Leveraging Functional Annotations and Multiethnic Data to Improve Polygenic Risk Prediction

Leveraging Functional Annotations and Multiethnic Data to Improve Polygenic Risk Prediction

A thesis presented

by

Carla Marquez Luna

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

September 2018

Dissertation Advisor: Professor Alkes Price Carla Marquez Luna

Leveraging Functional Annotations and Multiethnic Data to Improve Polygenic Risk Prediction

Abstract

Polygenic risk prediction is a widely-investigated topic because of its potential clinical application as well as its utility to have a better understanding of the genetic architecture of complex traits. Methods to perform polygenic risk prediction can be divided into 2 categories: methods that use only summary statistics such as pruning+thresholding[1,2] and LDpred[3]; and methods that require individual level data for both genotypes and phenotypes (BLUP and its variations). Polygenic risk prediction can achieve substantial accuracy when training data is available at large sample sizes. Due to restrictions of sharing individual-level data, methods that use summary statistics only are of special interest. In this work we focus on summary statistics based methods to perform polygenic risk prediction. The first chapter, presents a method that increases polygenic risk prediction accuracy in non-European populations. In the second chapter, we introduce a method that leverages trait-specific functional enrichments to increase prediction accuracy. In the third chapter, we develop a method that increases association power in meta-analysis.

In chapter one, we develop a multiethnic polygenic risk score that increases prediction accuracy in non-European population. To date, most available training data involves samples of European ancestry, and it is currently unclear how to accurately predict in other populations. Previous studies, have used either training data from European samples or training from the target population. Here, we introduce a multiethnic polygenic risk score that leverages training data from European samples and training data from the target population. The method takes advantage of both the accuracy that can be achieved with large training samples[4,5] and the accuracy that can be achieved with training data containing the same LD patterns as the target population. In application to predict type 2

diabetes (T2D) in Latino target samples in the SIGMA T2D data set[6], we attained a $> 70\%$ relative improvement in prediction accuracy (from R^2 = 0.027 to 0.047) compared to methods that use only one source of training data. We attained similar relative improvements in simulations. We also obtained a $> 70\%$ relative improvement in an analysis to predict T2D in a South Asian UK Biobank cohort, and a 30% relative improvement in an analysis to predict height in an African UK Biobank cohort.

In chapter two, we introduce a new method for polygenic risk prediction, LDpred-funct that leverages trait-specific functional enrichments to increase prediction accuracy. We fit functional priors using our recently developed baseline-LD model[7], which includes coding, conserved, regulatory and LD-related annotations. LDpred-funct first analytically estimates posterior mean causal effect sizes, accounting for functional priors and LD between variants. LDpred-funct then uses cross-validation within validation samples to regularize causal effect size estimates in bins of different magnitude, improving prediction accuracy for sparse architectures. We applied our method to predict 16 highly heritable traits in the UK Biobank. We used association statistics from British-ancestry samples as training data (avg N=365K) and samples of other European ancestries as validation data (avg N=22K), to minimize confounding. LDpred-funct attained a $+27\%$ relative improvement in prediction accuracy (avg prediction $R^2 = 0.173$; highest $R^2 = 0.417$ for height) compared to existing methods that do not incorporate functional information, consistent with simulations.

In chapter three, we introduce a summary statistic based extension of mixed model association method (Meta-LMM) that increases association power in meta-analysis. Meta-analysis of genome-wide summary statistics has been a succesful strategy to discover genetic risk variants. The most commonly used method is using inverse-variance weighting fixed effects meta-analysis, due to limitations of sharing individual-level data, most meta-analysis only share summary statistics. On the other hand, linear mixed model association approaches gain power by reducing phenotypic noise by conditioning out on known casual variants or using leave-one-chromosome-out scheme[8,9]. This method aims to increase power by reducing the phenotypic noise within each cohort by conditioning out using a leave-one-chromose-out scheme and using the other cohorts summary

statistics as training. We use the UK Biobank dataset to construct 10 independent cohorts ($N = 33K$ each), and applied Meta-LMM to 14 UK Biobank traits. Meta-LMM substantially outperformed fixed-effects meta-analysis, with a $+15\%$ median increase in χ^2 statistics (averaged across traits), consistent with simulations. And we show that on average 20% more loci were identified with Meta-LMM compared to fixed-effects meta-analysis. Our results show that this method outperforms most commonly used mehtods for meta-analysis.

Contents

Multi-ethnic polygenic risk scores improve risk prediction in diverse populations

Carla Márquez-Luna[1], Po-Ru Loh[2,3], South Asian Type 2 Diabetes (SAT2D) Consortium,
The SIGMA Type 2 Diabetes Consortium, Alkes L. Price[1,2,3]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA.
[2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.
[3]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA.
[4]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.
[5]23andMe Inc., Mountain View, CA, USA.

Abstract

Abstract Methods for genetic risk prediction have been widely investigated in recent years. However, most available training data involves European samples, and it is currently unclear how to accurately predict disease risk in other populations. Previous studies have used either training data from European samples in large sample size or training data from the target population in small sample size, but not both. Here, we introduce a multi-ethnic polygenic risk score that combines training data from European samples and training data from the target population. We applied this approach to predict type 2 diabetes (T2D) in a Latino cohort using both publicly available European summary statistics in large sample size (N_{eff}=40k) and Latino training data in small sample size (N_{eff}=8k). Here, we attained a $> 70\%$ relative improvement in prediction accuracy (from $R^2 = 0.027$ to $R^2 = 0.047$) compared to methods that use only one source of training data, consistent with large relative improvements in simulations. We observed a systematically lower load of T2D risk alleles in Latino individuals with more European ancestry, which could be explained by polygenic selection in ancestral European and/or Native American populations. We predict T2D in a South Asian UK Biobank cohort using European (N_{eff} =40k) and South Asian (N_{eff}=16k) training data and attained a $> 70\%$ relative improvement in prediction accuracy, and application to predict height in an African UK Biobank cohort using European (N=113k) and African (N=2k) training data attained a 30% relative improvement. Our work reduces the gap in polygenic risk prediction accuracy between European and non-European target populations.

KEY WORDS: genome-wide association study; polygenic prediction; height; type 2 diabetes

Introduction

Genetic risk prediction is an important and widely investigated topic because of its potential clinical application as well as its application to better understand the genetic architec-

ture of complex traits[10]. Many polygenic risk prediction methods have been developed and applied to complex traits. These include polygenic risk scores (PRS)[1–5,11–13], which use summary association statistics as training data, and Best Linear Unbiased Predictor (BLUP) methods and their extensions[14–21], which require individual-level genotype and phenotype data.

However, all of these methods are inadequate for polygenic risk prediction in non-European populations, because they consider training data from only a single population. Existing training data sets have much larger sample sizes in European populations, but the use of European training data for polygenic risk prediction in non-European populations reduces prediction accuracy, due to different patterns of linkage disequilibrium (LD) (or potentially due to different causal effects)[1,3,22,23]. For example, ref. 3 reported a relative decrease of 53-89% in schizophrenia risk prediction accuracy in Japanese and African-American populations compared to Europeans when applying PRS methods using European training data. An alternative is to use training data from the same population as the target population, but this would generally imply a much lower sample size, reducing prediction accuracy.

To tackle this problem, we developed an approach that combines PRS based on European training data with PRS based on training data from the target population. The method takes advantage of both the accuracy that can be achieved with large training samples[4,5] and the accuracy that can be achieved with training data containing the same LD patterns as the target population. In application to predict type 2 diabetes (T2D) in Latino target samples in the SIGMA T2D data set[6], we attained a >70% relative improvement in prediction accuracy (from R^2 =0.027 to R^2 =0.047) compared to methods that use only one source of training data. We attained similar relative improvements in simulations. We also obtained a >70% relative improvement in an analysis to predict T2D in a South Asian UK Biobank cohort, and a 30% relative improvement in an analysis to predict height in an African UK Biobank cohort.

Materials and Methods

Polygenic risk score using a single training population

Polygenic risk scores are constructed using SNP effect sizes estimated from genome-wide association studies, which perform marginal regression of the phenotype of interest on each SNP in turn. Explicitly, for continuous traits, we estimate effect sizes (where $i = 1, \ldots, M$ indexes genetic markers) using the model $y = b_0 + b_i g_i + b_{PC} PC + \epsilon$, where g_i denotes genotypes at marker i, PC denotes one or more principal components used to adjust for ancestry, and ϵ denotes environmental noise. For binary traits, we use the analogous logistic model $logit[P(y = 1)] = b_0 + b_i g_i + b_{PC} PC + \epsilon$.

Given a vector of estimated effect sizes $\widehat{b}_l$ from a genome-wide association study performed on a set of training samples, the polygenic risk score[1] (PRS) for a target individual with genotypes g_i is defined as $\widehat{y} = \sum_{i=1}^{M} \widehat{b}_l g_i$. In practice, rather than computing the PRS using estimated effect sizes for all available genetic markers, the PRS is computed on a subset of genetic markers obtained via informed LD-pruning[2] (also known as LD-clumping) followed by P-value thresholding[1]. Specifically, this "pruning + thresholding" strategy has two parameters, R^2_{LD} and P_T, and proceeds as follows. First, we prune the SNPs based on a pairwise threshold R^2_{LD}, removing the less significant SNP in each pair (using PLINK; see Web Resources). Second, we restrict to SNPs with an association P-value below the significance threshold P_T.

The parameters R^2_{LD} and P_T are commonly tuned using on validation data to optimize prediction accuracy[1,2]. While in theory this procedure is susceptible to overfitting, in practice, validation sample sizes are typically large, and R^2_{LD} and P_T are selected from a small discrete set of parameter choices, so overfitting is considered to have a negligible effect. Accordingly, in this work, we consider $R^2_{LD} \in \{0.1, 0.2, 0.5, 0.8\}$ and $P_T \in \{1.0, 0.8, 0.5, 0.4, 0.3, 0.2, 0.1, 0.08, 0.05, 0.02, 0.01, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$, and we always report results corresponding to the best choices of these parameters. In all of our primary analyses involving two training populations (see below), values of R^2_{LD} and P_T were optimized based only on PRS in a single training population, to ensure that PRS using two training populations did not gain any relative advantage from the optimization

of these parameters.

In this work, we specifically consider PRS built using European (EUR), Latino (LAT), South Asian (SAS), or African (AFR) training samples. We use the notation to denote PRS built using European samples, and analogously for the other populations.

Polygenic risk score using two training populations

Given a pair of polygenic risk scores computed as above using two distinct training populations, we define the multi-ethnic PRS with mixing weights α_1 and α_2 as the linear combination of the two PRS with these weights: e.g., for EUR and LAT, we define $PRS_{EUR+LAT} = \alpha_1 PRS_{EUR} + \alpha_2 PRS_{LAT}$. We employ two different approaches to avoid overfitting. In our primary analyses, we estimate mixing weights α_1 and α_2 using validation data and compute adjusted R^2 to account for the additional degree of freedom. In our secondary analyses, we estimate mixing weights α_1 and α_2 using cross-validation (see Assessment of methods below).

For comparison purposes in analyses of real phenotypes, we also evaluated a meta-analysis PRS (e.g. EUR-LAT-meta) using a sample size weighted average of estimated effect sizes in each population[24]; for dichotomous phenotypes we weighted by effective sample size $N_{eff} = 4/(1/N_{case} + 1/N_{control})$. We performed LD-pruning and P-value thresholding using P-values obtained from the meta-analysis, using the LD reference panel from the population that achieved the highest prediction accuracy.

Polygenic risk score using one or two training populations and genetic ancestry

We further define polygenic risk scores that include an ancestry predictor, namely, the top principal component in a given data set, computed using the union of all available (training and validation) samples from that population. (We considered only the top PC in each data set that we analyzed, because lower PCs had a squared correlation with phenotype lower than 0.005 in each case; we recommend that ancestry predictors restrict to PCs with squared correlation with phenotype of 0.005 or larger.) We define a polygenic risk score LAT+ANC with mixing weights α_1 and α_2 as $PRS_{LAT+ANC} = \alpha_1 PRS_{LAT} +$

$\alpha_2 PC$, and we define a polygenic risk score EUR+LAT+ANC with mixing weights α_1, α_2 and α_3 as $PRS_{EUR+LAT+ANC} = \alpha_1 PRS_{EUR} + \alpha_1 PRS_{LAT} + \alpha_3 PC$. As above, we employ two different approaches to avoid overfitting: in our primary analyses, we estimate mixing weights using validation data and compute adjusted R^2; in our secondary analyses, we estimate mixing weights using cross-validation.

Assessment of methods

We assessed the accuracy of polygenic risk scores in validation samples (independent from samples used to estimate effect sizes). We used adjusted R^2 as the accuracy metric for continuous traits and liability-scale adjusted R^2 (ref. 25) for binary traits. Adjusted R^2 is defined as $\widehat{R}^2 - (1 - \widehat{R}^2)\frac{p}{n-p-1}$, where $p \in \{1, 2, 3\}$ is the number of PRS or ANC components in the mixture, n is the number of validation samples, and $\widehat{R}^2$ is the raw (unadjusted) R^2. The adjusted R^2 metric roughly corrects for increased model complexity in multi-component PRS, so in our primary analyses, we report accuracy as adjusted R^2 using best-fit mixing weights $\widehat{\alpha}_k$ estimated using the validation data.

To verify that this metric provides robust model comparisons, we also performed auxiliary analyses in which we used 10-fold cross-validation: specifically, for each left-out fold in turn, we estimated mixing weights using the other 9 folds and evaluated adjusted R^2 for PRS computed using these weights on the left-out fold. We then computed average adjusted R^2 across the 10 folds. (When analyzing data from an unbalanced case-control study with #cases $<<$ #controls, we used stratified 10-fold cross-validation, selecting the folds such that each fold had the same case-control ratio; this applies only to the South Asian UK Biobank T2D analysis.)

Finally, for analyses in which we needed to use samples from the same cohort for both building PRS (i.e., estimating effect sizes $\widehat{b}_i$) and validation, we also used cross-validation. In our primary analyses, we employed 10-fold cross-validation, using 90% of the cohort to estimate $\widehat{b}_i$ and the remaining 10% of the cohort to validate predictions (using the adjusted R^2 metric with best-fit mixture weights $\widehat{\alpha}_k$). In our secondary analyses, we employed 10×9-fold cross-validation, in which 90% of the cohort was used to estimate both $\widehat{b}_i$ and $\widehat{\alpha}_k$ and the remaining 10% of the cohort was used to validate predictions. To estimate $\widehat{\alpha}_k$,

we iteratively split the 90% set of training samples into an 80% training-training set and a 10% training-test set; we estimated $\widehat{b}_i$ in the 80% training-training set and computed a PRS for the 10% training-test set for each of the 9 training-test folds, and we then performed a single regression of phenotype against each PRS across the entire 90% set of training samples to estimate $\widehat{\alpha}_k$. Finally, we re-estimated $\widehat{b}_i$ for the final test prediction using the entire 90% set of training samples.

Simulations

We simulated quantitative phenotypes using real genotypes from European (WTCCC2) and Latino (SIGMA) data sets (see below). We fixed the proportion of causal markers at 1% and fixed SNP-heritability h_g^2 at 0.5, and sampled normalized effect sizes β_i from a normal distribution with variance equal to h_g^2 divided by the number of causal markers. We calculated per-allele effect sizes b_i as $b_i = \frac{\beta_i}{\sqrt{2*p_i(1-p_i)}}$, where p_i is the minor allele frequency of SNP i in the European data set. We simulated phenotypes as $Y_j = \sum_{i=1}^{M} b_i g_{ij} + \epsilon_j$, where $\epsilon_j \sim N(0, 1 - h_g^2)$.

In our primary simulations, we discarded the causal SNPs and used only the non-causal SNPs as input to the prediction methods (i.e. we simulated untyped causal SNPs, which we believe to be realistic). As an alternative, we also considered simulations in which we included the causal SNPs as input to the prediction methods (i.e., a scenario in which causal SNPs are typed). We performed simulations using all available European (WTCCC2) and Latino (SIGMA) training data (approximately a 2:1 ratio). We also performed simulations using training data in which Europeans were subsampled to attain a 1:1 ratio, as the relative performance of different methods may depend on relative training sample sizes; we considered different training sample sizes rather than different validation sample sizes, because the validation sample size does not (in expectation) impact the prediction accuracy.

We also performed simulations in which Latino phenotypes were explicitly correlated to ancestry (population stratification). In these simulations, we added a constant multiple of PC1 (representing European vs. Native American ancestry, with positive values representing higher European ancestry) to the Latino phenotypes such that the correlation

between phenotype and PC1 was equal to -0.11, which is the correlation between the T2D phenotype and PC1 in the SIGMA data set.

We performed simulations under 4 different scenarios: (i) using all chromosomes, (ii) using chromosomes 1-4, (iii) using chromosomes 1-2, and (iv) using chromosome 1 only. The motivation for performing simulations with a subset of chromosomes was to increase N/M, extrapolating to performance at larger sample sizes, as in previous work[3].

Simulation data sets: WTCCC2 and SIGMA

Our simulations used real genotypes from the WTCCC2 and SIGMA data sets (rows 1-2 of Table 1.1). The WTCCC2 data set consists of 15,622 unrelated European samples from a multiple sclerosis study genotyped at 360,557 SNPs after QC[8,26] (see Web Resources). The SIGMA data set consists of 8,214 unrelated Latino samples genotyped at 2,440,134 SNPs after QC[6] (see Web Resources). We restricted our simulations to 232,629 SNPs present in both data sets (with matched reference and variant alleles) after removing A/T and C/G SNPs to eliminate potential strand ambiguity.

Training and validation data sets for predicting type 2 diabetes in Latinos: DIAGRAM, SIGMA and UK Biobank

Our analyses of type 2 diabetes in Latinos used summary association statistics from the DIAGRAM data set and genotypes and phenotypes from the SIGMA data set (row 3 of Table 1.1). The DIAGRAM data set consists of 12,171 cases and 56,862 controls of European ancestry for which summary association statistics at 2,473,441 imputed SNPs are publicly available (see Web Resources)[27]. As noted above, the SIGMA data set consists of 8,214 unrelated Latino samples (3,848 type 2 diabetes cases and 4,366 controls) genotyped at 2,440,134 SNPs after QC. QC procedures are reported in ref. 6, and include the removal of one individual from each pair of relatives with relatedness greater than 10% ($n = 532$), as well as a PCA analysis using EIGENSTRAT[28] (see Web Resources) to identify and remove samples with evidence of high African or East Asian ancestry ($n = 181$).

SIGMA association statistics were computed with adjustment for 2 PCs, as in ref. 6. We restricted our analyses of type 2 diabetes to 776,374 SNPs present in both data sets (with

Table 1.1: List of data sets used in simulations and analyses of real phenotypes. We list the training and validation data sets and validation procedures used in simulations (rows 1-2), predicting T2D in Latinos (rows 3-4), predicting T2D in South Asians (row 5) and predicting height in Africans (row 6). N refers to sample size (continuous traits), Neff refers to effective sample size $4/(1/N_{case} + 1/N_{control})$ (dichotomous traits). $*$: sample size in each training fold. $**$: sample size in union of validation folds.

Target population	Trait	European training	Target population training	Target population validation	Validation procedure (primary)	Validation procedure (secondary)
Latino	2:01	WTCCC2	SIGMA	SIGMA	10-fold cross validation	NA
	Simulations	(N=15,622)	(N=7,393$*$)	(N=8,214$**$)		
Latino	1:01	WTCCC2	SIGMA	SIGMA	10-fold cross validation	NA
	Simulations	(N=7,393)	(N=7,393$*$)	(N=8,214$**$)		
Latino	T2D	DIAGRAM	SIGMA	SIGMA	10-fold cross validation	10x9-fold cross-validation
		(N_{eff} =40,101)	(N_{eff}=7,363$*$)	(N_{eff} =8,181$**$)		
Latino	T2D	UK Biobank	SIGMA	SIGMA	10-fold cross validation	NA
		(N_{eff} =19,842)	(N_{eff}=7,363$*$)	(N_{eff} =8,181$**$)		
South Asian	T2D	DIAGRAM	SAT2D	UK Biobank	In-sample fit	10-fold cross validation
		(N_{eff} =40,101)	(N_{eff} =16,065)	(N_{eff}=919)		
African	Height	UK Biobank	N'Diaye et al.	UK Biobank	In-sample fit	10-fold cross validation
		(N=113,660)	(N=20,427)	(N=1,745)		

matched reference and variant alleles) after removing A/T and C/G SNPs to eliminate potential strand ambiguity. For the SIGMA data set, we used the top 2 PCs as computed in ref. 6. We also performed an analysis of type 2 diabetes using imputed genotypes from the SIGMA T2D data set[6], restricting to 2,062,617 SNPs present in both data sets (with matched reference and variant alleles) after removing A/T and C/G SNPs to eliminate potential strand ambiguity.

We performed a secondary analysis using 113,851 British samples from UK Biobank[29] (see Web Resources) as European training data (5,198 type 2 diabetes cases and 108,653 controls) (row 4 of Table 1.1). UK Biobank association statistics were computed with adjustment for 10 PCs[29], estimated using FastPCA[30] (see Web Resources). We computed summary statistics for 608,878 genotyped SNPs from UK Biobank after removing A/T and C/G SNPs to eliminate potential strand ambiguity. We analyzed 187,142 SNPs present in the SIGMA and UK Biobank data sets. We defined type 2 diabetes cases in UK Biobank as "any diabetes" with "age of diagnosis > 30". We note that the p-values at two top type 1 diabetes (T1D) loci (rs2476601, rs9268645) were only nominally significant ($p \sim 0.05$) for this T2D phenotype, indicating low contamination with T1D cases.

Training and validation data sets for predicting type 2 diabetes in South Asians: DIAGRAM, SAT2D and UK Biobank

Our analysis of type 2 diabetes in South Asians used European summary association statistics from the DIAGRAM data set (described above), South Asian summary statistics data from the South Asian Type 2 Diabetes (SAT2D) Consortium[31], and South Asian genotypes and phenotypes from UK Biobank (see Web Resources) as test data (row 5 of Table 1.1). The SAT2D data set consists of 5,561 South Asian type 2 diabetes cases and 14,458 South Asian controls for which we summary statistics for 2,646,472 imputed SNPs were available. The UK Biobank test data consists of 1,756 unrelated samples of South Asian ancestry (272 type 2 diabetes cases and 1,484 controls), genotyped at 608,878 SNPs after QC, with the following self-reported ethnicity distribution: 52 Bangladeshi, 1,301 Indian and 403 Pakistani. We removed one individual from each pair of relatives with relatedness greater than 20% (n=30). We performed a PCA analysis using EIGENSTRAT[28]

(see Web Resources) to identify and remove genetic outliers, but did not identify any outliers. We analyzed 208,400 SNPs present in the DIAGRAM, SAT2D and UK Biobank data sets after removing A/T and C/G SNPs to eliminate potential strand ambiguity.

Training and validation data sets for predicting height in Africans: UK Biobank and NDiaye et al.

Our analyses of height in Africans used European summary association statistics from UK Biobank (see Web Resources), African summary statistics from ref. 32 and African genotypes and phenotypes from UK Biobank (row 6 of Table 1.1). European summary statistics from UK Biobank were computed using 113,660 British samples for which height phenotypes were available with adjustment for 10 PCs[29], estimated using FastPCA[30] (see Web Resources). The ref. 32 data set consists of 20,427 samples of African ancestry with summary association statistics at 3,254,125 imputed SNPs. The UK Biobank data set consists of 1,745 unrelated samples of African ancestry, genotyped at 608,878 SNPs after QC, with the following self-reported ethnicity distribution: 743 African, 1,002 Caribbean. We removed one individual from each pair of relatives with relatedness greater than 20% (n=32). We performed a PCA analysis using EIGENSTRAT[28] (see Web Resources) to identify and remove genetic outliers, but did not identify any outliers. We restricted our analysis to 232,182 SNPs present in the UK Biobank and ref. 32 data sets after removing A/T and C/G SNPs to eliminate potential strand ambiguity.

Results

Simulations

We performed simulations using real genotypes and simulated phenotypes (row 1 of Table 1.1). We simulated continuous phenotypes under a non-infinitesimal model with 1% of markers chosen to be causal with the same effect size in all samples and SNP-heritability $h_g^2 = 0.5$ (see Methods); we report the average adjusted R^2 and standard errors over 100 simulations. We used WTCCC2[8,26] data (15,622 samples after QC; see Methods) as the European training data, and the SIGMA data[6] (8,214 samples) as the Latino training

and validation data (with 10-fold cross-validation). We simulated phenotypes using the 232,629 SNPs present in both data sets and built predictions from these SNPs excluding the causal SNPs, modeling the causal SNPs as untyped (see Methods).

Prediction accuracies (adjusted R^2) and optimal weights for the 5 main methods (EUR, LAT, LAT+ANC, EUR+LAT, EUR+LAT+ANC) are reported in Table 1.2A. In each case, the best prediction accuracy was attained using LD-pruning threshold $R^2_{LD} = 0.8$ (results using different LD-pruning thresholds are reported in S1 Table); the median value of the optimal P-value threshold PT was equal to 0.01 for EUR and 0.05 for LAT. On average, the EUR method performed only 23% better than the LAT method, despite having twice as much training data. This reflects a tradeoff between the larger training sample size for EUR and the target-matched LD patterns for LAT. EUR+LAT attained $64\% - 101\%$ relative improvements vs. EUR and LAT respectively (and used a slightly larger weight for EUR than for LAT), highlighting the advantages of incorporating multiple sources of training data. When including an ancestry predictor, EUR+LAT+ANC attained a 10% relative improvement vs. EUR+LAT ($\geq$ 80% relative improvement vs. EUR or LAT), reflecting small genetic effects of ancestry on phenotype that can arise from random genetic drift between populations at causal markers (which is better-captured by ancestry components than by SNPs used in a PRS).

For comparison purposes, we also performed simulations using training data in which Europeans were subsampled to attain a 1:1 ratio (row 2 of Table 1.1); prediction accuracies and optimal weights for the 5 main methods are reported in Table 1.2B. On average, the LAT method performed 190% better than the EUR method, again demonstrating the advantages of target-matched LD patterns. EUR+LAT attained 24%-260% relative improvements vs. LAT and EUR respectively (and used a larger weight for LAT than for EUR), again highlighting the advantages of incorporating multiple sources of training data.

Predictions using Latino effect sizes that were not adjusted for genetic ancestry (LAT_{unadj}, $EUR + LAT_{unadj}$, $EUR + LAT_{unadj} + ANC$, as compared to LAT, EUR+LAT, EUR+LAT+ANC) were much less accurate (S2 Table), as in previous work[33]; this is consistent with the fact that LATunadj predictions were dominated by genetic ancestry (ad-

Table 1.2: Accuracy of main prediction methods in simulations. We report results for A) 2:1 training sample size ratio (row 1 of Table 1.1) and B) 1:1 training sample size ratio (row 2 of Table 1.1). We report average adjusted R^2 over 100 simulations for each of the 5 main prediction methods. We also report normalized weights, defined as the mixing weight $\widehat{\alpha}_k$ (see Methods) multiplied by the standard deviation of the PRS.

A)

Model	Average weight (s.e.) associated to each predictor		Average adj. R^2(s.e.)
	European PRS	Latino PRS	
EUR	0.19449 (0.004)		0.03927 (0.002)
LAT		0.17780 (0.003)	0.03200 (0.001)
LAT+ANC		0.17613 (0.002)	0.04115 (0.002)
EUR+LAT	0.17847 (0.004)	0.15784 (0.003)	0.06441 (0.002)
EUR+LAT+ANC	0.19098 (0.004)	0.15578 (0.002)	0.07053 (0.002)

B)

Model	Average weight (s.e.) associated to each predictor		Average adj. R^2 (s.e.)
	European PRS	Latino PRS	
EUR	0.08715 (0.007)		0.01156 (0.001)
LAT		0.18239 (0.003)	0.03391 (0.001)
LAT+ANC		0.17815 (0.002)	0.04202 (0.002)
EUR+LAT	0.07494 (0.008)	0.17485 (0.002)	0.04211 (0.001)
EUR+LAT+ANC	0.09070 (0.005)	0.17464 (0.002)	0.04751 (0.002)

justed $R^2 = 0.37$; S3 Table). We also observed a modest correlation (adjusted $R^2 = 0.025$) between the EUR prediction and genetic ancestry (S3 Table), again reflecting small genetic effects of ancestry on phenotype that can arise from random genetic drift between populations at causal markers. The relative performance of the different prediction methods was similar in simulations in which phenotypes explicitly contained an ancestry term, representing environmentally-driven stratification (S4 Table).

We extrapolated the results in Table 1.2 to larger sample sizes by limiting the simulations to subsets of chromosomes, as in previous work[3] (Figure 1.1 and S5 Table). EUR+LAT+ANC was the best performing method in each of these experiments. We also performed simulations using predictions constructed using all SNPs including the causal SNPs (S1 Figure and S6 Table). In these experiments, EUR+LAT+ANC was once again the best performing method, and EUR performed much better than LAT, consistent with the larger training sample size for EUR and the fact that differential tagging of causal SNPs is of reduced importance when causal SNPs are typed.

Analyses of type 2 diabetes in Latinos

We applied the same methods to predict T2D in Latino target samples from the SIGMA T2D data set (row 3 of Table 1.1). We used publicly available European summary statistics from DIAGRAM[27] (12,171 cases and 56,862 controls; effective sample size $= 4/(1/N_{case} + 1/N_{control}) = 40,101$) as European training data and SIGMA T2D genotypes and phenotypes[6] (3,848 cases and 4,366 controls; effective sample size = 8,181) as Latino training and validation data, employing 10-fold cross-validation.

Prediction accuracies (adjusted R^2 on the liability scale[25], assuming 8% prevalence[2] and optimal weights for the 5 main methods (EUR, LAT, LAT+ANC, EUR+LAT, EUR+LAT+ANC) are reported in Table 1.3 (other prediction metrics are reported in S7 Table). In each case, the best prediction accuracy was obtained using LD-pruning threshold R^2_{LD}=0.8 (results using different LD-pruning thresholds are reported in S8 Table); the value of the optimal P-value threshold PT was equal to 0.05 for EUR and 0.2 for LAT. EUR performed only 33% better than LAT despite the much larger training sample size, again reflecting a tradeoff between sample size and target-matched LD patterns.

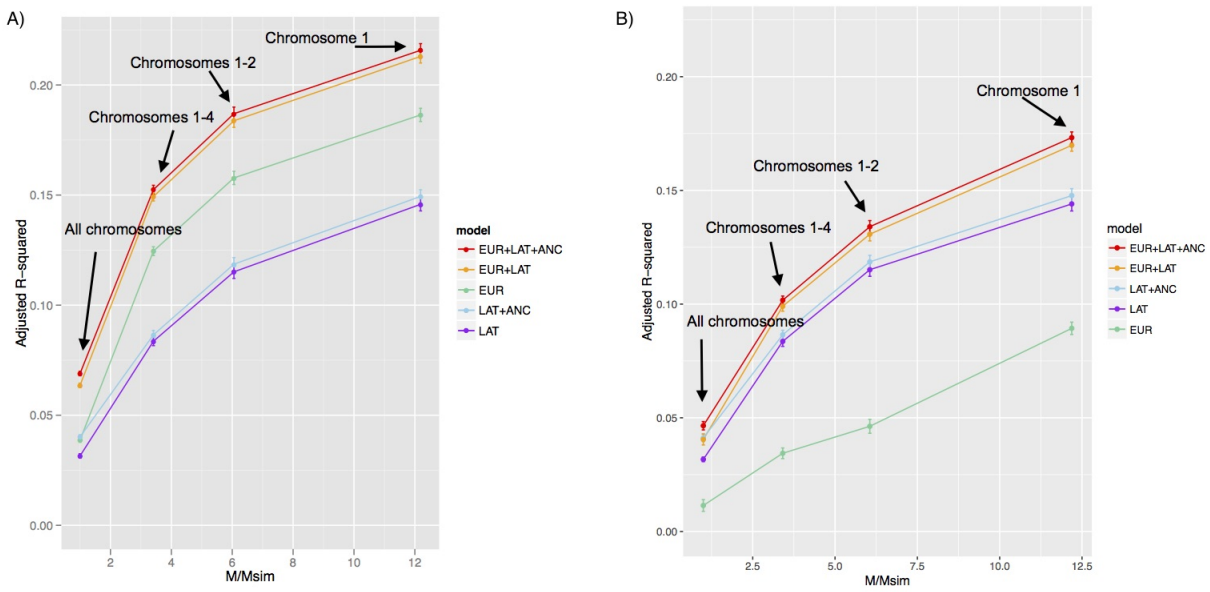


Figure 1.1: Accuracy of main prediction methods in simulations using subsets of chromosomes. We report results for A) 2:1 training sample size ratio (row 1 of Table 1.1) and B) 1:1 training sample size ratio (row 2 of Table 1.1). We report prediction accuracies for each of the 5 main prediction methods as a function of M/Msim, where M=232,629 is the total number of SNPs and Msim is the actual number of SNPS used in each simulation: 232,629 (all chromosomes), 68,188 (chromosomes 1-4), 38,412 (chromosomes 1-2), and 19,087 (chromosome 1). Numerical results are provided in S5 Table.

EUR+LAT attained 75%-133% relative improvements vs. EUR and LAT respectively (and used a slightly larger weight for EUR than for LAT), again highlighting the advantages of incorporating multiple sources of training data. We also evaluated a meta-analysis PRS (EUR-LAT-meta) and determined that EUR+LAT attained a 19% relative improvement vs. EUR-LAT-meta (Table 1.3; also see S2 Figure), highlighting the advantages of optimizing mixing weights distinct from meta-analysis weights. Although adding an ancestry predictor to LAT produced a substantial improvement (LAT+ANC vs. LAT), adding an ancestry predictor to EUR+LAT produced an insignificant change in accuracy for EUR+LAT+ANC compared to EUR+LAT; this can be explained by the large negative correlation between the European PRS (EUR) and the proportion of European ancestry within Latino samples ($R = -0.75$; S9 Table), such that any predictor that includes EUR already includes effects of genetic ancestry. This correlation is far larger than analogous correlations due to random genetic drift in our simulations (S3 Table), suggesting that

this systematically lower load of T2D risk alleles in Latino individuals with more European ancestry could be due to polygenic selection[34,35] in ancestral European and/or Native American populations; previous studies using top GWAS-associated SNPs have also reported continental differences in genetic risk for T2D[36,37]. We observed a similar correlation ($R = -0.77$) when using British UK Biobank type 2 diabetes samples as European training data (row 4 of Table 1.1; see Methods), confirming that this negative correlation is not caused by population stratification in DIAGRAM. As in our simulations, predictions using Latino effect sizes that were not adjusted for genetic ancestry (LAT_{unadj}, $EUR + LAT_{unadj}$, $EUR + LAT_{unadj} + ANC$, as compared to LAT, EUR+LAT, EUR+LAT+ANC) were much less accurate (S10 Table), consistent with the fact that these predictions were dominated by genetic ancestry (S9 Table). We also computed predictions for each method using imputed SNPs from the SIGMA T2D data set; this did not improve prediction accuracy, but predicting using two training populations still achieved the highest accuracy (S11 Table).

Table 1.3: Accuracy of main prediction methods in analyses of type 2 diabetes in a Latino cohort. We report adjusted R^2 on the liability scale for each of the 5 main prediction methods, as well as EUR-LAT-meta. We obtained similar relative results using Nagelkerke R^2, R^2 on the observed scale and AUC (S7 Table). P-values are from likelihood ratio tests comparing models EUR and LAT to the null model, model LAT+ANC to LAT, model EUR+LAT to EUR, and EUR+LAT+ANC to EUR+LAT. For the EUR model we used R^2_{LD} =0.8 and P_T =0.05, for LAT we used R^2_{LD} =0.8 and P_T =0.2, and for EUR-LAT-meta we used R^2_{LD} =0.8 and P_T =1. We also report normalized weights, defined as the mixing weight $\widehat{\alpha}_k$ (see Methods) multiplied by the standard deviation of the PRS.

Model	Weights associated to each predictor		Average adj. R^2(s.e.)	P-value for improvement over simpler model
	European PRS	Latino PRS		
EUR	0.1649		0.027	$< 10^{-49}$
LAT		0.14332	0.0203	$< 10^{-37}$
LAT+ANC		0.14623	0.03362	$< 10^{-24}$
EUR+LAT	0.16344	0.14164	0.04735	$< 10^{-37}$
EUR+LAT+ANC	0.17629	0.14108	0.04736	0.3
EUR-LAT-meta	0.16404	0.03012	0.0377	NA

We investigated how the prediction accuracy of each method varied as a function of P-

value thresholds, by varying either the EUR P-value threshold (Figure 1.2A and S12A Table) or the LAT P-value threshold (Figure 1.2B and S12B Table) between 10^{-8} and 1. In both cases, permissive P-value thresholds performed best, reflecting the relatively small sample sizes analyzed. However, the prediction accuracy of EUR+LAT+ANC was relatively stable, with prediction adjusted $R^2 > 0.037$ across all EUR P-value thresholds (Figure 1.2A) and adjusted $R^2 > 0.033$ across all LAT P-value thresholds (Figure 1.2B). In Figure 1.2A, we observe that as the EUR P-value threshold becomes more stringent, the difference in prediction accuracy between EUR+LAT+ANC and EUR+LAT increases, because EUR is less able to capture polygenic ancestry effects (see above).

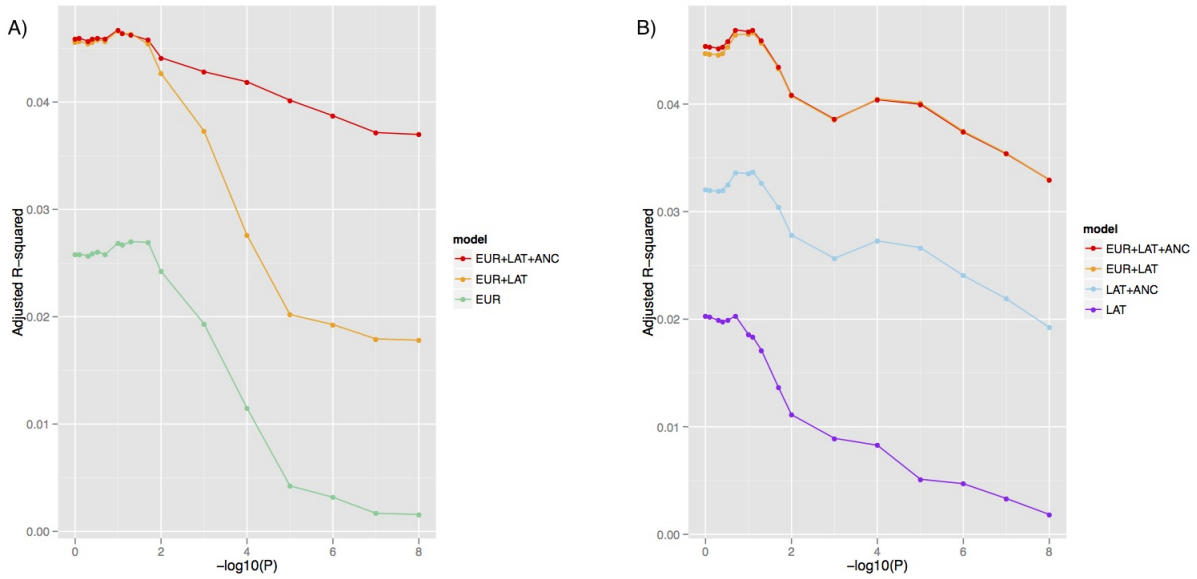


Figure 1.2: Accuracy of main prediction methods in analyses of type 2 diabetes in a Latino cohort as a function of P-value thresholds. We report prediction accuracies for each of the 5 main prediction methods as a function of (A) EUR P-value threshold, where applicable (with optimized LAT P-value threshold, where applicable) and (B) LAT P-value threshold, where applicable (with optimized EUR P-value threshold, where applicable). Numerical results are provided in S12a Table and S12b Table.

In the above results (Table 1.3 and Figure 1.2), we allowed each prediction method to optimize its mixing weights via an in-sample fit in the target sample. This procedure could in principle be susceptible to overfitting[38,39]. We did not expect overfitting to be a concern given the small number of mixing weights optimized (at most 3) relative to

the target sample size (8,181) and given our use of adjusted R^2 as the evaluation metric, but to verify this expectation, we repeated our analyses using 10×9-fold cross-validation (see Methods). Methods that use two training populations remained much more accurate than single ancestry methods, as prediction accuracy decreased only very slightly (2-4% relative decrease vs. Table 1.3) for each method (S13 Table). These slight decreases are expected, since mixing weights optimized within 10×9 cross-validation are slightly sub-optimal (due to reduced training data) and prediction accuracy is mildly sensitive to the choice of mixing weights (S2 Figure).

Analyses of type 2 diabetes in South Asians

We applied the same methods to predict T2D in South Asian target samples from the UK Biobank (row 5 of Table 1.1). We used publicly available European summary statistics from DIAGRAM (12,171 cases and 56,862 controls; effective sample size = 40,101) as European training data, South Asian summary statistics from SAT2D[31] (5,561 cases and 14,458 controls; effective sample size = 16,065) as South Asian training data, and UK Biobank genotypes and phenotypes (272 cases and 1,484 controls; effective sample size = 919) as South Asian validation data (see Methods).

Prediction accuracies (adjusted R^2 on the liability scale[25], assuming sample prevalence 15%) and optimal weights for the 5 main methods (EUR, SAS, SAS+ANC, SAS+LAT, EUR+SAS+ANC) are reported in Table 1.4 (other prediction metrics are reported in S14 Table). In each case, the best prediction accuracy was obtained using LD-pruning threshold $R^2_{LD} = 0.8$ (results using different LD-pruning thresholds are reported in S15 Table); the value of the optimal P-value threshold P_T was equal to 10^{-3} for EUR and 0.8 for SAS. EUR performed only 14% better than SAS despite the larger training sample size, again reflecting a tradeoff between sample size and target-matched LD patterns. EUR+SAS attained 72%-95% relative improvements vs. EUR and SAS respectively (and used a slightly larger weight for EUR than for SAS). In addition, EUR+SAS attained a 44% relative improvement vs. EUR-SAS-meta (Table 1.4), again highlighting the advantages of optimizing mixing weights distinct from meta-analysis weights. Adding an ancestry predictor to EUR+SAS produced an insignificant change in accuracy for EUR+ SAS +ANC compared

to EUR+SAS; we note a modest correlation between each prediction method and the proportion of European-related ancestry[40] within South Asian samples (see S16 Table). We repeated our analyses using stratified 10-fold cross-validation to estimate mixing weights (see Methods). We observed that methods that use two training populations continued to substantially outperform PRS using a single training population despite a decrease in prediction adjusted R^2 (vs. Table 1.4) for each method, consistent with the limited sample size for estimating mixing weights (S17 Table).

Table 1.4: Accuracy of main prediction methods in analyses of type 2 diabetes in a South Asian cohort. We report adjusted R^2 on the liability scale for each of the 5 main prediction methods, as well as EUR-SAS-meta. We obtained similar relative results using Nagelkerke R^2, R^2 on the observed scale and AUC (S14 Table). P-values are from likelihood ratio tests comparing models EUR and SAS to the null model, model SAS+ANC to SAS, model EUR+SAS to EUR, and EUR+LAT+ANC to EUR+SAS. For the EUR model we used R^2_{LD} =0.8 and $P_T = 10^{-3}$, for SAS we used R^2_{LD} =0.8 and P_T =0.8, and for EUR-SAS-meta we used R^2_{LD} =0.8 and $P_T = 10^{-3}$. We also report normalized weights, defined as the mixing weight $\widehat{\alpha}_k$ (see Methods) multiplied by the standard deviation of the PRS.

Model	Weights associated to each predictor		Average adj. R^2(s.e.)	P-value for improvement over simpler model
	European PRS	SAS PRS		
EUR	0.09001		0.01767	$< 10^{-3}$
SAS		0.08488	0.01556	$< 10^{-3}$
SAS+ANC		0.08821	0.01572	0.28
EUR+SAS	0.08309	0.07746	0.03031	$< 10^{-2}$
EUR+SAS+ANC	0.08138	0.07989	0.02968	0.46
EUR-SAS-meta	0.08695	0.00497	0.02098	NA

Analyses of height in Africans

We applied the same methods to predict height in African target samples from the UK Biobank (row 6 of Table 1.1). We used European summary statistics from UK Biobank (113,660 samples; British ancestry only) as European training data, African summary statistics from ref. 32 (20,427 samples) as African training data, and African UK Biobank genotypes and phenotypes (1,745 samples) as African validation data.

Prediction accuracies (adjusted R^2) and optimal weights for the 5 main methods (EUR,

AFR, AFR+ANC, EUR+AFR, EUR+AFR+ANC) are reported in Table 1.5. For EUR and AFR, the best prediction accuracy was obtained using $R^2_{LD} = 0.2$ and $R^2_{LD} = 0.8$ respectively, thus we used these respective values of R^2_{LD} for EUR and AFR in each PRS in all primary analyses (results using different LD thresholds are reported in S18 Table); the value of the optimal P-value threshold P_T was equal to 10^{-3} for EUR and 0.05 for AFR. EUR performed much better than AFR, consistent with the far larger training sample size. Nevertheless, EUR+AFR attained a 30% improvement vs. EUR (using a larger weight for EUR than for AFR). EUR+AFR also attained a small relative improvement (7%) vs. EUR-AFR-meta (Table 1.5). Adding an ancestry predictor to EUR+AFR produced an insignificant change in accuracy for EUR+AFR+ANC compared to EUR+AFR; we note a modest correlation between each prediction method and the proportion of European-related ancestry[40] within African samples (see S19 Table). We repeated our analyses using stratified 10-fold cross-validation to estimate mixing weights (see Methods). We observed that methods that use two training populations continued to substantially outperform PRS using a single training population despite a decrease in prediction adjusted R^2 (vs. Table 1.5) for each method, consistent with the limited sample size for estimating mixing weights (S20 Table).

Table 1.5: We report adjusted R^2 on the observed scale for each of the 5 main prediction methods, as well as EUR-AFR-meta. P-values are from likelihood ratio tests comparing models EUR and AFR to the null model, model AFR+ANC to AFR, model EUR+AFR to EUR, and EUR+LAT+ANC to EUR+AFR. For the EUR model we used R^2_{LD} =0.2 and $P_T = 10^{-3}$, for AFR we used R^2_{LD} =0.8 and P_T =0.05 and for EUR-AFR-meta we used R^2_{LD} =0.2 and $P_T = 10^{-6}$. We also report normalized weights, defined as the mixing weight (see Methods) multiplied by the standard deviation of the PRS.

Model	Weights associated to each predictor		Average adj. R^2(s.e.)	P-value for improvement over simpler model
	European PRS	AFR PRS		
EUR	0.164		0.02618	$< 10^{-11}$
AFR		0.106	0.01074	$< 10^{-5}$
AFR+ANC		0.124	0.01331	0.01
EUR+AFR	0.155	0.092	0.03397	$< 10^{-3}$
EUR+AFR+ANC	0.15	0.102	0.03443	0.17
EUR-AFR-meta	0.15064	0.02707	0.03158	NA

Discussion

We have shown that combining training data from European samples and training data from the target population attains a $> 70\%$ relative improvement in prediction accuracy for type 2 diabetes in both Latino and South Asian cohorts compared to prediction methods that use training data from a single population. In addition, this approach attains 30% relative improvement in prediction accuracy for height in an African cohort. These relative improvements are robust to overfitting, consistent with simulations and reduce the documented gap in risk prediction accuracy between European and non-European target populations[1,3,22,23,41,42]; we note that there are at least 35 phenotypes for which there are published GWAS data sets in Europeans and at least one non-European population (with minimum sample size of 8,000) that are listed in the NHGRI-EBI GWAS Catalog[43], where our approach could potentially be valuable (S21 Table). Intuitively, our approach leverages both large training sample sizes and training data with target-matched LD patterns. We note that the effects of differential tagging (or different causal effect sizes) in different populations can potentially be quantified using cross-population genetic correlation[44–46], and that leveraging data from a different population to improve predictions is a natural analogue to leveraging data from a correlated trait[16].

Despite these advantages, our work is subject to limitations and leaves several questions open for future exploration. First, although we have demonstrated large relative improvements in prediction accuracy, absolute prediction accuracies are currently not large enough to achieve clinical utility, which will require larger sample sizes[4,5]; our simulations suggest that multi-ethnic polygenic risk scores will continue to produce improvements at larger sample sizes (Figure 1.1). Second, while our focus here was on prediction without using individual-level training data, when such data is available it may be possible to attain higher prediction accuracy using methods that fit all markers simultaneously, such as Best Linear Unbiased Predictor (BLUP) methods and their extensions[14–21]. Third, our LDpred risk prediction method[3], which analyzes summary statistics in conjunction with LD information from a reference panel, is more accurate in European populations than the informed LD-pruning + P-value thresholding approach employed here; we did

not employ LDpred due to the complexities of admixture-LD in analyses of admixed populations that explicitly model LD[47], but extending LDpred to handle these complexities could further improve accuracy. Fourth, we note that in our application to real phenotypes adding an ancestry predictor produced insignificant changes in prediction accuracy, primarily because ancestry effects are captured by the polygenic risk scores; adding an ancestry predictor only improves prediction when we use a stringent P-value threshold to build the polygenic risk score (Figure 1.2). Fifth, we have not considered here how to improve prediction accuracy in data sets with related individuals[19]. Sixth, we did not incorporate local ancestry, which could potentially improve prediction accuracy in admixed populations[48]. Seventh, we did not incorporate data from the X chromosome, which is likely to harbor additional heritability that could improve prediction accuracy[49]. Finally, we focused our analyses on common variants, but future work may wish to consider rare variants as well.

Web Resources

PLINK: `https://www.cog-genomics.org/plink2`

WTCCC2 data set: `http://www.wtccc.org.uk/ccc2`

SIGMA data set: `http://www.type2diabetesgenetics.org`

DIAGRAM summary association statistics: `http://www.diagram-consortium/org/`

UK Biobank data set: `https://www.ukbiobank.ac.uk`

FastPCA (EIGENSOFT version 6.1.4): `http://www.hsph.harvard.edu/alkes-price/software/`

EIGENSTRAT (EIGENSOFT version 6.0.1): `http://www.hsph.harvard.edu/alkes-price/software/`

Acknowledgements

We are grateful to B. Vilhjalmsson and L. Liang for helpful discussions. We are grateful to G. Lettre for assistance with data from ref. 32. This research has been conducted using the

UK Biobank Resource (Application Number: 16549). This research was funded by NIH grant R01 GM105857 (A.L.P.).

Consortia

South Asian Type 2 Diabetes (SAT2D) Consortium. Jaspal S Kooner, Danish Saleheen, Xueling Sim, Joban Sehmi, Weihua Zhang, Philippe Frossard, Latonya F Been, Kee-Seng Chia, Antigone S Dimas, Neelam Hassanali, Tazeen Jafar, Jeremy BM Jowett, Xinzhing Li, Venkatesan Radha, Simon D Rees, Fumihiko Takeuchi, Robin Young, Tin Aung, Abdul Basit, Manickam Chidambaram, Debashish Das, Elin Grunberg, Asa K Hedman, Zafar I Hydrie, Muhammed Islam, Chiea-Chuen Khor, Sudhir Kowlessur, Malene M Kristensen, Samuel Liju, Wei-Yen Lim, David R Matthews, Jianjun Liu, Andrew P Morris, Alexandra C Nica, Janani M Pinidiyapathirage, Inga Prokopenko, Asif Rasheed, Maria Samuel, Nabi Shah, A Samad Shera, Kerrin S Small, Chen Suo, Ananda R Wickremasinghe, Tien Yin Wong, Mingyu Yang, Fan Zhang, DIAGRAM, MuTHER, Goncalo R Abecasis, Anthony H Barnett, Mark Caulfield, Panos Deloukas, Tim Frayling, Philippe Froguel, Norihiro Kato, Prasad Katulanda, M Ann Kelly, Junbin Liang, Viswanathan Mohan, Dharambir K Sanghera, James Scott, Mark Seielstad, Paul Z Zimmet, Paul Elliott, Yik Ying Teo, Mark I McCarthy, John Danesh, E Shyong Tai, and John C Chambers

The SIGMA Type 2 Diabetes Consortium. Amy L. Williams, Suzanne B. R. Jacobs, Hortensia Moreno-Macas, Alicia Huerta-Chagoya, Claire Churchouse, Carla Mrquez-Luna, Humberto Garca-Ortz, Mara Jos Gmez-Vzquez, Stephan Ripke, Alisa K. Manning, Benjamin Neale, David Reich, Daniel O. Stram, Juan Carlos Fernndez-Lpez, Nick Patterson, Suzanne B. R. Jacobs, Claire Churchhouse, Shuba Gopal, James A. Grammatikos, Ian C. Smith, Kevin H. Bullock, Amy A. Deik, Amanda L. Souza, Kerry A. Pierce, Clary B. Clish, Anglica Martnez-Hernndez, Francisco Barajas-Olmos, Federico Centeno-Cruz, Elvia Mendoza-Caamal, Cecilia Contreras-Cubas, Cristina Revilla-Monsalve, Sergio Islas-Andrade, Emilio Crdova, Xavier Sobern, Mara Elena Gonzlez-Villalpando, Brian E. Henderson, Kristine Monroe, Lynne Wilkens, Laurence N. Kolonel, and Loic Le Marchand,

Laura Riba, Mara Luisa Ordez-Snchez, Rosario Rodrguez-Guilln, Ivette Cruz-Bautista, Maribel Rodrguez-Torres, Linda Liliana Muoz-Hernndez, Donaj Gmez, Ulises Alvirde, Olimpia Arellano, Robert C. Onofrio, Wendy M. Brodeur, Diane Gage, Jacquelyn Murphy, Jennifer Franklin, Scott Mahan, Kristin Ardlie, Andrew T. Crenshaw, Wendy Winckler, Maria L. Cortes, Nol P. Burtt, Carlos A. Aguilar-Salinas, Clicerio Gonzlez-Villalpando, Jose C. Florez, Lorena Orozco, Christopher A. Haiman, Teresa Tusi-Luna, David Altshuler

Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets

Carla Márquez-Luna[1], Steven Gazal[2,3], Po-Ru Loh[3,4], Nicholas Furlotte[5], Adam Auton[5], 23andMe Research Team[5], Alkes L. Price[1,2,3]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA.
[2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.
[3]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA.
[4]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.
[5]23andMe Inc., Mountain View, CA, USA.

Abstract

Genetic variants in functional regions of the genome are enriched for complex trait heritability. Here, we introduce a new method for polygenic prediction, LDpred-funct, that leverages trait-specific functional enrichments to increase prediction accuracy. We fit priors using the recently developed baseline-LD model, which includes coding, conserved, regulatory and LD-related annotations. We analytically estimate posterior mean causal effect sizes and then use cross-validation to regularize these estimates, improving prediction accuracy for sparse architectures. LDpred-funct attained higher prediction accuracy than other polygenic prediction methods in simulations using real genotypes. We applied LDpred-funct to predict 16 highly heritable traits in the UK Biobank. We used association statistics from British-ancestry samples as training data (avg N=365K) and samples of other European ancestries as validation data (avg N=22K), to minimize confounding. LDpred-funct attained a +27% relative improvement in prediction accuracy (avg prediction R^2=0.173; highest R^2=0.417 for height) compared to existing methods that do not incorporate functional information, consistent with simulations. For height, meta-analyzing training data from UK Biobank and 23andMe cohorts (total N=1107K; higher heritability in UK Biobank cohort) increased prediction R^2 to 0.429. Our results show that modeling functional enrichment substantially improves polygenic prediction accuracy, bringing polygenic prediction of complex traits closer to clinical utility.

Introduction

Genetic variants in functional regions of the genome are enriched for complex trait heritability[50–55]. In this study, we aim to leverage functional enrichment to improve polygenic prediction[10]. Several studies have shown that incorporating prior distributions on causal effect sizes can improve prediction accuracy[3,17,18,21], compared to standard Best Linear Unbiased Prediction (BLUP) or Pruning+Thresholding methods[1,2,56]. Recent efforts to incorporate functional information have produced promising results[13,57], but may be limited by dichotomizing between functional and non-functional variants[13] or restricting their

analyses to genotyped variants[57].

Here, we introduce a new method, LDpred-funct, for leveraging trait-specific functional enrichments to increase polygenic prediction accuracy. We fit functional priors using our recently developed baseline-LD model[7], which includes coding, conserved, regulatory and LD-related annotations. LDpred-funct first analytically estimates posterior mean causal effect sizes, accounting for functional priors and LD between variants. LDpred-funct then uses cross-validation within validation samples to regularize causal effect size estimates in bins of different magnitude, improving prediction accuracy for sparse architectures. We show that LDpred-funct attains higher polygenic prediction accuracy than other methods in simulations with real genotypes, analyses of 16 highly heritable UK Biobank traits, and meta-analyses of height using training data from UK Biobank and 23andMe cohorts.

Material and Methods

Polygenic prediction methods

We compared 5 main prediction methods: Pruning+Thresholding[1,2] (P+T), LDpred-inf[3], P+T with functionally informed LASSO shrinkage[13] (P+T-funct-LASSO), and our new the LDpred-funct-inf method, and our new LDpred-funct method. P+T and LDpred-inf are polygenic prediction methods that do not use functional annotations. P+T-funct-LASSO is a modification of P+T that corrects marginal effect sizes for winner's curse, accounting for functional annotations. LDpred-funct-inf is an improvement of LDpred-inf that incorporates functionally informed priors on causal effect sizes. LDpred-funct is an improvement of LDpred-funct-inf that uses cross-validation to regularize posterior mean causal effect size estimates, improving prediction accuracy for sparse architectures. Each method is described in greater detail below. In both simulations and analyses of real traits, we used squared correlation (R^2) between predicted phenotype and true phenotype in a held-out set of samples as our primary measure of prediction accuracy.

P+T. The P+T method builds a polygenic risk score (PRS) using a subset of independent SNPs obtained via informed LD-pruning[2] (also known as LD-clumping) followed

by P-value thresholding[1]. Specifically, the method has two parameters, R^2_{LD} and P_T, and proceeds as follows. First, the method prunes SNPs based on a pairwise threshold R^2_{LD}, removing the less significant SNP in each pair. Second, the method restricts to SNPs with an association P-value below the significance threshold P_T. Letting M be the number of SNPs remaining after LD-clumping, polygenic risk scores (PRS) are computed as

$$PRS(P_T) = \sum_{i=1}^{M} \mathbb{1}_{\{P_i < P_T\}} \tilde{\beta}_i g_i, \tag{2.1}$$

where $\tilde{\beta}_i$ are normalized marginal effect size estimates and g_i is a vector of normalized genotypes for SNP i. The parameters R^2_{LD} and P_T are commonly tuned using validation data to optimize prediction accuracy[1,2]. While in theory this procedure is susceptible to overfitting, in practice, validation sample sizes are typically large, and R^2_{LD} and P_T are selected from a small discrete set of parameter choices, so that overfitting is considered to have a negligible effect[1,2,10,58]. Accordingly, in this work, we consider $R^2_{LD} \in \{0.1, 0.2, 0.5, 0.8\}$ and $P_T \in \{1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, 3 * 10^{-4}, 10^{-4}, 3 * 10^{-5}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$, and we always report results corresponding to the best choices of these parameters. The P+T method is implemented in the PLINK software (see Web Resources).

LDpred-inf. The LDpred-inf method estimates posterior mean causal effect sizes under an infinitesimal model, accounting for LD[3]. The infinitesimal model assumes that normalized causal effect sizes have prior distribution $\beta_i \sim N(0, \sigma^2)$, where $\sigma^2 = h^2_g/M$, h^2_g is the SNP-heritability, and M is the number of SNPs. The posterior mean causal effect sizes are

$$E(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}, \mathbf{D}) = (\frac{N}{1-h^2_l} * \mathbf{D} + \frac{1}{\sigma^2}\mathbf{I})^{-1} N * \tilde{\boldsymbol{\beta}}, \tag{2.2}$$

where $\mathbf{D}$ is the LD matrix between markers, $\mathbf{I}$ is the identity matrix, N is the training sample size, $\tilde{\boldsymbol{\beta}}$ is the vector of marginal association statistics, and $h^2_l \approx kh^2/M$ is the heritability of the k SNPs in the region of LD; following ref. 3 we use the approximation $1 - h^2_l \approx 1$, which is appropriate when $M >> k$. $\mathbf{D}$ is typically estimated using validation data, restricting to non-overlapping LD windows. We determined that an LD window

size corresponding to approximately 0.15% of all (genotyped and imputed) SNPs is sufficiently large in practice. h_g^2 can be estimated from raw genotype/phenotype data[59,60] (the approach that we use here; see below), or can be estimated from summary statistics using the aggregate estimator as described in ref. 3. To approximate the normalized marginal effect size ref. 3 uses the p-values to obtain absolute Z scores and then multiplies absolute Z scores by the sign of the estimated effect size. When sample sizes are very large, p-values may be rounded to zero, in which case we approximate normalized marginal effect sizes $\widehat{\beta}_i$ by $\widehat{b}_i \frac{\sqrt{2*p_i*(1-p_i)}}{\sqrt{\sigma_Y^2}}$, where $\widehat{b_i}$ is the per-allele marginal effect size estimate, p_i is the minor allele frequency of SNP i, and σ_Y^2 is the phenotypic variance in the training data. This applies to all the methods that use normalized effect sizes.

Although the published version of LDpred-inf requires a matrix inversion (Equation 3.2), we have implemented a computational speedup that computes the posterior mean causal effect sizes by efficiently solving[61] the system of linear equations $(\frac{1}{\sigma^2}\mathbf{I}+N*\mathbf{D})E(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}},\mathbf{D}) = N\tilde{\boldsymbol{\beta}}$.

LDpred[3] is an extension of LDpred-inf that uses a point-normal prior to estimate posterior mean effect sizes via Markov Chain Monte Carlo (MCMC). In this work, we do not include LDpred in our main analyses; we determined in our secondary analyses that LDpred performs worse than LDpred-inf when applied to the UK Biobank data set that we analyze here (see Results).

P+T-funct-LASSO. Ref. 13 proposed an extension of P+T that corrects the marginal effect sizes of SNPs for winner's curse and incorporates external functional annotation data (P+T-funct-LASSO). The winner's curse correction is performed by applying a LASSO shrinkage to the marginal association statistics of the PRS:

$$PRS_{LASSO}(P_T) = \sum_{i=1}^{M} sign(\tilde{\beta}_i)||\tilde{\beta}_i| - \lambda(P_T)|\mathbb{1}_{\{P_i<P_T\}}g_i, \tag{2.3}$$

where $\lambda(P_T) = \Phi^{-1}(1-\frac{P_T}{2})sd(\tilde{\beta}_i)$, where Φ^{-1} is the inverse standard normal CDF.

Functional annotations are incorporated via two disjoint SNPs sets, representing "high-prior" SNPs (HP) and "low-prior" SNPs (LP), respectively. We define the HP SNP set for P+T-funct-LASSO as the set of SNPs in the top 10% of expected per-SNP

heritability under the baseline-LD model[7], the baseline-LD model includes coding, conserved, regulatory and LD-related annotations, whose enrichments are jointly estimated using stratified LD score regression[7,54] (see Baseline-LD model annotations section). We also performed secondary analyses using the top 5% (P+T-funct-LASSO-top5%). We define $PRS_{LASSO,HP}(P_{HP})$ to be the PRS restricted to the HP SNP set, and $PRS_{LASSO,LP}(P_{LP})$ to be the PRS restricted to the LP SNP set, where P_{HP} and P_{LP} are the optimal significance thresholds for the HP and LP SNP sets, respectively. We define $PRS_{LASSO}(P_{HP}, P_{LP}) = PRS_{LASSO,HP}(P_{HP}) + PRS_{LASSO,LP}(P_{LP})$. We also performed secondary analyses were we allow an additional regularization to the two PRS, that is: $PRS_{LASSO}(P_{HP}, P_{LP}) = \alpha_1 PRS_{LASSO,HP}(P_{HP}) + \alpha_2 PRS_{LASSO,LP}(P_{LP})$, we refer to this method as P+T-funct-LASSO-weighted.

LDpred-funct-inf. We modify LDpred-inf to incorporate functionally informed priors on causal effect sizes using the baseline-LD model[7], which includes coding, conserved, regulatory and LD-related annotations, whose enrichments are jointly estimated using stratified LD score regression[7,54]. Specifically, we assume that normalized causal effect sizes have prior distribution $\beta_i \sim N(0, c*\sigma_i^2)$, where σ_i^2 is the expected per-SNP heritability under the baseline-LD model (fit using training data only) and c is a normalizing constant such that $\sum_{i=1}^{M} \mathbb{1}_{\{\sigma_i^2>0\}} c\sigma_i^2 = h_g^2$; SNPs with $\sigma_i^2 \leq 0$ are removed, which is equivalent to setting $\sigma_i^2 = 0$. The posterior mean causal effect sizes are

$$E[\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}, \mathbf{D}, \sigma_1^2, \ldots, \sigma_{M_+}^2] = \mathbf{W}^{-1} N * \tilde{\boldsymbol{\beta}} = \left[N * \mathbf{D} + \frac{1}{c} \begin{pmatrix} \frac{1}{\sigma_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_{M_+}^2} \end{pmatrix} \right]^{-1} N * \tilde{\boldsymbol{\beta}}, \quad (2.4)$$

where M_+ is the number of SNPs with $\sigma_i^2 > 0$.

The posterior mean causal effect sizes are computed by solving the system of linear equations $\mathbf{W} E[\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}, \mathbf{D}, \sigma_1^2, \ldots, \sigma_M^2] = N * \tilde{\boldsymbol{\beta}}$. h_g^2 is estimated as described above (see LDpred-inf). $\mathbf{D}$ is estimated using validation data, restricting to windows of size $0.15\% M_+$.

LDpred-funct. We modify LDpred-funct-inf to regularize posterior mean causal effect

sizes using cross-validation. We partition the posterior mean causal effect sizes into K bins (similar to reference 62), where each bin has roughly the same sum of squared posterior mean effect sizes. Let $S = \sum_i E[\beta_i|\tilde{\beta}_i]^2$. To define each bin, we first rank the posterior mean effect sizes based on their squared values $E[\beta_i|\tilde{\beta}_i]^2$. We define bin b_1 as the smallest set of top SNPs with $\sum_{i \in b_1} E[\beta_i|\tilde{\beta}_i]^2 \geq \frac{S}{K}$, and iteratively define bin b_k as the smallest set of additional top SNPs with $\sum_{i \in b_1,\ldots,b_k} E[\beta_i|\tilde{\beta}_i]^2 \geq \frac{kS}{K}$. Let $PRS(k) = \sum_{i \in b_k} E[\beta_i|\tilde{\beta}_i]g_i$. We define

$$PRS_{LDpred-funct} = \sum_{k=1}^{K} \alpha_k PRS(k), \tag{2.5}$$

where the bin-specific weights α_k are optimized using validation data via 10-fold cross-validation. For each held-out fold in turn, we estimate the weights α_k using the samples from the other nine folds and compute PRS on the held-out fold using these weights. We then compute the average prediction R^2 across the 10 held-out folds. We set the number of bins (K) to be between 1 and 100, such that the number of samples used to estimate the K weights in each fold is $\sim$300 times larger than K:

$$K = \min(100, \lceil \frac{0.9N}{300} \rceil), \tag{2.6}$$

where N is the number of validation samples. Thus, if there are $\sim$300 validation samples or fewer, LDpred-funct reduces to the LDpred-funct-inf method. In simulations, we set K to 20 (based on 8,441 validation samples; see below), approximately concordant with Equation 2.6.

Simulations

We simulated quantitative phenotypes using real genotypes from the UK Biobank interim release (see below). We used up to 50,000 unrelated British-ancestry samples as training samples, and 8,441 samples of other European ancestries as validation samples (see below). We made these choices to minimize confounding due to shared population stratification or cryptic relatedness between training and validation samples (which, if present, could overstate the prediction accuracy that could be obtained in independent samples[39]), while preserving a large number of training samples. We restricted our simulations to 459,284 imputed SNPs on chromosome 1 (see below), fixed the number of

causal SNPs at 2,000 or 5,000 (we also performed secondary simulations with 1,000 or 10,000 causal variants), and fixed the SNP-heritability h_g^2 at 0.5. We sampled normalized causal effect sizes β_i for causal SNPs from a normal distribution with variance equal to $\frac{\sigma_i^2}{p}$, where p is the proportion of causal SNPs and σ_i^2 is the expected causal per-SNP heritability under the baseline-LD model[7], fit using stratified LD score regression (S-LDSC)[7,54] applied to height summary statistics computed from unrelated British-ancestry samples from the UK Biobank interim release (N=113,660). We computed per-allele effect sizes b_i as $b_i = \frac{\beta_i}{\sqrt{2p_i(1-p_i)}}$, where p_i is the minor allele frequency for SNP i estimated using the validation genotypes. We simulated phenotypes as $Y_j = \sum_i^M b_i g_{ij} + \epsilon_j$, where $\epsilon_j \sim N(0, 1-h_g^2)$. We set the training sample size to either 10,000, 20,000 or 50,000. The motivation to perform simulations using one chromosome is to be able to extrapolate performance at larger sample sizes[3] according to the ratio N/M, where N is the training sample size. We compared each of the five methods described above. For LDpred-funct-inf and LDpred-funct, we set baseline-LD model parameters for each functional annotation equal to the baseline-LD model parameters used to generate the data, representing a best-case scenario for LDpred-funct-inf and LDpred-funct. For LDpred-funct, we report adjusted-R^2 defined as $R^2 - (1 - R^2)\frac{K}{N-K-1}$, with N is the number of validation samples and K the number of bins.

Full UK Biobank data set

The full UK Biobank data set includes 459,327 European-ancestry samples and ∼20 million imputed SNPs[63] (after filtering as in ref. 59, excluding indels and structural variants). We selected 16 UK Biobank traits with phenotyping rate $> 80\%$ ($> 80\%$ of females for age at menarche, $> 80\%$ of males for balding), SNP-heritability $h_g^2 > 0.2$, and low correlation between traits (as described in ref. 59). We restricted training samples to 409,728 British-ancestry samples[63], including related individuals (avg N=365K phenotyped training samples; see Table S22). As in our simulations, we computed association statistics from training samples using BOLT-LMM v2.3[59]. We have made these association statistics publicly available (see Web Resources). We restricted validation samples to 25,112 samples of non-British European ancestry, after removing validation samples that were

related (> 0.05) to training samples and/or other validation samples (avg N=22K phenotyped validation samples; see Table S22). As in our simulations, we made these choices to minimize confounding due to shared population stratification or cryptic relatedness between training and validation samples (which, if present, could overstate the prediction accuracy that could be obtained in independent samples[39]), while preserving a large number of training samples. We analyzed 6,334,603 genome-wide imputed SNPs, after removing SNPs with minor allele frequency $< 1\%$, removing SNPs with imputation accuracy < 0.9, and removing A/T and C/G SNPs to eliminate potential strand ambiguity. We used h_g^2 estimates from BOLT-LMM v2.3[59] as input to LDpred-inf, LDpred-funct-inf and LDpred-funct.

UK Biobank interim release

The UK Biobank interim release includes 145,416 European-ancestry samples[64]. We used the UK Biobank interim release both in simulations using real genotypes, and in a subset of analyses of height phenotypes (to investigate how prediction accuracy varies with training sample size).

In our analyses of height phenotypes, we restricted training samples to 113,660 unrelated (≤ 0.05) British-ancestry samples for which height phenotypes were available. We computed association statistics by adjusting for 10 PCs[29], estimated using FastPCA[30] (see Web Resources). For consistency, we used the same set of 25,030 validation samples of non-British European ancestry with height phenotypes as defined above. We analyzed 5,957,957 genome-wide SNPs, after removing SNPs with minor allele frequency $< 1\%$, removing SNPs with imputation accuracy < 0.9, removing SNPs that were not present in the 23andMe height data set (see below), and removing A/T and C/G SNPs to eliminate potential strand ambiguity. We analyzed the same set of 5,957,957 SNPs both in the height meta-analysis of interim UK Biobank and 23andMe data sets and in the height meta-analysis of full UK Biobank and 23andMe data sets.

In our simulations, we restricted training samples to up to 50,000 of the 113,660 unrelated British-ancestry samples, and restricted validation samples to 8,441 samples of non-British European ancestry, after removing validation samples that were related (> 0.05) to

training samples and/or other validation samples. We restricted the 5,957,957 genome-wide SNPs (see above) to chromosome 1, yielding 459,284 SNPs after QC.

23andMe height summary statistics

The 23andMe data set consists of summary statistics computed from 698,430 European-ancestry samples (23andMe customers who consented to participate in research) at 9,898,287 imputed SNPs, after removing SNPs with minor allele frequency $< 1\%$ and that passed QC filters (which include filters on imputation quality, avg.rsq< 0.5 or min.rsq< 0.3 in any imputation batch, and imputation batch effects). Analyses were restricted to the set of individuals with $> 97\%$ European ancestry, as determined via an analysis of local ancestry[65]. Summary association statistics were computed using linear regression adjusting for age, gender, genotyping platform, and the top five principal components to account for residual population structure. The summary association statistics will be made available to qualified researchers (see Web Resources).

We analyzed 5,957,935 genome-wide SNPs, after removing SNPs with minor allele frequency $< 1\%$, removing SNPs with imputation accuracy < 0.9, removing SNPs that were not present in the full UK Biobank data set (see above), and removing A/T and C/G SNPs to eliminate potential strand ambiguity.

Meta-analysis of full UK Biobank and 23andMe height data sets

We meta-analyzed height summary statistics from the full UK Biobank and 23andMe data sets. We define

$$PRS_{meta} = \gamma_1 PRS_1 + \gamma_2 PRS_2, \tag{2.7}$$

where PRS_i is the PRS obtained using training data from cohort i. The PRS can be obtained using P+T, P+T-funct-LASSO, LDpred-inf or LDpred-funct. The meta-analysis weights γ_i can either be specified via fixed-effect meta-analysis (e.g. $\gamma_i = \frac{N_i}{\sum N_i}$) or optimized using validation data[58]. We use the latter approach, which can improve prediction accuracy (e.g. if the cohorts differ in their heritability as well as their sample size). In our primary analyses, we fit the weights γ_i in-sample and report prediction accuracy using

adjusted R^2 to account for in-sample fitting[66]. We also report results using 10-fold cross-validation: for each held-out fold in turn, we estimate the weights γ_i using the other nine folds and compute PRS on the held-out fold using these weights. We then compute the average prediction R^2 across the 10 held-out folds.

When using LDpred-funct as the prediction method, we perform the meta-analysis as follows. First, we use LDpred-funct-inf to fit meta-analysis weights γ_i. Then, we use γ_i to compute (meta-analysis) weighted posterior mean causal effect sizes (PMCES) via $PMCES = \gamma_1 PMCES_1 + \gamma_2 PMCES_2$, which are binned into k bins. Then, we estimate bin-specific weights α_k (used to compute (meta-analysis + bin-specific) weighted posterior mean causal effect sizes $\sum_{k=1}^{K} \alpha_k PMCES(k)$) using validation data via 10-fold cross validation.

Baseline-LD model annotations.

The baseline-LD model contains a broad set of 75 functional annotations (including coding, conserved, regulatory and LD-related annotations), whose enrichments are jointly estimated using stratified LD score regression[7,54]. For each trait, we used the τ_c values estimated for that trait to compute σ_i^2, the expected per-SNP heritability of SNP i under the baseline-LD model, as

$$\sigma_i^2 = \sum_c a_c(i)\tau_c, \tag{2.8}$$

where $a_c(i)$ is the value of annotation c at SNP i.

Joint effect sizes τ_c for each annotation c are estimated via

$$E[\chi_i^2] = N \sum_c \tau_c l(i,c) + 1, \tag{2.9}$$

where $l(i,c)$ is the LD score of SNP i with respect to annotation a_c and χ_i^2 is the chi-square statistic for SNP i. We note that τ_c quantifies effects that are unique to annotation c. In all analyses of real phenotypes, τ_c and σ_i^2 were estimated using training samples only.

In our primary analyses, we used 489 unrelated European samples from phase 3 of the 1000 Genomes Project[67] as the reference data set to compute LD scores, as in ref. 7.

To verify that our 1000 Genomes reference data set produces reliable LD estimates, we repeated our LDpred-funct analyses using S-LDSC with 3,567 unrelated individuals from

UK10K[68] as the reference data set (as in ref. 69), ensuring a closer ancestry match with British-ancestry UK Biobank samples. We also repeated our LDpred-funct analyses using S-LDSC with the baseline-LD+LDAK model (instead of the baseline-LD model), with UK10K as the reference data set. The baseline-LD+LDAK model (introduced in ref. 69) consists of the baseline-LD model plus one additional continuous annotation constructed using LDAK weights[70], which has values $(p_j(1-p_j))^{1+\alpha} w_j$, where $\alpha = -0.25$, p_j is the allele frequency of SNP j, and w_j is the LDAK weight of SNP j computed using UK10K data.

Results

Simulations

We performed simulations using real genotypes from the UK Biobank interim release and simulated phenotypes (see Material and Methods). We simulated continuous phenotypes with SNP-heritability $h_g^2 = 0.5$, using 476,613 imputed SNPs from chromosome 1. We selected either 2,000 or 5,000 variants to be causal; we refer to these as "sparse" and "polygenic" architectures, respectively. We sampled normalized causal effect sizes from normal distributions with variances based on expected causal per-SNP heritabilities under the baseline-LD model[7], fit using stratified LD score regression (S-LDSC)[7,54] applied to height summary statistics from British-ancestry samples from the UK Biobank interim release. We randomly selected 10,000, 20,000 or 50,000 unrelated British-ancestry samples as training samples, and we used 8,441 samples of non-British European ancestry as validation samples. By restricting simulations to chromosome 1 ($\approx 1/10$ of SNPs), we can extrapolate results to larger sample sizes ($\approx$ 10x larger; see Application to 16 UK Biobank traits), analogous to previous work[3].

We compared prediction accuracies (R^2) for five main methods: P+T[1,2], LDpred-inf[3], P+T-funct-LASSO[13], LDpred-funct-inf and LDpred-funct (see Material and Methods). Results are reported in Figure 2.1, Figure S3, Table S23 and Table S24. Among methods that do not use functional information, the prediction accuracy of LDpred-inf was similar to P+T for the sparse architecture and superior to P+T for the polygenic architecture, consistent with

previous work[3]. Incorporating functional information via LDpred-funct-inf produced a 13.6% (resp. 13.4%) relative improvement for the sparse (resp. polygenic) architecture, compared to LDpred-inf. Accounting for sparsity using LDpred-funct further improved prediction accuracy, particularly for the sparse architecture, resulting in a 24.8 % (resp. 18.8%) relative improvement, compared to LDpred-inf. LDpred-funct performed slightly better than P+T-funct-LASSO for the sparse architecture and much better than P+T-funct-LASSO for the polygenic architecture. The difference in prediction accuracy between LDpred-inf and each other method, as well as the difference in prediction accuracy between LDpred-funct and each other method, was statistically significant in most cases (see Table S24). Although LDpred-funct used K=20 posterior mean causal effect size bins to regularize effect sizes in our main simulations, results were not sensitive to this parameter (Table S25); K=50 bins consistently performed slightly better, but we did not optimize this parameter. Simulations with 1,000 or 10,000 causal variants generally recapitulated these findings, although P+T-funct-LASSO performed better than LDpred-funct for the extremely sparse architecture (Table S23). Our simulations are supportive of the potential advantages of LDpred-funct-inf and LDpred-funct. However, we caution that all of our simulations use the same model (the baseline-LD model) to simulate phenotypes and to compute predictions. Thus, our simulations should be viewed as a best case scenario for LDpred-funct-inf and LDpred-funct; a more realistic assessment of the advantages of these methods can only be obtained by analyzing real traits.

Application to 16 UK Biobank traits

We applied P+T, LDpred-inf, P+T-funct-LASSO, LDpred-funct-inf and LDpred-funct to 16 UK Biobank traits. We selected the 16 traits based on phenotyping rate $> 80\%$, SNP-heritability $h_g^2 > 0.2$, and low correlation between traits (as described in ref. 59). We analyzed training samples of British ancestry (avg N=365K; see Table S22) and validation samples of non-British European ancestry (avg N=22K). We included 6,334,603 imputed SNPs in our analyses (see Material and Methods). We computed summary statistics and h_g^2 estimates from training samples using BOLT-LMM v2.3[59] (see Table S26). We estimated trait-specific functional enrichment parameters for the baseline-LD model[7] by running S-

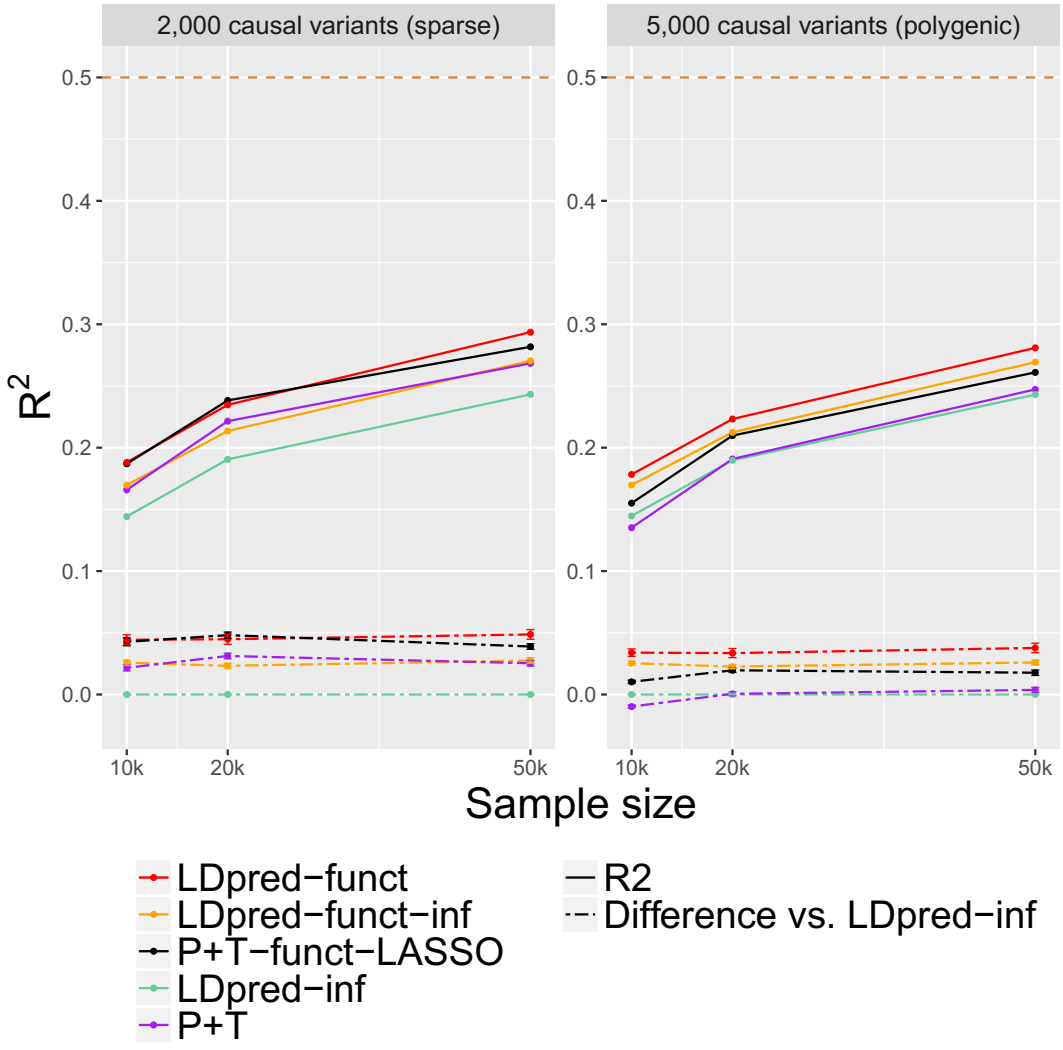


Figure 2.1: Accuracy of 5 polygenic prediction methods in simulations using UK Biobank genotypes. We report results for P+T, LDpred-inf, P+T-funct-LASSO, LDpred-funct-inf and LDpred-funct in chromosome 1 simulations with 2,000 causal variants (sparse architecture) and 5,000 causal variants (polygenic architecture). Results are averaged across 100 simulations. Top dashed line denotes simulated SNP-heritability of 0.5. Bottom dashed lines denote differences vs. LDpred-inf; error bars represent 95% confidence intervals. Results for other values of the number of causal variants are reported in Figure S3, and numerical results are reported in Table S23 and Table S24.

LDSC[7,54] on these summary statistics.

Results are reported in Figure 2.2 and Table S27, Table S28 and Table S29. Among methods that do not use functional information, LDpred-inf outperformed P+T (average relative improvement: $+4\%$), consistent with simulations under a polygenic architecture. We previously developed a different method, LDpred[3], which uses a point-normal prior to

estimate posterior mean effect sizes via Markov Chain Monte Carlo (MCMC), but we determined that LDpred performs worse than LDpred-inf in UK Biobank data (Table S29).

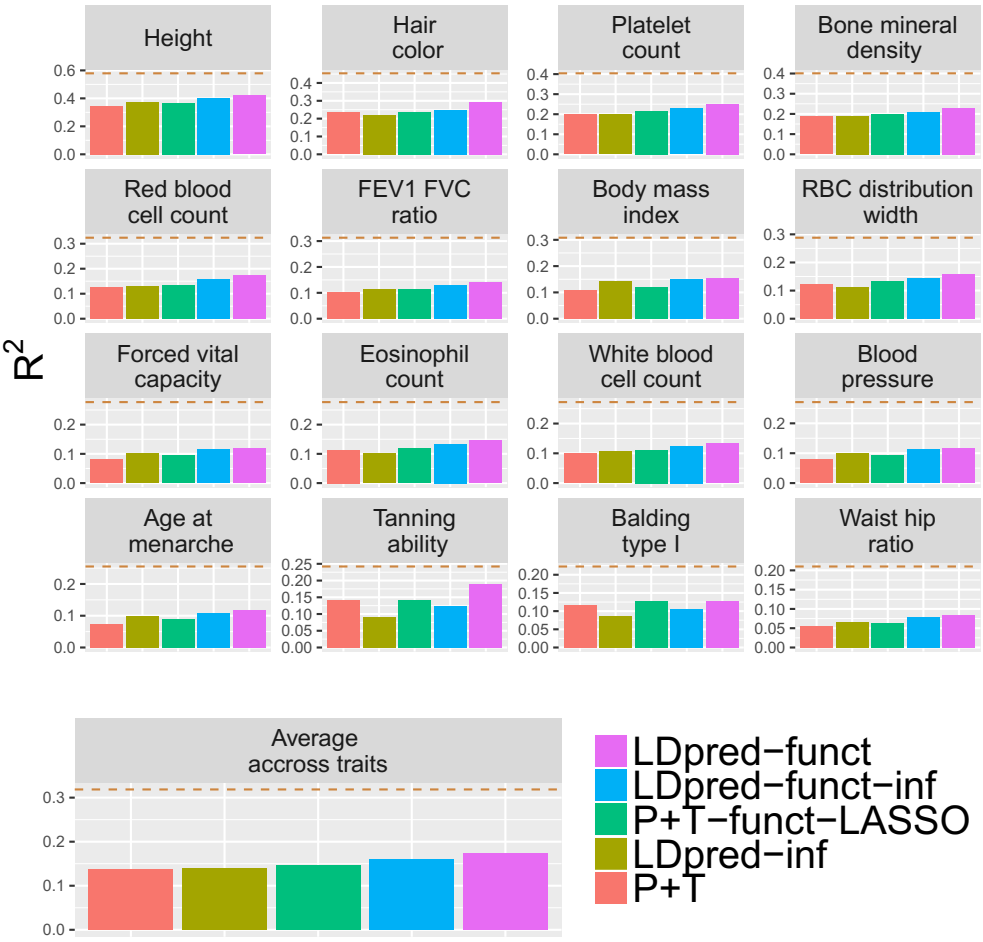


Figure 2.2: Accuracy of 5 polygenic prediction methods across 16 UK Biobank traits. We report results for P+T, LDpred-inf, P+T-funct-LASSO, LDpred-funct-inf and LDpred-funct. Dashed lines denote estimates of SNP-heritability. Numerical results are reported in Table S27 and Table S29. Jackknife s.e. for differences vs. LDpred-inf are reported in Table S28; for Average across traits, each jackknife s.e. is < 0.0009.

Incorporating functional information via LDpred-funct-inf produced a $+17\%$ average relative improvement, consistent with simulations (relative improvements ranged from $+6\%$ for body mass index to $+35\%$ for tanning ability). Accounting for sparsity using LDpred-funct further improved prediction accuracy (avg prediction R^2=0.173; highest R^2=0.417 for height), resulting in a $+27\%$ average relative improvement compared to

LDpred-inf, consistent with simulations under a polygenic architecture (relative improvements ranged from $+5\%$ for body mass index to $+104\%$ for tanning ability). LDpred-funct also performed substantially better than P+T-funct-LASSO ($+18\%$ average relative improvement), consistent with simulations under a polygenic architecture. Although LDpred-funct used an average of $K = 67$ posterior mean causal effect size bins to regularize effect sizes in these analyses (see Equation 2.6), results were not sensitive to this parameter (Table S30); K=100 bins consistently performed slightly better, but we did not optimize this parameter. In addition, although our main analyses involved very large validation sample sizes (up to 25,032; Table S22), which aids the regularization step of LDpred-funct, the bulk of the improvement of LDpred-funct vs. LDpred-funct-inf remained when restricting to smaller validation sample sizes (as low as 1,000; see Table S31). We also evaluated a modification of P+T-funct-LASSO in which different weights were allowed for the two predictors (P+T-funct-LASSO-weighted; see Material and Methods), but results were little changed $+4\%$ average relative improvement vs. P+T-funct-LASSO (see Table S29). Similar results were also obtained when defining the "high-prior" (HP) SNP set for P+T-funct-LASSO using the top 5% of SNPs with the highest per-SNP heritability, instead of the top 10% (see Table S29).

We performed several secondary analyses using LDpred-funct-inf. First, we determined that incorporating baseline-LD model functional enrichments that were meta-analyzed across traits (31 traits from ref. 7), instead of the trait-specific functional enrichments used in our primary analyses, slightly reduced prediction accuracy (Table S29). Second, we determined that using our previous baseline model[54], instead of the baseline-LD model[7], slightly reduced prediction accuracy (Table S29). Third, we determined that inferring functional enrichments using only the SNPs that passed QC filters and were used for prediction had no impact on prediction accuracy (Table S29). Fourth, we determined that using UK10K (instead of 1000 Genomes) as the LD reference panel had virtually no impact on prediction accuracy (Table S29). Additional secondary analyses are reported in the Discussion section.

Application to height in meta-analysis of UK Biobank and 23andMe cohorts

We applied P+T, LDpred-inf, P+T-funct-LASSO, LDpred-funct-inf and LDpred-funct to predict height in a meta-analysis of UK Biobank and 23andMe cohorts (see Material and Methods). Training sample sizes were equal to 408,092 for UK Biobank and 698,430 for 23andMe, for a total of 1,106,522 training samples. For comparison purposes, we also computed predictions using the UK Biobank and 23andMe training data sets individually, as well as a training data set consisting of 113,660 British-ancestry samples from the UK Biobank interim release. (The analysis using the 408,092 UK Biobank training samples was nearly identical to the analysis of Figure 2.2, except that we used a different set of 5,957,935 SNPs, for consistency throughout this set of comparisons; see Material and Methods.) We used 25,030 UK Biobank samples of non-British European ancestry as validation samples in all analyses.

Results are reported in Figure 2.3 and Table S32. The relative improvements attained by LDpred-funct-inf and LDpred-funct were broadly similar across all four training data sets (also see Figure 2.2), implying that these improvements are not specific to the UK Biobank data set. Interestingly, compared to the full UK Biobank training data set (R^2=0.416 for LDpred-funct), prediction accuracies were only slightly higher for the meta-analysis training data set (R^2=0.429 for LDpred-funct), and were lower for the 23andMe training data set (R^2=0.343 for LDpred-funct), consistent with the $\approx 30\%$ higher heritability in UK Biobank as compared to 23andMe and other large cohorts[7,59,60]; the higher heritability in UK Biobank could potentially be explained by lower environmental heterogeneity. We note that in the meta-analysis, we optimized the meta-analysis weights using validation data (similar to ref. 66), instead of performing a fixed-effect meta-analysis. This approach accounts for differences in heritability as well as sample size, and attained a $> 3\%$ relative improvement compared to fixed-effects meta-analysis (see Table S32).

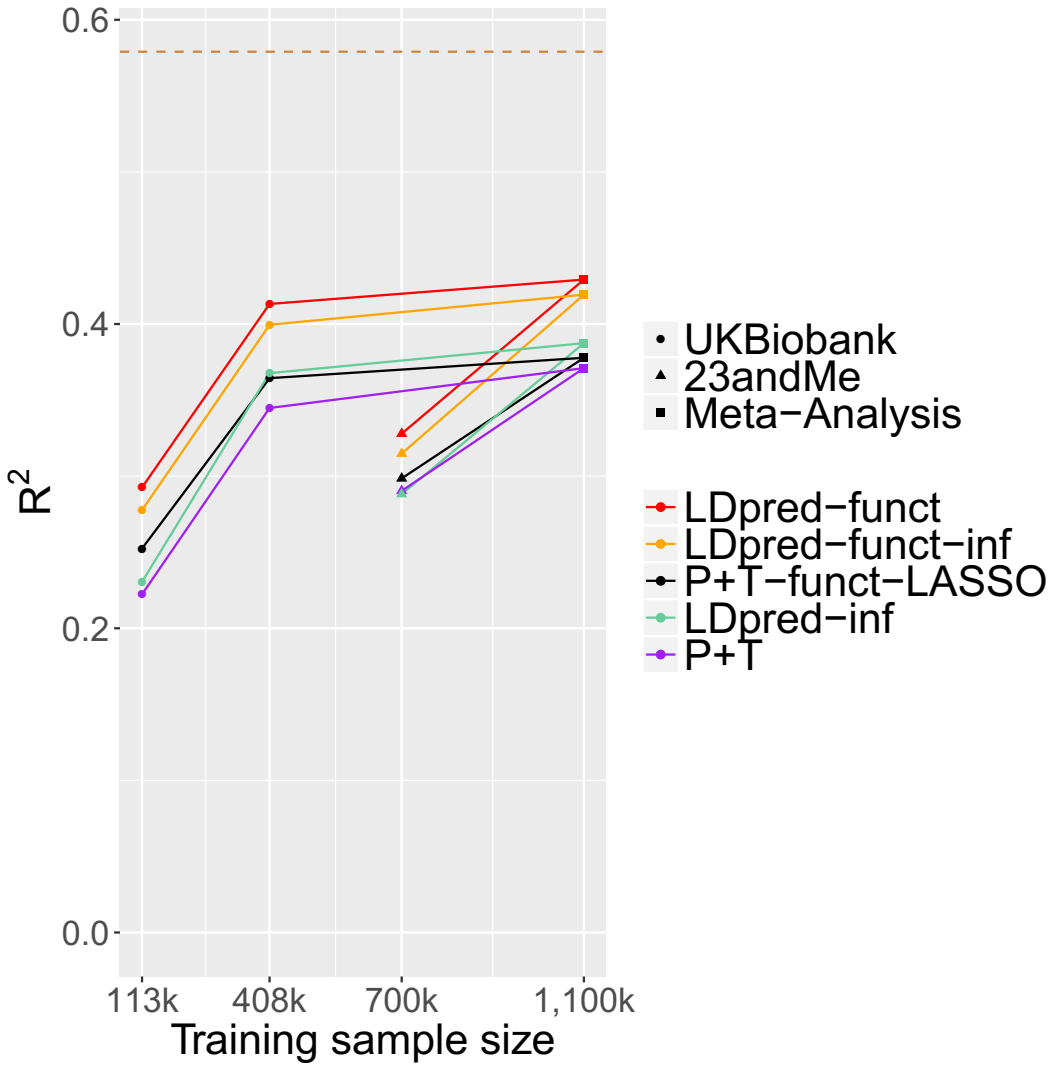


Figure 2.3: Accuracy of 5 prediction methods in height meta-analysis of UK Biobank and 23andMe cohorts. We report results for P+T, LDpred-inf, P+T-funct-LASSO, LDpred-funct-inf and LDpred-funct, for each of 4 training data sets: UK Biobank interim release (113,660 training samples), UK Biobank (408,092 training samples), 23andMe (698,430 training samples) and meta-analysis of UK Biobank and 23andMe (1,107,430 training samples). Nested training data sets are connected by solid lines. Dashed line denotes estimate of SNP-heritability in UK Biobank. Numerical results are reported in Table S32.

Discussion

We have shown that leveraging trait-specific functional enrichments inferred by S-LDSC with the baseline-LD model[7] substantially improves polygenic prediction accuracy. Across 16 UK Biobank traits, we attained a +17% average relative improvement using a method that leverages functional enrichment (LDpred-funct-inf) and a +27% average rel-

ative improvement using a method that performs an additional regularization step to account for sparsity (LDpred-funct), compared to the most accurate method tested that does not model functional enrichment (LDpred-inf).

Previous work has highlighted the potential advantages of leveraging functional enrichment to improve prediction accuracy[13,57]. We included one such method[13] (which we call P+T-funct-LASSO) in our analyses, determining that LDpred-funct attains a +18% average relative improvement vs. P+T-funct-LASSO across 16 UK Biobank traits. Another method of interest is the AnnoPred method of ref. 57, which is closely related to LDpred-funct-inf. However, ref. 57 considers only genotyped variants and binary annotations. We determined that functional enrichment information is far less useful when restricting to genotyped variants (+1% improvement for LDpred-funct-inf (typed) vs. LDpred-inf (typed); Table S29), likely because tagging variants may not belong to enriched functional annotations; also, as noted above, the additional regularization step of LDpred-funct substantially improves prediction accuracy.

Our work has several limitations. First, LDpred-funct analyzes summary statistic training data (which are publicly available for a broad set of diseases and traits[71]), but methods that use raw genotypes/phenotypes as training data have the potential to attain higher accuracy[59]; incorporating functional enrichment information into prediction methods that use raw genotypes/phenotypes as training data remains a direction for future research. Second, the regularization step employed by LDpred-funct to account for sparsity relies on heuristic cross-validation instead of inferring posterior mean causal effect sizes under a prior sparse functional model; we made this choice because the appropriate choice of sparse functional model is unclear, and because inference of posterior means via MCMC may be subject to convergence issues. As a consequence, the improvement of LDpred-funct over LDpred-funct-inf is contingent on the number of validation samples available for cross-validation; in particular, for small validation samples, the number of cross-validation bins is equal to 1 (Equation 2.6) and LDpred-funct is identical to LDpred-funct-inf. Third, we have considered only single-trait analyses, although leveraging genetic correlations among traits has considerable potential to improve prediction accuracy[16,72]. Fourth, we have not considered how to leverage functional enrichment for

polygenic prediction in related individuals[19]. Fifth, we have not investigated the application of our methods to polygenic prediction in diverse populations[66], for which very similar functional enrichments have been reported[73,74]. Finally, the improvements in prediction accuracy that we reported are a function of the baseline-LD model[7], but there are many possible ways to improve this model, e.g. by incorporating tissue-specific enrichments[50–55,75–78], modeling MAF-dependent architectures[79,80], and/or employing alternative approaches to modeling LD-dependent effects[70]; we anticipate that future improvements to the baseline-LD model will yield even larger improvements in prediction accuracy. As an initial step to explore alternative approaches to modeling LD-dependent effects, we repeated our analyses using the baseline-LD+LDAK model (introduced in ref. 69), which consists of the baseline-LD model plus one additional continuous annotation constructed using LDAK weights[70]. (Recent work has shown that incorporating LDAK weights increases polygenic prediction accuracy in analyses that do not include the baseline-LD model[81].) We determined that results were virtually unchanged (avg prediction R^2=0.1600 for baseline-LD+LDAK vs. 0.1601 for baseline-LD using LDpred-funct-inf with UK10K SNPs; see Table S29 and Table S33). Despite these limitations and open directions for future research, our work unequivocally demonstrates that leveraging functional enrichment using the baseline-LD model substantially improves polygenic prediction accuracy.

Acknowledgements

We thank the research participants and employees of 23andMe for making this work possible. We are grateful to S. Sunyaev, S. Chun, L. O'Connor, O. Weissbrod and H. Finucane for helpful discussions. This research was conducted using the UK Biobank Resource under Application #16549 and was funded by NIH grants R01 GM105857, R01 MH101244 and U01 HG009379.

Collaborators for the 23andMe research team are: Michelle Agee, Babak Alipanahi, Robert K. Bell, Katarzyna Bryc, Sarah L. Elson, Pierre Fontanillas, David A. Hinds, Jennifer C. McCreight, Karen E. Huber, Aaron Kleinman, Nadia K. Litterman, Matthew H.

McIntyre, Joanna L. Mountain, Elizabeth S. Noblin, Carrie A.M. Northover, Steven J. Pitts, J. Fah Sathirapongsasuti, Olga V. Sazonova, Janie F. Shelton, Suyash Shringarpure, Chao Tian, Joyce Y. Tung, Vladimir Vacic, and Catherine H. Wilson.

Web Resources

Software implementing the LDpred-funct-inf and LDpred-funct methods will be released prior to publication as a publicly available, open-source software package: `https://www.hsph.harvard.edu/alkes-price/software`

LDscore regression software: `https://github.com/bulik/ldsc`

UK Biobank Resource: `http://www.ukbiobank.ac.uk/`

BOLT-LMM v2.3 software `http://data.broadinstitute.org/alkesgroup/BOLT-LMM/`

BOLT-LMM v2.3 association statistics: `https://data.broadinstitute.org/alkesgroup/UKBB/UKBB_409K/`

23andMe height association statistics: The full summary statistics for the 23andMe height GWAS will be made available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Please visit `https://research.23andme.com/collaborate/#publication` for more information and to apply to access the data.

Summary statistic based extension of mixed model association method to increase meta-analysis power

Carla Márquez-Luna[1], Alkes L. Price[1,2,3]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA.
[2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.
[3]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA.

Abstract

Meta-analysis of genome-wide summary statistics has been a succesful strategy to discover genetic risk variants. The most commonly used method is using inverse-variance weighting fixed effects meta-analysis, due to limitations of sharing individual-level data, most meta-analysis only share summary statistics. Here we introduce a summary statistic based extension of mixed model association method (Meta-LMM) that increases association power in meta-analysis. This method aims to increase power by reducing the phenotypic noise by conditioning out using a leave-one-chromose-out scheme. We use the UK Biobank dataset to construct 10 independent cohorts ($N = 33K$ each), and applied Meta-LMM to 14 UK Biobank traits. Meta-LMM substantially outperformed fixed-effects meta-analysis, with a $+15\%$ median increase in χ^2 statistics (averaged across traits), consistent with simulations. And we show that on average 20% more loci were identified with Meta-LMM compared to fixed-effects meta-analysis. Our results show that this method outperforms most commonly used mehtods for meta-analysis.

Introduction

Meta-analysis of genome-wide summary statistics is an important method for discovering genetic risk variants[82]. And has been one of the most succesful approaches to discover new disease risk loci for several complex traits[83–85]. Due to restrictions of sharing individual-level data, methods developed for meta-analysis only use summary statistics data. Typically, these studies use inverse-variance-weighting fixed effects meta-analysis. Which is a method that assumes that the true effect for each allele is the same in each data set, and weight each cohort using the appropiate weights, typically proportional to the sample size of each cohort[86].

Linear mixed model association approaches gain power by reducing phenotypic noise by conditioning out on known casual variants or using leave-one-chromosome-out scheme[8,9]. These methods maximize power by running association analysis on a residualized phenotype using BLUP[9,18,56] predictions. Here we introduce a summary statistic

based extension of mixed model association method (Meta-LMM) that increases association power in meta-analysis. This method aims to increase power by residualizing the phenotypes for each cohort using polygenic risk score predictions, and estimating for each cohort a "more powerful" set of summary statistics and further combine them using fixed-effects meta-analysis. We show that Meta-LMM attains higher association power compared to fixed effects meta-analysis in simulations with real genotypes, and in analyses of 14 highly heritable UK Biobank traits.

Methods

Meta-LMM method

The Meta-LMM method consists of 5 main steps (see Figure 3.1 and Figure S4): (1a) Compute association statistics within each cohort; (1b) Share association statistics across cohorts; (2a) Residualize phenotypes within each cohort using association statistics from other cohorts (2b) Recompute association statistics within each cohort using residualized phenotypes; (3) Meta-analyze association statistics from Step (2b) across cohorts. We assume that we have C independent cohorts, with no related samples or duplicated samples between cohorts, and individual-level data cannot be shared accross cohorts.

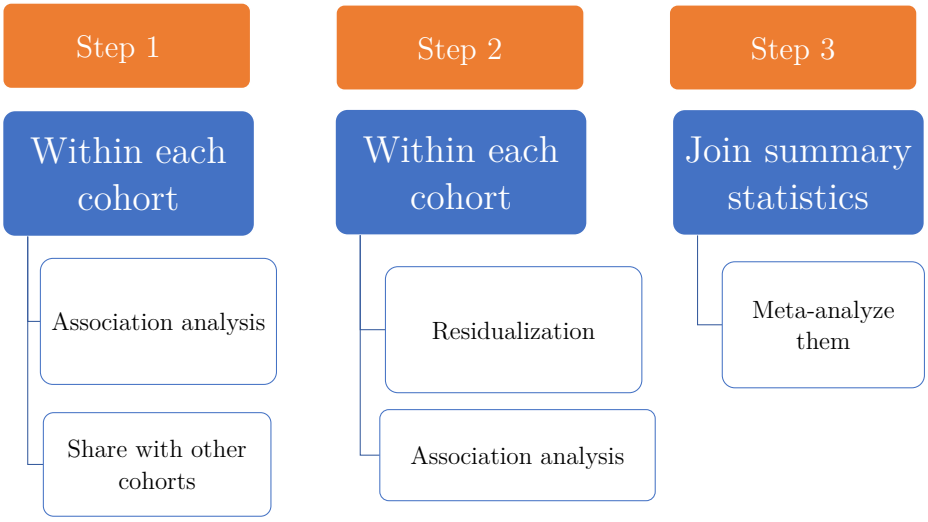


Figure 3.1: Primary steps to compute Meta-LMM summary statistics. In this figure, we show the necessary steps to get Meta-LMM summary statistics.

In step (1a), we compute mixed-model association statistics within each cohort using BOLT-LMM[9,59], an effective method for maximizing power (within each cohort) and minimizing confounding. However, other methods for computing association statistics within each cohort could also be accommodated.

Step (2a) consists of 3 steps within each cohort c:

(i) Meta-analyze association statistics from Step (1a) using fixed-effect meta-analysis, restricting to other cohorts (we exclude the association statistics from cohort c in order to prevent overfitting the phenotypes in the prediction step, which would cause true signal from the target chromosome to be removed in the residualization step; see below).

(ii) For each target chromosome chr, compute polygenic risk scores (PRS) for each individual in cohort c using association statistics from (i), restricted to other chromosomes (leave-one-chromosome-out scheme). PRS are computed using either Pruning+Thresholding[1,2] with optimal weight (Meta-LMM-P+T), LDpred-inf[3] with optimal weight (Meta-LMM-LDpred-inf), or a combined method with optimal weights (Meta-LMM). For Meta-LMM-P+T, we define the PRS via

$$PRS_{P+T} = \sum_{i=1}^{M} \mathbb{1}_{\{P_i<P_T\}} \tilde{\beta}_{i,c} g_i, \tag{3.1}$$

where $\tilde{\beta}_i$ are normalized marginal effect size estimates and g_i is a vector of normalized genotypes for SNP i and M is the total number of SNPs. The parameters R^2_{LD} and P_T are commonly tuned using validation data to optimize prediction accuracy[1,2].

For Meta-LMM-LDpred-inf, we define the PRS via

$$PRS_{LDpred-inf} = \sum_{i=1}^{M} E(\beta_{i,c}|\tilde{\beta}_c, D) g_i, \tag{3.2}$$

is the posterior mean causal effect size is defined as a function of the snp heritability h^2_g, the training sample size N, the marginal effect size $\tilde{\beta}$ and the LD matrix between markers D.

For Meta-LMM, we define the PRS via

$$PRS_{c-chr} = \widehat{\alpha}_{1,c-chr} PRS_{P+T} + \widehat{\alpha}_{2,c-chr} PRS_{LDpred-inf}, \tag{3.3}$$

In each case, optimal weights are fit in-sample using individuals from cohort c and all chromosomes. We recommend Meta-LMM as the primary PRS method, as it produces the best results (see Results), but we also provide results for Meta-LMM-P+T and Meta-LMM-LDpred-inf for completeness.

(iii) For each target chromosome chr, compute residualized phenotypes via $Y_{residual,c-chr} = Y - PRS_{c-chr}$.

In step (2b), we compute summary association statistics SS'_{meta_c} using the residualized $Y_{residual-c-chr}$ for each cohort c and target chromosome chr. We use linear regression with 20 principal components (PCs); we note that BOLT-LMM does not allow for different phenotypes for each target chromosome.

In step (3), we meta-analyze association statistics from Step (2b) using fixed-effects meta-analysis.

Fixed effects meta-analysis

Fixed effects meta-analysis is the most commonly used method to peform meta-analysis of GWAS data. It assumes that the true effect of each risk allele is the same accross datasets, and combines summary statistics by using inverse variance weighting. The fixed effects meta-analysis beta is defined as

$$\beta_{meta} = \frac{\sum_{c=1}^{C} N_c q_c * (1 - q_c)\beta_c}{\sum_{c=1}^{C} N_c q_c * (1 - q_c)}, \tag{3.4}$$

where C is the total number of cohorts, N_c is the total sample size for cohort c and q_c is the minor allele frequency associated to the SNP on cohort c. We computed fixed effect meta-analysis using Plink2 (see Web resources).

Simulations

We simulated quantitative phenotypes using real genotypes from the UK Biobank dataset (see below). We restricted our analysis to 337,538 unrelated British-ancestry samples[63]. We analyzed 616,214 genome-wide SNPs, after removing SNPs with minor allele frequency $< 1\%$. We divided the total sample into 10 different cohorts of the same sample size each ($N_c = 33K$). We sampled normalized causal effect sizes β_i for causal SNPs from

a normal distribution with variance equal to $\frac{h_g^2}{M_{causal}}$, where M_{causal} the total number of causal variants. We restricted causal variant to be only in the odd chromosomes, to facilitate null calibration analyses on even chromosomes. We fixed the proportion of causal variants to be 0.1% and 5%, and fixed the SNP-heritability h_g^2 at 0.5. We simulated 20 phenotypes as $Y_j = \sum_i^M b_i g_{ij} + \epsilon_j$, where $\epsilon_j \sim N(0, 1 - h_g^2)$, with M equat to the total number of causal variants.

We computed association statistics using 5 different methods: fixed effects meta-analysis and four different variations of our method, Meta-LMM, Meta-LMM-P+T and Meta-LMM-LDpred-inf, and Meta-LMM-True. Meta-LMM-True is a cheating method where we use the true effect sizes b_i to compute the polygenic risk score PRS_{c-chr}; this method is not applicable to real traits, but is included for comparison purposes.

Power analyses. We use 3 different metrics to assess power. First, for each method we compute the average χ^2 restricting to only the true causal variants. Second, for each method we compute the average χ^2 restricted to variants that have $\chi^2 > 30$ accross all methods. Third, we compute the average χ^2 of all the variants in the odd chromosomes.

Null calibration analyses. To asses null calibration we compute the average χ^2 of all the variants in the even chromosomes. Given that there are no causal variants in the even chromosome the average χ^2 is expected to be ~ 1.

As secondary analyses, we assessed the impact of different methods of association in Step (1a). Specifically, we used linear regression + 10 PCs and linear regression + 20 PCs as alternatives to BOLT-LMM.

UK Biobank data set

The full UK Biobank data set includes 459,327 European-ancestry samples and $824,283$ genotyped SNPs[63]. We selected 14 UK Biobank traits with phenotyping rate $> 80\%$ (excluding sex-specific traits), SNP-heritability $h_g^2 > 0.2$, and low correlation between traits (as described in ref. 59). We restricted our analysis to 337,538 unrelated British-ancestry samples[63] (avg N=321K phenotyped samples; see Table S34). We use the remainder 121,789 samples as an external discovery sample. We analyzed 616,214 genome-wide

SNPs, after removing SNPs with minor allele frequency $< 1\%$. We used h^2_g estimates from BOLT-LMM v2.3[59] as input to LDpred-inf.

We computed association statistics using 6 different methods: fixed-effect meta-analysis (Meta-Fixed), Meta-LMM, Meta-LMM-P+T, Meta-LMM-LDpred-inf, BOLT-LMM-inf and BOLT-LMM. For the first 4 methods we divided the total sample into 10 different cohorts of 33,754 samples each. For BOLT-LMM-inf and BOLT-LMM we analyzed the 337K samples together; these analyses would not be possible in the case of large meta-analyses in which raw genotypes/phenotypes cannot be shared across cohorts, but are included for comparison purposes.

Power analyses. We assessed statistical power using three different metrics. For the first metric, we take the set of SNPs that have a $\chi^2 > 30$ accross the 6 different methods (similar as in ref. 59). We compute un-informed LD pruning on these set of SNPs to obtain a set of independent variants. We used a 500kb window and r^2 threshold of 0.1 for LD pruning. We use un-informed LD pruning instead of LD-clumping (or informed LD-pruning) to avoid giving any preference to a particular method. And we report the median of ratios between χ^2 statistic estimated using method X and fixed effects meta-analysis, where method X can be any of the other 5 methods listed above. For the second metric, we use the 121,789 samples from the UK Biobank that were not included in the main sample and compute summary statistics using BOLT-LMM. We select the set of independent SNPs that have $\chi^2 > 30$ and the R^2 between any two SNPs is < 0.1 (in this case, we do use LD-clumpling). And report the median ratios between χ^2 statistic estimated using method X and fixed effects meta-analysis, as in the first metric. The third metric, we report the total number of independent genome-wide significant variants. We use PLINK LD-clumpling tool using LD computed from on of the cohorts. We used a 500kb window and r^2 threshold of 0.01 for LD clumping, and we further collapsed associated SNPs within 100kb of each other.

Null calibration analyses. As in ref 59, we used the *attenuation ratio* defined as (LDSC intercept -1) / (mean χ^2 -1) to asses calibration. We used the LDSC software to run LD score regression on each set of association statistics using the baselineLD model.

Results

Simulations

We performed simulations using real genotypes from UK Biobank and simulated phenotypes (see Materials and Methods). We simulated continuous phenotypes with SNP-heritability $h_g^2 = 0.5$, using 616,214 genome-wide SNPs. We selected either 0.1% or 5% of variants to be causal; we refer to these as "sparse" and "polygenic" architectures, respectively. We selected causal SNPs randomly from the odd chromosomes, so that the even chromosomes contain only non-causal SNPs to assess null calibration. We randomly divided 337,538 unrelated British-ancestry samples into 10 cohorts of equal size. We evaluated 4 main methods: Meta-Fixed, Meta-LMM-P+T, Meta-LMM-LDpred-inf and Meta-LMM. For comparison purposes, we also evaluated a cheating method that uses true effect sizes to residualize phenotypes (Meta-LMM-True).

We first assessed null calibration. We computed mean χ^2 statistics across SNPs on the even chromosomes, which contain only non-causal (null) SNPs. Results are reported in Figure 3.2A and Table S35. We determined that all methods are well-calibrated, as the average χ^2 statistic for null SNPs was ≈ 1.

We next assessed power to detect true associations. For each method, we computed mean χ^2 statistics across simulated causal SNPs. We compared these means across the different methods. Results are reported in Figure 3.2B and Table S36. Meta-LMM substantially outperformed Meta-Fixed in these simulations, with a $+36\%$ (resp. $+21\%$) increase in average χ^2 statistics compared to Meta-Fixed for the sparse (resp. polygenic) architecture. Among meta-analysis methods that use a single prediction method to residualize phenotypes, Meta-LMM-P+T outperformed Meta-LMM-LDpred-inf for both architectures (although we note that our simulation approach of placing all causal SNPs on odd chromosomes limits effective polygenicity, even for the polygenic architecture). The improvements in average χ^2 closely tracked the accuracy of the predictions used to residualize phenotypes (see Table S36), consistent with previous work[9,59]; as expected, Meta-LMM-True (a cheating method with prediction R^2=100%) performed best. We obtained similar results using two other metrics, average χ^2 for SNPs with $\chi^2 > 30$ across all methods and average χ^2

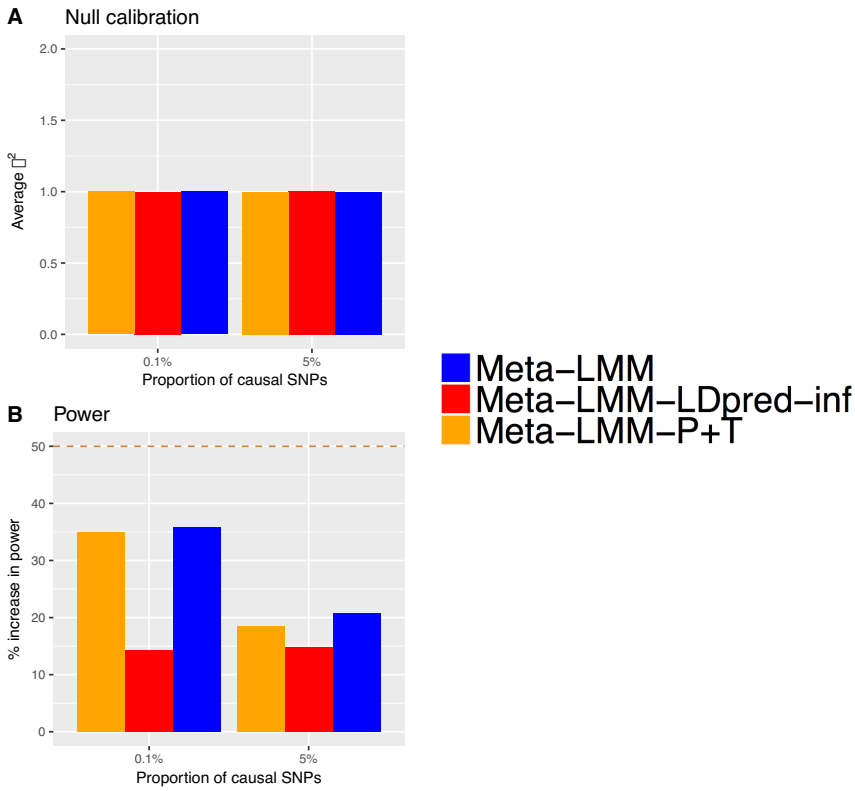


Figure 3.2: Power and calibration analyses of 5 meta-analysis methods in simulations using UK Biobank genotypes, for 2 different genetics architectures. A) Null calibration is assesed as the average χ^2 statistics restricted to SNPs in even chromosomes, with s.e. ≤ 0.002 accross diferent scenarios. Results are reported over 20 simulations. For comparison purposes we report calibration values for Meta-Fixed and Meta-LMM-True in Table S35. **B)** Percent increase in power is reported as the ratio between the average χ^2 statistics restricted to true causal SNPs in Method-X over Meta-Fixed, where Method-X can be: Meta-LMM-P+T, Meta-LMM-LDpred-inf, Meta-LMM. We also provide % improvent for Meta-LMM-True, and results are reported in Table S36. Golden dashed line represents the boost in power obtained using Meta-LMM-True. Numerical values for Figure 3.2 A) and B) are reported in Table S35 and Table S36, respectively.

for all SNPs on odd chromosomes (see Table S37).

Finally, we assessed the impact of not fully correcting for population stratification in the initial set of association statistics used to compute predictions for residualizing phenotypes (Step 1a). We determined that incomplete correlation for stratification in this step (e.g. < 10 PCs) can lead to severely inflated Meta-LMM statistics (see Table S38). We hypothesize that uncorrected population stratification in Step 1a can dominate polygenic predictions computed in Step 2a (see ref. 33), resulting in severe inflation of association statistics computed using the resulting residuals.

Application to UK Biobank traits

We analyzed 14 UK Biobank traits. We selected the 14 traits based on phenotyping rate $> 80\%$, SNP-heritability $h_g^2 > 0.2$, and low correlation between traits (as described in ref. 59). We analyzed 337,538 unrelated samples of British ancestry (avg N=321K phenotyped samples; see Table S34). We included 616,214 genotyped SNPs in our analyses (see Material and Methods). We evaluated 4 main methods: Meta-Fixed, Meta-LMM-P+T, Meta-LMM-LDpred-inf and Meta-LMM. For comparison purposes, we also evaluated two mixed model association methods, BOLT-LMM-inf and BOLT-LMM[9,59] (applied to the full set of samples), which are not applicable in settings where only summary statistics can be shared across cohorts.

We first assessed null calibration. For each method, we computed the LDSC attenuation ratio, defined as (LDSC intercept -1)/(Average $\chi^2 - 1$) (refs. 59 and 47). Results are reported in Figure S5 and Table S39. The attenuation ratios were very similar for Meta-LMM and Meta-Fixed, as well as the other methods, and were relatively small (avg. 0.091 for Meta-LMM vs. 0.089 for Meta-Fixed), confirming that Meta-LMM statistics were approximately well-calibrated.

We next assessed power to detect true associations. As our primary metric, we computed the median ratio of χ^2 statistics for each method vs. Meta-Fixed, restricted to independent SNPs with $\chi^2 > 30$ across all methods (analogous to previous work[59]). Results are reported in Figure 3.3 and Table S40. Meta-LMM substantially outperformed Meta-Fixed, with a $+15\%$ median increase in χ^2 statistics (averaged across traits); Meta-LMM outperformed Meta-Fixed for all traits except systolic blood pressure. Meta-LMM-P+T and Meta-LMM-LDpred-inf also performed well, with a $> 12\%$ improvement vs. Meta-Fixed in each case. These improvements closely tracked the accuracy of the predictions used to residualize phenotypes (Table S41), as in our simulations. Meta-LMM captured nearly all of the improvement of BOLT-LMM-inf and the bulk of the improvement of BOLT-LMM, a gold standard method that requires a merged set of raw genotypes/phenotypes. We obtained similar results when restricting χ^2 statistics to independent SNPs that were genome-wide significant in a non-overlapping discovery sample (see Methods;

Figure S6 and Table S42), with a $+15\%$ improvement for Meta-LMM vs. Meta-Fixed and an improvement for all traits. We also obtained similar results using the number of independent genome-wide significant loci (see Methods; Figure S7 and Table S43), with a $+19\%$ improvement for Meta-LMM vs. Meta-Fixed and an improvement for all traits.

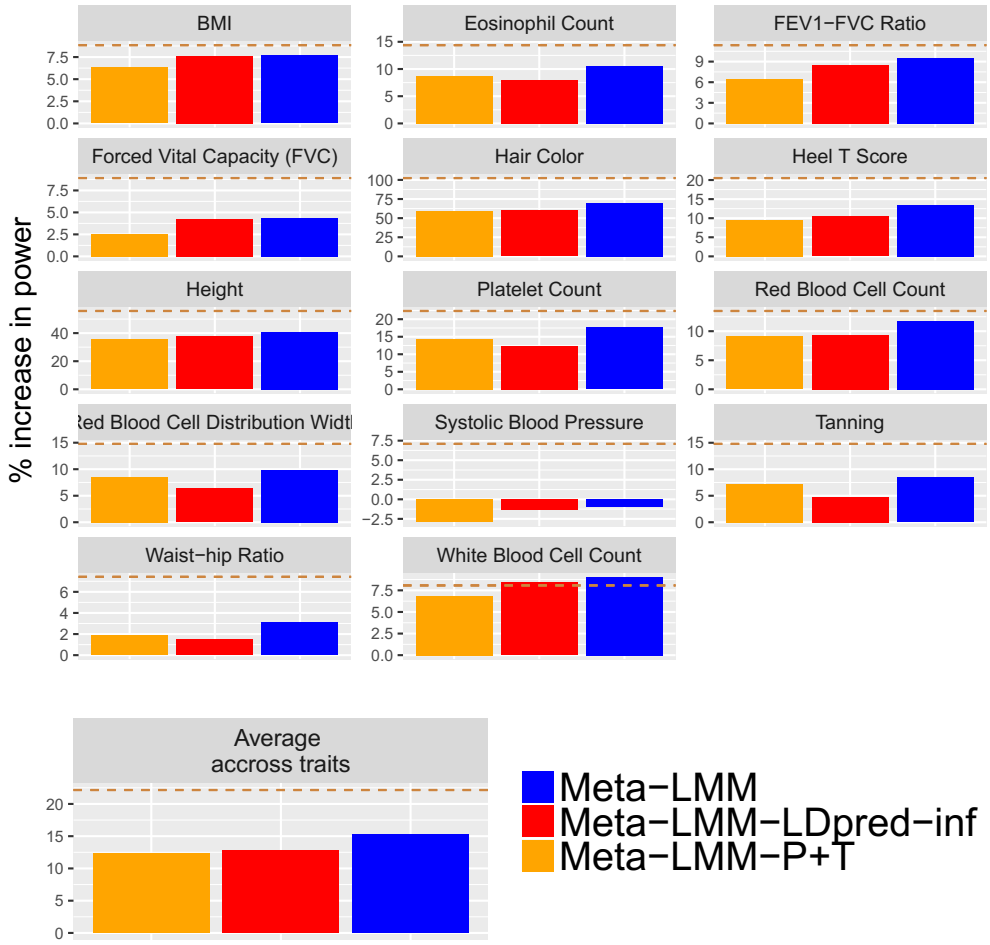


Figure 3.3: Percent improvement in power for 3 meta-analyses methods relative to fixed-effects meta-analysis when applied to 14 UK Biobank. We report the median of ratios between χ^2 statistics estimated using Method X and Meta-Fixed, where method-X can be Meta-LMM, Meta-LMM-P+T, and Meta+LMM+LDpred-inf. We also report in Table S40 analogous results using BOLT-LMM-inf and BOLT-LMM, wich respesents the best case scenenario for increasing association power. Golden dashed line represents the boost in power obtained using BOLT-LMM. We restrict calculations to SNPs that have $\chi^2 > 30$ accross all the 6 methods being compared. Numerical values are in Table S40.

Discussion

We have described a method that increases power in meta-analyses by redusing the noise the association statistics by an out-of chromosome residualization. This method is applicable in settings where only summary statistics can be shared across cohorts. We have shown both in simulations and real traits that our method increased association power over fixed-effects meta-analysis, which is the most common method for meta-analysis in genome-wide association studies. Across 14 UK Biobank traits, we attained a $+15\%$ average increase in power compared to fixed-effects meta-analysis, this improvement was validated with other two different metrics of power. Our method could be used as well to increase association power within a single cohort of moderate sample size. We could use publicly available summary statistics from the same trait or a correlated trait estimated using an independent cohort, and use them to residualize the phenotype. And then meta-analyzed the summary statistics of the two cohorts.

Although Meta-LMM increases association power compared to fixed-effects meta-analysis, it still has several limitations. First, our method assumes that the cohorts being meta-analyzed are independent between each other, which is a common assumption for most meta-analysis. If there are overlapping subjects or related individuals across cohorts we would risk to overfit the phenotype in the residualization step and loose power of association in the following step. Second, in this study we consider that all the cohorts come from the same continental population, have similar population structure and SNP heritability. We note that as long as the cohorts belong to the same continental population we do not expect a decrease in power due to the residualization step. An additional challenge would be to consider how to do meta-analysis in cohorts with different h^2_g, and weight each population accordingly. Third, another limitation is that we are not residualizing within cohort. In our analyses we consider that we have 10 cohorts of moderate sample size in which case we have a sufficiently large training data for the residualization step. For a smaller number of cohorts we expect a decrease in power due to moderate sample size used as training, one way to increase it would be to add an additional layer of cross-validation in the residualization step and incorporate within summary statistics. If we

only have two cohorts, the we running BOLT-LMMv2.3 within each cohort and further meta-analyze might be enough. Fourth, if we have sufficiently large cohorts, it is possible that it will suffice to running BOLT-LMMv2.3 within each cohort and further apply a fixed-effects meta-analysis. In this case, a posible future research direction could be to modify BOLT-LMM so it can incorporate summary statistics from other studies, or add modify BOLT-LMM so it can take different phenotypes for different chromosomes. Fifth, we did not applied our method to case-control association studies, in principle we could apply our method to analyze case-control studies; although there are some well documented pitfalls if we do not account appropiately for disease prevelance and case-control ascertainment[87,88]. Sixth, a reduced power in the residualization step will be expected if doing trans-ethnic meta-analyses, but our method offers flexibilty to use different prediction methods; although, trans-ethnic meta-analyses entails additional complexities due to the genetic heterogeneity between populations[89,90]. In principle, in presence of heterogeneity effects, we could change the fixed-effects meta-analysis for another method that accounts for heterogeneity[89]. Seventh, we limit our analysis to only genotyped variants but in principle, it is possible apply our method to imputed data in the same way as described here. For analysis of imputed variants, one option would be to use only the genotyped variants to construct the residualized phenotype and then run the association analysis using all genotyped/imputed variants (as in ref. 9). If it is the case that diferent SNP arrays are used across cohorts, then we recomend to all genotyped/imputed variants in all the required steps.

Web Resources

UK Biobank Resource: `http://www.ukbiobank.ac.uk/`

BOLT-LMM v2.3 software `http://data.broadinstitute.org/alkesgroup/BOLT-LMM/`

Plink2: `https://www.cog-genomics.org/plink/2.0/`

LDpred: `https://www.hsph.harvard.edu/alkes-price/software/`

EIGENSTRAT (EIGENSOFT version 6.0.1): `https://www.hsph.harvard.edu/`

alkes-price/software/

Bibliography

[1] International Schizophrenia Consortium, Shaun M. Purcell, Naomi R. Wray, Jennifer L. Stone, Peter M. Visscher, Michael C. O'Donovan, Patrick F. Sullivan, and Pamela Sklar. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, August 2009.

[2] Eli A. Stahl, Daniel Wegmann, Gosia Trynka, Javier Gutierrez-Achury, Ron Do, Benjamin F. Voight, Peter Kraft, Robert Chen, Henrik J. Kallberg, Fina A. S. Kurreeman, Diabetes Genetics Replication and Meta-analysis Consortium, Myocardial Infarction Genetics Consortium, Sekar Kathiresan, Cisca Wijmenga, Peter K. Gregersen, Lars Alfredsson, Katherine A. Siminovitch, Jane Worthington, Paul I. W. de Bakker, Soumya Raychaudhuri, and Robert M. Plenge. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.*, 44(5):483–489, May 2012.

[3] Bjarni J. Vilhjlmsson, Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindstrm, Stephan Ripke, Peter Kraft, Nick Patterson, and Alkes L. Price. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, 97(4):576–592, October 2015.

[4] Nilanjan Chatterjee, Bill Wheeler, Joshua Sampson, Patricia Hartge, Stephen J. Chanock, and Ju-Hyun Park. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.*, 45(4):400–405, 405e1–3, April 2013.

[5] Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.*, 9(3):e1003348, March 2013.

[6] SIGMA Type 2 Diabetes Consortium, Amy L. Williams, Suzanne B. R. Jacobs, Hortensia Moreno-Macas, Alicia Huerta-Chagoya, Claire Churchhouse, Carla Mrquez-Luna, Humberto Garca-Ortz, Mara Jos Gmez-Vzquez, Nol P. Burtt, Carlos A. Aguilar-Salinas, Clicerio Gonzlez-Villalpando, Jose C. Florez, Lorena Orozco, Christopher A. Haiman, Teresa Tusi-Luna, and David Altshuler. Sequence variants in SLC16a11 are a common risk factor for type 2 diabetes in Mexico. *Nature*, 506(7486):97–101, February 2014.

[7] Steven Gazal, Hilary K. Finucane, Nicholas A. Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, Brendan Bulik-Sullivan, Benjamin M. Neale, Alexander Gusev, and Alkes L. Price. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat Genet*, advance online publication, September 2017.

[8] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed model association methods. *Nature genetics*, 46(2):100–106, 02 2014.

[9] Po-Ru Loh, George Tucker, Brendan K. Bulik-Sullivan, Bjarni J. Vilhjlmsson, Hilary K. Finucane, Rany M. Salem, Daniel I. Chasman, Paul M. Ridker, Benjamin M. Neale, Bonnie Berger, Nick Patterson, and Alkes L. Price. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, 47(3):284–290, March 2015.

[10] Nilanjan Chatterjee, Jianxin Shi, and Montserrat Garca-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet*, 17(7):392–406, July 2016.

[11] Luigi Palla and Frank Dudbridge. A Fast Method that Uses Polygenic Scores to Estimate the Variance Explained by Genome-wide Marker Panels and the Proportion of Variants Affecting a Trait. *Am. J. Hum. Genet.*, 97(2):250–259, August 2015.

[12] Sonia Shah, Marc J. Bonder, Riccardo E. Marioni, Zhihong Zhu, Allan F. McRae,

Alexandra Zhernakova, Sarah E. Harris, Dave Liewald, Anjali K. Henders, Michael M. Mendelson, Chunyu Liu, Roby Joehanes, Liming Liang, BIOS Consortium, Daniel Levy, Nicholas G. Martin, John M. Starr, Cisca Wijmenga, Naomi R. Wray, Jian Yang, Grant W. Montgomery, Lude Franke, Ian J. Deary, and Peter M. Visscher. Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. *Am. J. Hum. Genet.*, 97(1):75–85, July 2015.

[13] Jianxin Shi, Ju-Hyun Park, Jubao Duan, Sonja T. Berndt, Winton Moy, Kai Yu, Lei Song, William Wheeler, Xing Hua, Debra Silverman, Montserrat Garcia-Closas, Chao Agnes Hsiung, Jonine D. Figueroa, Victoria K. Cortessis, Nria Malats, Margaret R. Karagas, Paolo Vineis, I.-Shou Chang, Dongxin Lin, Baosen Zhou, Adeline Seow, Keitaro Matsuo, Yun-Chul Hong, Neil E. Caporaso, Brian Wolpin, Eric Jacobs, Gloria M. Petersen, Alison P. Klein, Donghui Li, Harvey Risch, Alan R. Sanders, Li Hsu, Robert E. Schoen, Hermann Brenner, MGS (Molecular Genetics of Schizophrenia) GWAS Consortium, GECCO (The Genetics and Epidemiology of Colorectal Cancer Consortium), The GAME-ON/TRICL (Transdisciplinary Research in Cancer of the Lung) GWAS Consortium, PRACTICAL (PRostate cancer Asso-Ciation group To Investigate Cancer Associated aLterations) Consortium, PanScan Consortium, The GAME-ON/ELLIPSE Consortium, Rachael Stolzenberg-Solomon, Pablo Gejman, Qing Lan, Nathaniel Rothman, Laufey T. Amundadottir, Maria Teresa Landi, Douglas F. Levinson, Stephen J. Chanock, and Nilanjan Chatterjee. Winner's Curse Correction and Variable Thresholding Improve Performance of Polygenic Risk Modeling Based on Genome-Wide Association Study Summary-Level Data. *PLOS Genetics*, 12(12):e1006493, December 2016.

[14] Gustavo de los Campos, Daniel Gianola, and David B. Allison. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat Rev Genet*, 11(12):880–886, December 2010.

[15] David Golan and Saharon Rosset. Effective genetic-risk prediction using mixed models. *Am. J. Hum. Genet.*, 95(4):383–393, October 2014.

[16] Robert Maier, Gerhard Moser, Guo-Bo Chen, Stephan Ripke, Cross-Disorder Working Group of the Psychiatric Genomics Consortium, William Coryell, James B. Potash, William A. Scheftner, Jianxin Shi, Myrna M. Weissman, Christina M. Hultman, Mikael Landn, Douglas F. Levinson, Kenneth S. Kendler, Jordan W. Smoller, Naomi R. Wray, and S. Hong Lee. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am. J. Hum. Genet.*, 96(2):283–294, February 2015.

[17] Gerhard Moser, Sang Hong Lee, Ben J. Hayes, Michael E. Goddard, Naomi R. Wray, and Peter M. Visscher. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet.*, 11(4):e1004969, April 2015.

[18] Doug Speed and David J. Balding. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.*, 24(9):1550–1557, September 2014.

[19] George Tucker, Po-Ru Loh, Iona M. MacLeod, Ben J. Hayes, Michael E. Goddard, Bonnie Berger, and Alkes L. Price. Two-Variance-Component Model Improves Genetic Prediction in Family Datasets. *Am. J. Hum. Genet.*, 97(5):677–690, November 2015.

[20] Omer Weissbrod, Dan Geiger, and Saharon Rosset. Multikernel linear mixed models for complex phenotype prediction. *Genome Res.*, June 2016.

[21] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.*, 9(2):e1003264, 2013.

[22] Noah A. Rosenberg, Lucy Huang, Ethan M. Jewett, Zachary A. Szpiech, Ivana Jankovic, and Michael Boehnke. Genome-wide association studies in diverse populations. *Nat. Rev. Genet.*, 11(5):356–366, May 2010.

[23] Marco Scutari, Ian Mackay, and David Balding. Using Genetic Distance to Infer the Accuracy of Genomic Prediction. *PLOS Genet*, 12(9):e1006288, September 2016.

[24] Eleftheria Zeggini, Laura J. Scott, Richa Saxena, Benjamin F. Voight, Jonathan L. Marchini, Tianle Hu, Paul I. W. de Bakker, Gonalo R. Abecasis, Peter Almgren, Gitte Andersen, Kristin Ardlie, Kristina Bengtsson Bostrm, Richard N. Bergman, Lori L. Bonnycastle, Knut Borch-Johnsen, Nol P. Burtt, Hong Chen, Peter S. Chines, Mark J. Daly, Parimal Deodhar, Chia-Jen Ding, Alex S. F. Doney, William L. Duren, Katherine S. Elliott, Michael R. Erdos, Timothy M. Frayling, Rachel M. Freathy, Lauren Gianniny, Harald Grallert, Niels Grarup, Christopher J. Groves, Candace Guiducci, Torben Hansen, Christian Herder, Graham A. Hitman, Thomas E. Hughes, Bo Isomaa, Anne U. Jackson, Torben Jrgensen, Augustine Kong, Kari Kubalanza, Finny G. Kuruvilla, Johanna Kuusisto, Claudia Langenberg, Hana Lango, Torsten Lauritzen, Yun Li, Cecilia M. Lindgren, Valeriya Lyssenko, Amanda F. Marvelle, Christa Meisinger, Kristian Midthjell, Karen L. Mohlke, Mario A. Morken, Andrew D. Morris, Narisu Narisu, Peter Nilsson, Katharine R. Owen, Colin N. A. Palmer, Felicity Payne, John R. B. Perry, Elin Pettersen, Carl Platou, Inga Prokopenko, Lu Qi, Li Qin, Nigel W. Rayner, Matthew Rees, Jeffrey J. Roix, Anelli Sandbaek, Beverley Shields, Marketa Sjgren, Valgerdur Steinthorsdottir, Heather M. Stringham, Amy J. Swift, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Nicholas J. Timpson, Tiinamaija Tuomi, Jaakko Tuomilehto, Mark Walker, Richard M. Watanabe, Michael N. Weedon, Cristen J. Willer, Wellcome Trust Case Control Consortium, Thomas Illig, Kristian Hveem, Frank B. Hu, Markku Laakso, Kari Stefansson, Oluf Pedersen, Nicholas J. Wareham, Inłs Barroso, Andrew T. Hattersley, Francis S. Collins, Leif Groop, Mark I. McCarthy, Michael Boehnke, and David Altshuler. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.*, 40(5):638–645, May 2008.

[25] Sang Hong Lee, Michael E Goddard, Naomi R Wray, and Peter M Visscher. A Better Coefficient of Determination for Genetic Profile Analysis. *Genet. Epidemiol.*, 36(3):214–224, April 2012.

[26] Stephen Sawcer, Garrett Hellenthal, Matti Pirinen, Chris C.A. Spencer, Nikolaos A. Patsopoulos, Loukas Moutsianas, Alexander Dilthey, Zhan Su, Colin Free-

man, Sarah E. Hunt, Sarah Edkins, Emma Gray, David R. Booth, Simon C. Potter, An Goris, Gavin Band, Annette Bang Oturai, Amy Strange, Janna Saarela, Cline Bellenguez, Bertrand Fontaine, Matthew Gillman, Bernhard Hemmer, Rhian Gwilliam, Frauke Zipp, Alagurevathi Jayakumar, Roland Martin, Stephen Leslie, Stanley Hawkins, Eleni Giannoulatou, Sandra Dalfonso, Hannah Blackburn, Filippo Martinelli Boneschi, Jennifer Liddle, Hanne F. Harbo, Marc L. Perez, Anne Spurkland, Matthew J Waller, Marcin P. Mycko, Michelle Ricketts, Manuel Comabella, Naomi Hammond, Ingrid Kockum, Owen T. McCann, Maria Ban, Pamela Whittaker, Anu Kemppinen, Paul Weston, Clive Hawkins, Sara Widaa, John Zajicek, Serge Dronov, Neil Robertson, Suzannah J. Bumpstead, Lisa F. Barcellos, Rathi Ravindrarajah, Roby Abraham, Lars Alfredsson, Kristin Ardlie, Cristin Aubin, Amie Baker, Katharine Baker, Sergio E. Baranzini, Laura Bergamaschi, Roberto Bergamaschi, Allan Bernstein, Achim Berthele, Mike Boggild, Jonathan P. Bradfield, David Brassat, Simon A. Broadley, Dorothea Buck, Helmut Butzkueven, Ruggero Capra, William M. Carroll, Paola Cavalla, Elisabeth G. Celius, Sabine Cepok, Rosetta Chiavacci, Franoise Clerget-Darpoux, Katleen Clysters, Giancarlo Comi, Mark Cossburn, Isabelle Cournu-Rebeix, Mathew B. Cox, Wendy Cozen, Bruce A.C. Cree, Anne H. Cross, Daniele Cusi, Mark J. Daly, Emma Davis, Paul I.W. de Bakker, Marc Debouverie, Marie Beatrice Dhooghe, Katherine Dixon, Rita Dobosi, Bndicte Dubois, David Ellinghaus, Irina Elovaara, Federica Esposito, Claire Fontenille, Simon Foote, Andre Franke, Daniela Galimberti, Angelo Ghezzi, Joseph Glessner, Refujia Gomez, Olivier Gout, Colin Graham, Struan F.A. Grant, Franca Rosa Guerini, Hakon Hakonarson, Per Hall, Anders Hamsten, Hans-Peter Hartung, Rob N. Heard, Simon Heath, Jeremy Hobart, Muna Hoshi, Carmen Infante-Duarte, Gillian Ingram, Wendy Ingram, Talat Islam, Maja Jagodic, Michael Kabesch, Allan G. Kermode, Trevor J. Kilpatrick, Cecilia Kim, Norman Klopp, Keijo Koivisto, Malin Larsson, Mark Lathrop, Jeannette S. Lechner-Scott, Maurizio A. Leone, Virpi Lepp, Ulrika Liljedahl, Izaura Lima Bomfim, Robin R. Lincoln, Jenny Link, Jianjun Liu, slaug R. Lorentzen, Sara Lupoli, Fabio Macciardi, Thomas Mack, Mark Marriott, Vittorio Martinelli, Deborah Mason, Jacob L. McCauley, Frank Mentch, Inger-Lise

Mero, Tania Mihalova, Xavier Montalban, John Mottershead, Kjell-Morten Myhr, Paola Naldi, William Ollier, Alison Page, Aarno Palotie, Jean Pelletier, Laura Piccio, Trevor Pickersgill, Fredrik Piehl, Susan Pobywajlo, Hong L. Quach, Patricia P. Ramsay, Mauri Reunanen, Richard Reynolds, John D. Rioux, Mariaemma Rodegher, Sabine Roesner, Justin P. Rubio, Ina-Maria Rckert, Marco Salvetti, Erika Salvi, Adam Santaniello, Catherine A. Schaefer, Stefan Schreiber, Christian Schulze, Rodney J. Scott, Finn Sellebjerg, Krzysztof W. Selmaj, David Sexton, Ling Shen, Brigid Simms-Acuna, Sheila Skidmore, Patrick M.A. Sleiman, Cathrine Smestad, Per Soelberg Srensen, Helle Bach Sndergaard, Jim Stankovich, Richard C. Strange, Anna-Maija Sulonen, Emilie Sundqvist, Ann-Christine Syvnen, Francesca Taddeo, Bruce Taylor, Jenefer M. Blackwell, Pentti Tienari, Elvira Bramon, Ayman Tourbah, Matthew A. Brown, Ewa Tronczynska, Juan P. Casas, Niall Tubridy, Aiden Corvin, Jane Vickery, Janusz Jankowski, Pablo Villoslada, Hugh S. Markus, Kai Wang, Christopher G. Mathew, James Wason, Colin N.A. Palmer, H-Erich Wichmann, Robert Plomin, Ernest Willoughby, Anna Rautanen, Juliane Winkelmann, Michael Wittig, Richard C. Trembath, Jacqueline Yaouanq, Ananth C. Viswanathan, Haitao Zhang, Nicholas W. Wood, Rebecca Zuvich, Panos Deloukas, Cordelia Langford, Audrey Duncanson, Jorge R. Oksenberg, Margaret A. Pericak-Vance, Jonathan L. Haines, Tomas Olsson, Jan Hillert, Adrian J. Ivinson, Philip L. De Jager, Leena Peltonen, Graeme J. Stewart, David A. Hafler, Stephen L. Hauser, Gil McVean, Peter Donnelly, and Alastair Compston. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219, August 2011.

[27] Andrew P. Morris, Benjamin F. Voight, Tanya M. Teslovich, Teresa Ferreira, Ayellet V. Segr, Valgerdur Steinthorsdottir, Rona J. Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, Inga Prokopenko, Hyun Min Kang, Christian Dina, Tonu Esko, Ross M. Fraser, Stavroula Kanoni, Ashish Kumar, Vasiliki Lagou, Claudia Langenberg, Jian'an Luan, Cecilia M. Lindgren, Martina Mller-Nurasyid, Sonali Pechlivanis, N. William Rayner, Laura J. Scott, Steven Wiltshire, Loic Yengo, Leena Kinnunen, Elizabeth J. Rossin, Soumya Raychaudhuri, Andrew D. Johnson, Antigone S.

Dimas, Ruth J. F. Loos, Sailaja Vedantam, Han Chen, Jose C. Florez, Caroline Fox, Ching-Ti Liu, Denis Rybin, David J. Couper, Wen Hong L. Kao, Man Li, Marilyn C. Cornelis, Peter Kraft, Qi Sun, Rob M. van Dam, Heather M. Stringham, Peter S. Chines, Krista Fischer, Pierre Fontanillas, Oddgeir L. Holmen, Sarah E. Hunt, Anne U. Jackson, Augustine Kong, Robert Lawrence, Julia Meyer, John R. B. Perry, Carl G. P. Platou, Simon Potter, Emil Rehnberg, Neil Robertson, Suthesh Sivapalaratnam, Alena Stankov, Kathleen Stirrups, Gudmar Thorleifsson, Emmi Tikkanen, Andrew R. Wood, Peter Almgren, Mustafa Atalay, Rafn Benediktsson, Lori L. Bonnycastle, Nol Burtt, Jason Carey, Guillaume Charpentier, Andrew T. Crenshaw, Alex S. F. Doney, Mozhgan Dorkhan, Sarah Edkins, Valur Emilsson, Elodie Eury, Tom Forsen, Karl Gertow, Bruna Gigante, George B. Grant, Christopher J. Groves, Candace Guiducci, Christian Herder, Astradur B. Hreidarsson, Jennie Hui, Alan James, Anna Jonsson, Wolfgang Rathmann, Norman Klopp, Jasmina Kravic, Kaarel Krjutkov, Cordelia Langford, Karin Leander, Eero Lindholm, Stphane Lobbens, Satu Mnnist, Ghazala Mirza, Thomas W. Mhleisen, Bill Musk, Melissa Parkin, Loukianos Rallidis, Jouko Saramies, Bengt Sennblad, Sonia Shah, Gunnar Sigursson, Angela Silveira, Gerald Steinbach, Barbara Thorand, Joseph Trakalo, Fabrizio Veglia, Roman Wennauer, Wendy Winckler, Delilah Zabaneh, Harry Campbell, Cornelia van Duijn, Andre G. Uitterlinden, Albert Hofman, Eric Sijbrands, Goncalo R. Abecasis, Katharine R. Owen, Eleftheria Zeggini, Mieke D. Trip, Nita G. Forouhi, Ann-Christine Syvnen, Johan G. Eriksson, Leena Peltonen, Markus M. Nthen, Beverley Balkau, Colin N. A. Palmer, Valeriya Lyssenko, Tiinamaija Tuomi, Bo Isomaa, David J. Hunter, Lu Qi, Wellcome Trust Case Control Consortium, Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Asian Genetic Epidemiology NetworkType 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Alan R. Shuldiner, Michael Roden, Ines Barroso, Tom Wilsgaard, John Beilby, Kees Hovingh, Jackie F. Price, James F. Wilson, Rainer Rauramaa, Timo A. Lakka, Lars Lind, George Dedoussis, Inger Njlstad, Nancy L. Pedersen, Kay-Tee Khaw, Nicholas J. Wareham, Sirkka M. Keinanen-Kiukaanniemi,

Timo E. Saaristo, Eeva Korpi-Hyvlti, Juha Saltevo, Markku Laakso, Johanna Kuusisto, Andres Metspalu, Francis S. Collins, Karen L. Mohlke, Richard N. Bergman, Jaakko Tuomilehto, Bernhard O. Boehm, Christian Gieger, Kristian Hveem, Stephane Cauchi, Philippe Froguel, Damiano Baldassarre, Elena Tremoli, Steve E. Humphries, Danish Saleheen, John Danesh, Erik Ingelsson, Samuli Ripatti, Veikko Salomaa, Raimund Erbel, Karl-Heinz Jckel, Susanne Moebus, Annette Peters, Thomas Illig, Ulf de Faire, Anders Hamsten, Andrew D. Morris, Peter J. Donnelly, Timothy M. Frayling, Andrew T. Hattersley, Eric Boerwinkle, Olle Melander, Sekar Kathiresan, Peter M. Nilsson, Panos Deloukas, Unnur Thorsteinsdottir, Leif C. Groop, Kari Stefansson, Frank Hu, James S. Pankow, Jose Dupuis, James B. Meigs, David Altshuler, Michael Boehnke, Mark I. McCarthy, and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.*, 44(9):981–990, September 2012.

[28] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–909, August 2006.

[29] Kevin J. Galinsky, Po-Ru Loh, Swapan Mallick, Nick J. Patterson, and Alkes L. Price. Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure. *The American Journal of Human Genetics*, 99(5):1130–1139, November 2016.

[30] Kevin J. Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J. Patterson, and Alkes L. Price. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1b in Europe and East Asia. *The American Journal of Human Genetics*, 98(3):456–472, March 2016.

[31] Jaspal S. Kooner, Danish Saleheen, Xueling Sim, Joban Sehmi, Weihua Zhang, Philippe Frossard, Latonya F. Been, Kee-Seng Chia, Antigone S. Dimas, Neelam Hassanali, Tazeen Jafar, Jeremy B. M. Jowett, Xinzhong Li, Venkatesan Radha, Si-

mon D. Rees, Fumihiko Takeuchi, Robin Young, Tin Aung, Abdul Basit, Manickam Chidambaram, Debashish Das, Elin Grundberg, Asa K. Hedman, Zafar I. Hydrie, Muhammed Islam, Chiea-Chuen Khor, Sudhir Kowlessur, Malene M. Kristensen, Samuel Liju, Wei-Yen Lim, David R. Matthews, Jianjun Liu, Andrew P. Morris, Alexandra C. Nica, Janani M. Pinidiyapathirage, Inga Prokopenko, Asif Rasheed, Maria Samuel, Nabi Shah, A. Samad Shera, Kerrin S. Small, Chen Suo, Ananda R. Wickremasinghe, Tien Yin Wong, Mingyu Yang, Fan Zhang, DIAGRAM, MuTHER, Goncalo R. Abecasis, Anthony H. Barnett, Mark Caulfield, Panos Deloukas, Timothy M. Frayling, Philippe Froguel, Norihiro Kato, Prasad Katulanda, M. Ann Kelly, Junbin Liang, Viswanathan Mohan, Dharambir K. Sanghera, James Scott, Mark Seielstad, Paul Z. Zimmet, Paul Elliott, Yik Ying Teo, Mark I. McCarthy, John Danesh, E. Shyong Tai, and John C. Chambers. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat. Genet.*, 43(10):984–989, October 2011.

[32] Amidou N'Diaye, Gary K. Chen, Cameron D. Palmer, Bing Ge, Bamidele Tayo, Rasika A. Mathias, Jingzhong Ding, Michael A. Nalls, Adebowale Adeyemo, Vronique Adoue, Christine B. Ambrosone, Larry Atwood, Elisa V. Bandera, Lewis C. Becker, Sonja I. Berndt, Leslie Bernstein, William J. Blot, Eric Boerwinkle, Angela Britton, Graham Casey, Stephen J. Chanock, Ellen Demerath, Sandra L. Deming, W. Ryan Diver, Caroline Fox, Tamara B. Harris, Dena G. Hernandez, Jennifer J. Hu, Sue A. Ingles, Esther M. John, Craig Johnson, Brendan Keating, Rick A. Kittles, Laurence N. Kolonel, Stephen B. Kritchevsky, Loic Le Marchand, Kurt Lohman, Jiankang Liu, Robert C. Millikan, Adam Murphy, Solomon Musani, Christine Neslund-Dudas, Kari E. North, Sarah Nyante, Adesola Ogunniyi, Elaine A. Ostrander, George Papanicolaou, Sanjay Patel, Curtis A. Pettaway, Michael F. Press, Susan Redline, Jorge L. Rodriguez-Gil, Charles Rotimi, Benjamin A. Rybicki, Babatunde Salako, Pamela J. Schreiner, Lisa B. Signorello, Andrew B. Singleton, Janet L. Stanford, Alex H. Stram, Daniel O. Stram, Sara S. Strom, Bhoom Suktitipat, Michael J. Thun, John S. Witte, Lisa R. Yanek, Regina G. Ziegler, Wei Zheng, Xiaofeng Zhu, Joseph M. Zmuda,

Alan B. Zonderman, Michele K. Evans, Yongmei Liu, Diane M. Becker, Richard S. Cooper, Tomi Pastinen, Brian E. Henderson, Joel N. Hirschhorn, Guillaume Lettre, and Christopher A. Haiman. Identification, Replication, and Fine-Mapping of Loci Associated with Adult Height in Individuals of African Ancestry. *PLOS Genet*, 7(10):e1002298, October 2011.

[33] Chia-Yen Chen, Jiali Han, David J. Hunter, Peter Kraft, and Alkes L. Price. Explicit Modeling of Ancestry Improves Polygenic Risk Scores and BLUP Prediction. *Genet. Epidemiol.*, 39(6):427–438, September 2015.

[34] Matthew R. Robinson, Gibran Hemani, Carolina Medina-Gomez, Massimo Mezzavilla, Tonu Esko, Konstantin Shakhbazov, Joseph E. Powell, Anna Vinkhuyzen, Sonja I. Berndt, Stefan Gustafsson, Anne E. Justice, Bratati Kahali, Adam E. Locke, Tune H. Pers, Sailaja Vedantam, Andrew R. Wood, Wouter van Rheenen, Ole A. Andreassen, Paolo Gasparini, Andres Metspalu, Leonard H. van den Berg, Jan H. Veldink, Fernando Rivadeneira, Thomas M. Werge, Goncalo R. Abecasis, Dorret I. Boomsma, Daniel I. Chasman, Eco J. C. de Geus, Timothy M. Frayling, Joel N. Hirschhorn, Jouke Jan Hottenga, Erik Ingelsson, Ruth J. F. Loos, Patrik K. E. Magnusson, Nicholas G. Martin, Grant W. Montgomery, Kari E. North, Nancy L. Pedersen, Timothy D. Spector, Elizabeth K. Speliotes, Michael E. Goddard, Jian Yang, and Peter M. Visscher. Population genetic differentiation of height and body mass index across Europe. *Nat Genet*, 47(11):1357–1362, November 2015.

[35] Michael C. Turchin, Charleston WK Chiang, Cameron D. Palmer, Sriram Sankararaman, David Reich, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, and Joel N. Hirschhorn. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat Genet*, 44(9):1015–1019, September 2012.

[36] Rong Chen, Erik Corona, Martin Sikora, Joel T. Dudley, Alex A. Morgan, Andres Moreno-Estrada, Geoffrey B. Nilsen, David Ruau, Stephen E. Lincoln, Carlos D. Bustamante, and Atul J. Butte. Type 2 Diabetes Risk Alleles Demonstrate Extreme Di-

rectional Differentiation among Human Populations, Compared to Other Diseases. *PLOS Genet*, 8(4):e1002621, April 2012.

[37] Erik Corona, Rong Chen, Martin Sikora, Alexander A. Morgan, Chirag J. Patel, Aditya Ramesh, Carlos D. Bustamante, and Atul J. Butte. Analysis of the Genetic Basis of Disease in the Context of Worldwide Human Relationships and Migration. *PLOS Genet*, 9(5):e1003447, May 2013.

[38] Charles Kooperberg, Michael LeBlanc, and Valerie Obenchain. Risk prediction using genome-wide association studies. *Genet. Epidemiol.*, 34(7):643–652, November 2010.

[39] Naomi R. Wray, Jian Yang, Ben J. Hayes, Alkes L. Price, Michael E. Goddard, and Peter M. Visscher. Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet*, 14(7):507–515, July 2013.

[40] David Reich, Kumarasamy Thangaraj, Nick Patterson, Alkes L. Price, and Lalji Singh. Reconstructing Indian population history. *Nature*, 461(7263):489–494, September 2009.

[41] Carlos D. Bustamante, Francisco M. De La Vega, and Esteban G. Burchard. Genomics for the world. *Nature*, 475(7355):163–165, July 2011.

[42] Alice B. Popejoy and Stephanie M. Fullerton. Genomics is failing on diversity. *Nature News*, 538(7624):161, October 2016.

[43] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, Zoe May Pendlington, Danielle Welter, Tony Burdett, Lucia Hindorff, Paul Flicek, Fiona Cunningham, and Helen Parkinson. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*, 45(Database issue):D896–D901, January 2017.

[44] Brielin C. Brown, Asian Genetic Epidemiology Network-Type 2 Diabetes (AGEN-T2D) Consortium , Chun Jimmie Ye, Alkes L. Price, and Noah Zaitlen. Transethnic

genetic correlation estimates from summary statistics. *American Journal of Human Genetics*.

[45] Teresa R. de Candia, S. Hong Lee, Jian Yang, Brian L. Browning, Pablo V. Gejman, Douglas F. Levinson, Bryan J. Mowry, John K. Hewitt, Michael E. Goddard, Michael C. O'Donovan, Shaun M. Purcell, Danielle Posthuma, International Schizophrenia Consortium, Molecular Genetics of Schizophrenia Collaboration, Peter M. Visscher, Naomi R. Wray, and Matthew C. Keller. Additive genetic variation in schizophrenia risk is shared by populations of African and European descent. *Am. J. Hum. Genet.*, 93(3):463–470, September 2013.

[46] Nicholas Mancuso, Nadin Rohland, Kristin A. Rand, Arti Tandon, Alexander Allen, Dominique Quinque, Swapan Mallick, Heng Li, Alex Stram, Xin Sheng, Zsofia Kote-Jarai, Douglas F. Easton, Rosalind A. Eeles, the PRACTICAL Consortium, Loic Le Marchand, Alex Lubwama, Daniel Stram, Stephen Watya, David V. Conti, Brian Henderson, Christopher A. Haiman, Bogdan Pasaniuc, and David Reich. The contribution of rare variation to prostate cancer heritability. *Nat Genet*, 48(1):30–35, January 2016.

[47] Brendan K. Bulik-Sullivan, Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*, 47(3):291–295, March 2015.

[48] Michael F. Seldin, Bogdan Pasaniuc, and Alkes L. Price. New approaches to disease mapping in admixed populations. *Nat Rev Genet*, 12(8):523–528, August 2011.

[49] Taru Tukiainen, Matti Pirinen, Antti-Pekka Sarin, Claes Ladenvall, Johannes Kettunen, Terho Lehtimki, Marja-Liisa Lokki, Markus Perola, Juha Sinisalo, Efthymia Vlachopoulou, Johan G. Eriksson, Leif Groop, Antti Jula, Marjo-Riitta Jrvelin, Olli T. Raitakari, Veikko Salomaa, and Samuli Ripatti. Chromosome X-Wide Association

Study Identifies Loci for Fasting Insulin and Height and Evidence for Incomplete Dosage Compensation. *PLOS Genetics*, 10(2):e1004127, February 2014.

[50] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, page 1222794, September 2012.

[51] Gosia Trynka, Cynthia Sandor, Buhm Han, Han Xu, Barbara E. Stranger, X. Shirley Liu, and Soumya Raychaudhuri. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.*, 45(2):124–130, February 2013.

[52] Joseph K. Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, 94(4):559–573, April 2014.

[53] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Ra-

jagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, February 2015.

[54] Hilary K. Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, Stephan Ripke, Felix R. Day, ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, The RACI Consortium, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R. B. Perry, Yukinori Okada, Soumya Raychaudhuri, Mark J. Daly, Nick Patterson, Benjamin M. Neale, and Alkes L. Price. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*, 47(11):1228–1235, November 2015.

[55] Kyle Kai-How Farh, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J. Housley, Samantha Beik, Noam Shoresh, Holly Whitton, Russell J. H. Ryan, Alexander A. Shishkin, Meital Hatan, Marlene J. Carrasco-Alfonso, Dita Mayer, C. John Luckey, Nikolaos A. Patsopoulos, Philip L. De Jager, Vijay K. Kuchroo, Charles B. Epstein, Mark J. Daly, David A. Hafler, and Bradley E. Bernstein. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, February 2015.

[56] Charles R Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447, 1975.

[57] Yiming Hu, Qiongshi Lu, Ryan Powles, Xinwei Yao, Can Yang, Fang Fang, Xinran

Xu, and Hongyu Zhao. Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLOS Computational Biology*, 13(6):1–16, 06 2017.

[58] Carla Mrquez-Luna, Po-Ru Loh, South Asian Type 2 Diabetes (SAT2D) Consortium, The SIGMA Type 2 Diabetes Consortium, and Alkes L. Price. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.*, 41(8):811–823, December 2017.

[59] Po-Ru Loh, Gleb Kichaev, Steven Gazal, Armin P. Schoech, and Alkes L. Price. Mixed-model association for biobank-scale datasets. *Nature Genetics*, 50(7):906–908, July 2018.

[60] Tian Ge, Chia-Yen Chen, Benjamin M. Neale, Mert R. Sabuncu, and Jordan W. Smoller. Phenome-wide heritability analysis of the UK Biobank. *PLOS Genetics*, 13(4):e1006711, April 2017.

[61] Gilbert Strang. *Linear Algebra and Its Applications*. Academic Press, Inc., 2nd edition, 1980.

[62] Sung Chun, Maxim Imakaev, Nathan O Stitziel, and Shamil R Sunyaev. Non-parametric polygenic risk prediction using partitioned gwas summary statistics. *bioRxiv*, 01 2018.

[63] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Gil McVean, Stephen Leslie, Peter Donnelly, and Jonathan Marchini. Genome-wide genetic data on ˜500,000 UK Biobank participants. *bioRxiv*, page 166298, July 2017.

[64] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An Open Access Resource for Identifying

the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3):e1001779, March 2015.

[65] Eric Y Durand, Chuong B Do, Joanna L Mountain, and J. Michael Macpherson. Ancestry composition: A novel, efficient pipeline for ancestry deconvolution. *bioRxiv*, 2014.

[66] Carla Marquez-Luna, The SIGMA Type 2 Diabetes Consortium, and Alkes L. Price. Multi-ethnic polygenic risk scores improve risk prediction in diverse populations. *bioRxiv*, page 051458, May 2016.

[67] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[68] UK10K Consortium et al. The uk10k project identifies rare variants in health and disease. *Nature*, 526(7571):82, 2015.

[69] Steven Gazal, Hilary K Finucane, and Alkes L Price. Reconciling s-ldsc and ldak functional enrichment estimates. *bioRxiv*, 2018.

[70] Doug Speed, Na Cai, the UCLEB Consortium, Michael R. Johnson, Sergey Nejentsev, and David J. Balding. Reevaluation of SNP heritability in complex human traits. *Nature Genetics*, 49(7):986–992, July 2017.

[71] Bogdan Pasaniuc and Alkes L. Price. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.*, 18(2):117–127, 2017.

[72] Robert M. Maier, Zhihong Zhu, Sang Hong Lee, Maciej Trzaskowski, Douglas M. Ruderfer, Eli A. Stahl, Stephan Ripke, Naomi R. Wray, Jian Yang, Peter M. Visscher, and Matthew R. Robinson. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nature Communications*, 9(1):989, March 2018.

[73] Gleb Kichaev, Gaurav Bhatia, Po-Ru Loh Loh, Steven Gazal, Kathryn Burch, Malika

Freund, Armin Schoech, Bogdan Pasaniuc, and Alkes L. Price. Leveraging polygenic functional enrichment to improve gwas power. *Submitted*.

[74] Masahiro Kanai, Masato Akiyama, Atsushi Takahashi, Nana Matoba, Yukihide Momozawa, Masashi Ikeda, Nakao Iwata, Shiro Ikegawa, Makoto Hirata, Koichi Matsuda, Michiaki Kubo, Yukinori Okada, and Yoichiro Kamatani. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature Genetics*, 50(3):390–400, March 2018.

[75] Diego Calderon, Anand Bhaskar, David A. Knowles, David Golan, Towfique Raj, Audrey Q. Fu, and Jonathan K. Pritchard. Inferring Relevant Cell Types for Complex Traits by Using Single-Cell Gene Expression. *Am. J. Hum. Genet.*, 101(5):686–699, November 2017.

[76] Halit Ongen, Andrew A. Brown, Olivier Delaneau, Nikolaos I. Panousis, Alexandra C. Nica, GTEx Consortium, and Emmanouil T. Dermitzakis. Estimating the causal tissues for complex traits and diseases. *Nat. Genet.*, 49(12):1676–1683, December 2017.

[77] Hilary K. Finucane, Yakir A. Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam Shoresh, Giulio Genovese, Arpiar Saunders, Evan Macosko, Samuela Pollack, Brainstorm Consortium, John R. B. Perry, Jason D. Buenrostro, Bradley E. Bernstein, Soumya Raychaudhuri, Steven McCarroll, Benjamin M. Neale, and Alkes L. Price. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.*, 50(4):621–629, April 2018.

[78] Daniel Backenroth, Zihuai He, Krzysztof Kiryluk, Valentina Boeva, Lynn Pethukova, Ekta Khurana, Angela Christiano, Joseph D. Buxbaum, and Iuliana Ionita-Laza. FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-Specific Functional Effects of Noncoding Variation: Methods and Applications. *Am. J. Hum. Genet.*, 102(5):920–942, May 2018.

[79] Armin Schoech, Daniel Jordan, Po-Ru Loh, Steven Gazal, Luke O'Connor, Daniel J. Balick, Pier F. Palamara, Hilary Finucane, Shamil R. Sunyaev, and Alkes L. Price. Quantification of frequency-dependent genetic architectures and action of negative selection in 25 UK Biobank traits. *bioRxiv*, page 188086, September 2017.

[80] Jian Zeng, Ronald de Vlaming, Yang Wu, Matthew R. Robinson, Luke R. Lloyd-Jones, Loic Yengo, Chloe X. Yap, Angli Xue, Julia Sidorenko, Allan F. McRae, Joseph E. Powell, Grant W. Montgomery, Andres Metspalu, Tonu Esko, Greg Gibson, Naomi R. Wray, Peter M. Visscher, and Jian Yang. Signatures of negative selection in the genetic architecture of human complex traits. *Nature Genetics*, 50(5):746–753, May 2018.

[81] Doug Speed and David Balding. Better estimation of SNP heritability from summary statistics provides a new understanding of the genetic architecture of complex traits. *bioRxiv*, page 284976, March 2018.

[82] Evangelos Evangelou and John P. A. Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.*, 14(6):379–389, June 2013.

[83] Adam E Locke, Ruth J F Loos, and Elizabeth K Speliotes. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 02 2015.

[84] Andrew R. Wood, Tonu Esko, Jian Yang, and Timothy M. Frayling. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.*, 46(11):1173–1186, November 2014.

[85] Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511:421 EP –, 07 2014.

[86] Cristen J Willer, Yun Li, and Gonçalo R Abecasis. Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 09 2010.

[87] Noah Zaitlen, Bogdan Paaniuc, Nick Patterson, Samuela Pollack, Benjamin Voight, Leif Groop, David Altshuler, Brian E. Henderson, Laurence N. Kolonel, Loic Le

Marchand, Kevin Waters, Christopher A. Haiman, Barbara E. Stranger, Emmanouil T. Dermitzakis, Peter Kraft, and Alkes L. Price. Analysis of casecontrol association studies with known risk variants. *Bioinformatics*, 28(13):1729–1737, 2012.

[88] Noah Zaitlen, Sara Lindstrm, Bogdan Pasaniuc, Marilyn Cornelis, Giulio Genovese, Samuela Pollack, Anne Barton, Heike Bickebller, Donald W. Bowden, Steve Eyre, Barry I. Freedman, David J. Friedman, John K. Field, Leif Groop, Aage Haugen, Joachim Heinrich, Brian E. Henderson, Pamela J. Hicks, Lynne J. Hocking, Laurence N. Kolonel, Maria Teresa Landi, Carl D. Langefeld, Loic Le Marchand, Michael Meister, Ann W. Morgan, Olaide Y. Raji, Angela Risch, Albert Rosenberger, David Scherf, Sophia Steer, Martin Walshaw, Kevin M. Waters, Anthony G. Wilson, Paul Wordsworth, Shanbeh Zienolddiny, Eric Tchetgen Tchetgen, Christopher Haiman, David J. Hunter, Robert M. Plenge, Jane Worthington, David C. Christiani, Debra A. Schaumberg, Daniel I. Chasman, David Altshuler, Benjamin Voight, Peter Kraft, Nick Patterson, and Alkes L. Price. Informed Conditioning on Clinical Covariates Increases Power in Case-Control Association Studies. *PLOS Genetics*, 8(11):e1003032, November 2012.

[89] Andrew P Morris. Transethnic meta-analysis of genomewide association studies. *Genetic Epidemiology*, 35(8):809–822, 12 2011.

[90] Marc A Coram, Sophie I Candille, Qing Duan, Kei Hang K Chan, Yun Li, Charles Kooperberg, Alex P Reiner, and Hua Tang. Leveraging multi-ethnic evidence for mapping complex traits in minority populations: An empirical bayes approach. *American Journal of Human Genetics*, 96(5):740–752, 05 2015.

Appendix A

Multi-ethnic polygenic risk scores improve risk prediction in diverse populations

Supplementary Tables:

Model	LD-pruning thresholds			
	0.1	**0.2**	**0.5**	**0.8**
EUR	0.02886 (0.001)	0.03089 (0.001)	0.03610 (0.002)	0.03927 (0.002)
LAT	0.02268 (0.002)	0.02516 (0.002)	0.02845 (0.003)	0.03200 (0.001)
LAT+ANC	0.03262 (0.006)	0.03486 (0.006)	0.03759 (0.006)	0.04115 (0.002)
EUR+LAT	0.05020 (0.002)	0.05338 (0.002)	0.05984 (0.002)	0.06441 (0.002)
EUR+LAT+ANC	0.05432 (0.002)	0.05739 (0.002)	0.06449 (0.002)	0.07053 (0.002)

S1 Table. Prediction accuracy of 5 prediction methods in simulations using different LD-pruning thresholds. Reported values are mean adjusted R^2 and s.e. over 100 simulations.

Model	Average weight (s.e.) associated to each predictor.		Average adj. R^2 (s.e.)	European training	Latino training
	European PRS	Latino PRS		Median P-value threshold	Median P-value threshold
EUR	0.19449 (0.004)		0.03927 (0.002)	0.01	
LAT_{unadj}		0.12577 (0.004)	0.01731 (0.001)		10^{-6}
LAT_{unadj}+ANC		0.18251 (0.01)	0.01814 (0.001)		10^{-6}
LAT		0.17780 (0.003)	0.03200 (0.001)		0.05
LAT+ANC		0.17613 (0.002)	0.04115 (0.002)		0.05
EUR+LAT_{unadj}	0.19436 (0.004)	0.07765 (0.006)	0.04865 (0.002)	0.01	10^{-6}
EUR+LAT_{unadj}+ANC	0.20419 (0.004)	0.15806 (0.009)	0.05106 (0.001)	0.01	10^{-6}
EUR+LAT	0.17847 (0.004)	0.15784 (0.003)	0.06441 (0.002)	0.01	0.05
EUR+LAT+ANC	0.19098 (0.004)	0.15578 (0.002)	0.07053 (0.002)	0.01	0.05

S2 Table. Accuracy of 9 prediction methods in simulations. We report prediction accuracies for methods using both ancestry-adjusted Latino effect sizes (LAT) and ancestry-unadjusted Latino effect sizes (LAT_{unadj}). Reported values are mean adjusted R^2 over 100 simulations. We also report normalized weights, defined as the mixing weight $\hat{\alpha}_k$ (see Methods) multiplied by the standard deviation of the PRS.

Model	Average R^2 (s.e.)
EUR	0.0254 (0.019)
LAT_{unadj}	0.3721 (0.034)
LAT_{unadj}+ANC	0.2205 (0.037)
LAT	0.0015 (0.007)
LAT+ANC	0.0437 (0.025)
EUR+LAT_{unadj}	0.0626 (0.02)
EUR+LAT_{unadj}+ANC	0.0337 (0.019)
EUR+LAT	0.0103 (0.016)
EUR+LAT+ANC	0.0178 (0.018)

S3 Table. R^2 with European ancestry for 9 prediction methods in simulations. European ancestry is represented by PC1 in the SIGMA data set. Reported values are mean R^2 over 100 simulations. The average R^2 between ancestry and phenotype was 0.011.

Model	Average weight (s.e.) associated to each predictor.		Average adj. R^2 (s.e.)	European training	Latino training
	European PRS	Latino PRS		Median P-value threshold	Median P-value threshold
EUR	0.19452 (0.004)		0.03927 (0.002)	0.01	
LAT_{unadj}		0.01353 (0.011)	0.01181 (0.001)		10^{-6}
LAT_{unadj}+ANC		0.24467 (0.016)	0.01359 (0.001)		10^{-6}
LAT		0.17866 (0.002)	0.03227 (0.001)		0.05
LAT+ANC		0.17650 (0.002)	0.04095 (0.002)		0.05
EUR+LAT_{unadj}	0.20402 (0.004)	0.01035 (0.009)	0.04587 (0.002)	0.01	10^{-6}
EUR+LAT_{unadj}+ANC	0.20671 (0.004)	0.19082 (0.014)	0.04760 (0.002)	0.01	10^{-6}
EUR+LAT	0.17729 (0.004)	0.15818 (0.002)	0.06426 (0.002)	0.01	0.05
EUR+LAT+ANC	0.19060 (0.004)	0.15681 (0.002)	0.06960 (0.002)	0.01	0.05

S4 Table. Accuracy of 9 prediction methods in simulations with ancestry-correlated phenotypes. We report prediction accuracies for methods using both ancestry-adjusted Latino effect sizes (LAT) and ancestry-unadjusted Latino effect sizes (LAT_{unadj}). Reported values are mean adjusted R^2 and s.e. over 100 simulations. We also report normalized weights, defined as the mixing weight $\hat{\alpha}_k$ (see Methods) multiplied by the standard deviation of the PRS.

A)

Model	Chr 1	Chr 1-2	Chr 1-4	Chr 1-22
EUR	0.18641 (0.003)	0.15778 (0.003)	0.12453 (0.002)	0.03927 (0.002)
LAT	0.14580 (0.003)	0.11512 (0.003)	0.08360 (0.002)	0.03200 (0.001)
LAT+ANC	0.14941 (0.003)	0.11859 (0.003)	0.08651 (0.002)	0.04115 (0.002)
EUR+LAT	0.21298 (0.003)	0.18374 (0.003)	0.14931 (0.002)	0.06441 (0.002)
EUR+LAT+ANC	0.21576 (0.003)	0.18695 (0.003)	0.15244 (0.002)	0.07053 (0.002)

B)

Model	Chr 1	Chr 1-2	Chr 1-4	Chr 1-22
EUR	0.08946 (0.003)	0.04638 (0.002)	0.03451 (0.001)	0.01156 (0.001)
LAT	0.14417 (0.003)	0.11523 (0.003)	0.08371 (0.002)	0.03391 (0.001)
LAT+ANC	0.14794 (0.003)	0.1188 (0.003)	0.08673 (0.002)	0.04202 (0.002)
EUR+LAT	0.17003 (0.003)	0.13095 (0.003)	0.09926 (0.002)	0.04211 (0.001)
EUR+LAT+ANC	0.17353 (0.003)	0.13436 (0.003)	0.10204 (0.002)	0.04751 (0.002)

S5 Table. Numerical values of results displayed in Fig 1A and 1B. We report results for A) 2:1 training sample size ratio (row 1 of Table 1) and B) 1:1 training sample size ratio (row 2 of Table 1). We report prediction accuracies for each of the 5 main prediction methods, for each subset of chromosomes. Reported values are mean adjusted R^2 and s.e. over 100 simulations.

A)

Model	Chr 1	Chr 1-2	Chr 1-4	Chr 1-22
EUR	0.277 (0.003)	0.247 (0.003)	0.207 (0.002)	0.079 (0.003)
LAT	0.143 (0.003)	0.130 (0.003)	0.113 (0.002)	0.042 (0.001)
LAT+ANC	0.158 (0.003)	0.141 (0.003)	0.120 (0.002)	0.052 (0.002)
EUR+LAT	0.295 (0.003)	0.267 (0.003)	0.232 (0.002)	0.106 (0.002)
EUR+LAT+ANC	0.301 (0.002)	0.275 (0.003)	0.243 (0.002)	0.122 (0.002)

B)

Model	Chr 1	Chr 1-2	Chr 1-4	Chr 1-22
EUR	0.166 (0.005)	0.080 (0.003)	0.069 (0.002)	0.022 (0.001)
LAT	0.142 (0.003)	0.130 (0.003)	0.113 (0.002)	0.044 (0.001)
LAT+ANC	0.156 (0.003)	0.141 (0.003)	0.119 (0.002)	0.053 (0.002)
EUR+LAT	0.229 (0.004)	0.169 (0.003)	0.148 (0.002)	0.060 (0.001)
EUR+LAT+ANC	0.238 (0.004)	0.178 (0.003)	0.155 (0.002)	0.067 (0.002)

S6 Table. Numerical values of results displayed in S1 Fig A and B. We report results for A) 2:1 training sample size ratio (row 1 of Table 1) and B) 1:1 training sample size ratio (row 2 of Table 1). We report prediction accuracies for each of the 5 main prediction methods, for each subset of chromosomes, in simulations including the causal SNPs. Reported values are mean adjusted R^2 and s.e. over 100 simulations.

Model	Observed-scale adj. R^2	Liability-scale adj. R^2	Nagelkerke R^2	AUC
EUR	0.02707	0.02700	0.03633	0.59012
LAT	0.02042	0.02030	0.02742	0.58175
LAT+ANC	0.03361	0.03362	0.04517	0.60342
EUR+LAT	0.04702	0.04735	0.06311	0.62375
EUR+LAT+ANC	0.04703	0.04736	0.06328	0.62416

S7 Table. Accuracy of 5 prediction methods in analyses of type 2 diabetes in a Latino cohort, using alternate prediction metrics. Liability-scale adjusted R^2 was computed assuming a disease prevalence of K=0.08.

Model	LD-pruning thresholds			
	0.1	**0.2**	**0.5**	**0.8**
EUR	0.02256	0.02339	0.02573	0.02700
LAT	0.01830	0.01842	0.01980	0.02030
LAT+ANC	0.03219	0.03148	0.03261	0.03362
EUR+LAT	0.04167	0.04229	0.04496	0.04735
EUR+LAT+ANC	0.04168	0.04226	0.04491	0.04736
EUR-LAT-meta	0.02556	0.02801	0.03270	0.03770

S8 Table. Prediction accuracy of main prediction methods in analyses of type 2 diabetes in a Latino cohort using different LD-pruning thresholds. Liability-scale adjusted R^2 was computed assuming a disease prevalence of K=0.08.

Model	R	R^2
EUR	-0.751	0.564
LAT_{unadj}	-0.995	0.990
LAT_{unadj}+ANC	-0.999	0.999
LAT	0.025	0.001
LAT+ANC	-0.607	0.369
EUR+LAT_{unadj}	-0.684	0.468
EUR+LAT_{unadj} +ANC	-0.671	0.450
EUR+LAT	-0.548	0.300
EUR+LAT+ANC	-0.513	0.263
T2D phenotype	-0.112	0.013

S9 Table. R and R^2 with European ancestry for 9 prediction methods and T2D phenotype in analyses of type 2 diabetes in a Latino cohort. European ancestry is represented by PC1 in the SIGMA data set.

Model	Weight associated to each predictor		Adjusted R^2	European training	Latino training
	European PRS	Latino PRS		P-value threshold	P-value threshold
EUR	0.16490		0.02700	0.05	
LAT_{unadj}		0.11151	0.01219		0.05
LAT_{unadj}+ANC		0.03866	0.01213		0.05
LAT		0.14332	0.02030		0.2
LAT+ANC		0.14623	0.03362		0.2
EUR+LAT_{unadj}	0.18268	-0.02398	0.02714	0.05	0.05
EUR+LAT_{unadj} +ANC	0.18736	0.13564	0.02728	0.05	0.05
EUR+LAT	0.16344	0.14164	0.04735	0.05	0.2
EUR+LAT+ANC	0.17629	0.14108	0.04736	0.05	0.2

S10 Table. Accuracy of 9 prediction methods in analyses of type 2 diabetes in a Latino cohort. We report adjusted R^2 on the liability scale for methods using both ancestry-adjusted Latino effect sizes (LAT) and ancestry-unadjusted Latino effect sizes (LAT_{unadj}). We also report normalized weights, defined as the mixing weight $\hat{\alpha}_k$ (see Methods) multiplied by the standard deviation of the PRS. We also report normalized weights, defined as the mixing weight $\hat{\alpha}_k$ (see Methods) multiplied by the standard deviation of the PRS.

Model	Weights associated to each predictor		Adjusted R^2
	EUR	LAT	
EUR	0.15625		0.02410
LAT		0.14062	0.01941
LAT+ANC		0.11329	0.02223
EUR+LAT	0.12754	0.10611	0.03469
EUR+LAT+ANC	0.13456	0.11083	0.03470

S11 Table. Accuracy of 5 prediction methods in analyses of type 2 diabetes in a Latino cohort using imputed genotypes. We report R^2 on the liability scale for each of the 5 main prediction methods. We also report normalized weights, defined as the mixing weight $\hat{\alpha}_k$ (see Methods) multiplied by the standard deviation of the PRS.

Model	10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	0.01	0.02	0.05	0.1	0.2	0.5	1
EUR	0.002	0.002	0.003	0.004	0.011	0.019	0.024	0.027	0.027	0.027	0.026	0.026	0.026
EUR+LAT	0.018	0.018	0.019	0.020	0.028	0.037	0.043	0.045	0.046	0.047	0.046	0.045	0.046
EUR+LAT+ANC	0.037	0.037	0.039	0.040	0.042	0.043	0.044	0.046	0.046	0.047	0.046	0.046	0.046

S12A Table. Numerical values for results displayed in Fig 2A. We report prediction adjusted R^2 for each of the 3 prediction methods that include the EUR predictor.

	P-value Threshold												
Model	**10^{-8}**	**10^{-7}**	**10^{-6}**	**10^{-5}**	**10^{-4}**	**10^{-3}**	**0.01**	**0.02**	**0.05**	**0.1**	**0.2**	**0.5**	**1**
LAT	0.002	0.003	0.005	0.005	0.008	0.009	0.011	0.014	0.017	0.019	0.020	0.020	0.020
LAT+ANC	0.019	0.022	0.024	0.027	0.027	0.026	0.028	0.030	0.033	0.034	0.034	0.032	0.032
EUR+LAT	0.033	0.035	0.037	0.040	0.040	0.039	0.041	0.043	0.046	0.046	0.046	0.045	0.045
EUR+LAT+ANC	0.033	0.035	0.037	0.040	0.040	0.039	0.041	0.043	0.046	0.047	0.047	0.045	0.045

S12B Table. Numerical values for results displayed in Fig 2B. We report prediction adjusted R^2 for each of the 4 prediction methods that include the LAT predictor.

Model	Average weight (s.d.) associated to each predictor.		Avg. adj. R^2 across folds (s.d.)	Adj. R^2 merging folds
	European PRS	Latino PRS		
EUR	0.165 (0.004)		0.02731 (0.014)	0.02650
LAT		0.133 (0.012)	0.01966 (0.006)	0.01997
LAT+ANC		0.130 (0.008)	0.03230 (0.009)	0.03267
EUR+LAT	0.158 (0.008)	0.125 (0.008)	0.04645 (0.014)	0.04646
EUR+LAT+ANC	0.177 (0.008)	0.125 (0.008)	0.04596 (0.014)	0.04593

S13 Table. Accuracy of 5 prediction methods in analyses of type 2 diabetes in a Latino cohort, using 10x9-fold cross-validation. We report adjusted R^2 on the liability scale for each of the 5 main prediction methods, and the average of adjusted R^2 within each fold. Adjusted R^2 merging folds is lower than average adjusted R^2 across folds because of miscalibration between folds. We used 10-fold cross-validation for EUR and 10x9-fold cross-validation for LAT, LAT+ANC, EUR+LAT and EUR+LAT+ANC (see Methods). We also report normalized weights, defined as the mixing weight $\hat{\alpha}_k$ (see Methods) multiplied by the standard deviation of the PRS.

Model	Observed-scale adj. R^2	Liability-scale adj. R^2	Nagelkerke R^2	AUC
EUR	0.00753	0.01767	0.01423	0.57453
SAS	0.00664	0.01556	0.01243	0.55606
SAS+ANC	0.00670	0.01572	0.01359	0.56153
EUR+SAS	0.01292	0.03031	0.02454	0.59155
EUR+SAS+ANC	0.01265	0.02968	0.02507	0.59366

S14 Table. Accuracy of 5 prediction methods in analyses of type 2 diabetes in a South Asian cohort, using alternate prediction metrics. Liability-scale adjusted R^2 was computed using the sample disease prevalence estimate of K=0.15.

Model	LD-pruning threshold			
	0.1	**0.2**	**0.5**	**0.8**
EUR	0.01064	0.01272	0.01380	0.01767
SAS	0.01212	0.00994	0.01196	0.01556
SAS+ANC	0.01220	0.01000	0.01203	0.01572
EUR+SAS	0.02213	0.02209	0.02456	0.03031
EUR+SAS+ANC	0.02157	0.02120	0.02366	0.02968

S15 Table. Prediction accuracy of 5 prediction methods in analyses of type 2 diabetes in a South Asian cohort using different LD-pruning thresholds. Liability-scale adjusted R^2 was computed using the sample disease prevalence estimate of K=0.15.

Model	*R* with PC1	R^2 with PC1
EUR	-0.08572	0.00735
SAS	0.13099	0.01716
SAS+ANC	-0.15702	0.02466
EUR+SAS	0.02550	0.00065
EUR+SAS+ANC	-0.11607	0.01347
T2D phenotype	-0.01390	0.00019

S16 Table. *R* and R^2 with European ancestry for 5 prediction methods and T2D phenotype in analyses of type 2 diabetes in a South Asian cohort. European ancestry is represented by PC1 in the data set.

Model	Weight EUR PRS	Weight SAS PRS	Avg. adj. R^2 across folds (s.d)	Adj. R^2 merging folds
EUR	0.09001 (0.007)		0.01681 (0.031)	0.01519
SAS		0.08487 (0.008)	0.01700 (0.035)	0.01257
SAS+ANC		0.08821 (0.008)	0.01572 (0.034)	0.01188
EUR+SAS	0.08310 (0.007)	0.07745 (0.008)	0.02785 (0.039)	0.02614
EUR+SAS+ANC	0.08140 (0.007)	0.07987 (0.008)	0.02642 (0.039)	0.02462

S17 Table. Accuracy of 5 prediction methods in analyses of type 2 diabetes in a South Asian cohort, using stratified 10-fold cross-validation. We report adjusted R^2 on the liability scale averaged over 500 different partitions of the data into 10 stratified folds, and the average of adjusted R^2 within each fold. Adjusted R^2 merging folds is lower than average adjusted R^2 across folds because of miscalibration between folds. We used 10-fold cross-validation for all methods, including EUR and SAS (see Methods). We also report normalized weights, defined as the mixing weight $\hat{\alpha}_k$ (see Methods) multiplied by the standard deviation of the PRS.

Model	LD-pruning threshold			
	0.1	**0.2**	**0.5**	**0.8**
EUR	0.01442	0.02619	0.02215	0.02235
AFR	0.00785	0.00877	0.01023	0.01075
AFR+ANC	0.00981	0.01081	0.01238	0.01332
EUR+AFR	0.02095	0.03319	0.03103	0.02940
EUR+AFR+ANC	0.02420	0.03344	0.03048	0.03019

S18 Table. Prediction accuracy of 5 prediction methods in analyses of height in an African cohort using different LD-pruning thresholds.

Model	*R* with PC1	R^2 with PC1
EUR	-0.12249	0.01500
AFR	0.29584	0.08752
AFR+ANC	-0.18300	0.03349
EUR+AFR	0.04358	0.00190
EUR+AFR+ANC	-0.11575	0.01340
Height	-0.02199	0.00048

S19 Table. *R* and R^2 with European ancestry for 5 prediction methods and height phenotype in analyses of height in an African cohort. European ancestry is represented by PC1 in the data set.

Model	Weight EUR PRS	Weight AFR PRS	Avg. adj. R^2 across folds (s.d.)	Adj. R^2 merging folds
EUR	0.16352 (0.008)		0.02653 (0.026)	0.02377
AFR		0.10635 (0.008)	0.01075 (0.017)	0.0085
AFR+ANC		0.12366 (0.008)	0.01253 (0.018)	0.01046
EUR+AFR	0.15485 (0.009)	0.09171 (0.008)	0.03358 (0.028)	0.03095
EUR+AFR+ANC	0.14969 (0.008)	0.10221 (0.008)	0.03347 (0.029)	0.03087

S20 Table. Accuracy of 5 prediction methods in analyses of height in an African cohort, using 10-fold cross validation. We report adjusted R^2 merging folds averaged over 500 different partitions of the data into 10 stratified folds, and the average of adjusted R^2 within each fold. Adjusted R^2 merging folds is lower than average adjusted R^2 across folds because of miscalibration between folds. We used 10-fold cross-validation for all methods, including EUR and AFR (see Methods). We also report normalized weights, defined as the mixing weight $\hat{\alpha}_k$ (see Methods) multiplied by the standard deviation of the PRS.

Table S21. Phenotypes for which GWAS have been published in Europeans and at least one non-European population with minimum sample size of 8,000.

STUDY.ACCESSION	PUBMEDID	FIRST.AUTHOR	DATE	STAGE	NUMBER.OF.INDIVDUALS	BROAD.ANCESTRAL.CATEGORY	COUNTRY.OF.ORIGIN	DISEASE.trait
GCST002245	24162737	European Alzheimer's Disease Initiative (EADI)	10/27/13	initial	55134	European	NR	Alzheimer's disease (late onset)
GCST002954	26049409	Hirano A	6/5/15	initial	8808	East Asian	NR	Alzheimer's disease (late onset)
GCST001026	21460841	Naj AC	4/3/11	initial	15675	European	NR	Alzheimer's disease (late onset)
GCST001709	23042114	Hirota T	10/7/12	initial	9443	East Asian	NR	Atopic dermatitis
GCST001363	22197932	Paternoster L	12/25/11	initial	26171	European	NR	Atopic dermatitis
GCST000602	20173747	Ellinor PT	2/21/10	initial	14179	European	NR	Atrial fibrillation
GCST000446	19597491	Gudbjartsson DF	7/13/09	initial	36137	European	NR	Atrial fibrillation
GCST000445	19597492	Benjamin EJ	7/13/09	initial	40518	European	NR	Atrial fibrillation
GCST001499	22544366	Ellinor PT	4/29/12	initial	59133	European	NR	Atrial fibrillation
GCST004373	28416822	Low SK	4/17/17	initial	36792	East Asian	NR	Atrial fibrillation
GCST001072	21572416	Kato N	5/15/11	initial	19608	East Asian	NR	Blood pressure
GCST002167	24001895	Kelly TN	9/3/13	initial	22275	East Asian	NR	Blood pressure
GCST002143	23972371	Francescheini N	8/20/13	initial	28190	African American or Afro-Caribbean	NR	Blood pressure
GCST001235	21909110	Wain LV	9/11/11	initial	74064	European	NR	Blood pressure
GCST001676	22982992	Yang J	9/12/12	initial	133154	European	NR	Body mass index
GCST000185	18454148	Loos RJ	5/4/08	initial	16876	European	NR	Body mass index

Table S21 (Continued)

GCST001415	22344219	Wen W	2/19/12	initial	22762	East Asian	NR	Body mass index
GCST000298	19079261	Willer CJ	12/14/08	initial	32387	European	Italy	Body mass index
GCST001967	23583978	Monda KL	4/14/13	initial	37956	African American or Afro-Caribbean	NR	Body mass index
GCST000022	17434869	Frayling TM	4/12/07	initial	10657	European	U.K., Republic of Ireland	Body mass index
GCST002227	24064335	Pei YF	10/8/13	initial	8463	European	NR	Body mass index
GCST002461	24861553	Wen W	5/26/14	initial	82438	East Asian	NR	Body mass index
GCST002783	25673413	Locke AE	2/12/15	initial	236781	European	NR	Body mass index
GCST000830	20935630	Speliotes EK	10/10/10	initial	123865	European	NR	Body mass index
GCST002021	23669352	Graff M	5/12/13	initial	13627	European	NR	Body mass index
GCST000037	17529974	Stacey SN	5/27/07	initial	13145	European	NR	Breast cancer
GCST001937	23535729	Michailidou K	4/1/13	initial	22627	European	NR	Breast cancer
GCST000811	20872241	Li J	9/26/10	initial	8428	European	NR	Breast cancer
GCST002537	25038754	Cai Q	7/20/14	initial	9450	East Asian	NR	Breast cancer
GCST000678	20453838	Turnbull C	5/9/10	initial	8556	European	NR	Breast cancer
GCST001683	22976474	Siddiq A	9/13/12	initial	32530	European	NR	Breast cancer
GCST003782	28171663	Huo D	9/4/16	initial	8112	African American or Afro-Caribbean	NR	Breast cancer
GCST001930	23535733	Garcia-Closas M	4/1/13	initial	39387	European	NR	Breast cancer
GCST003842	27117709	Couch FJ	4/27/16	initial	19291	European	NR	Breast cancer
GCST003520	27354352	Han MR	6/27/16	initial	13905	East Asian	NR	Breast cancer
GCST000933	21196492	Okada Y	12/31/10	initial	10112	East Asian	NR	C-reactive protein
GCST000430	19567438	Elliott P	7/1/09	initial	17967	South Asian, European	NR	C-reactive protein

Table S21 (Continued)

GCST001650	22939635	Reiner AP	8/28/12	initial	8280	African American or Afro-Caribbean	NR	C-reactive protein
GCST001787	23266556	Peters U	12/21/12	initial	27809	European	NR	Colorectal cancer
GCST003017	26151821	Schumacher FR	7/7/15	initial	37955	European	NR	Colorectal cancer
GCST002454	24836286	Zhang B	5/18/14	initial	8270	East Asian	NR	Colorectal cancer
GCST003799	26965516	Zeng C	3/8/16	initial	21096	East Asian	NR	Colorectal cancer
GCST002411	24737748	Whiffin N	4/15/14	initial	13443	European	NR	Colorectal cancer
GCST002919	25990418	Al-Tassan NA	5/20/15	initial	17556	European	NR	Colorectal cancer
GCST001544	22634755	Dunlop MG	5/27/12	initial	17780	European	NR	Colorectal cancer
GCST002586	25187374	Hwang JY	9/3/14	initial	24740	East Asian	NR	Fasting plasma glucose
GCST000276	19060907	Prokopenko I	12/1/08	initial	35812	European	NR	Fasting plasma glucose
GCST000303	19096518	Pare G	12/19/08	initial	14618	European	NR	Glycated hemoglobin levels
GCST002390	24647736	Chen P	3/19/14	initial	17290	East Asian	NR	Glycated hemoglobin levels
GCST000803	20858683	Soranzo N	9/21/10	initial	46368	European	NR	Glycated hemoglobin levels
GCST000431	19570815	Estrada K	7/1/09	initial	10074	European	NR	Height
GCST000644	20397748	Liu JZ	4/1/10	initial	11536	European	NR	Height
GCST000372	19343178	Soranzo N	4/3/09	initial	12611	European	NR	Height
GCST000174	18391952	Weedon MN	4/6/08	initial	13665	European	NR	Height
GCST000817	20881960	Lango Allen H	9/29/10	initial	133653	European	NR	Height
GCST000176	18391950	Lettre G	4/6/08	initial	15821	European	NR	Height
GCST000611	20189936	Okada Y	2/26/10	initial	19633	East Asian	NR	Height

Table S21 (Continued)

GCST001263	21998595	N'Diaye A	10/6/11	initial	20427	African American or Afro-Caribbean, African unspecified	NR	Height
GCST002647	25282103	Wood AR	10/5/14	initial	253288	European	NR	Height
GCST000175	18391951	Gudbjartsson DF	4/6/08	initial	30968	European	NR	Height
GCST002702	25429064	He M	11/26/14	initial	36227	East Asian	NR	Height
GCST001290	22021425	Carty CL	10/21/11	initial	8149	African American or Afro-Caribbean	NR	Height
GCST000522	19893584	Kim JJ	11/6/09	initial	8842	East Asian	NR	Height
GCST000398	19430479	Levy D	5/10/09	initial	29136	European	NR	Hypertension
GCST004143	28273873	Park YM	3/5/17	initial	8839	East Asian	NR	Hypertension
GCST001506	22570627	van Koolwijk LM	5/3/12	initial	11972	European	NR	Intraocular pressure
GCST002580	25173106	Hysi PG	8/31/14	initial	27558	European	NR	Intraocular pressure
GCST002767	25637523	Springelkamp H	1/30/15	initial	8105	NR	NR	Intraocular pressure
GCST002466	24880342	Wang Y	6/1/14	initial	27209	European	NR	Lung cancer
GCST000257	18978787	Wang Y	11/2/08	initial	10295	European	NR	Lung cancer
GCST001740	23143601	Lan Q	11/11/12	initial	10054	East Asian	NR	Lung cancer
GCST003325	26732429	Wang Z	1/4/16	initial	13154	East Asian	NR	Lung cancer
GCST001638	22899653	Timofeeva MN	8/16/12	initial	44385	European	NR	Lung cancer
GCST001335	22139419	Gieger C	11/30/11	initial	18600	European	Italy, Germany	Mean platelet volume
GCST001439	22423221	Qayyum R	3/8/12	initial	16388	African American or Afro-Caribbean	NR	Mean platelet volume
GCST000400	19448622	Sulem P	5/15/09	initial	15297	European	NR	Menarche (age at onset)
GCST002013	23667675	Tanikawa C	5/7/13	initial	15495	East Asian	NR	Menarche (age at onset)
GCST000404	19448620	Perry JR	5/17/09	initial	17510	European	NR	Menarche (age at onset)

Table S21 (Continued)

GCST001973	23599027	Demerath EW	4/17/13	initial	18089	African American or Afro-Caribbean	NR	Menarche (age at onset)
GCST000880	21102462	Elks CE	11/21/10	initial	87802	European	Italy, Netherlands	Menarche (age at onset)
GCST002541	25231870	Perry JR	7/23/14	initial	182413	European	NR	Menarche (age at onset)
GCST001436	22399527	Kristiansson K	3/7/12	initial	10564	European	NR	Metabolic syndrome
GCST002732	25705158	Shim U	12/31/14	initial	8842	East Asian	NR	Metabolic syndrome
GCST002544	25064009	Nalls MA	7/27/14	initial	108990	European	NR	Parkinson's disease
GCST001126	21738487	Do CB	6/23/11	initial	33050	European	NR	Parkinson's disease
GCST001430	22451204	Pankratz N	3/1/12	initial	8477	European	NR	Parkinson's disease
GCST000959	21292315	Nalls MA	2/1/11	initial	17352	European	NR	Parkinson's disease
GCST003922	28011712	Foo JN	12/22/16	initial	14006	East Asian	NR	Parkinson's disease
GCST003383	26805783	Schick UM	1/21/16	initial	12491	Hispanic or Latin American	NR	Platelet count
GCST002186	24026423	Shameer K	9/12/13	initial	13582	European	NR	Platelet count
GCST002733	25705162	Oh JH	12/31/14	initial	8842	East Asian	NR	Platelet count
GCST001735	23139255	Butler AM	11/8/12	initial	13415	African American or Afro-Caribbean	NR	PR interval
GCST000562	20062060	Pfeufer A	1/10/10	initial	28517	European	NR	PR interval
GCST001746	23166209	Smith JG	11/19/12	initial	13105	African American or Afro-Caribbean	NR	QT interval
GCST000363	19305408	Newton-Cheh C	3/22/09	initial	13685	European	NR	QT interval
GCST000364	19305409	Pfeufer A	3/22/09	initial	15842	European	NR	QT interval
GCST002500	24952745	Arking DE	6/22/14	initial	71061	European	Italy, Germany	QT interval

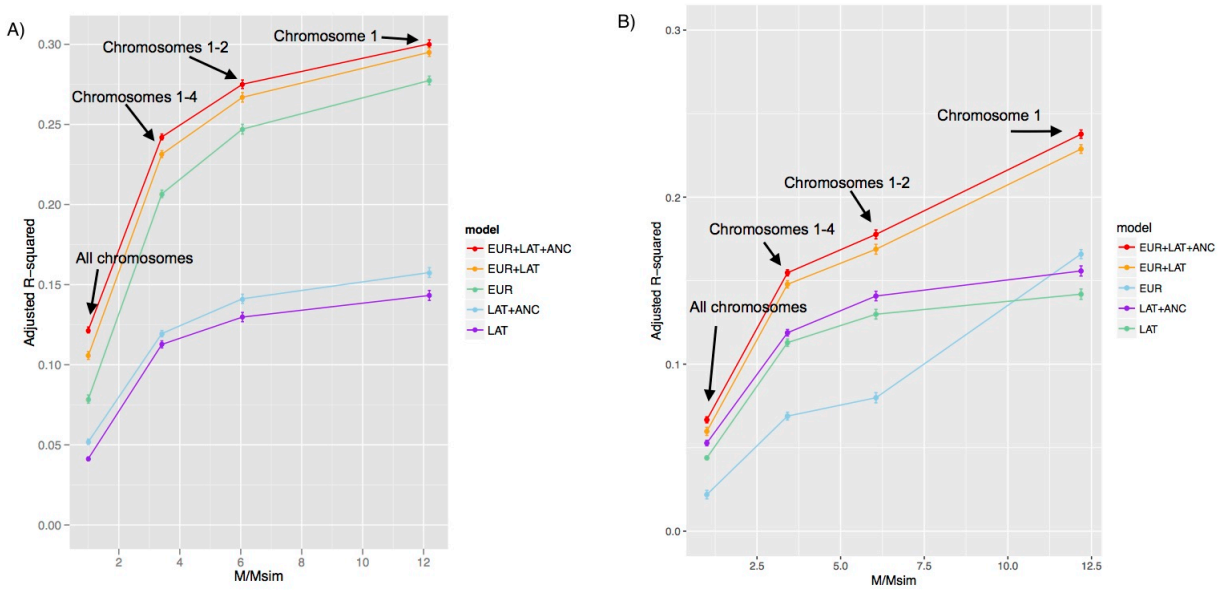
Table S21 (Continued)								
GCST003818	27798624	Eppinga RN	10/31/16	initial	127919	European	U.K.	Resting heart rate
GCST001748	23183192	Deo R	11/23/12	initial	13372	African American or Afro-Caribbean	NR	Resting heart rate
GCST000731	20639392	Eijgelsheim M	7/16/10	initial	38991	European	NR	Resting heart rate
GCST002323	24449572	Orozco G	1/1/14	initial	8305	European	NR	Rheumatoid arthritis
GCST000232	18794853	Raychaudhuri S	9/14/08	initial	15853	European	NR	Rheumatoid arthritis
GCST001454	22446963	Okada Y	3/25/12	initial	20965	East Asian	NR	Rheumatoid arthritis
GCST000679	20453842	Stahl EA	5/9/10	initial	25708	European	NR	Rheumatoid arthritis
GCST001851	23894747	Aberg KA	2/1/13	initial	21953	European	NR	Schizophrenia
GCST000435	19571808	Stefansson H	7/1/09	initial	16161	European	NR	Schizophrenia
GCST001301	22037555	Shi Y	10/30/11	initial	10218	East Asian	NR	Schizophrenia
GCST002539	25056061	Ripke S	7/22/14	initial	82315	European	Portugal, U.K., Republic of Ireland, Denmark	Schizophrenia
GCST003880	27922604	Yu H	12/6/16	initial	10154	East Asian	NR	Schizophrenia
GCST003048	26198764	Goes FS	7/21/15	initial	150064	NR	NR	Schizophrenia
GCST001242	21926974	Ripke S	9/18/11	initial	21856	European	NR	Schizophrenia
GCST001696	23049750	Kumasaka N	9/25/12	initial	11696	East Asian	NR	Smoking behavior
GCST000667	20418888	Thorgeirsson TE	4/25/10	initial	31266	European	NR	Smoking behavior
GCST000668	20418889	Liu JZ	4/25/10	initial	41150	European	NR	Smoking behavior
GCST001286	22006218	Yoon D	10/18/11	initial	8842	East Asian	NR	Smoking behavior
GCST001539	22832964	David SP	5/22/12	initial	32389	African American or Afro-Caribbean	NR	Smoking behavior

Table S21 (Continued)

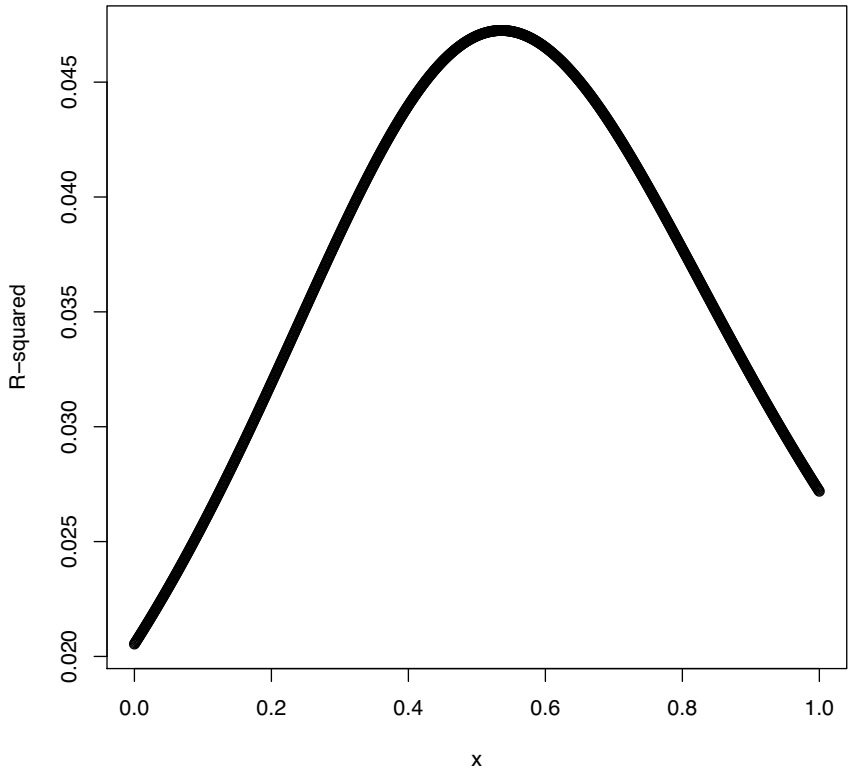
GCST000666	20418890	The Tobacco and Genetics Consortium	4/25/10	initial	74035	European	NR	Smoking behavior
GCST000379	19369658	Ikram MA	4/15/09	initial	19602	European	NR	Stroke
GCST001400	22306652	Bellenguez C	2/5/12	initial	9520	European	NR	Stroke
GCST002988	26089329	Carty CL	6/18/15	initial	14519	African American or Afro-Caribbean	NR	Stroke
GCST002630	25249183	Lu X	9/23/14	initial	11816	East Asian	NR	Systolic blood pressure
GCST004279	28135244	Warren HR	1/30/17	initial	140882	European	NR	Systolic blood pressure
GCST000394	19430483	Newton-Cheh C	5/10/09	initial	34433	European	NR	Systolic blood pressure
GCST001234	21909109	Kim YJ	9/11/11	initial	12545	East Asian	NR	Triglycerides
GCST003217	26582766	Lu X	11/18/15	initial	8344	East Asian	NR	Triglycerides
GCST002216	24097068	Willer CJ	10/6/13	initial	94595	European	NR	Triglycerides
GCST000027	17460697	Steinthorsdottir V	4/26/07	initial	8686	European	NR	Type 2 diabetes
GCST003400	26818947	Imamura M	1/28/16	initial	41646	East Asian	NR	Type 2 diabetes
GCST002317	24390345	Williams AL	12/25/13	initial	8214	Hispanic or Latin American	NR	Type 2 diabetes
GCST000167	18372903	Zeggini E	3/30/08	initial	10128	European	NR	Type 2 diabetes
GCST001213	21874001	Kooner JS	8/28/11	initial	20019	South Asian	India, Sri Lanka, Pakistan, Bangladesh	Type 2 diabetes
GCST002128	23945395	Hara K	8/14/13	initial	26805	East Asian	NR	Type 2 diabetes
GCST001351	22158537	Cho YS	12/11/11	initial	15000	East Asian	NR	Type 2 diabetes
GCST003619	27189021	Cook JP	5/18/16	initial	56799	European	NR	Type 2 diabetes
GCST000712	20581827	Voight BF	6/27/10	initial	47117	European	U.K.	Type 2 diabetes
GCST002560	25102180	Ng MC	8/7/14	initial	23827	African American or Afro-Caribbean	NR	Type 2 diabetes

Table S21 (Continued)								
GCST000242	18834626	Dehghan A	10/1/08	initial	11847	European	NR	Urate levels
GCST000818	20884846	Yang Q	9/30/10	initial	28283	European	NR	Urate levels
GCST001163	21768215	Tin A	7/18/11	initial	8651	African American or Afro-Caribbean	NR	Urate levels
GCST000427	19557197	Heard-Costa NL	6/26/09	initial	31373	European	NR	Waist circumference
GCST003337	26785701	Wen W	1/20/16	initial	39869	East Asian	NR	Waist circumference
GCST002138	23966867	Liu CT	8/15/13	initial	19744	African American or Afro-Caribbean	NR	Waist-hip ratio
GCST000829	20935629	Heid IM	10/10/10	initial	77167	European	NR	Waist-hip ratio
GCST003564	27195708	Scott WR	5/19/16	initial	10318	South Asian	India, Sri Lanka, Pakistan, Bangladesh	Waist-to-hip ratio adjusted for body mass index
GCST002782	25673412	Shungin D	2/12/15	initial	142762	European	NR	Waist-to-hip ratio adjusted for body mass index
GCST001302	22037903	Crosslin DR	10/30/11	initial	12046	European	NR	White blood cell count
GCST001133	21738479	Reiner AP	6/30/11	initial	16388	African American or Afro-Caribbean	NR	White blood cell count
GCST001137	21738480	Nalls MA	7/1/11	initial	19509	European	NR	White blood cell count
GCST004126	28158719	Jain D	2/1/17	initial	11809	Hispanic or Latin American	NR	White blood cell count

Supplementary Figures.



S1 Fig. Accuracy of 5 prediction methods in simulations using subsets of chromosomes, including the causal SNPs. We report results for A) 2:1 training sample size ratio (row 1 of Table 1) and B) 1:1 training sample size ratio (row 2 of Table 1). We report prediction accuracies for each of the 5 main prediction methods as a function of M/Msim, where M=232,629 is the total number of SNPs and Msim is the actual number of SNPS used in each simulation: 232,629 (all chromosomes), 68,188 (chromosomes 1-4), 38,412 (chromosomes 1-2), and 19,087 (chromosome 1). Numerical results are provided in S6 Table.



S2 Fig. Sensitivity to mixing weights in analyses of type 2 diabetes in a Latino cohort. We report the prediction R^2 of xEUR + (1-x)LAT, with x varying between 0 and 1. As expected, the prediction accuracy at x=0.8 is similar to the prediction accuracy of EUR-LAT-meta (Table 3).

Appendix B

Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets

Supplementary Figures

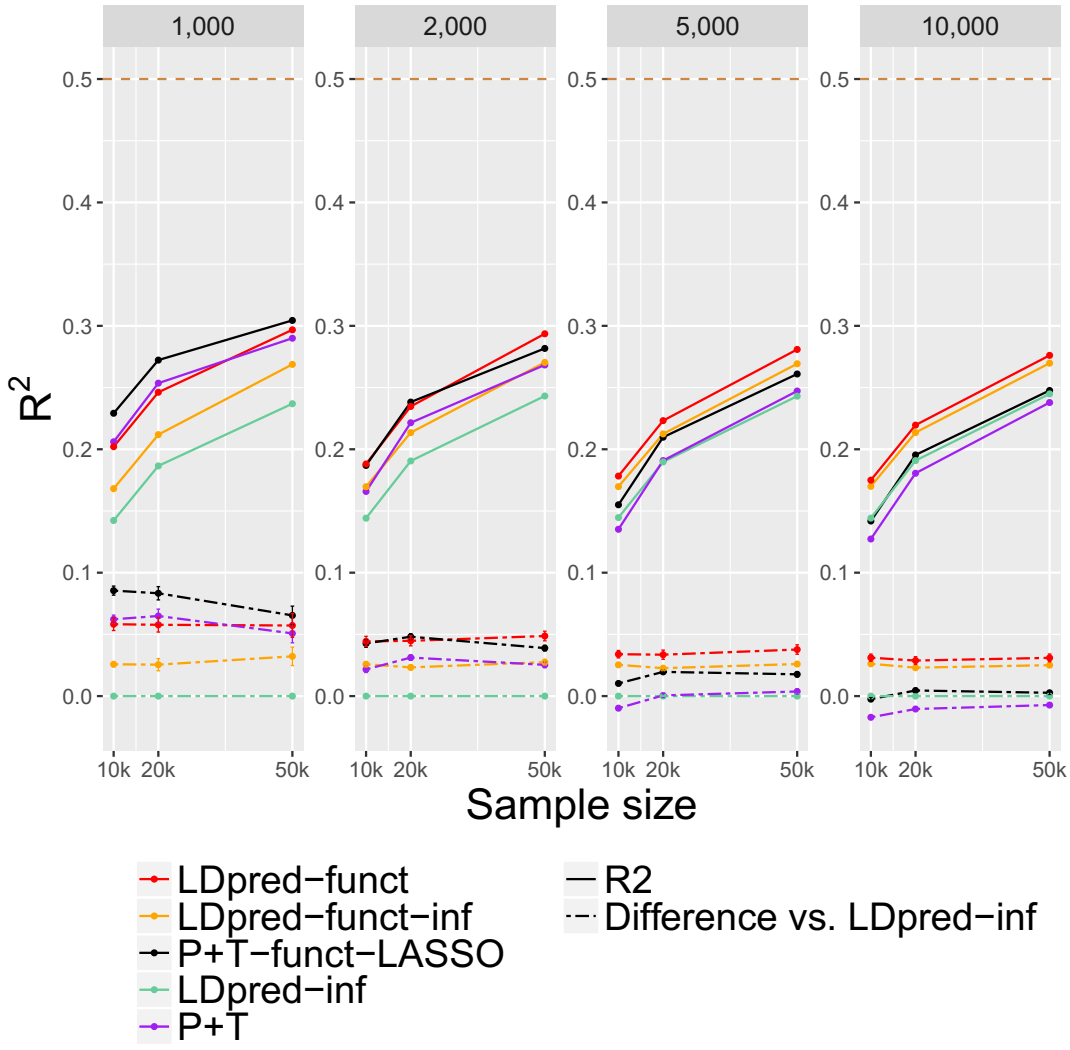


Figure S3: Accuracy of 5 polygenic prediction methods in simulations using UK Biobank genotypes, for 4 values of the number of causal variants. We report results for P+T, LDpred-inf, P+T-funct-LASSO, LDpred-funct-inf and LDpred-funct in chromosome 1 simulations with 1,000 causal variants (extremely sparse architecture), 2,000 causal variants (sparse architecture), 5,000 causal variants (polygenic architecture) and 10,000 causal variants (extremely polygenic architecture). Results are averaged across 100 simulations. Top dashed line denotes simulated SNP-heritability of 0.5. Bottom dashed lines denote differences vs. LDpred-inf; error bars represent 95% confidence intervals. Numerical results are reported in Table S23 and Table S24.

Supplementary Tables

Table S22: List of 16 UK Biobank traits. We list the training sample size and validation sample size for each trait.

	Trait	Training N	Validation N (ancestry distribution)	
1	Height	408092	25030	(43.5% Irish, 56.5% Other)
2	Hair color	403024	24773	(43.5% Irish, 56.5% Other)
3	Platelet count	395747	24277	(43.5% Irish, 56.5% Other)
4	Bone mineral density	397274	24167	(43.6% Irish, 56.4% Other)
5	Red blood cell count	396464	24305	(43.5% Irish, 56.5% Other)
6	FEV1-FVC ratio	331786	19929	(42.5% Irish, 57.5% Other)
7	Body mass index	407667	25000	(43.5% Irish, 56.5% Other)
8	RBC distribution width	394258	24175	(43.5% Irish, 56.5% Other)
9	Eosinophil count	391787	24030	(43.4% Irish, 56.6% Other)
10	Forced vital capacity	331786	19929	(42.5% Irish, 57.5% Other)
11	White blood cell count	395835	24293	(43.5% Irish, 56.5% Other)
12	Blood pressure	376437	23127	(43.2% Irish, 56.8% Other)
13	Age at menarche	214860	13999	(39.7% Irish, 60.3% Other)
14	Tanning ability	400721	24608	(43.5% Irish, 56.5% Other)
15	Balding type I	186506	10578	(48.9% Irish, 51.1% Other)
16	Waist hip ratio	408196	25032	(43.5% Irish, 56.5% Other)

Table S23: Accuracy of 5 polygenic prediction methods in simulations using UK Biobank genotypes, for 4 values of the number of causal variants. We report results for P+T, LDpred-inf, P+T-funct-LASSO, LDpred-funct-inf and LDpred-funct in chromosome 1 simulations with 1,000 causal variants (extremely sparse architecture), 2,000 causal variants (sparse architecture), 5,000 causal variants (polygenic architecture) and 10,000 causal variants (extremely polygenic architecture). Results are averaged across 100 simulations.

		Training sample size		
# Causal		10,000	20,000	50,000
variants	Model	Average $R^2(s.e.)$	Average $R^2(s.e.)$	Average $R^2(s.e.)$
1,000	P+T	0.2061 (0.0022)	0.2536 (0.0021)	0.2900 (0.0019)
	LDpred-inf	0.1423 (0.0020)	0.1865 (0.0031)	0.2369 (0.0045)
	P+T-funct-LASSO	0.2292 (0.0024)	0.2723 (0.0024)	0.3044 (0.002)
	LDpred-funct-inf	0.1681 (0.0024)	0.2119 (0.0028)	0.2688 (0.0033)
	LDpred-funct	0.2021 (0.0021)	0.2462 (0.0019)	0.2968 (0.0025)
2,000	P+T	0.1658 (0.0022)	0.2215 (0.0026)	0.2683 (0.0029)
	LDpred-inf	0.1442 (0.0019)	0.1905 (0.0023)	0.2432 (0.0028)
	P+T-funct-LASSO	0.1869 (0.0026)	0.2383 (0.0028)	0.2817 (0.0031)
	LDpred-funct-inf	0.1697 (0.0022)	0.2135 (0.0026)	0.2703 (0.003)
	LDpred-funct	0.1881 (0.0017)	0.2347 (0.0019)	0.2936 (0.0016)
5,000	P+T	0.1352 (0.0016)	0.1909 (0.0020)	0.2472 (0.0024)
	LDpred-inf	0.1447 (0.0017)	0.1898 (0.0022)	0.2430 (0.0027)
	P+T-funct-LASSO	0.1550 (0.0018)	0.2098 (0.0021)	0.2610 (0.0026)
	LDpred-funct-inf	0.1698 (0.0019)	0.2125 (0.0022)	0.2693 (0.0027)
	LDpred-funct	0.1783 (0.0012)	0.2232 (0.0013)	0.2809 (0.0015)
10,000	P+T	0.1273 (0.0015)	0.1806 (0.002)	0.2379 (0.0024)
	LDpred-inf	0.1442 (0.0017)	0.1908 (0.0021)	0.2449 (0.0026)
	P+T-funct-LASSO	0.1419 (0.0017)	0.1954 (0.0022)	0.2477 (0.0026)
	LDpred-funct-inf	0.1700 (0.0020)	0.2136 (0.0023)	0.2698 (0.0028)
	LDpred-funct	0.1750 (0.0012)	0.2196 (0.0012)	0.2761 (0.0013)

Table S24: Differences between polygenic prediction methods in simulations using UK Biobank genotypes, for 4 values of the number of causal variants. We report results for P+T, LDpred-inf, P+T-funct-LASSO, LDpred-funct-inf and LDpred-funct in chromosome 1 simulations with 1,000 causal variants (extremely sparse architecture), 2,000 causal variants (sparse architecture), 5,000 causal variants (polygenic architecture) and 10,000 causal variants (extremely polygenic architecture). Results are averaged across 100 simulations. (a) Difference between R^2 for each method vs. R^2 for LDpred-inf. (b) Difference between R^2 for LDpred-funct vs. R^2 for each method.

(a)

# Causal variants	Model	Training sample size 10,000 Diff. $R^2(s.e.)$	20,000 Diff. $R^2(s.e.)$	50,000 Diff. $R^2(s.e.)$
1,000	P+T	0.0622 (0.0017)	0.0649 (0.0028)	0.0508 (0.0038)
	LDpred-inf	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
	P+T-funct-LASSO	0.0855 (0.0018)	0.0833 (0.0027)	0.0654 (0.0038)
	LDpred-funct-inf	0.0258 (0.0010)	0.0255 (0.0025)	0.0322 (0.0038)
	LDpred-funct	0.0583 (0.0026)	0.0578 (0.0030)	0.0572 (0.0048)
2,000	P+T	0.0216 (0.0012)	0.0312 (0.0011)	0.0253 (0.0011)
	LDpred-inf	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
	P+T-funct-LASSO	0.0427 (0.0016)	0.0481 (0.0012)	0.0389 (0.0011)
	LDpred-funct-inf	0.0258 (0.0010)	0.0233 (0.0010)	0.0275 (0.0011)
	LDpred-funct	0.0443 (0.0021)	0.0448 (0.0021)	0.0487 (0.0020)
5,000	P+T	-0.0098 (0.0006)	0.0006 (0.0008)	0.0037 (0.0010)
	LDpred-inf	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
	P+T-funct-LASSO	0.0103 (0.0007)	0.0196 (0.0008)	0.0177 (0.0011)
	LDpred-funct-inf	0.0254 (0.0008)	0.0226 (0.0008)	0.026 (0.0009)
	LDpred-funct	0.0339 (0.0015)	0.0336 (0.0019)	0.0377 (0.0019)
10,000	P+T	-0.0172 (0.0007)	-0.0104 (0.0007)	-0.0072 (0.0008)
	LDpred-inf	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
	P+T-funct-LASSO	-0.0024 (0.0007)	0.0046 (0.0008)	0.0027 (0.0009)
	LDpred-funct-inf	0.0262 (0.0008)	0.0230 (0.0008)	0.0250 (0.0007)
	LDpred-funct	0.0311 (0.0015)	0.0288 (0.0016)	0.031 (0.0016)

(b)

# Causal variants	Model	Training sample size 10,000 Diff. $R^2(s.e.)$	20,000 Diff. $R^2(s.e.)$	50,000 Diff. $R^2(s.e.)$
1,000	P+T	-0.004 (0.0029)	-0.0071 (0.0027)	0.0064 (0.0034)
	LDpred-inf	0.0583 (0.0026)	0.0578 (0.003)	0.0572 (0.0048)
	P+T-funct-LASSO	-0.0272 (0.003)	-0.0255 (0.0028)	-0.0082 (0.0035)
	LDpred-funct-inf	0.0325 (0.0028)	0.0323 (0.0025)	0.025 (0.0034)
	LDpred-funct	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
2,000	P+T	0.0227 (0.0024)	0.0136 (0.0023)	0.0234 (0.0023)
	LDpred-inf	0.0443 (0.0021)	0.0448 (0.0021)	0.0487 (0.002)
	P+T-funct-LASSO	0.0017 (0.0026)	-0.0033 (0.0023)	0.0098 (0.0023)
	LDpred-funct-inf	0.0185 (0.0021)	0.0215 (0.002)	0.0212 (0.0022)
	LDpred-funct	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
5,000	P+T	0.0437 (0.0016)	0.033 (0.0018)	0.034 (0.0018)
	LDpred-inf	0.0339 (0.0015)	0.0336 (0.0019)	0.0377 (0.0019)
	P+T-funct-LASSO	0.0237 (0.0016)	0.0139 (0.0018)	0.0201 (0.0019)
	LDpred-funct-inf	0.0086 (0.0015)	0.0109 (0.0017)	0.0118 (0.0018)
	LDpred-funct	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
10,000	P+T	0.0483 (0.0014)	0.0393 (0.0015)	0.0382 (0.0016)
	LDpred-inf	0.0311 (0.0015)	0.0288 (0.0016)	0.031 (0.0016)
	P+T-funct-LASSO	0.0336 (0.0015)	0.0243 (0.0016)	0.0283 (0.0017)
	LDpred-funct-inf	0.0049 (0.0015)	0.0058 (0.0016)	0.006 (0.0017)
	LDpred-funct	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)

Table S25: Sensitivity of LDpred-funct results to number of bins used for regularization in simulations using UK Biobank genotypes. We report results with the number of posterior mean causal effect size bins used for regularization (K) set to 10, 20, 50 or 100. LDpred-funct-K denotes each respective value of K. We also report results for LDpred-funct-inf, which is identical to LDpred-funct with K set to 1. Results are averaged across 100 simulations.

		Training sample size		
# Causal		10,000	20,000	50,000
variants	Model	Average $R^2(s.e.)$	Average $R^2(s.e.)$	Average $R^2(s.e.)$
1,000	LDpred-funct-inf	0.1681 (0.0024)	0.2119 (0.0028)	0.2688 (0.0033)
	LDpred-funct-10	0.1958 (0.002)	0.2402 (0.0019)	0.2937 (0.0019)
	LDpred-funct-20	0.2021 (0.0021)	0.2462 (0.0019)	0.2968 (0.0025)
	LDpred-funct-50	0.2130 (0.0021)	0.2561 (0.0021)	0.3089 (0.0021)
	LDpred-funct-100	0.2243 (0.0022)	0.2647 (0.0025)	0.2976 (0.0074)
2,000	LDpred-funct-inf	0.1697 (0.0022)	0.2135 (0.0026)	0.2703 (0.0030)
	LDpred-funct-10	0.1840 (0.0024)	0.2296 (0.0027)	0.2912 (0.0015)
	LDpred-funct-20	0.1881 (0.0024)	0.2347 (0.0028)	0.2936 (0.0015)
	LDpred-funct-50	0.1978 (0.0025)	0.2439 (0.0028)	0.3005 (0.0017)
	LDpred-funct-100	0.2054 (0.0028)	0.2528 (0.0028)	0.3019 (0.0054)
5,000	LDpred-funct-inf	0.1698 (0.0019)	0.2125 (0.0022)	0.2693 (0.0027)
	LDpred-funct-10	0.1758 (0.0019)	0.2206 (0.0023)	0.2788 (0.0028)
	LDpred-funct-20	0.1783 (0.0019)	0.2232 (0.0023)	0.2809 (0.0028)
	LDpred-funct-50	0.1836 (0.0019)	0.229 (0.0024)	0.2861 (0.0028)
	LDpred-funct-100	0.1899 (0.002)	0.2344 (0.0026)	0.2915 (0.0028)
10,000	LDpred-funct-inf	0.1700 (0.0020)	0.2136 (0.0023)	0.2698 (0.0028)
	LDpred-funct-10	0.1746 (0.0012)	0.2199 (0.0012)	0.2746 (0.0028)
	LDpred-funct-20	0.1750 (0.002)	0.2196 (0.0023)	0.2761 (0.0028)
	LDpred-funct-50	0.1799 (0.002)	0.2240 (0.0024)	0.2800 (0.0028)
	LDpred-funct-100	0.1849 (0.0021)	0.2289 (0.0024)	0.2835 (0.0029)

Table S26: Parameter values for 16 UK Biobank traits. For each trait, we list the training sample size, h_g^2 estimate (from BOLT-LMM v2.3; used by LDpred-inf, LDpred-funct-inf and LDpred-funct) and c parameter (used by LDpred-funct-inf and LDpred-funct).

	Trait	Training N	h_g^2	c
1	Height	408092	0.58	0.45
2	Hair color	403024	0.45	0.23
3	Platelet count	395747	0.40	0.30
4	Bone mineral density	397274	0.40	0.27
5	Red blood cell count	396464	0.32	0.22
6	FEV1-FVC ratio	331786	0.31	0.24
7	Body mass index	407667	0.31	0.28
8	RBC distribution width	394258	0.29	0.20
9	Eosinophil count	391787	0.28	0.19
10	Forced vital capacity	331786	0.28	0.22
11	White blood cell count	395835	0.27	0.22
12	Blood pressure	376437	0.27	0.21
13	Age at menarche	214860	0.26	0.20
14	Tanning ability	400721	0.24	0.09
15	Balding type I	186506	0.22	0.11
16	Waist hip ratio	408196	0.21	0.16

Table S27: Accuracy of 5 polygenic prediction methods across 16 UK Biobank traits. We report results for P+T, LDpred-inf, P+T-funct-LASSO, LDpred-funct-inf and LDpred-funct. Jackknife s.e. for differences vs. LDpred-inf are reported in Table S28. Results for Average across traits are reported in Table S29.

	Trait	h2g	P+T	LDpred-inf	P+T-funct-LASSO	LDpred -funct-inf	LDpred -funct
1	Height	0.579	0.3462	0.3717	0.3667	0.4019	0.4167
2	Hair color	0.454	0.2339	0.2191	0.2389	0.2472	0.2883
3	Platelet count	0.404	0.1994	0.1982	0.2150	0.2290	0.2460
4	Bone mineral density	0.401	0.1871	0.1887	0.1993	0.2105	0.2232
5	Red blood cell count	0.324	0.1247	0.1291	0.1326	0.1572	0.1673
6	FEV1-FVC ratio	0.313	0.1029	0.1139	0.1142	0.1306	0.1345
7	Body mass index	0.308	0.1087	0.1407	0.1189	0.1501	0.1481
8	RBC distribution width	0.288	0.1237	0.1118	0.1346	0.1429	0.1525
9	Eosinophil count	0.277	0.1131	0.1026	0.1189	0.1336	0.1394
10	Forced Vital Capacity	0.277	0.0817	0.1002	0.0935	0.1148	0.1136
11	White blood cell count	0.272	0.0994	0.1054	0.1109	0.1249	0.1282
12	Blood pressure	0.271	0.0802	0.0991	0.0919	0.1111	0.1111
13	Age at menarche	0.255	0.0747	0.0989	0.0899	0.1071	0.1120
14	Tanning ability ability	0.242	0.1405	0.0913	0.1430	0.1234	0.1864
15	Balding type I	0.223	0.1158	0.0874	0.1269	0.1065	0.1235
16	Waist hip ratio	0.210	0.0567	0.0664	0.0645	0.0786	0.0789

Table S28: Differences between polygenic prediction methods across 16 UK Biobank traits. We report results for P+T, LDpred-inf, P+T-funct-LASSO, LDpred-funct-inf and LDpred-funct. We report the difference between R^2 for each method vs. R^2 for LDpred-inf.

	Trait	h^2_g	P+T	LDpred-inf	P+T-funct-LASSO	LDpred-funct-inf	LDpred-funct
1	Height	0.58	-0.0256 (0.0033)	0.0000	-0.0108 (0.0030)	0.0302 (0.0018)	0.0448 (0.0025)
2	Hair color	0.45	0.0148 (0.0038)	0.0000	0.0212 (0.0034)	0.0281 (0.0021)	0.0816 (0.0034)
3	Platelet count	0.40	0.0013 (0.0033)	0.0000	0.0168 (0.0032)	0.0308 (0.0019)	0.0472 (0.0027)
4	Bone mineral density	0.40	-0.0016 (0.0035)	0.0000	0.0106 (0.0030)	0.0217 (0.0016)	0.0342 (0.0024)
5	Red blood cell count	0.32	-0.0044 (0.0033)	0.0000	0.0034 (0.0027)	0.0281 (0.0016)	0.0381 (0.0024)
6	FEV1-FVC ratio	0.31	-0.0110 (0.0035)	0.0000	0.0004 (0.0028)	0.0167 (0.0016)	0.0182 (0.0022)
7	Body mass index	0.31	-0.0320 (0.0025)	0.0000	-0.0242 (0.0024)	0.0094 (0.0014)	0.0077 (0.0016)
8	RBC distribution width	0.29	0.0120 (0.0031)	0.0000	0.0182 (0.0027)	0.0311 (0.0018)	0.0402 (0.0026)
9	Eosinophil count	0.28	0.0105 (0.0031)	0.0000	0.0163 (0.0026)	0.0310 (0.0018)	0.0368 (0.0025)
10	Forced vital capacity	0.28	-0.0185 (0.0029)	0.0000	-0.0067 (0.0025)	0.0146 (0.0015)	0.0101 (0.0018)
11	White blood cell count	0.27	-0.0060 (0.0026)	0.0000	0.0055 (0.0025)	0.0195 (0.0016)	0.0223 (0.0021)
12	Blood pressure	0.27	-0.0189 (0.0026)	0.0000	-0.0071 (0.0024)	0.0120 (0.0014)	0.0117 (0.0018)
13	Age at menarche	0.26	-0.0242 (0.0036)	0.0000	-0.0091 (0.0033)	0.0082 (0.0016)	0.0123 (0.0025)
14	Tanning ability	0.24	0.0492 (0.0033)	0.0000	0.0519 (0.0030)	0.0321 (0.0016)	0.0946 (0.0036)
15	Balding type I	0.22	0.0284 (0.0055)	0.0000	0.0312 (0.0041)	0.0190 (0.0020)	0.0356 (0.0037)
16	Waist hip ratio	0.21	-0.0098 (0.0022)	0.0000	-0.0019 (0.0021)	0.0122 (0.0012)	0.0121 (0.0017)
	Average across traits		-0.0022 (0.0009)	0.0000	0.0072 (0.0008)	0.0215 (0.0004)	0.0342 (0.0006)

Table S29: Accuracy of secondary polygenic prediction methods across 16 UK Biobank traits. For each method, we report the average prediction R^2 across 16 UK Biobank traits. Rows 1-5 correspond to the "Average across traits" panel of Figure 2. Rows 6-8 are methods that analyze only genotyped SNPs (601,728 genotyped SNPs after QC). Rows 9-10 are slightly modified versions of P+T-funct-LASSO. Row 11 uses baseline-LD model functional enrichments that were meta-analyzed across 31 traits. Row 12 uses the baseline model, instead of the baseline-LD model. Row 13 restricts the baseline-LD model to the 6,334,603 SNPs that passed QC filters and were used for prediction. Row 14 infers baseline-LD model parameters using UK10K SNPs, instead of 1000 Genomes SNPs. Row 15 uses UK10K SNPs and uses the baseline-LD+LDAK model, instead of the baseline-LD model.

	Method	Average R^2
1	P+T	0.1368
2	LDpred-inf	0.1390
3	P+T-funct-LASSO	0.1475
4	LDpred-funct-inf	0.1606
5	LDpred-funct	0.1739
6	LDpred-inf (typed)	0.1360
7	LDpred-funct-inf (typed)	0.1378
8	LDpred (typed)	0.1117
9	P+T-funct-LASSO-weighted	0.1549
10	P+T-funct-LASSO (5%)	0.1538
11	LDpred-funct-inf (meta31)	0.1560
12	LDpred-funct-inf(baseline)	0.1573
13	LDpred-funct-inf(QCfilters)	0.1606
14	LDpred-funct-inf(UK10K)	0.1601
15	LDpred-funct-inf(UK10K, baseline-LD+LDAK)	0.1600

Table S30: Sensitivity of LDpred-funct results to number of bins used for regularization across 16 UK Biobank traits. We report results with the number of posterior mean causal effect size bins used for regularization (K) set to 10, 20, 50, 75 or 100. LDpred-funct-K denotes each respective value of K. We also report results for LDpred-funct-inf, which is identical to LDpred-funct with K set to 1. For each trait, the column with highest prediction R^2 is denoted in bold font.

	Trait	LDpred-funct-inf	LDpred-funct-10	LDpred-funct-20	LDpred-funct-50	LDpred-funct-75	LDpred-funct-100
1	Height	0.4019	0.4147	0.4154	0.4153	**0.4161**	0.4152
2	Hair color	0.2472	0.2848	0.2869	**0.2934**	0.2883	0.3035
3	Platelet count	0.2290	0.2448	0.2452	0.2458	**0.2464**	0.2460
4	Bone mineral density	0.2105	0.2213	0.2225	**0.2237**	0.2224	0.2212
5	Red blood cell count	0.1572	0.1669	0.1677	0.1675	0.1681	**0.1682**
6	FEV1-FVC ratio	0.1306	**0.1353**	0.1348	0.1343	0.1336	0.1315
7	Body mass index	0.1501	0.1501	**0.1504**	0.1494	0.1481	0.1473
8	RBC distribution width	0.1429	0.1523	**0.1533**	0.1532	0.1525	0.1508
9	Eosinophil count	0.1336	0.1412	**0.1412**	0.1403	0.1397	0.1386
10	Forced vital capacity	0.1148	**0.1160**	0.1155	0.1145	0.1128	0.1118
11	White blood cell count	0.1249	0.1291	**0.1295**	0.1285	0.1279	0.1262
12	Blood pressure	0.1111	**0.1125**	0.1119	0.1118	0.1108	0.1105
13	Age at menarche	0.1071	0.1118	0.1116	**0.1122**	0.1112	0.1070
14	Tanning ability	0.1234	0.1720	0.1796	0.1858	0.1875	**0.1878**
15	Balding type I	0.1065	0.1217	**0.1235**	0.1220	0.1198	0.1185
16	Waist hip ratio	0.0786	**0.0818**	0.0810	0.0804	0.0798	0.0782
	Average across traits	0.1606	0.1723	0.1731	0.1736	0.1728	0.1726

Table S31: Sensitivity of LDpred-funct results to number of validation samples across 16 UK Biobank traits. We report results with the number of validation samples set to 1,000, 2,000, 5,000, 10,000 (the number of regularization bins is proportional to the number of validation samples; see Equation 2.6. Results are averaged across 20 random subsets of each size. ALL denotes results of LDpred-funct using the total number of validation samples (reported in Table S22). We also report results for LDpred-funct-inf, which is equivalent to LDpred-funct in the limit of a very small number of validation samples.

			Validation sample size				
	Trait	LDpred-funct-inf	1000	2000	5000	10000	ALL
1	Height	0.4019	0.4007 (0.0052)	0.4171 (0.0026)	0.4162 (0.0019)	0.4154 (0.0016)	0.4167
2	Hair color	0.2472	0.2692 (0.0053)	0.2752 (0.0040)	0.2763 (0.0025)	0.2874 (0.0016)	0.3009
3	Platelet count	0.2290	0.2463 (0.0050)	0.2477 (0.0044)	0.2418 (0.0014)	0.2436 (0.0013)	0.2460
4	Bone mineral density	0.2105	0.2235 (0.0049)	0.2219 (0.0033)	0.2232 (0.0017)	0.2247 (0.0013)	0.2232
5	Red blood cell count	0.1572	0.1579 (0.0047)	0.1743 (0.0039)	0.1667 (0.0016)	0.1672 (0.0011)	0.1673
6	FEV1-FVC ratio	0.1306	0.1373 (0.0055)	0.1348 (0.0026)	0.136 (0.0017)	0.1351 (0.0007)	0.1345
7	Body mass index	0.1501	0.1596 (0.0055)	0.1501 (0.0034)	0.1482 (0.0018)	0.1491 (0.0011)	0.1481
8	RBC distribution width	0.1429	0.1598 (0.0052)	0.1503 (0.0028)	0.1492 (0.0016)	0.1519 (0.0012)	0.1525
9	Eosinophil count	0.1336	0.1492 (0.0052)	0.1439 (0.0042)	0.1402 (0.0014)	0.1406 (0.001)	0.1394
10	Forced vital capacity	0.1148	0.1198 (0.0031)	0.1196 (0.0029)	0.1152 (0.0015)	0.1139 (0.001)	0.1136
11	White blood cell count	0.1249	0.1322 (0.0040)	0.1335 (0.0036)	0.1249 (0.0018)	0.1289 (0.0012)	0.1282
12	Blood pressure	0.1111	0.1170 (0.0033)	0.1114 (0.0020)	0.1112 (0.0013)	0.1100 (0.0009)	0.1111
13	Age at menarche	0.1071	0.1175 (0.0040)	0.1139 (0.0029)	0.1102 (0.0013)	0.1112 (0.0011)	0.1120
14	Tanning ability	0.1234	0.1397 (0.0045)	0.1429 (0.0029)	0.1703 (0.0020)	0.1833 (0.0011)	0.1864
15	Balding type I	0.1065	0.1218 (0.0038)	0.1176 (0.0025)	0.1209 (0.0013)	0.1228 (0.0003)	0.1235
16	Waist hip ratio	0.0786	0.0866 (0.0031)	0.0811 (0.0023)	0.0791 (0.0019)	0.0790 (0.0008)	0.0789
17	Average across traits	0.1606	0.1711	0.1710	0.1706	0.1728	0.1739

Table S32: Accuracy of 5 prediction methods in height meta-analysis of UK Biobank and 23andMe cohorts. We report results for P+T, LDpred-inf, P+T-funct-LASSO, LDpred-funct-inf and LDpred-funct, for each of 4 training data sets: UK Biobank interim release (113,660 training samples), UK Biobank (408,092 training samples), 23andMe (698,430 training samples) and meta-analysis of UK Biobank and 23andMe (1,107,430 training samples). We also report results for a fixed-effect meta-analysis of UK Biobank and 23andMe.

Data Set	Training N	P+T	LDpred-inf	P+T-funct -LASSO	LDpred -funct-inf	LDpred -funct
UK Biobank interim release	113,660	0.2223	0.2305	0.2524	0.2777	0.2926
UK Biobank	408,092	0.3448	0.3677	0.3644	0.3995	0.4132
23andMe	698,430	0.2903	0.2882	0.2985	0.3148	0.3279
Meta-analysis of UK Biobank and 23andMe	1,107,430	0.3710	0.3874	0.3778	0.4193	0.4292
Fixed-effect meta-analysis	1,107,430	0.3687	0.3653	0.3663	0.3965	0.4051

Table S33: Accuracy of LDpred-funct-inf(1000G), LDpred-funct-inf(UK10K) and LDpred-funct-inf(UK10K, baseline-LD+LDAK) across 16 UK Biobank traits. We report results for each trait. Results for Average across traits are reported in Table S29.

	Trait	h_g^2	LDpred-funct-inf under different priors: baselineLD (1000G)	baselineLD (UK10K)	baselineLD + LDAK (UK10K)
1	Height	0.579	0.4019	0.4011	0.4018
2	Hair color	0.454	0.2472	0.2501	0.2501
3	Platelet count	0.404	0.2290	0.2294	0.2298
4	Bone mineral density	0.401	0.2105	0.2122	0.2117
5	Red blood cell count	0.324	0.1572	0.1566	0.1544
6	FEV1-FVC ratio	0.313	0.1306	0.1309	0.1323
7	Body mass index	0.308	0.1501	0.1503	0.1502
8	RBC distribution width	0.288	0.1429	0.1432	0.1451
9	Eosinophil count	0.277	0.1336	0.1335	0.1342
10	Forced vital capacity	0.277	0.1148	0.1147	0.1140
11	White blood cell count	0.272	0.1249	0.1246	0.1251
12	Blood pressure	0.271	0.1111	0.1113	0.1136
13	Age at menarche	0.255	0.1071	0.0995	0.0930
14	Tanning ability	0.242	0.1234	0.1206	0.1190
15	Balding type I	0.223	0.1065	0.1040	0.1070
16	Waist hip ratio	0.210	0.0786	0.0793	0.0785

Appendix C

Summary statistic based extension of mixed model association method to increase meta-analysis power

Supplementary Figures

Step 1).
for each cohort c **do**
1a) Compute summary association statistics using BOLT-LMM, call them SS_c.
1b) Share with other cohorts the summary statistics SS_c.

Step 2)
for each cohort c **do**
2a) i. Meta-analyze summary statistics using all cohorts except for summary statistics from cohort c. Call these summary statistics SS_{meta_c}.
for each chromosome chr **do**
2a) ii. Compute a PRS_{c-chr} using recommended prediction method, using summary statistics SS_{meta_c} from all chromosomes except from chromosome chr.
Compute residual $Y_{residual-c-chr}$, where $Y_{residual-c-chr} = Y - \widehat{\alpha} PRS_{c-chr}$.
2b) Compute summary association statistics using $Y_{residual-c-chr}$ as outcome. Call these summary statistics $SS'_{meta_{c-chr}}$.

Share summary association statistics SS'_{meta_c} across cohorts.

Step 3) Meta-analyze summary association statistics SS'_{meta_c} using all cohort, and call them $SS_{Meta-LMM}$.

Figure S4: **Pseudocode to compute Meta-LMM summary statistics.** In this figure, we show the necessary steps to get Meta-LMM summary statistics.

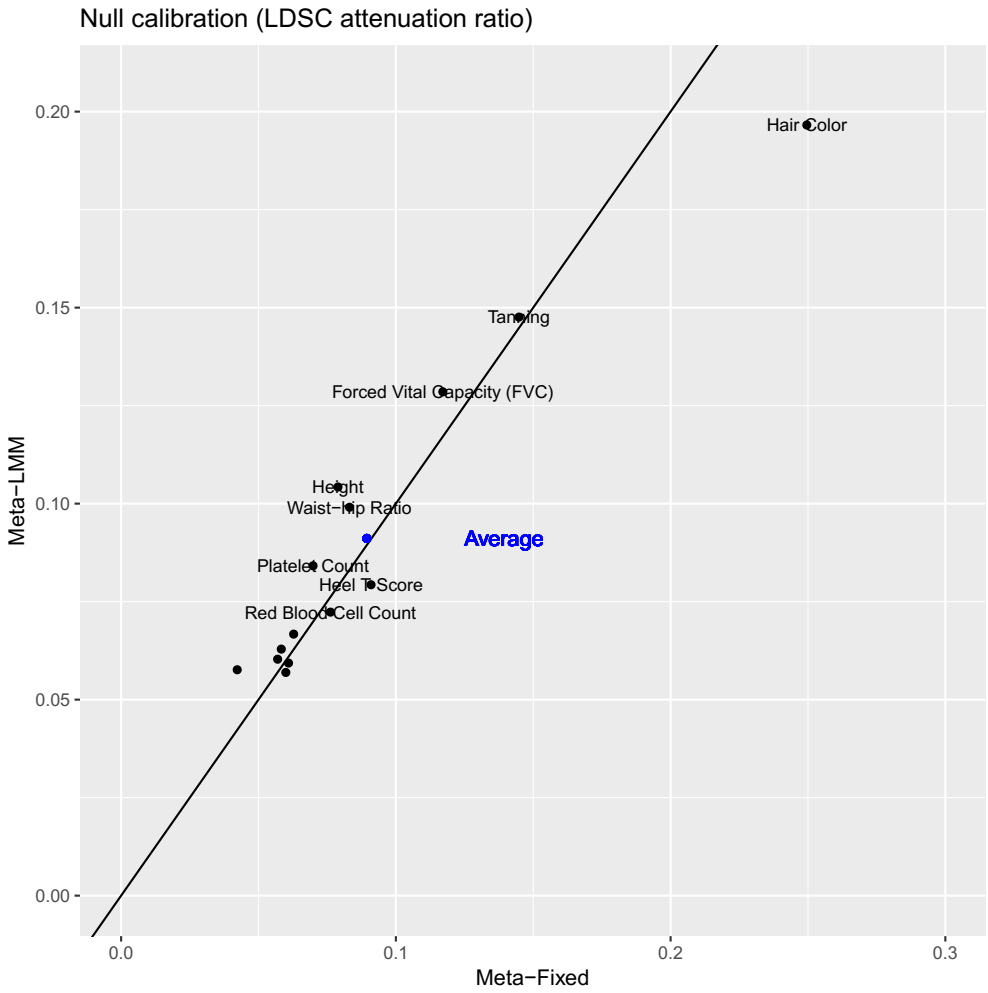


Figure S5: Test statistic for calibration of Meta-LMM vs. Meta-Fixed when applied to 14 UK Biobank traits. We compare the attenuation ratios from LD score regresssion, numerical values are reported in Table S39.

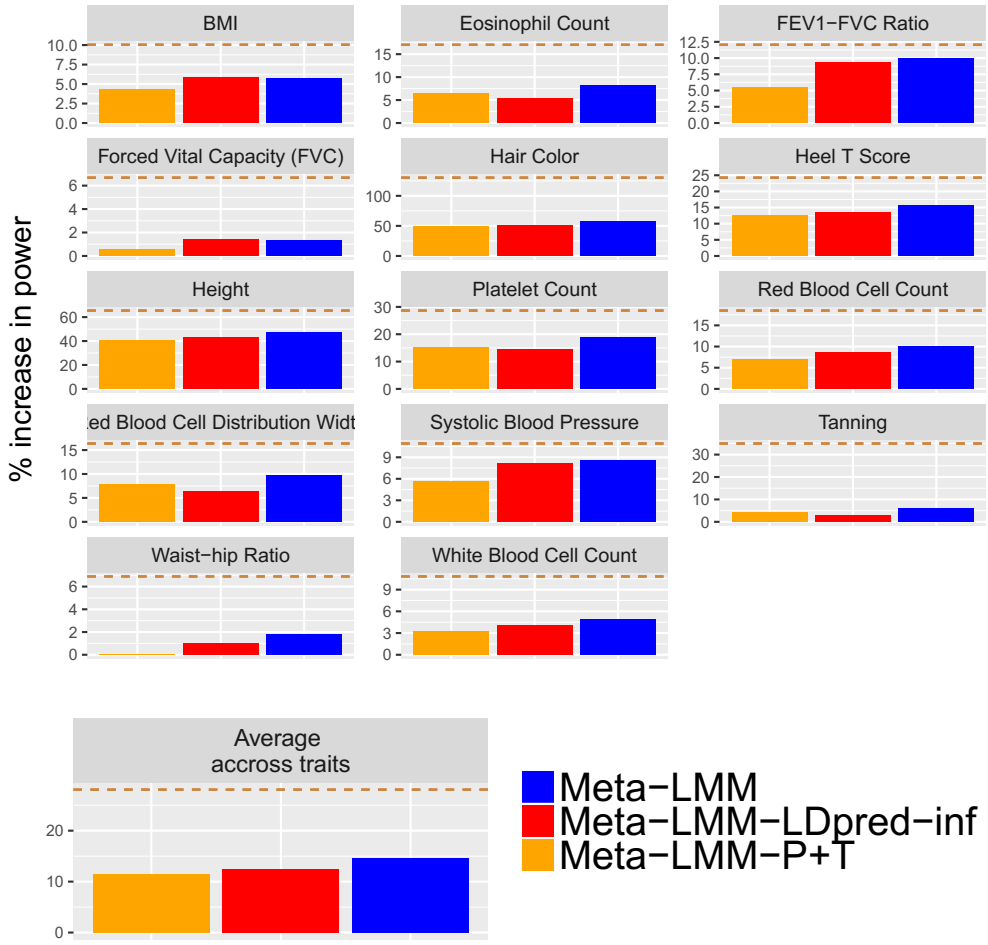


Figure S6: Power analyses for 3 meta-analyses methods relative to fixed-effects meta-analysis when applied to 14 UK Biobank. We report percent increases in χ^2 statistics defined as the median of ratios between χ^2 statistics estimated using Method X and Meta-Fixed, where method-X can be Meta-LMM, Meta-LMM-P+T, and Meta+LMM+LDpred-inf. We restrict calculations to SNPs that have $\chi^2 > 30$ in an additional independet sample of 120K Brish Europeans. We also report in Table S42 analogous results using BOLT-LMM-inf and BOLT-LMM, wich respesents the best case scenenario for increasing association power. Golden dashed line represents the boost in power obtained using BOLT-LMM. Numerical values are in Table S42.

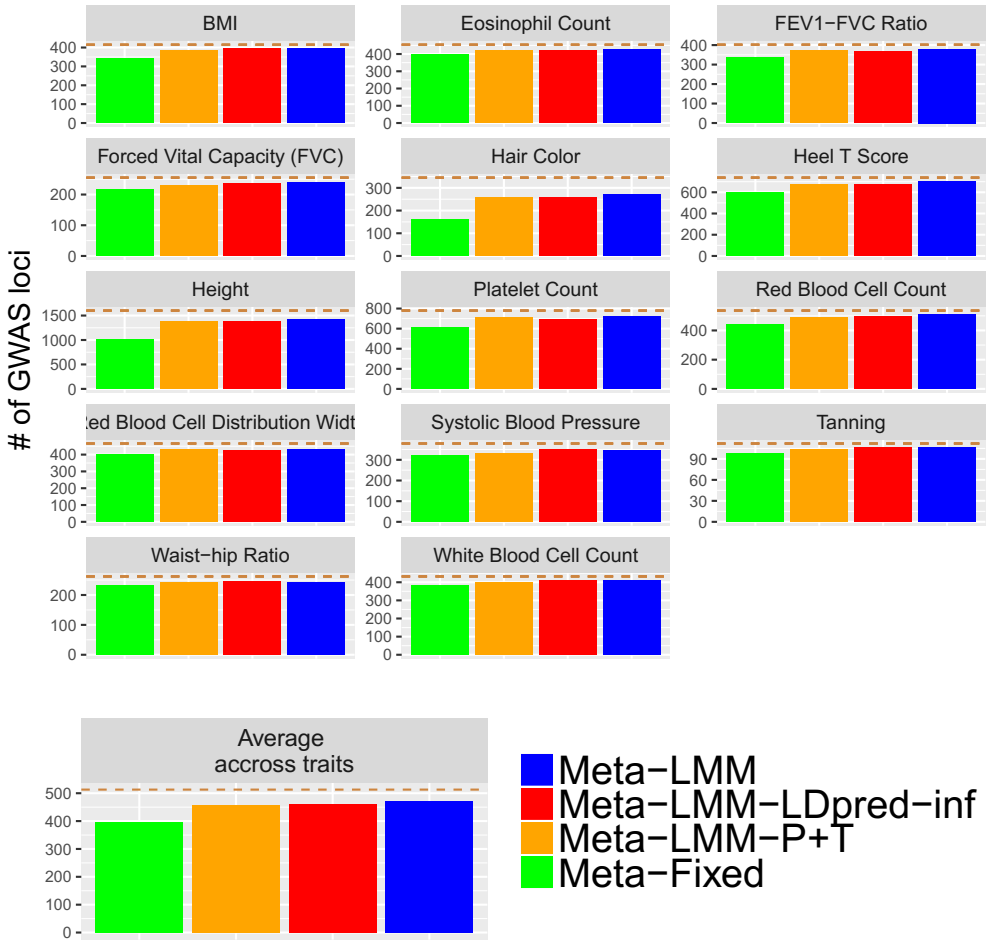


Figure S7: Number of independent genome-wide significant associations ($p < 5\times10^8$) identified by 4 meta-analysis methods applied to 14 traits in the UK Biobank. We obtain the number on indpendent loci using LD-cumpling, and also report the average accross traits. Dashed golden bar is the number of independendent GWAS loci obtained using BOLT-LMM. Numerical values are in Table S43.

Supplementary Tables

Table S34: List of 14 UK Biobank traits. We list the total sample size for each trait.

	trait	N
1	Eosinophil Count	323392
2	Platelet Count	326702
3	Red Blood Cell Distribution Width	325480
4	Red Blood Cell Count	327286
5	White Blood Cell Count	326797
6	Heel T Score	327848
7	BMI	336458
8	Height	336816
9	Waist-hip Ratio	336902
10	Systolic Blood Pressure	310820
11	FEV1-FVC Ratio	274961
12	Forced Vital Capacity (FVC)	274961
13	Hair Color	332690
14	Tanning	330797

Table S35: Null calibration analyses of 5 meta-analysis methods in simulations using UK Biobank genotypes, for 2 different genetics architectures. We report the average χ^2 statistic of SNPs in restricted to even chromosomes obtained using the following methods: Meta-Fixed, Meta-LMM-P+T, Meta-LMM-LDpred-inf, Meta-LMM, Meta-LMM-True. Results are averaged across 20 simulations.

p	SNP set	Meta-Fixed	Meta-LMM-P+T	Meta-LMM-LDpred-inf	Meta-LMM	Meta-LMM-True
0.1%	even chromosomes	1.004 (0.002)	1.000 (0.002)	1.001 (0.002)	1.001 (0.002)	1.000 (0.002)
5%	even chromosomes	1.002 (0.001)	1.002 (0.001)	1.000 (0.001)	1.000 (0.001)	0.997 (0.001)

Table S36: Relationship between prediction R^2 and $\%$ increase in power of 3 meta-analysis methods compared to Meta-Fixed in simulations using UK Biobank genotypes, for 2 different genetics architectures. We report the prediction R^2 obtained using 3 different prediction methods (P+T, LDpred-inf and P+T + LDpred-inf). We optimal weights for each PRS in the joint model P+T + LDpredinf are 0.56 (resp. 0.325) for P+T and 0.081 (resp. 0.212) for simulated genetic architecture with 0.1% (resp. 5%) causal variants. We report the ratio of the average χ^2 statistic of method X vs Meta-Fixed, where method X can be Meta-LMM-LDpred-inf, Meta-LMM-P+T and Meta-LMM. Results are averaged across 20 simulations.

p	models	% increase in power	R^2
0.1%	Meta-LMM-LDpred-inf	14.335	0.201
	Meta-LMM-P+T	35.022	0.376
	Meta-LMM	35.854	0.380
	Meta-LMM-TRUE	50.297	0.498
5%	Meta-LMM-LDpred-inf	14.839	0.205
	Meta-LMM-P+T	18.524	0.232
	Meta-LMM	20.747	0.253
	Meta-LMM-TRUE	64.570	0.498

Table S37: Power analyses of 5 meta-analysis methods in simulations using UK Biobank genotypes, for 2 different genetics architectures. We report results for average χ^2 statistics restricted to the 3 different SNP sets obtained using the following methods: Meta-Fixed, Meta-LMM-P+T, Meta-LMM-LDpred-inf, Meta-LMM, Meta-LMM-True. Results are averaged across 20 simulations.

p	SNP set	Meta-Fixed	Meta-LMM-P+T	Meta-LMM-LDpred-inf	Meta-LMM	Meta-LMM-True
0.1%	true effects	169.117	192.422	227.057	228.722	254.179
	genome-wide significant	113.067	132.190	164.300	165.512	196.262
	odd chromosomes	3.578	3.992	4.660	4.690	5.279
5%	true effects	8.888	10.206	10.533	10.730	14.627
	genome-wide significant	50.295	57.002	58.742	59.903	81.119
	odd chromosomes	3.634	4.067	4.176	4.242	5.553

Table S38: Null calibration analyses of 4 meta-analysis methods in simulations using UK Biobank genotypes assumin 0.1% of causal variants. We report the average χ^2 statistic of SNPs in restricted to even chromosomes over 20 simulations. Meta-Fixed (linreg. + X PCs) and Meta-LMM-LDpredinf (linreg. + X PCs) refers to methods were in step 1a) we use linear regression plus X PCs to compute association statistics.

Method	Average χ^2 in even chromosomes (s.e.)
Meta-Fixed (BOLT-LMM default)	1.004 (0.001)
Meta-Fixed (LR + 10 PCs)	1.002 (0.002)
Meta-Fixed (LR + 20 PCs)	1.002 (0.002)
Meta-LMM (BOLT-LMM default)	1.001 (0.003)
Meta-LMM-LDpredinf (LR + 10 PCs)	1.024 (0.008)
Meta-LMM-LDpredinf (LR + 20 PCs)	1.001 (0.002)

Table S39: Calibration analysis using LDSC attenuation ratio using 6 association methods when applied to 14 UK Biobank traits. LDSC attenuation is defined as $(\text{LDSC intercept} - 1)/\text{mean}\chi^2 - 1$. We include attenuation ratio obtained for BOLT-LMM and BOLT-LMM-inf for comparison purposes.

	Trait	Meta-Fixed	Meta-LMM-P+T	Meta-LMM-LDpred-inf	Meta-LMM	BOLT-LMM-inf	BOLT-LMM
1	Eosinophil Count	0.058	0.064	0.063	0.063	0.061	0.067
2	Platelet Count	0.070	0.082	0.076	0.084	0.076	0.074
3	Red Blood Cell Distribution Width	0.057	0.062	0.062	0.060	0.055	0.051
4	Red Blood Cell Count	0.076	0.073	0.076	0.072	0.076	0.065
5	White Blood Cell Count	0.060	0.055	0.055	0.057	0.063	0.066
6	Heel T Score	0.091	0.078	0.084	0.079	0.089	0.083
7	BMI	0.061	0.054	0.053	0.059	0.056	0.058
8	Height	0.079	0.104	0.101	0.104	0.104	0.094
9	Waist-hip Ratio	0.083	0.093	0.091	0.099	0.084	0.086
10	Systolic Blood Pressure	0.063	0.063	0.064	0.067	0.067	0.065
11	FEV1-FVC Ratio	0.042	0.055	0.055	0.058	0.048	0.042
12	Forced Vital Capacity (FVC)	0.117	0.124	0.123	0.129	0.099	0.097
13	Hair Color	0.250	0.212	0.221	0.197	0.224	0.226
14	Tanning	0.145	0.137	0.142	0.148	0.194	0.202
15	Average across traits	0.089	0.090	0.090	0.091	0.093	0.091

Table S40: Power analyses for 5 association methods relative to fixed-effects meta-analysis when applied to 14 UK Biobank. We report percent increases in χ^2 statistics defined as the median of ratios between χ^2 statistics estimated using Method X and Meta-Fixed, where method-X can be Meta-LMM, Meta-LMM-P+T, and Meta+LMM+LDpred-inf, BOLT-LMM-inf and BOLT-LMM. We restrict calculations to SNPs that have $\chi^2 > 30$ accross all the 6 methods being compared. We include results for BOLT-LMM-inf and BOLT-LMM for comparison purposes, as these methods assumes that individual-level data for the 10 cohorts can be analyzed together.

	trait	Meta-LMM-P+T	Meta-LMM-LDpred-inf	Meta-LMM	BOLT-LMM-inf	BOLT-LMM
1	Eosinophil Count	1.0869	1.0806	1.1062	1.086	1.1439
2	Platelet Count	1.1436	1.1243	1.1764	1.1379	1.2235
3	Red Blood Cell Distribution Width	1.085	1.0641	1.0983	1.0785	1.1477
4	Red Blood Cell Count	1.0907	1.0938	1.1176	1.0997	1.1348
5	White Blood Cell Count	1.0684	1.084	1.0906	1.0698	1.0807
6	Heel T Score	1.0949	1.1048	1.1339	1.141	1.2052
7	BMI	1.0632	1.076	1.0768	1.075	1.0885
8	Height	1.3531	1.3768	1.4074	1.48	1.5576
9	Waist-hip Ratio	1.0189	1.0151	1.0311	1.0504	1.0742
10	Systolic Blood Pressure	0.9707	0.9865	0.9908	1.0584	1.0708
11	FEV1-FVC Ratio	1.0645	1.0855	1.0952	1.0862	1.114
12	Forced Vital Capacity (FVC)	1.0255	1.0422	1.0436	1.0833	1.0891
13	Hair Color	1.5926	1.6017	1.6947	1.6784	2.0243
14	Tanning	1.072	1.0475	1.0841	1.044	1.1478
15	Average across traits	1.1236	1.1274	1.1533	1.1549	1.2216

Table S41: Accuracy of 3 different polygenic prediction methods used in the residualization step of Meta-LMM applied to 14 UK Biobank traits. We report the optimal weights assign to each normalized PRS when modelling jointy P+T and LDpred-inf.

	trait	P+T	LDpred-inf	P+T+LDpred-inf	P+T PRS weight	LDpred-inf PRS weight
1	Eosinophil Count	0.099	0.092	0.112	0.199	0.162
2	Platelet Count	0.185	0.174	0.215	0.273	0.233
3	Red Blood Cell Distribution Width	0.121	0.101	0.137	0.245	0.164
4	Red Blood Cell Count	0.12	0.119	0.139	0.203	0.199
5	White Blood Cell Count	0.082	0.088	0.100	0.154	0.188
6	Heel T Score	0.165	0.164	0.200	0.248	0.244
7	BMI	0.09	0.109	0.110	0.076	0.265
8	Height	0.262	0.287	0.309	0.238	0.349
9	Waist-hip Ratio	0.05	0.054	0.069	0.14	0.161
10	Systolic Blood Pressure	0.066	0.082	0.083	0.064	0.232
11	FEV1-FVC Ratio	0.084	0.096	0.112	0.161	0.21
12	Forced Vital Capacity (FVC)	0.059	0.076	0.077	0.005	0.272
13	Hair Color	0.208	0.21	0.262	0.283	0.291
14	Tanning	0.095	0.073	0.108	0.228	0.141
15	Average across traits	0.120	0.123	0.145	0.180	0.222

Table S42: Power analyses for 5 association methods relative to fixed-effects meta-analysis when applied to 14 UK Biobank. We report percent increases in χ^2 statistics defined as the median of ratios between χ^2 statistics estimated using Method X and Meta-Fixed, where method-X can be Meta-LMM, Meta-LMM-P+T, and Meta+LMM+LDpred-inf, BOLT-LMM-inf and BOLT-LMM. We restrict calculations to SNPs that have $\chi^2 > 30$ in an additional independet sample of 120K Brish Europeans. We include results for BOLT-LMM-inf and BOLT-LMM for comparison purposes, as these methods assumes that individual-level data for the 10 cohorts can be analyzed together.

	trait	Meta-LMM-P+T	Meta-LMM-LDpred-inf	Meta-LMM	BOLT-LMM-inf	BOLT-LMM
1	Eosinophil Count	1.0638	1.054	1.0814	1.1073	1.1706
2	Platelet Count	1.1533	1.144	1.1899	1.1936	1.2866
3	Red Blood Cell Distribution Width	1.0781	1.0644	1.0977	1.0865	1.1637
4	Red Blood Cell Count	1.0695	1.0868	1.0998	1.1422	1.1849
5	White Blood Cell Count	1.0318	1.0408	1.049	1.0819	1.1082
6	Heel T Score	1.1244	1.1361	1.1574	1.1726	1.2426
7	BMI	1.0427	1.0586	1.0575	1.0925	1.1005
8	Height	1.4061	1.4329	1.4692	1.5725	1.6547
9	Waist-hip Ratio	1.0006	1.0102	1.0178	1.0437	1.0689
10	Systolic Blood Pressure	1.0562	1.0822	1.0855	1.0956	1.1092
11	FEV1-FVC Ratio	1.0552	1.0928	1.0989	1.1042	1.1206
12	Forced Vital Capacity (FVC)	1.0058	1.0137	1.0136	1.0567	1.0668
13	Hair Color	1.4849	1.504	1.566	1.9177	2.3029
14	Tanning	1.0433	1.0307	1.0602	1.2304	1.3495
15	Average across traits	1.1154	1.1251	1.146	1.207	1.2807

Table S43: Number of independent GWAS loci obtained using 6 different association methods when applied to 14 traits in the UK Biobank. We obtain the number on indpendent loci using LD-cumpling.

Trait	Meta-Fixed	Meta-LMM-P+T	Meta-LMM-LDpred-inf	Meta-LMM	BOLT-LMM-inf	BOLT-LMM
Waist-hip Ratio	232	242	245	242	253	262
Tanning	98	104	106	106	106	112
White Blood Cell Count	381	401	410	410	421	432
Forced Vital Capacity (FVC)	215	230	237	239	256	255
Eosinophil Count	395	419	423	428	427	454
Systolic Blood Pressure	322	332	351	348	372	379
FEV1-FVC Ratio	336	370	369	380	379	402
BMI	342	383	395	393	401	415
Red Blood Cell Distribution Width	403	430	427	434	444	466
Red Blood Cell Count	443	490	494	509	509	536
Heel T Score	600	672	677	706	695	738
Platelet Count	612	712	694	726	715	780
Hair Color	162	258	256	269	279	345
Height	1012	1387	1383	1426	1513	1602
Average across traits	397	459	462	473	484	513