Supplementary Material

cvBLUPs TO COMPUTE LINEAR MIXED-MODEL SHRINK PARAMETERS

Here, we examine the use of cvBLUPs (cross-validated Best Linear Unbiased Predictors) in estimating linear mixed-model (LMM) shrink parameters. The BLUP effect size estimates from a linear mixed model are ''shrunk'' to be smaller in magnitude to exploit a bias-variance trade-off that reduces their mean squared error. With the estimated genetic variance component $\hat{\sigma}_g^2$, the estimated distribution of effect estimates $b_j \sim \mathcal{N}\left(0, \frac{\hat{\sigma}_g^2}{M}\right)$ is an empirical prior on the effect sizes, which yields estimates that are biased toward the prior mean of 0. Equivalently, the BLUP effect size estimates are ridge regression estimates with a penalty parameter λ related to the LMM variance components: $\lambda = \frac{1-h^2}{h^2} = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_\varepsilon^2}$.

When single-nucleotide polymorphisms (SNPs) are independent, the LMM shrinkage estimates of the effect sizes b for given estimates of the variance components are found to be as given in a study (Vilhjálmsson et al., 2015).

$$\hat{b}_j \sim \mathcal{N}\left(\frac{h^2}{h^2 + \frac{M}{N}} b, \quad \frac{h^2}{h^2 + \frac{M}{N}} \frac{\hat{\sigma}_g^2}{M}\right) \tag{1}$$

and the shrink, $\frac{h^2}{h^2 + \frac{M}{N}}$, can be calculated directly from the number of individuals and SNPs. However, when SNPs are in linkage disequilibrium, it is more complicated to estimate the effective number of SNPs (Patterson et al., 2006) and hence the shrink. Here we show how cvBLUPs can be used to estimate the shrink directly.

First, in the case of independent SNPs, the LMM estimates of the SNP effects $\hat{b}$ would be reduced in magnitude and variance relative to the true values b, as in Equation 1, and there would be a corresponding decrease in the variance of polygenic risk scores for a new individual with normalized genotypes Z_{new}, calculated using $\hat{b}$ as weights rather than the true effect sizes b:

$$\begin{aligned} \text{var}\left(Z_{\text{new}}\hat{b}\right) &= \left(\frac{h^2}{h^2 + \frac{M}{N}}\right) \text{var}(Z_{\text{new}} b) \\ &= \left(\frac{h^2}{h^2 + \frac{M}{N}}\right) \sigma_g^2 \end{aligned} \tag{2}$$

Equation 2 suggests the direct estimation of the shrink by taking the ratio of $\hat{\sigma}_g^2$ to $\text{var}\left(Z_{\text{new}}\hat{b}\right)$. We confirm the approach is approximately unbiased via simulation (Supplementary Table S1). For multiple settings of N (number of subjects), M (number of SNPs), and h^2(heritability), heritability and variance components were estimated and the value of the standard independent-SNP model for the shrink is compared with the empirical BLUP shrink $\frac{\text{var(BLUP)}}{\hat{\sigma}_g^2}$ and the empirical cvBLUP shrink $\frac{\text{var(cvBLUP)}}{\hat{\sigma}_g^2}$. In each simulation setting, the BLUP shrink is much larger than the independent model value or the cvBLUP shrink due to overfitting by the standard BLUPs. However, the cvBLUP formula and the independent-SNP model are consistent for all parameter settings.

Because this approach does not require identification of an effective number of SNPs, it extends directly to the case where there is linkage disequilibrium. We applied this approach to simulated data from the metabolic syndrome in men (METSIM) cohort used above (Supplementary Table S2). We estimated the shrink parameter for simulated phenotypes based on the real genotypes at 609131 SNPs with minor allele frequencies greater than 0.01 on the 6263 unrelated (at the 0.05 level) subjects. Twenty simulations were run using fractions of causal SNPs between 0.0001 and 1.0. Causal SNPs were chosen by simple random sampling with equal probability from the genotyped SNPs. SNP effect sizes were normally distributed and

trait heritability was 50%. The shrink formula for independent SNPs suggested a shrink of about 0.0045–0.0055, while the cvBLUP shrink—the ratio of the variance of the cvBLUPs to the estimated genetic variance component, ranged 0.06–0.08.

In this setting, with LD between the SNPs, the independent SNP formula for the shrink is invalid. In particular, the effective number of SNPs $\tilde{M}$ is much smaller than the total number of SNPs leading to a shrink estimate over ten times smaller than the the cvBLUP-based estimate.

RELATIONSHIP BETWEEN PRINCIPAL COMPONENTS AND CROSS-VALIDATED BLUPS

Genetic principal components are routinely used as quantitative measures of study of participants' ancestry (Patterson et al., 2006; McVean, 2009) and as such are used as adjustment covariates in association studies to block confounding by population structure (Patterson et al., 2006). Linear mixed models provide another framework for controlling potential confounding by population structure (Yang et al., 2014). Both genetic principal component (PC) adjusted regression and linear mixed models (LMMs) are methods that account for ancestry and other forms of genetic structure in a data set by analyses of the genetic relatedness matrix (GRM) $K = \frac{1}{m} ZZ^T$. Specifically, PCA involves calculating some number of eigenvectors of the GRM—or equivalently left singular vectors of the scaled genotype matrix Z–and using them as adjustment covariates in regression analyses, while LMMs model observed outcomes as nonindependent, with the random effects that contribute to the outcome y correlated to a degree related to the amount of shared genetic variation between each pair of subjects: $\mathrm{cov}(y) = \Sigma = \sigma_g^2 K + \sigma_\varepsilon^2 I_n$.

In principal component-adjusted analyses, some number of PCs are used as adjustment covariates in linear regression. With linear mixed models, generalized least squares with an estimate of the sample covariance $\widehat{\Sigma}$ are used.

Principal component-adjusted analysis has several disadvantages. First, it is not clear how many principal components should be used. In genome-wide association studies (GWASs), it has become conventional to use a standard number of PCs, say 10, but it is generally not clear whether that will be enough to account for the components of genetic structure that are confounded with nongenetic factors in a study. Second, it is not clear which PCs should be included. Conventionally, PCs are sorted in descending order by the magnitude of the corresponding eigenvalues, and PCs with the largest eigenvalues are used. However, selected PCs may not be associated with any nongenetic factors and may not relieve any confounding. Finally, in association testing applications, tests needlessly lose power due to overadjustment when PCs that do not adjust for confounding are included as covariates. By construction, PCs represent axes of variation in the genotype data, so some may be highly correlated with a test SNP, and inclusion of correlated covariates increases the standard errors around test SNP effect estimates.

The spectral decomposition of the GRM K is as follows:

$$K = U \Lambda U^T \tag{3}$$

where U is a unitary matrix whose columns are the genetic PCs or eigenvectors of K, and Λ is a diagonal matrix of the corresponding eigenvalues.

Principal component-adjusted regression includes U_k (first k columns of U) as regression covariates to improve the estimation or testing of β_g:

$$y = X\beta + g\beta_g + U_k\gamma + \varepsilon \tag{4}$$

This leaves open the question of how many or which PCs should be included. As an alternative to standard practice, we can try using all of the PCs, and use regularization to keep the model estimable. We can also minimize problems due to overadjustment by calculating the PCs U using a GRM K^* or scaled genotype matrix Z^* that does not include SNPs in linkage disequilibrium with or on the same chromosome as the test-SNP g.

Toward a connection to cvBLUPs, rescale the columns of U by the corresponding singular values:

$$U_s = U\sqrt{\Lambda} \tag{5}$$

Supplementary Table S1. Estimation of the Shrink Parameters for BLUPs and cvBLUPs in Simulations with Independent SNPs

N	*M*	h^2	$\widehat{h^2}$	*Independent shrink*	*BLUP shrink*	*cvBLUP shrink*
400	400	0.5	0.512 (0.102)	0.335 (0.003)	0.627 (0.007)	0.389 (0.006)
400	800	0.5	0.502 (0.119)	0.200 (0.003)	0.560 (0.010)	0.217 (0.005)
400	1200	0.5	0.494 (0.128)	0.139 (0.003)	0.530 (0.011)	0.150 (0.004)
800	400	0.5	0.492 (0.062)	0.496 (0.002)	0.695 (0.004)	0.573 (0.004)
800	800	0.5	0.501 (0.066)	0.334 (0.002)	0.619 (0.005)	0.383 (0.004)
800	1200	0.5	0.500 (0.075)	0.248 (0.002)	0.581 (0.006)	0.280 (0.004)
1200	400	0.5	0.505 (0.052)	0.600 (0.002)	0.770 (0.003)	0.700 (0.003)
1200	800	0.5	0.501 (0.051)	0.426 (0.002)	0.661 (0.003)	0.494 (0.003)
1200	1200	0.5	0.501 (0.051)	0.332 (0.002)	0.614 (0.004)	0.378 (0.003)
400	400	0.1	0.105 (0.080)	0.089 (0.006)	0.199 (0.010)	0.106 (0.006)
400	800	0.1	0.114 (0.096)	0.052 (0.004)	0.174 (0.011)	0.060 (0.004)
400	1200	0.1	0.114 (0.104)	0.035 (0.003)	0.170 (0.011)	0.044 (0.003)
800	400	0.1	0.102 (0.044)	0.165 (0.006)	0.238 (0.008)	0.169 (0.006)
800	800	0.1	0.101 (0.056)	0.090 (0.005)	0.178 (0.008)	0.094 (0.004)
800	1200	0.1	0.105 (0.066)	0.063 (0.004)	0.160 (0.008)	0.067 (0.004)
1200	400	0.1	0.098 (0.032)	0.222 (0.006)	0.283 (0.007)	0.225 (0.006)
1200	800	0.1	0.101 (0.035)	0.130 (0.004)	0.208 (0.006)	0.132 (0.004)
1200	1200	0.1	0.098 (0.041)	0.088 (0.003)	0.170 (0.006)	0.090 (0.004)

The independent shrink is that derived in Formula 1 for independent SNPs.

BLUP, Best Linear Unbiased Predictor; cvBLUP, cross-validated Best Linear Unbiased Predictor; SNP, single-nucleotide polymorphism.

Rescaling U to U_s puts a higher prior on PCs with larger eigenvalues. Now compress the contribution of the principal components to the outcome by calculating a vector of phenotypic predictions using ridge regression with a penalty λ on the principal components:

$$\hat{y}_{pc} = U_s U_s^T (U_s U_s^T + \lambda I_n)^{-1} (y - X\hat{\beta}) \tag{6}$$

If we define $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_g^2} = \frac{1-h^2}{h^2}$ and recall that $U_s = U\sqrt{(\Lambda)}$, so that $U_s U_s^T = U\Lambda U^T = K^*$, we have the following:

$$\hat{y}_{pc} = \sigma_g^2 K (\sigma_g^2 K^* + \sigma_\varepsilon^2 I_n)^{-1} (y - X\hat{\beta}) \tag{7}$$

However, this is just a BLUP. So, BLUPs arise from a limiting case of trying to do PC-adjusted regression with all PCs. Therefore, cross-validated predictions from ridge regression on all PCs are cvBLUPs.

Supplementary Table S2. Estimates of the Shrink Parameter for Simulated Phenotypes Based on the Real Genotypes, METSIM Cohort, and Various Genetic Architectures or Fractions of Causal SNPs

Fraction causal SNPs	*Independent shrink*	*BLUP shrink*	*cvBLUP shrink*
1.0	0.0050 (1.60E-04)	0.514 (0.015)	0.076 (0.002)
0.1	0.0053 (1.71E-04)	0.539 (0.016)	0.079 (0.002)
0.01	0.0055 (1.73E-04)	0.556 (0.016)	0.080 (0.002)
0.001	0.0047 (5.45E-05)	0.477 (0.005)	0.067 (0.001)
0.0001	0.0040 (1.20E-04)	0.409 (0.011)	0.058 (0.001)

The independent shrink is that derived in Formula 1 for independent SNPs.

METSIM, metabolic syndrome in men.

cvBLUPs as adjustment covariates are similar to a compression of all PCs into a single covariate, with the PCs given prior weights that emphasize the PCs with larger eigenvalues, but do not exclude any. The PCs are also weighted by their relevance to the outcome because they represent predictions from a ridge regression model that implicitly has ridge regression effect sizes for the association of the PC with the outcome. Overfitting in the ridge regression step is avoided by the leave-one-out cross-validation. Finally, loss of power by overadjustment is avoided by excluding the chromosome or SNPs in linkage disequilibrium, with test SNP from the GRM used for calculation of the cvBLUPs. In fact, unlike PC adjustment but similar to standard LMM analyses, cvBLUP adjustment boosts the power of association studies by modeling genetic contributions to the phenotype other than the SNP of interest, thereby increasing the signal-to-noise ratio.

FORMULA FOR EFFICIENT LEAVE-ONE-OUT CROSS-VALIDATION OF THE LINEAR MIXED MODEL

Let

$$y = X\beta + Zb + \varepsilon \tag{8}$$

represent a linear mixed model with continuous outcome y, fixed effect covariates X, fixed effect sizes β, additively coded genotypes Z, random genetic effect sizes b, and unmodeled or environmental factors ε. The genetic effect sizes b and environmental factors ε are modeled as i.i.d. normally distributed random variables with variances $\frac{1}{M}\sigma_g^2$ and σ_ε^2, respectively:

$$\begin{aligned} b &\sim \mathcal{N}\left(0, \frac{\sigma_g^2}{M} I_M\right) \\ \varepsilon &\sim \mathcal{N}\left(0, \sigma_\varepsilon^2 I_N\right) \end{aligned} \tag{9}$$

By convention, the genotypes are scaled to have mean 0 and variance 1, and the SNP effect sizes are assumed to have effect sizes independent of minor allele frequency (MAF) on this scale. So, the total genetic contributions to the phenotype Zb are normally distributed with mean 0 and variance σ_g^2. The scaled genotypes Z are used to calculate an SNP-based GRM K:

$$K = \frac{1}{M} ZZ^T. \tag{10}$$

The observations y in the linear mixed model are normally distributed with a covariance Σ that depends on K and the variance components σ_g^2 and σ_ε^2

$$\begin{aligned} y &\sim \mathcal{N}(X\beta, \Sigma) \\ \Sigma &= \sigma_g^2 K + \sigma_\varepsilon^2 I_N \\ \widehat{\Sigma} &= \hat{\sigma}_g^2 K + \hat{\sigma}_\varepsilon^2 I_N \\ H &= \hat{\sigma}_g^2 K \widehat{\Sigma}^{-1} \\ \hat{\beta} &= \left(X^T \widehat{\Sigma}^{-1} X\right)^{-1} \left(X^T \widehat{\Sigma}^{-1} y\right) \\ \hat{b} &= \frac{\hat{\sigma}_g^2}{M} Z^T \widehat{\Sigma}^{-1} (y - X\hat{\beta}) \\ \hat{y} &= \text{BLUP} \\ &= Hy \\ &= \frac{\hat{\sigma}_g^2}{M} ZZ^T \widehat{\Sigma}^{-1} (y - X\hat{\beta}) \\ &= Z\hat{b} \end{aligned} \tag{11}$$

To simplify notation, let fixed effect sizes be β and fixed effects $X\beta = 0$:

$$\begin{aligned} \hat{b} &= \frac{\hat{\sigma}_g^2}{M} Z^T \widehat{\Sigma}^{-1} y \\ \hat{y} &= \frac{\hat{\sigma}_g^2}{M} Z Z^T \widehat{\Sigma}^{-1} y \\ &= Z\hat{b} \end{aligned} \tag{12}$$

Exclude observation i from the genetic predictive model fit and then make an out-of-sample (oos) prediction for observation i:

$$\begin{aligned} \hat{b}_{-i} &= \frac{\hat{\sigma}_g^2}{M} Z_{-i}^T \widehat{\Sigma}_{-i}^{-1} y_{-i} \\ \hat{y}_{i,\,oos} &= Z_i \hat{b}_{-i} \end{aligned} \tag{13}$$

Now generate the augmented vector y_{+i} as the vector of outcomes y with observation y_i replaced with its out-of-sample prediction from a model trained on the remaining observations, $\hat{y}_{i,\,oos}$.

Consider the ridge regression interpretation of the mixed model with the ridge penalty $\lambda = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_g^2} = \frac{1-h^2}{h^2}$. Assume the variance components or heritability has been estimated in a prior step or are known.

For the reduced data set with the outcome for observation i removed we have:

$$\hat{b}_{-i} = \arg\min_b \frac{1}{2} \sum_{k \neq i} (y_{-i} - Z_{-i} b)^2 + \frac{\lambda}{2} ||b||_2^2 \tag{14}$$

For the augmented data set with the outcome for observation i replaced by an out-of-sample prediction we have:

$$\hat{b}_{+i} = \arg\min_b \frac{1}{2} \sum_{k \neq i} (y_k - Z_k b)^2 + (\hat{y}_{i,\,oos} - Z_i b)^2 + \frac{\lambda}{2} ||b||_2^2 \tag{15}$$

For the augmented model, $(\hat{y}_{i,\,oos} - Z_i \hat{b}) = 0$, and the remaining terms in the expression for $\hat{b}_{+i}$ are the same as for the reduced model $\hat{b}_{-i}$, so $\hat{b}_{+i} = \hat{b}_{-i}$.

Now consider the differences in the predictions for the ith value from the augmented model and the model with all observations:

$$\begin{aligned} \hat{y}_{+i} - \hat{y}_i &= Z_i \hat{b}_{+i} - Z_i \hat{b} \\ &= Z_i \frac{\hat{\sigma}_g^2}{M} Z^T \widehat{\Sigma}^{-1} y_{+i} - Z_i \frac{\hat{\sigma}_g^2}{M} Z^T \widehat{\Sigma}^{-1} y \\ &= Z_i \frac{\hat{\sigma}_g^2}{M} Z^T \widehat{\Sigma}^{-1} (y_{+i} - y) \\ &= H_{i,.} (y_{+i} - y) \end{aligned} \tag{16}$$

Here, $Z_i \frac{\hat{\sigma}_g^2}{m} Z^T \widehat{\Sigma}^{-1}$ is the ith row of the matrix H or $H_{i,.}$. The vectors y_{+i} and y only differ at the ith element, so:

$$\begin{aligned} \hat{y}_{+i} - \hat{y}_i &= H_{i,.} (y_{+i} - y) \\ &= H_{i,i} \left(Z_i \hat{b}_{-i} - y_i \right) \end{aligned} \tag{17}$$

Supplementary Table S3. Correlations of Measured HDL-Cholesterol Levels in Blood, cvBLUPs, and Polygenic Risk Scores Based on External GWAS Results

	Raw.HDLc	*Quantile.HDLc*	*cvBLUP.HDLc*	*PRS.HDLc*	*PRS5k.HDLc*
Raw.HDLc	1.00	0.98	0.12	0.23	0.07
Quantile.HDLc	0.98	1.00	0.13	0.24	0.07
cvBLUP.HDLc	0.12	0.13	1.00	0.11	0.03
PRS.HDLc	0.23	0.24	0.11	1.00	0.29
PRS5k.HDLc	0.07	0.07	0.03	0.29	1.00

Correlations are calculated for 4047 unrelated men in the METSIM cohort. Correlations based both on raw cholesterol measurements in mg/ml and quantile-normalized levels are shown. The cvBLUPs were calculated using heritability estimates and cross-validated genetic predictions for the quantile normalized cholesterol levels in this subset of the METSIM cohort. The PRSs were calculated using results of a GWAS for the quantile normalized levels of HDL cholesterol in the U.K. Biobank phase-2 males. About 147K subjects were used for each SNP's association test and effect size estimate. To simulate the properties of PRSs based on smaller reference data sets, the United Kingdom Biobank (UKBB) results were perturbed to represent the results of a GWAS with only 5k subjects, and the modified GWAS results were used to generate the polygenic risk scores with the PRS5k label.

GWAS, genome-wide association studies; HDL, high-density lipoprotein; METSIM, metabolic syndrome in men; UKBB.

Finally,

$$
\begin{aligned}
Z_i\hat{b}_{-i} - Z_i\hat{b} &= H_{i,i}\left(Z_i\hat{b}_{-i} - y_i\right) \\
&= H_{i,i}Z_i\hat{b}_{-i} - H_{i,i}y_i \\
(1-H_{i,i})Z_i\hat{b}_{-i} &= Z_i\hat{b} - H_{i,i}y_i \\
Z_i\hat{b}_{-i} &= \frac{Z_i\hat{b} - H_{i,i}y_i}{1-H_{i,i}} \\
Z_i\hat{b}_{-i} &= \frac{Z_i\hat{b} - H_{i,i}y_i}{1-H_{i,i}} \\
\hat{y}_{i,oos} &= Z_i\hat{b}_{-i} \\
&= \frac{Z_i\frac{\hat{\sigma}_g^2}{M}Z^T\widehat{\Sigma}^{-1}y - H_{i,i}y}{1-H_{i,i}} \\
&= \frac{\hat{y}_i - H_{i,i}y}{1-H_{i,i}} \\
&= \frac{H_{i,.} - H_{i,i}}{1-H_{i,i}}y
\end{aligned}
\tag{18}
$$

Supplementary Table S4. Correlations of Measured LDL-Cholesterol Levels in Blood, cvBLUPs, and Polygenic Risk Scores Based on External GWAS Results

	Raw.LDLc	*Quantile.LDLc*	*cvBLUP.LDLc*	*PRS.LDLc*	*PRS5k.LDLc*
Raw.LDLc	1.00	1.00	0.10	0.21	0.06
Quantile.LDLc	1.00	1.00	0.10	0.21	0.06
cvBLUP.LDLc	0.10	0.10	1.00	0.06	0.03
PRS.LDLc	0.21	0.21	0.06	1.00	0.09
PRS5k.LDLc	0.06	0.06	0.03	0.09	1.00

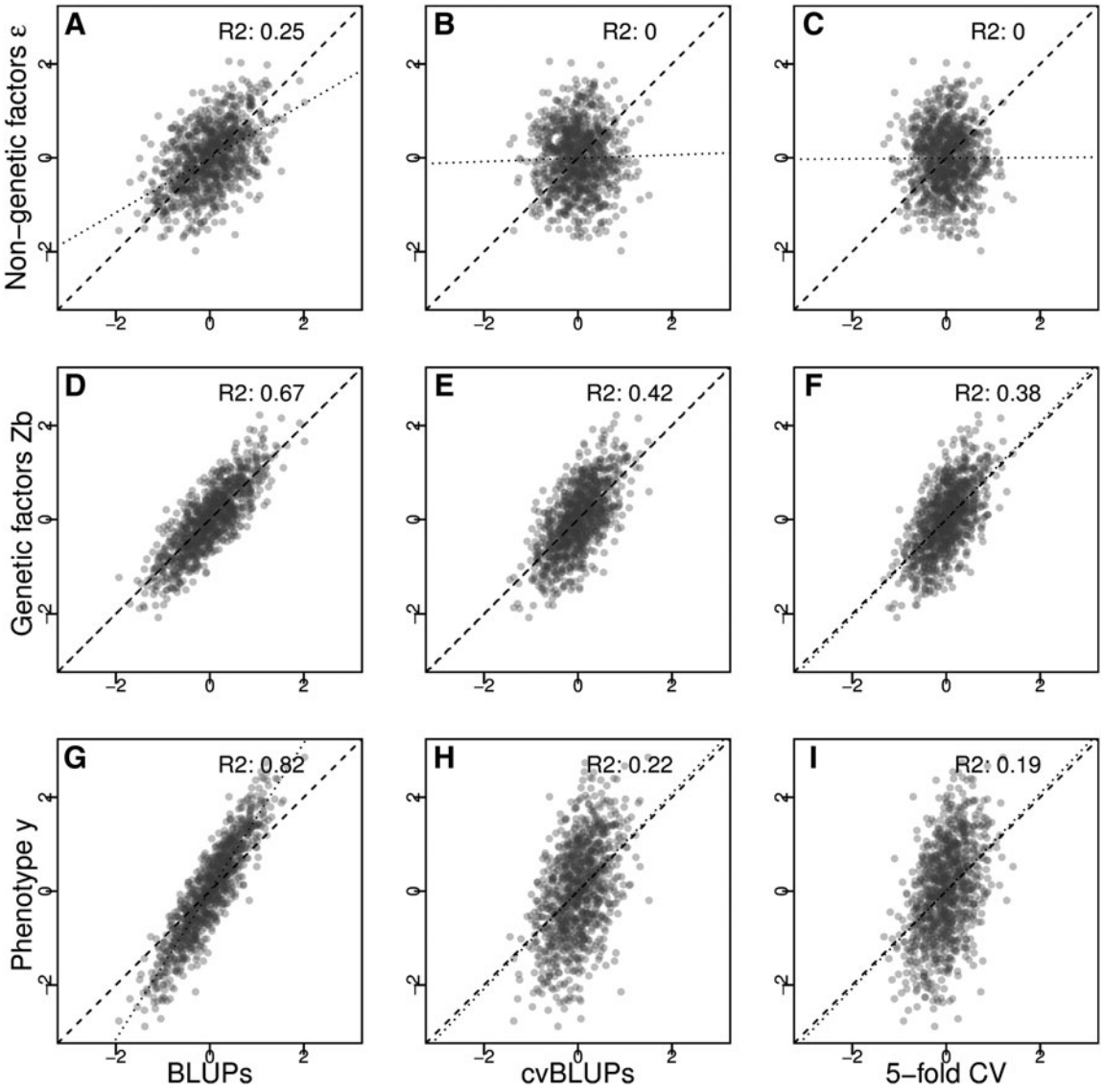
Correlations are calculated for 4047 unrelated men in the METSIM cohort.

Correlations based on both raw cholesterol measurements in mg/ml and quantile-normalized levels are shown. The cvBLUPs were calculated using heritability estimates and cross-validated genetic predictions for the quantile normalized cholesterol levels in this subset of the METSIM cohort. The PRSs were calculated using results of a GWAS for the quantile normalized levels of LDL cholesterol in the U.K. Biobank phase-2 males. About 159K subjects were used for each SNP's association test and effect size estimate. To simulate the properties of PRSs based on smaller reference data sets, the United Kingdom Biobank (UKBB) results were perturbed to represent the results of a GWAS with only 5k subjects, and the modified GWAS results were used to generate the polygenic risk scores with the PRS5k label.

LDL, low-density lipoprotein.

CORRELATIONS OF PHENOTYPES, CVBLUPS, AND STANDARD PRSs

Supplementary Tables S3 and S4 show the correlations of measured high-density lipoprotein (HDL) and low-density lipoprotein (LDL) cholesterol levels, cvBLUPs, and standard polygenic risk scores based on external GWAS results. The measurements and genetic scores are calculated in 4047 unrelated men from the METSIM cohort (Laakso et al., 2017) who had not been diagnosed with diabetes or prescribed statins. The cvBLUPs were calculated by cross-validation of linear mixed models in this same group of subjects using the quantile-normalized measurements as the outcome. The polygenic risk scores PRS.HDLc and PRS.LDLc were calculated by a standard LD-pruning and significance-filtering of external GWAS results using PRSice-2 (Euesden et al., 2014). The external reference GWASs used to identify causal SNPs, and their weights were the association studies for quantile-normalized HDL and LDL cholesterol in the men from the U.K. Biobank study, round 2 (Bycroft et al., 2018), with the summary statistics downloaded from [http://www.nealelab.is/uk-biobank].



SUPPLEMENTARY FIG. S1. Comparison of cvBLUPs and fivefold cross-validated genetic predictions. Correlations of genetic predictions [BLUP **(A,D,G)**, cvBLUP **(B,E,H)**, and fivefold cross-validated predictions **(C,F,I)**], with true genetic factors *Zb* **(A,B,C)**, independent environmental factors ε, **(D,E,F)** and the phenotype *y* **(G,H,I)** in a simulation of a continuous phenotype with $h^2 = 50\%$, 1000 subjects, and 1000 independent SNPs having random effect sizes. BLUPs are correlated with ε, while cvBLUPs are not. Dashed lines are diagonal, and dotted lines are from linear regression fits to the scatter plots. R^2 values for the linear fits to the scatter plots are shown. BLUP, Best Linear Unbiaded Predictor; cvBLUP, cross-validated Best Linear Unbiased Predictor; LMM, linear mixed model; SNP, single-nucleotide polymorphism.

The GWAS for HDL cholesterol in the U.K. Biobank had about 147K male subjects available for the analysis of each SNP. For the LDL analysis, there were about 159K subjects available for analysis of each SNP. By contrast, the cvBLUPs are calculated using only 4047 subjects from the METSIM cohort. With almost 40 times more data available, we expect more accurate identification of SNPs strongly associated with cholesterol levels and more accurate estimation of the effect sizes in U.K. Biobank (United Kingdom Biobank, UKBB) than in the relatively small METSIM cohort. So, unsurprisingly, for both phenotypes, the PRSs based on the UKBB results (PRS.HDLc and PRS.LDLc) are more highly correlated with measured cholesterol levels than the cvBLUPs.

A particular benefit of cvBLUPs is their applicability in situations were there is no suitable external reference data set for use as a source of accurately identified trait-associated SNPs and their accurately estimated effect sizes. To provide context for the accuracy of cvBLUPs as genetic prediction while accounting for the available study size, the results of the GWASs in UKBB were perturbed to simulate results of GWASs with only 5000 subjects. So do this, the standard errors for each effect estimate were multiplied by the factor $\left(\frac{N}{5000}\right)^{\frac{1}{2}}$ where N is the original GWAS sample size, and effect estimates were perturbed by addition of normally distributed noise with variance equal to the new standard error squared minus the original standard error squared to represent the greater estimation error in the smaller data set. The perturbed GWAS results were used to generate PRSs using PRSice-2 as before, yielding the scores PRS5k.HDLc and PRS5k.LDLc. Notably, these scores based on comparable numbers of observations as the cvBLUPs have correlations with the phenotypes that are only about 60% as strong as the cvBLUPs.

ADDITIONAL SIMULATION RESULTS

Supplementary Figure S1 shows the correlations of three types of genetic predictions with the true genetic contribution to the trait Zb, nongenetic contributions to the trait, ε, and the total phenotype $y = Zb + \varepsilon$. This plot is similar to Figure 1, except that a third type of genetic prediction has included out-of-sample predictions from fivefold cross-validation of a linear model. The cvBLUPs, based on efficient leave-one-out cross-validation of the LMM and the fivefold CV predictions, have similar performance in that they have negligible correlation with nongenetic factors ε, and closely matched correlations to the true genetic contribution to the phenotype Zb and the actual phenotype y.

The cvBLUPs can have a computational advantage over k-fold cross-validation of a mixed model. In standard BLUP calculations from an LMM with N subjects and M SNPs, as in Equations 4 and 5, given that GRM K has already been calculated and that estimates of the variance components $\hat{\sigma}_g^2$ and $\hat{\sigma}_\varepsilon^2$ have already been made, then the computational effort to calculate the standard BLUPs as in Equation 5 will be $O\left(N^3\right)$. In k-fold cross-validation, the BLUP effect sizes $\hat{b}$ will have to be explicitly calculated, and then, the genetic predictions will be made of the held out set. The slow step is the decomposition or inversion of a covariance matrix, which takes time $O\left(\left(\frac{k-1}{k}\right)^3 N^3\right)$.

Repeating this operation k times requires time $O\left(k\left(\frac{k-1}{k}\right)^3 N^3\right)$. So, k-fold cross-validation is $\frac{(k-1)^3}{k^2}$ fold or approximately k-fold slower than a single calculation of BLUPs or cvBLUPs.

SUPPLEMENTARY REFERENCES

Bycroft, C., Freeman, C., Petkova, D., et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203.

Euesden, J., Lewis, C.M., and O'Reilly, P.F. 2014. Prsice: Polygenic risk score software. *Bioinformatics* 31, 1466–1468.

Laakso, M., Kuusisto, J., Stancakova, A., et al. 2017. Metabolic syndrome in men (metsim) study: A resource for studies of metabolic and cardiovascular diseases. *J. Lipid Res.* 58, 481–493.

McVean, G. 2009. A genealogical interpretation of principal components analysis. *PLoS Genet.* 5, e1000686.

Patterson, N., Price, A.L., and Reich, D. 2006. Population structure and eigenanalysis. *PLoS Genet.* 2, e190.

Vilhjálmsson, B.J., Yang, J., Finucane, H.K., et al. 2015. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97, 576–592.

Yang, J., Zaitlen, N.A., Goddard, M.E., et al. 2014. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* 46, 100.