^k\mathbf{A_{ij}})\right)\right] \tag{6}$$

$$(\mathrm{cov}_e(i,j))^{k+1} = (\mathrm{cov}_e(i,j))^k \left[1 + \frac{(\mathrm{cov}_e(i,j))^k}{m}\left(\mathbf{y'P}^k\mathbf{I_{ij}P}^k\mathbf{y} - \mathrm{Tr}(\mathbf{P}^k\mathbf{I_{ij}})\right)\right] \tag{7}$$

for iteration k+1 in terms of iteration k. We iterated until the rate of change of the log likelihood function was less than about 10^{-4}. We also checked that the rate of change of the square of the covariance predictions was less than 10^{-8}. We checked our results against the software developed by Yang et al. (GCTA) [7] for the univariate model.

The multivariate model for 7 traits can be quite slow. To speed up the computation, we transformed to a coordinate system where the covariance matrices were diagonal [3] to speed up the computation. Let $\mathbf{z_k}$ be the set of phenotypes for individual k. We used the canonical transformation $\tilde{\mathbf{z}}_\mathbf{k} = \mathbf{Qz_k}$ such that $\mathbf{QGQ'} = \mathbf{\Lambda}$ and $\mathbf{QRQ'} = \mathbf{I_t}$. $\mathbf{Q}$ can be computed from the formula $\mathbf{Q} = \sqrt{\mathbf{SRS'}}\mathbf{S}^{-1}$ where $\mathbf{SGR}^{-1} = \mathbf{\Lambda S}$ ($\mathbf{S}$ is the matrix of left eigenvectors of $\mathbf{GR}^{-1}$). The transformed genetic covariances are given by $\mathbf{\Lambda}$ and the residual covariances are $\mathbf{I_t}$. Each step consisted of taking a single step with the univariate EM algorithm for the transformed additive genetic and residual variance followed by a transformation back to the original coordinates. We iterated until the maximum of the magnitudes of the components of the gradient of the log likelihood function was less than approximately 5×10^{-4} .

In our computations, we used both the direct EM algorithm and the canonically transformed algorithm because even though the transformed algorithm was in principle faster, it sometimes had poor convergence properties if the initial guess was not sufficiently close to the maximum likelihood value. We ensured that both give the same results. For computational efficiency, the results shown are computed from the bivariate model for the different trait pairs. We confirmed our results with a multivariate model that included all traits.

Our error estimates were given by the inverse of the Fisher information matrix $\mathbf{F}$, which we computed by evaluating the Hessian of the log likelihood at the maximum likelihood predictions. $\mathbf{F}$ is a symmetric $t(t+1) \times t(t+1)$ dimensional matrix with rows (columns) corresponding to the genetic and residual variances and covariances with block elements (that are not all contiguous) for traits i, j, k, and l given by

$$\mathbf{F_{ij,kl}} = \frac{1}{2}\begin{pmatrix} \mathrm{Tr}(\mathbf{PA_{ij}PA_{kl}}) & \mathrm{Tr}(\mathbf{PA_{ij}PI_{kl}}) \\ \mathrm{Tr}(\mathbf{PA_{ij}PI_{kl}}) & \mathrm{Tr}(\mathbf{PI_{ij}PI_{kl}}) \end{pmatrix} \tag{8}$$

for $i \le j$ and $k \le l$.

References

[1] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common snps explain a large proportion of the heritability for human height. Nat Genet 42: 565-9.

[2] Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, et al. (2011) Genome partitioning of genetic variation for complex traits using common snps. Nat Genet 43: 519-25.

[3] Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sunderland, Mass.: Sinauer.

[4] Deary IJ, Yang J, Davies G, Harris SE, Tenesa A, et al. (2012) Genetic contributions to stability and change in intelligence from childhood to old age. Nature 482: 212-5.

[5] Lee SH, Wray NR, Goddard ME, Visscher PM (2011) Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88: 294-305.

[6] Price AL, Helgason A, Thorleifsson G, McCarroll SA, Kong A, et al. (2011) Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genetics 7.

[7] Yang J, Lee SH, Goddard ME, Visscher PM (2011) Gcta: a tool for genome-wide complex trait analysis. Am J Hum Genet 88: 76-82.