Genetic determinants of polygenic prediction accuracy within a population

Tianyuan Lu (tianyuan.lu@mail.mcgill.ca)
McGill University https://orcid.org/0000-0002-5664-5698

Vincenzo Forgetta
Lady Davis Institute

J. Brent Richards
Lady Davis Institute for Medical Research, Jewish General Hospital; King's College London
https://orcid.org/0000-0002-3746-9086

Celia Greenwood
Department of Human Genetics, McGill University, Montreal, Quebec https://orcid.org/0000-0002-2427-5696

Article

Keywords:

Posted Date: January 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1157621/v1

Genetic determinants of polygenic prediction accuracy within a population

Tianyuan Lu[1,2], Vincenzo Forgetta[1], J Brent Richards[1,3,4], and Celia M T Greenwood[1,3,5,6]

[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada
[2]Quantitative Life Sciences Program, McGill University, Montreal, Canada
[3]Department of Human Genetics, McGill University, Montreal, Canada
[4]Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom
[5]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
[6]Gerald Bronfman Department of Oncology, McGill University, Montreal, Canada

Correspondence to
Celia M T Greenwood celia.greenwood@mcgill.ca; Address: Jewish General Hospital, Room H-414, 3755 Chemin de la Côte-Sainte-Catherine, Montréal, Québec, H3T 1E2, Canada; Tel: 514-340-8222 ext. 28397

Abstract

Genomic risk prediction is on the emerging path towards personalized medicine. However, the accuracy of polygenic prediction varies strongly in different individuals. In this study, based on up to 352,277 White British participants in the UK Biobank, we constructed polygenic risk scores for 15 physiological and biochemical quantitative traits after performing genome-wide association studies (GWASs). We identified 185 polygenic prediction variability quantitative trait loci (pvQTLs) for 11 traits by Levene's test among 254,376 unrelated individuals. We validated the effects of pvQTLs using an independent test set of 58,927 individuals. A score aggregating 51 pvQTL SNPs for triglycerides had the strongest Spearman correlation of 0.185 (p-value < 1.0×10^{-300}) with the squared prediction errors. We found a strong enrichment of complex genetic effects conferred by pvQTLs compared to risk loci identified in GWASs, including 89 pvQTLs exhibiting dominance effects. Incorporation of dominance effects into polygenic risk scores significantly improved polygenic prediction for triglycerides, low-density lipoprotein cholesterol, vitamin D, and platelet. After including 87 dominance effects for triglycerides, the adjusted R^2 for the polygenic risk score had an 8.1% increase on the test set. In addition, 108 pvQTLs had significant interaction effects with measured environmental or lifestyle exposures. In conclusion, we have discovered and validated genetic determinants of polygenic prediction variability for 11 quantitative biomarkers, and partially profiled the underlying complex genetic effects. These findings may assist interpretation of genomic risk prediction in various contexts, and encourage novel approaches for constructing polygenic risk scores with complex genetic effects.

Introduction

In recent years, large-scale genome-wide association studies (GWASs) have begun to reveal genetic architecture of complex traits [1]. Accurately estimated effects of genetic determinants have rendered construction of polygenic risk scores possible [2, 3]. These polygenic risk scores aggregate multiple genetic variants associated with the target traits across the genome and may be able to capture a significant proportion of trait heritability [2, 4]. It has been recognized that polygenic risk scores can contribute importantly both clinically and in research, by enabling risk stratification in large populations [2, 5-7], informing on risk factors [8, 9], assisting diagnosis for complex diseases [10], and suggesting potential therapeutic targets [11].

Despite its ever-increasing efficiency, polygenic prediction has important shortcomings. Inter-population heterogeneity, especially genetic ancestry discrepancies, may lead to substantial attenuation in the predictive performance of polygenic risk scores [12, 13]. However, even within the same population, the prediction accuracy of polygenic risk scores can be highly variable [14]. For instance, amongst White British participants in the UK Biobank [15], polygenic risk scores have been demonstrated to predict body mass index (BMI) more accurately in middle-aged adults than in older adults, and to predict years of schooling more accurately amongst individuals in lower socioeconomic status groups [14]. Analogous to trait variability, such prediction variability may itself be partially under genetic control and may have genetic architecture distinct from that of the trait. For example, a variability quantitative trait locus in the *FTO* gene has been found to be associated with the variability of BMI, with approximately 7% difference in the population

variance between the opposite homozygous genotypes [16], although most known risk loci for the phenotypic mean of BMI do not have detectable effects on the phenotypic variance [17].

In addition, to date, most existing polygenic risk scores include linear additive effects of common genetic variants. While more complex genetic effects, including dominance effects and interaction effects, exist for many complex traits [18-21], they are difficult to rigorously model in polygenic risk scores due to their relatively weak effect sizes [22, 23]. Therefore, profiling the genetic determinants of prediction variability within a population may help better understand intra-population heterogeneity, interpret results of genomic risk predictions, and further improve methods for developing polygenic risk scores.

In this study, leveraging resources from the UK Biobank [15], we construct polygenic risk scores for 15 vital physiological and biochemical quantitative traits. We then systematically search for polygenic prediction variability quantitative trait loci (pvQTLs) acting on the residuals after correcting for the standard linear PRS and other known covariates, and validate their effects on the accuracy of polygenic prediction. We assess dominance effects and interaction effects with environmental and lifestyle exposures underlying these pvQTLs. Lastly, we seek to improve polygenic prediction by incorporating dominance effects into polygenic risk scores.

Results

Characterization of pvQTLs

An overview of this study is provided in **Figure 1** and detailed in **Methods**. Briefly, White British participants in the UK Biobank were randomly assigned into a discovery set (80%), a linkage disequilibrium (LD) reference set (1.5%), and a test set (18.5%) (**Supplementary Tables S1** and **S2**). On the UK Biobank discovery set, GWAS and conditional and joint multiple SNP (COJO) analyses identified QTLs for 15 physiological and biochemical traits. These traits demonstrated high polygenicity, with glucose having the smallest number of near-independent QTL SNPs (60), and height having the largest number (1,317) (**Supplementary Tables S3** and **S4**). Although less powered than the largest meta-analyses of GWASs [24], these in-sample GWASs should ensure homogeneity of estimated genetic effects in the discovery and test sets. Polygenic risk scores constructed using these near-independent SNPs captured a non-trivial proportion of the total phenotypic variance, up to 16.93% for platelet counts (**Supplementary Table S3**).

By performing the median-based Levene's test, we then identified genetic variants significantly associated with polygenic prediction variability (**Methods**). In total, 185 near-independent pvQTLs were identified for 11 out of the 15 quantitative traits under investigation (**Figure 2A**). No pvQTLs were identified for diastolic blood pressure (DBP), height, waist-to-hip ratio (WHR), and nucleated red blood cell (nRBC) count. However, plasma lipid traits demonstrated the most pvQTL, with 52 for low-density lipoprotein (LDL) cholesterol and 51 for triglycerides (**Supplementary Table S5**). Based on genomic annotations, 108 of these 185 pvQTLs were located in 81 known genes (**Supplementary Table S6**). Furthermore, we found 51 pvQTLs that seemed to be located in upstream regulatory regions of known genes, and 26 pvQTLs in downstream regulatory regions (**Supplementary Table S6**).

Subsequently, we estimated the heritability of the polygenic prediction variability of each trait. Although a considerable proportion of trait variance is explained by GWAS SNPs (**Figure 2B**, and previously shown by many authors http://www.nealelab.is/uk-biobank/), none of these 15 traits exhibited a strong heritability of prediction variability. Traits that had more pvQTLs in general demonstrated higher heritability of prediction variability, though the highest heritability was estimated to be only 5.73% (SD = 1.21%) for triglycerides, followed by 4.46% (SD = 0.30%) for BMI, and 3.83% for vitamin D (SD = 1.73%; **Figure 2C**). On the contrary, four traits that did not have identified pvQTLs (DBP, height, WHR, and nRBC) and calcium which had one pvQTL, had the lowest prediction variability heritability estimates from LD score regression, all below 0.70% (**Figure 2C**).

Our observations of genetic correlations in phenotypic variability were limited to traits that were biologically related (**Supplementary Figure S1** and **Supplementary Table S7**). As expected, LDL, triglycerides, and glucose had pairwise significant genetically correlated prediction variability, with the strongest correlation of 0.38 (SD = 0.09; p-value = 1.2×10^{-5}) between LDL and triglycerides. Meanwhile, prediction variability of BMI was genetically correlated with prediction variability of WHR (correlation = 0.35; SD = 0.07; p-value = 7.4×10^{-7}), heel bone mineral density (correlation = 0.33; SD = 0.08; p-value = 8.9×10^{-6}), plasma glucose (correlation = 0.38; SD = 0.07; p-value = 2.9×10^{-8}), hemoglobin A1c (HbA1c; correlation = 0.29; SD = 0.05; p-value = 2.4×10^{-9}), and LDL (correlation = 0.14; SD = 0.05; p-value = 7.1×10^{-3}).

Validation of pvQTL effects on polygenic prediction

We examined whether the identified pvQTLs had effects on polygenic prediction (**Methods**). For eight traits that had more than one pvQTL, on the test set, a pvQTL score constructed based on the discovery set demonstrated significant association with the squared prediction errors of the polygenic risk scores (**Figure 3**). For example, the pvQTL score for triglycerides based on 51 pvQTL SNPs had a Spearman correlation of 0.185 (p-value < 1.0×10^{-300}) with the squared prediction errors; the pvQTL score for vitamin D based on ten pvQTL SNPs had a Spearman correlation of 0.109 (p-value = 4.1×10^{-141}) with the squared prediction errors (**Figure 3**). Although only one pvQTL was identified for heel bone mineral density (BMD), this pvQTL SNP also demonstrated significant association with the squared prediction errors (Spearman correlation = 0.036; p-value = 7.2×10^{-11}; **Figure 3**). On the other hand, the single pvQTL for systolic blood pressure (SBP) had the weakest association with the squared prediction errors and was deemed insignificant (**Figure 3**).

Enrichment of dominance effects and SNP-exposure interaction effects amongst pvQTLs

The majority of QTLs did not demonstrate impact on polygenic prediction variability, since there was little evidence of colocalization with pvQTL signals (colocalization posterior probability, i.e. CLPP < 0.01 [25]; **Figure 4A** and **Supplementary Table S8**). Compared to QTLs, pvQTLs were more likely to have dominance effects on the corresponding traits (**Figure 4B**), as 89 (48.1%) of the 185 pvQTL SNPs had at least one allele with a p-value < 1.4×10^{-4} (Bonferroni threshold accounting for 370 tests) for dominance effect (**Methods**). Interestingly, QTLs with a higher CLPP, suggesting a stronger effect on prediction variability, were also more likely to have a detectable dominance

effect. Specifically, 92 (20.2% out of 455) QTLs with a CLPP > 0.1 had a p-value < $1.4x10^{-4}$ for dominance effect, compared to 62 (7.9% out of 787) among those with a CLPP between 0.01 and 0.1, 53 (4.0% out of 1,328) among those with a CLPP between 0.001 and 0.01, and 10 (4.5% out of 223) among those with a CLPP ≤ 0.001 (**Figure 4B**).

On the discovery set, at least one significant pvQTL SNP-exposure interaction effect with a p-value < $3.0x10^{-5}$ (Bonferroni threshold accounting for 9 exposures x 185 pvQTL SNPs = 1,665 tests) was detected for each of the nine exposures, while a total of 108 (58.4%) pvQTL SNPs had interaction effects with at least one exposure (**Supplementary Figure S2** and **Supplementary Table S9**). Notably, pvQTL SNPs for LDL, triglycerides, glucose and HbA1c had interaction effects with the use of cholesterol-lowering drug or anti-hypertensive drug; pvQTL SNPs for glucose and HbA1c further had interaction effects with type 1 and type 2 diabetes (**Supplementary Figure S2**). As expected, these interaction effects also appeared to be more enriched amongst pvQTLs than GWAS-identified QTLs (**Figure 4C**). Full summary statistics of dominance effects and SNP-exposure interaction effects are provided in **Supplementary Table S8** and **Supplementary Table S9**.

Improved polygenic prediction by incorporating dominance effects

By modelling dominance effects on the discovery set (**Methods**), polygenic risk scores incorporating additional dominance effects may have improved predictive performance on the test set (**Figure 5** and **Supplementary Table S10**). For instance, after incorporating 87 dominance effects with a false discovery rate (FDR) < 0.05 for triglycerides, the adjusted R^2 for the polygenic

risk score increased from 0.1193 to 0.1290 (8.1% relative increment) including covariate effects; with 64 additional dominance effects in an LDL polygenic risk score, the adjusted R^2 increased from 0.0950 to 0.1024 (7.8% relative increment; **Figure 5**). Significant improvements in predictive performance were also observed for polygenic risk scores for vitamin D (1.1% relative increment) and platelet (0.5% relative increment), despite their smaller magnitude (**Figure 5**). These improvements were consistent if a more stringent Bonferroni threshold was implemented to pre-select dominance effects to be incorporated into the polygenic risk scores (**Figure 5** and **Supplementary Table S11**). However, predictive performance of polygenic risk scores for other traits did not exhibit evident improvements.

Discussion

In the past decade, discovery of genetic determinants for complex traits enabled by well-established cohorts has empowered polygenic risk scores to achieve considerable prediction accuracy [2, 3, 26, 27]. However, one of the major obstacles yet to be overcome before polygenic risk scores can be universally utilized in health care is the highly variable predictive performance amongst different groups of individuals [14]. Following one line of investigation as to why these performance differences occur, in this work, we sought to identify and characterize pvQTLs for 15 key physiological and biochemical quantitative traits based on White British participants in the UK Biobank.

In total, 185 pvQTLs were identified for 11 out of the 15 traits under investigation. Although the overall genetic contributions to polygenic prediction variability were small in number and

magnitude, these pvQTLs affected pivotal genes in the biological pathways pertaining to the target traits. For instance, pvQTLs were identified in the *APOB*, *LDLR*, and *PCSK9* genes for LDL, which are three clinically actionable genes for familial hyperlipidemia [28-30]. Meanwhile, as expected, some pvQTLs were previously known to be associated with both the phenotypic mean and phenotypic variance of the target traits, including but not limited to pvQTLs identified in the *FTO* gene for BMI [16], in the *CHRNA3* gene for ratio of the forced expiratory volume in the first one second to the forced vital capacity of the lungs (FEV1/FVC ratio) [17], and in the *TCF7L2* gene for glucose and HbA1c levels [17]. Apart from SBP, the aggregated effects of pvQTLs for ten traits were verified to be significantly associated with squared polygenic prediction errors on an independent test set. These identified pvQTLs hence provide important resources for characterizing intra-population heterogeneity and are candidates for explaining complex genetic effects.

Dominance effects were found to be strongly enriched among pvQTLs compared to GWAS-identified QTLs. By incorporating dominance effects conferred by pvQTLs, polygenic prediction accuracy was improved for LDL, triglycerides, vitamin D, and platelets, but not for other traits where pvQTLs showed more limited effects. Although the improvements in accuracy were not large, our results encourage analysts to explicitly model complex genetic effects in addition to linear additive effects for common variants when developing polygenic risk scores, particularly when polygenic prediction accuracy demonstrates a high variability between sub-populations having different demographic characteristics or exposed to different risk factors. Genome-wide scans for dominance effects have been difficult in traditional GWASs due to limited statistical

power. Yet, , loci harboring dominance effects have sometimes been identified in large cohorts, such as *KSR2* and *ZNF507-LOC400684* for coronary artery disease [31]. We thus also anticipate our analytical scheme could be generalized to identify dominance effects with a significantly reduced number of association tests, and could bring a new avenue towards understanding the genetics of complex traits.

SNP-exposure interactions are another important source of genotype-dependent phenotypic variance heterogeneity [17, 32, 33]. Among the identified pvQTLs, nearly 60% interacted with at least one of the nine environmental or lifestyle exposures that represented crucial clinical conditions. These interaction effects implied that the same genetic variant could have condition-specific effects on the corresponding traits. For instance, the FEV1/FVC ratio pvQTL (rs56077333 in the *CHRNA3* gene, which is associated with lung functions and smoking behaviors [34, 35]) had an effect on FEV1/FVC ratio specific to ever-smokers; use of cholesterol-lowering drug and anti-hypertensive drug, as well as type 1 diabetes and type 2 diabetes widely altered the effects of pvQTL SNPs on blood lipids and glycemic traits. Notably, although we did not attempt to incorporate these interaction effects into polygenic risk scores in order to avoid potential confounding effects, constructing condition-specific polygenic risk scores may also be desired in future prospective studies with larger sample sizes.

Our study has important limitations. First of all, our findings should be considered population-specific. It has been revealed in multiple studies that genetic, particularly polygenic effects on complex traits may be subject to alterations in the underlying genetic architectures [12, 13],

238 including but not limited to changes in minor allele frequencies, LD patterns, interactions with
239 population-specific environmental exposures as well as trait heritability. We posit that these
240 inter-population differences also have strong impacts on pvQTL effects. Due to the limited
241 sample sizes of non-European ancestry populations present in the UK Biobank, we refrained from
242 screening for pvQTLs in these populations. We expect future work in emerging non-European
243 cohorts to extend our findings, potentially leveraging population-specific polygenic risk scores.
244 Second, following Wang et al. [17], we adopted Levene's test for pvQTL discovery. Compared to
245 other classical approaches, such as Bartlett's test [36] and Fligner-Killeen test [37] that also assess
246 violation of variance homogeneity assumption in linear regression, Levene's test has been shown
247 to have a considerably lower false positive rate, especially when the effects on variance are weak
248 [17]. However, because the model assumptions of Levene's test is violated in the presence of
249 genetic relatedness, related individuals were discarded from our analyses. A recently developed
250 double generalized linear model that performs a dispersion effect test is able to incorporate
251 random effect modelling to account for genetic relatedness [38], while achieving comparable
252 performance as Levene's test on the same samples [17]. This method could probably have
253 enhanced power by including related individuals, but was not utilized in this study to perform
254 genome-wide scanning for pvQTL due to its high computational cost [17, 38]. It is worth noting
255 also that Levene's test targets variance heterogeneity in quantitative traits. It remains
256 challenging to efficiently identify and characterize pvQTL for binary traits, such as disease
257 outcomes. In addition, our analyses were restricted to common SNPs to ensure statistical power.
258 More rigorous methods, such as region-based tests [39], may allow for integration of rare
259 variants that also have strong effects on phenotypic variability. Lastly, although we identified

significant dominance effects and SNP-exposure interaction effects underlying some pvQTL, there were still many pvQTLs that did not display dominance effects and did not interact with any of the exposures examined in this study. While arguably these pvQTLs could interact with exposures not tested in the current study, we note that they may also arise from other mechanisms [33], such as genetically controlled homeostatic regulation and epistasis. More investigations are needed to evaluate if these mechanisms are plausible.

In summary, we have discovered and validated genetic variants associated with the polygenic prediction variability of 11 vital physiological and biochemical traits. We have elucidated the nature of some of the pvQTL effects and the added value of incorporating complex genetic effects into polygenic risk scores. These findings warrant future investigations into the interpretation of polygenic risk prediction in different contexts, as well as novel approaches for developing polygenic risk scores for improved personalized medicine.

Methods

Study cohort

We utilized the UK Biobank [15], one of the largest genotyped cohorts to ensure statistical power. Between 2006-2010, the UK Biobank recruited and genotyped approximately 500,000 middle aged and older participants at multiple assessment centers located in the United Kingdom. Though participants in the UK Biobank were healthier, less obese, and less likely to smoke and consume alcohol compared to the general population [40], this cohort has facilitated extensive investigations on the associations between the genetics, environmental and lifestyle exposures,

and health outcomes. The UK Biobank performed genome-wide genotyping using Affymetrix arrays based on DNA extracted from blood samples provided by the participants. The genotypes were imputed to the Haplotype Reference Consortium reference panel [41].

Deep phenotyping of the UK Biobank participants was conducted upon the initial assessment visit [15], including a wide variety of anthropometric measurements, blood and urine biomarkers, etc. We hereby focused on 15 quantitative traits that were important biomarkers in health care practice or research while having a complex genetic architecture. These traits were BMI, WHR, standing height, heel BMD, plasma total calcium, plasma vitamin D, FEV1/FVC ratio, plasma glucose, HbA1c, plasma LDL cholesterol, plasma triglycerides, DBP, SBP, nRBC count, and platelet count.

Because the genetic architectures of complex traits might differ across populations of different genetic ancestries [12, 13], we only included 440,346 genotyped White British ancestry participants. We randomly split this cohort into three datasets: a discovery set (80.0%), an LD reference set (1.5%), and a test set (18.5%). Furthermore, to reduce the confounding effects of genetic relatedness, we first derived pairwise genetic relationship based on autosomal SNPs using the Genome-wide Complex Trait Analysis (GCTA) software [42] with default settings. We then randomly removed one individual in each pair of third-degree or closer relatives that had a kinship > 0.0442 [17]. In total, 318,071 largely unrelated individuals were retained (**Supplementary Table S1**), including 254,376 in the discovery set, 4,768 in the LD reference set, and 58,927 in the test set. Apart from the linear mixed model-based GWASs performed on the

discovery set, all downstream analyses were conducted using the datasets that had excluded related individuals.

Genome-wide association studies and construction of polygenic risk scores

On the discovery set that included related individuals, we retained 6,708,723 common SNPs (minor allele frequency > 0.05) that were genotyped or imputed with an imputation quality score (INFO) > 0.8. We next conducted GWASs for each of the 15 quantitative traits using the linear mixed model implemented in fastGWA [43], adjusted for age, sex, recruitment center, genotyping array, as well as the first 20 genetic principal components. The genetic principal components provided by the UK Biobank were calculated using genotyped SNPs of high quality after LD-pruning [15]. For each trait, we excluded individuals whose phenotypes were more than 5 standard deviations (SD) away from the phenotypic mean after adjusting for these covariates.

We then performed COJO analysis [44] of the GWAS summary statistics of each trait to identify near-independent quantitative trait loci (QTLs; p-value < $5x10^{-8}$) using the GCTA software [42] with default settings. The LD reference set was used as the LD reference panel to account for LD at each given locus. For each trait, following Yengo et al. [24], we constructed polygenic risk scores as the sum of allele dosage of the near-independent QTL SNPs, weighted by their effect sizes on the corresponding trait estimated in GWAS. We evaluated the proportion of total phenotypic variance explained by each polygenic risk score in the discovery, LD reference, and test sets separately.

Levene's test for identifying prediction variability quantitative trait loci

In the discovery set excluding related individuals, for each trait, we regressed out the effects of age, sex, recruitment center, genotyping array, the first 20 genetic principal component, and the polygenic effect summarized by a polygenic risk score, using linear regression. With the regression residuals (prediction errors) as responses, we performed median-based Levene's test on a per-SNP basis, as implemented in the OmicS-data-based Complex trait Analysis (OSCA) software [45]. Again, we excluded individuals whose residualized phenotypes were more than 5 SD away from the mean. The per-allele effect of each SNP on the squared prediction errors was derived from z-statistics [45, 46]. Notably, we did not impose any transformation of the phenotypes in either the GWASs or in pvQTL discovery, because (1) Levene's test is robust to distribution of the phenotypes, and (2) it has been shown that non-linear transformation, such as the logarithm transformation or the rank-based inverse normal transformation, may lead to an inflated false positive rate in Levene's test [17].

COJO analysis [44] was performed on the pvQTL summary statistics of each trait to identify near-independent pvQTL SNPs using the GCTA software [42] with default settings and with the LD reference set as the LD reference panel.

To account for testing multiple correlated traits, we determined the effective number of traits by an eigen-decomposition approach [17]. Specifically, we calculated the variance-covariance matrix of all 15 traits based on the UK Biobank discovery set excluding related individuals. The

eigen-decomposition of this variance-covariance matrix yielded 15 ordered eigenvalues (λ_1, …, λ_{15}). The effective number was estimated as

$$\frac{\left(\sum_{k=1}^{15}\lambda_k\right)^2}{\sum_{k=1}^{15}\lambda_k^2} \approx 10.8$$

Therefore, we set the corrected genome-wide significance threshold as p-value < $5\times10^{-8}/10.8 \approx 4.6\times10^{-9}$. All identified pvQTLs were annotated to their nearest genes based on the hg19 genome assembly.

Estimating heritability and cross-trait genetic correlation of polygenic prediction variability

We next estimated the heritability of polygenic prediction variability by performing LD score regression [47] leveraging the pvQTL summary statistics for each trait. We also performed cross-trait LD score regression [48] for each pair of the 15 traits to quantify the genetic correlation of polygenic prediction variability. Genetic correlation with an FDR < 0.05 was considered significant. LD score regression and cross-trait LD score regression were implemented in the LDSC software [47] with default settings. LD scores were provided by the LDSC software, and were obtained from the participants of the 1000 Genomes Project of European ancestry [49, 50].

Predicting the variability of prediction errors

To estimate the pvQTL effects on polygenic prediction variability, we constructed pvQTL scores based on the per-allele effect of each near-independent pvQTL SNP on the prediction errors estimated above. On the test set, for traits with identified pvQTLs, we first regressed out the effects of the corresponding polygenic risk scores, age, sex, recruitment center, genotyping array,

and the first 20 genetic principal components. We examined whether the squared residuals (prediction errors) were associated with the pvQTL scores, by estimating the Spearman's rank correlation coefficient (ρ).

Colocalization analyses

We subsequently evaluated whether the QTLs identified in GWAS and used for constructing standard polygenic risk scores had impact on the prediction variability by performing colocalization analyses using the eCAVIAR software [25]. We retrieved GWAS summary statistics and pvQTL summary statistics for all SNPs in a 100-SNP window [25] centered around each near-independent QTL SNP identified in GWASs. The UK Biobank LD reference set was used as the LD reference panel. Evidence for colocalization was assessed by the colocalization posterior probability, i.e. CLPP [25].

Detecting dominance effects

In the discovery set, but excluding related individuals, we tested whether the identified pvQTLs had significant dominance effects. Significance of the dominance effects was assessed for each near-independent pvQTL SNP by comparing the following two linear regression models using a likelihood ratio test:

$$Y \sim \beta_0 + D\beta_D + \boldsymbol{Q\gamma}$$

and

$$Y \sim \beta_0 + \boldsymbol{Q\gamma}$$

where Y represents the trait with phenotypic mean β_0; D is an indicator of carriers of the chosen allele with effect β_D (the two alleles of each pvQTL SNP were tested in turn for dominance effects); $\boldsymbol{Q}$ represents covariates (age, sex, recruitment center, genotyping array, the first 20 genetic principal component, as well as the polygenic risk score) with effects $\boldsymbol{\gamma}$.

Detecting SNP-exposure interaction effects

Furthermore, we examined whether the pvQTL SNPs showed evidence for interaction with measured environmental or lifestyle exposures, using the discovery set. Nine common exposures were investigated, including age, sex, smoking, self-reported alcohol intake frequency, type 2 diabetes, type 1 diabetes, use of cholesterol-lowering drug, use of anti-hypertensive drug, and sedentary activity duration. Although type 2 diabetes and type 1 diabetes are not traditional exposure variables, they were included due to the widespread impact on metabolism.

Smoking status was determined according to self-reported smoking history, and included ever-smokers and never-smokers. Type 2 diabetes was ascertained using a combination of self-reported physician-made diagnosis, self-reported use of anti-diabetic drugs (insulin, sulfonylureas, metformin, etc.) and International Classification of Diseases (ICD-10) diagnosis codes (E11-E14). Type 1 diabetes cases were identified based on ICD-10 diagnosis code of E10. Individuals with type 1 diabetes were not considered type 2 diabetes cases. Cholesterol-lowering drugs reported by UK Biobank participants included statin (atorvastatin, fluvastatin, pravastatin, rosuvastatin, simvastatin, etc.), cholesterol absorption inhibitors (ezetimibe, etc.), and others. Anti-hypertensive drugs included diuretics (amiloride, bumetanide, etc.), calcium channel

blockers (amlodipine, nifedipine, diltiazem, verapamil, etc.), angiotensin-converting enzyme inhibitors (enalapril, lisinopril, ramipril, etc.), angiotensin II receptor antagonists (losartan, valsartan, etc.), adrenergic receptor antagonists (atenolol, metoprolol, nadolol, etc.) and others. Sedentary activity duration was the sum of self-reported number of hours spent on driving, using computer, or watching television [17]. Summaries of these environmental factors are provided in **Supplementary Table S2**.

Significance of the SNP-exposure interaction effects was assessed for each pair of near-independent pvQTL SNP and candidate exposure by comparing the two following linear regression models using a likelihood ratio test:

$$Y \sim \beta_0 + G\beta_G + E\beta_E + G \times E\beta_{G\times E} + \boldsymbol{Q\gamma}$$

and

$$Y \sim \beta_0 + G\beta_G + E\beta_E + \boldsymbol{Q\gamma}$$

where Y represents the phenotype with phenotypic mean β_0; G and E represent the genotype at an identified pvQTL SNP, and one of the nine candidate exposures with effects β_G and β_E respectively; $G \times E$ stands for the SNP-exposure interaction with effect $\beta_{G\times E}$; $\boldsymbol{Q}$ represents other covariates (age, sex, recruitment center, genotyping array, the first 20 genetic principal component, as well as the polygenic risk score) with effects $\boldsymbol{\gamma}$.

We repeated the analyses of dominance effects and SNP-exposure interaction effects for all near-independent QTL SNPs identified in GWASs.

Incorporating dominance effects into polygenic risk scores

To potentially improve polygenic predictions, for each trait, we built a multivariate linear regression model on the discovery set to combine the polygenic risk score with all pvQTLs demonstrating significant dominance effects, adjusted for age, sex, recruitment center, genotyping array, and the first 20 genetic principal components. To account for multiple testing (185 pvQTLs x 2 alleles = 370 tests), we experimented with both an FDR threshold and a Bonferroni threshold for calling significant dominance effects. Using the estimated regression coefficients and the Test dataset, we then compared the predictive performance of a polygenic risk score including dominance effects with the original polygenic risk score. We bootstrapped the test set 1,000 times to obtain a confidence interval for the predictions.

We refrained from incorporating SNP-exposure interaction effects into polygenic risk scores due to potential confounding and the possibility of reverse causation.

Acknowledgements

This research has been conducted using the UK Biobank resource under Application Number 27449 and 60755. This study was enabled in part by support provided by Calcul Québec and Compute Canada. C.M.T.G. is supported by a Canadian Institutes of Health Research grant (CIHR; PJT-148620). The Richards research group is supported by the Canadian Institutes of Health Research (365825; 409511), the Lady Davis Institute of the Jewish General Hospital, the Canadian Foundation for Innovation, the NIH Foundation, Cancer Research UK, Genome Québec, the Public Health Agency of Canada and the Fonds de Recherche Québec Santé (FRQS). J.B.R. is supported by a FRQS Clinical Research Scholarship Merite. T.L. has been supported by a Vanier Canada Graduate Scholarship, an FRQS Doctoral Training Fellowship and a McGill University Faculty of Medicine Scholarship.

Conflict of Interest

J.B.R. is the founder of 5 Prime Sciences, and has served as a consultant to GlaxoSmithKline and Deerfield Capital for their genetics programs. The other authors have no relevant disclosures.

Data availability

Full summary statistics of genome-wide discovery of pvQTLs and supplementary tables are deposited in a figshare repository https://figshare.com/s/fc68bcca435bb0a29787, currently available via this private link and will be made publicly available after publication. TL had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Restrictions apply to the availability of individual-level data from the UK Biobank to preserve patient confidentiality. These data are available from UK Biobank upon successful project application to the research committee. Computational scripts to conduct analyses included in this research will be available upon reasonable request made to the corresponding author.

Author contributions

T.L., C.M.T.G., and J.B.R. designed and directed the study. T.L. and C.M.T.G. designed analytical framework. V.F. managed data and computational software. T.L. performed the analyses and wrote the initial manuscript. All authors revised and approved the manuscript.

Table Legends

Table S1: Cohort characteristics of unrelated UK Biobank White British participants.

Table S2: Summary of environmental and lifestyle exposures in the UK Biobank discovery set (N = 254,376).

Table S3: Proportion of phenotypic variance explained by polygenic risk scores in the UK Biobank. Proportion of variance explained (adjusted R2) was assessed by linear regression.

Table S4: Summary of all near-independent QTL SNPs used for constructing polygenic risk scores. Results of COJO analyses are included.

Table S5: Summary of all near-independent pvQTL SNPs. Results of COJO analyses are included.

Table S6: Annotation of pvQTLs.

Table S7: Summary statistics of genetic correlation estimates based on QTLs and pvQTLs respectively.

Table S8: Summary of dominance effects and SNP-exposure interaction effects tested for near-independent QTL SNPs, and QTL-pvQTL colocalization.

Table S9: Summary of dominance effects and SNP-exposure interaction effects tested for near-independent pvQTL SNPs.

Table S10: Additional dominance effects with an FDR < 0.05 combined with polygenic risk scores. Coefficients obtained from multivariate linear regression are provided. Polygenic risk scores were standardized to have zero mean and unit variance.

Table S11: Additional dominance effects with Bonferroni-corrected p-value < 0.05 combined with polygenic risk scores. Coefficients obtained from multivariate linear regression are provided. Polygenic risk scores were standardized to have zero mean and unit variance.

Figure Legends

Figure 1. Study overview. The White British participants in the UK Biobank were randomly split into three disjoint subsets. Third-degree or more closely related individuals were removed from the LD reference set and the test set. Genome-wide association studies were performed on the discovery set including related individuals, using linear mixed model regression. Identification of pvQTL was conducted on the discovery set excluding related individuals.

Figure 2. Identification of pvQTLs for 15 quantitative traits. (A) Manhattan plots demonstrate genetic associations with polygenic prediction variability. Blue dashed lines indicated a corrected genome-wide significance threshold of p-value < 4.6x10-9. (B) Comparison of estimated trait heritability based on GWASs and estimated heritability of polygenic prediction variability based on pvQTL studies. Heritability estimates and 95% confidence intervals are indicated for each trait.

Figure 3. pvQTLs predict polygenic prediction errors on an out-of-sample test set. pvQTL scores were derived for BMI, triglycerides, LDL, glucose, HbA1c, platelet, and vitamin D which had at least nine pvQTLs. Median squared prediction errors (dots) and inter-quartile ranges (error bars) are compared across pvQTL score deciles. FEV1/FVC ratio had two pvQTLs, resulting in nine pvQTL score values. Median squared prediction errors and inter-quartile ranges are summarized according to the ranks. Calcium, heel BMD, and SBP had one pvQTL each, thus median squared prediction errors and inter-quartile ranges are summarized with respect to genotypes of the corresponding pvQTL SNP. Spearman correlation estimates are indicated for each trait.

Figure 4. Enrichment of complex genetic effects amongst pvQTLs. (A) Number of pvQTLs or QTLs identified for each trait. (B) Comparison of p-value distribution for dominance effect tests

between pvQTLs and QTLs. For each genetic variant, the smaller p-value obtained from two dominance effect tests conducted using two different alleles was used to represent evidence of dominance effect for the corresponding pvQTL or QTL. (C) Comparison of p-value distribution for SNP-exposure interaction effect tests between pvQTLs and QTLs. All QTLs were categorized with respect to evidence of colocalization with pvQTL signals. QTLs with a higher CLPP were more likely to have effects on polygenic prediction variability.

Figure 5. Improving polygenic prediction by modelling dominance effects. An FDR threshold and a Bonferroni threshold were implemented separately to pre-select dominance effects to be incorporated into the polygenic risk scores. Likelihood ratio tests were performed on the discovery set to assess the significance of the jointly added dominance effects (ANOVA p-value). Adjusted R^2 metrics including covariate effects, based on the baseline models without dominance effects and the complex models with dominance effects respectively, were derived on the test set. Distributions of the relative predictive performance of the two models were obtained based on the 1,000 bootstrap samples. Paired t-tests were performed to evaluate whether adding dominance effects significantly improved the predictive performance.

Figure S1. Genetic correlation between 15 quantitative traits. The upper diagonal matrix represents genetic correlation estimated based on pvQTL studies; The lower diagonal matrix represents genetic correlation estimated based on GWASs. Significantly correlated pairs with an FDR < 0.05 are colored by estimated genetic correlation. Full summary statistics are available in Supplementary Table S7.

Figure S2. Illustration of SNP-exposure interaction effects. Each pvQTL was tested for interaction effects with nine environmental or lifestyle exposures (y-axis) on the corresponding traits (x-axis). Only pvQTLs having significant interaction effects with at least one exposure are included. Significant interaction effects with a p-value < 3.0×10^{-5} are colored by z-scores in linear regression. Full summary statistics of interaction effect tests are provided in Supplementary Table S9. FFR: FEV1/FVC ratio; TG: triglycerides.

References

1. Visscher, P.M., et al., *10 Years of GWAS Discovery: Biology, Function, and Translation.* Am J Hum Genet, 2017. **101**(1): p. 5-22.
2. Khera, A.V., et al., *Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations.* Nat Genet, 2018. **50**(9): p. 1219-1224.
3. Wand, H., et al., *Improving reporting standards for polygenic scores in risk prediction studies.* Nature, 2021. **591**(7849): p. 211-219.
4. Khera, A.V., et al., *Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood.* Cell, 2019. **177**(3): p. 587-596 e9.
5. Inouye, M., et al., *Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention.* J Am Coll Cardiol, 2018. **72**(16): p. 1883-1893.
6. Lu, T., et al., *A polygenic risk score to predict future adult short stature amongst children.* J Clin Endocrinol Metab, 2021.
7. Lu, T., et al., *Improved prediction of fracture risk leveraging a genome-wide polygenic risk score.* Genome Med, 2021. **13**(1): p. 16.
8. Lu, T., et al., *Polygenic risk for coronary heart disease acts through atherosclerosis in type 2 diabetes.* Cardiovasc Diabetol, 2020. **19**(1): p. 12.
9. Lewis, C.M. and E. Vassos, *Polygenic risk scores: from research tools to clinical instruments.* Genome Med, 2020. **12**(1): p. 44.
10. Lu, T., et al., *Individuals with common diseases but with a low polygenic risk score could be prioritized for rare variant screening.* Genet Med, 2021. **23**(3): p. 508-515.
11. Ritchie, S.C., et al., *Integrative analysis of the plasma proteome and polygenic risk of cardiometabolic diseases.* Nat Metab, 2021.
12. Martin, A.R., et al., *Clinical use of current polygenic risk scores may exacerbate health disparities.* Nat Genet, 2019. **51**(4): p. 584-591.
13. Martin, A.R., et al., *Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations.* Am J Hum Genet, 2017. **100**(4): p. 635-649.
14. Mostafavi, H., et al., *Variable prediction accuracy of polygenic scores within an ancestry group.* Elife, 2020. **9**.
15. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. **562**(7726): p. 203-209.
16. Yang, J., et al., *FTO genotype is associated with phenotypic variability of body mass index.* Nature, 2012. **490**(7419): p. 267-72.
17. Wang, H., et al., *Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank.* Sci Adv, 2019. **5**(8): p. eaaw3538.
18. Varona, L., et al., *Non-additive Effects in Genomic Selection.* Front Genet, 2018. **9**: p. 78.
19. Huber, C.D., et al., *Gene expression drives the evolution of dominance.* Nat Commun, 2018. **9**(1): p. 2750.
20. Aschard, H., et al., *Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases.* Am J Hum Genet, 2012. **90**(6): p. 962-72.
21. Kerin, M. and J. Marchini, *Inferring Gene-by-Environment Interactions with a Bayesian Whole-Genome Regression Model.* Am J Hum Genet, 2020. **107**(4): p. 698-713.

22. Zhu, Z., et al., *Dominance genetic variation contributes little to the missing heritability for human complex traits.* Am J Hum Genet, 2015. **96**(3): p. 377-85.
23. Sulc, J., et al., *Quantification of the overall contribution of gene-environment interaction for obesity-related traits.* Nature communications, 2020. **11**(1): p. 1-13.
24. Yengo, L., et al., *Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry.* Hum Mol Genet, 2018. **27**(20): p. 3641-3649.
25. Hormozdiari, F., et al., *Colocalization of GWAS and eQTL Signals Detects Target Genes.* Am J Hum Genet, 2016. **99**(6): p. 1245-1260.
26. Lambert, S.A., G. Abraham, and M. Inouye, *Towards clinical utility of polygenic risk scores.* Hum Mol Genet, 2019. **28**(R2): p. R133-R142.
27. Escribe, C., et al., *Block coordinate descent algorithm improves variable selection and estimation in error-in-variables regression.* Genet Epidemiol, 2021.
28. Miller, D.T., et al., *ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG).* Genet Med, 2021.
29. Rehm, H.L., et al., *ACMG clinical laboratory standards for next-generation sequencing.* Genet Med, 2013. **15**(9): p. 733-47.
30. Richards, S., et al., *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.* Genet Med, 2015. **17**(5): p. 405-24.
31. Nikpay, M., et al., *A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease.* Nat Genet, 2015. **47**(10): p. 1121-1130.
32. Franks, P.W. and G. Pare, *Putting the Genome in Context: Gene-Environment Interactions in Type 2 Diabetes.* Curr Diab Rep, 2016. **16**(7): p. 57.
33. Paré, G., et al., *On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study.* PLoS Genet, 2010. **6**(6): p. e1000981.
34. Kaur-Knudsen, D., B.G. Nordestgaard, and S.E. Bojesen, *CHRNA3 genotype, nicotine dependence, lung function and disease in the general population.* Eur Respir J, 2012. **40**(6): p. 1538-44.
35. Stevens, V.L., et al., *Nicotinic receptor gene variants influence susceptibility to heavy smoking.* Cancer Epidemiology and Prevention Biomarkers, 2008. **17**(12): p. 3517-3525.
36. Bartlett, M.S., *Properties of sufficiency and statistical tests.* Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences, 1937. **160**(901): p. 268-282.
37. Fligner, M.A. and T.J. Killeen, *Distribution-free two-sample tests for scale.* Journal of the American Statistical Association, 1976. **71**(353): p. 210-213.
38. Young, A.I., F.L. Wauthier, and P. Donnelly, *Identifying loci affecting trait variability and detecting interactions in genome-wide association studies.* Nat Genet, 2018. **50**(11): p. 1608-1614.
39. Soave, D. and L. Sun, *A generalized Levene's scale test for variance heterogeneity in the presence of sample correlation and group uncertainty.* Biometrics, 2017. **73**(3): p. 960-971.

40. Fry, A., et al., *Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population.* Am J Epidemiol, 2017. **186**(9): p. 1026-1034.
41. McCarthy, S., et al., *A reference panel of 64,976 haplotypes for genotype imputation.* Nat Genet, 2016. **48**(10): p. 1279-83.
42. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.
43. Jiang, L., et al., *A resource-efficient tool for mixed model association analysis of large-scale data.* Nat Genet, 2019. **51**(12): p. 1749-1755.
44. Yang, J., et al., *Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits.* Nat Genet, 2012. **44**(4): p. 369-75, S1-3.
45. Zhang, F., et al., *OSCA: a tool for omic-data-based complex trait analysis.* Genome Biol, 2019. **20**(1): p. 107.
46. Zhu, Z., et al., *Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets.* Nat Genet, 2016. **48**(5): p. 481-7.
47. Bulik-Sullivan, B.K., et al., *LD Score regression distinguishes confounding from polygenicity in genome-wide association studies.* Nat Genet, 2015. **47**(3): p. 291-5.
48. Bulik-Sullivan, B., et al., *An atlas of genetic correlations across human diseases and traits.* Nat Genet, 2015. **47**(11): p. 1236-41.
49. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
50. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

Figures

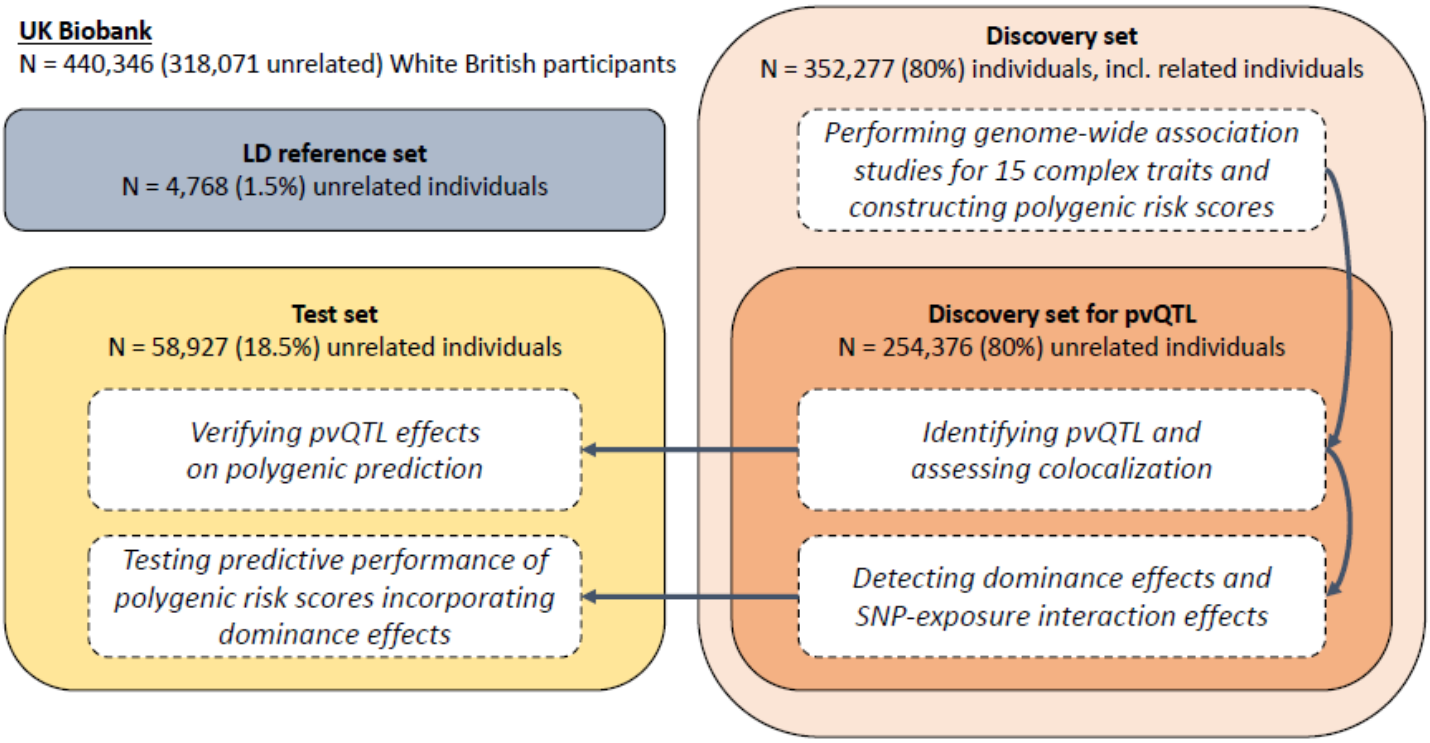

Figure 1

Study overview. The White British participants in the UK Biobank were randomly split into three disjoint subsets. Third-degree or more closely related individuals were removed from the LD reference set and the test set. Genome-wide association studies were performed on the discovery set including related individuals, using linear mixed model regression. Identification of pvQTL was conducted on the discovery set excluding related individuals.

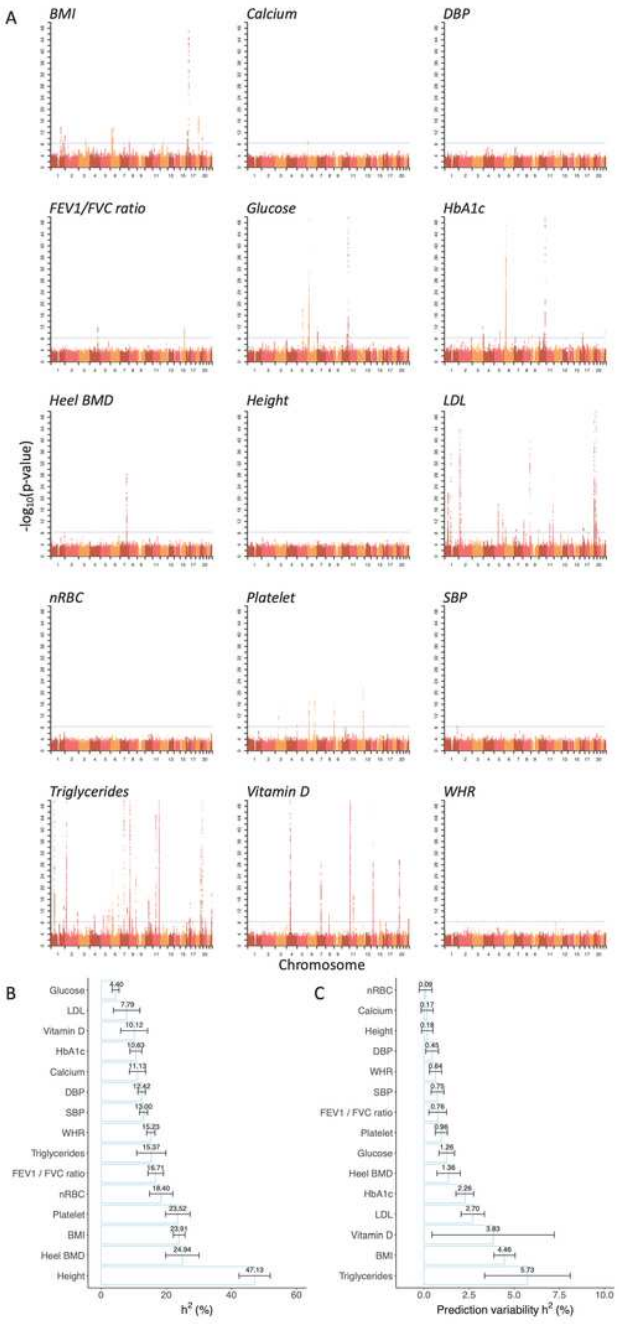

Figure 2

Identification of pvQTLs for 15 quantitative traits. (A) Manhattan plots demonstrate genetic associations with polygenic prediction variability. Blue dashed lines indicated a corrected genome-wide significance threshold of

p-value < 4.6x10-9. Comparison of (B) estimated trait heritability based on GWASs and (C) estimated heritability of polygenic prediction variability based on pvQTL studies. Heritability estimates and 95% confidence intervals are obtained from LD score regression and are indicated for each trait.

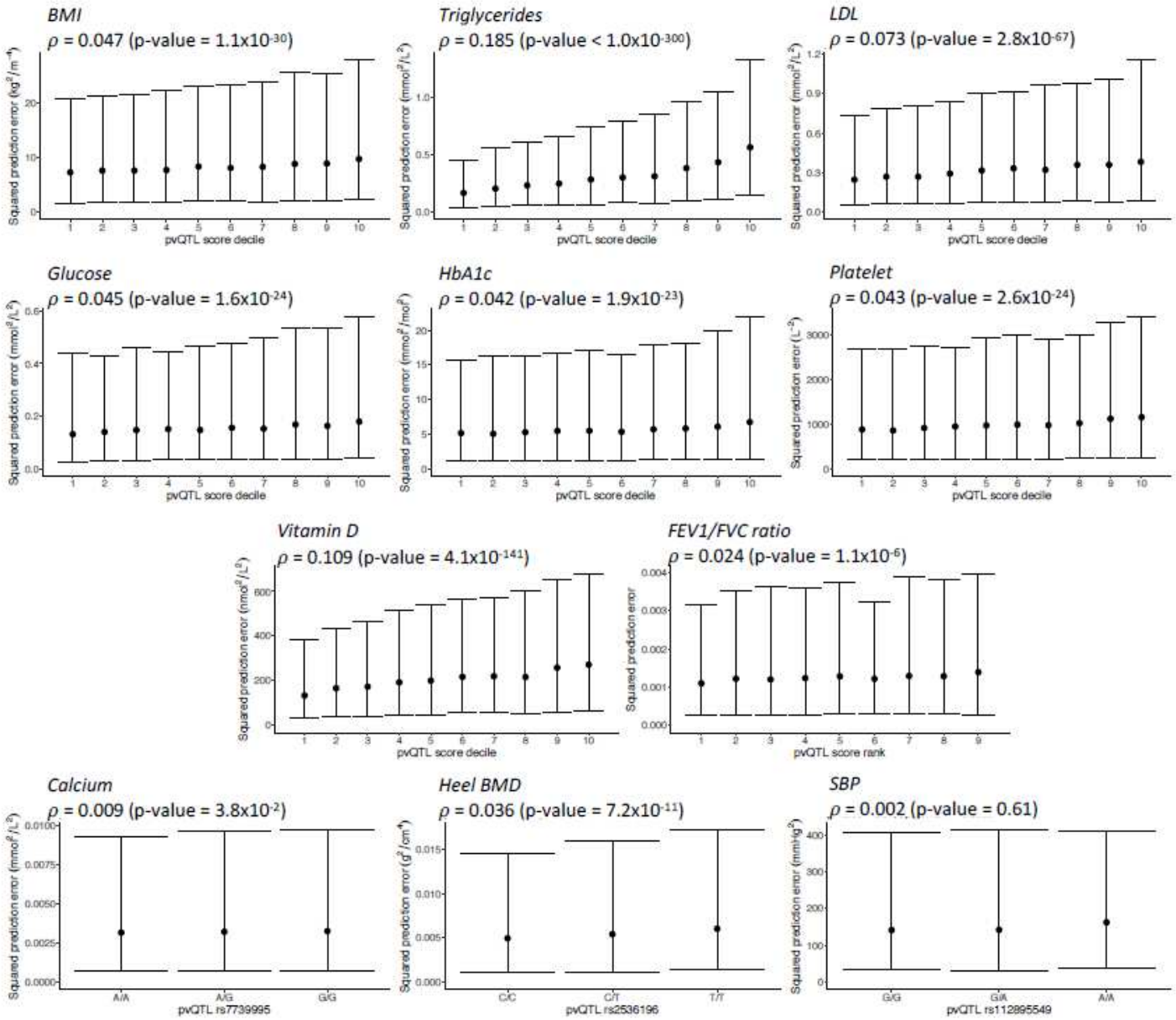

Figure 3

pvQTLs predict polygenic prediction errors on an out-of-sample test set.

pvQTL scores were derived for BMI, triglycerides, LDL, glucose, HbA1c, platelet, and vitamin D which had at least nine pvQTLs. Median squared prediction errors (dots) and inter-quartile ranges (error bars) are compared across pvQTL score deciles. FEV1/FVC ratio had two pvQTLs, resulting in nine pvQTL score values. Median squared prediction errors and inter-quartile ranges are summarized according to the ranks. Calcium, heel BMD, and SBP had one pvQTL each, thus median squared prediction errors and inter-quartile ranges are summarized with respect to genotypes of the corresponding pvQTL SNP. Spearman correlation estimates are indicated for each trait.

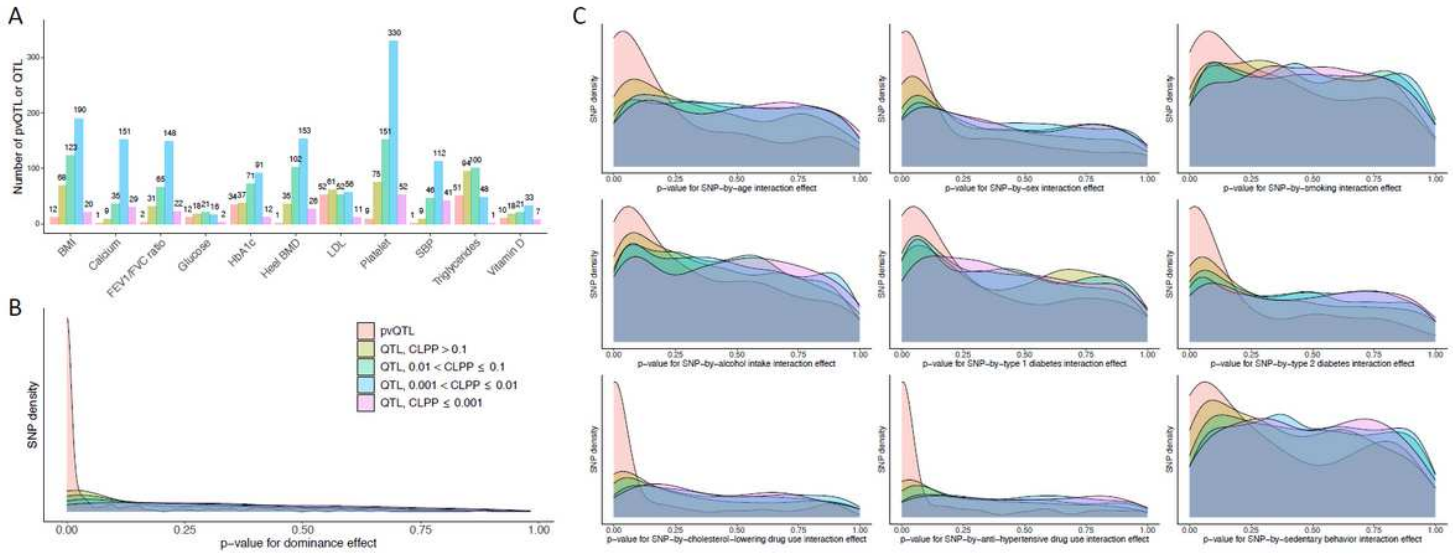

Figure 4

Enrichment of complex genetic effects amongst pvQTLs. (A) Number of pvQTLs or QTLs identified for each trait. (B) Comparison of p-value distribution for dominance effect tests between pvQTLs and QTLs. For each genetic variant, the smaller p-value obtained from two dominance effect tests conducted using two different alleles was used to represent evidence of dominance effect for the corresponding pvQTL or QTL. (C) Comparison of p-value distribution for SNP-exposure interaction effect tests between pvQTLs and QTLs. All QTLs were

categorized with respect to evidence of colocalization with pvQTL signals. QTLs with a higher CLPP were more likely to have effects on polygenic prediction variability.

Trait	# added effects	ANOVA p-value	Adjusted R^2 – Baseline model	Adjusted R^2 – Complex model	Paired t-test p-value
FDR threshold					
BMI	3	3.3×10^{-5}	0.0585	0.0588	3.7×10^{-4}
Triglycerides	87	2.1×10^{-242}	0.1193	0.1290	$< 1.0\times10^{-350}$
LDL	64	5.8×10^{-133}	0.0950	0.1024	1.0×10^{-321}
Glucose	13	2.4×10^{-36}	0.0260	0.0262	9.2×10^{-3}
HbA1c	20	3.1×10^{-44}	0.0968	0.0969	0.82
Platelet	9	1.3×10^{-110}	0.2210	0.2222	1.7×10^{-8}
Vitamin D	11	9.1×10^{-40}	0.0748	0.0756	4.5×10^{-16}
FEV1/FVC ratio	1	1.2×10^{-2}	0.1228	0.1227	0.58
Calcium	1	4.4×10^{-3}	0.0681	0.0681	0.95
Heel BMD	2	1.6×10^{-8}	0.1634	0.1636	0.15
SBP	0				
Bonferroni threshold					
BMI	2	3.0×10^{-13}	0.0585	0.0586	0.17
Triglycerides	63	$< 1.0\times10^{-350}$	0.1181	0.1297	$< 1.0\times10^{-350}$
LDL	40	1.7×10^{-167}	0.0961	0.1022	8.0×10^{-292}
Glucose	4	1.1×10^{-29}	0.0255	0.0257	1.0×10^{-2}
HbA1c	8	4.7×10^{-43}	0.0965	0.0965	0.62
Platelet	9	1.3×10^{-110}	0.2210	0.2222	1.7×10^{-8}
Vitamin D	5	1.3×10^{-32}	0.0747	0.0751	7.7×10^{-6}
FEV1/FVC ratio	0				
Calcium	0				
Heel BMD	1	2.7×10^{-6}	0.1634	0.1636	0.17
SBP	0				

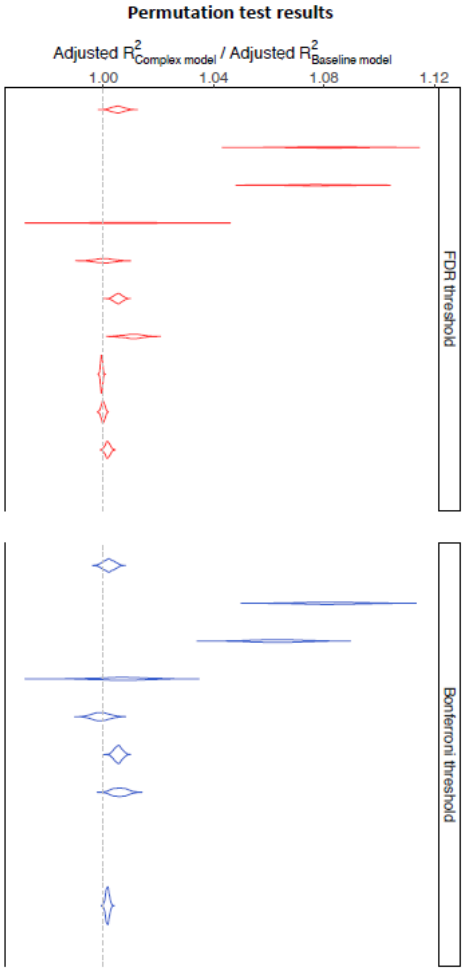

Figure 5

Improving polygenic prediction by modelling dominance effects. An FDR threshold and a Bonferroni threshold were implemented separately to pre-select dominance effects to be incorporated into the polygenic risk scores. Likelihood ratio tests were performed on the discovery set to assess the significance of the jointly added dominance effects (ANOVA p-value). Adjusted R2 metrics including covariate effects, based on the baseline models without dominance effects and the complex models with dominance effects respectively, were derived on the test set. Distributions of the relative predictive performance of the two models were obtained based on the 1,000 bootstrap samples. Paired t-tests were performed to evaluate whether adding dominance effects significantly improved the predictive

performance.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesS1S3.docx
- TableS4.xlsx
- TableS5.xlsx
- TableS6.xlsx
- TableS7.xlsx
- TableS8.xlsx
- TableS9.xlsx
- TableS10.xlsx
- TableS11.xlsx