Explicit modeling of ancestry improves polygenic risk scores and BLUP prediction

Chia-Yen Chen[1,*], **Jiali Han**[1,2,3], **David J. Hunter**[1,2,4], **Peter Kraft**[1,2,4,5], and **Alkes L. Price**[1,4,5,*]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115

[2]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115

[3]Department of Epidemiology, Richard M. Fairbanks School of Public Health, Simon Cancer Center, Indiana University, Indianapolis, Indiana 46202

[4]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142

[5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115

Abstract

Polygenic prediction using genome-wide SNPs can provide high prediction accuracy for complex traits. Here, we investigate the question of how to account for genetic ancestry when conducting polygenic prediction. We show that the accuracy of polygenic prediction in structured populations may be partly due to genetic ancestry. However, we hypothesized that explicitly modeling ancestry could improve polygenic prediction accuracy. We analyzed three GWAS of hair color, tanning ability and basal cell carcinoma (BCC) in European Americans (sample size from 7,440 to 9,822) and considered two widely used polygenic prediction approaches: polygenic risk scores (PRS) and Best Linear Unbiased Prediction (BLUP). We compared polygenic prediction without correction for ancestry to polygenic prediction with ancestry as a separate component in the model. In 10-fold cross-validation using the PRS approach, the R^2 for hair color increased by 66% (0.0456 to 0.0755; $p<10^{-16}$), the R^2 for tanning ability increased by 123% (0.0154 to 0.0344; $p<10^{-16}$) and the liability-scale R^2 for BCC increased by 68% (0.0138 to 0.0232; $p<10^{-16}$) when explicitly modeling ancestry, which prevents ancestry effects from entering into each SNP effect and being over-weighted. Surprisingly, explicitly modeling ancestry produces a similar improvement when using the BLUP approach, which fits all SNPs simultaneously in a single variance component and causes ancestry to be underweighted. We validate our findings via simulations, which show that the differences in prediction accuracy will increase in magnitude as sample sizes increase. In summary, our results show that explicitly modeling ancestry can be important in both PRS and BLUP prediction.

*Corresponding authors: Alkes L. Price, Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, 665 Huntington Avenue Building 2, Room 211, Boston, Massachusetts 02115., Phone: 617-432-2262, aprice@hsph.harvard.edu. Chia-Yen Chen, Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street, Boston, MA 02114., Phone: 617-643-5405, cychen@mail.harvard.edu.

Introduction

Genome-wide association studies (GWAS) have identified many single nucleotide polymorphisms (SNPs) associated with complex traits [Visscher et al. 2012]. Genetic prediction based on genome-wide significant SNPs from GWAS provides some predictive ability [Wray et al. 2007; Meigs et al. 2008; Kraft and Hunter 2009; So et al. 2011], but polygenic prediction, using genetic variants that do not achieve genome-wide significance, can improve prediction accuracy [de los Campos et al. 2010; Chatterjee et al. 2013; Dudbridge 2013]. Polygenic prediction has been applied to many complex traits, such as schizophrenia and height [Evans et al. 2009; Purcell et al. 2009; Wei et al. 2009; Bush et al. 2010; Lango Allen et al. 2010; Machiela et al. 2011; Makowsky et al. 2011; Stahl et al. 2012; Vazquez et al. 2012; Wei et al. 2013; Smoller et al. 2013; Rietveld et al. 2013; Ripke et al. 2014; Wood et al. 2014].

Polygenic prediction accuracy may be partly due to genetic ancestry, especially for complex traits that are known to be associated with ancestry. For example, differences in height between Northern Europeans and Southern Europeans (e.g. due to natural selection) explain 5% of the variance of height in European Americans from the Framingham Heart Study, accounting for most of the polygenic prediction accuracy in 1,997 unrelated individuals from this sample [Makowsky et al. 2011; Wray et al. 2013; Turchin et al. 2012]. Other traits that are well-known to be associated with ancestry in European-ancestry populations include hair color, eye color, skin pigmentation, skin cancer, multiple sclerosis, rheumatoid arthritis, type 1 diabetes, Crohn's disease, Alzheimer's disease, coagulation factor VII (FVII) plasma level, hemoglobin disorders, and educational attainment [Candille et al. 2012; Nan et al. 2009a; Rosati 2001; Cimmino et al. 1998; Patterson et al. 2009; Kenny et al. 2012; Panza et al. 2003; Bernardi et al. 1997; Angastiniotis and Modell 1998; Borjas 1994]. Thus, it is possible that polygenic prediction studies of these traits may likewise owe some of their accuracy to genetic ancestry [Evans et al. 2009; Wei et al. 2009; Bush et al. 2010; Makowsky et al. 2011; Stahl et al. 2012; Vazquez et al. 2012; Wei et al. 2013; Rietveld et al. 2013]. In genetic association studies, ancestry is a confounder that may cause false positive associations with no biological significance [Price et al. 2010]. On the other hand, if the ultimate goal is to conduct prediction in the same population, polygenic predictions that are partly due to ancestry may be useful. However, it is important to understand how much of polygenic prediction accuracy is due to ancestry, as this will affect prediction accuracy in other populations.

An open question is how to make use of ancestry information in polygenic prediction. Most studies of polygenic prediction ignore ancestry information, but this may not be optimal. We propose to correct for ancestry in the polygenic model using principal components, while incorporating associations between ancestry and the trait as a separate component in the

prediction model. We hypothesize that explicit modeling of ancestry in this fashion can improve polygenic prediction accuracy, by preventing ancestry from entering into each SNP effect and being overweighted. In this study, we investigated the impact of explicit modeling of genetic ancestry on polygenic prediction accuracy by analyzing GWAS samples of European Americans for three pigmentation-related traits, including natural hair color (HC), childhood and adolescent tanning ability (TA), and basal cell carcinoma (BCC). We employed cross-validation schemes to ensure that improvements in prediction accuracy were not caused by over-fitting. We determined that explicitly modeling ancestry can significantly improve prediction accuracy across different polygenic prediction approaches, such as polygenic risk scores and best linear unbiased predictors. We note that although polygenic risk scores may not be the optimal prediction method when raw genotypes are available, we have chosen to include them as one of the main focuses of our analyses, because they are currently the most widely used approach for polygenic risk prediction in humans.

Materials and Methods

GWAS data

We performed analyses of 3 genome-wide association studies (GWAS) of hair color (HC), tanning ability (TA), and basal cell carcinoma (BCC) nested in the Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS). The sample collection, genotyping and quality control for each GWAS were described previously [Hunter et al. 2007; Nan et al. 2011; Qi et al. 2010; Rimm et al. 1991; Curhan and Taylor 2008; Taylor et al. 2005]. All samples are self-reported European Americans. We included only SNPs genotyped with call rate >99% and minor allele frequency >1% and excluded A/T and C/G SNPs to avoid strand ambiguity. The missing genotypes were mean-imputed in the subsequent analyses. We extracted samples with complete phenotype and self-reported ancestry information into 3 GWAS sets. We obtained 7,440 samples with HC, 9,822 samples with TA, and 2,086 BCC cases and 6,173 controls. We applied LD pruning by genomic position with a R^2 threshold of 0.2 to obtain 71,557, 71,549 and 71,575 independent SNPs for HC, TA, and BCC, respectively.

In addition to the genotyped SNPs, we obtained imputed genotypes for variants that showed genome-wide significant association with HC, TA and BCC. This included 12 SNPs and one two-SNP haplotype associated with each of HC, TA and BCC and 6 additional SNPs associated with BCC only (Table S1) [Gerstenblith et al. 2010]. Note that some of these genome-wide significant SNPs were identified in the samples analyzed here, especially for TA; therefore, higher prediction accuracy based these genome-wide significant SNPs is expected [Han et al. 2008; Nan et al. 2009a; Nan et al. 2009b]. The imputation was done with using BEAGLE (for phasing) and MaCH using the 1000 Genomes Project Interim Phase I release as the reference panel [The 1000 Genomes Project Consortium 2012; Li et al. 2010; Browning and Browning 2007].

Phenotype definitions

The phenotype information was collected from prospective questionnaires. The natural hair color at 20 years of age was collected in 4 levels: blonde, light brown, dark brown, and black. We excluded samples with red hair color, which is a distinct trait from other hair colors and is mainly determined by the *MC1R* gene [Han et al. 2008]. The childhood and adolescent tanning ability with more than 2 hours of sun light exposure was collected in 4 levels: practically none, light tan, average tan, and deep tan. A previous study showed that deeper tan denotes a better response to sun light exposure and is associated with lower risk of skin cancer [Nan et al. 2009a]. Both hair color and tanning ability were analyzed as quantitative traits. The BCC cases were identified through biennial self-report; the validation rate of BCC self-reports was 90% in NHS and HPFS [Han et al. 2006; Hunter et al. 1990].

Self-reported ancestry

The participants were assigned to 1 of the 4 self-reported ancestry groups: Scandinavian, Southern European, Ashkenazi Jewish, and Other European (which includes samples with no self-reported European sub-population ancestry information). We calculated the mean and standard deviation for each trait by self-reported ancestry groups. We also tested for association between self-reported ancestry and each trait with a Wald test for HC and TA and a χ^2 test and odds ratios with 95% confidence intervals for BCC.

Inference of genetic ancestry

To infer genetic ancestry, we performed principal component analysis (PCA) with EIGENSTRAT software to obtain principal components as measure of genetic ancestry [Price et al. 2006; Patterson et al. 2006]. We performed PCA on the 3 GWAS and obtained the top 10 PCs specific to each GWAS. These PCs were used in all primary analyses. We also performed additional analyses in which we derived the top PCs independently for training and validation sets with an external European American reference panel (see Discussion). We plotted PC1 and PC2 to show the structure of these GWAS samples with color-coding by self-reported ancestry. To further investigate population structure, we projected the samples from Eastern European countries in the Population Reference Sample (POPRES) to the PC space derived from the TA GWAS samples, which contained most of our merged GWAS samples. The collections and methods for the Population Reference Sample (POPRES) are described by Nelson *et al.* (2008). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgistudy_id=phs000145.v1.p1 through dbGaP accession number phs000145.v1.p1. The PC projection was done by applying the SNP weights, which were calculated with principal components and eigenvalues derived from the TA GWAS samples, to the POPRES samples [Chen et al. 2013]. To compare the structure of POPRES samples and our GWAS samples, we plotted the projected PC1 and PC2 of POPRES samples with the PC1 and PC2 of the TA GWAS samples.

Prediction using genetic and self-reported ancestry

We demonstrated the prediction ability of ancestry by building prediction models with 1) self-reported ancestry, 2) PC1 and PC2, 3) the top 5 PCs, and 4) the top 5 PCs plus self-

reported ancestry in all samples. To control for over-fitting, we also employed 10-fold cross-validation: we built prediction models by estimating model coefficients with the training set which includes 90% of the total sample size and applied the estimated model coefficients to the validation set, which includes 10% of the total sample size excluded from the training set, and calculated predicted phenotypes for the validation samples. A predicted phenotype was calculated for each validation sample as $\hat{Y}=\sum_{k=1}^{K}\beta_k x_k$ where k indexes the predictors, including PCs and/or self-reported ancestry, β_k is the regression coefficient estimated in the training set and x_k is the covariate value for the validation set. We repeated this process 10 times and pooled the validation samples with predicted phenotypes to calculate out-of-sample R^2 between predicted and true phenotypes. We note that the PCs used in all primary analyses were derived with all samples together.

R^2 was calculated for each model and a likelihood ratio test (LRT) was used to compare models. We used R^2 on the observed scale for HC and TA and R^2 on the liability scale for BCC. The R^2 on the observed scale were obtained by fitting linear regression models with HC and TA as continuous outcomes and with self-reported ancestry and PCs as predictors. To obtain the R^2 on liability scale for BCC, we first obtained R^2 on the observed scale for BCC by treating the BCC case-control status as a continuous outcome where all BCC cases take value 1 and controls take value 0 and fitted linear regression models with the continuous BCC outcome and the predictors. We then transformed the R^2 on the observed scale for BCC to the R^2 on the liability scale by applying the method based on the liability threshold model described in Lee et al. 2012. The transformation of observed scale R^2 to liability scale R^2 assumes a disease liability that follows the standard normal distribution and assumes all disease cases have disease liability exceeding a certain threshold value. We determined the threshold based on the population prevalence of the disease estimated in previous studies [Lear and Smith 1997; Lee et al. 2012]. The transformation is

$$R^2_{liability}=R^2_{observed}C/(1+R^2_{observed}\theta C) \text{ where } C=\frac{K(1-K)}{Z^2}\frac{K(1-K)}{P(1-P)} \text{ and }$$

$\theta=m\frac{P-K}{1-K}(m\frac{P-K}{1-K}-t)$. K is the population prevalence of BCC (0.3); t is the threshold according to K (0.52); Z is the height of a normal density at the threshold (0.35); m is the mean liability for BCC cases (1.16); P is the proportion of BCC cases in the case-control sample (0.25). In all analyses, the observed scale and liability scale R^2 were obtained using the same method. To obtain the LRT p-values, we used linear regression for HC and TA and logistic regression for BCC.

Prediction using Genetic Risk Scores (GRS) based on genome-wide significant SNPs

We built prediction models based on genetic risk scores (GRS) using genome-wide significant SNPs associated to HC, TA and BCC and compared GRS models to models based on both GRS and ancestry. We considered prediction models including 1) GRS_{unadj}, 2) GRS_{adj}, 3) GRS_{adj} and the top 5 PCs, and 4) GRS_{adj} and the top 5 PCs and self-reported ancestry (described in detail below). We employed both 10-fold cross-validation, which gives in-sample R^2, and 10×9-fold nested cross-validation which gives out-of-sample R^2 that accounts for model over-fitting.

For 10-fold cross-validation, we estimated effect sizes of the known associated SNPs in the training sets by fitting the following models in the 90% training samples:

$$Y = \beta_0 + \beta_1 GWAS + \beta_{unadj,j} SNP_j + \varepsilon \text{ for HC and TA} \quad (1)$$

or

$$logit[P(Y=1)] = \beta_0 + \beta_1 GWAS + \beta_{unadj,j} SNP_j \text{ for BCC} \quad (2)$$

where SNP_j is the genotype of genome-wide significant SNP j and GWAS is a 6×1 vector of 6 indicator variables indicating which of the 7 individual GWAS the sample is from. We then applied the effect size estimates to the validation samples to construct GRS_{unadj}. We calculated GRS_{unadj} for the 10% validation samples by

$$GRS_{unadj,i} = \sum_{j=1}^{M} \hat{\beta}_{unadj,j} SNP_{i,j} \quad GRS_i = \sum_{j=1}^{M} \hat{\beta}_{1,j} SNP_{i,j} \quad GRS_i = \sum_{j=1}^{M} \hat{\beta}_{1,j} SNP_{i,j} \quad (3)$$

where $\hat{\beta}_{unadj,j}$ is the estimated regression coefficient using the training samples from model (1) and (2) and $SNP_{i,j}$ is the genotype of validation sample i at SNP j. We also calculated GRS_{adj} by fitting models:

$$Y = \beta_0 + \beta_1 GWAS + \beta_2 PC + \beta_{unadj,j} SNP_j + \varepsilon \text{ for HC and TA} \quad (4)$$

or

$$logit[P(Y=1)] = \beta_0 + \beta_1 GWAS + \beta_2 PC + \beta_{unadj,j} SNP_j. \text{ for BCC} \quad (5)$$

where SNP_j and GWAS are as previously defined and PC is a 5×1 vector of the top 5 PCs. The GRS_{adj} was then calculated as

$$GRS_{adj,i} = \sum_{j=1}^{M} \hat{\beta}_{adj,j} SNP_{i,j} \quad (6)$$

where $\hat{\beta}_{adj,j}$ were estimated from model (4) and (5). We then pooled the samples from 10 mutually exclusive validation sets with GRS_{unadj} and GRS_{adj} constructed with estimates from their corresponding training samples and fitted models. In-sample R^2 and LRT p-values were calculated in the validation set for these models, where GRS and other predictors (PCs/self-reported ancestry) we treated as univariate predictors.

For the 10×9-fold nested cross-validation, the GWAS samples were split into 2 parts with 90% of the total samples as the training set and 10% of the total samples as validation set in each fold of the nested cross-validation. For the 90% training set, we further split the samples into 2 training sets. The first training set had 80% of the total samples and the second training set had 10% of the total samples. We used the first training set to estimate

effect sizes of the genome-wide significant SNPs and used those effect sizes to construct the GRS for the second training set as well as the validation set. We then fit prediction models that include GRS (and PCs/self-reported ancestry) in the second training set. The predicted phenotype was calculated for each validation sample as $\hat{Y} = \sum_{k=1}^{K} \beta_k x_k$ where k index the predictors, including GRS and/or PCs/self-reported ancestry, β_k is the regression coefficient estimated in the second training set and x_k is the covariate value for the validation sample. For each of the 10 folds of validation samples, we repeated this model building process for each of the 9 folds of first and second training sets and averaged the predicted phenotypes of validation samples across the 9 folds. With the pooled validation samples from the 10×9-fold nested cross-validation, we calculated out-of-sample R^2 between true phenotypes and predicted phenotypes. The out-of-sample R^2 were always calculated between a single predicted phenotype ($\hat{Y}$) and the true phenotype of validation samples and the potential inflation of R^2 due to increased number of predictors was eliminated. We note that the PCs used in all primary analyses were derived with all samples together.

Polygenic prediction using Polygenic Risk Scores (PRS) based on genome-wide SNPs

For the polygenic model with explicit modeling of ancestry, we adjusted for the top 5 PCs when constructing the PRS and added the top 5 PCs to the polygenic model as additional covariates. We compared prediction models constructed with 1) PRS_{unadj}, 2) PRS_{adj}, 3) PRS_{adj} and the top 5 PCs, and 4) PRS_{adj} and the top 5 PCs and self-reported ancestry. We first did the analyses with 10-fold cross-validation using R^2 and likelihood ratio test as figures of merit and then repeated the analyses in 10×9-fold nested cross-validation. Both 10-fold cross-validation and 10×9-fold nested cross-validation were performed in a similar fashion as described above for GRS model comparisons.

We adopted the polygenic risk score method to build polygenic prediction models [Evans et al. 2009; Purcell et al. 2009]. We used around 72,000 LD-pruned, independent genome-wide SNPs in PRS. To construct PRS without ancestry correction (PRS_{unadj}), we fitted regression models (1) and (2) for each of the independent genome-wide SNPs individually using the 90% training samples. To construct PRS with ancestry correction (PRS_{adj}), we fitted models (4) and (5) for each independent genome-wide SNP individually. The $\beta_{unadj, j}$ and $\hat{\beta}_{adj, j}$ were used as weights to construct PRS_{unadj} and PRS_{adj}, respectively. Models (3) and (6) were used to construct PRS, except that we included all independent genome-wide SNPs in the PRS. We also constructed PRS_{unadj} and PRS_{adj} using different numbers of genome-wide SNPs determined by the association p-value (calculated for each training set) for the SNP and the target phenotype: 5×10^{-8}, 1×10^{-7}, 1×10^{-6}, 1×10^{-5}, 1×10^{-4}, 0.001, 0.01, 0.05, 0.1, 0.5, and 1.0. For BCC, we did not construct PRS with p-value threshold smaller than 1×10^{-5} because no SNP showed association p-value smaller than 1×10^{-5} for BCC in our data.

We further extended the polygenic models above to include 1) both GRS_{unadj} and $PRS_{unadj,-gs}$, 2) both GRS_{adj} and $PRS_{adj,-gs}$, 3) both GRS_{adj} and $PRS_{adj,-gs}$ plus the top 5 PCs, and 4) both GRS_{unadj} and $PRS_{unadj,-gs}$ plus the top 5 PCs. Here, $PRS_{unadj,-gs}$ and $PRS_{adj,-gs}$ are PRS that excluded SNPs with genome-wide significant association for the target trait in previous GWAS (which are included in GRS), and SNPs in LD with those

genome-wide significant SNPs (pairwise $R^2 > 0.2$). The GRS and PRS were combined in the prediction model by either summing the GRS and PRS to make one score or including GRS and PRS as two separate components of the model.

We used in-sample R^2 and LRT to compare these models in the validation set in the 10-fold cross-validation, where all predictors were treated as univariate predictors in the model. We calculated out-of-sample R^2 in the 10×9-fold nested cross-validation, where the predicted phenotypes for each validation sample were calculated as $\hat{Y} = \sum_{k=1}^{K} \beta_k x_k$ where k index the predictors, including PRS and/or PCs/self-reported ancestry, β_k is the regression coefficient estimated in the second training set and x_k is the covariate value for the validation sample. We note that the PCs used in all primary analyses were derived with all samples together.

Polygenic prediction using Best Linear Unbiased Predictors (BLUP) based on genome-wide SNPs

In addition to PRS, we performed polygenic prediction based on BLUP by fitting linear mixed models to estimate genome-wide SNP effect sizes simultaneously [Yang et al. 2010]. We compared prediction models based on 1) $BLUP_{unadj}$, 2) $BLUP_{adj}$, 3) $BLUP_{adj}$ plus the top 5 PCs, 4) $BLUP_{unadj,-gs}$, 4) $BLUP_{adj,-gs}$, and 5) $BLUP_{adj,-gs}$ plus the top 5 PCs, where –gs denotes excluding genome-wide significant SNPs from the random effects of linear mixed model and entered the model as fixed effects. We note that the PCs used in all primary analyses were derived with all samples together. We adopted 10-fold cross-validation to fit linear mixed models in the 90% training samples and output the random effects and fixed effects estimates with GCTA. The linear mixed model we used to obtain effect size estimates takes the general form:

$$Y = X\beta + g + \varepsilon \text{ with } Var(Y) = A\sigma_g^2 + I\sigma_\varepsilon^2 \quad (7)$$

where β is a vector of fixed effects, g is the total genetic effects of the training samples with $g \sim N(0, A\sigma_g^2)$, A is the genetic relationship matrix derived from SNP genotypes, ε is the residual with $\varepsilon \sim N(0, I\sigma_\varepsilon^2)$ [Yang et al. 2010]. In our case, the fixed effects included individual GWAS study indicator, PCs, and/or genome-wide significant SNPs. The effect sizes of SNPs used to calculate PRS were calculated as $\left(\sigma_g^2/m\right) XV^{-1}Y$ where V^{-1} is the inverse of the phenotypic covariance matrix of the training samples and m is the number of SNPs used to calculated the genetic relationship matrix. We then apply the effect estimates to the 10% validation samples to construct PRS, GRS, and PCs. The calculation of PRS in the validation samples is the same as models (3) and (6). We used in-sample R^2 and LRT to compare these models in the 10-fold cross-validation with the pooled validation samples, where all predictors were treated as univariate predictors in the model.

Simulations comparing polygenic prediction using polygenic risk scores

It is widely known that polygenic prediction accuracy is predominantly determined by the training sample size. In our analysis using real GWAS samples, the training sample size is fixed. To examine the performance of different PRS approaches under different training

sample sizes, we performed the following simulation study. We simulated a quantitative trait with 100,000 independent SNPs, where all SNPs were causal and the proportion of phenotypic variance explained by these SNPs was 0.5. We simulated sets of 1,000, 5,000, 10,000, 50,000, 100,000, 500,000, 1,000,000 training samples and 10,000 validation samples. Training samples and validation samples were each drawn from two subpopulations POP1 and POP2, with minor allele frequency equal to 0.50 in POP1 and drawn from a normal distribution with mean 0.50 and standard deviation 0.05 in POP2, so that the fixation index F_{ST}(POP1, POP2)=0.005. The signs of marker effect sizes were flipped so that with probability 0.51, the allele with higher frequency in POP2 had positive effect, with the result that POP2 had systematically higher trait mean and subpopulation ancestry had an R^2 of 0.044 with the trait. We compared three approaches: PRS_{unadj}, PRS_{adj}, and PRS_{adj} + ancestry, which is equivalent to PRS_{adj} + top PC except that ancestry is assumed to be inferred with perfect accuracy. (We did not consider additional PCs because the data was simulated to contain only one dimension of population structure.) Results were averaged across 100 simulations. We compared the simulation results to the results from our PRS analysis using real GWAS samples.

Results

Inference of genetic ancestry

We performed analyses in three genome-wide association studies (GWAS) of hair color (HC), tanning ability (TA), and basal cell carcinoma (BCC) with samples from the Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS). We investigated the population structure in our GWAS samples using principal component analysis (PCA) [Price et al. 2006; Patterson et al. 2006], and compared the top principal components (PCs) to self-reported European ancestry. We first plotted PC1 and PC2 to investigate the population structure in the three GWAS (Figure 1). The PCA plot was color-coded with self-reported European ancestry, which has four categories: Scandinavian, Southern European, Ashkenazi Jewish, and Other European. We found that samples of self-reported Southern European ancestry had relatively lower PC1 values than samples of self-reported Scandinavian ancestry, while self-reported Ashkenazi Jewish samples had the lowest PC1 on average. In addition, the average PC1 and PC2 of Other European samples are close to the average PC1 and PC2 of Scandinavian samples, suggesting that Other European samples are predominantly of Northern European ancestry. The structure we identified in our GWAS sample is different from the canonical PC1 and PC2 in European American samples shown in Price *et al.* (2008) perhaps because the current data set has more samples of Eastern European ancestry than the samples analyzed in that paper. By projecting POPRES samples onto PC1 and PC2, we determined that samples with high PC2 value represent Eastern European ancestry (Figure S1) [Nelson et al. 2008]. This result suggests that lower PCs may capture additional structure. In the prediction analyses below, we include either PC1 and PC2 or the top 5 PCs; including more than 5 PCs (up to 10 PCs) did not change any of the results.

Prediction using genetic and self-reported ancestry

Pigmentation-related traits are associated with ancestry, due to positive selection on lighter pigmentation in populations in regions of higher latitude [Sulem et al. 2007]. Indeed, our data showed that HC, TA and BCC are strongly associated with genome-wide ancestry and self-reported ancestry (Table 1). Compared with self-reported Scandinavian samples, Southern European samples on average have darker hair color (2.60 in Southern European vs. 2.08 in Scandinavian), increased tanning ability (2.67 in Southern European vs. 2.36 in Scandinavian), which corresponds to better protection against ultraviolet radiation, and lower risk of BCC (odds ratio [OR] for Scandinavian vs. Southern European (ref) = 1.075, 95% confidence interval [CI]: 1.036, 1.116). Ashkenazi Jewish samples on average had the darkest hair color (2.76), which is consistent with their lower PC1 value vs. South Europeans, but also had the lowest tanning ability (2.30) and highest risk of BCC among 4 self-reported ancestry groups (OR = 1.123, 95% CI: 1.074, 1.174), which are not consistent with their lower PC1 value vs. South Europeans. The Scandinavian samples had similar but slightly more extreme phenotypes than the Other European samples, consistent with the similar but slightly higher average PC1 for Scandinavian vs. Other European samples.

Motivated by these observations, we fitted prediction models based on ancestry. We first examined models based on self-reported ancestry only, where the in-sample R^2 was 0.0397 for HC, 0.0081 for TA, and 0.0072 for BCC (Table 2). We then examined prediction models based on PCs only. The model based on top 5 PCs achieved R^2 of 0.0724 for HC, 0.0350 for TA, and 0.0202 for BCC in the in-sample analysis (Table 2). We also showed that adding PC3 to PC5 in addition to PC1 and PC2 improved model prediction accuracy significantly for all 3 phenotypes ($p < 10^{-16}$ for HC and TA; 1.93×10^{-15} for BCC). This suggests that including the subtle structure captured by later PCs provided additional value in predicting these phenotypes. However, including more than 5 PCs did not change any of the results. The R^2 obtained by using self-reported ancestry are lower than the R^2 obtained by using the top 5 PCs. Interestingly, we found including self-reported ancestry in the models in addition to the top 5 PCs further improved prediction accuracy (Table 2). The improvement was significant for HC ($p = 1.32 \times 10^{-9}$) and TA ($p = 5.70 \times 10^{-3}$) but not for BCC ($p = 0.18$) in the in-sample analysis. In addition, we showed that the added information in self-reported ancestry for HC (where this effect is strongest) is primarily due to self-reported Scandinavian ancestry, which is not completely captured by the Northwest-Southeast European cline of PC1 (Table S2). We also performed out-of-sample analyses using 10-fold cross-validation and obtained similar results (Table S3). Overall, we showed that both PCs and self-reported ancestry explained part of the phenotypic variation and can be used to build prediction models for these pigmentation-related traits. Furthermore, while models based on PCs have higher prediction accuracy than models based on self-reported ancestry, self-reported ancestry can still improve prediction accuracy over PCs.

Prediction using Genetic Risk Scores (GRS) based on genome-wide significant SNPs

To investigate the potential improvement of genetic risk prediction models by incorporating ancestry information, we compared genetic risk prediction models based on genetic risk scores (GRS) [Meigs et al. 2008]. We used imputed genotypes of SNPs shown to be genome-wide significant in previous GWAS to create GRS for each of the 3 phenotypes

(Table S1). The models we compared included GRS without ancestry correction ($\text{GRS}_{\text{unadj}}$), GRS with ancestry correction (adjusted for the top 5 PCs) (GRS_{adj}), and GRS with explicit modeling of ancestry (GRS_{adj}+PCs). In 10-fold cross-validation, $\text{GRS}_{\text{unadj}}$ attained an R^2 of 0.2236 for HC ($p < 10^{-16}$), 0.1311 for TA ($p < 10^{-16}$), and 0.0266 ($p < 10^{-16}$) for BCC (Table 3). The R^2 of GRS_{adj} were lower than the R^2 of $\text{GRS}_{\text{unadj}}$ for HC and BCC, but slightly higher for TA ($R^2 = 0.1324$). In 10×9-fold nested cross-validation, the R^2 of GRS_{adj} were consistently lower than R^2 of $\text{GRS}_{\text{unadj}}$ for all 3 phenotypes (Table S3). The R^2 of $\text{GRS}_{\text{unadj}}$ and GRS_{adj} were much higher than the models based on ancestry. However, we showed that the prediction accuracy was improved (relative to either $\text{GRS}_{\text{unadj}}$ or GRS_{adj}) by including PCs in the GRS_{adj} model (Table 3). We repeated the $\text{GRS}_{\text{unadj}}$, GRS_{adj} and GRS_{adj}+PCs model comparison by fitting all known genome-wide significant SNPs in one model simultaneously to obtain effect size estimates for constructing GRS (Table S4). The results showed similar patterns to the results from fitting genome-wide significant SNPs individually (Table 3). The improvement also remains for the out-of-sample R^2 from 10×9-fold nested cross-validation (Table S3). We also calculated the R^2 between the top 5 PCs and the predicted phenotypes based on $\text{GRS}_{\text{unadj}}$ and GRS_{adj} (Table S5). We found that the R^2 between PCs and GRS_{adj} were only slightly lower than the R^2 between PCs and $\text{GRS}_{\text{unadj}}$, which suggests that the genome-wide significant SNPs used to calculate GRS have systematically different allele frequencies across European subpopulations, e.g. due to natural selection [Turchin et al. 2012]. In addition to PCs, we also showed that including self-reported ancestry further improved prediction accuracy on top of GRS_{adj} + PCs for HC (R^2 from 0.2378 to 0.2423; $p = 1.24 \times 10^{-9}$) and TA (R^2 from 0.1384 to 0.1397; $p = 2.73 \times 10^{-3}$), but not BCC (R^2 from 0.0372 to 0.0385; $p = 0.108$) (Table S6).

Polygenic prediction using Polygenic Risk Scores (PRS) based on genome-wide SNPs

We investigated the impact of explicitly modeling ancestry on polygenic prediction by comparing 3 polygenic models: 1) the polygenic risk score (PRS) without ancestry correction ($\text{PRS}_{\text{unadj}}$), 2) a polygenic risk score with correction for ancestry (PRS_{adj}), and 3) a polygenic model with explicit modeling of ancestry, where we corrected for ancestry but added a separate ancestry component (PRS_{adj}+PCs) [Purcell et al. 2009; Evans et al. 2009]. We compared PRS models with PRS that included all available SNPs across the genome after LD pruning. Results are displayed in Table 4. Comparing PRS_{adj}+PCs to $\text{PRS}_{\text{unadj}}$, the R^2 improved 67% (from 0.0473 to 0.0789) for hair color, 114% (from 0.0162 to 0.0347) for tanning ability and 62% (from 0.0145 to 0.0235) for BCC. The model improvements were significant with p-value $< 10^{-16}$. We found that PRS_{adj} attained a lower R^2 than the other models (0.0055 for HC; 0.0003 for TA; 0.0029 for BCC). Notably, the improvement in R^2 of PRS_{adj}+PCs over PRS_{adj} was much larger than the R^2 of $\text{PRS}_{\text{unadj}}$ in each case, suggesting that $\text{PRS}_{\text{unadj}}$ suffered from a mis-specified weighting of ancestry effects, which enter into each SNP effect in the unadjusted model. The low R^2 of PRS_{adj} is consistent with the relatively small sample size in our analyses [Chatterjee et al. 2013; Dudbridge 2013], and indicates that the prediction accuracy of $\text{PRS}_{\text{unadj}}$ was largely due to ancestry. Indeed, we observed R^2 as large as 0.9250 between PC1 and $\text{PRS}_{\text{unadj}}$ (Table S5). This is consistent with our simulations, in which the prediction R^2 for $\text{PRS}_{\text{unadj}}$ was consistently biased toward the same number, 0.044, which is the R^2 between ancestry and the target trait. However, the prediction R^2 of PRS_{adj} increased from 0.003 to 0.439 as the training sample size increased

from 1,000 to 1,000,000, and PRS_{adj}+ancestry consistently outperformed both PRS_{unadj} and PRS_{adj} (Table 5). Results for PRS_{adj} were close to the values predicted by theory in the absence of ancestry effects (see Equation (1) of Wray et al. 2013 Nat Rev Genet) [Wray et al. 2013]. These simulations show that explicit modeling of genetic ancestry can improve polygenic prediction accuracy, and that the magnitude of this improvement can become extremely large as sample sizes grow.

We also compared the 3 types of polygenic models with GRS_{unadj} or GRS_{adj} included as a component separate from PRS_{unadj} or PRS_{adj} in the model (PRS_{unadj} + GRS_{unadj}, PRS_{adj} + GRS_{adj}, and PRS_{adj} + GRS_{adj} + PCs). We excluded genome-wide significant SNPs (and SNPs in LD with known associated SNPs) from PRS_{unadj} and PRS_{adj}, since GRS based on genome-wide significant SNPs were included in the model as a separate component. Comparing the models with both PRS + GRS to the models with PRS only, the R^2 in general improved and the improvements were consistent with the R^2 from the GRS models (Table 4). The improvement in R^2 due to explicit modeling of ancestry remained but with a smaller magnitude, similar to the analysis using GRS only (Table 3); similar relative results were obtained when using GRS_{adj} in each model. For the PRS + GRS analyses, we also constructed GRS by fitting all known genome-wide significant SNPs in one model simultaneously to obtain effect size estimates (Table S7). The results showed similar patterns to the results from fitting genome-wide significant SNPs individually (Table 4). We also performed out-of-sample analyses using 10×9-fold nested cross-validation and obtained similar results (Table S3).

In addition to the 3 polygenic models described above, we evaluated the models of PRS_{unadj}+PCs and PRS_{unadj}+GRS_{unadj}+PCs and found that these models attained lower R^2 than PRS_{adj} + PCs and PRS_{adj}+GRS_{adj}+PCs (Table 4). This result may be caused by the fact that PRS_{unadj} and PCs are correlated due to ancestry information in both, which is not optimally modeled. Finally, we considered models including self-reported ancestry in addition to PCs as ancestry measures. The polygenic models with explicit modeling of ancestry were further improved by self-reported ancestry for HC and TA, but not BCC (Table S6), consistent with predictions using ancestry only (Table 2).

In addition to using all independent SNPs across the genome to construct PRS, we applied a p-value threshold based on association tests between individual SNPs and corresponding phenotypes to select SNPs entering PRS. We used p-value thresholds from $5×10^{-8}$ to 1 and constructed PRS corresponding to each p-value threshold [Evans et al. 2009; Purcell et al. 2009]. We fitted the 3 types of polygenic models with PRS only and the 3 types of polygenic models with both PRS and GRS, using PRS subject to different p-value thresholds (Figure 2). The prediction accuracy for the polygenic models varied with p-value thresholds and showed different patterns for the 3 traits. The R^2 of PRS_{adj}+PCs model for HC and TA increased with more stringent p-value threshold, while R^2 of PRS_{adj}+PCs model for BCC remained stable. Overall, polygenic models with explicit modeling of ancestry had higher R^2 than models without ancestry correction across all p-value thresholds applied to PRS. In addition, the difference between R^2 from the model with explicit modeling of ancestry and the model without ancestry correction remained constant across p-value thresholds. For the PRS only models, the R^2 changed with the p-value threshold for PRS but the magnitude of

change was relatively small. For the models with both PRS and GRS components, the R^2 were still improved with a smaller magnitude by explicitly modeling ancestry, compared with PRS only models. We also repeated the analyses in 10×9-fold nested cross-validation and the results were similar to 10-fold simple cross-validation (Figure S2). Overall, we showed that explicit modeling of genetic ancestry improves polygenic prediction accuracy, and that this pattern persists when including known associated SNPs via a GRS, and when using p-value thresholds.

Polygenic prediction using Best Linear Unbiased Predictors (BLUP) based on genome-wide SNPs

As a comparison to PRS, which use SNP effect sizes obtained by fitting models for each SNP individually, we compared the 3 types of polygenic models based on best linear unbiased predictors (BLUP), which use SNP effect sizes obtained by fitting all SNPs using a linear mixed model [Henderson 1975; Haile-Mariam et al. 2013; de los Campos et al. 2013; Habier et al. 2013; Bolormaa et al. 2013]. We created 3 polygenic models: BLUP without ancestry correction ($BLUP_{unadj}$); BLUP with ancestry correction ($BLUP_{adj}$); and BLUP with explicit modeling of ancestry ($BLUP_{adj}$+PCs); in $BLUP_{adj}$ and $BLUP_{adj}$+PCs, ancestry was modeled using fixed effects (see Methods). In general, the R^2 of BLUP exhibited similar patterns as the R^2 of PRS (Table 6). In particular, $BLUP_{adj}$+PCs substantially outperformed $BLUP_{unadj}$. As with PRS, the improvement in R^2 of $BLUP_{adj}$+PCs over $BLUP_{adj}$ was much larger than the R^2 of $BLUP_{unadj}$ in each case, suggesting that $BLUP_{unadj}$ suffered from a mis-specified weighting of ancestry effects; although BLUP fits all SNPs simultaneously, it does not allow for different weights for different ancestry effects (e.g. PCs). $BLUP_{adj}$ had low R^2 due to small sample size, indicating that the prediction accuracy of $BLUP_{unadj}$ was largely due to ancestry (Table S5). In models with BLUP and GRS as 2 separate components (where genome-wide significant SNPs were included in the GRS but not in BLUP), $BLUP_{adj}$+GRS_{adj}+PCs had significantly higher R^2 than $BLUP_{unadj}$+GRS_{unadj} (Table 6). Adding self-reported ancestry in the $BLUP_{adj}$+GRS_{adj}+PCs and the $BLUP_{adj}$+PCs models improves the prediction accuracy significantly for HC and TA, but not BCC (Table S6).

Discussion

Polygenic prediction using genome-wide SNPs has been widely applied to many complex traits. The most widely used polygenic prediction method is to construct a polygenic risk score (PRS), which is the sum of allele counts weighted by effect sizes estimated individually for each SNP, optionally restricting to SNPs with association p-value below a certain threshold [Chatterjee et al. 2013; Dudbridge 2013]. (It has previously been noted (Chatterjee *et al.* 2013) that other model selection and shrinkage methods, such as least absolute shrinkage and selection operator (LASSO)-type threshold methods that analyze all SNPs simultaneously, may modestly outperform the standard polygenic prediction approach). In previous studies using PRS, some of them adjusted for ancestry (PRS_{adj}) [Purcell et al. 2009; Stahl et al. 2012; Smoller et al. 2013; Rietveld et al. 2013, PGC 2014 Nature], and some of them did not (PRS_{unadj}) [Evans et al. 2009; Bush et al. 2010; Lango Allen et al. 2010; Machiela et al. 2011; Wood et al. 2014 Nat Genet]. However, a separate

point is that modeling genetic ancestry as a separate component (PRS_{adj}+PCs) can improve polygenic prediction accuracy relative to either PRS_{unadj} or PRS_{adj}, as we showed for HC, TA and BCC; the same result is expected to hold for other traits associated to ancestry in subtly structured populations; this could explain why recent GWAS of height, which have reported that their association results may be impacted by very subtle population stratification [Lango Allen et al. 2010; Wood et al. 2014 Nat Genet], have attained prediction R2 using PRS that are lower than predicted by theory [Daetwyler et al. 2008 PLoS ONE]. In addition, the result still holds when including known associated SNPs via a genetic risk score (GRS), and when using p-value thresholds. In the European American data we analyzed, using 5 PCs to model ancestry improved our results relative to using 2 PCs. Surprisingly, including self-reported ancestry in addition to 5 PCs improved results further, as self-reported ancestry contains information about Scandinavian ancestry that is not well-captured by the PCs. Thus, PCs are not guaranteed to capture all ancestry information in subtly structured populations.

When applying methods that do not model genetic ancestry as a separate component (e.g. PRS_{unadj} or PRS_{adj}), it is still of interest to understand how much of polygenic prediction accuracy is due to ancestry, as this will affect prediction accuracy in other populations. For example, Rietveld et al. reported a polygenic prediction R^2 of 0.02 ($p < 10^{-16}$) in an analysis of unrelated samples but a polygenic prediction R^2 of only 0.003 ($p = 0.001$) in a within-family analysis immune to ancestry effects (see Table S25 of Rietveld et al.) [Rietveld et al. 2013]. Unmodeled population structure provides one plausible explanation for this discrepancy. We note that although Rietveld et al. adjusted for 4 PCs in their analyses (PRS_{adj}), there is no guarantee that PCs capture all ancestry information in subtly structure populations. In addition, it is possible that PRS_{adj}+PCs might attain a higher prediction R^2 than PRS_{adj} in unrelated samples in the Rietveld et al. data, as we showed for HC, TA and BCC.

A previous study analyzed GWAS data from the Framingham Heart Study to compare prediction models for skin cancer risk using either pedigree information, or PCs, or polygenic prediction using Bayesian Lasso [Vazquez et al. 2012]. This study reported that polygenic prediction attained the highest prediction accuracy, PCs attained the next highest prediction accuracy, and pedigree information attained the lowest prediction accuracy. However this study did not explicitly model ancestry in the polygenic prediction model. Based on our results, it is possible that polygenic prediction with explicit modeling of ancestry might attain higher accuracy than the polygenic prediction approach employed by ref. [Vazquez et al. 2012]. It is also possible that combining pedigree information with polygenic predictions might improve prediction accuracy, although we have not explored that approach here [Haile-Mariam et al. 2013].

Alternatives to the widely used PRS approach for polygenic prediction include Best Linear Unbiased Prediction (BLUP) and its Bayesian extensions, which fit all markers simultaneously [de los Campos et al. 2013; Habier et al. 2013; Zhou et al. 2013; Bolormaa et al. 2013]. The potential benefits of explicitly modeling ancestry in this context may depend on the prior distribution of effect sizes that is assumed. Under an infinitesimal (Gaussian) prior, fitting all marker effects simultaneously may implicitly capture the effects of ancestry.

However, our results showed that BLUP still showed an improvement in prediction accuracy when explicitly modeling ancestry. Although previous work has argued against including PCs as separate fixed effects in a mixed model when estimating components of heritability [Janss et al. 2012], and has shown that including PCs as separate fixed effects is generally not necessary when using standard mixed model methods for association [Yang et al. 2014], those studies did not consider prediction accuracy. Our recommendations for prediction are different from those recommendations for heritability estimation and association.

A limitation of our study is that the increased accuracy of PRS_{adj}+PCs (relative to PRS_{adj}) is contingent on validation samples having the same population structure as the training samples, as was the case in our cross-validation experiments. If validation samples do not have the same population structure then prediction accuracy will be lower, although PRS_{adj}+PCs is still an appropriate strategy. (For example, if validation samples are from a homogeneous population, then PRS_{adj}+PCs will produce predictions virtually identical to PRS_{adj}.) Another potential limitation of our study is that it may not be feasible to compute PCs by running PCA on raw genotypes from the union of training and validation samples, as was done here; for example, this will not be feasible if only summary statistics are available for training samples. However, in this case, the top PCs for validation samples can be computed using weights derived from external reference panels (as projected PCs), as we have described previously [Chen et al. 2013]. We repeated the main analyses with top projected PCs derived for both training and validation samples using an external European American reference panel and obtained similar results to the results using PCs derived from raw genotypes of training and validation samples (Tables S8–S12). (We note that our main analyses use only training sample genotypes and phenotypes and validation sample genotypes to predict validation sample phenotypes, which is appropriate.)

We conclude by emphasizing the dependence of polygenic prediction accuracy on sample size. Both PRS_{adj} (with or without P-value thresholding to impose model selection) and $BLUP_{adj}$ attained low R2 in our analyses at modest sample sizes (<10,000), consistent with theoretical derivations [Daetwyler et al. 2008; Wray et al. 2013] and simulations suggesting that very large training sample sizes (100,000 or more) are needed to achieve high prediction accuracy [Chatterjee et al. 2013; Dudbridge 2013]. This limitation is in no way specific to the PRS approach. At larger sample sizes, it is likely that polygenic prediction will have significant clinical ramifications [Chatterjee et al. 2013; Dudbridge 2013]; indeed, prediction R2 has improved in recent very large studies [Ripke et al. 2014]. Notably, our simulations specifically show that the impact of explicit modeling of genetic ancestry will become extremely large as sample sizes grow.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Angastiniotis M, Modell B. Global epidemiology of hemoglobin disorders. Ann N Y Acad Sci. 1998; 850:251–69. [PubMed: 9668547]

Bernardi F, Arcieri P, Bertina RM, Chiarotti F, Corral J, Pinotti M, Prydz H, Samama M, Sandset PM, Strom R, et al. Contribution of Factor VII Genotype to Activated FVII Levels: Differences in Genotype Frequencies Between Northern and Southern European Populations. Arterioscler Thromb Vasc Biol. 1997; 17:2548–553. [PubMed: 9409226]

Bolormaa S, Pryce JE, Kemper K, Savin K, Hayes BJ, Barendse W, Zhang Y, Reich CM, Mason BA, Bunch RJ, et al. Accuracy of prediction of genomic breeding values for residual feed intake, carcass and meat quality traits in Bos taurus, Bos indicus and composite beef cattle. J Anim Sci. 2013; 91:3088–104. [PubMed: 23658330]

Borjas GJ. The Economics of Immigration. J Econ Lit. 1994; 32:1667–717.

Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007; 81:1084–97. [PubMed: 17924348]

Bush WS, Sawcer SJ, de Jager PL, Oksenberg JR, McCauley JL, Pericak-Vance Ma, Haines JL. Evidence for polygenic susceptibility to multiple sclerosis--the shape of things to come. Am J Hum Genet. 2010; 86:621–5. [PubMed: 20362272]

Candille SI, Absher DM, Beleza S, Bauchet M, McEvoy B, Garrison Na, Li JZ, Myers RM, Barsh GS, Tang H, et al. Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four European populations. PloS One. 2012; 7(10):e48294. [PubMed: 23118974]

Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park J-H. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet. 2013; 45(4):400–5. 405e1–3. [PubMed: 23455638]

Chen C-Y, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL. Improved ancestry inference using weights from external reference panels. Bioinformatics. 2013; 29:1399–406. [PubMed: 23539302]

Cimmino, Ma; Parisi, M.; Moggiana, G.; Mela, GS.; Accardo, S. Prevalence of rheumatoid arthritis in Italy: the Chiavari Study. Ann Rheum Dis. 1998; 57:315–8. [PubMed: 9741317]

Curhan GC, Taylor EN. 24-H Uric Acid Excretion and the Risk of Kidney Stones. Kidney Int. 2008; 73:489–96. [PubMed: 18059457]

Daetwyler HD, Villanueva B, Woolliams Ja. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PloS One. 2008; 3:e3395. [PubMed: 18852893]

Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. PLoS Genet. 2013; 9(3):e1003348. [PubMed: 23555274]

Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet. 2009; 18:3525–31. [PubMed: 19553258]

Gerstenblith MR, Shi J, Landi MT. Genome-wide association studies of pigmentation and skin cancer: a review and meta-analysis. Pigment Cell Melanoma Res. 2010; 23:587–606. [PubMed: 20546537]

Habier D, Fernando RL, Garrick DJ. Genomic-BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics. 2013; 194:597–607. [PubMed: 23640517]

Haile-Mariam M, Nieuwhof GJ, Beard KT, Konstatinov KV, Hayes BJ. Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. J Anim Breed Genet. 2013; 130:20–31. [PubMed: 23317062]

Han J, Kraft P, Colditz GA, Wong J, Hunter DJ. Melanocortin 1 receptor variants and skin cancer risk. Int J Cancer. 2006; 119:1976–84. [PubMed: 16721784]

Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, Hankinson SE, Hu FB, Duffy DL, Zhao ZZ, et al. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genet. 2008; 4:e1000074. [PubMed: 18483556]

Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometrics. 1975; 31:423–47. [PubMed: 1174616]

Hunter DJ, Colditz GA, Stampfer MJ, Rosner B, Willett WC, Speizer FE. Risk factors for basal cell carcinoma in a prospective cohort of women. Ann Epidemiol. 1990; 1:13–23. [PubMed: 1669486]

Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet. 2007; 39:870–4. [PubMed: 17529973]

Janss L, de Los Campos G, Sheehan N, Sorensen D. Inferences from genomic models in stratified populations. Genetics. 2012; 192:693–704. [PubMed: 22813891]

Kenny EE, Pe'er I, Karban A, Ozelius L, Mitchell AA, Ng SM, Erazo M, Ostrer H, Abraham C, Abreu MT, et al. A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. PLoS Genet. 2012; 8:e1002559. [PubMed: 22412388]

Kraft P, Hunter DJ. Genetic risk prediction--are we there yet? N Engl J Med. 2009; 360:1701–3. [PubMed: 19369656]

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–8. [PubMed: 20881960]

Lear JT, Smith AG. Basal cell carcinoma. Postgrad Med J. 1997; 73:538–42. [PubMed: 9373591]

Lee SH, Goddard ME, Wray NR, Visscher PM. A better coefficient of determination for genetic profile analysis. Genet Epidemiol. 2012; 36:214–24. [PubMed: 22714935]

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010; 34:816–34. [PubMed: 21058334]

De los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet. 2010; 11:880–6. [PubMed: 21045869]

De los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. PLoS Genet. 2013; 9:e1003608. [PubMed: 23874214]

Machiela MJ, Chen C-Y, Chen C, Chanock SJ, Hunter DJ, Kraft P. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. Genet Epidemiol. 2011; 35:506–14. [PubMed: 21618606]

Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de los Campos G. Beyond missing heritability: prediction of complex traits. PLoS Genet. 2011; 7:e1002051. [PubMed: 21552331]

Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PWF, D'Agostino RB, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med. 2008; 359:2208–19. [PubMed: 19020323]

Nan H, Kraft P, Hunter DJ, Han J. Genetic variants in pigmentation genes, pigmentary phenotypes, and risk of skin cancer in Caucasians. Int J Cancer. 2009a; 125:909–17. [PubMed: 19384953]

Nan H, Kraft P, Qureshi AA, Guo Q, Chen C, Hankinson SE, Hu FB, Thomas G, Hoover RN, Chanock S. Genome-wide association study of tanning phenotype in a population of European ancestry. J Invest Dermatol. 2009b; 129:2250–7. [PubMed: 19340012]

Nan H, Xu M, Kraft P, Qureshi AA, Chen C, Guo Q, Hu FB, Curhan G, Amos CI, Wang L-E, et al. Genome-wide association study identifies novel alleles associated with risk of cutaneous basal cell carcinoma and squamous cell carcinoma. Hum Mol Genet. 2011; 20:3718–24. [PubMed: 21700618]

Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Vollenweider P. The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research. Am J Hum Genet. 2008; 83:347–58. [PubMed: 18760391]

Panza F, Solfrizzi V, D'Introno A, Colacicco AM, Capurso C, Capurso A, Kehoe PG. Shifts in angiotensin I converting enzyme insertion allele frequency across Europe: implications for Alzheimer's disease risk. J Neurol Neurosurg Psychiatry. 2003; 74:1159–61. [PubMed: 12876265]

Patterson CC, Dahlquist GG, Gyürüs E, Green A, Soltész G. Incidence trends for childhood type 1 diabetes in Europe during 1989–2003 and predicted new cases 2005–20: a multicentre prospective registration study. Lancet. 2009; 373:2027–33. [PubMed: 19481249]

Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2:e190. [PubMed: 17194218]

Price AL, Butler J, Patterson N, Capelli C, Pascali VL, Scarnicci F, Ruiz-Linares A, Groop L, Saetta Aa, Korkolopoulou P, et al. Discerning the Ancestry of European Americans in Genetic Association Studies. PLoS Genet. 2008; 4:e236. [PubMed: 18208327]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick Na, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–9. [PubMed: 16862161]

Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010; 11:459–63. [PubMed: 20548291]

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460:748–52. [PubMed: 19571811]

Qi L, Cornelis MC, Kraft P, Stanya KJ, Linda Kao WH, Pankow JS, Dupuis J, Florez JC, Fox CS, Paré G, et al. Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Hum Mol Genet. 2010; 19:2706–15. [PubMed: 20418489]

Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW, Westra H-J, Shakhbazov K, Abdellaoui A, Agrawal A, et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science. 2013; 340:1467–71. [PubMed: 23722424]

Rimm EB, Giovannucci EL, Willett WC, Colditz Ga, Ascherio A, Rosner B, Stampfer MJ. Prospective study of alcohol consumption and risk of coronary disease in men. Lancet. 1991; 338:464–8. [PubMed: 1678444]

Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511:421–7. [PubMed: 25056061]

Rosati G. The prevalence of multiple sclerosis in the world: an update. Neurol Sci. 2001; 22:117–39. [PubMed: 11603614]

Smoller JW, Craddock N, Kendler K, Lee PH, Neale BM, Nurnberger JI, Ripke S, Santangelo S, Sullivan PF. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet. 2013; 381:1371–9. [PubMed: 23453885]

So H, Kwan JSH, Cherny SS, Sham PC. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. Am J Hum Genet. 2011; 88:548–65. [PubMed: 21529750]

Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, Kraft P, Chen R, Kallberg HJ, Kurreeman FS, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet. 2012; 44:483–9. [PubMed: 22446960]

Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet. 2007; 39:1443–52. [PubMed: 17952075]

Taylor E, Stampfer M, Curhan G. Obesity, weight gain, and the risk of kidney stones. J Am Med Assoc. 2005; 293:455–62.

The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]

Turchin MC, Chiang CWK, Palmer CD, Sankararaman S, Reich D, Hirschhorn JN. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat Genet. 2012; 44:1015–9. [PubMed: 22902787]

Vazquez AI, de Los Campos G, Klimentidis YC, Rosa GJM, Gianola D, Yi N, Allison DB. A comprehensive genetic approach for improving prediction of skin cancer risk in humans. Genetics. 2012; 192:1493–502. [PubMed: 23051645]

Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. Am J Hum Genet. 2012; 90:7–24. [PubMed: 22243964]

Wei Z, Wang K, Qu H-Q, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R, et al. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet. 2009; 5:e1000678. [PubMed: 19816555]

Wei Z, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, Kim C, Mentch F, Steen KV, Visscher PM, et al. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. Am J Hum Genet. 2013; 92:1008–12. [PubMed: 23731541]

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014; 46:1173–86. [PubMed: 25282103]

Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 2007; 17:1520–8. [PubMed: 17785532]

Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013; 14:507–15. [PubMed: 23774735]

Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2010; 88:76–82. [PubMed: 21167468]

Yang J, Zaitlen N, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014; 46:100–6. [PubMed: 24473328]

Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 2013; 9:e1003264. [PubMed: 23408905]
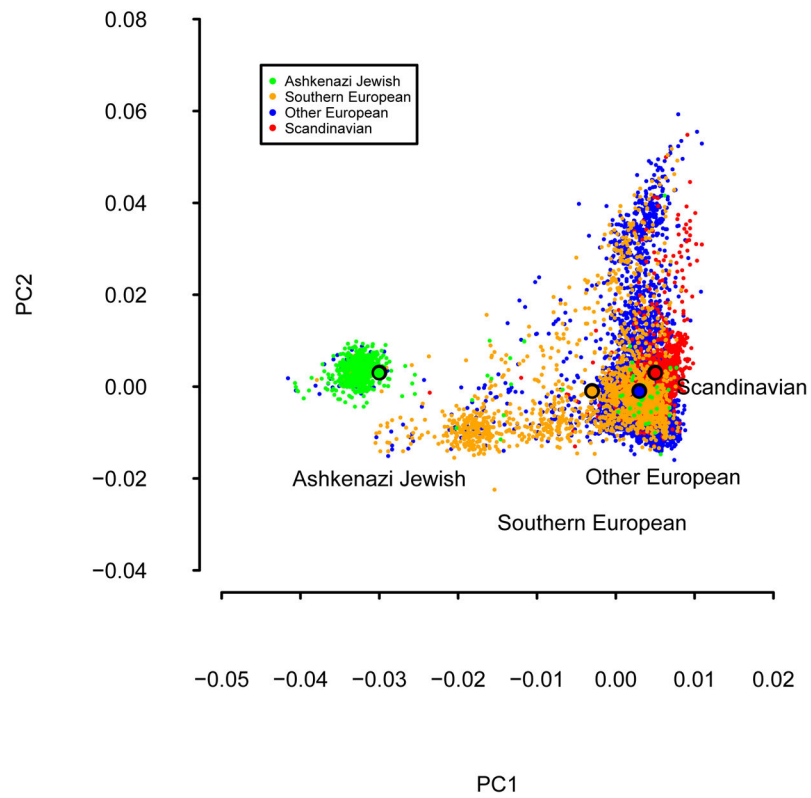
Figure 1. PCA plots for tanning ability GWAS samples, color-coded with self-reported ancestry
Black circles with color-coded interior denote average PC values for each self-reported ancestry. The tanning ability GWAS had the largest samples size among the 3 GWAS analyzed. Results for hair color and basal cell carcinoma GWAS samples, which are largely overlapped with the tanning ability GWAS samples, were extremely similar.
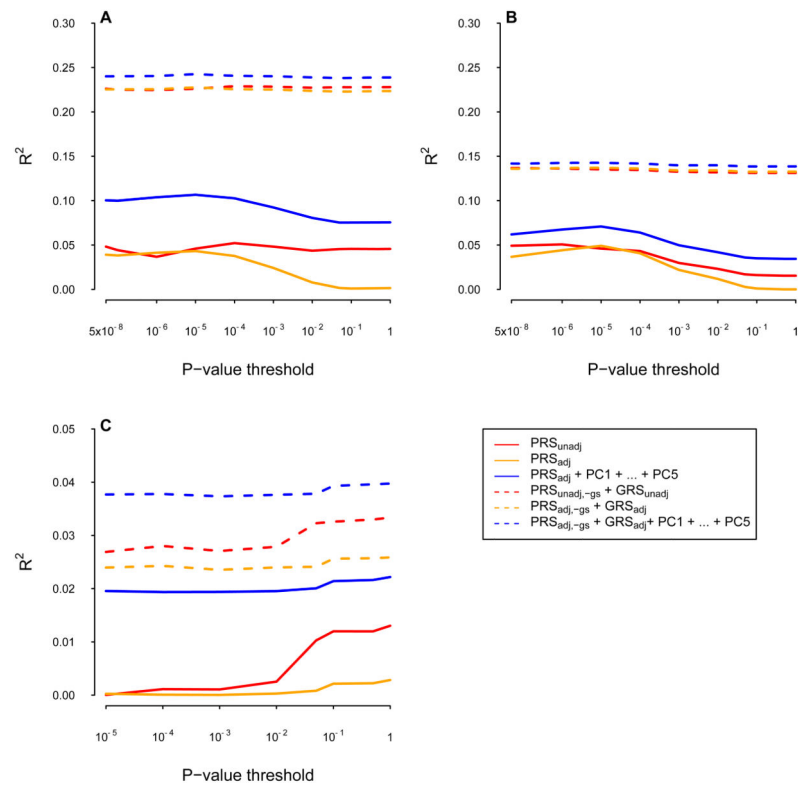
Figure 2. R^2 for polygenic prediction using PRS with different association p-value thresholds, GRS and PCs in 10-fold cross-validation

(A): Hair color (B): Tanning ability (C): Basal cell carcinoma

Table 1

Average PCs and phenotype values for self-reported ancestry groups in GWASs on hair color, tanning ability, and basal cell carcinoma

Self-reported European ancestry	Average PCs (SD)[a]		Hair color		Tanning ability		Basal cell carcinoma		
	1	2	N	Average	N	Average	Cases	Control	OR (95% CI)
Ashkenazi Jewish	−0.030 (0.009)	0.003 (0.004)	425	2.76 (0.73)	579	2.30 (0.87)	338	155	1.123 (1.074, 1.174)
Southern European	−0.003 (0.010)	−0.001 (0.011)	1325	2.60 (0.74)	1700	2.67 (0.90)	1151	285	Reference
Other European[b]	0.003 (0.005)	−0.001 (0.010)	4963	2.39 (0.75)	6559	2.54 (0.91)	4074	1419	1.062 (1.035, 1.089)
Scandinavian	0.005 (0.003)	0.003 (0.009)	727	2.08 (0.80)	984	2.36 (0.91)	610	227	1.075 (1.036, 1.116)
Total			7440		9822		6173	2086	

[a] Average PCs are derived from tanning ability GWAS samples, which has the largest sample size among the 3 GWAS analyzed here. Samples of the 3 GWAS are largely overlapped.

[b] Other European refers to European American samples without self-reported European sub-population ancestry information.

Table 2

R^2 and p-values from likelihood ratio test for prediction models using PCs and self-reported European ancestry in in-sample analysis using all samples

Model		Hair color		Tanning ability		Basal cell carcinoma[a]	
		R^2	P-value[b]	R^2	P-value[b]	R^2	P-value[b]
1	Self-reported ancestry	0.0397	$<10^{-16}$	0.0081	$<10^{-16}$	0.0072	9.19×10^{-7}
2	PC1 + PC2	0.0468	$<10^{-16}$	0.0146	$<10^{-16}$	0.0036	5.01×10^{-4}
3	PC1 + … + PC5	0.0724	$<10^{-16}$	0.0350	$<10^{-16}$	0.0202	1.93×10^{-15}
4	PC1 + … + PC5 + self-reported ancestry	0.0779	1.32×10^{-9}	0.0363	5.70×10^{-3}	0.0213	1.81×10^{-1}

[a]R^2 on liability scale

[b]P-values are from likelihood ratio tests comparing model 1 and 2 to the null model, model 3 to model 2, and model 4 to model 3.

Table 3

R^2 and p-values from likelihood ratio test for prediction models using GWAS identified SNPs and PCs in 10-fold cross-validation

Model		Hair color		Tanning ability		Basal cell carcinoma[a]	
		R^2	P-value[b]	R^2	P-value[b]	R^2	P-value[b]
1	GRS_{unadj}	0.2236	$<10^{-16}$	0.1311	$<10^{-16}$	0.0266	$<10^{-16}$
2	GRS_{adj}	0.2227	$<10^{-16}$	0.1324	$<10^{-16}$	0.0234	$<10^{-16}$
3	$GRS_{adj} + PC1 + \ldots + PC5$	0.2378	$<10^{-16}$	0.1384	1.93×10^{-13}	0.0372	6.82×10^{-13}

[a]R^2 on liability scale

[b]P-values are from likelihood ratio tests comparing model 1 and 2 to the null model, and model 3 to model 2.

Table 4

R^2 and p-values from likelihood ratio test for polygenic prediction using PRS, GRS and PCs in 10-fold cross-validation

(a) PRS included all SNPs available from genome-wide SNP chip after LD-pruning

Model		Hair color		Tanning ability		Basal cell carcinoma[a]	
		R^2	P-value[b]	R^2	P-value[b]	R^2	P-value[b]
1	PRS$_{unadj}$	0.0473	<10^{-16}	0.0162	<10^{-16}	0.0145	3.34 × 10^{-15}
2	PRS$_{adj}$	0.0055	1.65 × 10^{-10}	0.0003	7.67 × 10^{-2}	0.0029	5.90 × 10^{-4}
3	PRS$_{adj}$ + PC1 + … + PC5	0.0789	<10^{-16}	0.0347	<10^{-16}	0.0235	<10^{-16}
4	PRS$_{unadj}$ + PC1 + … + PC5	0.0745	<10^{-16}	0.0344	<10^{-16}	0.0226	1.36 × 10^{-6}

(b) GRS included known genome-wide significant SNPs; PRS excluded SNPs in LD with known genome-wide significant SNPs but included all other SNPs

Model		Hair color		Tanning ability		Basal cell carcinoma[a]	
		R^2	P-value[b]	R^2	P-value[b]	R^2	P-value[b]
5	PRS$_{unadj,-gs}$ + GRS$_{unadj}$[c]	0.2280	8.82 × 10^{-11}	0.1312	0.1776	0.0333	1.44 × 10^{-8}
6	PRS$_{adj,-gs}$ + GRS$_{adj}$	0.2235	7.43 × 10^{-3}	0.1326	0.1820	0.0259	6.23 × 10^{-4}
7	PRS$_{adj,-gs}$ + GRS$_{adj}$ + PC1 + … + PC5	0.2388	<10^{-16}	0.1386	2.45 × 10^{-13}	0.0397	5.46 × 10^{-13}
8	PRS$_{unadj,-gs}$ + GRS$_{unadj}$ + PC1 + … + PC5	0.2376	<10^{-16}	0.1378	1.39 × 10^{-14}	0.0398	6.00 × 10^{-6}

[a] R^2 on liability scale

[b] P-values are from likelihood ratio tests comparing model 1 and 2 to the null model, model 3 to model 2, model 4 to model 1, model 5 and 6 to the corresponding models with only GRS, and model 7 to model 6, and model 8 to model 5.

[c] The R^2 for PRS$_{unadj,-gs}$ + GRS$_{adj}$ were 0.2284, 0.1327 and 0.0308 for hair color, tanning ability, basal cell carcinoma.

-gs denotes excluding genome-wide significant SNPs.

Table 5

Comparison between prediction R^2 of PRS_{unadj}, PRS_{adj}, and PRS_{adj} + ancestry in simulations with different training sample sizes

The fluctuations in R^2 of PRS_{unadj} are not statistically significant based on the number of simulations performed.

Training sample size (N)	R^2		
	PRS_{unadj} (SE)	PRS_{adj} (SE)	PRS_{adj} + ancestry (SE)
1,000	0.041 (0.0016)	0.003 (0.0002)	0.042 (0.0016)
5,000	0.041 (0.0020)	0.014 (0.0004)	0.051 (0.0019)
10,000	0.043 (0.0019)	0.027 (0.0005)	0.064 (0.0018)
50,000	0.045 (0.0020)	0.111 (0.011)	0.139 (0.0019)
100,000	0.044 (0.0020)	0.182 (0.0011)	0.201 (0.0017)
500,000	0.042 (0.0020)	0.380 (0.0015)	0.384 (0.0016)
1,000,000	0.042 (0.0018)	0.439 (0.0012)	0.440 (0.0013)

SE: standard error

Table 6

R^2 and p-values from likelihood ratio test for polygenic prediction using BLUP, GRS and PCs in 10-fold cross-validation

(a) BLUP included all SNPs available from genome-wide SNP chip after LD-pruning

Model		Hair color		Tanning ability		Basal cell carcinoma[a]	
		R^2	P-value[b]	R^2	P-value[b]	R^2	P-value[b]
1	BLUP_{unadj}	0.0588	$<10^{-16}$	0.0131	$<10^{-16}$	0.0125	2.93×10^{-13}
2	BLUP_{adj}	0.0045	5.86×10^{-9}	0.0001	0.385	0.0013	0.0112
3	$\text{BLUP}_{adj} + \text{PC1} + \ldots + \text{PC5}$	0.0781	$<10^{-16}$	0.0345	$<10^{-16}$	0.0218	$<10^{-16}$

(b) GRS including genome-wide significant SNPs; BLUP excluded SNPs in LD with genome-wide significant SNPs but all other SNPs

Model		Hair color		Tanning ability		Basal cell carcinoma[a]	
		R^2	P-value[b]	R^2	P-value[b]	R^2	P-value[b]
4	$\text{BLUP}_{unadj,-gs} + \text{GRS}_{unadj}$	0.2350	$<10^{-16}$	0.1221	8.07×10^{-1}	0.0275	6.49×10^{-7}
5	$\text{BLUP}_{adj,-gs} + \text{GRS}_{adj}$	0.2258	8.10×10^{-3}	0.1233	2.97×10^{-6}	0.0208	5.52×10^{-2}
6	$\text{BLUP}_{adj,-gs} + \text{GRS}_{adj} + \text{PC1} + \ldots + \text{PC5}$	0.2413	$<10^{-16}$	0.1308	$<10^{-16}$	0.0359	3.83×10^{-15}

[a] R^2 on liability scale

[b] P-values are from likelihood ratio tests comparing model 1 and 2 to the null model, model 3 to model 2, model 4 and 5 to the corresponding models with only GRS, and model 6 to model 5.

-gs denotes excluding genome-wide significant SNPs.