Significant Sparse Polygenic Risk Scores across 813 traits in UK Biobank

Authors and affiliations

Yosuke Tanigawa[1,4,*], Junyang Qian[2], Guhan Venkataraman[1], Johanne Marie Justesen[1], Ruilin Li[3], Robert Tibshirani[1,2], Trevor Hastie[1,2], Manuel A. Rivas[1,*]

1. Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, United States
2. Department of Statistics, Stanford University, Stanford, CA 94305, United States
3. Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, United States
4. Current address: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, United States

* Correspondence: MAR mrivas@stanford.edu, YT tanigawa@mit.edu

Abstract

We present a systematic assessment of polygenic risk score (PRS) prediction across more than 1,500 traits using genetic and phenotype data in the UK Biobank. We report 813 sparse PRS models with significant ($p < 2.5 \times 10^{-5}$) incremental predictive performance when compared against the covariate-only model that considers age, sex, types of genotyping arrays, and the principal component loadings of genotypes. We report a significant correlation between the number of genetic variants selected in the sparse PRS model and the incremental predictive performance (Spearman's $\rho = 0.61$, $p = 2.2 \times 10^{-59}$ for quantitative traits, $\rho = 0.21$, $p = 9.6 \times 10^{-4}$ for binary traits). The sparse PRS model trained on European individuals showed limited transferability when evaluated on non-European individuals in the UK Biobank. We provide the PRS model weights on the Global Biobank Engine (https://biobankengine.stanford.edu/prs).

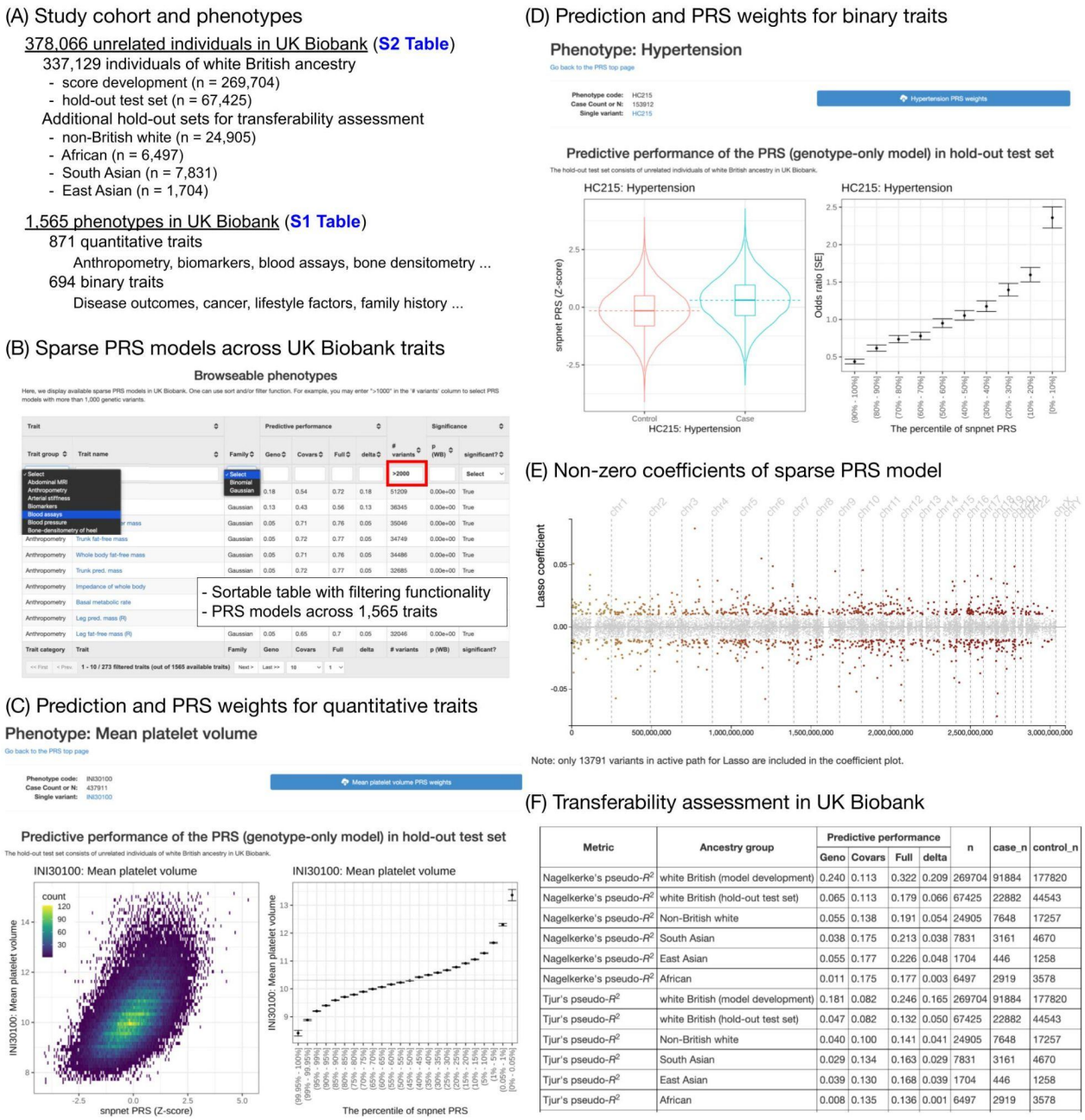
Author summary

Polygenic risk scores (PRSs), an approach to estimate genetic predisposition on disease liability by aggregating the effects across multiple genetic variants, has attracted increasing research interest. While there have been improvements in the predictive performance of PRS for some traits, the applicability of PRS models across a wide range of human traits has not been clear. Here, applying penalized regression using Batch Screening Iterative Lasso (BASIL) algorithm to more than 269,000 individuals of white British ancestry in UK Biobank, we systematically characterize PRS models across more than 1,500 traits. We report 813 traits with PRS models of statistically significant predictive performance. While the statistical significance does not necessarily directly translate into clinical relevance, we investigate the properties of the 813 significant PRS models and report a significant correlation between predictive performance and estimated SNP-based heritability. We find that the number of genetic variants selected in our sparse PRS model is significantly correlated with the incremental predictive performance in both quantitative and binary traits. Our transferability assessment of PRS models in UK Biobank revealed that the sparse PRS models trained on individuals of European ancestry had a lower predictive performance for individuals of African and Asian ancestry groups.

Introduction

Polygenic risk score (PRS), an estimate of an individual's genetic liability to a trait or disease, has been proposed for disease risk prediction with potential clinical relevance for some traits[1,2]. Due to training data sample size increase and methods development advances for variable selection and effect size estimation, PRS predictive performance has improved[3–17]. However, it has not been clear what would be the predictive performance of PRS models when it is applied to a wide range of traits and their transferability across ancestry groups. Rich phenotypic information in large-scale genotyped cohorts provides an opportunity to address this question.

Here, we present significant sparse PRSs across 813 traits in the UK Biobank[18,19]. We applied the recently developed batch screening iterative lasso (BASIL) algorithm implemented in the R *snpnet* package[10] across more than 1,500 traits consisting of binary outcomes and quantitative traits, including disease outcomes and biomarkers, respectively (**Fig. 1**, **S1 Table**). As opposed to most of the recently developed PRS methods that take genome-wide association study (GWAS) summary statistics as input, BASIL/*snpnet* is capable of performing variable selection and effect size estimation simultaneously from individual-level genotype and phenotype data. BASIL/*snpnet* results in sparse PRS models, meaning that most genetic variants in the input dataset have zero coefficient. For example, the snpnet PRS for standing height, a classic example of polygenic traits, includes 51,209 variants, which has non-zero coefficients for 4.7% of 1,080,968 genetic variants and allelotypes present in the input genetic data. Moreover, this approach does not require the explicit specification of the underlying genetic architecture of traits, suitable for a phenome-wide application of PRS modeling. Using individuals in a hold-out test set, we evaluated their predictive performance and their statistical significance, resulting in 813 significant ($p < 2.5 \times 10^{-5}$) PRS models. We find a significant correlation between the number of the genetic variants selected in the model and the incremental predictive performance compared to the covariate-only models across quantitative traits and binary traits. We assess the transferability of the PRS models across ancestry groups using individuals from non-British white, African, South Asian, and East Asian ancestry in the UK Biobank. We make the coefficients of the PRS models publicly available via the PRS map web application on the Global Biobank Engine[20] (https://biobankengine.stanford.edu/prs).



(F) Transferability assessment in UK Biobank

Metric	Ancestry group	Predictive performance				n	case_n	control_n
		Geno	Covars	Full	delta			
Nagelkerke's pseudo-R^2	white British (model development)	0.240	0.113	0.322	0.209	269704	91884	177820
Nagelkerke's pseudo-R^2	white British (hold-out test set)	0.065	0.113	0.179	0.066	67425	22882	44543
Nagelkerke's pseudo-R^2	Non-British white	0.055	0.138	0.191	0.054	24905	7648	17257
Nagelkerke's pseudo-R^2	South Asian	0.038	0.175	0.213	0.038	7831	3161	4670
Nagelkerke's pseudo-R^2	East Asian	0.055	0.177	0.226	0.048	1704	446	1258
Nagelkerke's pseudo-R^2	African	0.011	0.175	0.177	0.003	6497	2919	3578
Tjur's pseudo-R^2	white British (model development)	0.181	0.082	0.246	0.165	269704	91884	177820
Tjur's pseudo-R^2	white British (hold-out test set)	0.047	0.082	0.132	0.050	67425	22882	44543
Tjur's pseudo-R^2	Non-British white	0.040	0.100	0.141	0.041	24905	7648	17257
Tjur's pseudo-R^2	South Asian	0.029	0.134	0.163	0.029	7831	3161	4670
Tjur's pseudo-R^2	East Asian	0.039	0.130	0.168	0.039	1704	446	1258
Tjur's pseudo-R^2	African	0.008	0.135	0.136	0.001	6497	2919	3578

Figure 1. Significant sparse polygenic risk scores (PRSs) across 813 traits in the UK Biobank. (A) We analyzed a total of more than 378,000 unrelated individuals and 1,565 traits in UK Biobank. We used 80 % of individuals of white British ancestry for score development. For evaluation, we used the remaining 20 % of individuals and additional individuals in other ancestry groups. (B) The full list of 1,565 traits with predictive performance is shown as a sortable table. (C) The predictive performance of PRS models for quantitative traits is summarized as a heatmap comparing the predicted risk score (Z-score) and observed trait value (left) and mean and standard error of trait values stratified by percentile bin (right). (D) The predictive performance of PRS models for binary traits is summarized as PRS score distribution stratified by case/control status (left) and odds ratio stratified by percentile bin (right). (E) The non-zero coefficients of the sparse PRS model are shown. (F) The predictive performance evaluation in training and test sets consist of individuals of white British ancestry, as well as additional sets consisting of individuals from non-British white, African, South Asian, and East Asian ancestry groups in the UK Biobank.

Results

Characterizing sparse PRS models with BASIL algorithm

To build sparse PRSs across a wide range of phenotypes, we compiled a total of 1,565 traits in the UK Biobank. We grouped them into trait categories, such as disease outcomes, anthropometry measures, and cancer phenotypes (**S1 Table**, **Methods**). We analyzed a total of 1,080,968 genetic variants and allelotypes from the directly-genotyped variants[19], imputed HLA allelotypes[21], and copy number variants[22]. Using 80% (n = 269,704) of unrelated individuals of white British ancestry, we applied batch screening iterative lasso (BASIL) implemented in the R *snpnet* package[10]. This recently developed method characterizes PRS models by simultaneously performing variable selection and effect size estimation. Applying different levels of penalization in the Lasso regression with penalty factors, we prioritized the medically relevant alleles in the PRS model. Specifically, we used the predicted consequence of the genotyped variants and the pathogenicity information in the ClinVar database. We prioritized protein-truncating variants, protein-altering variants, imputed HLA allelotype, and known pathogenic and likely-pathogenic variants by assigning lower penalty factors (**Methods**). As unpenalized covariates, we included age, sex, and the loadings of the top ten principal components (PCs) of genotypes. For 35 blood and urine biomarker traits, we took the snpnet PRS models from a recently published study[23], where the PRS models were characterized with the same methods on the same set of individuals following the adjustment for an extensive list of technical covariates, including fasting time and dilution factors, as well as for age, sex, and genotype PCs.

To evaluate the predictive performance (R^2 for quantitative traits and observed-scale Nagelkerke's pseudo-R^2 [also known as Cragg and Uhler's pseudo-R^2][24,25] for binary traits) and its significance of PRS models, we focused on the remaining 20 % of unrelated individuals in the hold-out test set (n = 67,425) as well as additional sets of unrelated individuals in the following ancestry groups in UK Biobank: non-British European (non-British white, n = 24,905), African (n = 6,497), South Asian (n = 7,831), and East Asian (n = 1,704) (**S2 Table**, **Methods**). We found 813 PRS models with significant ($p < 2.5 \times 10^{-5} = 0.05/2{,}000$, adjusted for multiple hypothesis testing with Bonferroni method) predictive performance in the hold-out test set of white British individuals (**Methods**). For the binary traits, we also evaluated the receiver operating characteristic area under the curve [ROC-AUC] and Tjur's Coefficient of Discrimination (Tjur's pseudo-R^2)[26].

The participants of the UK Biobank were genotyped on two different arrays: about 10 % of participants were genotyped on the UK BiLEVE Axiom array, whereas the rest were genotyped on the UK Biobank Axiom array[19]. To account for the potential biases correlated with the types of arrays, we evaluated the predictive performance of the PRS by accounting for the types of the arrays in addition to the age, sex, and the top ten genotype PCs. We found the identity of the UK Biobank assessment centers mostly has a non-significant impact on the predictive performance (**S1 Fig**, **Methods**).

To assess the degree of prioritization of the medically relevant alleles, we selected standing height, body mass index (BMI), high cholesterol, and asthma. We compared the predictive performance and the number of genetic variants for each functional category. For the four selected traits, we found a little difference in the predictive performance (R^2 = 0.177 vs. 0.176 for the PRS model with penalty factor and without penalty factor, respectively, for standing height, R^2 = 0.111 vs. 0.111 for BMI, AUC = 0.620 vs. 0.619 for high cholesterol, and AUC = 0.617 vs. 0.617 for asthma) (**S2 Fig**) while we saw an

enrichment of the number of the medically relevant alleles with non-zero coefficients in the PRS model with prioritization (2.14 fold enrichment standing height, 2.75 fold for BMI, 4.14 fold for high cholesterol, and 4.33 fold asthma) (**Table 1**, **S3 Table**), highlighting the flexibility of the BASIL/snpnet in assigning different levels of penalization based on the variant-level information.

With the same set of four traits, we asked whether including the imputed genetic variants could improve the predictive performance. We saw some gain in the predictive performance in three traits but not for standing height (**S3 Fig**). Based on those results, we decided to move on to the phenome-wide application of the BASIL algorithm implemented in the R snpnet packages on the directly genotyped variants, imputed allelotypes, and copy number variants while prioritizing the medically relevant alleles with penalty factors.

Table 1. The prioritization of the medically relevant alleles with penalty factors. The numbers of the genetic variants or allelotypes with non-zero coefficient values are shown for the four traits. The denominator represents the total number of variables included in the model. The numerator represents the number of the medically relevant alleles, which are one of the following: protein-truncating variants, protein-altering variants, imputed HLA allelotypes, the pathogenic or likely-pathogenic variants in the ClinVar database. The enrichment of the medically relevant variants is also shown.

Trait	Number of selected genetic variants or allelotypes		
	with penalty factor	without penalty factor	enrichment
Standing height	4187 / 51209	2129 / 55937	2.15
Body mass index	2543 / 27126	977 / 28667	2.75
High cholesterol	969 / 5987	215 / 5506	4.14
Asthma	1022 / 6430	250 / 6819	4.34

Significance and estimated effect size of sparse PRS models

We estimated the SNP-based heritability by applying linkage disequilibrium (LD) score regression (LDSC)[27] on genome-wide association study (GWAS) summary statistics. We compared it against the predictive performance (R^2 for quantitative traits and Nagelkerke's pseudo-R^2 for binary traits) of the significant PRS models (**Fig 2**). Across 244 binary traits and 569 quantitative traits with significant PRS models, we found higher estimated observed-scale heritability for quantitative traits. Overall, we found a significant correlation between the estimated SNP-based observed-scale heritability and predictive performance (Spearman's rank correlation coefficient ρ = 0.44, p-value = 3.5 x 10^{-13} for binary traits, ρ = 0.46, p-value = 1.4 x 10^{-31} for quantitative traits).

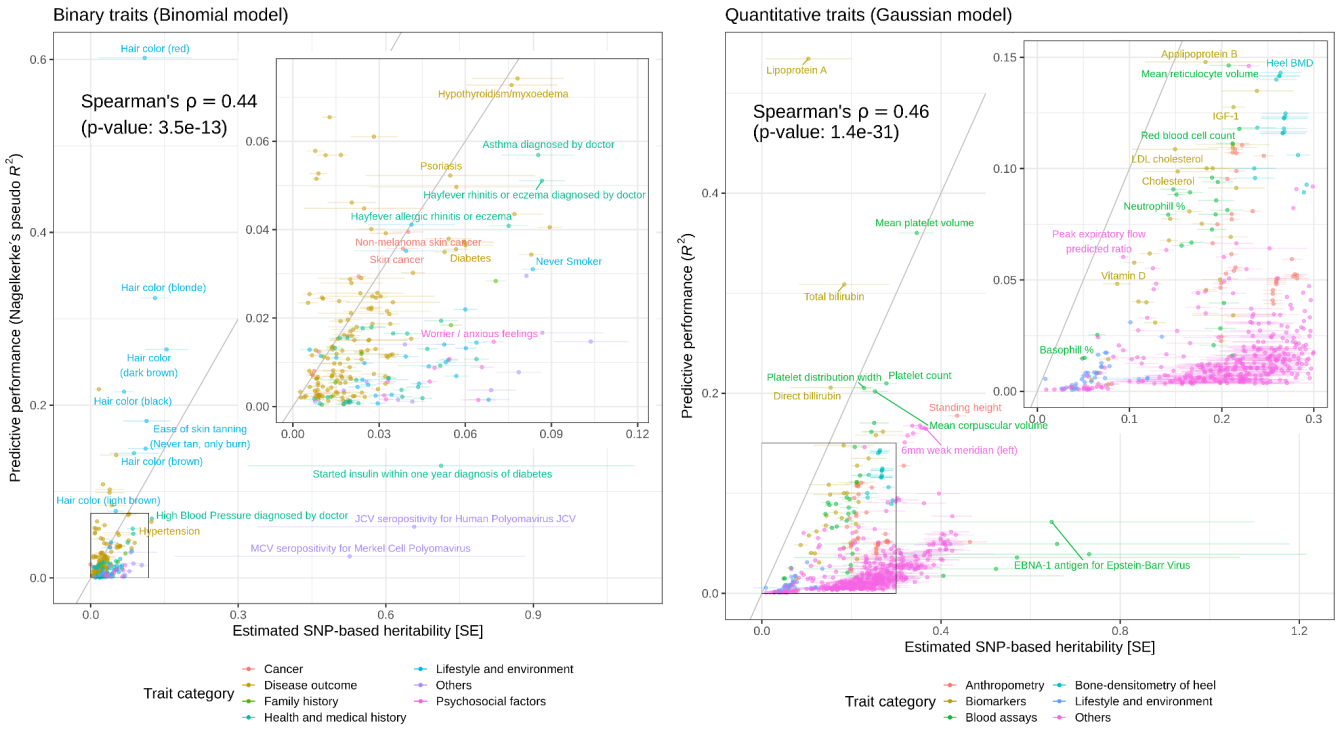


Figure 2. Comparison of the estimated SNP-based heritability and predictive performance across the 813 traits with significant PRSs. The predictive performance (Nagelkerke's pseudo-R^2 for 244 binary traits [left] and R^2 for 569 quantitative traits [right]) of the PRS models that only consider genetic variants are compared against the estimated SNP-based heritability. The solid gray lines represent y = x. We show the points on the bottom left corners in the inset plots. The error bars represent standard error. BMD: Bone mineral density.

The basic covariates alone are already informative for phenotype prediction. To assess the incremental utility of PRSs, we quantified the incremental predictive performance by comparing the predictive performance of the full model that considers both genotypes and covariates and that of the covariate-only model across the 813 traits with significant sparse PRS. We found most traits have a modest increase in the effect sizes of the prediction with a few notable exceptions, such as celiac disease (Nagelkerke's pseudo-R^2 = 0.149 in the full model vs 0.006 in the covariate-only model, p = 3.8 x 10^{-162}), hair color (red) (Nagelkerke's pseudo-R^2 = 0.603 vs. 0.008, p < 1 x 10^{-300}), mean platelet volume (R^2 = 0.36 vs. 0.001, p < 1 x 10^{-300}), heel bone mineral density (R^2 = 0.20 vs. 0.06, p < 1 x 10^{-300}), and blood and urine biomarker traits[23] (**Fig 3**, **Fig 4**).

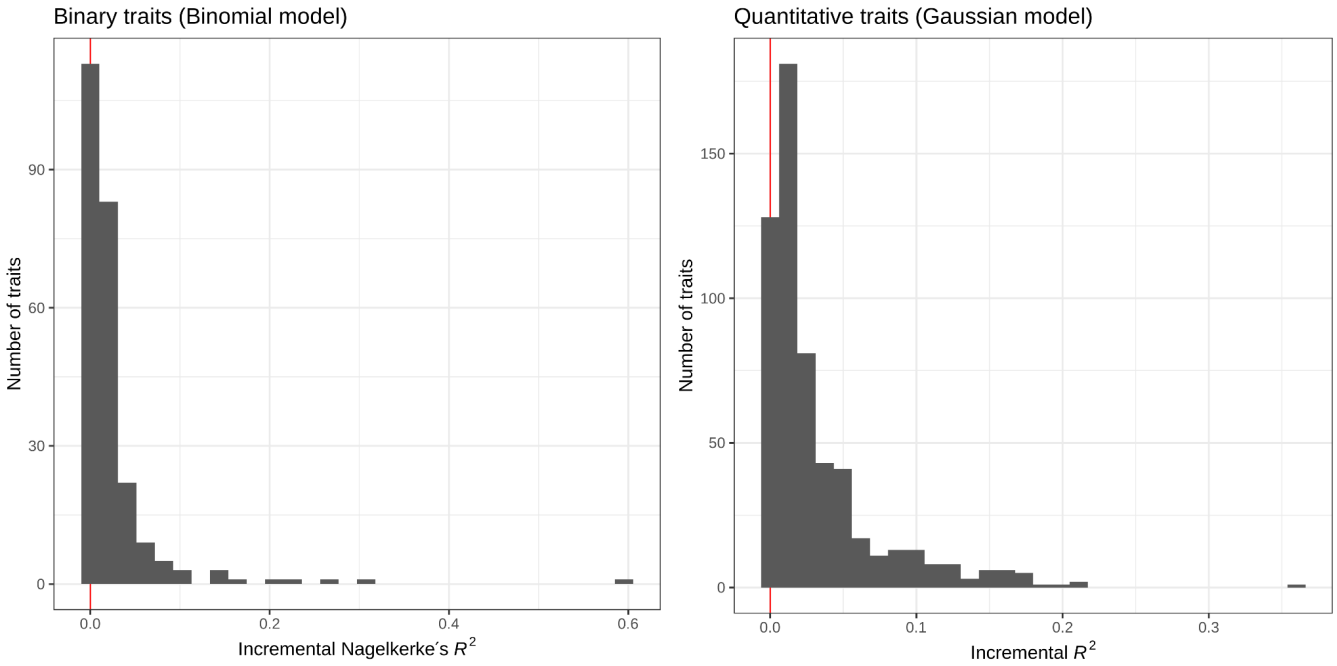


Figure 3. Incremental predictive performance of the PRS model across the 813 traits with significant predictive performance in the hold-out test set individuals of white British ancestry. The predictive performance (Nagelkerke's pseudo-R^2 for 244 binary traits [left] and R^2 for 569 quantitative traits [right]) of the full models that consider both the genotype and covariates are compared against that of the covariate-only models, and their difference (the incremental predictive performance) are shown as a histogram.

Sparse PRS models offer an interpretation of genomic loci underlying the polygenic risk

Celiac disease is an autoimmune disorder that affects the small intestine from gluten consumption. The sparse PRS model for this trait, for example, consists of 428 variants that contain the imputed HLA allelotypes and variants near the MHC region in chromosome 6[19,21]. The PRS model also contains genetic variants in all other autosomes, including a previously implicated missense variant in chromosome 12 (rs3184504, log(OR) = 0.15 in multivariate PRS model) in *SH2B3*. This gene encodes SH2B adaptor protein 3, which is involved in cellular signaling, hematopoiesis, and cytokine receptors[28] (**Fig 4**).

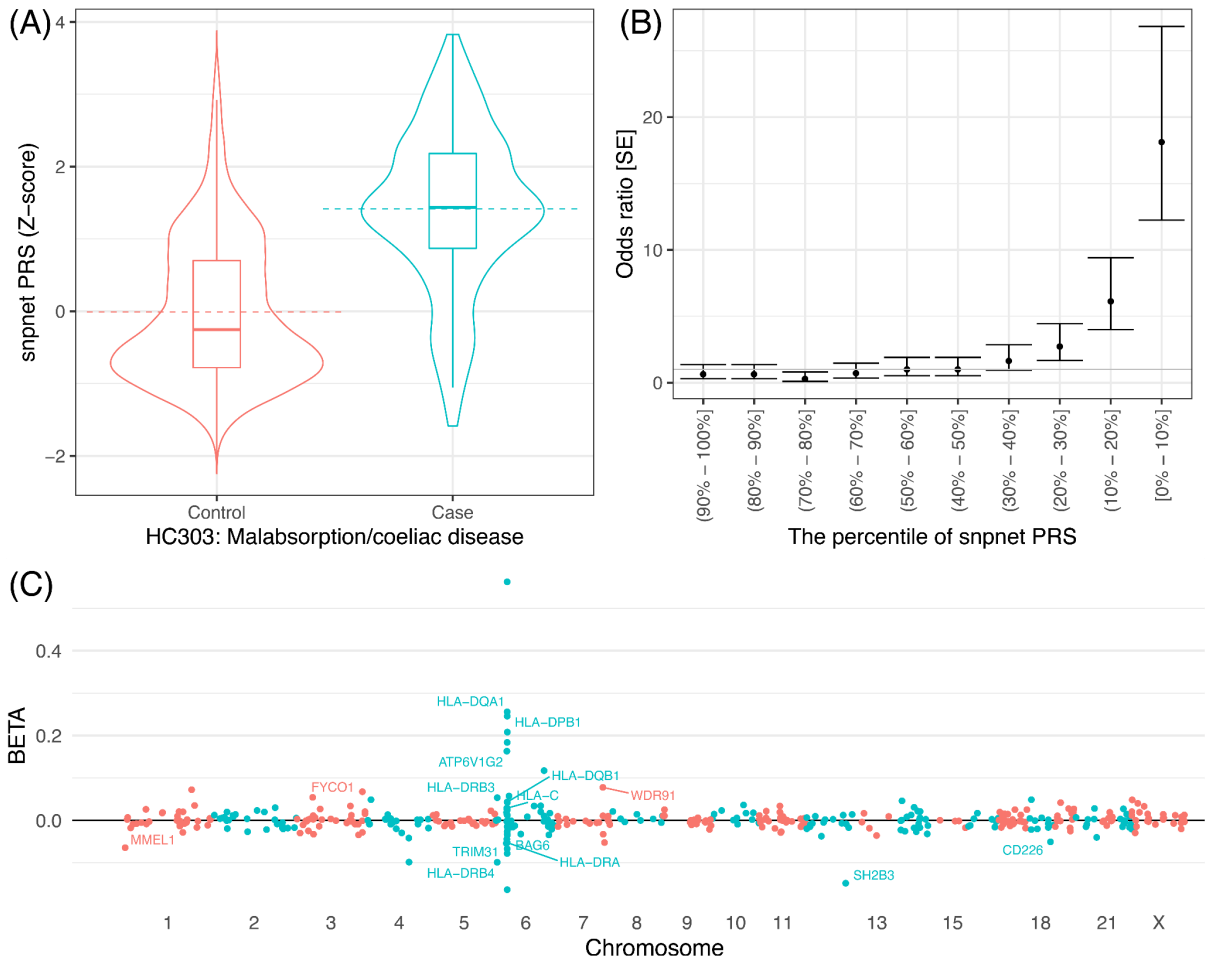


Figure 4. The sparse PRS model and their predictive performance for celiac disease. (A, B) the predictive performance of celiac disease PRS. (A) the celiac disease PRS distribution (y-axis) in a hold-out test set stratified by the disease case status (x-axis). The dashed lines represent the mean and the quantiles are shown as box plots. (B) The disease prevalence odds ratio compared to the individuals with middle (40-60 th percentile) PRS score stratified by PRS percentile bins. The error bars represent standard error (SE). (C) the coefficients of the celiac disease PRS model. The estimated effect size (y-axis) for each genetic variant (x-axis) is shown. The gene symbols are annotated in the plot for coding variants and HLA allelotypes with large effect size estimates.

The size of the PRS model is correlated with the incremental predictive performance

The significant PRS models have a wide range of the number of variables selected in the model, ranging from only one variable for iritis PRS (HLA allelotype, HLA-B*27:05, at the well-established HLA-B*27 locus[29,30]) to 51,209 variants selected for standing height PRS (**Fig 5**). We examined whether there is a relationship between the number of active variables in the significant PRS model and the incremental predictive performance. The significant correlation between the two quantities is stronger in quantitative (Spearman's rank correlation coefficient $\rho = 0.61$, $p = 2.2 \times 10^{-59}$) traits than in binary ($\rho = 0.21$, $p = 9.6 \times 10^{-4}$), reflecting the difference in power between binary and quantitative traits[31].

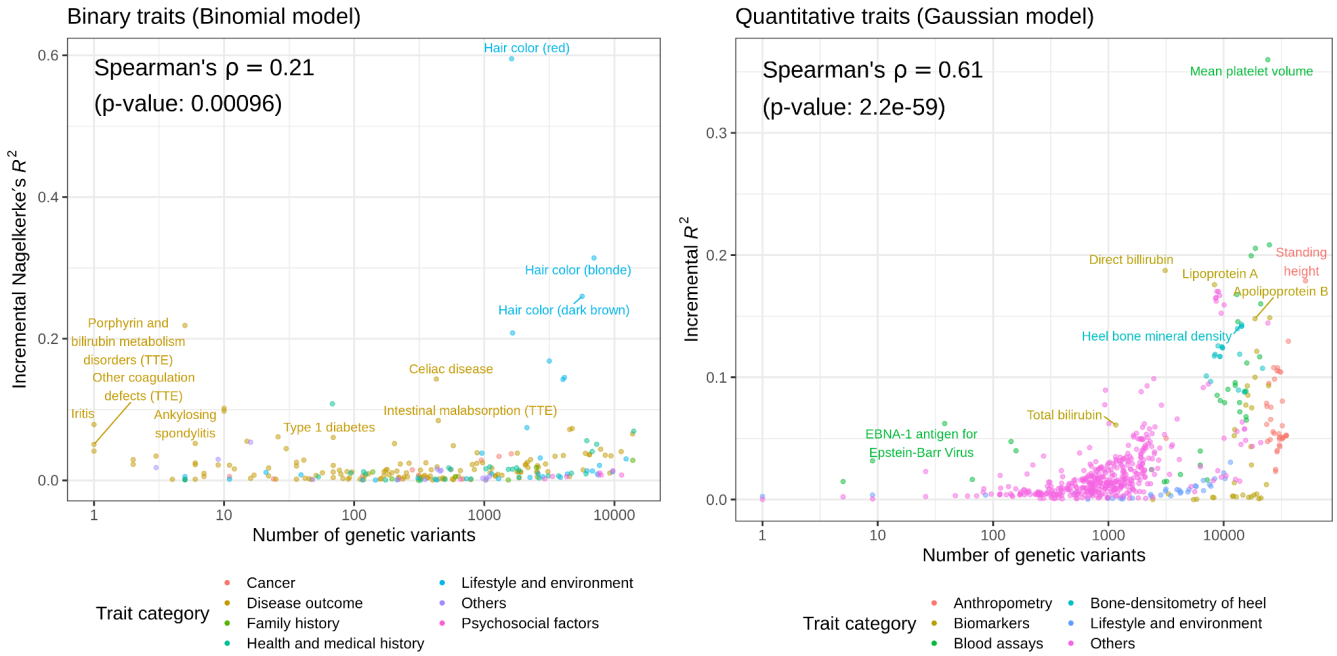


Figure 5. Comparison of the effect size and the model size of sparse PRS. The number of the genetic variants included in the model (size of the model, x-axis) and the incremental predictive performance (effect size of the model, y-axis) are shown for 244 binary traits (left) and 569 quantitative traits (right). TTE: time-to-event phenotype.

Sparse PRS models exhibit limited transferability across ancestry groups

While the majority of the participants in the UK Biobank are of European ancestry, the inclusion of individuals from African and Asian ancestry enables an assessment of the transferability of the PRS models across ancestry groups in UK Biobank. In addition to the hold-out test set that we derived from the white British population, we focused on additional sets of individuals from non-British European (non-British white), African, South Asian, and East Asian ancestry groups and compared the incremental predictive performance with that in white British hold-out test set (**Fig 6**). For quantitative traits, the models predicted well for non-British white (linear regression fit of the incremental predictive performance: $y = 0.91x$), but they suffer limited transferability for the non-European ancestry groups ($y = 0.56x$, $y = 0.47x$, and $y = 0.13x$ for South Asian, East Asian, and African, respectively). Similarly, in binary traits, the non-British white showed higher transferability ($y = 0.80\ x$) than the non-European ancestry groups ($y = 0.027x$, $y = 0.059x$, and $y = -0.145x$ for South Asian, East Asian, and African, respectively).

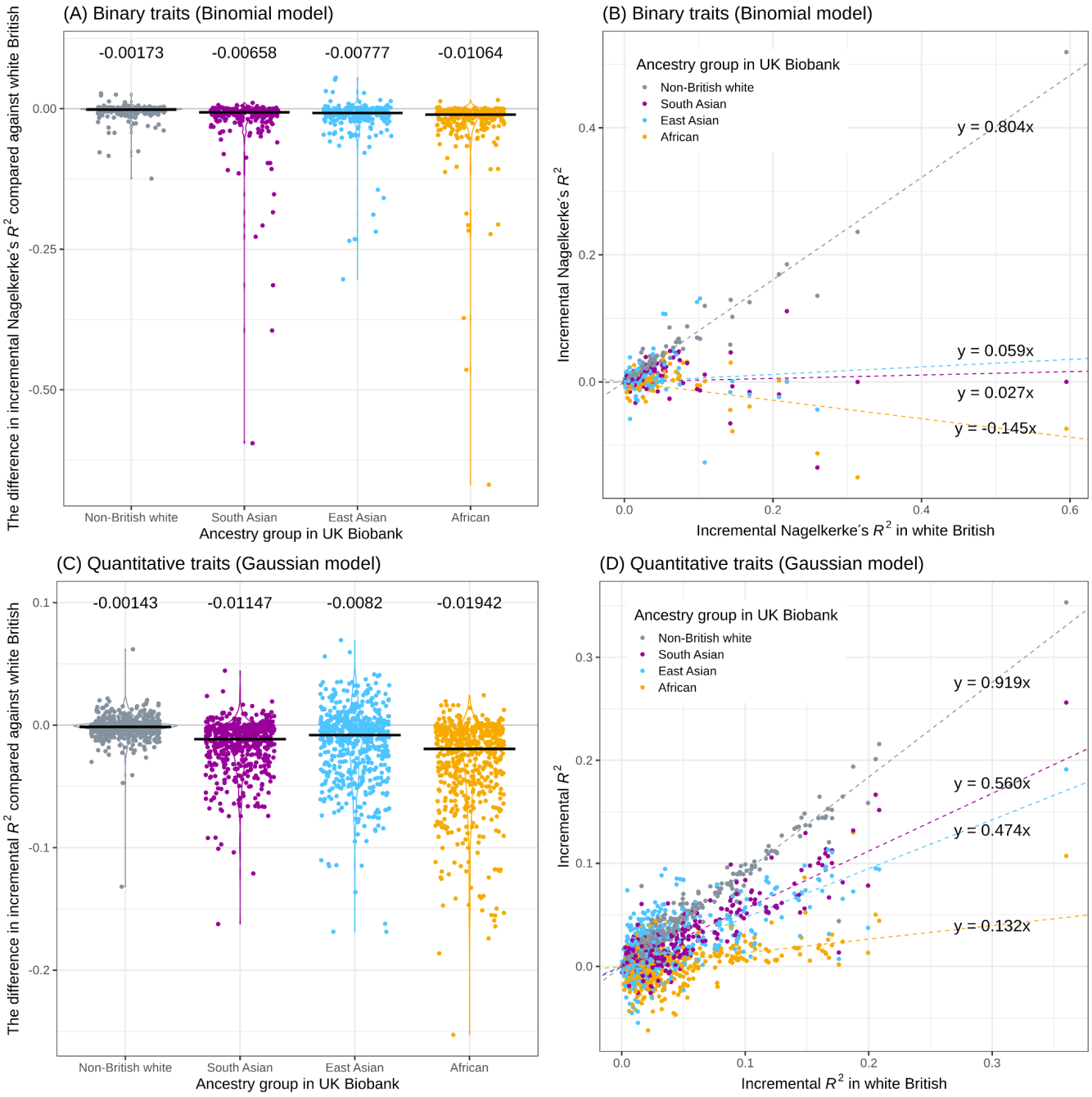


Figure 6. Transferability assessment of PRS models across ancestry groups in the UK Biobank. The incremental predictive performance (Nagelkerke's pseudo-R^2 for 244 binary traits (A, B) and incremental R^2 for 569 quantitative traits (C, D)) was quantified in individuals in different ancestry groups in the UK Biobank and was compared against that in the hold-out test set constructed from the individuals in white British ancestry group. (A, C) the difference in the incremental predictive performance between the target group (x-axis, double-coded with color) and the source white British cohort. The median values are shown as black horizontal bars and numbers. (B, D) comparison of the incremental predictive performance in the target group (color) and the test set. A simple linear regression fit was shown for each ancestry group with the dashed lines. The slopes of the regression lines were also shown.

Discussion

In this study, we performed a systematic scan of polygenic prediction across more than 1,500 traits and reported 813 significant sparse PRS models. We found a correlation between the predictive performance of the significant PRS models and SNP-based heritability estimates. We assessed the effect size of the PRS model by quantifying the incremental predictive performance, which we define as the difference in the predictive performance between the covariate-only model and the full model consisting of both covariates and genetics. In both quantitative and binary traits, we find a significant correlation between the number of independent loci included in the model and their incremental predictive performance.

Our study is complementary to many other studies that focus on fewer traits to construct PRS models from GWAS meta-analysis and mixed models. While the sample size in our study is sufficiently large to observe statistical significance in predictive performance across hundreds of traits, it does not necessarily mean the clinical relevance of the PRS models. Moreover, population-based recruitment in UK Biobank may not be the best strategy to achieve the highest predictive performance for some traits. A disease-focused study[6,32–34] would be an attractive alternative strategy, especially when multiple genotyped cohorts recruited for the same disease are available or the disease of interest has a low prevalence. Our study, instead, focused on the phenome-wide application of PRS across hundreds of traits in a single cohort by applying BASIL algorithm with readily available implementation in R *snpnet* package[10], which does not require explicit modeling of underlying genetic architecture across a wide variety of traits.

For binary traits, we used observed-scale pseudo-R^2 and observed-scale SNP-based heritability estimates, given that population prevalence is available for only a subset of binary traits considered in the present study. Conversion to liability-scale estimates will further enhance the validity of the comparison[35] and is of interest for future investigation.

Like other PRS approaches that consider datasets from one source population in the PRS training, our sparse model trained on the individual-level data of white British showed limited transferability across diverse ancestry groups[36–38]. The sample sizes of non-European ancestry groups in UK Biobank are smaller than that of European ancestry groups. In general, that will result in larger uncertainties in predictive performance assessment. Nonetheless, when we assess the incremental predictive performance across ancestry groups by comparing the full model consisting of the genetic data and basic covariates and the covariate-only model, we found the binary traits, including disease outcomes, have lower transferability compared to quantitative traits, including biomarkers, blood measurements, and anthropometric traits. The power difference between binary and quantitative traits[31], limitation in power for some traits, especially for the binary traits with limited case counts, and differences in heritability may be the contributing factors of the observed difference. Improvements of PRS models with high transferability across ancestry groups and the admixed individuals are of interest for future research.

Given the medical relevance[39–51], we prioritized pathogenic and likely-pathogenic variants reported in ClinVar[52] as well as predicted protein-truncating and protein-altering variants (**Methods**). Our analysis focusing on four traits suggests that prioritizing the medically relevant alleles does not necessarily improve the predictive performance. While our sparse PRS models show enrichment in the number of selected medically relevant alleles, there is no guarantee that the genetic variants included

in the sparse PRS models were causal. It warrants further follow-up analysis with statistical fine-mapping and detailed functional characterization at each locus.

The increased availability of PRS models across multiple traits[17] exhibits a wide range of applications, including the improved genetic risk prediction of disease[23,53] and the identification of causal relationships across complex traits[54]. We provide the results on the Global Biobank Engine (https://biobankengine.stanford.edu/prs) as well as on the PGS catalog[17] and envision the resource will serve as an important basis to understand the polygenic basis of complex traits.

Methods

Compliance with ethical regulations and informed consent

This research has been conducted using the UK Biobank Resource under Application Number 24983, "Generating effective therapeutic hypotheses from genomic and hospital linkage data" (http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf). Based on the information provided in Protocol 44532, the Stanford IRB has determined that the research does not involve human subjects as defined in 45 CFR 46.102(f) or 21 CFR 50.3(g). All participants of the UK Biobank provided written informed consent (more information is available at https://www.ukbiobank.ac.uk/2018/02/gdpr/).

Study population and genetic data

UK Biobank is a population-based cohort study collected from multiple sites across the United Kingdom[18]. To minimize the variabilities due to population structure in our dataset, we restricted our analyses to unrelated individuals based on the following four criteria[46,55] reported by the UK Biobank in sample QC file, "ukb_sqc_v2.txt": (1) used to compute principal components ("used_in_pca_calculation" column); (2) not marked as outliers for heterozygosity and missing rates ("het_missing_outliers" column); (3) do not show putative sex chromosome aneuploidy ("putative_sex_chromo- some_aneuploidy" column); and (4) have at most ten putative third-degree relatives ("excess_relatives" column).

Using a combination of genotype principal components (PCs), the self-reported ancestry (UK Biobank Field ID 21000, https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=21000), and "in_white_British_ancestry_subset" column in the sample QC file from UK Biobank, we subsequently focused on people of self-identified white British (n = 337,129), self-identified non-British white (n = 24,905), African (n = 6,497), South Asian (n = 7,831), and East Asian (n = 1,704) ancestry as described elsewhere[23]. Briefly, we used a two-step procedure to define the five groups. We first used the genotype principal component loadings of the individuals and set thresholds on component 1 and component 2 as follows: (1) self-identified White British: $-20 \leq PC1 \leq 40$ and $-25 \leq PC2 \leq 10$ and in_white_British_ancestry_subset == 1; (2) self-identified non-British White: $-20 \leq PC1 \leq 40$, $-25 \leq PC2 \leq 10$, has a self-reported ancestry of White, and does not identify themselves as White British; (3) African: $260 \leq PC1$, $50 \leq PC2$, and does not identify themselves as any of the following: Asian, White, Mixed, or Other population groups; (4) South Asian: $40 \leq PC1 \leq 120$, $-170 \leq PC2 \leq -80$, and does not identify themselves as any of the following: Black, White, Mixed, or Other population groups; and (5) East Asian: $130 \leq PC1 \leq 170$, $PC2 \leq -230$, and does not identify themselves as any of the following: Black, White, Mixed, or Other population groups. To refine the population definition by removing the outliers, we computed population-specific genotype PCs using approximately LD independent ($R^2 < 0.5$) common (population-specific minor allele frequency > 5%) biallelic variants outside of the major histocompatibility complex region[23]. We applied following thresholds[23]: (1) South Asian: $-0.02 \leq$ population-specific $PC1 \leq 0.03$, $-0.05 \leq$ population-specific $PC2 \leq 0.02$; and (2) East Asian: $-0.01 \leq$ population-specific $PC1 \leq 0.02$, $-0.02 \leq$ population-specific $PC2 \leq 0$.

We randomly split the white British cohort into 70% training (n = 235,991), 10% validation (to select the optimal sparsity level) (n = 33,713), and 20% test (n = 67,425) sets[23,56]. We used the same split of

training, validation, and test set for all tested traits. The non-British white, African, South Asian, and East Asian samples were only used as test sets.

Variant quality control and variant annotation

We used genotype datasets (release version 2 for the directly genotyped variants and the imputed HLA allelotype datasets)[19], the CNV dataset[22], and the hg19 human genome reference for the main PRS analyses in the study. Additionally, we considered imputed variants (release version 3) to investigate whether the imputed variants would improve the predictive performance. We annotated the directly-genotyped variants using Ensembl's Variant Effect Predictor (VEP) (version 101)[57,58] with the LOFTEE plugin (https://github.com/konradjk/loftee)[49], for which we created a Docker container image (https://github.com/yk-tanigawa/docker-ensembl-vep-loftee). Using ClinVar (version 20200914)[28], we annotated "pathogenic" and "likely pathogenic" variants.

We performed variant quality control as described elsewhere[23,46,55]. Briefly, we focused on the variants passing the following criteria: (1) outside of the major histocompatibility complex (MHC) region (hg19 chr6:25477797-36448354); (2) the missingness of the variant is less than 1%, considering that the two genotyping arrays (the UK BiLEVE Axiom array and the UK Biobank Axiom array) cover a slightly different set of variants[19]; (3) the minor-allele frequency is greater than 0.01%; (4) Hardy-Weinberg disequilibrium test p-value is less than 1.0×10^{-7}; (5) Passed the comparison of minor allele frequency with the gnomAD dataset (version 2.0.1) as described before[46,49]; (6) We manually investigated the cluster plots for a subset of variants and removed 11 variants that have unreliable genotype calls[46].

We grouped the VEP-predicted consequence of the variants into six groups: protein-truncating variants (PTVs), protein-altering variants (PAVs), proximal coding variants (PCVs), Intronic variants (Intronic), variants in the untranslated region (Intronic), and other variants (Others). Our grouping rule of the VEP-predicted consequence is summarized in (**S4 Table**).

We included the imputed copy number variants (CNVs)[22] and imputed HLA allelotypes[21]. The CNVs were called using PennCNV (v.1.0.4)[59] on raw signal intensity data from each genotyping array as described elsewhere[22]. Because the precise location of the CNVs is not identified, we did not infer the functional consequences of CNVs with variant annotation. The HLA allelotypes at HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRB1, -DRB3, -DRB4, and -DRB5 loci were imputed using the HLA*IMP:02 and imputed dosage file is provided by the UK Biobank. We included 156 alleles across all 11 loci that had a frequency of 0.1% or greater in the white British. We rounded allele dosage when they were within plus or minus 0.1 of 0, 1, or 2. We excluded the remaining nonzero entries. We also excluded erroneous total allele counts post-rounding[21].

When evaluating whether the inclusion of the imputed variants would improve the predictive performance of the PRS models, we focused on the 5,931,362 imputed variants[19] based on the following criteria: (1) imputation INFO score is greater than 0.7, (2) minor allele frequency computed across the entire ~500k genotyped samples (UK Biobank Resource 1967, https://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=1967) is greater than 0.01, (3) biallelic variants, (4) the variant is not present in the directly genotyped variants, and (5) missingness is less than 1%. We subsequently combined the imputed variant dataset with the directly genotyped variants, imputed HLA allotypes, and copy number variants.

Phenotype definitions in the UK Biobank

We analyzed a wide variety of traits in the UK Biobank, including disease outcome[46,60], family history [46,60], cancer registry data[46], blood and urine biomarkers[23], hematological measurements, and other binary and quantitative phenotypes[55,56]. Some phenotype information collected at UK Biobank's assessment center contains up to four instances, each of which corresponds to (1) the initial assessment visit (2006-2010), (2) first repeat assessment visit (2012-2013), and (3) imaging visit (2014-), and (4) first repeat imaging visit (2019-). Briefly, for binary traits, we performed manual curation of phenotypic definitions and assigned "case" status if the participants are classified as the case in at least one of their visits and "control" otherwise. For quantitative traits, we took the median of non-NA values, as described elsewhere[55].

Previously, we analyzed blood and urine biomarker traits, investigating the effects of covariates on the biomarker levels and derived covariate-adjusted biomarker values[23]. Briefly, we used a linear regression model to account for the covariate effects on the log-transformed measurement values from UK Biobank and adjusted for principal component loadings of genotype, age, sex, age by sex interactions, self-identified ancestry group, self-identified ancestry group by sex interactions, fasting time, estimated sample dilution factor, assessment center indicators, genotyping batch indicators, time of sampling during the day, the month of assessment, and day of the assay. We used the PRS models trained for the covariate-adjusted traits[23]. To quantify the incremental predictive performance against the covariate-only models, we quantified predictive performance against the original measurement values, except eGFR, AST/ALT ratio, and non-albumin protein, where we used the covariate-adjusted trait values. Those three traits are derived from covariate-adjusted biomarkers[23] and do not have raw measurement values.

The list of 1,565 traits with at least 100 cases (for binary traits) or non-NA measurements (for quantitative traits) analyzed in this study is listed in (**S1 Table**).

Construction of sparse PRS models

Using the batch screening iterative Lasso (BASIL) algorithm implemented in the R *snpnet* package[10], we constructed the sparse PRS models for the 1,565 traits. We used the Gaussian family and the R^2 metric for quantitative traits, whereas we used the binomial family and the AUC-ROC metric for the binary traits[10]. For each trait, we fit a series of regression models with a varying degree of sparsity on the training set, consisting of 70% (n = 235,991) of unrelated individuals of white British ancestry. The predictive performance of each of the models is evaluated on the validation set, which consists of 10% (n = 33,713) of unrelated individuals of white British ancestry to guide the selection of the optional level of sparsity. We selected the sparsity that maximizes the predictive performance in the validation set. We subsequently refit the penalized regression model using the individuals in the combined training and validation set individuals (n = 269,704), which we denote as score development set, to maximize the power in the regression model[10]. We used the same training, validation, and test set split across all the PRS models analyzed in this study.

As opposed to many PRS methods that operate on the GWAS summary statistics[3–9,13–15], our method takes individual-level genotype and phenotype data. Using L_1 penalized regression (also known as Lasso), BASIL simultaneously performs variable selection and effect size estimation of the selected variants. We included age, sex, and top ten population-specific genotype PC loadings computed for the

white British individuals[23] as unpenalized covariates. Thanks to the L_1 penalty term in the objective function that penalizes the number of features of non-zero regression coefficients, the resulting models will be sparse, meaning that they will have fewer genetic variants than unpenalized models[10].

To prioritize coding variants over non-coding variants in linkage, we assigned three levels of penalty factors (also known as penalty scaling parameter)[61]: 0.5 for pathogenic variants in ClinVar[52] or protein-truncating variants according to VEP-based variant annotation[58]; 0.75 for likely pathogenic variants in ClinVar, VEP-predicted protein-altering variants, or imputed allelotypes; and 1.0 for all other variants. The assignment rules of the penalty factors are summarized in (**S5 Table**). The variants with lower values of penalty factors are prioritized in the L_1 penalized regression. To assess the degree of prioritization of the medically relevant alleles and their impacts on the predictive performance, we focused on four traits (standing height, BMI, high cholesterol, and asthma) and fit a separate model without penalty factors. We compared the number of selected variants and the predictive performance.

Predictive performance and transferability of PRS models

We evaluated the predictive performance (R^2 for quantitative traits and Nagelkerke's pseudo-R^2 [also known as Cragg and Uhler's pseudo-R^2][24,25] for binary traits) of PRS models (**S6 Table**). For the binary traits, we also evaluated the receiver operating characteristic area under the curve [ROC-AUC] and Tjur's Coefficient of Discrimination (Tjur's pseudo-R^2)[26]. For R^2 and ROC-AUC, we evaluated the 95% confidence interval of predictive performance using approximate standard error of R^2[62,63] and DeLong's method[64], respectively. We used the individuals in the hold-out test set (n = 67,425) of white British ancestry as well as additional sets of individuals in non-British white (n = 24,905), African (n = 6,497), South Asian (n = 7,831), and East Asian (n = 1,704) ancestry groups. We evaluated the predictive performance of (1) the genotype-only model, (2) the covariate-only model, and (3) the full model that considers both covariates and genotypes. We computed the difference between the full model and the covariate-only model to derive the incremental predictive performance.

To evaluate the predictive performance of the covariate-only model in the hold-out test set of white British ancestry, we fit a generalized regression model, trait ~ age + sex + array + Genotype PCs, using the individuals in the score development set. We subsequently computed the risk scores based on the covariate terms for the individuals in the hold-out test set. The array is an indicator variable denoting the types of the genotyping array (either the UK BiLEVE Axiom array or the UK Biobank Axiom array). For the individuals in non-British white, African, South Asian, and East Asian ancestry groups, we took the ancestry group-specific PCs computed for each set[23] and fit the same regression model for each group. We did not use the array indicator variable for African, South Asian, and East Asian because all individuals in those ancestry groups were genotyped on the UK Biobank Axiom Array (**S2 Table**).

To evaluate the predictive performance of the genotype-only model, we computed the polygenic risk score for the sets of individuals for evaluation using the --score command implemented in plink2[65]. We quantified evaluation metrics (R^2, Nagelkerke's pseudo-R^2, ROC-AUC, and Tjur's pseudo-R^2).

To evaluate the predictive performance of the full model, we fit a model, trait ~ 1 + covariate-only score + PRS, using the covariate-only score and PRS described above. The constant term accounts for the potential differences in the trait mean (for quantitative traits) or case prevalence (for binary traits) between the score development population and the target population. We looked at the p-value reported for the PRS term for the statistical significance of the PRS model. We used $p < 2.5 \times 10^{-5}$ (=

0.05/2000, adjusted for multiple hypothesis testing using the Bonferroni method for the number of traits analyzed in the study) as the significance threshold.

We also computed the difference in R^2 or Nagelkerke's pseudo-R^2 between the full and covariate-only models to derive the incremental predictive performance.

SNP-based Heritability estimation

To compare the incremental predictive performance of the PRS models with SNP-based heritability, we applied genome-wide association analysis with PLINK. Specifically, we applied --glm command in PLINK[65] v2.00-alpha with age, sex, array, the number of CNVs, the length of CNVs, and the top ten genotype PC loadings as covariates. The array is an indicator variable denoting whether the UK Biobank Axiom array or UK BiLEVE Axiom array was used in the genotyping. We included this term if the variants were directly measured on both arrays. The number and the length of the CNVs are described elsewhere[22]. The genotype PCs are the principal component (PC) loadings of individuals. We computed the population-specific PCs using the unrelated individuals in white British and used the first 10 PCs[23]. In the regression analysis, we standardized the variance of the covariates (--covar-variance-standardize option) and applied quantile normalization for the quantitative phenotype (--pheno-quantile-normalize option). Note, we did not perform quantile normalization in the PRS analysis. We used “cc-residualize” and “firth-residualize” options that implement the approximation[66] for efficient computation of GWAS p-values. We subsequently applied linkage disequilibrium (LD) score regression (LDSC)[27] and characterized the SNP-based heritability (**S7 Table**). We compared the predictive performance of the PRS models and the LDSC-based heritability estimates.

Correlation analysis of the number of genetic variants and predictive performance of PRS models

We applied Spearman’s correlation test implemented in R to assess the rank correlation between the size (the number of genetic variants included in the model) and the effect size (the incremental predictive performance) of the PRS model.

Statistics

For computational and statistical analysis, we used Jupyter Notebook[67], R[68], R tidyverse package[69], and GNU parallel[70]. The p-values were computed from two-sided tests unless otherwise specified.

References

1. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020;12: 44. doi:10.1186/s13073-020-00742-5

2. Wray NR, Lin T, Austin J, McGrath JJ, Hickie IB, Murray GK, et al. From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. JAMA Psychiatry. 2021;78: 101–109. doi:10.1001/jamapsychiatry.2020.3049

3. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015;97: 576–592. doi:10.1016/j.ajhg.2015.09.001

4. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. Genet Epidemiol. 2017;41: 469–480. doi:10.1002/gepi.22050

5. Zhu X, Stephens M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. Ann Appl Stat. 2017;11: 1561–1592. doi:10.1214/17-aoas1046

6. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50: 1219–1224. doi:10.1038/s41588-018-0183-z

7. Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019;10: 1776. doi:10.1038/s41467-019-09718-5

8. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience. 2019;8: giz082. doi:10.1093/gigascience/giz082

9. Lloyd-Jones LR, Zeng J, Sidorenko J, Yengo L, Moser G, Kemper KE, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat Commun. 2019;10: 5086. doi:10.1038/s41467-019-12653-0

10. Qian J, Tanigawa Y, Du W, Aguirre M, Chang C, Tibshirani R, et al. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. PLoS Genet. 2020;16: e1009141. doi:10.1371/journal.pgen.1009141

11. Li R, Chang C, Justesen JM, Tanigawa Y, Qiang J, Hastie T, et al. Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank. Biostatistics. 2020; kxaa038. doi:10.1093/biostatistics/kxaa038

12. Li R, Chang C, Tanigawa Y, Narasimhan B, Hastie T, Tibshirani R, et al. Fast Numerical Optimization for Genome Sequencing Data in Population Biobanks. Bioinformatics. 2021; btab452. doi:10.1093/bioinformatics/btab452

13. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. Bioinformatics. 2020;36: 5424–5431. doi:10.1093/bioinformatics/btaa1029

14. Choi SW, Mak TS-H, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;15: 2759–2772. doi:10.1038/s41596-020-0353-1

15. Ojavee SE, Kousathanas A, Trejo Banos D, Orliac EJ, Patxot M, Läll K, et al. Genomic architecture and prediction of censored time-to-event phenotypes with a Bayesian genome-wide analysis. Nat Commun. 2021;12: 2337. doi:10.1038/s41467-021-22538-w

16. Wand H, Lambert SA, Tamburro C, Iacocca MA, O'Sullivan JW, Sillari C, et al. Improving reporting standards for polygenic scores in risk prediction studies. Nature. 2021;591: 211–219. doi:10.1038/s41586-021-03243-6

17. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat Genet. 2021;53: 420–425.

doi:10.1038/s41588-021-00783-5

18. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12: e1001779. doi:10.1371/journal.pmed.1001779

19. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562: 203–209. doi:10.1038/s41586-018-0579-z

20. McInnes G, Tanigawa Y, DeBoever C, Lavertu A, Olivieri JE, Aguirre M, et al. Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. Bioinformatics. 2018;35: 2495–2497. doi:10.1093/bioinformatics/bty999

21. Venkataraman GR, Olivieri JE, DeBoever C, Tanigawa Y, Justesen JM, Dilthey A, et al. Pervasive additive and non-additive effects within the HLA region contribute to disease risk in the UK Biobank. bioRxiv. 2020. p. 2020.05.28.119669. doi:10.1101/2020.05.28.119669

22. Aguirre M, Rivas MA, Priest J. Phenome-wide Burden of Copy-Number Variation in the UK Biobank. Am J Hum Genet. 2019;105: 373–383. doi:10.1016/j.ajhg.2019.07.001

23. Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars N, Benner C, Aguirre M, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. Nat Genet. 2021;53: 185–194. doi:10.1038/s41588-020-00757-z

24. Cragg JG, Uhler RS. The Demand for Automobiles. Can J Econ. 1970;3: 386–406. doi:10.2307/133656

25. Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika. 1991;78: 691–692. doi:10.1093/biomet/78.3.691

26. Tjur T. Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination. Am Stat. 2009;63: 366–372. doi:10.1198/tast.2009.08210

27. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47: 1228–1235. doi:10.1038/ng.3404

28. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet. 2011;43: 1193–1201. doi:10.1038/ng.998

29. Chang JH, McCluskey PJ, Wakefield D. Acute anterior uveitis and HLA-B27. Surv Ophthalmol. 2005;50: 364–388. doi:10.1016/j.survophthal.2005.04.003

30. Qi J, Li Q, Lin Z, Liao Z, Wei Q, Cao S, et al. Higher risk of uveitis and dactylitis and older age of onset among ankylosing spondylitis patients with HLA-B*2705 than patients with HLA-B*2704 in the Chinese population. Tissue Antigens. 2013;82: 380–386. doi:10.1111/tan.12254

31. Yang J, Wray NR, Visscher PM. Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. Genet Epidemiol. 2010;34: 254–257. doi:10.1002/gepi.20456

32. Nikpay M, Goel A, Won H-H, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015;47: 1121–1130. doi:10.1038/ng.3396

33. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. Genomic risk prediction of coronary artery disease in 480,000 adults: Implications for primary prevention. J Am Coll Cardiol. 2018;72: 1883–1893. doi:10.1016/j.jacc.2018.07.079

34. Mars N, Koskela JT, Ripatti P, Kiiskinen TTJ, Havulinna AS, Lindbohm JV, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat Med. 2020;26: 549–557. doi:10.1038/s41591-020-0800-0

35. Lee SH, Goddard ME, Wray NR, Visscher PM. A better coefficient of determination for genetic profile analysis. Genet Epidemiol. 2012;36: 214–224. doi:10.1002/gepi.21614

36. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet. 2017;100: 635–649. doi:10.1016/j.ajhg.2017.03.004

37. Kim MS, Patel KP, Teng AK, Berens AJ, Lachance J. Genetic disease risks can be misestimated across global populations. Genome Biol. 2018;19: 179. doi:10.1186/s13059-018-1561-7

38. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51: 584–591. doi:10.1038/s41588-019-0379-x

39. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet. 2005;37: 161–165. doi:10.1038/ng1509

40. Cohen JC, Boerwinkle E, Mosley TH Jr, Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med. 2006;354: 1264–1272. doi:10.1056/NEJMoa054013

41. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011;43: 1066–1073. doi:10.1038/ng.952

42. Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, et al. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science. 2015;348: 666–669. doi:10.1126/science.1261877

43. Rivas MA, Graham D, Sulem P, Stevens C, Desch AN, Goyette P, et al. A protein-truncating R179X variant in RNF186 confers protection against ulcerative colitis. Nat Commun. 2016;7: 12342. doi:10.1038/ncomms12342

44. Narasimhan VM, Hunt KA, Mason D, Baker CL, Karczewski KJ, Barnes MR, et al. Health and population effects of rare gene knockouts in adult humans with related parents. Science. 2016;352: 474–477. doi:10.1126/science.aac8624

45. Saleheen D, Natarajan P, Armean IM, Zhao W, Rasheed A, Khetarpal SA, et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature. 2017;544: 235–239. doi:10.1038/nature22034

46. DeBoever C, Tanigawa Y, Lindholm ME, McInnes G, Lavertu A, Ingelsson E, et al. Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. Nat Commun. 2018;9: 1612. doi:10.1038/s41467-018-03910-9

47. Emdin CA, Khera AV, Chaffin M, Klarin D, Natarajan P, Aragam K, et al. Analysis of predicted loss-of-function variants in UK Biobank identifies variants protective for disease. Nature Communications. 2018. doi:10.1038/s41467-018-03911-8

48. Tanigawa Y, Wainberg M, Karjalainen J, Kiiskinen T, Venkataraman G, Lemmelä S, et al. Rare protein-altering variants in ANGPTL7 lower intraocular pressure and protect against glaucoma. PLoS Genet. 2020;16: e1008682. doi:10.1371/journal.pgen.1008682

49. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581: 434–443. doi:10.1038/s41586-020-2308-7

50. Lam BYH, Williamson A, Finer S, Day FR, Tadross JA, Gonçalves Soares A, et al. MC3R links nutritional state to childhood growth and the timing of puberty. Nature. 2021. doi:10.1038/s41586-021-04088-9

51. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature. 2021. doi:10.1038/s41586-021-04103-z

52. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46: D1062–D1067. doi:10.1093/nar/gkx1153

53. Chung W, Chen J, Turman C, Lindstrom S, Zhu Z, Loh P-R, et al. Efficient cross-trait penalized regression increases prediction accuracy in large cohorts using secondary phenotypes. Nat Commun. 2019;10: 569. doi:10.1038/s41467-019-08535-0

54. Richardson TG, Harrison S, Hemani G, Davey Smith G. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. Elife. 2019;8: e43657. doi:10.7554/eLife.43657

55. Tanigawa Y, Li J, Justesen JM, Horn H, Aguirre M, DeBoever C, et al. Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. Nat Commun. 2019;10: 4064. doi:10.1038/s41467-019-11953-9

56. Aguirre M, Tanigawa Y, Venkataraman GR, Tibshirani R, Hastie T, Rivas MA. Polygenic risk modeling with latent trait-related genetic components. Eur J Hum Genet. 2021;29: 1071–1081. doi:10.1038/s41431-021-00813-0

57. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020;48: D682–D688. doi:10.1093/nar/gkz966

58. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17: 122. doi:10.1186/s13059-016-0974-4

59. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17: 1665–1674. doi:10.1101/gr.6861907

60. DeBoever C, Tanigawa Y, Aguirre M, McInnes G, Lavertu A, Rivas MA. Assessing Digital Phenotyping to Enhance Genetic Studies of Human Diseases. Am J Hum Genet. 2020;106: 611–622. doi:10.1016/j.ajhg.2020.03.007

61. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33: 1–22. doi:10.18637/jss.v033.i01

62. Olkin I, Finn JD. Correlations redux. Psychol Bull. 1995;118: 155–164. doi:10.1037/0033-2909.118.1.155

63. Cohen J, Cohen P, West SG, Aiken LS. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Routledge; 2013. Available: https://play.google.com/store/books/details?id=fAnSOgbdFXIC

64. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44: 837–845. doi:10.2307/2531595

65. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4: 7. doi:10.1186/s13742-015-0047-8

66. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet. 2021;53: 1097–1103. doi:10.1038/s41588-021-00870-7

67. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks – a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press; 2016. pp. 87–90. doi:10.3233/978-1-61499-649-1-87

68. R Core Team. R: A language and environment for statistical computing. 2019. Available: https://www.R-project.org/

69. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the tidyverse. J Open

Source Softw. 2019;4: 1686. doi:10.21105/joss.01686

70. Tange O. GNU Parallel 2018. 2018. doi:10.5281/zenodo.1146014

Supporting Information

[PUBLIC] PRSmap Supplementary Tables (medRxiv v3)

S1 Table. List of traits analyzed in the study and the predictive performance of the corresponding PRS models. For the 1,565 traits analyzed in the study, the following information is shown: trait category, the phenotype ID in Global Biobank Engine (GBE ID), trait name, the types of link functions in a generalized linear model (Gaussian for quantitative traits and Binomial for binary traits), the predictive performance of the genotype-only model, covariate-only model, the full model that considers both genotype and covariates, as well as the incremental predictive performance (Delta[Full, covariates-only]), the number of genetic variants included in the PRS model, the statistical significance of the incremental predictive performance in a hold-out test set consists of a subset of white British individuals in the UK Biobank, whether the p-value is significant after multiple-hypothesis correction ($p < 2.5 \times 10^{-5}$), whether the trait is included in the correlation analysis between the number of selected variants and the incremental predictive performance, the score ID in polygenic score (PGS) catalog, the experimental factor ontology term ID of the mapped traits in PGS catalog, and the label of the mapped traits in PGS catalog.

S2 Table. The cohort characteristics. For each ancestry group in UK Biobank, the number of individuals (n), age (mean and standard deviation [sd]), sex (percentage of individuals in male), the fraction of individuals genotyped on the UK Biobank Axiom Array. The statistics for the white British ancestry group were shown for the 70% training set, 10% validation set, and 20% test set.

S3 Table. The variant consequence grouping. We grouped the Ensembl's variant effect predictor (VEP)-predicted consequence of the genetic variants into six groups (Consequence group): protein-truncating variants (PTVs), protein-altering variants (PAVs), protein-coding variants (PCVs), intronic variants (Intronic), variants in untranslated region (UTR), and other non-coding variants (Others). The links to the sequence ontology (SO) term detailing the definition of each of the predicted consequences are shown.

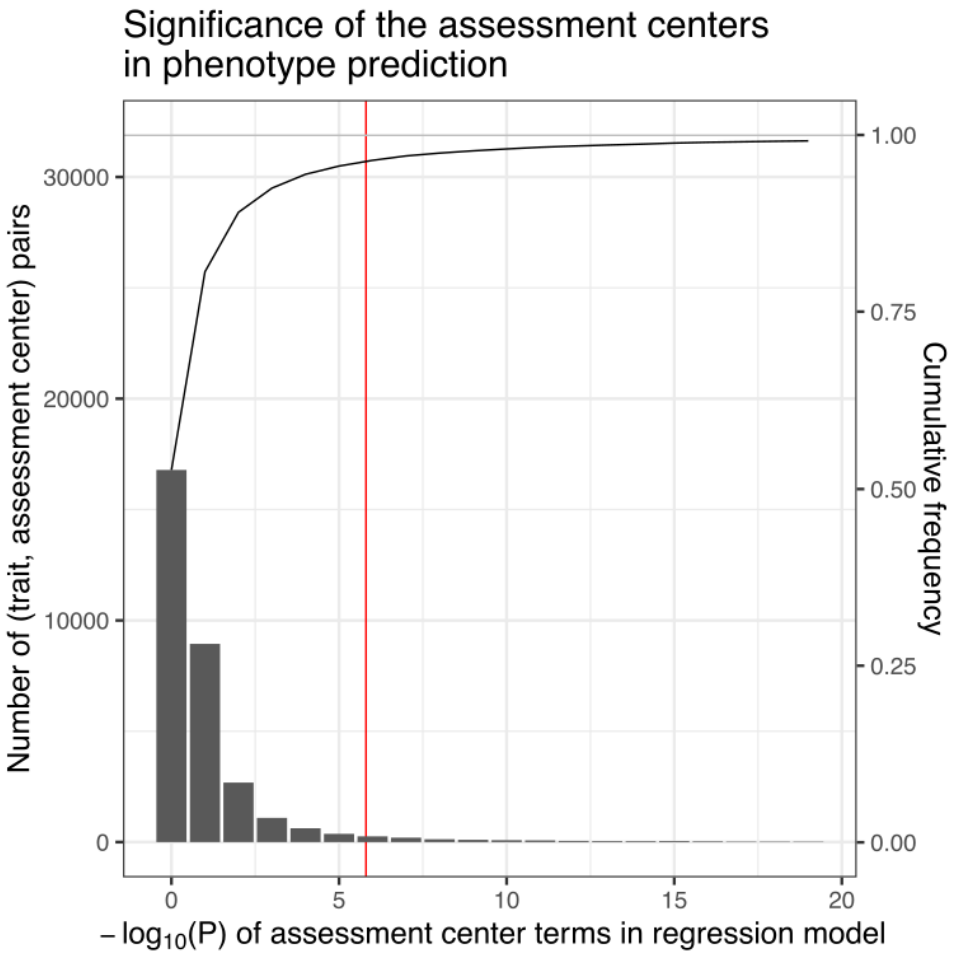
S4 Table. The penalty factor assignment rule. We used the VEP-predicted consequence and ClinVar annotation to prioritize protein-truncating, protein-altering, and (likely) pathogenic variants by assigning lower penalty factor values. The penalty factor and the number of variants stratified by genetic variants (genotype or allelotype), predicted consequence, and ClinVar annotation is shown.

S5 Table. The number of variants with non-zero BETAs is shown across four traits. For each trait, we compared two models: without and with penalty factors to prioritize the medically relevant alleles.

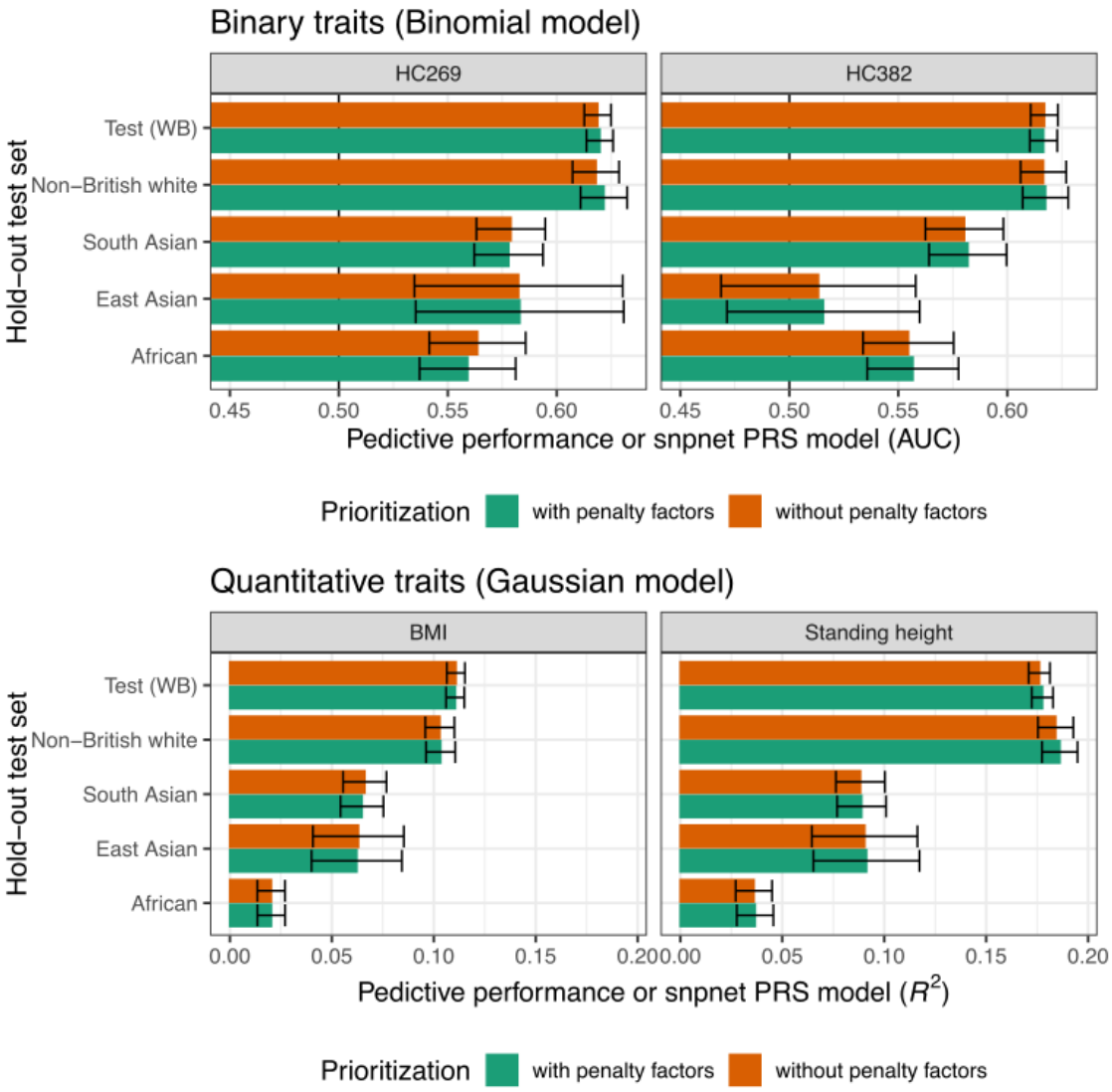
S6 Table. The predictive performance of PRS models. For each trait (Trait category, GBE_ID, and Trait Name), we show the types of link functions in a generalized linear model (GLM family column, Gaussian for quantitative traits and binomial for binary traits), the population split (population), the types of the predictive model (model column), the types of evaluation metric (R^2 [R2], Nagelkerke's pseudo-R^2

[NagelkerkeR2], AUROC [AUC], or Tjur's Coefficient of Discrimination [TjurR2]), the value of the specified metric and its lower and upper bound of 95% confidence interval, and the statistical significance (p-value).

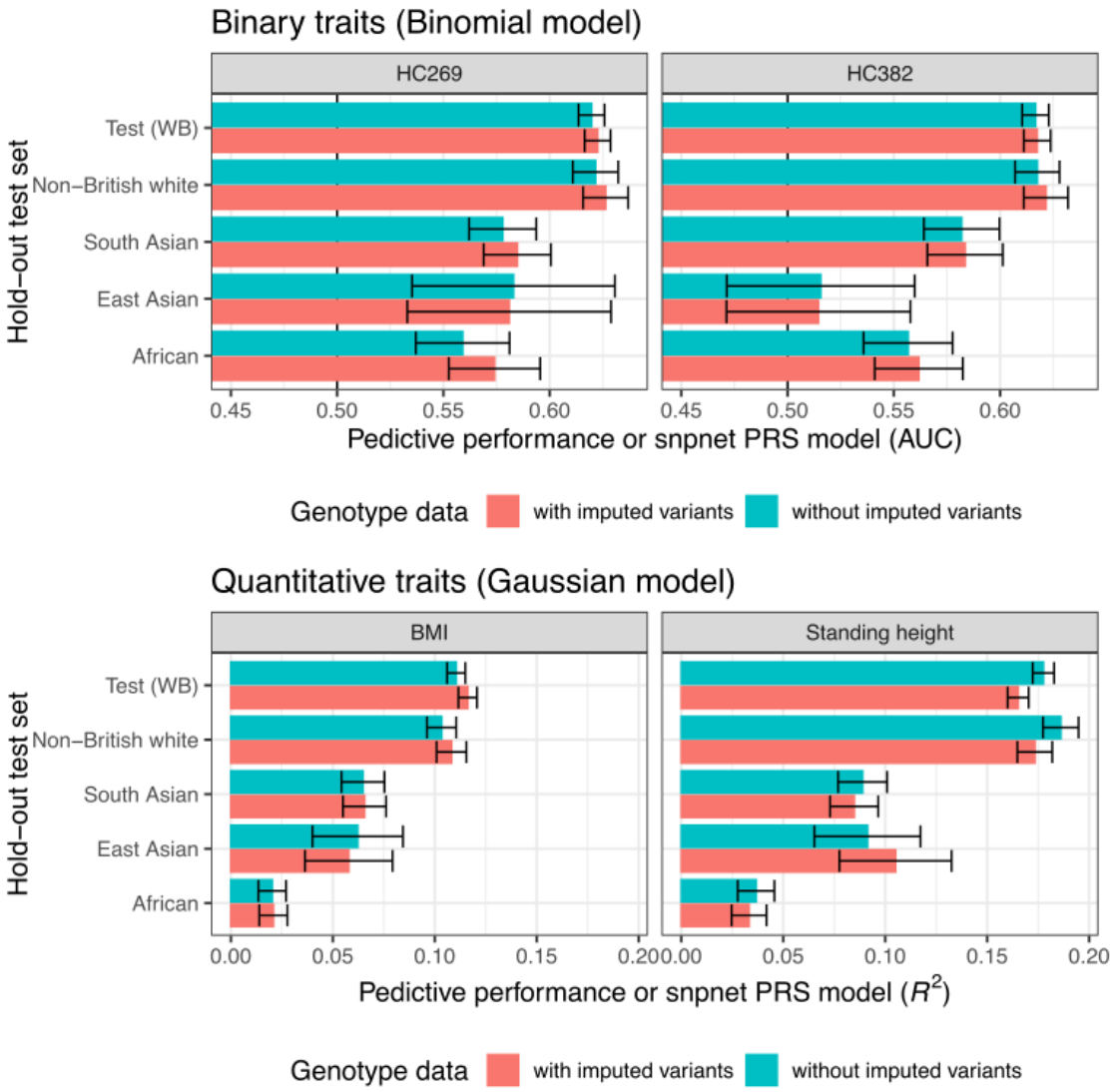
S7 Table. Estimated SNP-based heritability. For each trait with a significant PRS model (trait, trait_name, and trait_category), we show the types of link functions in a generalized linear model (family column, Gaussian for quantitative traits and binomial for binary traits), estimated SNP-based observed-scale heritability with standard error (h2_obs and h2_obs_se), lambda GC (lambda_GC), mean chi-square statistic (mean_chi2), LD score regression intercept and its standard error (intercept and intercept_se), and the proportion of the inflation attributed to the LD score regression intercept, defined by (intercept -1)/(mean(chi-square)-1), and its standard error (ratio and ratio_se).



S1 Figure. Statistical significance of the assessment center terms in phenotype prediction. We fit a regression model on age, sex, the types of genotyping arrays, polygenic risk score, and assessment centers for each of the 1,565 traits analyzed in the study. The frequency of the statistical significance ($-\log_{10}(P)$) of assessment center variables was shown. The cumulative frequency was shown on the secondary axis on the right. The statistical significance after the Bonferroni correction was shown as a red vertical line.



S2 Figure. The impact of prioritizing the medically relevant alleles with penalty factors on the predictive performance of snpnet PRS models. The predictive performance (AUC for binary traits and R^2 for quantitative traits) evaluated across hold-out test set individuals of different ancestry groups in UK Biobank are shown for four traits. The error bars represent the 95% confidence interval.



S3 Figure. The impact of the imputed genetic variants on the predictive performance of snpnet PRS models. The predictive performance (AUC for binary traits and R^2 for quantitative traits) evaluated across hold-out test set individuals of different ancestry groups in UK Biobank are shown for four traits. The error bars represent the 95% confidence interval.

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 24983, “Generating effective therapeutic hypotheses from genomic and hospital linkage data” (http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf). Based on the information provided in Protocol 44532, the Stanford IRB has determined that the research does not involve human subjects as defined in 45 CFR 46.102(f) or 21 CFR 50.3(g). All participants of UK Biobank provided written informed consent (more information is available at https://www.ukbiobank.ac.uk/2018/02/gdpr/). Some of the computing for this project was performed on the Sherlock cluster. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results.

Author information

Author contributions

M.A.R. conceived and designed the study. Y.T. carried out the statistical and computational analyses with advice from J.Q, G.V., J.M.J., R.L., R.T., T.H., and M.A.R. Y.T., G.V., and J.M.J. contributed to the quality control of the data with advice from M.A.R. Y.T. developed the web application with advice from M.A.R. The manuscript was written by Y.T. and M.A.R. and revised by all the co-authors. All co-authors have approved the final version of the manuscript.

Funding

This work has been supported by the Funai Foundation for Information Technology [to Y.T.]; Stanford University School of Medicine [to Y.T., R.L., and M.A.R.]; National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) [R01HG010140 to M.A.R.]; NIH center for Multi and Trans-ethnic Mapping of Mendelian and Complex Diseases [5U01 HG009080 to M.A.R]; NIH [5R01 EB 001988-21 to T.H., 5R01 EB001988-16 to R.T.]; and National Science Foundation (NSF) [DMS-1407548 to T.H., 19 DMS1208164 to R.T.]. The authors of this manuscript have received the following salary support: NHGRI of NIH [R01HG010140 to Y.T. and M.A.R., R01HG008155 to Y.T.], NIH [5U01 HG009080 to M.A.R.], and the National Institute on Aging of NIH [R01AG067151 to Y.T.]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies; funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The corresponding author(s) have read the journal's policy and the authors of this manuscript have the following competing interests: M.A.R is a consultant at MazeTx and is currently on leave at HiBio. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data and Code Availability Statement

The sparse PRS model weights generated from this study are available on the Global Biobank Engine (https://biobankengine.stanford.edu/prs). The significant PRS models are also available at the PGS catalog (https://www.pgscatalog.org/publication/PGP000244/ and https://www.pgscatalog.org/publication/PGP000128/, score IDs are listed in **S1 Table**). The BASIL algorithm implemented in the R *snpnet* package was used in the PRS analysis, which is available at https://github.com/rivas-lab/snpnet. The analyses presented in this study were based on data accessed through the UK Biobank: https://www.ukbiobank.ac.uk.

(A) Study cohort and phenotypes

<u>378,066 unrelated individuals in UK Biobank</u> (**S2 Table**)

337,129 individuals of white British ancestry
- score development (n = 269,704)
- hold-out test set (n = 67,425)

Additional hold-out sets for transferability assessment
- non-British white (n = 24,905)
- African (n = 6,497)
- South Asian (n = 7,831)
- East Asian (n = 1,704)

<u>1,565 phenotypes in UK Biobank</u> (**S1 Table**)

871 quantitative traits

Anthropometry, biomarkers, blood assays, bone densitometry ...

694 binary traits

Disease outcomes, cancer, lifestyle factors, family history ...

(B) Sparse PRS models across UK Biobank traits

Browseable phenotypes

Here, we display available sparse PRS models in UK Biobank. One can use sort and/or filter function. For example, you may enter ">1000" in the '# variants' column to select PRS models with more than 1,000 genetic variants.

Trait			Predictive performance					Significance	
Trait group	Trait name	Family	Geno	Covars	Full	delta	# variants	p (WB)	significant?
Select (Abdominal MRI, Anthropometry, Arterial stiffness, Biomarkers, Blood assays, Blood pressure, Bone-densitometry of heel)		Select (Binomial, Gaussian)					>2000		Select
			0.18	0.54	0.72	0.18	51209	0.00e+00	True
		Gaussian	0.13	0.43	0.56	0.13	36345	0.00e+00	True
	er mass	Gaussian	0.05	0.71	0.76	0.05	35046	0.00e+00	True
Anthropometry	Trunk fat-free mass	Gaussian	0.05	0.72	0.77	0.05	34749	0.00e+00	True
Anthropometry	Whole body fat-free mass	Gaussian	0.05	0.71	0.76	0.05	34486	0.00e+00	True
Anthropometry	Trunk pred. mass	Gaussian	0.05	0.72	0.77	0.05	32685	0.00e+00	True
Anthropometry	Impedance of whole body								
Anthropometry	Basal metabolic rate								
Anthropometry	Leg pred. mass (R)								
Anthropometry	Leg fat-free mass (R)	Gaussian	0.05	0.65	0.7	0.05	32046	0.00e+00	True
Trait category	Trait	Family	Geno	Covars	Full	delta	# variants	p (WB)	significant?

<< First | < Prev. | 1 - 10 / 273 filtered traits (out of 1565 available traits) | Next > | Last >> | 10 | 1

- Sortable table with filtering functionality
- PRS models across 1,565 traits

(C) Prediction and PRS weights for quantitative traits

Phenotype: Mean platelet volume

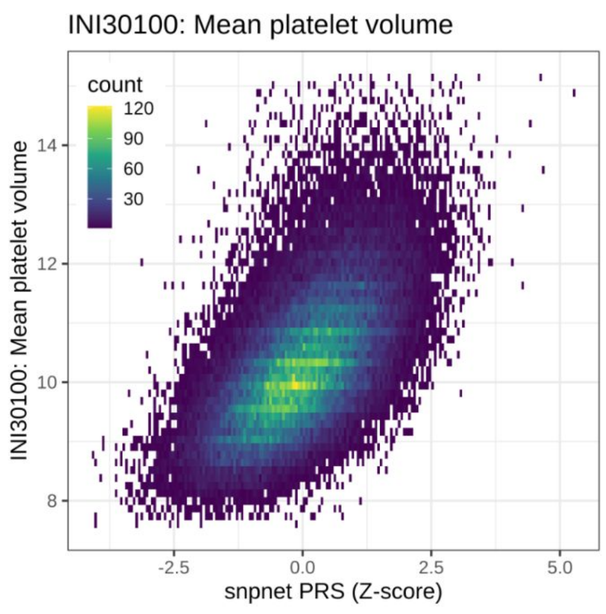
Go back to the PRS top page

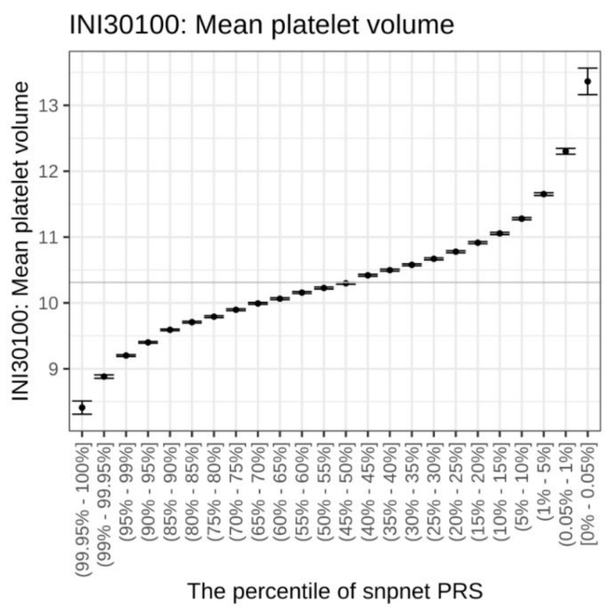
Phenotype code: INI30100
Case Count or N: 437911
Single variant: INI30100

Mean platelet volume PRS weights

Predictive performance of the PRS (genotype-only model) in hold-out test set

The hold-out test set consists of unrelated individuals of white British ancestry in UK Biobank.





(D) Prediction and PRS weights for binary traits

Phenotype: Hypertension

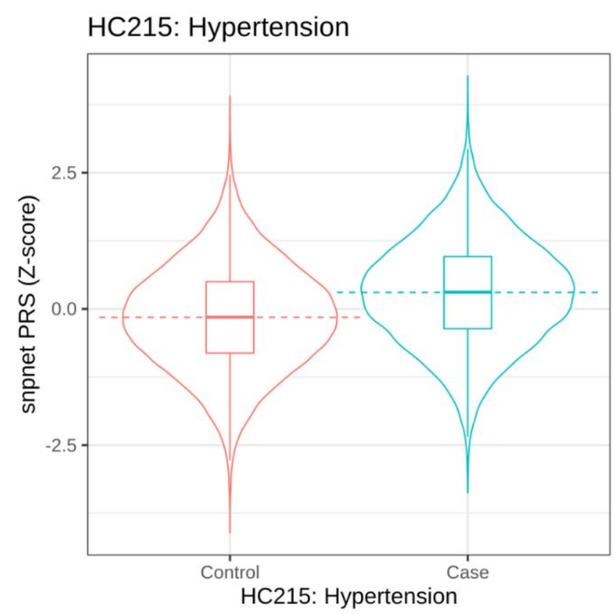
Go back to the PRS top page

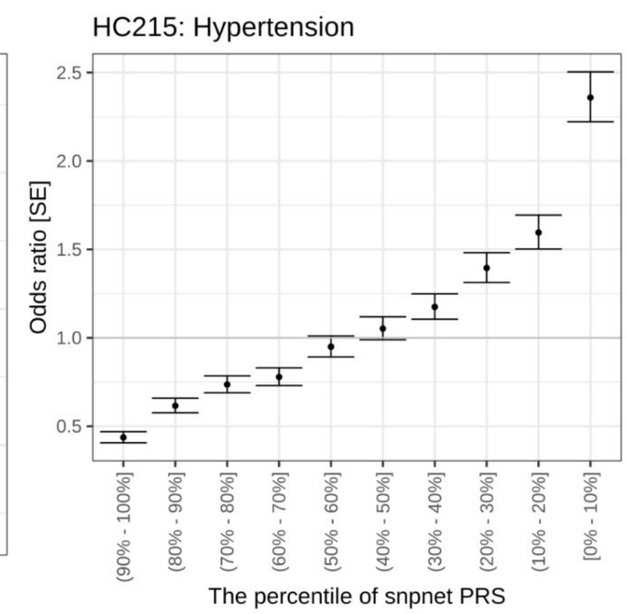
Phenotype code: HC215
Case Count or N: 153912
Single variant: HC215

Hypertension PRS weights

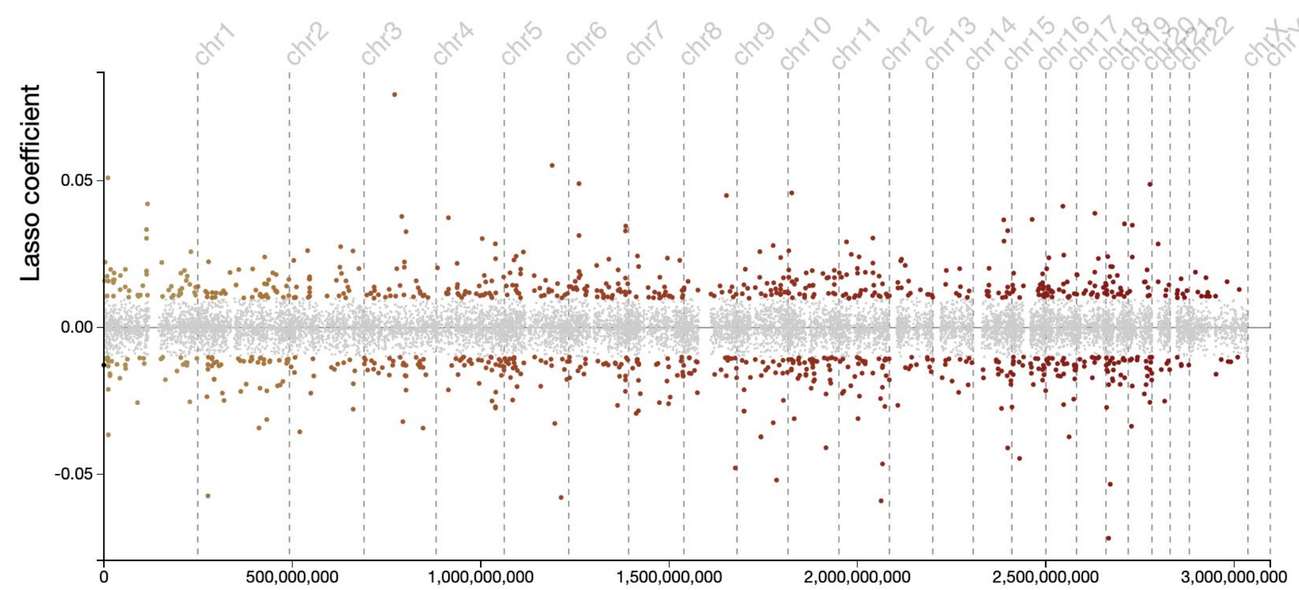
Predictive performance of the PRS (genotype-only model) in hold-out test set

The hold-out test set consists of unrelated individuals of white British ancestry in UK Biobank.





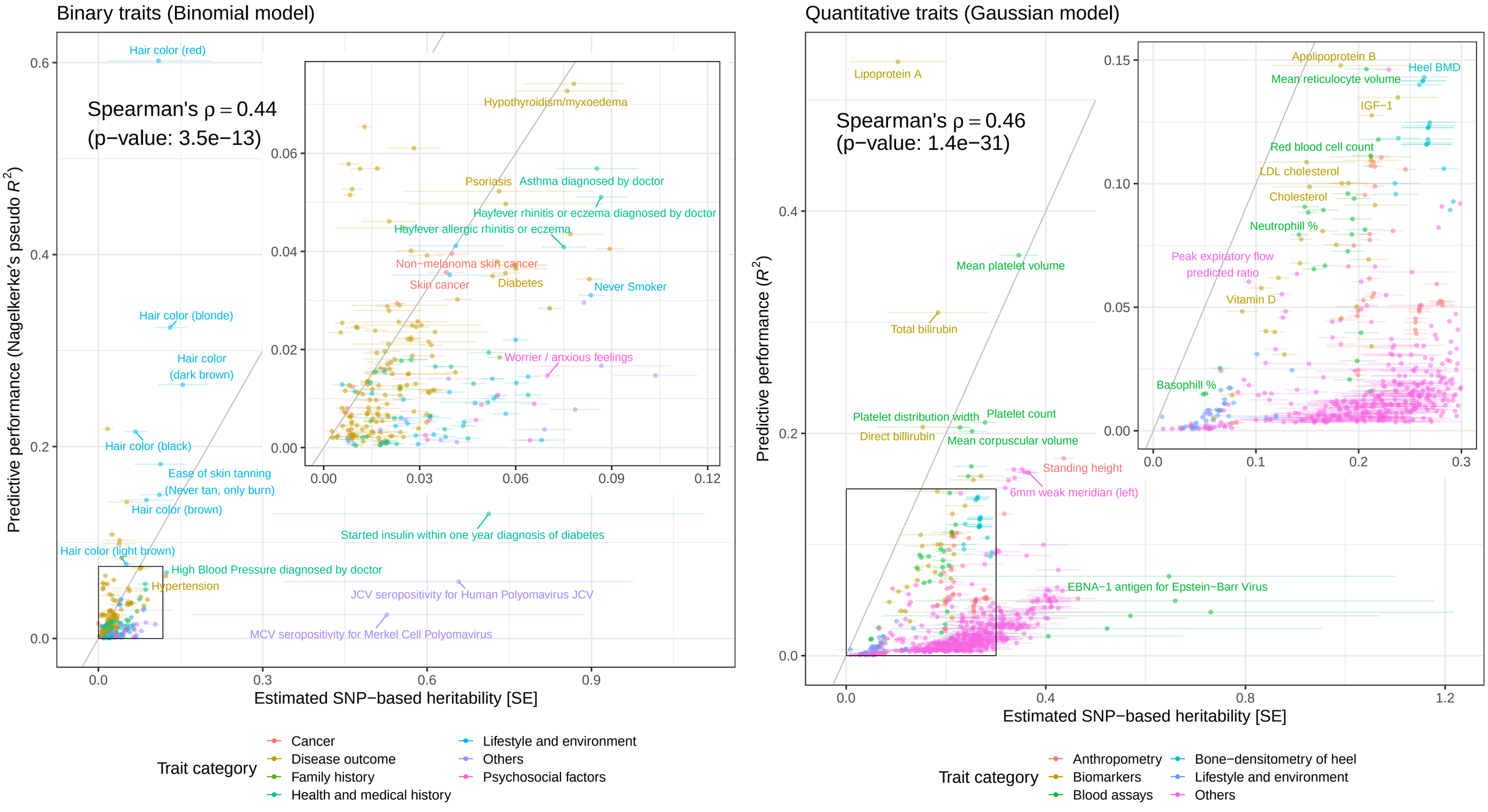
(E) Non-zero coefficients of sparse PRS model

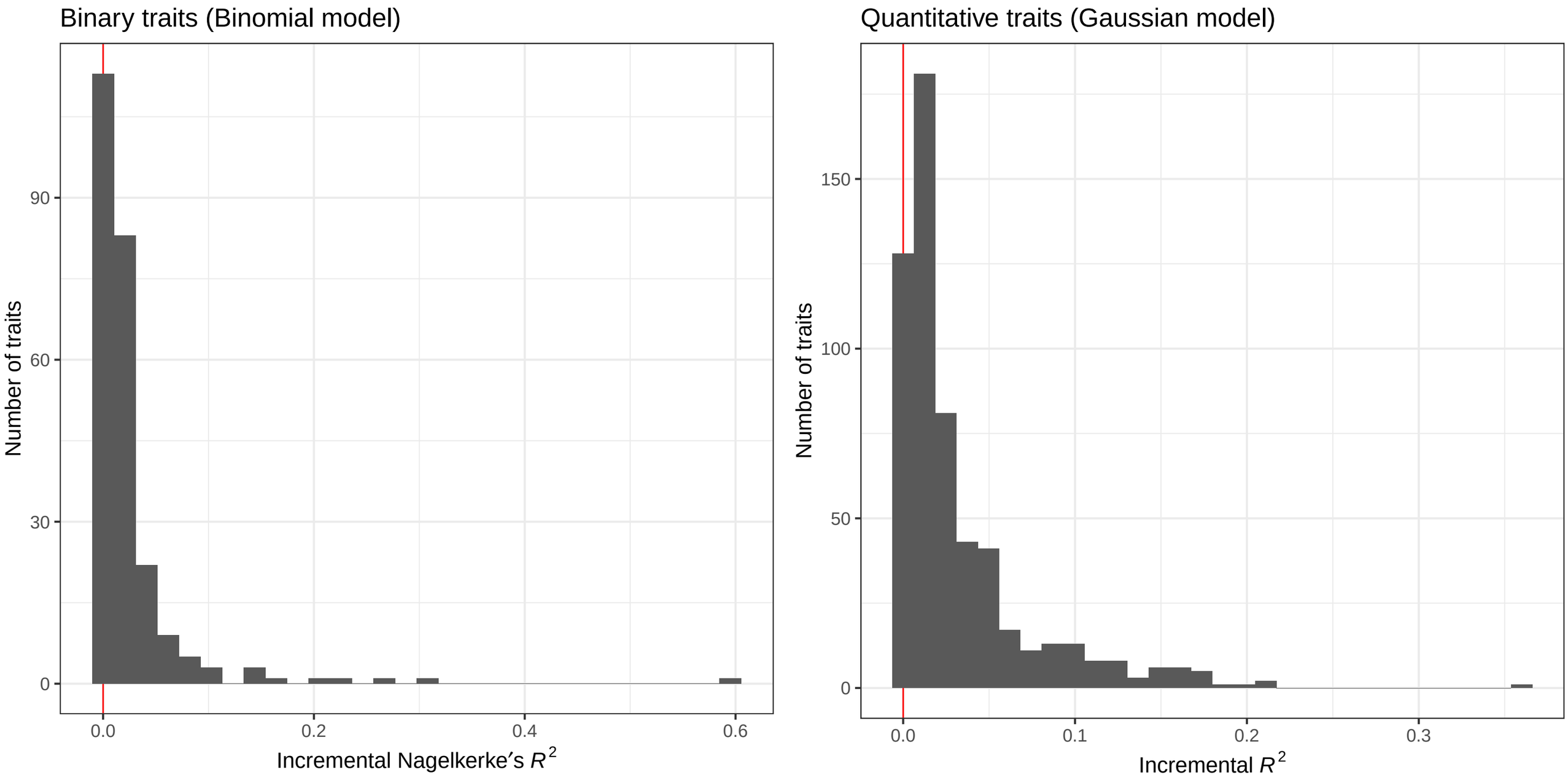


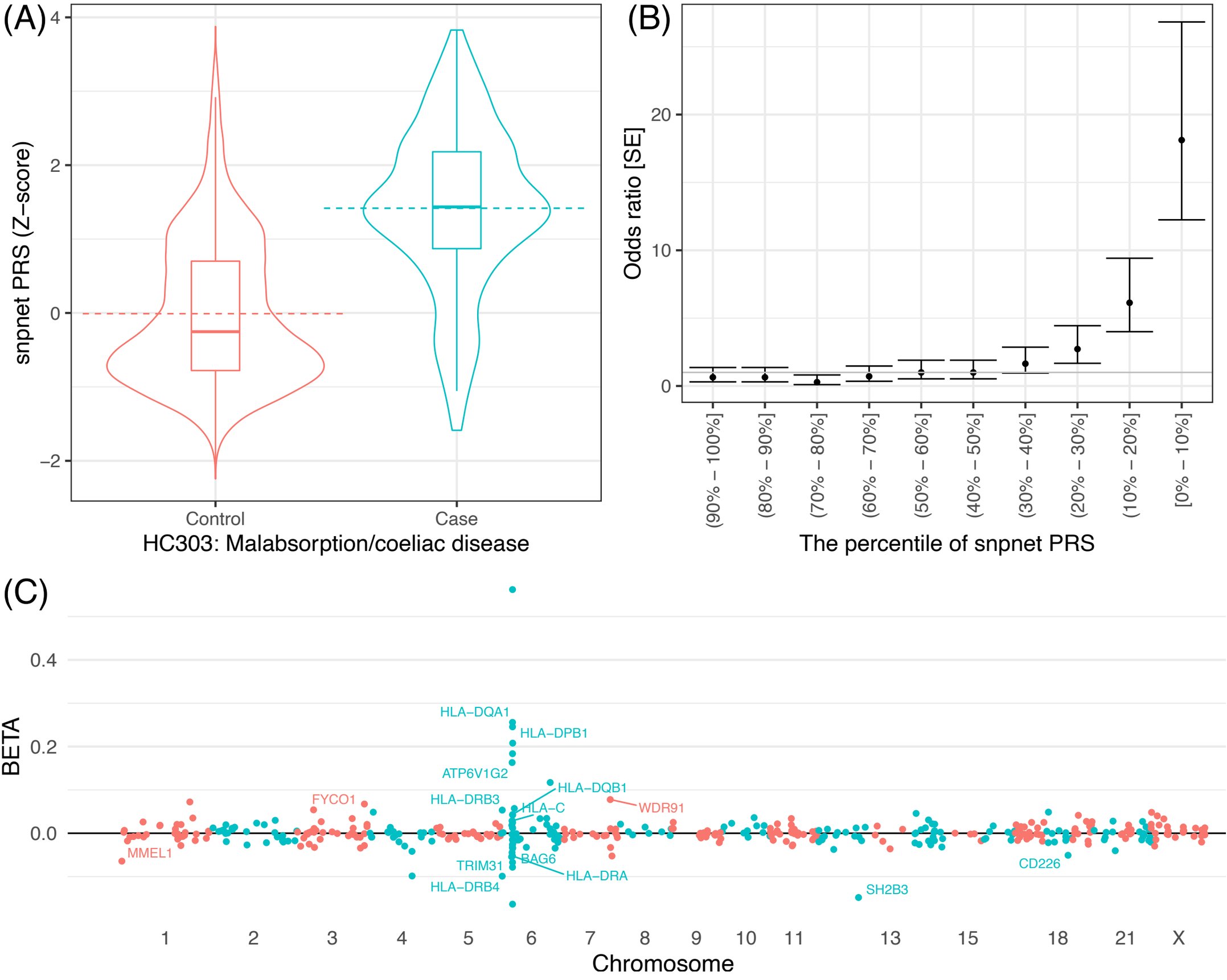
Note: only 13791 variants in active path for Lasso are included in the coefficient plot.

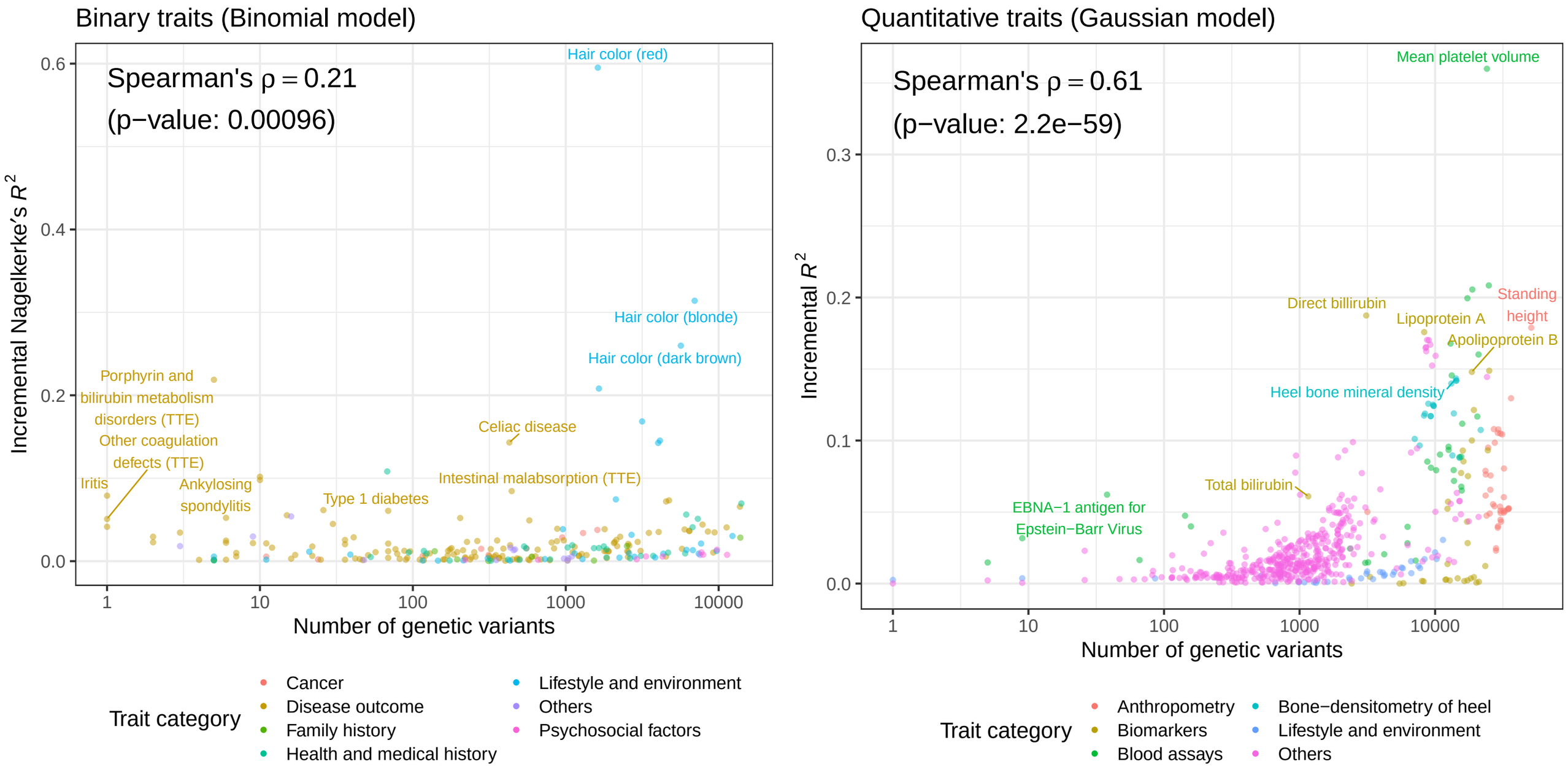
(F) Transferability assessment in UK Biobank

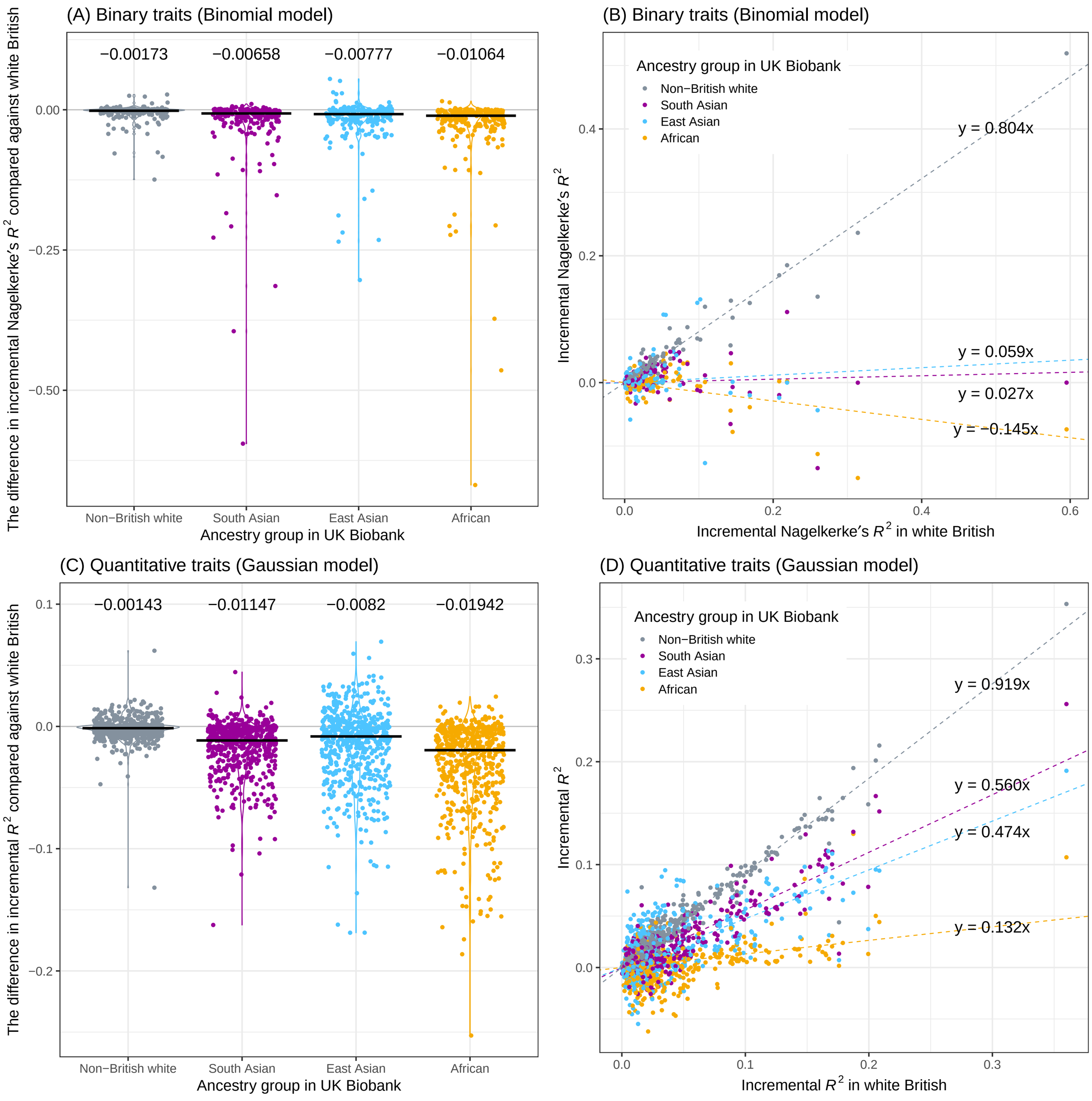
Metric	Ancestry group	Predictive performance				n	case_n	control_n
		Geno	Covars	Full	delta			
Nagelkerke's pseudo-R^2	white British (model development)	0.240	0.113	0.322	0.209	269704	91884	177820
Nagelkerke's pseudo-R^2	white British (hold-out test set)	0.065	0.113	0.179	0.066	67425	22882	44543
Nagelkerke's pseudo-R^2	Non-British white	0.055	0.138	0.191	0.054	24905	7648	17257
Nagelkerke's pseudo-R^2	South Asian	0.038	0.175	0.213	0.038	7831	3161	4670
Nagelkerke's pseudo-R^2	East Asian	0.055	0.177	0.226	0.048	1704	446	1258
Nagelkerke's pseudo-R^2	African	0.011	0.175	0.177	0.003	6497	2919	3578
Tjur's pseudo-R^2	white British (model development)	0.181	0.082	0.246	0.165	269704	91884	177820
Tjur's pseudo-R^2	white British (hold-out test set)	0.047	0.082	0.132	0.050	67425	22882	44543
Tjur's pseudo-R^2	Non-British white	0.040	0.100	0.141	0.041	24905	7648	17257
Tjur's pseudo-R^2	South Asian	0.029	0.134	0.163	0.029	7831	3161	4670
Tjur's pseudo-R^2	East Asian	0.039	0.130	0.168	0.039	1704	446	1258
Tjur's pseudo-R^2	African	0.008	0.135	0.136	0.001	6497	2919	3578


Binary traits (Binomial model)
Spearman's ρ = 0.44
(p−value: 3.5e−13)
Predictive performance (Nagelkerke's pseudo R^2)
Estimated SNP−based heritability [SE]
Hair color (red)
Hair color (blonde)
Hair color (dark brown)
Hair color (black)
Ease of skin tanning (Never tan, only burn)
Hair color (brown)
Hair color (light brown)
High Blood Pressure diagnosed by doctor
Hypertension
Hypothyroidism/myxoedema
Psoriasis
Asthma diagnosed by doctor
Hayfever rhinitis or eczema diagnosed by doctor
Hayfever allergic rhinitis or eczema
Non−melanoma skin cancer
Skin cancer
Diabetes
Never Smoker
Worrier / anxious feelings
Started insulin within one year diagnosis of diabetes
JCV seropositivity for Human Polyomavirus JCV
MCV seropositivity for Merkel Cell Polyomavirus
Trait category
Cancer
Disease outcome
Family history
Health and medical history
Lifestyle and environment
Others
Psychosocial factors
Quantitative traits (Gaussian model)
Spearman's ρ = 0.46
(p−value: 1.4e−31)
Predictive performance (R^2)
Estimated SNP−based heritability [SE]
Lipoprotein A
Mean platelet volume
Total bilirubin
Platelet distribution width
Platelet count
Direct billirubin
Mean corpuscular volume
Standing height
6mm weak meridian (left)
EBNA−1 antigen for Epstein−Barr Virus
Apolipoprotein B
Heel BMD
Mean reticulocyte volume
IGF−1
Red blood cell count
LDL cholesterol
Cholesterol
Neutrophill %
Peak expiratory flow predicted ratio
Vitamin D
Basophill %
Trait category
Anthropometry
Biomarkers
Blood assays
Bone−densitometry of heel
Lifestyle and environment
Others


Binary traits (Binomial model)
Number of traits
90
60
30
0
0.0
0.2
0.4
0.6
Incremental Nagelkerke's R^2
Quantitative traits (Gaussian model)
Number of traits
150
100
50
0
0.0
0.1
0.2
0.3
Incremental R^2


(A)
snpnet PRS (Z−score)
4
2
0
−2
Control
Case
HC303: Malabsorption/coeliac disease
(B)
Odds ratio [SE]
20
10
0
(90% − 100%]
(80% − 90%]
(70% − 80%]
(60% − 70%]
(50% − 60%]
(40% − 50%]
(30% − 40%]
(20% − 30%]
(10% − 20%]
[0% − 10%]
The percentile of snpnet PRS
(C)
BETA
0.4
0.2
0.0
HLA−DQA1
HLA−DPB1
ATP6V1G2
HLA−DQB1
FYCO1
HLA−DRB3
HLA−C
WDR91
MMEL1
TRIM31
BAG6
HLA−DRA
HLA−DRB4
CD226
SH2B3
1
2
3
4
5
6
7
8
9
10
11
13
15
18
21
X
Chromosome


Binary traits (Binomial model)
Spearman's ρ = 0.21
(p−value: 0.00096)
Incremental Nagelkerke's R^2
Number of genetic variants
Hair color (red)
Hair color (blonde)
Hair color (dark brown)
Porphyrin and bilirubin metabolism disorders (TTE)
Other coagulation defects (TTE)
Iritis
Ankylosing spondylitis
Type 1 diabetes
Celiac disease
Intestinal malabsorption (TTE)
Trait category
Cancer
Disease outcome
Family history
Health and medical history
Lifestyle and environment
Others
Psychosocial factors
Quantitative traits (Gaussian model)
Spearman's ρ = 0.61
(p−value: 2.2e−59)
Incremental R^2
Number of genetic variants
Mean platelet volume
Direct billirubin
Lipoprotein A
Standing height
Apolipoprotein B
Heel bone mineral density
Total bilirubin
EBNA−1 antigen for Epstein−Barr Virus
Trait category
Anthropometry
Biomarkers
Blood assays
Bone−densitometry of heel
Lifestyle and environment
Others


(A) Binary traits (Binomial model)
−0.00173
−0.00658
−0.00777
−0.01064
0.00
−0.25
−0.50
The difference in incremental Nagelkerke's R^2 compared against white British
Non−British white
South Asian
East Asian
African
Ancestry group in UK Biobank
(B) Binary traits (Binomial model)
Ancestry group in UK Biobank
Non−British white
South Asian
East Asian
African
y = 0.804x
y = 0.059x
y = 0.027x
y = −0.145x
0.4
0.2
0.0
Incremental Nagelkerke's R^2
0.0
0.2
0.4
0.6
Incremental Nagelkerke's R^2 in white British
(C) Quantitative traits (Gaussian model)
−0.00143
−0.01147
−0.0082
−0.01942
0.1
0.0
−0.1
−0.2
The difference in incremental R^2 compared against white British
Non−British white
South Asian
East Asian
African
Ancestry group in UK Biobank
(D) Quantitative traits (Gaussian model)
Ancestry group in UK Biobank
Non−British white
South Asian
East Asian
African
y = 0.919x
y = 0.560x
y = 0.474x
y = 0.132x
0.3
0.2
0.1
0.0
Incremental R^2
0.0
0.1
0.2
0.3
Incremental R^2 in white British