Contents

0.1 Covariance estimation

The LIMIX framework enables multi-trait LMM analyses for more than two traits and provides methods to decompose phenotypic variance [?]. In single-variant LMMs, typically two phenotypic variance components are assumed: i) a genetic component g and ii) a noise component ψ (Equations **??** and **??**). To estimate the variance decomposition (VD) into g and ψ, the phenotype is described as the null model of the LMM:

$$\mathbf{y} = \mathbf{g} + \boldsymbol{\psi}, \ \mathbf{g} \sim \mathcal{N}\left(0, \mathbf{C_g} \otimes \mathbf{R}\right), \ \psi \sim \mathcal{N}\left(0, \mathbf{C_n} \otimes \mathbf{I_n}\right) \tag{1}$$

The sample-by-sample covariances of the genetic and noise term, R and I_N, are provided and the model estimates the trait-by-trait covariance matrices C_g and C_n. The complexity of the VD is $O(N^2 + t(NP^2 + NP^4))$ with N the number of samples, P the number of traits, and t the number of iterations of Broyden's method for optimising the restricted marginal likelihood (RML) of the parameter estimates. From this equation, it becomes evident that as the number of traits increases, the complexity increases by a power of four and explains why the LMM is not feasable for large trait sets. In order to to allow for multi-trait mapping of large trait sets, a bootstrapping-based approach was investigated. Instead of modeling all traits at the same time, a subset of s traits is randomly selected, the VD computed and the $s \times s$ covariance matrices C_g^* and C_n^* recorded. This random selection with replacement is repeated n times such that each two traits are drawn together at least ten times, with n depending on the overall trait size P and the sampling size s. The challenge after the VD is to combine the bootstrap results in a way, that the resulting C_g and C_n are true covariance matrices, i.e. positive semi-definite matrices. Simply averaging values over all bootstrapping results leads to matrices with negative eigenvalues which have to be regularized to achieve positive semi-definiteness. The regularization leads to an overestimation of the trait variances (diagonal) compared to the trait-trait covariances (off-diagonal). In order to circumvent this overestimation, the unregularized averaged C_g and C_n are used as an initial guess to fit a residual sum of squares model over all bootstrap runs. The model makes use of cholesky decomposition of the matrix to be fitted, resulting in $\frac{1}{2}P(P+1)$ model parameters to be fitted. The fitting is achieved by using the BFGS optimizer, a limited memory quasi-Newton algorithm for solving large nonlinear optimization problems [?]. C_g and C_n are fitted separately. The complexity of LiMMBo is $O(N^2 + nt_1(Ns^2 + Ns^4) + t_2(\frac{1}{2}P(P+1))$, which is the sum of the complexity of the VD as described above for the subset of s traits and the complexity of fitting the BFGS algorithm t_2 times for the full trait set size P.

0.2 Simulated dataset

In order to assess calibration and power of LiMMBo, genotypes and phenotypes were simulated similar to strategies described in [**??**].

Genotypes. The synthetic genotypes were generated based on real genotype data from four European ancestry populations of the 1000 Genomes (1KG) Project (populations: CEU, FIN, GBR, TSI) [Abecasis *et al.* 2012]. Depending on the cohort structure, each newly synthesised individual is assigned to N ancestors from the original 1KG Project and their genome split into blocks of 1,000 SNPs (thereby retaining realistic LD structure between SNPs). For each SNP block, an ancestor is chosen at random and its genotype is copied to the individuals genome. Low numbers of N introduce relatedness among indivduals, whereas high numbers of N lead to low levels of structure and relatedness. By allowing random selection of the ancestors from the four subpopulations, population structure will be low, whereas predefining subpopulation of the ancestor can give rise to structured populations. Three distinct cohorts of 1,000 individuals each were generated: i) unrelated individuals, with population structure (unrelatedPopStructure): $N = 10$, prior assigment to ancestorial population , ii) unrelated individuals, no population structure (unrelatedNoPopStructure): $N = 10$, no prior assigment to ancestorial population and iii) related individuals, no population structure (relatedNoPopStructure): $N = 2$, no prior assigment to ancestorial population. The genetic relatedness matrices and scatter plots of the first two principal components for each cohort are shown in Figure 1.

Phenotypes. The phenotypes $\mathbf{Y} \in \mathcal{R}^{N,P}$ of N samples and P traits were generated as the sum of four components: i) fixed genetic effects $\mathbf{U} \in \mathcal{R}^{N,P}$, ii) random genetic effects $\mathbf{G} \in \mathcal{R}^{N,P}$, iii) fixed noise effects$\mathbf{C} \in \mathcal{R}^{N,P}$ and iv) random noise effects $\mathbf{\Psi} \in \mathcal{R}^{N,P}$. For each component, a certain percentage of variance is shared across all traits (*shared*) and the remainder is independent (*ind*) across traits.

1. *Fixed genetic effects* $\mathbf{U}$. For the fixed genetic effects, S random SNPs for N samples were drawn from the simulated genotypes. From the S random SNPs, a proportion θ was selected to be causal across all traits. $\mathbf{U}^{shared} \in \mathcal{R}^{N,P}$ was simulated as the matrix product of this shared causal SNP matrix $\mathbf{X}^{shared} \in \mathcal{R}^{N,\theta\text{x}S}$ and a $\mathbf{B}^{shared} \in \mathcal{R}^{\theta\text{x}S,P}$ shared effect size matrix. $\mathbf{B}^{shared}$ in turn is the matrix product of the two normally distributed vectors $b_s \in \mathcal{R}^{\theta\text{x}S,1}$ and $b_p^T \in \mathcal{R}^{1,P}$. The remaining $(1-\theta)\text{x}S$ SNPs were simulated to have an idependent effect across a limited number of traits p^{ind}. To realise this structure, $\mathbf{B}^{ind} \in \mathcal{R}^{(1-\theta)\text{x}S,P}$ is initialised with normally distributed entries. Subsequently, $1 - p^{ind}$ traits are randomly selected and the row entries for $\mathbf{B}^{ind}$ at these traits set to zero. $\mathbf{U}^{ind} \in \mathcal{R}^{N,P}$ is the matrix product of

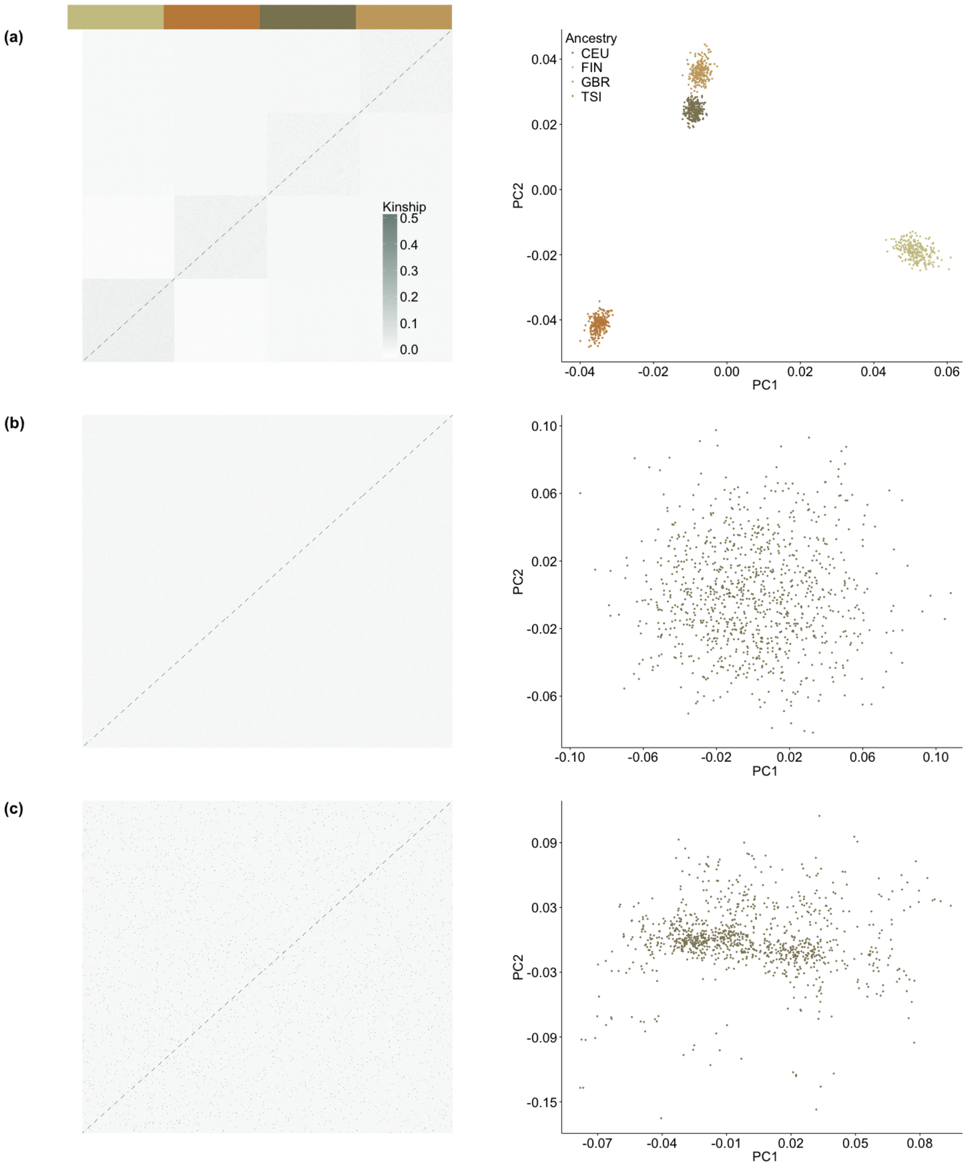


Figure 1: Kinship matrices and principle components of three simulated European ancestry cohorts. The genotypes were simulated based in genotype data from four Europen ancestry population (ancestry color key in panel a). Depending on the choice and number of ancestors for the sampling of chromosomes to simulate an individulas genotype, cohorts with differing levels of population and relatedness structure will arise. (a) unrelated individuals, with population structure: $N = 10$, prior assigment to ancestorial population. (b) unrelated individuals, no population structure: $N = 10$, no prior assigment to ancestorial population. (c) related individuals, no population structure: $N = 2$, no prior assigment to ancestorial population.

$\mathbf{X}^{ind} \in \mathcal{R}^{N,(1-\theta)\text{x}S}$ and $\mathbf{B}^{ind}$. The fixed genetic effect $\mathbf{U}$ is the sum of $\mathbf{U}^{shared}$ and $\mathbf{U}^{ind}$.

2. *Fixed noise effects* $\mathbf{C}$ The fixed noise effects$\mathbf{C}$ are based on K normally distributed confounders $\mathbf{F} \in \mathcal{R}^{N,K}$, with a proportion γ being shared across all traits yielding the shared confounder matrix $\mathbf{F}^{shared} \in \mathcal{R}^{\gamma\text{x}K,P}$. The proportion of $1-\gamma$ confounder

that are independent build the independent confounder matrix $\mathbf{F}^{ind} \in \mathcal{R}^{(1-\gamma)\text{x}K,P}$. The effect size matrices $\mathbf{A}^{shared} \in \mathcal{R}^{(\gamma)\text{x}K,P}$ and $\mathbf{A}^{ind} \in \mathcal{R}^{(1-\gamma)\text{x}K,P}$ were designed in analogy to the ones for the fixed genetic effects. The total fixed noise effect is then $\mathbf{C} = \mathbf{K}^{shared}\mathbf{A}^{shared} + \mathbf{K}^{ind}\mathbf{A}^{ind}$.

3. *Random genetic effects* The random genetic effects $\mathbf{G} \in \mathcal{R}^{N,P}$ are modeled as a matrix-normally distributed random variable, defined by its mean $\mathbf{M} \in \mathcal{R}^{N,P}$, its column covariance $\mathbf{C} \in \mathcal{C}^{P,P}$ and its row covariance $\mathbf{R} \in \mathcal{R}^{N,N}$.

$$\mathbf{G} \sim MN_{N,P}(0, \mathbf{R}, \mathbf{C}) \tag{2}$$

The NxN kinship matrix K, estimated according to Equation **??** from the SNP genotypes of simulated samples represents $\mathbf{R}$. The structure of the trait-to-trait covariance $\mathbf{C}$ depends on the design ot the effect, shared or independent. To construct $\mathbf{G}$, assume a matrix-normally distributed random variable $\mathbf{Z}$ with $\mathbf{M} = 0$ and $\mathbf{R} = \mathbf{K}$:

$$\mathbf{Z} \sim MN_{N,P}(0, \mathbf{K}, \mathbf{C}) \tag{3}$$

$\mathbf{Z}$ can be expressed in terms of a multivariate normal distribution

$$vec(Z) \sim N_{N\text{x}P}(0, \mathbf{C} \otimes \mathbf{K}) \tag{4}$$

With the cholesky decompositon of $\mathbf{K}$ and $\mathbf{C}$ into $\mathbf{K} = \mathbf{B}\mathbf{B}^T$ and $\mathbf{C} = \mathbf{A}\mathbf{A}^T$

$$vec(Z) \sim N_{N\text{x}P}(0, \mathbf{A}\mathbf{A}^T \otimes \mathbf{B}\mathbf{B}^T) \tag{5}$$

which can be rearranged as

$$\begin{aligned} vec(\mathbf{Z}) &\sim N(0, (\mathbf{A} \otimes \mathbf{B})I(\mathbf{A}^T \otimes \mathbf{B}^T) \\ vec(\mathbf{Z}) &\sim N(0, (\mathbf{A} \otimes \mathbf{B})I(\mathbf{A} \otimes \mathbf{B})^T) \end{aligned} \tag{6}$$

Using the property of

$$\begin{aligned} \mathbf{Y} &\sim N(0, \Sigma) \\ w\mathbf{Y} &\sim N(0, w\Sigma w^T) \end{aligned} \tag{7}$$

we can let $vec(Z) = (\mathbf{A} \otimes \mathbf{B})vec(\mathbf{Y})$ such that

$$(\mathbf{A} \otimes \mathbf{B})vec(\mathbf{Y}) \sim N(0, (\mathbf{A} \otimes \mathbf{B})I(\mathbf{A} \otimes \mathbf{B})^T) \tag{8}$$

Using [**?**]: Lemma 4.3.1, we get

$$(\mathbf{A} \otimes \mathbf{B}) vec(\mathbf{Y}) = vec(\mathbf{B}\mathbf{Y}\mathbf{A}^T) \tag{9}$$

For the independent effect, $\mathbf{A}^{ind}$ is a diagonal matrix with normally distributed entries: $\mathbf{A}^{ind,T} = \text{diag}(a_1, a_2, \ldots, a_P) \sim N(0,1)$, such that $\mathbf{G}^{shared} = vec(\mathbf{B}\mathbf{Y}\mathbf{A}^{ind,T})$. $\mathbf{A}^{shared}$ of the shared effect is a matrix of row rank one, with normally distributed entries in row 1 and zeros elsewhere: $a_{1,j} \sim N(0,1)$ and $a_{i \neq 1,j} = 0$ such that $\mathbf{G}^{shared} = vec(\mathbf{B}\mathbf{Y}\mathbf{A}^{shared,T})$. The total random genetic effect $\mathbf{G}$ is $\mathbf{G} = \mathbf{G}^{shared} + \mathbf{G}^{ind}$.

4. *Random noise effects* The random noise effects $\mathbf{\Psi}$ are simulated as the sum of two random noise effects, a shared and an independent one. The shared random effect $\mathbf{\Psi}^{shared}$ is simulated as $vec(\mathbf{\Psi}^{shared}) \sim N(0,1)$ The independent random effect $\mathbf{\Psi}^{ind}$ is simulated as the matrix product of two normal distributions $\mathbf{a} \sim N_N(0,1)$ and $\mathbf{b} \sim N_P(0,1)$: $\mathbf{\Psi}^{ind} = \mathbf{a}\mathbf{b}^T$.

Each of the phenotype components is rescaled such that their average column variance explains x percent of the total variance. The scale factor a is derived is follows: Let X be a random variable with expected value $E[X] = \mu_x$ and variance $V[X] = E[(X - \mu_x)^2]$ and let $Y = aX$. Then

$$\begin{aligned} E[Y] &= a\mu_x \\ V[Y] &= E[(Y - \mu_y)^2] \\ V[Y] &= E[(aX - a\mu_x)^2] \\ &= a^2 E[(X - \mu_x)^2] \end{aligned} \tag{10}$$

To scale the phenotype components such that their average column variance $\overline{V_{col}} = \frac{V_1 + \ldots + V_p}{p}$ explains a specified percentage x of the total variance, choose the scaling factor a such that:

$$\begin{aligned} x &= a^2 \times \overline{V_{col}} \\ a &= \sqrt{\frac{x}{\overline{V_{col}}}} \end{aligned} \tag{11}$$

The final simulated phenotype is expressed as

$$\mathbf{Y} = \mathbf{U}^{scaled} + \mathbf{C}^{scaled} + +\mathbf{G}^{scaled} + \mathbf{\Psi}^{scaled} \tag{12}$$

For each of the genotype cohorts described in 0.2, multiple phenotype scenarios depending on the percentage of variance explained by genetics h_2 (sum of fixed and random genetic effects) and number of traits P were simulated: i) $h_2 = 0.2, 0.5, 0.8$ and ii) $P = 10, 20, ..., 100$. Table 1 introduces the parameter symbols for the the different phenotype components. In Table2, their values and value ranges used in the different phenotype simulation scenarios are summarised.

Table 1: Parameters for phenotype simulation.

		variance explained	shared	independent
	total	h_2		
genetic effects	fixed	$h_2 h_2^s$	θ	1-θ
	random	$h_2 h_2^g$	η	1-η
	total	(1-h_2)		
noise effects	fixed	(1-h_2)δ	γ	1-γ
	random	(1-h_2)(1-δ)	α	1-α

Table 2: Parameter values for simulated phenotypes.

Parameter	Values
h_2	0.8, 0.5, 0.2
h_2^s	0.0125, 0.02, 0.05
h_2^g	0.9875, 0.98, 0.5
$1 - h_2$	0.2, 0.5, 0.8
δ	0.4
$1 - \delta$	0.6
θ	0.6
η	0.8
γ	0.6
α	0.8

0.3 LiMMBo yields covariance estimates comparable to RML estimates

For trait set sizes of up to 30 traits, the RML estimates of C_g and C_n are possible. In order to compare the estimates derived from pure RML and from the combination of RML estimates of p-sized subset matrices (LiMMBo), the VD of the simulated phenotypes with $P = (10, 20, 30)$ was estimated both via RML and LiMMBo for all phenotype setups (Section **??**). For $P = 10$, subsets of $p = 5$ were drawn, otherwise $p = 10$. C_g and C_n estimates of both methods were used in a any effect multi-variate LMM (Equation **??**) across all genome-wide SNPs (mtGWAS). Statistical calibration of the mtGWAS was estimated by counting the number of tests that exceed a given threshold α divided by the overall number of tests conducted (number of SNPs). Figure 2 shows the comparison of the false discovery rate (FDR) estimates of mtGWAS depending on the method of VD estimation, trait set size and genetic architecture. If the model is well calibrated, the

FDR (depicted as bar charts in different transparency for both estimates) reaches as far or beyond the vertical line for the applied α threshold. mtGWAS with LiMMBo-derived C_g and C_n estimates yields equally well calibrated results compared to the results from RLM estimates.

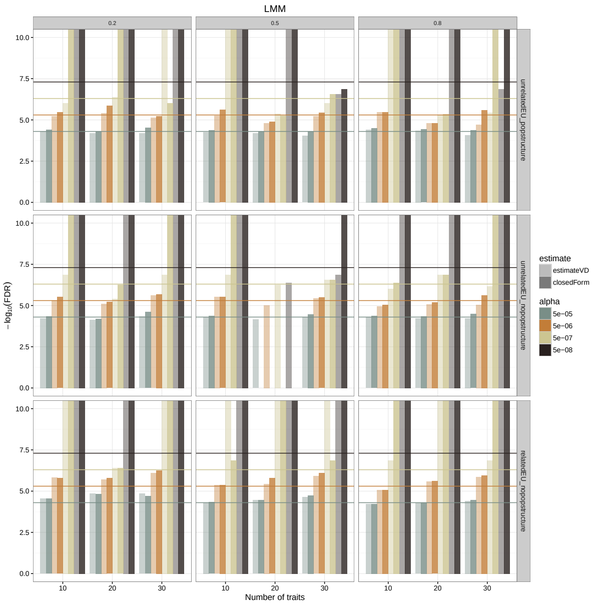
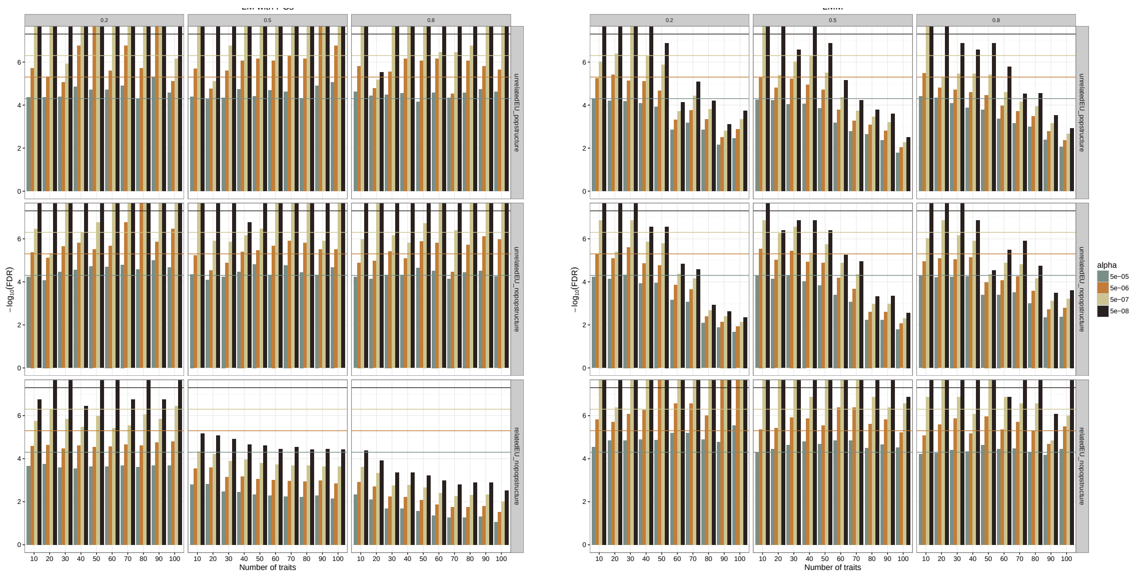


Figure 2: Comparison of LMM based on RLM trait-trait covariance estimates to LiMMBo-derived estimates. For each genetic architecture (unrelatedEU_popstructure, unrelatedEU_nopopstructure, relatedEU_nopopstructure) and percentage of variance explained by genetics (0.2, 0.5, 0.8; purely by background genetic effects), three different trait set sizes (10, 20, 30) were simulated and the calibration of the model was assesed by mtGWAS with LiMMBo-derived (estimateVD) RLM estimates (closedForm) of C_g and C_n. For calibrated models, the FDR (depicted as bar charts in different transparency for both estimates) reaches as far or beyond the vertical line for the applied α threshold. Overall, estimates from both models yield equally well calibrated results.

0.4 Model choice for mtGWAS depends on relatedness and population structure

Figure 3 shows the calibration of the mtGWAS for either the LM with the first ten principle components (LM-PC) of the kinship matrix as covariates F (Figure 3a; Equation **??**) or the LMM (Figure 3b; Equation **??**). The model calibration strongly depends on the underlying genetic architecture of the cohort. In cohorts without relatedness and structure, LM-PC is much better calibrated than the LMM, whereas the LMM outperforms the LM-PC in cohorts with related individuals.



(a) Linear model with PCs

(b) Linear mixed model

Figure 3: Calibration of different models depending on population structure, number of traits and percentage of genetic variance. (a) The LM set-up is reasonably well-calibrated for populations of unrelated individuals (upper two panels) across all trait sizes but is no calibrated for populations wtih high inter-individual relatedness especially in cases with a strong underlying genetic cause of the trait. (b) The LMM is well-calibrated up to trait set sizes of 30-40 traits for unrelated populations, but loses calibration for larger trait set sizes (upper two panels). It stays well calibrated across all trait sizes for populations with related individuals (lower panel).

Bibliography

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. & McVean, G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65. doi:10.1038/nature11632

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P. & Zondervan, K.T. (2010) Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564–73. doi:10.1038/nprot.2010.116

Delaneau, O., Marchini, J. & Zagury, J.F. (2012) A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–81. doi:10.1038/nmeth.1785

Delaneau, O., Zagury, J.F. & Marchini, J. (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, 10(1):5–6. doi:10.1038/nmeth.2307

Howie, B., Marchini, J. & Stephens, M. (2011) Genotype imputation with thousands of genomes. *G3 (Bethesda, Md.)*, 1(6):457–70. doi:10.1534/g3.111.001198

Howie, B.N., Donnelly, P. & Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6):e1000529. doi:10.1371/journal.pgen.1000529

Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–13. doi:10.1038/ng2088

Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. (2012) Improved heritability estimation from genome-wide SNPs. *American journal of human genetics*, 91(6):1011–21. doi:10.1016/j.ajhg.2012.10.010

Teo, Y.Y., Inouye, M., Small, K.S., Gwilliam, R., Deloukas, P., Kwiatkowski, D.P. & Clark, T.G. (2007) A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics (Oxford, England)*, 23(20):2741–6. doi:10.1093/bioinformatics/btm443

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, 437(7063):1299–320. doi:10.1038/nature04226

UK10K Consortium (2014) UK10K Project. *http://www.uk10k.org*