Please cite this article in press as: Coram et al., Leveraging Multi-Ethnic Evidence for Risk Assessment of Quantitative Traits in Minority Populations, The American Journal of Human Genetics (2017), http://dx.doi.org/10.1016/j.ajhg.2017.06.015

ARTICLE

Leveraging Multi-Ethnic Evidence for Risk Assessment of Quantitative Traits in Minority Populations

Marc A. Coram,[1,5] Huaying Fang,[2] Sophie I. Candille,[2] Themistocles L. Assimes,[3,4] and Hua Tang[2,*]

An essential component of precision medicine is the ability to predict an individual's risk of disease based on genetic and non-genetic factors. For complex traits and diseases, assessing the risk due to genetic factors is challenging because it requires knowledge of both the identity of variants that influence the trait and their corresponding allelic effects. Although the set of risk variants and their allelic effects may vary between populations, a large proportion of these variants were identified based on studies in populations of European descent. Heterogeneity in genetic architecture underlying complex traits and diseases, while broadly acknowledged, remains poorly characterized. Ignoring such heterogeneity likely reduces predictive accuracy for minority individuals. In this study, we propose an approach, called XP-BLUP, which ameliorates this ethnic disparity by combining trans-ethnic and ethnic-specific information. We build a polygenic model for complex traits that distinguishes candidate trait-relevant variants from the rest of the genome. The set of candidate variants are selected based on studies in any human population, yet the allelic effects are evaluated in a population-specific fashion. Simulation studies and real data analyses demonstrate that XP-BLUP adaptively utilizes trans-ethnic information and can substantially improve predictive accuracy in minority populations. At the same time, our study highlights the importance of the continued expansion of minority cohorts.

Introduction

An important component of precision medicine is to incorporate the genetic variation of an individual in disease risk assessment, as well as in optimizing disease prevention and treatment strategies. In the context of complex traits, such as blood glucose level, lipid concentration, and body mass index, this goal has proven challenging for two reasons: first, these traits are strongly influenced by an array of genetic, environmental, and lifestyle risk factors, most of which are not systematically measured; second, the genetic components of complex traits feature polygenic architecture, meaning hundreds or thousands of genes influence an individual's phenotype, of which our current knowledge—largely derived from regions reaching statistical significance in genome-wide association studies (GWASs)—represents only the tip of the iceberg.[1] At the same time, studies with expanding cohorts continue to uncover variants associated with a variety of traits and diseases, indicating moderate to high heritability.[2–5] Therefore, while genetic variation alone is insufficient for accurate disease or trait prediction, it is reasonable to expect that a summary score of all trait-relevant variants can meaningfully quantify the heritable component that underlies variation in complex traits or disease risks. Furthermore, this genetic score complements conventional non-genetic risk factors, and integration of genetic and non-genetic risk factors may lead to more accurate health assessment. The goal of this paper is to develop an analytical approach to quantifying the genetic risk score that influences the trait; the term "trait prediction" is used as shorthand, but it should be emphasized that the output of this "trait prediction" aims to capture only the genetic risk factors, and is meant to be combined with non-genetic risk factors such as age, gender, or smoking status.

Although a number of methods have been developed for complex trait prediction, most of these methods are designed for and applied to populations of European descent (EUR). These methods fall into two categories. One group of methods builds regularized regression models based on established or suggestive trait loci.[6] In its simplest form, such a method selects independent SNPs reaching a pre-specified p value threshold. At each SNP, one allele is designated as the trait-increasing allele, while the other allele is a trait-decreasing allele. The genetic risk simply tallies the number of trait-increasing alleles across all selected SNPs.[7,8] More sophisticated methods have been developed to weigh SNPs by the corresponding allelic effects, which are estimated from an independent *training* cohort.[9,10] These methods are computationally efficient and perform well in populations where large cohorts are available. However, for minority individuals, who are under-represented in GWASs, these methods are sub-optimal because much less information is available for selecting relevant SNPs and for estimating allelic effects. Consequentially, under this framework, trait prediction may be considerably less precise in minority populations.

[1]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA 94305, USA; [2]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA; [3]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA; [4]Cardiovascular Institute, Stanford University School of Medicine, Stanford, CA 94305, USA
[5]Present address: Google Inc., Mountain View, CA 94043, USA
*Correspondence: huatang@stanford.edu
http://dx.doi.org/10.1016/j.ajhg.2017.06.015.

Please cite this article in press as: Coram et al., Leveraging Multi-Ethnic Evidence for Risk Assessment of Quantitative Traits in Minority Populations, The American Journal of Human Genetics (2017), http://dx.doi.org/10.1016/j.ajhg.2017.06.015

The second group of methods is based on best linear unbiased prediction (BLUP) for predicting random effects in linear mixed effects models (LMMs).[11,12] Originally applied to animal and plant breeding,[13,14] LMMs have become a powerful method to estimate SNP-based heritability using GWAS data.[15,16] By assuming a common distribution that describes the allelic effects at all SNPs and focusing on the aggregated genetic effects rather than individual genetic effects, LMMs and BLUP are particularly attractive for modeling polygenic architecture. On the other hand, to capture minute effects spread over a majority of the genome, this assumed common distribution—usually a Gaussian distribution—fails to capture major trait loci with large effects efficiently. To cope with this problem, a number of methods have been proposed that use multiple random effect terms with distinct variances to represent classes of SNPs with varying allelic effects.[17–19] External biological annotation offers a natural grouping of SNPs into classes that correspond to the multiple random effects terms; Speed and Balding[17] developed an algorithm that groups genomic regions based on their observed regional heritability. In parallel, Bayesian hierarchical models have been developed, which consider the underlying allelic effects as drawn from a mixture of Gaussian distributions, thus providing a probabilistic framework that allocates a sparse fraction of the SNPs to have moderate to large effects.[18–20] Although these methods differ in how the SNP classes are defined, simulation and real data analyses results consistently suggest that very large GWAS sample sizes are needed to achieve good predictive performance. Under favorable simulation settings, a sample size of ~100,000 is recommended.[16] While such a sample size is becoming feasible for some common traits in European populations, it remains infeasible for minority populations.

In this paper, we present a trans-ethnic framework for assessing genetic risk for minority individuals. Previously we have shown that complex traits genetic architecture overlaps substantially between ethnicities and that harnessing trans-ethnic information leads to substantial improvement in our ability to discover trait loci that are relevant in minority populations.[21,22] Therefore, we reason that trait prediction in minority populations may also benefit from integrating much larger GWAS data generated from populations of European descent. Our proposed approach, termed cross-population BLUP (XP-BLUP), is a multiple-component LMM model designed specifically to address the need for efficient prediction in minority populations. The SNPs are placed into classes defined using GWAS evidence from any ethnicity, while the variances of the random-effects and BLUP are computed using population-specific data. We use simulation to illustrate the usefulness of trans-ethnic information, as well as the importance of minority-specific GWAS data. Application to real data of lipid concentration indicates that the genetic risk factor formulated by XP-BLUP on a moderate sized African American (AA) cohort achieves similar predictive accuracy as commonly measured non-genetic risk factors, such as age and BMI.

Material and Methods

LMM and BLUP for Complex Traits

Our proposed trans-ethnic prediction approach is based on LMM and BLUP; the statistical theory and computation are well established. To facilitate the comparison and to set up notations, we briefly summarize the standard LMM and BLUP for complex traits. Details about the computation and the underlying model assumptions have been described in the context of SNP-based heritability estimation.[15,17,23]

Let $\boldsymbol{X}$ be the N × M genotype matrix of N individuals and M markers, and $\boldsymbol{Y}$ be the vector of length N, representing the quantitative phenotype centered at 0. Following convention, let $\boldsymbol{Z}$ be the centered and scaled genotype matrix: if $X_{jm} \in (0, 1, 2)$ denotes genotypes of individual j at SNP m, then $Z_{jm} = (X_{jm} - 2f_m)/\sqrt{2f_m(1-f_m)}$ where f_m is the allele frequency.[15] $\boldsymbol{Y}$ is modeled as the sum of the genetic ($\boldsymbol{g}$) and non-genetic ($\boldsymbol{e}$) contributions: $\boldsymbol{Y} = \boldsymbol{g} + \boldsymbol{e}$. (In what follows, all boldfaced symbols represent vectors or matrices.) For cohorts in which individuals are not close relatives, it is convenient to assume that the non-genetic contribution e_j is drawn independently from a Gaussian distribution, $N(0, \sigma_e^2)$. Under a polygenic architecture, the additive genetic component can be represented as a linear combination of the standardized genotypes ($\boldsymbol{Z}$) weighted by the corresponding allelic effects, $\boldsymbol{u}$: $\boldsymbol{g} = \boldsymbol{Z}\boldsymbol{u}$. In its simplest form, we consider the allelic effects, u_m, as independently sampled from a Gaussian distribution with mean 0 and variance σ^2/M, $N(0, \sigma^2/M)$. It can be shown that $\boldsymbol{g} \sim N(0, \boldsymbol{A}\sigma^2)$, where $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{Z}'/M$ is a genetic relationship matrix (GRM). Covariates, such as age and gender, can be included in the model as fixed effects terms. We will refer to this basic model as the single-component model.

Building upon the single-component model, a number of studies have aimed to use multiple random effects terms to represent groups of SNPs with different allelic effects.[18,20,24] Let $\boldsymbol{Z}_k$ denote the sub-matrix (columns of matrix $\boldsymbol{Z}$) that corresponds to the SNPs in group k. The general form of a K-component LMM can be written as:

$$\boldsymbol{Y} = \sum_{k=1}^{K} \boldsymbol{g}_k + \boldsymbol{e} \qquad \text{(Equation 1)}$$

where $\boldsymbol{g}_k = \boldsymbol{Z}_k\boldsymbol{u}_k$ represents the genetic contribution of group k. This model allows SNPs in group k to have allelic effects drawn from a group-specific Gaussian distribution, $u \sim N(0, \sigma_k^2/|C_k|)$, where $|C_k|$ is the number of markers in group k. Analogous to the single-component model, $\boldsymbol{g}_{\boldsymbol{k}} \sim N(0, \boldsymbol{A}_{\boldsymbol{k}}\sigma_k^2)$, where $\boldsymbol{A}_{\boldsymbol{k}} = \boldsymbol{Z}_k\boldsymbol{Z}_{\boldsymbol{k}}'/|C_k|$ can be thought of as the GRM based on SNPs in group k. It should be noted that the SNP groups do not need to be disjoint; a SNP belonging to both groups k_1 and k_2 has an effect size drawn from $N(0, \sigma_{k_1}^2/|C_{k_1}| + \sigma_{k_2}^2/|C_{k_2}|)$.

The variance parameters $\sigma_1^2, \ldots, \sigma_K^2$ and σ_e^2 are estimated by restricted maximum likelihood (REML).[25,26] Given that the variance parameters have been estimated, the BLUP of allelic effects for SNPs in class k is:

$$\widehat{\boldsymbol{u}}_k = \widehat{\sigma}_k^2\boldsymbol{A}_k\boldsymbol{Z}_{\boldsymbol{k}}'\boldsymbol{V}^{-1}\boldsymbol{Y}$$

where $\boldsymbol{V} = \sum_k \widehat{\sigma}_k^2\boldsymbol{A}_k + \widehat{\sigma}_e^2\boldsymbol{I}_N$, and $\boldsymbol{I}_N$ is the identity matrix of dimension N.

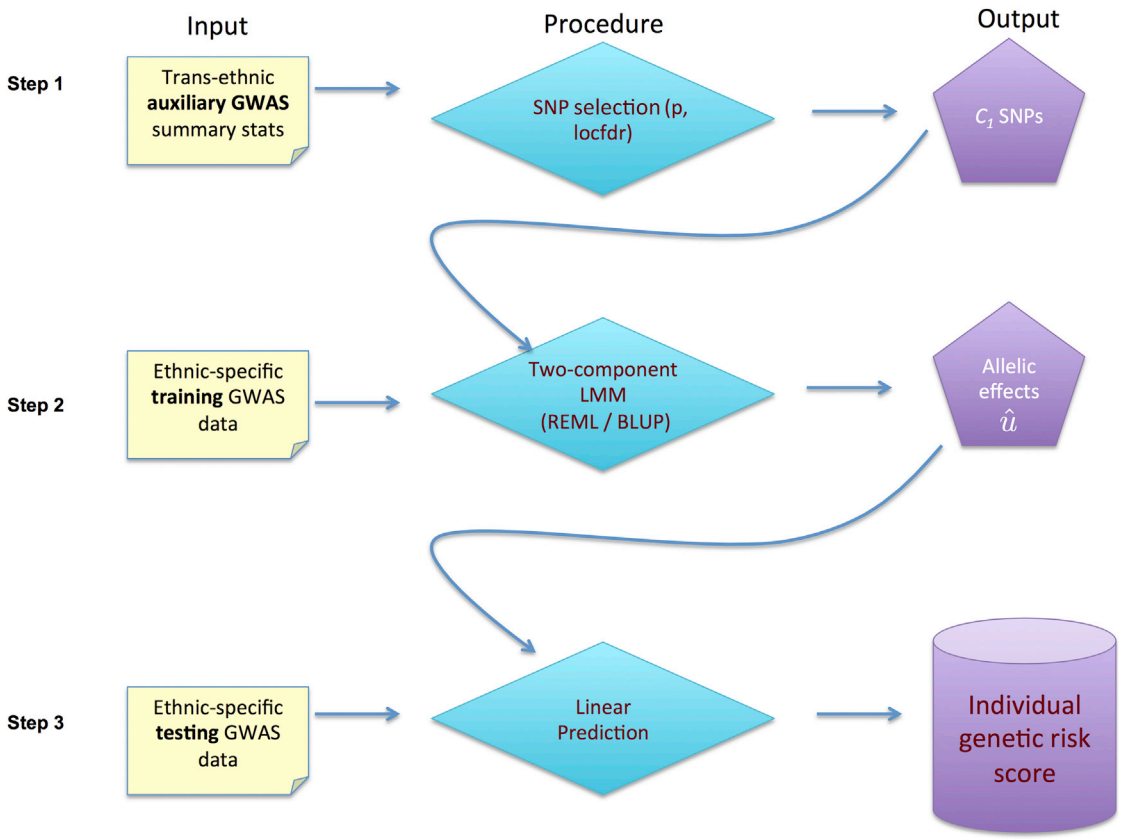

Figure 1. Overview of the XP-BLUP Procedure
Step 1: SNPs are sorted into two classes based on summary statistics in a trans-ethnic auxiliary GWAS. Step 2: A two-component linear mixed-effects model (LMM) and ethnic-specific training GWAS data are used to compute the allelic effects associated with each variant in the target population. Step 3: Trait value of an individual of unknown phenotype can be estimated using his/her genotypes and the allelic effects estimated in step 2.

Finally, the predicted trait values of individuals, with genotype $\boldsymbol{Z}^{\text{test}}$, are computed by

$$\hat{\boldsymbol{Y}}^{\text{test}} = \sum_{k=1}^{K} \boldsymbol{Z}_k^{\text{test}} \hat{\boldsymbol{u}}_k. \qquad \text{(Equation 2)}$$

Trans-ethnic Genetic Prediction

We now describe the problem of trans-ethnic trait prediction. Consider a setting in which GWASs have been conducted for a trait in a *target* population of primary interest (e.g., AA) and the goal is to predict the trait value of other individuals from the same population. We will refer to the first set of individuals, for whom both genotype data and trait values are used to build the prediction model, as the training set; the second set of individuals, for whom we wish to predict trait values using genotype data, is the testing set. In simulation studies, the (simulated) true trait values of the test set are used as the gold standard to evaluate the predictive accuracy. For trans-ethnic auxiliary information, we use the summary-level statistics—specifically SNP-level p values—from an *independent* auxiliary GWAS of matching trait (e.g., EUR). We make no assumption with regard to the genetic similarity between the target and the auxiliary GWAS—in fact, the auxiliary GWAS can be trans-ethnic or multi-ethnic; the "independent" assumption merely requires that the target GWAS and the auxiliary GWAS consist of non-overlapping individuals.

Figure 1 shows a flow chart of the trans-ethnic prediction approach, which utilizes a two-component LMM (K = 2 in Equation 1). The grouping of the SNPs is based on the summary statistics of the auxiliary GWAS. One component, denoted as C_2, consists of all genotyped SNPs; the other component, C_1, includes only SNPs showing evidence of association in the auxiliary GWAS. The variance parameters of the LMM and the BLUP of the allelic effects are estimated using genotypes of the training set in the target population. The predicted trait values on the testing individuals are computed using their genotype, according to Equation 2. Note that individual-level data from the auxiliary GWAS are not needed in the trans-ethnic prediction.

To select SNPs with strong evidence of association in the base GWAS, we use program XPEB to compute the local false discovery rate (locfdr) defined by Equation 5 in Coram et al.[22] In brief, this approach assumes that the observed test statistics, $\boldsymbol{S}$ (chi-square statistics that can be computed from the p values), arise from a mixture of null and non-null distributions. By maximizing the joint likelihood of all observed test statistics in the base GWAS, the mixture proportions can be estimated. The locfdr at a marker, m, is then computed as the posterior probability that S_m is drawn from the null component. We pre-specify a locfdr threshold, υ^*, such that a SNP is included in C_1 only if its locfdr falls below υ^*. For all simulations presented below, $\upsilon^* = 0.05$. Alternatively, one can pre-specify a p value threshold, τ^*, such that a SNP is included in C_1 if its p value in the base GWAS falls

below τ^*. Since there is a monotonic relationship between the p value and locfdr, using a locfdr criterion is equivalent to using some p value threshold. In the Discussion, we propose approaches that adaptively choose υ^* (equivalently, τ^*).

Simulation

We evaluate the proposed prediction algorithm using simulated genotype and trait data generated in a previous study.[22] The genotype data consist of ~727,000 autosomal SNPs on the Illumina 1M SNP array with minor allele frequencies greater than 1%. It is simulated using program HAPGEN2 and HapMap Phase 3 genotypes and includes eight CEU-like cohorts (simEUR) and one YRI-like cohort (simAFR) of 10,000 individuals each. To simulate the phenotypes, 1,000 SNPs are sampled as causal variants in each ethnic population; the proportions of causal variants that overlap between the simEUR and the simAFR cohorts (δ) range from 0.001 to 1, with $\delta = 0.001$ representing independent genetic architecture and $\delta = 1$ representing complete overlap in causal variants. The true allelic effects at these causal variants are sampled from a Gaussian distribution; the correlation between the allelic effects in simEUR and simAFR is 0.7. Lastly, trait values are simulated using the allelic effects, the simulated genotypes, and a random non-genetic component, whose variance is calibrated to achieve a heritability ranging from 40% to 90%. For each value of δ, 20 sets of trait values are simulated, with causal variants independently sampled from run to run. In practice, not all causal variants are genotyped. To consider such situations, we additionally analyzed simulated data removing all causal variants in both simEUR and simAFR.

To apply the proposed trans-ethnic prediction model, the eight simEUR cohorts are meta-analyzed (total N = 80,000) and treated as the auxiliary GWAS. Within simAFR, we randomly sample 8,000 individuals to form the training dataset, leaving the remaining 2,000 individuals as the testing set for evaluating the predictive accuracy, for which we focus on correlation (r) between the simulated true trait values and the predicted values. For each simulated trait, we compute the locfdr that the SNP is associated with the trait in simEUR using meta-analysis summary statistics. SNPs with locfdr below 0.05 are selected to C_1; all SNPs are included in set C_2. Next, we compute the two-component LMM and BLUP on the simAFR training set, recording the BLUP of allelic effects. These BLUP estimates are then used to predict the corresponding trait values in the 2,000 testing individuals. The accuracy measure reported is $cor(\boldsymbol{y}^*, \widehat{\boldsymbol{y}}^*)$, where $\widehat{\boldsymbol{y}}^*$ is defined by Equation 2.

We compare the proposed method with two alternative approaches. First, we consider a single-component LMM on the simAFR training set (N = 8,000) alone. This comparison assesses the contribution of trans-ethnic information. Second, we consider a two-component LMM that uses only European individuals. For this analysis, a ninth simEUR cohort is simulated as described above and used as the training dataset. For all comparisons, the testing set is the 2,000 simAFR individuals. These comparisons are designed to assess the importance of trans-ethnic and population-specific training data.

Predicting Genetic Scores for Lipid Concentration in African American Women

As a real data example, we analyze lipid concentration in African American females in the Women's Health Initiative SNP Health Association Resource (WHI-SHARe), which includes 8,153 postmenopausal AA women. All WHI participants have provided written informed consent, and institutional review board approval has been obtained at each of the 40 WHI clinical centers and at the clinical coordinating center at the Fred Hutchinson Cancer Research Center. For analyses described in this manuscript, which uses de-identified data, institutional review board approval has been obtained at Stanford University. We randomly sample four non-overlapping sets of 2,000 individuals as test sets; for each test set the remaining individuals are used to compute the BLUP. Prediction accuracy reported in Results are averaged across the four testing sets. For auxiliary GWAS, we use the meta-analysis of the Global Lipids Genetics Consortium (GLGC), which consists largely of individuals of European descent, with no overlap with WHI-SHARe.[3] In GLGC, 47, 37, and 32 loci meet the genome-wide significance threshold of 5×10^{-8} and are associated with high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, and triglyceride (TG) levels, respectively. Using the p value threshold of 5×10^{-8} selects approximately the same SNPs as using a locfdr of 0.05. In the WHI-SHARe AA cohort, genome-wide African ancestry proportion explains less than 1% of the observed phenotypic variance in HDL, LDL, and TG. Nonetheless, to eliminate the effects of admixture, we first regress out the estimated proportion of African ancestry from each lipid trait and use the residuals in XP-BLUP. To assess the effect of the pre-specified threshold, we also apply XP-BLUP varying the p value threshold from 5×10^{-3} to 5×10^{-8}. To compare the predictive accuracy of XP-BLUP with other risk factors, we use the same training individuals to estimate the predictive models that are based on (1) age or (2) BMI. To compare XP-BLUP with a conventional polygenic risk scores (PRSs) approach, we performed single-marker GWAS analysis using training individuals (independent loci are defined as being at least 250 kb apart). PRSs in testing individuals are computed using PLINK's score option, which weighs the index SNPs by the corresponding estimated allelic effects.[27]

Computation

For all results reported in the next section, we used GCTA to perform REML and BLUP estimation, although a number of available software programs can perform similar tasks.[12,17,28] The BLUP output of GCTA is re-processed to account for the two classes of SNPs using a customized script, and the predicted trait values on the testing set are produced in PLINK.[27]

Results

Predictive Accuracy on Simulated Data

Figure 2 compares the predictive accuracy, measured by correlation coefficients, of the proposed trans-ethnic prediction (XP-BLUP) with two alternative LMM methods. Since the simulation model assumes a genetic architecture with a heritability of 0.7, the maximum achievable correlation between the predicted and observed traits is $\sqrt{0.7} = 0.84$. Using the standard, single-component LMM and simAFR data alone (N = 8,000), the mean predictive accuracy on the out-of-sample testing individuals is 0.14 (Figure 2, blue box). As expected, the predictive accuracy of the proposed XP-BLUP increases with the degree of overlap between the auxiliary simEUR and the target simAFR population. When the trait loci completely overlap in

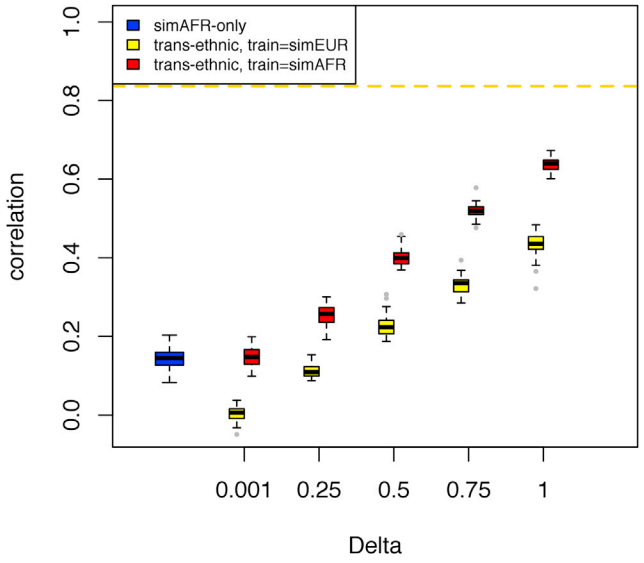

Figure 2. Comparison of Predictive Correlation
XP-BLUP (red) compares favorably to a single-component linear mixed-effects model (LMM, blue) and a two-component LMM model trained only on simulated European GWAS data (yellow). Delta specifies the proportion of overlapping loci between ethnicities. Dashed line is the theoretical maximum.

the two populations, the predictive accuracy is 0.64, indicating that the summary-level statistics from simEUR GWAS provide substantial information. On the other hand, for any level of genetic overlap, the predictive accuracy trained entirely on simEUR (Figure 2, yellow boxes) is lower than that trained on simAFR, emphasizing the importance of ethnically matched training cohorts. Additional simulations confirm that the same trend holds for various levels of underlying heritability and when some or all causal variants are not genotyped (Figures S1 and S2).

It should be noted that the improved predictive accuracy of XP-BLUP does not require that all true trait loci in the target population are included in C_1 class. This is because, even with completely overlapping genetic architecture ($\delta = 1$) and a sample size of 80,000, not all simEUR trait loci can be detected. Over all simulations, the average number of loci discovered in the simEUR GWAS is 550 out of 1,000 (range 523–581). The moderate simAFR training samples and the incomplete knowledge of all trait loci in the simAFR population account for the discrepancy between the observed predictive accuracy (0.64) and the theoretical maximum possible accuracy (0.84) at $\delta = 1$. On the other hand, XP-BLUP does not assume all SNPs in C_1 are indeed trait relevant in simAFR. For instance, at $\delta = 0.5$, half of the SNPs in C_1 are expected to bear no predictive information in simAFR, yet XP-BLUP still outperforms the single-component LMM. Lastly, when the genetic architecture is independent in simEUR and simAFR ($\delta = 0.001$), the auxiliary GWAS provides no useful information. In this scenario, XP-BLUP achieves similar accuracy as a single-component LMM, demonstrating its ability to ignore irrelevant auxiliary information. In contrast, a two-component LMM trained on simEUR can perform substantially worse than a single-component LMM when $\delta = 0.001$. Taken together, this set of analyses demonstrates two features of the XP-BLUP framework: first, the predictive model needs to be trained in a population-specific context; second, regardless of the underlying overlap in genetic architecture, trans-ethnic auxiliary information will not negatively impact the predictive accuracy.

BLUP of Allelic Effects

To better understand why the two-component LMM outperforms the single-component LMM, it is instructive to examine the BLUP allelic effects under the two approaches (Figures 3 and S3). Focusing on loci that affect trait values in simAFR (e.g., $\boldsymbol{u}_k \neq 0$), we form two groups based on whether a variant is selected into C_1 in the XP-BLUP analysis, which requires that locfdr < 0.05 in simEUR. The two histograms in the first row (Figures 3A and 3B) represent the true simulated allelic effects in simAFR for each of the two groups. The slight flattening of the histogram representing C_1 (Figure 3A) is due to requiring the variant to pass a significance threshold in the simEUR.

The predicted allelic effects using XP-BLUP and a single-component LMM are shown in the second (Figures 3C and 3D) and third (Figures 3E and 3F) rows, respectively. Under a single-component LMM, the allelic effects of all variants are assumed to have come from a common Gaussian distribution. The REML estimate of this common Gaussian distribution features a small variance of $\widehat{\sigma}_G^2/M$. In consequence, it heavily shrinks the estimated allelic effects toward 0. Such shrinkage captures the reality that a majority of the variants in the genome have zero or infinitesimal effects, in which case shrinkage is an effective strategy to reduce sampling variability of the prediction. Undesirably, however, true trait variants with moderate or even large effects are similarly shrunk (Figures 3E and S3). As the true trait variants are relatively sparse, the noise from non-associated variants overwhelms the true signal, leading to poor predictions.

In contrast, XP-BLUP allows the allelic effects of C_1 SNPs to be drawn from a different Gaussian distribution than the rest of the variants. Consider that for $\delta = 0.75$, approximately 75% of the SNPs in C_1 are associated with the trait in simAFR and the remaining 25% are simEUR specific. The allelic effects of C_1 SNPs estimated by XP-BLUP are much less attenuated (Figure 3C), compared with the corresponding estimates using a single-component LMM (Figure 3E). On the other hand, simAFR trait variants that are not included in C_1, either because they are simAFR specific or because the SNPs did not reach genome-wide significance in the simEUR GWAS, show similar level of shrinkage as those in a single-component LMM (Figures 3D and 3F). This observation suggests that the success of XP-BLUP depends on constructing C_1 SNPs enriched for variants near trait loci in simAFR. At $\delta = 0.001$, the SNPs in C_1 are not enriched for simAFR trait variants compared to the rest of the genome; in this situation, XP-BLUP

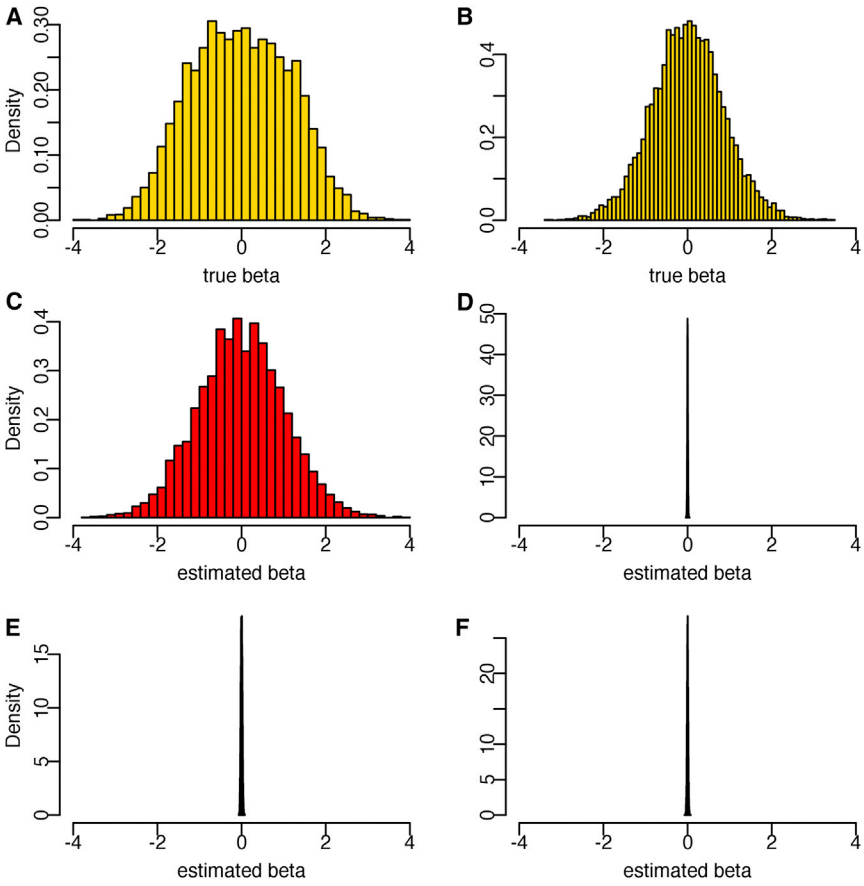

Figure 3. True and BLUP of Trait-Relevant SNP Effects in Simulated Data
True SNP effects (A and B), BLUP under trans-ethnic model (C and D), and BLUP under a single-component model (E and F). SNPs in C_1 (A, C, and E); SNPs not in C_1 (B, D, and F).

effectively ignores the simEUR auxiliary information and shrinks the C_1 SNPs similarly to the rest of the genome (Figure S3).

Predicting Genetic Scores for Lipid Concentration in African American Women

The Women's Health Initiative is a US-based study of common health issues in post-menopausal women. Over a period of 15 years, the study has enrolled a total of more than 160,000 women aged 50–79 years old. Among these, 8,153 African American participants have been genotyped as part of the Women's Health Initiative SNP Health Association Resource (WHI-SHARe). Previously, we have examined the lipid concentrations in this cohort to characterize overlapping genetic architecture.[21] Here we apply the proposed XP-BLUP method to predict lipid concentrations. For LDL, HDL, and TG, the total estimated heritability on the training set is 0.33, 0.28, and 0.26, respectively. As such, the theoretical maximal predictive correlation coefficients are 0.57, 0.53, and 0.51, respectively. Table 1 summarizes the predictive accuracy achieved, when C_1 is defined by conventional genome-wide significance threshold of 5×10^{-8}. Table 1 also includes the corresponding predictive accuracy for a number of epidemiologic risk factors available in this data, such as age and BMI. We find that XP-BLUP achieves higher accuracy than age and BMI for all three lipid traits. Further, XP-BLUP achieves higher accuracy compared to a polygenic risk score approach that includes only the index SNPs of genome-wide significant loci detected in the training dataset. Finally, combining XP-BLUP scores, age, age^2, smoking, BMI, and fasting status (for LDL only), the overall predicted risk scores achieve a correlation of 0.22, 0.30, and 0.25 with LDL, HDL, and TG, respectively.

The definition of $p < 5 \times 10^{-8}$ or locfdr < 0.05 may seem arbitrary. To investigate the sensitivity of the performance of XP-BLUP with respect to this threshold definition, we re-analyze the lipid traits by varying the threshold ranging from 5×10^{-3} to 5×10^{-8}. As the inclusion threshold is relaxed from 5×10^{-8} to 5×10^{-5}, the number of SNPs included in C_1 increases by a factor of ten, but the predictive correlation remains relatively stable (Tables S1–S3). We reason that this phenomenon reflects a tradeoff between true signal and noise: although a bigger set of C_1 SNPs likely includes more trait variants, an increasing proportion of added SNPs are false positives. Consequentially, the BLUP for C_1 SNPs are increasingly shrunk toward zero. This reasoning is supported by the observation that the phenotypic variance explained by C_1 SNPs does not always increase as more SNPs are included (column VG_1 in Tables S1–S3). The optimal threshold that balances this trade-off depends on the underlying genetic architecture as well as the sample size of training set. With much larger training samples, one can choose an optimal threshold using the following cross-validation (CV) approach.[29] The training set is divided into a number of equal parts; to be specific, we describe a ten-fold CV. At each p value threshold, the two-component LMM model is computed on nine parts of the data, and the predictive accuracy is evaluated on the remaining part, termed the validation set. In turn, each of the ten parts is treated as the validation set once; the optimal threshold value is the one that achieves the best predictive accuracy averaging over all ten validation sets. However, at the sample size of WHI-SHARe, we found no substantial differences in the predictive accuracy when the threshold ranges from more than three orders of magnitudes (Tables S1–S3). Therefore, for

Table 1. Correlation Coefficients between Predicted Scores and Observed Lipid Concentration in African Americans of the Women's Health Initiative SNP Health Association Resource

Trait	$\lvert C_1 \rvert$	VG/VP	XP-BLUP	LMM	Age	BMI	FE
LDL	413	0.33	0.18	0.095	0.085	0.044	0.16
HDL	569	0.28	0.22	0.086	0.042	0.21	0.17
TG	413	0.26	0.17	0.078	0.066	0.11	0.13

All correlations are evaluated on the African American testing set individuals. Abbreviations are as follows: $\lvert C_1 \rvert$, number of SNPs in C_1 set; VG/VP, SNP-heritability estimated by program GCTA; XP-BLUP, proposed trans-ethnic best linear unbiased prediction; C_1 SNPs are defined by $p < 5 \times 10^{-8}$ in Global Lipids Genetics Consortium (GLGC) meta-analysis; LMM, standard linear mixed effects model computed by GCTA; age, fixed effects model with age as a single predictor; BMI, fixed effects model with BMI as a single predictor; FE, fixed effects model that includes only index SNPs at genome-wide significant loci in the African American training set.

current studies in minority populations, we recommend applying XP-BLUP with a pre-specified threshold of $p < 5 \times 10^{-8}$.

Discussion

We have investigated XP-BLUP, a computational approach for assessing the genetic risk of complex traits in minority populations. The underlying idea is to use a two-component LMM, in which one component is highly enriched for trait-associated SNPs that are derived from independent, well powered, but possibly ethnically unmatched GWASs. Although an LMM with multiple random effects is a generally useful approach for integrating external knowledge of candidate regions or SNPs, this framework is particularly appealing for leveraging trans-ethnic GWAS results because it benefits from partial overlap in genetic architectures across populations, without assuming transferability of all trait loci between populations, as we explain in detail next.

First, both simulation (Figure 1) and the lipid data analysis (Table 1) emphasize that the performance of a multi-component LMM for prediction strongly depends on the level of enrichment in SNP set C_1. The ideal C_1 is the set that includes all and only trait-associating SNPs in the target population. The predictive accuracy diminishes when C_1 is heavily contaminated with SNPs not relevant in the target population, or it misses trait-associating SNPs. In theory, it is possible to choose C_1 and compute BLUP using the same GWAS data;[17,18] however, for minority studies with moderate sample size, a set constructed this way is likely less accurate. In contrast, for a variety of complex traits studied to date, accumulating evidence suggests considerable numbers of overlap loci between ethnicities.[30–33] For these traits, a large proportion of SNPs selected based on a large European GWAS are expected to be relevant in non-European populations. Therefore, a set selected based on a large GWAS in a non-matching population is expected to harbor higher fraction of variants for the target population than a set selected on a small dataset in matching population. Likewise, we expect that the trans-ethnic GWAS data provides more specific information than trait-independent biological annotation, such as chromatin accessibility. It should be noted that C_1 does not need to exclusively contain the causal variants and should be interpreted accordingly. For the purpose of prediction, any SNPs in linkage disequilibrium with causal variants can serve as useful predictors. This feature makes XP-BLUP easier to use in practice than a fixed-effects PRS model that requires "independent" risk SNPs, because defining independent trait loci and identifying likely causal variants, which may vary across populations, are challenging on their own.

Second, while our XP-BLUP approach takes advantage of the strong overlap between ethnicities, it does not assume *all* SNPs in C_1 to be associated with the trait in the minority population. In other words, the method accommodates incomplete overlap in genetic architecture. This is important because it is difficult to assess, a priori, the degree of overlap for a particular trait. Our simulation results suggest that XP-BLUP is robust to contamination of mis-specified SNPs. SNPs that are strongly associated in European populations but not in the target population tend to have small BLUP estimates because these BLUPs are estimated using training data in the target population. When C_1 includes a substantial fraction of non-associated SNPs, the prediction accuracy deteriorates due to the shrinkage effect; however, even when C_1 consists of randomly sampled SNPs ($\delta = 0.001$), the performance of XP-BLUP is not worse than a standard one-component LMM model. Furthermore, we choose not to make any assumption regarding the agreement, in sign or magnitude, between the estimated allelic effects in the European GWAS and those in the target population. This choice is based on the empirical evidence that the correlation between allelic effects across ethnicities varies by trait, and currently we do not have sufficient information to accurately model such correlation structure.

The XP-BLUP approach described here can be extended in several directions. First, it is anticipated that GWASs with expanded cohorts will continue to discover loci associated with traits and diseases. As the number of SNPs becomes sufficiently large and includes variants of infinitesimal effects, it may be useful to divide C_1 SNPs into multiple sets, representing variants with varying level of allelic effects. A practical approach is to split C_1 SNPs on the basis of putative contributions of each variant, defined by $2\widehat{\beta}^2 p(1-p)$, where $\widehat{\beta}$ is the estimated allelic effects in the auxiliary GWAS and p is allele frequency in the target population. At the same time, expanded minority cohorts may enable us to discover target population-specific variants. These variants can be incorporated in XP-BLUP framework as fixed effects. Second, our current implementation focuses on one trait and makes use of auxiliary GWASs of the matching trait. Recent studies have demonstrated widespread pleiotropy: genetic loci that simultaneously influence related and sometimes

seemingly unrelated traits.[34] Integrating pleiotropic information to jointly predict related traits has shown promise.[35] Extending XP-BLUP to leverage both trans-ethnic and cross-trait information may be particularly useful for under-studied traits. Third, we have focused on quantitative traits and have described XP-BLUP as a linear mixed-effects model. For a binary disease outcome, a commonly adopted approach is the liability-threshold model, which postulates that genetic factors contribute additively to an underlying quantitative trait, liability, and the observed disease status is a dichotomized version of the liability at a specific threshold. However, in most case-control GWASs, affected individuals are over-sampled and therefore the underlying liability in the cohorts may be strongly skewed. Computational strategies for estimating heritability and predicting disease status in case-control designs have been developed and may be adapted for trans-ethnic disease prediction.[36,37]

While trans-ethnic information can substantially improve predictive accuracy, our work emphasizes the importance of estimating the allelic effects in ethnically matched cohorts. Therefore, sustained efforts should be dedicated to building large and well-phenotyped cohorts in minority populations. In the coming years, a number of large cohorts, such as the US National Institutes of Health Precision Medicine Initiative and the Millions Veterans Program, will enable researchers to investigate the genetic basis of a wide range of complex traits and diseases in diverse populations. We anticipate that approaches leveraging trans-ethnic information will continue to play an important role in these studies.

Supplemental Data

Supplemental Data include three figures and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.06.015.

Acknowledgments

This study was supported by US NIH grant R01 GM073059 (M.A,C., H.F., S.I.C., H.T.). M.A.C. is an employee of Google, Inc. We thank three anonymous reviewers for their constructive comments and the Genetics Bioinformatics Service Center at Stanford University for computational resources. The WHI program is funded by the National Heart, Lung, and Blood Institute, NIH, US Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C.

Received: February 8, 2017
Accepted: June 29, 2017
Published: July 27, 2017

Web Resources

GCTA, http://www.complextraitgenomics.com/software/gcta/
PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/
dbGaP, http://www.ncbi.nlm.nih.gov/gap
XPEB program, http://med.stanford.edu/tanglab/software/
XP-BLUP script, https://github.com/tanglab/XP-BLUP

References

1. Park, J.H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat. Genet. *42*, 570–575.
2. 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.
3. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707–713.
4. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; and LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.
5. Berndt, S.I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M.F., Justice, A.E., Monda, K.L., Croteau-Chonka, D.C., Day, F.R., et al. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat. Genet. *45*, 501–512.
6. Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., and Aittokallio, T. (2014). Regularized machine learning in the genetic prediction of complex traits. PLoS Genet. *10*, e1004754.
7. Wacholder, S., Hartge, P., Prentice, R., Garcia-Closas, M., Feigelson, H.S., Diver, W.R., Thun, M.J., Cox, D.G., Hankinson, S.E., Kraft, P., et al. (2010). Performance of common genetic variants in breast-cancer risk models. N. Engl. J. Med. *362*, 986–993.
8. Weedon, M.N., Lango, H., Lindgren, C.M., Wallace, C., Evans, D.M., Mangino, M., Freathy, R.M., Perry, J.R., Stevens, S., Hall, A.S., et al.; Diabetes Genetics Initiative; Wellcome Trust Case Control Consortium; and Cambridge GEM Consortium (2008). Genome-wide association analysis identifies 20 loci that influence adult height. Nat. Genet. *40*, 575–583.
9. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P.; and International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
10. Shi, J., Park, J.H., Duan, J., Berndt, S.T., Moy, W., Yu, K., Song, L., Wheeler, W., Hua, X., Silverman, D., et al.; MGS (Molecular Genetics of Schizophrenia) GWAS Consortium; GECCO (The Genetics and Epidemiology of Colorectal Cancer Consortium); GAME-ON/TRICL (Transdisciplinary Research in Cancer of the Lung) GWAS Consortium; PRACTICAL (PRostate cancer AssoCiation group To Investigate Cancer Associated aLterations) Consortium; PanScan Consortium; and GAME-ON/ELLIPSE Consortium (2016). Winner's curse correction and variable thresholding improve performance

of polygenic risk modeling based on genome-wide association study summary-level data. PLoS Genet. *12*, e1006493.

11. de los Campos, G., Gianola, D., and Allison, D.B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. *11*, 880–886.
12. Canela-Xandri, O., Law, A., Gray, A., Woolliams, J.A., and Tenesa, A. (2015). A new tool called DISSECT for analysing large genomic data sets using a big data approach. Nat. Commun. *6*, 10162.
13. Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. Biometrics *31*, 423–447.
14. 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.
15. 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.
16. 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.
17. Speed, D., and Balding, D.J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. *24*, 1550–1557.
18. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264.
19. Guan, Y., and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann. Appl. Stat. *5*, 1780–1815.
20. 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.
21. Coram, M.A., Duan, Q., Hoffmann, T.J., Thornton, T., Knowles, J.W., Johnson, N.A., Ochs-Balcom, H.M., Donlon, T.A., Martin, L.W., Eaton, C.B., et al. (2013). Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. Am. J. Hum. Genet. *92*, 904–916.
22. Coram, M.A., Candille, S.I., Duan, Q., Chan, K.H., Li, Y., Kooperberg, C., Reiner, A.P., and Tang, H. (2015). Leveraging multi-ethnic evidence for mapping complex traits in minority populations: an empirical Bayes approach. Am. J. Hum. Genet. *96*, 740–752.
23. Makowsky, R., Pajewski, N.M., Klimentidis, Y.C., Vazquez, A.I., Duarte, C.W., Allison, D.B., and de los Campos, G. (2011). Beyond missing heritability: prediction of complex traits. PLoS Genet. *7*, e1002051.
24. 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.
25. Corbeil, R.R., and Searle, S.R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. Technometrics *18*, 31–38.
26. McCulloch, C.E., Searle, S.R., and Neuhaus, J.M. (2001). Generalized Linear Mixed Models (Wiley).
27. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
28. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
29. Hastie, T., Tibshirani, R., and Friedman, J. (2011). The Elements of Statistical Learning: Data Mining (Inference, and Prediction Springer).
30. Carty, C.L., Johnson, N.A., Hutter, C.M., Reiner, A.P., Peters, U., Tang, H., and Kooperberg, C. (2012). Genome-wide association study of body height in African Americans: the Women's Health Initiative SNP Health Association Resource (SHARe). Hum. Mol. Genet. *21*, 711–720.
31. Franceschini, N., Fox, E., Zhang, Z., Edwards, T.L., Nalls, M.A., Sung, Y.J., Tayo, B.O., Sun, Y.V., Gottesman, O., Adeyemo, A., et al.; Asian Genetic Epidemiology Network Consortium (2013). Genome-wide association analysis of blood-pressure traits in African-ancestry individuals reveals common associated genes in African and non-African populations. Am. J. Hum. Genet. *93*, 545–554.
32. Dumitrescu, L., Carty, C.L., Taylor, K., Schumacher, F.R., Hindorff, L.A., Ambite, J.L., Anderson, G., Best, L.G., Brown-Gentry, K., Bůžková, P., et al. (2011). Genetic determinants of lipid traits in diverse populations from the population architecture using genomics and epidemiology (PAGE) study. PLoS Genet. *7*, e1002138.
33. Carlson, C.S., Matise, T.C., North, K.E., Haiman, C.A., Fesinmeyer, M.D., Buyske, S., Schumacher, F.R., Peters, U., Franceschini, N., Ritchie, M.D., et al.; PAGE Consortium (2013). Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. PLoS Biol. *11*, e1001661.
34. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. (2013). Pleiotropy in complex traits: challenges and strategies. Nat. Rev. Genet. *14*, 483–495.
35. Li, C., Yang, C., Gelernter, J., and Zhao, H. (2014). Improving genetic risk prediction by leveraging pleiotropy. Hum. Genet. *133*, 639–650.
36. Golan, D., and Rosset, S. (2014). Effective genetic-risk prediction using mixed models. Am. J. Hum. Genet. *95*, 383–393.
37. Golan, D., Lander, E.S., and Rosset, S. (2014). Measuring missing heritability: inferring the contribution of common variants. Proc. Natl. Acad. Sci. USA *111*, E5272–E5281.