Two-Variance-Component Model Improves Genetic Prediction in Family Datasets

George Tucker,[1,2,3,10] Po-Ru Loh,[2,3,10] Iona M. MacLeod,[4,5] Ben J. Hayes,[5,6,7] Michael E. Goddard,[4,6] Bonnie Berger,[1,8] and Alkes L. Price[2,3,9,*]

Genetic prediction based on either identity by state (IBS) sharing or pedigree information has been investigated extensively with best linear unbiased prediction (BLUP) methods. Such methods were pioneered in plant and animal-breeding literature and have since been applied to predict human traits, with the aim of eventual clinical utility. However, methods to combine IBS sharing and pedigree information for genetic prediction in humans have not been explored. We introduce a two-variance-component model for genetic prediction: one component for IBS sharing and one for approximate pedigree structure, both estimated with genetic markers. In simulations using real genotypes from the Candidate-gene Association Resource (CARe) and Framingham Heart Study (FHS) family cohorts, we demonstrate that the two-variance-component model achieves gains in prediction r^2 over standard BLUP at current sample sizes, and we project, based on simulations, that these gains will continue to hold at larger sample sizes. Accordingly, in analyses of four quantitative phenotypes from CARe and two quantitative phenotypes from FHS, the two-variance-component model significantly improves prediction r^2 in each case, with up to a 20% relative improvement. We also find that standard mixed-model association tests can produce inflated test statistics in datasets with related individuals, whereas the two-variance-component model corrects for inflation.

Introduction

Mixed linear models (MLMs) are widely used for genetic prediction and association testing in genome-wide association studies (GWASs). In prediction, MLMs produce best linear unbiased predictions (BLUPs); BLUP and its extensions were first developed in agricultural genetics[1–4] and have since been applied to human genetics.[5–10] In association testing, MLMs model relatedness and population stratification, correcting for confounding and increasing power over linear regression (essentially by testing association of the residual from BLUP).[11–16] Mixed-model methods harness information from either genetic markers (identity by state [IBS] sharing) or known pedigree relationships. Recent work on the estimation of components of heritability[17] has demonstrated the advantages of a model with two variance components: one component for IBS sharing (corresponding to SNP heritability, h_g^2 [18,19]) and one for approximate pedigree structure, estimated via IBS sharing above a threshold (corresponding to total narrow-sense heritability, h^2 [20]). However, the potential advantages of this model for genetic prediction (or mixed-model association) have not been explored.

Through systematic simulations and analyses of quantitative phenotypes in the Candidate-gene Association Resource (CARe)[21] and Framingham Heart Study (FHS)[22,23] cohorts, we show that the two-variance-component model improves prediction r^2 over single-variance-component (standard BLUP) methods. Our simulations demonstrate that this improvement is achieved both at current sample sizes and for larger sample sizes, and our analyses of real CARe and FHS phenotypes confirm relative improvements in prediction r^2 of up to 20%. We also consider the situation in which phenotypes are available for ungenotyped individuals who are related to the genotyped cohort (e.g., via family history[24,25]) and show that leveraging this additional information for genetic prediction within a two-variance-component model achieves similar gains.

Additionally, we investigate the utility of the two-variance-component model for association testing. We evaluate the standard prospective MLM association statistic[15] in the context of familial relatedness and observe inflation of test statistics over a range of simulation parameters, contrary to previous findings.[11,13–15,26] We show that the two-variance-component model substantially reduces the inflation in simulations and in GWASs of CARe and FHS phenotypes.

Material and Methods

Overview of Methods

We use the two-variance-component model described in previous work on the estimation of components of heritability.[17] The first variance component is the usual genetic relationship matrix (GRM) computed from genetic markers (corresponding to h_g^2).[18] The second variance component is a thresholded version of the

[1]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; [2]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA; [3]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [4]Faculty of Veterinary and Agricultural Science, University of Melbourne, Melbourne, VIC 3010, Australia; [5]Dairy Futures Cooperative Research Centre, La Trobe University, Bundoora, VIC 3083, Australia; [6]BioSciences Research Division, Department of Environment and Primary Industries, Melbourne, VIC 3083, Australia; [7]Biosciences Research Centre, La Trobe University, Melbourne, VIC 3083, Australia; [8]Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA; [9]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA
[10]These authors contributed equally to this work
*Correspondence: aprice@hsph.harvard.edu

GRM in which pairwise relationship estimates smaller than a threshold t are set to zero. The idea is to capture strong relatedness structure, similarly to a pedigree relationship matrix. (If full pedigree information is available, the pedigree relationship matrix can also be used directly.) Explicitly modeling relatedness in this way allows the two-variance-component mixed model to capture additional heritability from untyped SNPs (corresponding to $h^2 - h_g^2$).[17] We used the two-variance-component model to compute genetic predictions via BLUP and to compute test association statistics via a Wald test.[1,11,27] (We note that best linear unbiased prediction, BLUP, is a general method for prediction that can be applied once a covariance model has been established, whether from one or many variance components. We will therefore use "standard BLUP" to refer to BLUP with the GRM as a single variance component, and we will use "BLUP" to more generally refer to BLUP with any number of variance components.) We further developed methods to treat the situation in which phenotypes for ungenotyped relatives are available; in brief, our approach uses pedigree information to impute the missing information.[28] Full mathematical details are provided below and in the Appendix, and we have released an open source Matlab implementation of these methods (FAMBLUP; see Web Resources).

Standard Mixed Model for Prediction

We begin by establishing notation and reviewing standard formulas for mixed-model prediction (i.e., standard BLUP) and association testing with one variance component.[1,11,27] Let N be the number of individuals in the study and M be the number of genotyped SNPs. Denote phenotypes by y, fixed-effect covariates by X, and normalized genotypes by W, all of which are mean-centered. We normalize each genotype by dividing by $\sqrt{2\widehat{p}(1-\widehat{p})}$, where $\widehat{p}$ is the empirical minor allele frequency (MAF).[18] We model phenotypes by using the following mixed model:

$$y = Xb + g + \epsilon, \qquad \text{(Equation 1)}$$

where $g \sim N(0, \Sigma_g)$ is a random-effect term modeling genetic effects, $\epsilon \sim N(0, \sigma_e^2 I)$ is a random-effect modeling noise, and b is a vector of coefficients for the fixed effects. In the standard marker-based mixed model, we assume $g = W\alpha$ is a linear combination of genotyped SNPs, where α is an M-vector of independent and identically distributed (iid) normal SNP effect sizes (the infinitesimal model), so that

$$y = Xb + W\alpha + \epsilon. \qquad \text{(Equation 2)}$$

Then the genetic covariance satisfies $\Sigma_g = \sigma_g^2 WW^T/M$, where WW^T/M is the GRM and σ_g^2 and σ_e^2 are variance parameters typically estimated via restricted maximum likelihood (REML).[29] In pedigree-based models that do not use marker information, $\Sigma_g = \sigma_h^2 \Theta$, where Θ is the pedigree relationship matrix, σ_h^2 and σ_e^2 are again estimated via REML.

These models naturally yield formulas for standard BLUP prediction.[1] Explicitly, if we denote training individuals (i.e., those with observed phenotypes) with subscript $-i$ and test individuals (i.e., those with phenotypes to be predicted) with subscript i, predictions are given by

$$\widehat{y}_i = \sigma_g^2 W_i W_{-i}^T \left(\sigma_g^2 W_{-i} W_{-i}^T + \sigma_e^2 I \right)^{-1} \left(y_{-i} - X_{-i} b \right) + X_i b.$$
$$\text{(Equation 3)}$$

Standard Mixed-Model Association Test

To test a candidate SNP (w) for association with the phenotype (y), we augment the marker-based model by including w as an additional fixed-effect covariate:

$$y = w\beta + Xb + W\alpha + \epsilon, \qquad \text{(Equation 4)}$$

where β is the coefficient for the SNP (w) and we wish to test whether $\beta \neq 0$. To do so, we estimate the variance parameters (σ_g^2, σ_e^2) by using REML and estimate the fixed-effect coefficients (β, b) by using maximum likelihood.[27] We then compute the Wald statistic to test $\beta \neq 0$ as follows. Let

$$V = \widehat{\sigma_g^2} WW^T/M + \widehat{\sigma_e^2} I \qquad \text{(Equation 5)}$$

denote the total phenotypic covariance and let $Q = [w; X]$ denote the combined fixed effects. Then $\widehat{\beta}$ is equal to the first entry of $(Q^T V^{-1} Q)^{-1} Q^T V^{-1} y$ and $\mathrm{var}(\widehat{\beta})$ is equal to the first entry of $(Q^T V^{-1} Q)^{-1}$. The Wald chi-square test statistic is given by

$$\chi_{\mathrm{Wald}}^2 = \frac{\widehat{\beta}^2}{\mathrm{var}\left(\widehat{\beta}\right)} \qquad \text{(Equation 6)}$$

and is asymptotically χ^2 distributed with 1 degree of freedom (df) under the null distribution. (We note that, theoretically, comparing the square root of Equation 6 to a t distribution is more precise, but in practice, the distinction is negligible at sample sizes of many thousands.)

We make one slight modification to the above association test to avoid proximal contamination (i.e., masking of the association signal by SNPs included in the random-effects term that are in linkage disequilibrium [LD] with the SNP being tested). Specifically, we use a leave-one-chromosome-out procedure in which, when testing SNP w, we exclude all SNPs on the same chromosome as w from the genotype matrix W used to model random genetic effects.[15,30,31]

Two-Variance-Component Mixed Model

Our use of a two-variance-component mixed model is motivated by the idea that in a sample containing related individuals, the pedigree relationship matrix (or an approximation thereof) can model additional heritable variance explained by untyped SNPs.[17] More precisely, consider expanding the marker-based model (Equation 2) to

$$y = Xb + W\alpha + U\gamma + \epsilon, \qquad \text{(Equation 7)}$$

where $U\gamma$ is the analog of $W\alpha$ for untyped SNPs, U, so that the total genetic effect is $g = W\alpha + U\gamma$. Ideally, we would use this model for prediction and its augmentation for association testing, but U is unobserved. Because the BLUP and Wald statistic formulas only require UU^T, however, we can still improve upon the standard model (Equation 2) by using an approximation of UU^T. If we let M_h denote the number of untyped SNPs, the matrix UU^T/M_h is the realized relationship matrix from untyped SNPs. Assuming a fixed pedigree relationship matrix (Θ) we have

$$E\left[UU^T/M_h\right] = \Theta, \qquad \text{(Equation 8)}$$

where the expectation is computed over possible realizations of genotypes passed down by descent (e.g., siblings share half of their genomes on average). When the study samples include close relatives, off-diagonal entries of Θ can be large, in which case these

entries are good approximations of the corresponding entries of UU^T/M_h and hold additional information not fully harnessed by models that use only the usual GRM from typed SNPs (WW^T/M). Substituting Θ for UU^T/M_h gives the model

$$y \sim N\left(Xb, \sigma_g^2 WW^T \middle/ M + \sigma_h^2 \Theta + \sigma_e^2 I\right). \qquad \text{(Equation 9)}$$

In our case, the pedigree relationship matrix (Θ) is also unavailable, so we need to make a further approximation in which we replace Θ with the estimator

$$\Theta \approx \left(WW^T/M\right)_{>t}, \qquad \text{(Equation 10)}$$

obtained from the usual GRM by keeping only those entries larger than a threshold t and setting all other entries to zero.[17] This approximation gives the model

$$y \sim N\left(Xb, \sigma_g^2 WW^T \middle/ M + \sigma_h^2 (WW^T/M)_{>t} + \sigma_e^2 I\right). \quad \text{(Equation 11)}$$

In theory, the optimal threshold (t) depends on M, N, and the amount of relatedness in the dataset, but in our genetic prediction analyses using human datasets, we found that the results were robust to the choice of t, so we set $t = 0.05$. For association testing, we found $t = 0.05$ to generally be robust (and we expect this choice to be appropriate in human genetics settings), but in more extreme simulation scenarios in which we built the GRM from only a few chromosomes, we observed that higher thresholds were required to model relatedness accurately enough to produce well-calibrated statistics. We therefore optimize t in all association analyses (all of which we conduct by using a leave-one-chromosome-out procedure[15,30,31]) by using the following approach. For each chromosome c in turn, we choose t to minimize the deviation between the thresholded GRM, $(W_{-c}W_{-c}^T/M_{-c})_{>t}$, computed with all chromosomes but c, and the GRM, $W_c W_c^T/M_c$, computed on the left-out chromosome c. We measure this deviation with the Frobenius norm

$$\left\|W_c W_c^T \middle/ M_c - \left(W_{-c}W_{-c}^T/M_{-c}\right)_{>t}\right\|_2^2, \qquad \text{(Equation 12)}$$

i.e., the sum of squared differences between matrix entries. Prediction and association testing proceed as before, once the threshold (t) has been set: we estimate σ_g^2, σ_h^2, and σ_e^2 by REML to enable calculation of BLUP predictions, and for association testing, we again introduce an additional fixed-effect term, $w\beta$, for the SNP being tested and construct a Wald statistic. (Again, for computational efficiency, we apply a leave-one-chromosome-out procedure within which we reuse variance parameters fitted once per left-out chromosome.[13,15,16]) We note that the computation of predictions, $\hat{y}$, can no longer be expressed as a simple matrix-vector product of genotypes with a vector ($\hat{\beta}$) of SNP weights, as is the case for standard (one-variance-component) genomic BLUP. Instead, the formula for $\hat{y}$ (given in the Appendix) involves two terms, only one of which has the above form; the other involves combining training and testing genotypes and has substantially greater computational cost ($O(N^3)$). Although performing prediction with the first term alone would be computationally efficient, we found that such an approach yields suboptimal results; see the Appendix for details.

We have released Matlab code that implements these methods in a stand-alone software package, FAMBLUP (see Web Resources). Although our implementation uses standard $O(N^3)$-time eigendecomposition-based variance components methods, we have taken care to optimize its central processing unit (CPU) and memory use:

for example, FAMBLUP association analysis of $N = 20,000$ samples requires ≈ 16 GB RAM and ≈ 1 single-threaded CPU day per chromosome. Memory usage scales with N^2, and computation time scales with N^3, so for $N = 30,000$ samples, the requirements are ≈ 36 GB RAM and ≈ 4 CPU days per chromosome. These computations are automatically multithreaded on multi-core machines and can be parallelized across chromosomes.

Extension to Ungenotyped Individuals

In the Appendix, we derive extensions of two-variance-component mixed-model prediction and association testing to make use of data available from additional phenotyped but ungenotyped relatives of genotyped individuals. In this setup, we assume that the full pedigree relationship matrix (containing both typed and untyped individuals) is known, whereas some entries of the SNP GRM—namely, those in rows or columns corresponding to untyped individuals—are unknown. Our procedure amounts to replacing the unobserved GRM with the expected GRM (based on known pedigree),[32–34] similar in spirit to regression imputation; mathematical derivations are presented in the Appendix.

CARe and FHS Datasets

We analyzed 8,367 African-American CARe samples from the ARIC, CARDIA, CFS, JHS, and MESA cohorts, comprising high-quality genotypes at 770,390 SNPs from an Affymetrix 6.0 array; the CARe dataset and quality-control procedures used to obtain the sample and SNP sets we analyzed are described in Lettre et al. and Pasaniuc et al.[21,35] We analyzed all samples in analyses of simulated phenotypes (for which real genotypes were used); in analyses of real CARe phenotypes—including BMI, height, high density lipoprotein cholesterol (HDL), and low density lipoprotein cholesterol (LDL), each available for 5,000–8,000 samples—we removed outlier individuals with phenotype values in the top or bottom 0.1%, individuals younger than 18 years old, and individuals with missing age or sex, given that regression results can be quite sensitive to outliers; we then applied a Box-Cox transformation to remove skewness. We analyzed 7,476 FHS SHARe samples with high-quality genotypes at 413,943 SNPs from an Affymetrix 500K array and with BMI and height phenotypes available; the FHS dataset and quality-control procedures are described in Dawber et al., Splansky et al., and Chen et al.[22,23,36] Our analyses were performed under the oversight of the Harvard institutional review board.

Genetic Prediction: Simulations with Real Genotypes

To assess the accuracy of genetic prediction methods, we simulated phenotypes based on genotypes from the CARe and FHS datasets; both CARe and FHS are family studies containing many close relatives. Because the CARe individuals are admixed, we projected out the first five principal components (equivalent to including them as fixed-effect covariates[29]) from genotypes and phenotypes in all analyses of both CARe and FHS data to avoid confounding from population structure.[37] We simulated phenotypes by generating causal effects for two subsets of SNPs: a set of M "observed SNPs," which we used for both phenotype simulation and BLUP prediction, and a set of M_h "untyped SNPs," which we used for phenotype simulation but did not provide to prediction methods. In this simulation framework, the standard GRM built by MLM methods accurately models variation due to observed SNPs, but direct or inferred pedigree information is necessary to capture variation due to untyped SNPs. We generated

effect sizes for observed and untyped SNPs from independent normal distributions $N(0, h_g^2/M)$ and $N(0, (h^2 - h_g^2)/M_h)$, respectively, where h_g^2 denotes heritability explained by observed SNPs and h^2 denotes total narrow-sense heritability. To build phenotypes, we multiplied the simulated effect sizes with the genotypes and added random noise from $N(0, (1 - h^2))$. We used SNPs on chromosome 1 as untyped SNPs and SNPs on varying subsets of chromosomes 2–22 as observed SNPs so as to simulate different values of N/M (which is a key quantity affecting performance of mixed-model prediction[38] and association[15]) and thereby estimate projected performance at larger N. We used $h_g^2 = 0.25$ and $h^2 = 0.5$ as typical values of these parameters.[39]

We note that under the above setup, untyped SNPs are completely untagged by typed SNPs, whereas in real data, untyped SNPs might be partially tagged by typed SNPs. In either case, the phenotype can be written as a sum of "genetic value explained by typed SNPs," "remaining genetic value," and "environmental value" (with variance parameters h_g^2, $h^2 - h_g^2$, and $1 - h^2$ corresponding to the same covariance structures in either case), so we expected that our results would be insensitive to this distinction. To verify this expectation, we also performed a set of simulations in which we selected the set of observed SNPs to be the 90% of SNPs with highest MAF and the set of untyped SNPs to be the 10% of SNPs with lowest MAF—similar to the simulation framework of Yang et al.[18]—to produce a realistic gap between h_g^2 and h^2 as a result of untyped rare variants. (The MAF cutoff corresponding to this split was 5.4%.)

Genetic Prediction: Simulations with Simulated Genotypes

To assess the potential performance of genetic prediction methods at extremely large sample sizes, we also simulated genotypes for sets of sib-pairs (relatedness = 0.5) with $M = 100$ SNPs and $N/M = 10, 20, \ldots, 100$. We generated unlinked markers for simplicity by randomly generating MAFs uniformly in the interval $[0.05, 0.5]$ and sampling genotypes of unrelated individuals from a binomial distribution with the generated MAF. For sib-pairs, with probability 0.5, the pair shared an allele drawn randomly; otherwise, the alleles for the pair were drawn independently. (We ran this procedure twice per SNP to create diploid genotypes.) We simulated phenotypes as above.

Genetic Prediction: Assessing Performance on Real Phenotypes

To compare the predictive performance of the two-variance-component model versus standard BLUP on real phenotypes, we performed cross-validation studies in which we repeatedly selected 10% of the phenotyped samples (from either CARe or FHS) as test data and used the remaining 90% of samples to train each predictor. For each training/test split (s) we thus obtained a pair of observed prediction r^2 values ($r_{2VC,s}^2$, $r_{BLUP,s}^2$). We then estimated the improvement of the two-variance-component model over standard BLUP as

$$\widehat{r_{2VC}^2 - r_{BLUP}^2} = \mathrm{mean}\left(r_{2VC,s}^2 - r_{BLUP,s}^2\right), \qquad \text{(Equation 13)}$$

where the mean is taken over the random splits (s). We estimated the SE of this quantity as:

$$\mathrm{SE}\left(\widehat{r_{2VC}^2 - r_{BLUP}^2}\right) \approx \mathrm{SD}\left(r_{2VC,s}^2 - r_{BLUP,s}^2\right)\Big/\sqrt{10}. \qquad \text{(Equation 14)}$$

The numerator is the SD of the per-split differences in r^2 (across random 90% training and 10% test set splits [s]), which measures the variability in observed performance differences between the two methods when assessed on 10% of the data. The division by $\sqrt{10}$ accounts for the 10× larger sample size of the full dataset. This estimate is approximate due to the complexities of estimating variance under cross-validation (specifically, the overlap among different test sets and among different training sets); in general, unbiased estimators of variance under cross-validation do not exist.[40]

Finally, in the Results section, we estimate relative improvements and SEs (i.e., we divide Equations 13 and 14 by the estimated baseline, mean($r_{BLUP,s}^2$)) to put our absolute estimates in context.

Association Testing: Simulations with Simulated Genotypes

We conducted a suite of mixed-model association simulations by using genotypes simulated in a similar manner as above. We systematically varied the number of related individuals, the degree of relatedness, the number of markers (M) in the genome, and the SNP heritability (h_g^2) and total heritability (h^2) of the simulated trait. Specifically, we simulated sets of $N = 1,000$ diploid individuals, in which $N_{rel} = 50, 125, 250,$ or 500 pairs of individuals who were related (leaving 900, 750, 500, or 0 unrelated individuals, respectively). Each pair of individuals shared a proportion, namely $p = 0, 0.1, 0.2, 0.3, 0.4,$ or 0.5, of their genomes in expectation. Additionally, we varied the number of markers, using $M = 1,000,$ 5,000, 10,000, or 20,000. We generated unlinked markers as above; for pairs of related individuals, with probability equal to the relatedness (p), the pair shared an allele drawn randomly; otherwise, the alleles for the pair were drawn independently. (As above, we ran this procedure twice per SNP to create diploid genotypes.) We also generated 100 additional candidate causal SNPs and 500 candidate null SNPs (at which to compute association test statistics) in the same way. We used an infinitesimal model to generate the phenotype: that is, we generated effect sizes for the observed SNPs from $N(0, h_g^2/M)$. We also generated effect sizes for the candidate causal SNPs from $N(0, (h^2 - h_g^2)/100)$. Because these SNPs are distinct from the M SNPs used for model building, they effectively served as untyped causal loci. Finally, we formed the phenotype by multiplying the effect sizes with the genotypes and adding independent noise distributed as $N(0, (1 - h^2)I)$.

Association Testing: Simulations with Real Genotypes

We also assessed mixed-model association methods in simulation studies by using simulated phenotypes based on genotypes from the CARe and FHS datasets. To avoid proximal contamination,[15,30,31] we tested SNPs on chromosomes 1 and 2 for association and used M observed SNPs on subsets of chromosomes 3–22 to compute GRMs, varying the number of chromosomes used in order to vary N/M. We generated quantitative phenotypes in which observed SNPs collectively explained 25% of variance and 250 causal SNPs from chromosome 1 explained another 25% of variance; all SNPs on chromosome 2 were null SNPs.

Results

Genetic Prediction: Simulations

To analyze the predictive power of the two-variance-component model, we simulated phenotypes based on

Table 1. Prediction Accuracy for Simulations Using CARe and FHS Genotypes

	Prediction $r^2(g)$		
Observed SNPs	BLUP	BLUP w/ Thresh.	2VC BLUP
CARe genotypes			
chr 2–22	0.062 (0.002)	0.061 (0.002)	0.071 (0.002)
chr 3–6	0.084 (0.002)	0.063 (0.002)	0.094 (0.002)
chr 3–4	0.098 (0.002)	0.059 (0.002)	0.108 (0.002)
FHS genotypes			
chr 2–22	0.225 (0.003)	0.225 (0.003)	0.238 (0.003)
chr 3–6	0.246 (0.003)	0.230 (0.003)	0.269 (0.003)
chr 3–4	0.263 (0.003)	0.231 (0.003)	0.291 (0.003)

Phenotypes were simulated to have $h^2 = 0.5$ and $h_g^2 = 0.25$, and we measured prediction $r^2(g)$ by using a random 90% of samples as training data and the remaining 10% as test data. Reported values are mean prediction $r^2(g)$ and SEM over 100 independent simulations (in which phenotypes were re-simulated and training-test splits resampled). "BLUP w/ Thresh." denotes BLUP prediction using the thresholded relationship matrix instead of the standard approach of using the GRM (denoted simply as "BLUP"). "2VC BLUP" denotes two-variance-component BLUP. "Prediction $r^2(g)$" denotes r^2 between predicted phenotypes and true genetic components of the simulated phenotypes.

genotypes from the CARe and FHS datasets as described in Material and Methods. In each simulation, we used the following procedure to measure the prediction accuracies of BLUP with the standard GRM as a single variance component, BLUP with the thresholded GRM as a single variance component, and BLUP with the two-variance-component model. First, we simulated phenotypes for all individuals (independently for each simulation replicate). Second, we randomly split the dataset, setting aside 90% of the individuals for training and 10% for testing. We then used each BLUP method to predict held-out test phenotypes by using the training samples to estimate genetic effects, and we calculated r^2 between the predicted phenotypes and the true genetic components of the simulated phenotypes (i.e., eliminating the added noise) on the test samples. (We chose to compute r^2 because it is a very widely used metric for assessing prediction accuracy;[2–4,6,7,9] however, other metrics such as mean square error are also sometimes used.[5]) We call this quantity "prediction $r^2(g)$"; on average, prediction $r^2(g)$ is $1/h^2$ times as large as standard prediction r^2, i.e., r^2 computed to simulated phenotypes that include both genetic and noise components. Relative performance of prediction methods is the same (on average) whether measured with prediction r^2 or prediction $r^2(g)$.

The two-variance-component model provided significant increases in $r^2(g)$ over both standard BLUP and BLUP using the thresholded GRM alone (Table 1), and the improvements were consistent across simulation replicates (Figure S1). We observed much larger prediction $r^2(g)$ values (across all methods) for the FHS simulations than for the CARe simulations, as expected given the much greater number of close relatives in the FHS dataset (18,415 pairs of individuals with genetic related-

ness > 0.2 among 7,476 FHS individuals versus 4,954 pairs among 8,367 CARe individuals). However, the relative improvements achieved by the two-variance-component model were fairly similar in these two distinct pedigree structures, and importantly, increasing values of N/M (mimicking larger sample sizes) also yielded similar relative improvements (Table 1). We also observed that the heritability parameter estimated by the standard mixed model was intermediate to h_g^2 and h^2, whereas the two-variance-component model accurately estimated h_g^2 and $h^2 - h_g^2$ (Table S1), as expected in samples with related individuals.[17] (We note that because the sum of the entries of the thresholded GRM is nonzero, we used the general formula given in Speed et al.[41] to estimate heritability parameters.) We also verified that in simulations with no untyped causal SNPs, the two-variance-component model produced no improvement over standard BLUP, indicating that our cross-validation scheme was immune to differences in model complexity (Table S2). Finally, we verified that simulations involving linkage disequilbrium between typed and untyped SNPs (achieved by setting typed SNPs to be the 90% of SNPs with highest MAF and untyped SNPs to be the 10% of SNPs with lowest MAF) produced similar results (Table S3). In these simulations, we also varied the fraction of heritability explained by typed versus untyped SNPs, and we observed that the two-variance-component model achieved larger gains for $h_g^2 \ll h^2$ and smaller gains for h_g^2 approaching h^2 (Table S3), consistent with our intuition that, if typed SNPs explain most of heritable variance, prediction using only typed SNPs achieves most of the available predictive power.

We further assessed the potential performance of the two-variance-component approach at very large values of N/M (up to 100) by simulating both genotypes and phenotypes (Material and Methods). (We note that human genotyping arrays typically contain $\approx$ 60,000 independent SNPs,[15,42] so $N/M = 8$ in this simulation corresponds to a dataset the size of UK Biobank, $N = 500,000$; see Web Resources.) In these simulations, we continued to observe gains when using the two-variance-component approach; two-variance-component prediction r^2 exceeded h_g^2 for very large N, whereas standard BLUP prediction r^2 was limited to less than h_g^2 (Figure S2).

Genetic Prediction: Real Phenotypes
Next, we evaluated the prediction accuracy of each method on CARe phenotypes—BMI, height, LDL, and HDL—and on FHS phenotypes—height and BMI. We adjusted phenotypes for age, sex, study center (for CARe phenotypes), and the top five principal components. (The complexities of the impact of ancestry on genetic prediction are discussed in Chen et al.[43]) To measure performance, we created 100 independent random 90%/10% splits of the dataset, as before, and calculated r^2 between predicted and true phenotypes on the test samples of each split. We observed that, for all phenotypes, the two-variance-component model increased prediction accuracy

Table 2. Prediction Accuracy for CARe and FHS Phenotypes

Phenotype	Prediction r^2			Prediction r^2 Relative to BLUP (SE)	
	BLUP	BLUP w/ Thresh.	2VC BLUP	BLUP w/ Thresh.	2VC BLUP
CARe prediction					
BMI	0.023	0.027	0.029	+14% (9%)	+18% (5%)
height	0.063	0.067	0.079	+5% (5%)	+20% (3%)
LDL	0.017	0.017	0.019	+2% (15%)	+11% (5%)
HDL	0.034	0.032	0.038	−7% (10%)	+11% (4%)
FHS prediction					
BMI	0.103	0.104	0.107	+1.0% (2.3%)	+3.5% (1.2%)
height	0.344	0.342	0.354	−0.7% (1.1%)	+2.9% (0.5%)
CARe prediction with genome-wide significant SNPs as fixed-effect covariates					
BMI	0.023	0.026	0.028	+14% (9%)	+19% (5%)
height	0.063	0.066	0.078	+5% (5%)	+20% (3%)
LDL	0.038	0.039	0.041	+3% (6%)	+6% (2%)
HDL	0.051	0.049	0.055	−4% (6%)	+7% (3%)
FHS prediction with genome-wide significant SNPs as fixed-effect covariates					
BMI	0.105	0.107	0.109	+1.2% (2.3%)	+3.5% (1.2%)
height	0.344	0.341	0.354	−0.8% (1.1%)	+2.8% (0.5%)

Prediction r^2 values are means over 100 random 90% training and 10% test data splits. Relative performance values reported are ratios of means minus 1; SEs are estimated as SDs of per-split differences in r^2 (over the random 10% test sets) divided by $\sqrt{10}$ (to account for the 10× larger sample size of the full dataset; see Material and Methods). "BLUP w/ Thresh." denotes BLUP prediction using the thresholded relationship matrix instead of the standard approach of using the GRM (denoted simply as "BLUP"). "2VC BLUP" denotes two-variance-component BLUP.

over both single-variance-component BLUP approaches, with a maximum relative improvement of 20% for height (Table 2); this improvement was consistent across different training/test splits (Figure S4). We observed no significant difference between the performance of the the two single-variance-component BLUP approaches (Table 2). As in our simulations, we observed a larger absolute prediction r^2 in FHS than in CARe, due to strong relatedness (consistent with de los Campos et al.[6]), and we observed that the heritability parameter estimated by the standard mixed model was intermediate to the heritability parameters $\widehat{h}_g^2$ and $\widehat{h}^2$ estimated by the two-variance-component model (Table S4). We verified in the CARe data that evaluating prediction accuracy by using the mean square error metric produced near-identical results (Table S5).

For phenotypes with a small number of large-effect loci, methods that explicitly model a non-infinitesimal genetic architecture can have substantially better prediction accuracy than standard BLUP.[2] A two-variance-component approach could be combined with such models, and as an initial exploration of this approach, we examined a non-infinitesimal extension of two-variance-component BLUP in which we included large-effect loci as fixed-effect covariates.[8] Explicitly, we first identified genome-wide significant SNPs ($p < 5 \times 10^{-8}$) according to a two-variance-component mixed-model association statistic. (As we show below, the standard MLM statistic is miscalibrated

in scenarios with pervasive relatedness, precluding its use.) We then added these SNPs as fixed-effect covariates in all of the models we previously compared and recomputed predictions (Table 2). Including large-effect loci resulted in substantial improvements in prediction r^2 achieved by each model for the CARe HDL and LDL phenotypes (Table 2), both of which are known to have several large-effect loci.[44] As before, for all phenotypes, we observed an increase in r^2 when using the two-variance-component model. We expect that the two-variance-component model will provide similar improvements in prediction r^2 if incorporated in more sophisticated non-infinitesimal models (e.g., Erbe et al. and Zhou et al. [3,5]).

Additionally, we explored the scenario in which some phenotypes are available for ungenotyped relatives of genotyped individuals. We simulated data with ungenotyped individuals by randomly masking the genotypes of 25% of the training individuals. Results on simulated and real phenotypes when using this masking were broadly consistent with the results reported above in which all individuals were typed (Tables S6–S9).

Association Testing

We next compared mixed-model association testing using the two-variance-component approach to standard MLM association testing[12,15] in datasets with related individuals, measuring calibration and power for each method. We

began by running a suite of tests using simulated genotypes and phenotypes, systematically varying the number of related individuals, the degree of relatedness, the number of markers in the genome, and the heritability of the simulated trait (see Material and Methods). Each simulation included both causal SNPs and "null SNPs," i.e., SNPs with no phenotypic effect. For null SNPs, Wald statistics computed by mixed-model association tests follow a 1 df chi-square distribution, under the assumption that the mixed model accurately models the phenotypic covariance. If the mixed model does not accurately model the covariance, as we expect for phenotypes with $h_g^2 < h^2$ in datasets containing relatedness, then the distribution of association statistics at null SNPs is miscalibrated, i.e., approximately follows a scaled 1 df chi-square.[45] We therefore measured calibration of MLM association methods by computing the mean Wald statistic over null SNPs. We measured power by dividing the mean Wald statistic over causal SNPs by the mean Wald statistic over null SNPs. Computing the ratio in the latter benchmark ensured that all methods, including those susceptible to inflation of test statistics, were equally calibrated before we compared power.

Contrary to previous work suggesting that mixed models fully correct for relatedness,[11,13–15,26] we found that for many parameter settings, standard MLM association analysis produced significantly inflated test statistics (up to 11% inflation, increasing with trait heritability, sample size, and extent of relatedness; Figure 1). In contrast, introducing a second variance component—either the thresholded GRM (Figure 1) or the true pedigree (Figure S4)—nearly eliminated the inflation. For all parameter settings, we observed that, compared to standard MLM association, the two-variance-component model maintained or slightly increased power (Figure S4).

Next, we simulated phenotypes based on genotypes from the CARe and FHS datasets (Material and Methods). Consistent with the previous simulations, standard MLM association produced inflated statistics (as measured in test statistics from chromosome 2, simulated to contain no causal SNPs) whereas the two-variance-component model alleviated inflation (Table 3; also see type I errors in Table S10). Importantly, these results suggest that the levels of relatedness that are required for inflation are present in real datasets.

Finally, we analyzed MLM association statistics for the CARe and FHS phenotypes (adjusted for covariates as before). Because we do not know the identity of causal and null SNPs in this case, we calculated the average Wald statistic over all SNPs by using leave-one-chromosome-out analysis,[15,30] noting that we expected the statistics to be slightly larger than 1 due to polygenicity.[15,42] Consistent with simulations, the average Wald statistics were higher for standard MLM association than for the two-variance-component method, suggesting that standard MLM statistics are slightly inflated, with an up to 1.05-fold inflation in FHS data (Table 4). Analysis of

genomic inflation factors λ_{GC}[46] corroborated these results (Table 4). We also compared our test statistics (which involve approximations, as in previous work[13,15,16]; see Material and Methods) to exact likelihood ratio test statistics under the two-variance-component model and verified that the approximate versus exact statistics were near identical ($r^2 = 0.999997$; Figure S5).

Discussion

We have shown that a mixed model with two variance components, one modeling genetic effects of typed SNPs and the other modeling phenotypic covariance from close relatives, offers increased prediction accuracy over standard BLUP and corrects miscalibration of standard mixed-model association analysis in human datasets containing strong relatedness. For current sample sizes and levels of relatedness, the absolute increase in prediction accuracy is modest (similar to other recent work on improving prediction accuracy for human complex traits[5,7–10] in contrast to agricultural traits[2–4]) and the inflation of standard mixed-model test statistics is small. However, our simulations suggest that, for larger sample sizes, the effects of relatedness will become more pronounced, so we expect the two-variance-component model to become increasingly relevant as sample sizes increase.

Although we are not aware of prior work in human genetics that involves using two variance components to model effects of typed SNPs as well as additional phenotypic covariance from close relatives, other methods for combining these two sources of information for prediction have been proposed; however, these methods either use only a limited number of genome-wide significant SNPs[24] or use only limited information about family history.[25] Separately, several studies have applied different multiple-variance-component models to improve mixed-model prediction and association in other ways. Widmer et al.[26] recently proposed a two-variance-component model that uses the standard GRM along with a GRM created from selected SNPs (as in FaST-LMM-Select[31]) that improves association power and calibration in family studies. (We note that, although Widmer et al. observe that standard mixed-model association is properly calibrated in their simulated family datasets, their simulations do not include untyped causal SNPs.) In another direction, Speed et al.[7] recently proposed a multiple variance component model that partitions SNPs into contiguous blocks, each used in a distinct variance component, and showed that this approach improves prediction accuracy. Incorporation of a variance component modeling relatedness—either from pedigree, thresholding the GRM, or other approaches[47]—into these methods or into recently proposed non-infinitesimal models for genetic prediction (e.g., weighted G-BLUP,[6] BayesR,[3,10] or BSLMM[5]) is a possible direction for future research.
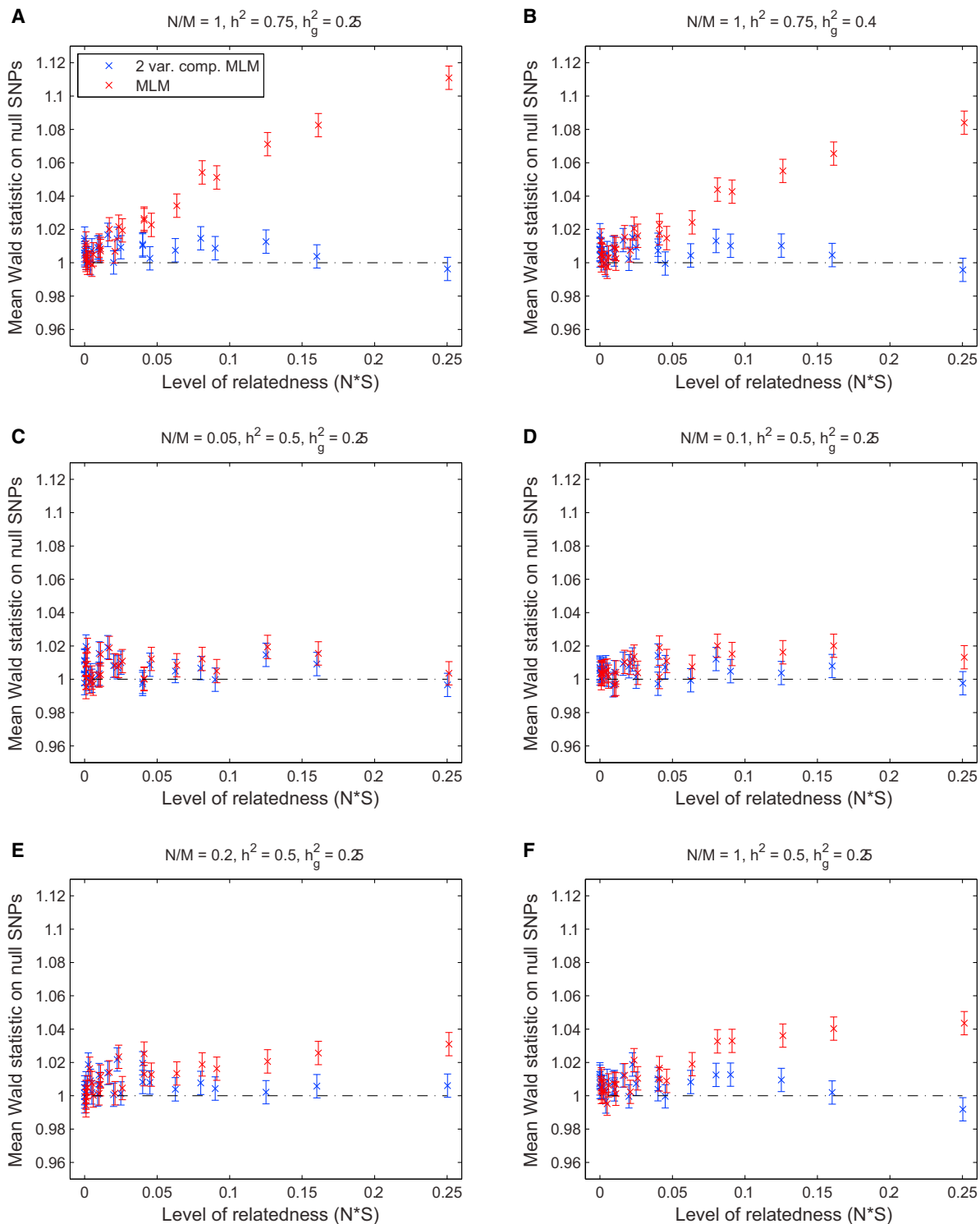
Figure 1. Calibration of Standard and Two-Variance-Component Mixed-Model Association Statistics on Simulated Genotypes and Phenotypes

We computed mean Wald statistics over null SNPs by using the standard mixed-model association test (MLM) and a two-variance-component model (2 var. comp. MLM) using GRM and thresholded GRM (i.e., approximate pedigree) components. Each panel shows results from a set of simulations with selected values of the simulation parameters N/M, h^2, and h_g^2. The set of simulations contained within each panel varies one additional parameter, NS, which measures the amount of relatedness in the simulated data. (S denotes the average squared off-diagonal entry of the pedigree relationship matrix.) Plotted values are mean Wald statistics and SEM over 100 simulations.

Table 3. Calibration of Standard and Two-Variance-Component Mixed-Model Association Statistics in CARe and FHS Simulations

Observed SNPs	No. of SNPs (M)	Standard Mixed Model Mean Wald	Two Variance Components	
			Mean Wald	Threshold (t)
CARe genotypes				
chr 3–22	615,445	1.013 (0.002)	1.000 (0.002)	0.024
chr 3–6	195,333	1.024 (0.002)	1.002 (0.002)	0.051
chr 3–4	99,690	1.028 (0.002)	1.003 (0.002)	0.081
chr 22	9,713	1.036 (0.002)	1.014 (0.002)	0.387
FHS genotypes				
chr 3–22	346,005	1.032 (0.003)	1.003 (0.003)	0.021
chr 3–6	110,203	1.071 (0.003)	1.008 (0.003)	0.040
chr 3–4	55,480	1.097 (0.003)	1.014 (0.003)	0.055
chr 22	5,277	1.189 (0.004)	1.055 (0.003)	0.258

Mean Wald statistics on candidate null SNPs for simulations with CARe or FHS genotypes and a trait with $h^2 = 0.5$, $h_g^2 = 0.25$. Reported values are means and SEM over 100 simulations. The two-variance-component model selected the specified threshold (t) to estimate the relatedness matrix. In simulations using only SNPs on chromosome 22 to compute GRMs, we observed slight inflation when using the two-variance-component model; given the large thresholds ($t > 0.25$) chosen by the model in these scenarios, we hypothesize that the number of SNPs was too small to distinguish relatedness from noise in the GRM, causing an incomplete correction. For corresponding type I error at different α levels, see Table S10.

A challenge facing all genetic prediction methods is the very large sample sizes that will be required to achieve clinically relevant prediction accuracy.[25,48] Indeed, in absolute terms, the prediction accuracy we achieved on real datasets of up to 8,000 samples was low, similar to other methods when applied to traits without large-effect loci.[5,6,10] Our simulations show that the two-variance-component approach we have proposed will maintain its relative improvement over standard BLUP as sample sizes increase; however, both of these methods face computational barriers at large N. (Standard BLUP does have the advantage that $O(N^3)$-time computation is required only for fitting the model but not for computing predictions on new samples; in contrast, a straightforward implementation of our two-variance-component method for prediction requires $O(N^3)$ time per REML iteration when estimating variance parameters as well as when computing predictions, a consequence of the need to combine training and target genotypes.) These limitations could potentially be overcome by using a combination of rapid relationship inference,[49] fast multiple-variance-component analysis (e.g., as implemented in BOLT-REML[50]), and iterative solution of the mixed-model equations.[51,52] Similarly, the computational challenge of large-scale two-variance-component association analysis could potentially be addressed by extending fast iterative methods for mixed-model association.[16] An alternative, computationally simple solution to inflation of association-test statistics is LD score regression;[53] however, this

approach might incur slight deflation as a result of attenuation bias.[16,53]

We also note four additional limitations of our two-variance-component approach. First, the method is only applicable to datasets with related individuals for which genotypes are available for analysis; however, large human datasets of this type are now being generated: deCODE Genetics has genotyped >30% of the Icelandic population,[54] the UK Biobank will soon have genotypes for $N = 500,000$ individuals (close to 1% of the UK population; see Web Resources), and 23andMe has assembled an even larger cohort.[55] Second, the improved predictive performance of the two-variance-component approach is a function of the relatedness structure. Our parallel work in cattle has reported improved prediction accuracy when using a two-variance-component model incorporating exact pedigree information[56] or breed information;[57] however, the two-variance-component model did not produce an improvement in analyses of Holstein dairy cattle (Table S11), perhaps because of the very small effective population size of this breed.[58] Third, although the intuition behind the two-variance-component model is to capture effects of rare variants not tagged by SNP arrays, our observed gains in prediction accuracy could also be partially explained by the approximate pedigree component picking up shared environment or epistasis; as such, care is needed in interpreting fitted variance parameters as heritability estimates.[17] Fourth, our approach does not address case-control ascertainment. Although many large family datasets are not ascertained for phenotype, investigating whether techniques employed by methods that do model ascertainment[8] can be integrated into our two-variance-component approach is a possible avenue for future work.

Appendix A

Formulas for Two-Variance-Component Mixed-Model Prediction

Here, we provide explicit formulas for computing best linear unbiased predictions under the two-variance-component model

$$y \sim N\left(0, WW^T \sigma_g^2 / M + (WW^T / M)_{>t} \sigma_h^2 + I\sigma_e^2\right),$$

combining Equation 9 with the approximation of pedigree by using thresholded IBS, Equation 10, and leaving out fixed effects for simplicity. We assume that we have a set of training individuals (denoted with subscript $-i$) for whom we have both genotype and phenotype information and that we have a set of testing individuals (denoted with subscript i) for whom we have only genotype information and wish to predict phenotypes. Thus, under this notation, W_i denotes the submatrix of genotypes from testing individuals and W_{-i} denotes the submatrix of genotypes from training individuals.

Under the assumption that the variance parameters σ_g^2, σ_h^2, and σ_e^2 have already been fitted (e.g., by using

Table 4. Calibration of Standard and Two-Variance-Component Mixed-Model Association Statistics for CARe and FHS Phenotypes

Phenotype	N	Standard Mixed Model			Two Variance Components			
		Mean Wald	λ_{GC}	$\widehat{h}_g^2$	Mean Wald	λ_{GC}	$\widehat{h}_g^2$	$\widehat{h}^2$
CARe phenotypes								
BMI	7,987	1.044	1.044	0.35	1.029	1.027	0.17	0.46
height	7,988	1.110	1.099	0.73	1.080	1.070	0.38	0.95
LDL	4,965	1.030	1.026	0.32	1.021	1.017	0.18	0.44
HDL	5,184	1.054	1.046	0.50	1.037	1.028	0.26	0.66
FHS phenotypes								
BMI	7,476	1.060	1.058	0.43	1.032	1.032	0.21	0.47
height	7,476	1.126	1.112	0.81	1.070	1.058	0.39	0.87

We report the number of individuals (*N*) phenotyped for each trait and the mean Wald statistics and heritability parameters computed by the standard and two-variance-component mixed models (averaged over 22 leave-one-chromosome-out runs).

REML on the training individuals), the BLUP predictions for the test phenotypes are given by

$$\widehat{y}_i = W_i\widehat{\beta} + \sigma_h^2\left(W_iW_{-i}^T/M\right)_{>t}\left(W_{-i}W_{-i}^T\sigma_g^2\Big/M \right.$$
$$\left. + \left(W_{-i}W_{-i}^T/M\right)_{>t}\sigma_h^2 + I\sigma_e^2\right)^{-1}y_{-i},$$

where

$$\widehat{\beta} = \sigma_g^2W_i^T\left(W_{-i}W_{-i}^T\sigma_g^2\Big/M + \left(W_{-i}W_{-i}^T/M\right)_{>t}\sigma_h^2 + I\sigma_e^2\right)^{-1}y_{-i}.$$

We note that the first term of the formula for $\widehat{y}_i$, namely $W_i\widehat{\beta}$, has the form of a simple matrix-vector product between genotypes of testing individuals and a vector ($\widehat{\beta}$) of SNP weights, as is the case for standard (one-variance-component) genomic BLUP. This term is easy to compute on testing individuals once we have estimated $\widehat{\beta}$ by using the training data, whereas the second term of the prediction formula requires more computation.

This observation suggests the possibility of performing prediction by using only the first term as a computationally efficient alternative to carrying out the full two-variance-component computation. We tested the performance of this approach on the CARe height phenotype but found no significant difference in its performance versus that of standard BLUP: –1% (SE 3%) change in prediction r^2, in contrast to the +20% (SE 3%) change in prediction r^2 of the full two-variance-component approach over standard BLUP (Table 2). This observation indicates that the gain in prediction accuracy achieved by the two-variance-component model is largely a result of capturing effects of rare variants and requires the use of both variance components.

Two-Variance-Component Mixed Model with Ungenotyped Individuals

Here, we derive extensions of two-variance-component mixed-model prediction and association testing to make use of data available from additional phenotyped but ungenotyped relatives of genotyped individuals. In this case, we assume that we are given the pedigree relationship matrix (Θ) among all individuals, both typed and untyped.

We will use subscripts *u* and *t* to denote submatrices of genotype and relationship matrices corresponding to untyped and typed individuals (e.g., W_u is the matrix of [unobserved] genotypes for the ungenotyped individuals, and W_t is the matrix of genotypes for the typed individuals, so that W_u and W_t together comprise the genotype matrix *W*). Because W_u is unknown, we need a distribution on *W* to describe the relationship between the genotypes of typed and untyped individuals. For modeling purposes, we assume that normalized SNPs in W_u are independently drawn from a multivariate normal distribution according to the pedigree structure: $N(0, \Theta)$.

To adapt our pedigree-based two-variance-component model (Equation 9),

$$y \sim N\left(Xb, WW^T\sigma_g^2\Big/M + \Theta\sigma_h^2 + I\sigma_e^2\right),$$

to deal with the fact that *W* now has a subset of unknown entries, we would ideally marginalize over the unobserved genotypes, assuming the above distribution on W_u. However, this approach leads to an intractable integral. Instead, we assume that *y* is normally distributed conditional on the observed genotypes, in which case it suffices to compute the mean and covariance of *y*. It is straightforward to see that the mean of *y* is *Xb* (assuming that we observe the covariates for all individuals), and the covariance of *y* is

$$V := \text{Cov}(y) = \sigma_g^2E\left[WW^T/M \mid W_t\right] + \sigma_h^2\Theta + \sigma_e^2I.$$

Therefore, this procedure amounts to replacing the unobserved GRM with the expected GRM,

$$E_{W_u}\left[WW^T/M \mid W_t\right],$$

where the expectation is over the unobserved genotypes. By using standard properties of the normal distribution, we can compute the required moments

$$E[W_u \mid W_t] = \Theta_{ut}\Theta_{tt}^{-1}W_t$$
$$E\left[W_u W_u^T \mid W_t\right] = M\Theta_{uu} - M\Theta_{ut}\Theta_{tt}^{-1}\Theta_{tu}$$
$$\quad\quad + E[W_u \mid W_t]E\left[W_u^T \mid W_t\right]$$
$$= M\Theta_{uu} + \Theta_{ut}\left(W_t W_t^T - M\Theta_{tt}^{-1}\right)\Theta_{tu}.$$

Thus,

$$E_{W_u}\left[WW^T/M \mid W_t\right] = \begin{bmatrix} \Theta_{uu} + \Theta_{ut}\left(W_t W_t^T/M - \Theta_{tt}^{-1}\right)\Theta_{tu} & \Theta_{ut}\Theta_{tt}^{-1}W_t W_t^T/M \\ W_t W_t^T \Theta_{tt}^{-1}\Theta_{tu}/M & W_t W_t^T/M \end{bmatrix}.$$

Legarra, Misztal, and Aguilar developed the same variance component to incorporate genetic marker information with pedigree information in the context of cattle phenotype prediction.[32–34]

The new phenotype model

$$y \sim N\left(Xb, \sigma_g^2 E[WW^T/M \mid W_t] + \sigma_h^2\Theta + \sigma_e^2 I\right) = N(Xb, V)$$

immediately enables BLUP prediction as before. However, for association testing, the test SNP $w = [w_u; w_t]$ is not completely specified, so we need a novel association statistic that accounts for the uncertainty in w. Assuming that $w \sim N(0, \Theta)$, the model for y simplifies to

$$y \sim N\left(\begin{bmatrix} \Theta_{ut}\Theta_{tt}^{-1}w_t \\ w_t \end{bmatrix}\beta + Xb, \begin{bmatrix} \Theta_{uu} - \Theta_{ut}\Theta_{tt}^{-1}\Theta_{tu} & 0 \\ 0 & 0 \end{bmatrix}\beta^2\right.$$
$$\left. + V\right).$$

and $\tilde{w} = [\Theta_{ut}\Theta_{tt}^{-1}w_t; w_t]$ can be interpreted as the BLUP imputation of the missing genotypes.

This distribution for y gives rise to a score statistic as follows. We start from the log likelihood function

$$\log p(y \mid V, X, W, \tilde{w}, \Theta, \beta) = -0.5\Big(\log \mid V + A\beta^2 \mid$$
$$+ (y - \tilde{w}\beta - Xb)^T\left(V + A\beta^2\right)^{-1}$$
$$\times (y - \tilde{w}\beta - Xb)\Big)$$

(up to a constant that does not depend on y or β), where

$$A = \begin{bmatrix} \Theta_{uu} - \Theta_{ut}\Theta_{tt}^{-1}\Theta_{tu} & 0 \\ 0 & 0 \end{bmatrix}$$

is the adjustment to the variance. Expanding around $\beta = 0$, the log likelihood simplifies to

$$-2\log p(y \mid V, X, W, \tilde{w}, \Theta, \beta) = \log \mid V \mid + \text{trace}\left(V^{-1}A\right)\beta^2$$
$$+ (y - \tilde{w}\beta - Xb)^T V^{-1}(y - \tilde{w}\beta - Xb)$$
$$- (y - Xb)^T\left(V^{-2}A\right)(y - Xb)\beta^2 + O(\beta^3),$$

so the score function is

$$U(\beta) = \frac{\partial \log p(y \mid V, X, W, \tilde{w}, \Theta, \beta)}{\partial \beta}$$
$$= -\Big(\text{trace}\left(V^{-1}A\right)\beta + (y - Xb)^T V^{-1}\tilde{w} + \tilde{w}^T V^{-1}\tilde{w}\beta$$
$$- (y - Xb)^T\left(V^{-2}A\right)(y - Xb)\beta\Big) + O(\beta^2),$$

and

$$\frac{\partial U}{\partial \beta} = -\Big(\text{trace}\left(V^{-1}A\right) + \tilde{w}^T V^{-1}\tilde{w}$$
$$- (y - Xb)^T\left(V^{-2}A\right)(y - Xb)\Big) + O(\beta)$$

Hence, the score statistic to test the hypothesis that $\beta = 0$ is

$$\frac{U(0)^2}{-\partial U/\partial \beta(0)} = \frac{\left(\tilde{w}^T V^{-1}(y - Xb)\right)^2}{\tilde{w}^T V^{-1}\tilde{w} + \text{trace}(V^{-1}A) - (y - Xb)^T(V^{-2}A)(y - Xb)},$$

where the nuisance parameters $(b, \sigma_g^2, \sigma_h^2, \sigma_e^2)$ are set to their maximum likelihood values when $\beta = 0$. The score statistic is asymptotically distributed as χ^2 with 1 df. It is possible to adjust the statistic slightly to produce a nearly equivalent statistic that is easier to compute and precisely χ^2-distributed under the null distribution. Observe that

$$E\left[(y - Xb)^T\left(V^{-2}A\right)(y - Xb)\right] =$$
$$E\left[\text{trace}\left((y - Xb)^T\left(V^{-2}A\right)(y - Xb)\right)\right]$$
$$= \text{trace}\left(E(y - Xb)(y - Xb)^T\right]V^{-2}A\right)$$
$$= \text{trace}(VV^{-2}A)$$
$$= \text{trace}(V^{-1}A),$$

resulting in the simplified statistic

$$\frac{\left(\tilde{w}^T V^{-1}(y - Xb)\right)^2}{\tilde{w}^T V^{-1}\tilde{w}},$$

which is χ^2-distributed under the null distribution. Notably, when V only incorporates pedigree information

(i.e., $V = \Theta\sigma_h^2 + I\sigma_e^2$), we recover a prospective analog to the retrospective MASTOR statistic.[28]

Alternatively, in a similar manner, we can compute a retrospective statistic by considering the score statistic produced by $\log p(w|y)$ instead. Analogous manipulations yield the retrospective statistic

$$\frac{(\tilde{w}^T V^{-1}(y - Xb))^2}{(y - Xb)^T V^{-1}\Theta_{\cdot t}\Theta_{tt}^{-1}\Theta_{t\cdot}V^{-1}(y - Xb)}.$$

Given that the results from the retrospective model and prospective model are similar (data not shown), we focused on the more commonly used prospective approach.

Supplemental Data

Supplemental Data include eleven tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.10.002.

Acknowledgments

Web Resources

The URLs for data presented herein are as follows:

FAMBLUP software (Matlab code), http://github.com/gjtucker/FAMBLUP

UK Biobank, http://www.ukbiobank.ac.uk/

References

1. Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. Biometrics *31*, 423–447.
2. Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics *157*, 1819–1829.
3. Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., Reich, C.M., Mason, B.A., and Goddard, M.E. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. *95*, 4114–4129.
4. Habier, D., Fernando, R.L., and Garrick, D.J. (2013). Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics *194*, 597–607.
5. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264.
6. de los Campos, G., Vazquez, A.I., Fernando, R., Klimentidis, Y.C., and Sorensen, D. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. *9*, e1003608.
7. Speed, D., and Balding, D.J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. *24*, 1550–1557.
8. Golan, D., and Rosset, S. (2014). Effective genetic-risk prediction using mixed models. Am. J. Hum. Genet. *95*, 383–393.
9. Maier, R., Moser, G., Chen, G.-B., Ripke, S., Coryell, W., Potash, J.B., Scheftner, W.A., Shi, J., Weissman, M.M., Hultman, C.M., et al.; Cross-Disorder Working Group of the Psychiatric Genomics Consortium (2015). Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am. J. Hum. Genet. *96*, 283–294.
10. Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLoS Genet. *11*, e1004969.
11. Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. *38*, 203–208.
12. Chen, W.-M., and Abecasis, G.R. (2007). Family-based association tests for genomewide association scans. Am. J. Hum. Genet. *81*, 913–926.
13. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348–354.
14. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.
15. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. *46*, 100–106.
16. Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.
17. Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., and Price, A.L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genet. *9*, e1003520.
18. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
19. de los Campos, G., Sorensen, D., and Gianola, D. (2015). Genomic heritability: what is it? PLoS Genet. *11*, e1005048.
20. Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era–concepts and misconceptions. Nat. Rev. Genet. *9*, 255–266.
21. Lettre, G., Palmer, C.D., Young, T., Ejebe, K.G., Allayee, H., Benjamin, E.J., Bennett, F., Bowden, D.W., Chakravarti, A., Dreisbach, A., et al. (2011). Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. PLoS Genet. *7*, e1001300.

22. Dawber, T.R., Meadors, G.F., and Moore, F.E., Jr. (1951). Epidemiological approaches to heart disease: the Framingham Study. Am. J. Public Health Nations Health *41*, 279–281.

23. Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., D'Agostino, R.B., Sr., Fox, C.S., Larson, M.G., Murabito, J.M., et al. (2007). The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. Am. J. Epidemiol. *165*, 1328–1335.

24. So, H.-C., Kwan, J.S., Cherny, S.S., and Sham, P.C. (2011). Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. Am. J. Hum. Genet. *88*, 548–565.

25. Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S.J., and Park, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat. Genet. *45*, 400–405, e1–e3.

26. Widmer, C., Lippert, C., Weissbrod, O., Fusi, N., Kadie, C., Davidson, R., Listgarten, J., and Heckerman, D. (2014). Further improvements to linear mixed models for genome-wide association studies. Sci. Rep. *4*, 6874.

27. Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics *178*, 1709–1723.

28. Jakobsdottir, J., and McPeek, M.S. (2013). MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals. Am. J. Hum. Genet. *92*, 652–666.

29. Patterson, H.D., and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. Biometrika *58*, 545–554.

30. Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat. Methods *8*, 833–835.

31. Listgarten, J., Lippert, C., Kadie, C.M., Davidson, R.I., Eskin, E., and Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. Nat. Methods *9*, 525–526.

32. Legarra, A., Aguilar, I., and Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. J. Dairy Sci. *92*, 4656–4663.

33. Misztal, I., Legarra, A., and Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. *92*, 4648–4655.

34. Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., and Lawlor, T.J. (2010). Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. *93*, 743–752.

35. Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G.K., Tandon, A., Kao, W.H., Ruczinski, I., Fornage, M., Siscovick, D.S., Zhu, X., et al. (2011). Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. PLoS Genet. *7*, e1001371.

36. Chen, C.-Y., Pollack, S., Hunter, D.J., Hirschhorn, J.N., Kraft, P., and Price, A.L. (2013). Improved ancestry inference using weights from external reference panels. Bioinformatics *29*, 1399–1406.

37. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

38. Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. *14*, 507–515.

39. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.

40. Bengio, Y., and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. J. Mach. Learn. Res. *5*, 1089–1105.

41. Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. *91*, 1011–1021.

42. Yang, J., Weedon, M.N., Purcell, S., Lettre, G., Estrada, K., Willer, C.J., Smith, A.V., Ingelsson, E., O'Connell, J.R., Mangino, M., et al.; GIANT Consortium (2011). Genomic inflation factors under polygenic inheritance. Eur. J. Hum. Genet. *19*, 807–812.

43. Chen, C.-Y., Han, J., Hunter, D.J., Kraft, P., and Price, A.L. (2015). Explicit modeling of ancestry improves polygenic risk scores and BLUP prediction. Genet. Epidemiol. *39*, 427–438.

44. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al.; Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.

45. Svishcheva, G.R., Axenovich, T.I., Belonogova, N.M., van Duijn, C.M., and Aulchenko, Y.S. (2012). Rapid variance components-based method for whole-genome association analysis. Nat. Genet. *44*, 1166–1170.

46. Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics *55*, 997–1004.

47. Crossett, A., Lee, A.B., Klei, L., Devlin, B., and Roeder, K. (2013). Refining genetically inferred relationships using treelet covariance smoothing. Ann. Appl. Stat. *7*, 669–690.

48. Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genet. *9*, e1003348.

49. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

50. Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H., Bulik-Sullivan, B., Pollack, S., P.-S. W., de Candia, T., Lee, S., Wray, N., et al.; Schizophrenia Working Group Psychiatric Genomics Consortium. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance components analysis. Nat. Genet. Published online November 2, 2015. http://dx.doi.org/10.1038/ng.3431.

51. Legarra, A., and Misztal, I. (2008). Technical note: Computing strategies in genome-wide selection. J. Dairy Sci. *91*, 360–366.

52. VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci. *91*, 4414–4423.

53. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

54. Styrkarsdottir, U., Thorleifsson, G., Sulem, P., Gudbjartsson, D.F., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Oddsson, A., Helgason, A., Magnusson, O.T., et al. (2013). Nonsense

mutation in the LGR4 gene is associated with several human diseases and other traits. Nature *497*, 517–520.

55. Do, C.B., Tung, J.Y., Dorfman, E., Kiefer, A.K., Drabant, E.M., Francke, U., Mountain, J.L., Goldman, S.M., Tanner, C.M., Langston, J.W., et al. (2011). Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genet. *7*, e1002141.

56. Haile-Mariam, M., Nieuwhof, G.J., Beard, K.T., Konstatinov, K.V., and Hayes, B.J. (2013). Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. J. Anim. Breed. Genet. *130*, 20–31.

57. Khansefid, M., Pryce, J.E., Bolormaa, S., Miller, S.P., Wang, Z., Li, C., and Goddard, M.E. (2014). Estimation of genomic breeding values for residual feed intake in a multibreed cattle population. J. Anim. Sci. *92*, 3270–3283.

58. Kemper, K.E., and Goddard, M.E. (2012). Understanding and predicting complex traits: knowledge from cattle. Hum. Mol. Genet. *21* (R1), R45–R51.

Two-Variance-Component Model

Improves Genetic Prediction in Family Datasets

George Tucker, Po-Ru Loh, Iona M. MacLeod, Ben J. Hayes, Michael E. Goddard, Bonnie Berger, and Alkes L. Price
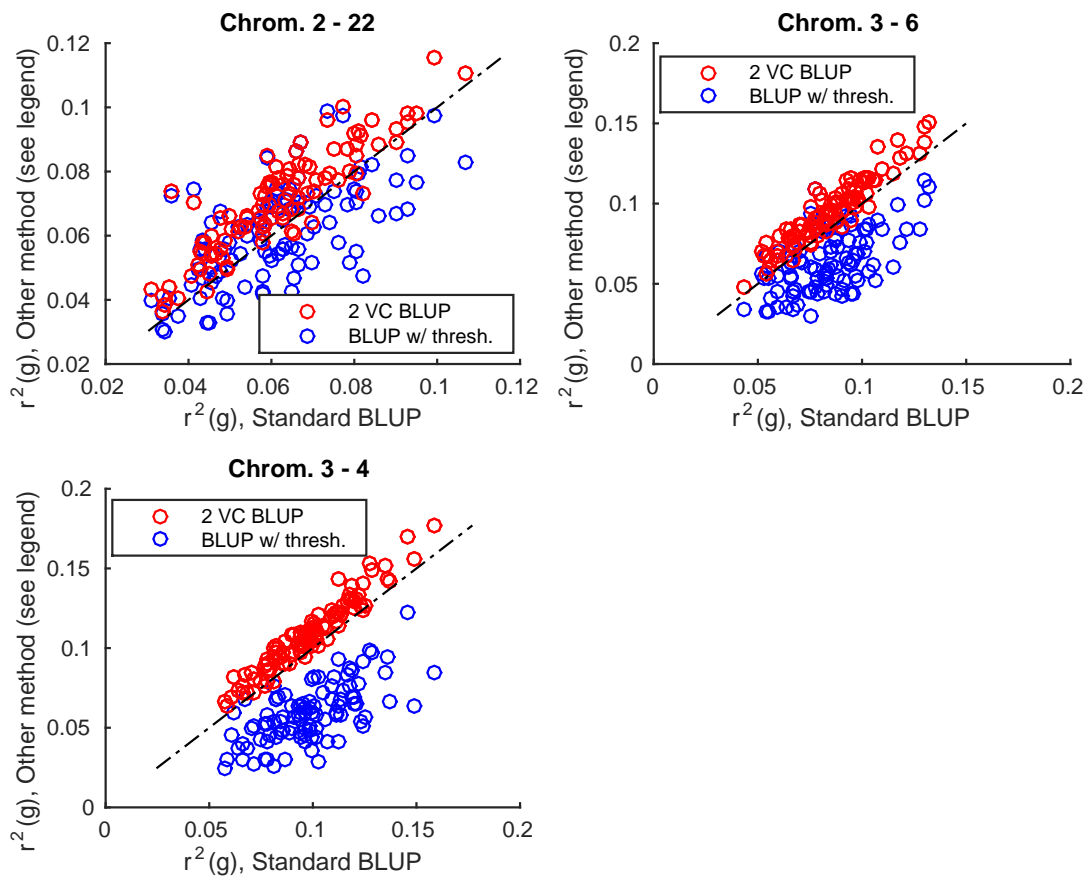
Figure S1. Comparison of prediction $r^2(g)$ **estimates for simulated phenotypes with CARe genotypes.** We compare two variance component BLUP (2 VC BLUP) and BLUP using the thresholded matrix (BLUP w/ thresh) vs. standard BLUP. Plotted values correspond to the 100 random $90/10$ train/test data splits summarized in Table 1a.
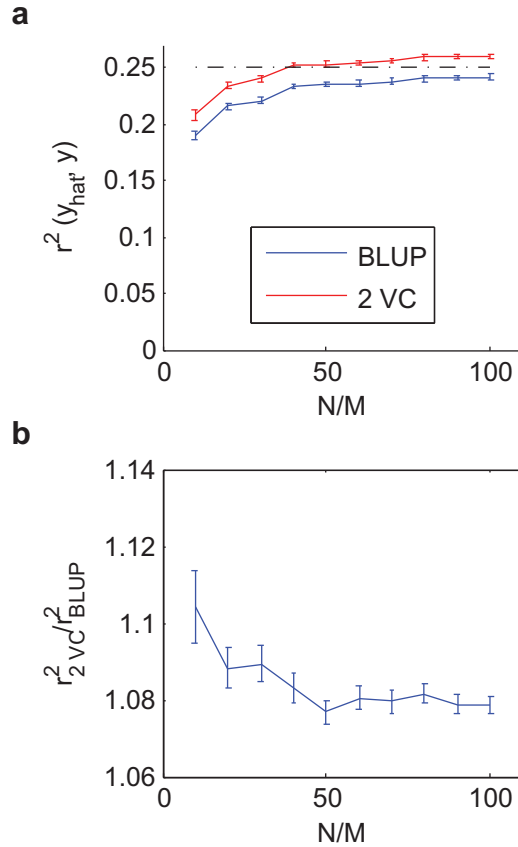
Figure S2. Performance of two variance component prediction vs. standard BLUP on simulated genotypes and phenotypes. We simulated genotypes for sets of sib-pairs (relatedness = 0.5) using the simulation procedure described in Material and Methods with M=100 SNPs and N/M=10,20,...,100. We simulated phenotypes with h_g^2=0.25 and h^2=0.5. We computed predictions using both standard BLUP and two variance component prediction for 10% of the data, using the remaining 90% for training. We ran standard BLUP using the genetic relationship relationship; for the two variance component approach, we included the true pedigree as a second variance component and also assumed the ratio of variance parameters was known to be equal to $(h^2 - h_g^2)/h_g^2 = 1$. These results are therefore an upper bound for the performance of the two variance components approach we have described in this manuscript (which approximates the true pedigree with a thresholded GRM and estimates the ratio of variance parameters). Plotted curves are means over 100 simulation replicates; error bars, s.e.m.
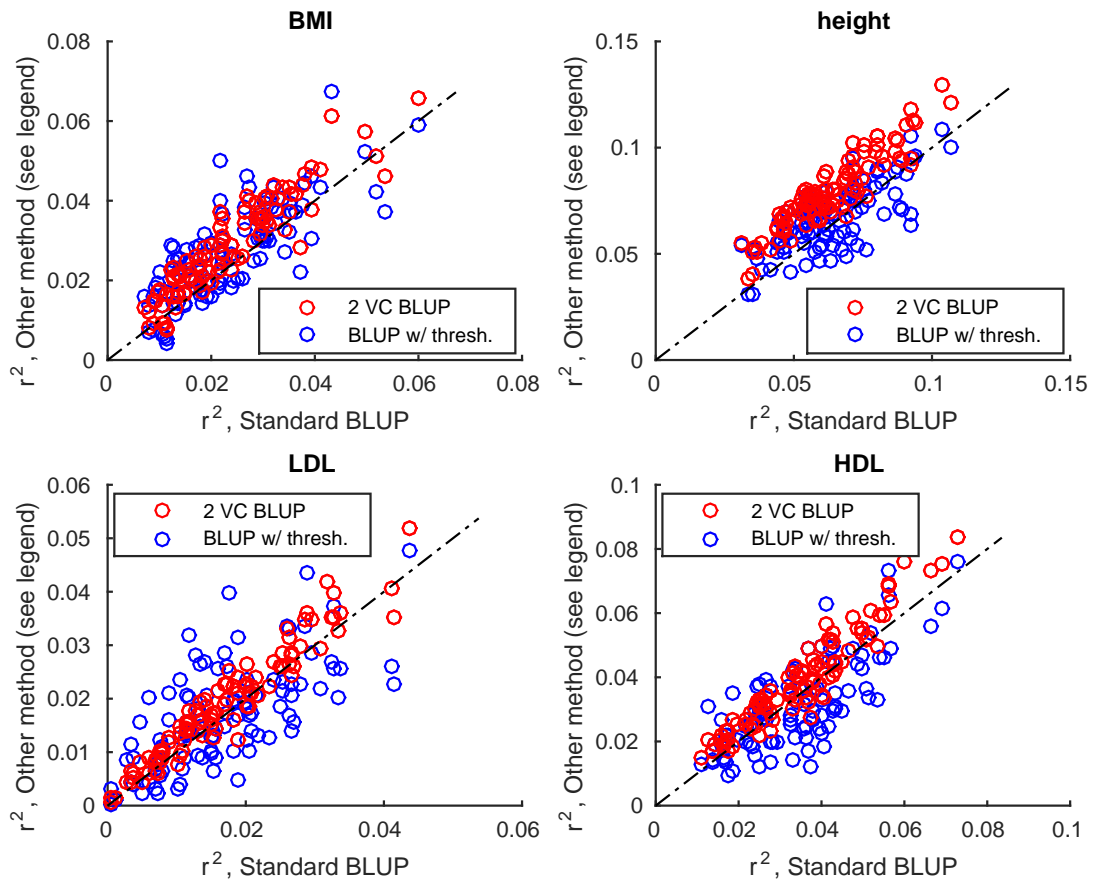
Figure S3. Comparison of prediction r^2 estimates for CARe phenotypes. We compare two variance component BLUP (2 VC BLUP) and BLUP using the thresholded matrix (BLUP w/ thresh) vs. standard BLUP. Plotted values correspond to the 100 random $90/10$ train/test data splits summarized in Table 2a.
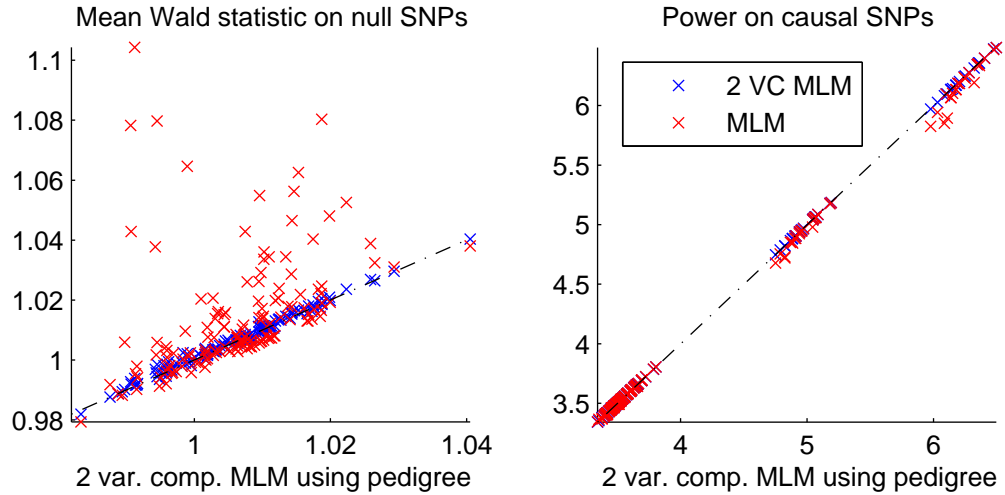
Figure S4. Inflation and power for mixed models on simulated genotypes and phenotypes.
Over a range of simulation parameters, we plot our metrics for inflation (left) and power (right) of association testing using the standard mixed model (MLM) and the two variance component model (2 var. comp. MLM) against a two variance component model that replaces the thresholded GRM with the true pedigree matrix. Each plotted point corresponds to a simulation parameter setting (i.e., choice of N/M, h^2, h_g^2, and NS) plotted in Figure 1. Plotted values are means over 50 simulations. The two variance component method produces near-identical results whether using the thresholded GRM or the true relatedness matrix, whereas standard MLM association produces inflated statistics in many cases and sometimes suffers decreased power.
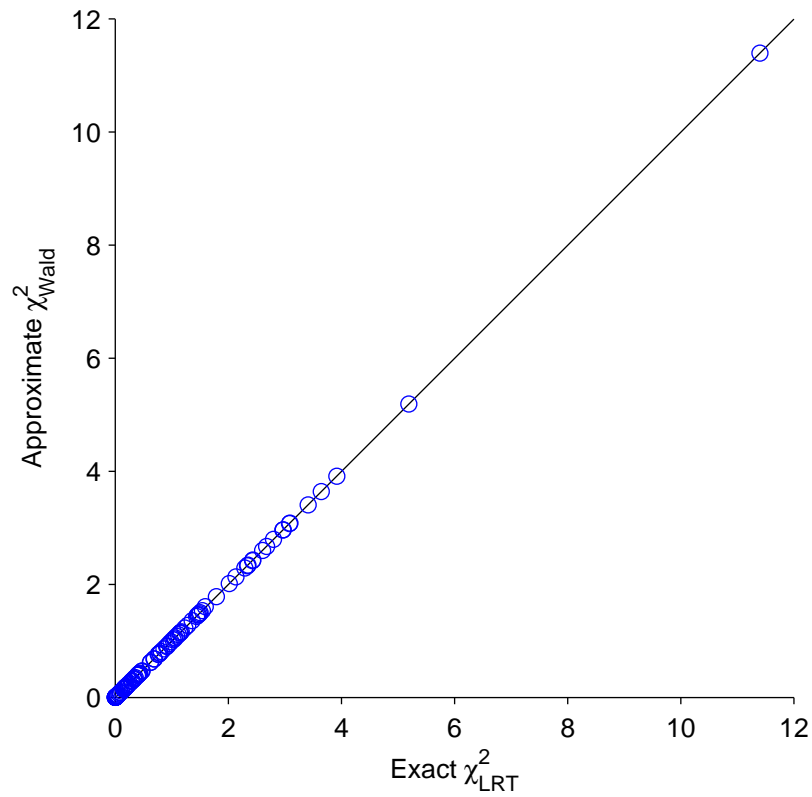
Figure S5. Comparison of approximate vs. exact two variance component association test statistics for CARe height phenotype. We computed exact likelihood ratio test statistics at 110 random SNPs (5 per chromosome) under the two variance component model and compared them to the approximate statistics we computed genome-wide. (The exact likelihood ratio test is computationally expensive, as it requires performing full maximum likelihood computations independently at each SNP.) We observed that our approximate method produced near-identical results ($r^2 = 0.999997$).

Table S1. Heritability parameters for simulations using CARe and FHS genotypes

(a) CARe genotypes

Observed SNPs	BLUP $\hat{h}_g^2$	BLUP w/ thresh. $\hat{h}_{>0.05}^2$	2 VC BLUP $\hat{h}_g^2$	$\hat{h}_{>0.05}^2$
Chrom. 2 - 22	0.380 (0.004)	0.507 (0.004)	0.225 (0.006)	0.285 (0.007)
Chrom. 3 - 6	0.323 (0.003)	0.491 (0.004)	0.238 (0.004)	0.256 (0.006)
Chrom. 3 - 4	0.303 (0.003)	0.482 (0.004)	0.244 (0.003)	0.242 (0.005)

(b) FHS genotypes

Observed SNPs	BLUP $\hat{h}_g^2$	BLUP w/ thresh. $\hat{h}_{>0.05}^2$	2 VC BLUP $\hat{h}_g^2$	$\hat{h}_{>0.05}^2$
Chrom. 2 - 22	0.440 (0.002)	0.495 (0.002)	0.250 (0.006)	0.247 (0.005)
Chrom. 3 - 6	0.393 (0.002)	0.489 (0.002)	0.243 (0.003)	0.248 (0.003)
Chrom. 3 - 4	0.364 (0.002)	0.475 (0.002)	0.231 (0.003)	0.248 (0.002)

Phenotypes were simulated to have $h^2 = 0.5$, $h_g^2 = 0.25$, and heritability parameters were estimated using a random 90% of samples as training data. Reported values are mean prediction r^2 and s.e.m. over 100 independent simulations (in which phenotypes were re-simulated and train/test splits resampled). BLUP w/ thresh. denotes BLUP prediction using the thresholded relationship matrix instead of the standard approach of using the GRM (denoted simply "BLUP").

Table S2. Prediction accuracy for simulations using CARe genotypes with no untyped causal SNPs

Observed SNPs	Prediction $r^2(g)$		
	BLUP	BLUP w/ thresh.	2 VC BLUP
Chrom. 2 - 22	0.097 (0.002)	0.053 (0.002)	0.097 (0.002)
Chrom. 3 - 6	0.172 (0.003)	0.055 (0.002)	0.172 (0.003)
Chrom. 3 - 4	0.244 (0.003)	0.058 (0.002)	0.243 (0.003)

Phenotypes were simulated to have $h^2 = 0.5$, $h_g^2 = 0.5$ (i.e., no untyped causal SNPs, so the two variance component model is expected to achieve no gain). Prediction $r^2(g)$ was measured using a random 90% of samples as training data and the remaining 10% as test data. Reported values are mean prediction $r^2(g)$ and s.e.m. over 100 independent simulations (in which phenotypes were re-simulated and train/test splits resampled). BLUP w/ thresh. denotes BLUP prediction using the thresholded relationship matrix instead of the standard approach of using the GRM (denoted simply "BLUP"). Prediction $r^2(g)$ denotes r^2 between predicted phenotypes and true genetic components of the simulated phenotypes.

Table S3. Prediction accuracy and heritability parameters for CARe simulations with LD between typed and untyped SNPs

h^2_{typed}	BLUP $\hat{h}^2_g$	2 VC BLUP $\hat{h}^2_g$	$\hat{h}^2_{>0.05}$	BLUP $r^2(g)$	2VC BLUP $r^2(g)$
0	0.348 (0.004)	0.188 (0.005)	0.321 (0.006)	0.064 (0.002)	0.077 (0.002)
0.05	0.362 (0.003)	0.213 (0.005)	0.291 (0.007)	0.065 (0.002)	0.077 (0.002)
0.1	0.378 (0.003)	0.244 (0.005)	0.263 (0.007)	0.068 (0.002)	0.078 (0.002)
0.15	0.392 (0.004)	0.271 (0.006)	0.232 (0.007)	0.073 (0.002)	0.080 (0.002)
0.2	0.409 (0.004)	0.296 (0.006)	0.213 (0.006)	0.073 (0.002)	0.079 (0.002)
0.25	0.415 (0.003)	0.310 (0.006)	0.193 (0.007)	0.082 (0.002)	0.086 (0.002)

We modified our simulations based on real CARe genotypes (Table 1a) to include LD between typed and untyped SNPs by setting typed SNPs to be the 90% of CARe SNPs with highest MAF and untyped SNPs to be the 10% of CARe SNPs with lowest MAF. (The MAF cutoff corresponding to this split was 5.4%.) As in our simulations without LD between typed and untyped SNPs, we simulated phenotypes with a total heritability of $h^2 = 0.5$; in these new simulations, we varied the fraction of variance directly explained by typed SNPs ($h^2_{typed} = 0, 0.05,$ 0.1, 0.15, 0.2, or 0.25), with the rest of the heritable variance ($h^2 - h^2_{typed}$) explained by untyped SNPs. We note that the fraction of variance attributed to typed SNPs (by both standard BLUP and 2VC BLUP) exceeds h^2_{typed} in these simulations because of the LD between typed and untyped SNPs; we varied h^2_{typed} from 0 to 0.25 for this reason. For the setting of h^2_{typed} in which 2VC BLUP partitions variance roughly equally between the GRM and thresholded GRM components ($h^2_{typed} = 0.1$)—matching the partitioning in our simulations without LD (Table S1a)—we observed that 2VC BLUP achieved an increase in prediction accuracy ($0.068 \rightarrow 0.078$) similar to our simulations without LD between typed and untyped SNPs ($0.062 \rightarrow 0.071$, Table 1a).

Table S4. Heritability parameters for CARe and FHS phenotypes

(a) CARe heritability parameters

Observed SNPs	BLUP $\hat{h}_g^2$	BLUP w/ thresh. $\hat{h}_{>0.05}^2$	2 VC BLUP $\hat{h}_g^2$	$\hat{h}_{>0.05}^2$
BMI	0.336 (0.002)	0.468 (0.002)	0.148 (0.002)	0.321 (0.004)
height	0.673 (0.002)	0.953 (0.002)	0.364 (0.002)	0.591 (0.003)
LDL	0.339 (0.002)	0.432 (0.003)	0.219 (0.003)	0.216 (0.004)
HDL	0.512 (0.002)	0.666 (0.003)	0.299 (0.003)	0.366 (0.004)

(b) FHS heritability parameters

Observed SNPs	BLUP $\hat{h}_g^2$	BLUP w/ thresh. $\hat{h}_{>0.05}^2$	2 VC BLUP $\hat{h}_g^2$	$\hat{h}_{>0.05}^2$
BMI	0.435 (0.001)	0.474 (0.001)	0.217 (0.002)	0.256 (0.002)
height	0.823 (0.001)	0.878 (0.001)	0.436 (0.002)	0.441 (0.002)

(c) CARe heritability parameters using genome-wide significant SNPs as fixed effect covariates

Observed SNPs	BLUP $\hat{h}_g^2$	BLUP w/ thresh. $\hat{h}_{>0.05}^2$	2 VC BLUP $\hat{h}_g^2$	$\hat{h}_{>0.05}^2$
BMI	0.336 (0.002)	0.470 (0.002)	0.143 (0.002)	0.326 (0.004)
height	0.672 (0.002)	0.953 (0.002)	0.363 (0.002)	0.592 (0.003)
LDL	0.339 (0.003)	0.452 (0.003)	0.195 (0.004)	0.258 (0.005)
HDL	0.503 (0.002)	0.662 (0.003)	0.291 (0.003)	0.370 (0.004)

(d) FHS heritability parameters using genome-wide significant SNPs as fixed effect covariates

Observed SNPs	BLUP $\hat{h}_g^2$	BLUP w/ thresh. $\hat{h}_{>0.05}^2$	2 VC BLUP $\hat{h}_g^2$	$\hat{h}_{>0.05}^2$
BMI	0.432 (0.001)	0.472 (0.001)	0.210 (0.002)	0.261 (0.002)
height	0.822 (0.001)	0.877 (0.001)	0.436 (0.002)	0.440 (0.002)

Heritability parameters are means over 100 random 90%-subsamples corresponding to the train/test splits used to estimate prediction r^2.

Table S5. Prediction accuracy for CARe and FHS phenotypes (1 – MSE)

(a) CARe prediction

| Phenotype | BLUP | 1 – MSE | | 1 – MSE relative to BLUP (s.e.) | |
		BLUP w/ thresh.	2 VC BLUP	BLUP w/ thresh.	2 VC BLUP
BMI	0.022	0.025	0.027	+14% (9%)	+19% (5%)
height	0.061	0.065	0.078	+7% (6%)	+22% (3%)
LDL	0.014	0.015	0.017	+1% (16%)	+13% (5%)
HDL	0.032	0.030	0.036	-8% (11%)	+12% (4%)

(b) CARe prediction using genome-wide significant SNPs as fixed effect covariates

| Phenotype | BLUP | 1 – MSE | | 1 – MSE relative to BLUP (s.e.) | |
		BLUP w/ thresh.	2 VC BLUP	BLUP w/ thresh.	2 VC BLUP
BMI	0.021	0.024	0.026	+15% (9%)	+20% (6%)
height	0.060	0.064	0.076	+6% (6%)	+22% (3%)
LDL	0.035	0.037	0.038	+4% (7%)	+7% (3%)
HDL	0.048	0.046	0.052	-6% (7%)	+8% (3%)

We normalized mean square errors by dividing by phenotypic variance and computed the mean of 1 – MSE over 100 random $90/10$ train/test data splits. Relative performance values reported are ratios of means minus 1; standard errors are estimated as standard deviations of per-split differences in 1 – MSE (over the random 10% test sets) divided by $\sqrt{10}$ (to account for the 10x larger sample size of the full data set; see Material and Methods). BLUP w/ thresh. denotes BLUP prediction using the thresholded relationship matrix instead of the standard approach of using the GRM (denoted simply "BLUP").

Table S6. Prediction accuracy for simulations with 25% untyped individuals using CARe and FHS genotypes

(a) CARe genotypes

Observed SNPs	Prediction $r^2(g)$		
	BLUP	BLUP w/ thresh.	2 VC BLUP
Chrom. 2 - 22	0.061 (0.002)	0.060 (0.002)	0.068 (0.002)
Chrom. 3 - 6	0.080 (0.002)	0.064 (0.002)	0.089 (0.002)
Chrom. 3 - 4	0.089 (0.002)	0.058 (0.002)	0.098 (0.002)

(b) FHS genotypes

Observed SNPs	Prediction $r^2(g)$		
	BLUP	BLUP w/ thresh.	2 VC BLUP
Chrom. 2 - 22	0.226 (0.003)	0.225 (0.003)	0.235 (0.003)
Chrom. 3 - 6	0.240 (0.003)	0.228 (0.003)	0.260 (0.003)
Chrom. 3 - 4	0.257 (0.003)	0.232 (0.003)	0.282 (0.003)

Phenotypes were simulated to have $h^2 = 0.5$, $h_g^2 = 0.25$, and prediction $r^2(g)$ was measured using a random 90% of samples as training data and the remaining 10% as test data. Reported values are mean prediction $r^2(g)$ and s.e.m. over 100 independent simulations (in which phenotypes were re-simulated and train/test splits resampled). BLUP w/ thresh. denotes BLUP prediction using the thresholded relationship matrix instead of the standard approach of using the GRM (denoted simply "BLUP"). Prediction $r^2(g)$ denotes r^2 between predicted phenotypes and true genetic components of the simulated phenotypes.

Table S7. Prediction accuracy for CARe and FHS phenotypes with 25% untyped individuals

(a) CARe prediction

Phenotype		Prediction r^2		Prediction r^2 relative to BLUP (s.e.)	
	BLUP	BLUP w/ thresh.	2 VC BLUP	BLUP w/ thresh.	2 VC BLUP
BMI	0.024	0.027	0.028	+12% (8%)	+15% (5%)
height	0.064	0.067	0.076	+4% (5%)	+16% (3%)
LDL	0.017	0.017	0.019	-1% (13%)	+7% (5%)
HDL	0.035	0.032	0.038	-9% (9%)	+8% (4%)

(b) FHS prediction

Phenotype		Prediction r^2		Prediction r^2 relative to BLUP (s.e.)	
	BLUP	BLUP w/ thresh.	2 VC BLUP	BLUP w/ thresh.	2 VC BLUP
BMI	0.103	0.104	0.106	+0.8% (2.2%)	+3.0% (1.1%)
height	0.344	0.342	0.352	-0.5% (1.1%)	+2.4% (0.5%)

(c) CARe prediction using genome-wide significant SNPs as fixed effect covariates

Phenotype		Prediction r^2		Prediction r^2 relative to BLUP (s.e.)	
	BLUP	BLUP w/ thresh.	2 VC BLUP	BLUP w/ thresh.	2 VC BLUP
BMI	0.023	0.027	0.028	+12% (8%)	+16% (5%)
height	0.064	0.066	0.076	+4% (5%)	+16% (3%)
LDL	0.036	0.036	0.037	+0% (6%)	+4% (3%)
HDL	0.052	0.049	0.055	-6% (6%)	+5% (3%)

(d) FHS prediction using genome-wide significant SNPs as fixed effect covariates

Phenotype		Prediction r^2		Prediction r^2 relative to BLUP (s.e.)	
	BLUP	BLUP w/ thresh.	2 VC BLUP	BLUP w/ thresh.	2 VC BLUP
BMI	0.104	0.105	0.107	+0.9% (2.2%)	+3.0% (1.1%)
height	0.344	0.342	0.352	-0.6% (1.1%)	+2.3% (0.5%)

Prediction r^2 values are means over 100 random $90/10$ train/test data splits. Relative performance values reported are ratios of means minus 1; standard errors are estimated as standard deviations of per-split differences in r^2 (over the random 10% test sets) divided by $\sqrt{10}$ (to account for the 10x larger sample size of the full data set; see Material and Methods). BLUP w/ thresh. denotes BLUP prediction using the thresholded relationship matrix instead of the standard approach of using the GRM (denoted simply "BLUP").

Table S8. Heritability parameters for simulations with 25% untyped individuals using CARe and FHS genotypes

(a) CARe genotypes

Observed SNPs	BLUP $\hat{h}_g^2$	BLUP w/ thresh. $\hat{h}_{>0.05}^2$	2 VC BLUP $\hat{h}_g^2$	$\hat{h}_{>0.05}^2$
Chrom. 2 - 22	0.421 (0.004)	0.507 (0.004)	0.228 (0.007)	0.280 (0.008)
Chrom. 3 - 6	0.355 (0.004)	0.489 (0.004)	0.239 (0.004)	0.251 (0.006)
Chrom. 3 - 4	0.328 (0.003)	0.482 (0.004)	0.244 (0.004)	0.240 (0.005)

(b) FHS genotypes

Observed SNPs	BLUP $\hat{h}_g^2$	BLUP w/ thresh. $\hat{h}_{>0.05}^2$	2 VC BLUP $\hat{h}_g^2$	$\hat{h}_{>0.05}^2$
Chrom. 2 - 22	0.457 (0.002)	0.495 (0.002)	0.253 (0.006)	0.242 (0.006)
Chrom. 3 - 6	0.417 (0.002)	0.489 (0.002)	0.240 (0.004)	0.250 (0.003)
Chrom. 3 - 4	0.386 (0.002)	0.475 (0.002)	0.229 (0.003)	0.249 (0.003)

Phenotypes were simulated to have $h^2 = 0.5$, $h_g^2 = 0.25$, and heritability parameters were estimated using a random 90% of samples as training data. Reported values are mean prediction r^2 and s.e.m. over 100 independent simulations (in which phenotypes were re-simulated and train/test splits resampled). BLUP w/ thresh. denotes BLUP prediction using the thresholded relationship matrix instead of the standard approach of using the GRM (denoted simply "BLUP").

Table S9. Heritability parameters for CARe and FHS phenotypes with 25% untyped individuals

(a) CARe heritability parameters

Observed SNPs	BLUP $\hat{h}_g^2$	BLUP w/ thresh. $\hat{h}_{>0.05}^2$	2 VC BLUP $\hat{h}_g^2$	2 VC BLUP $\hat{h}_{>0.05}^2$
BMI	0.372 (0.002)	0.468 (0.002)	0.142 (0.005)	0.326 (0.006)
height	0.781 (0.003)	0.953 (0.002)	0.374 (0.004)	0.580 (0.004)
LDL	0.371 (0.003)	0.432 (0.003)	0.234 (0.006)	0.201 (0.007)
HDL	0.563 (0.003)	0.666 (0.003)	0.311 (0.006)	0.354 (0.007)

(b) FHS heritability parameters

Observed SNPs	BLUP $\hat{h}_g^2$	BLUP w/ thresh. $\hat{h}_{>0.05}^2$	2 VC BLUP $\hat{h}_g^2$	2 VC BLUP $\hat{h}_{>0.05}^2$
BMI	0.448 (0.001)	0.474 (0.001)	0.221 (0.004)	0.252 (0.004)
height	0.843 (0.001)	0.878 (0.001)	0.448 (0.005)	0.427 (0.005)

(c) CARe heritability parameters using genome-wide significant SNPs as fixed effect covariates

Observed SNPs	BLUP $\hat{h}_g^2$	BLUP w/ thresh. $\hat{h}_{>0.05}^2$	2 VC BLUP $\hat{h}_g^2$	2 VC BLUP $\hat{h}_{>0.05}^2$
BMI	0.372 (0.002)	0.468 (0.002)	0.140 (0.005)	0.328 (0.006)
height	0.780 (0.003)	0.953 (0.002)	0.373 (0.004)	0.582 (0.004)
LDL	0.376 (0.003)	0.446 (0.003)	0.213 (0.006)	0.234 (0.007)
HDL	0.559 (0.003)	0.662 (0.003)	0.314 (0.006)	0.347 (0.007)

(d) FHS heritability parameters using genome-wide significant SNPs as fixed effect covariates

Observed SNPs	BLUP $\hat{h}_g^2$	BLUP w/ thresh. $\hat{h}_{>0.05}^2$	2 VC BLUP $\hat{h}_g^2$	2 VC BLUP $\hat{h}_{>0.05}^2$
BMI	0.448 (0.001)	0.474 (0.001)	0.219 (0.004)	0.254 (0.004)
height	0.843 (0.001)	0.878 (0.001)	0.448 (0.005)	0.427 (0.005)

Heritability parameters are means over 100 random 90%-subsamples corresponding to the train/test splits used to estimate prediction r^2.

Table S10. Type I error of standard and two-variance-component mixed model association statistics in CARe and FHS simulations

(a) CARe genotypes

Observed SNPs	Standard mixed model		Two variance components	
	α=0.01	α=0.0001	α=0.01	α=0.0001
Chrom. 3 - 22	1.05e-02	1.13e-04	1.00e-02	1.02e-04
Chrom. 3 - 6	1.10e-02	1.26e-04	1.01e-02	1.07e-04
Chrom. 3 - 4	1.11e-02	1.30e-04	1.01e-02	1.07e-04
Chrom. 22	1.14e-02	1.35e-04	1.05e-02	1.16e-04

(b) FHS genotypes

Observed SNPs	Standard mixed model		Two variance components	
	α=0.01	α=0.0001	α=0.01	α=0.0001
Chrom. 3 - 22	1.12e-02	1.40e-04	1.01e-02	1.09e-04
Chrom. 3 - 6	1.28e-02	1.84e-04	1.04e-02	1.20e-04
Chrom. 3 - 4	1.40e-02	2.11e-04	1.07e-02	1.25e-04
Chrom. 22	1.81e-02	3.57e-04	1.21e-02	1.50e-04

Type I error of Wald statistics on candidate null SNPs for simulations with CARe or FHS genotypes and a trait with $h^2 = 0.5, h_g^2 = 0.25$ (see Table 3 for details). Reported values are aggregated over 100 simulations testing null SNPs on chromosome 2 (63,077 SNPs for CARe, 34,608 SNPs for FHS).

Table S11. Prediction accuracy for a range of dairy cattle traits

Trait	Num. of Records Training (Validation)	BLUP	Prediction r^2 BLUP w/ pedigree	2 VC BLUP
Fat Yield	8820 (1053)	0.360	0.169	0.359
Milk Yield	8820 (1053)	0.490	0.267	0.506
Protein Yield	8820 (1053)	0.442	0.265	0.453
Teat Length	2500 (360)	0.312	0.203	0.315
Temperament	5543 (734)	0.110	0.053	0.109
Fertility	8428 (838)	0.225	0.122	0.225

We analyzed four dairy cattle traits from the Holstein breed using a data set previously described in [59]. We added two further phenotypes which had been recorded for the same animals: temperament score and teat length (both have an influence on the ease of milking). Animals had 632,002 SNP genotypes. We tested three BLUP prediction methods: standard mixed model BLUP, BLUP using the pedigree, and two variance component BLUP using both the GRM and pedigree. (We used the full pedigree relationship matrix in place of the thresholded GRM because extensive pedigree records are typically available in dairy cattle.) We analyzed corrected phenotypes including both progeny tested bulls and cows with repeat records in a weighted analysis (with weights calculated from the effective number of records per animal) as described in [59]. We implemented the analyses using ASReml software [60]. The training/test data split (approximately 90/10) was based on a date of birth cutoff, with the youngest bulls used for the test set. Only bull data was used for the test set because their phenotypes (progeny test with ≥20 daughter records) are considerably more accurate than those of cows.

We did not observe a consistent advantage in prediction accuracy using the two variance component model compared to standard BLUP. Possible reasons for the difference between cattle and human results are:

1. In Holstein dairy cattle, linkage disequilibrium (LD) decays much more slowly with physical distance between variants compared to humans because of their recent sharp decline to a very small effective population size [61].

2. The recent very small effective population size results in fewer rare variants segregating compared to some human populations which have relatively large effective population size and have undergone recent expansion. This means there is likely to be a lower proportion of rare causal variants in cattle compared to human populations.

3. The training population is very closely related to the test population.

These three factors combined suggest that similarly dense SNP genotypes may more accurately track the variance due to causal mutations within a single cattle breed compared to the human data in this study. We might still therefore expect some improvement in accuracy from the 2 VC model in cattle if using a less dense SNP chip (e.g., 50K) or mixed breed analysis because a more significant proportion of causal mutations may not be in high LD with typed SNPs.