Controlling for off-target genetic effects using polygenic scores improves the power of genome-wide association studies

Declan Bennett[1,], **Derek Morris**[2], **and Cathal Seoighe**[1,*]

[1]National University of Ireland, Galway, School of mathematics, Statistics and applied Mathematics, Galway, H91 TK33, Ireland
[2]National University of Ireland, Galway, Department of Biochemistry, Galway, H91 TK33, Ireland
[*]cathal.seoighe@nuigalway.ie

ABSTRACT

Ongoing increases in the size of human genotype and phenotype collections offer the promise of improved understanding of the genetics of complex diseases. In addition to the biological insights that can be gained from the nature of the variants that contribute to the genetic component of complex trait variability, these data have brought forward the prospect of predicting complex traits and the risk of complex genetic diseases from genotype data. Optimal realization of these objectives requires ongoing methodological developments, designed to identify true trait-associated variants and accurately predict phenotype from genotype. These methods must be computationally efficient, in order to remain tractable in the context of high variant densities and very large sample sizes. Here we show that the power of linear mixed models that are in widespread use for GWAS can be increased significantly by modeling off-target genetic effects using polygenic scores derived from SNPs that are not on the same chromosome as the target SNP. Using simulated and real data we found that this can result in a substantial increase in the number of variants passing genome-wide significance thresholds. This increase in power to detect trait-associated variants also translates into an increase in the accuracy with which the resulting polygenic score predicts the phenotype from genotype data. Our results suggest that advances in methods for phenotype prediction can be exploited to improve the control of off-target genetic effects, leading to more accurate GWAS results and further improvements in phenotype prediction.

Introduction

Linear mixed effects models (LMMs) are routinely applied to detect associations between SNPs and phenotypes in genome-wide association studies (GWAS). The contribution of a given SNP to the phenotype is modelled as a fixed effect, while the contribution of all other SNPs is modelled as a random effect, with the covariance structure of this random effect corresponding to the genetic relationships between the individuals in the study. In order to avoid including the component of the genetic relatedness contributed by variants in linkage disequilibrium (LD) with the tested variant, the genetic relationship matrix (GRM) is calculated using an LD-pruned set of variants[1,2]. Originally these methods assumed that a large number of SNPs may each make small contributions to the phenotypic variation, such that the phenotypic contribution of the genetic background could be modelled as a multivariate normal (with covariance matrix proportional to the GRM)[2–5]. Newer implementations of these models have allowed for the possibility of a combination of SNPs with effect sizes close to zero as well as some SNPs with large effect sizes. This can be done by modelling the effect sizes of background SNPs as a mixture of normal distributions, with a proportion, $p << 1$, of SNPs with large effect sizes and the remainder with effect sizes close to zero. This results in a 'spike and slab' mixture model for the SNP effect sizes[6] (the spike corresponds to a normal component with small variance and the slab corresponding to a second normal component with large variance). Implementations of these normal mixture models have

been developed that can be applied to biobank-scale data[7]. In addition to controlling for the direct genetic effects of other SNPs on the phenotype LMM methods have also been shown to account well for subtle population structure, and cryptic relatedness between individuals in the sample population[1–4,8].

Various options have been explored for which SNPs to include in the calculation of the GRM[9]. Including SNPs in LD with the target SNP results in loss of power, as the effect of the target SNP is partially accounted for by the random effect through the GRM. This has been referred to as proximal contamination[10]. On the other hand, including all (or most) SNPs that are not in LD with the target SNP, e.g. using a Leave One Chromosome Out (LOCO) approach, can result in dilution of the extent to which the relevant part of the genetic background is captured by the GRM. In the latter case, SNPs that are not relevant, in that they do not capture direct genetic effects or tag relevant population structure effects, effectively add noise to the GRM[10]. Alternatively, the GRM can be built from only the SNPs that are found using a linear model to be associated with the phenotype. Although this results in an increase in statistical power[5,9,11], it does not fully control for population structure and is not recommended if population structure is of substantial concern[6,9]. Methods have been developed that incorporate principal components into the GRM calculation built from significant SNPs; however, most of these methods are not suited to large biobank-scale data, without access to cloud computing or large compute farms[12–14]. Off-target SNPs can also be included in the statistical model as fixed effects and this is the recommended approach when there are SNPs with large effect sizes[9]. A model fitting approach to determine the SNPs to include as fixed effects has been developed, and this also results in increased power in GWAS[10].

As the genomic architecture of complex diseases is uncovered with the help of large biobanks, the application of GWAS results in a clinical setting is increasing in importance[15], including through the use of polygenic scores (PGS). PGS are constructed from weighted sums of allele dosages, with the weights corresponding to the effects size of the variants. Risk variants (variants associated with the phenotype) are typically inferred from the largest available GWAS, generally a meta-analysis. The clinical potential of PGS has already been shown in complex diseases such as coronary artery disease (CAD), diabetes and cancer[16–18]. In CAD, the identification of individuals with similar risk to those with rare high-risk monogenic variants has been reported[16]. Similarly, in breast cancer, pathogenic variants in BRCA1/2 account for 25% of familial risk of the disease with genome wide variants accounting for a further 18% of the risk[19,20]. European bias is a significant concern in clinical application of PGS, resulting from over-represented of individuals of European ancestry in GWAS and poor generalization of PGS when applied to individuals with genetic ancestry distinct to the population from which the score was derived[21,22].

Here, we set out to investigate whether including an additive PGS, estimated using a LOCO scheme, as a fixed effect in a linear mixed model has the potential to improve the power of GWAS. Our hypothesis is that the contribution of background genetic variation may not be adequately captured by a random effect, with covariance proportional to the GRM. In particular, variation resulting from SNPs with relatively large contributions to phenotypic variation may be better captured by the PGS. We tested this hypothesis in two ways. Firstly, using simulated data we tested for an improvement in power on the task of recovering known causal variants as a function of study size, number of causal variants and trait heritability. In addition, we

applied the method to standing height data from the UK Biobank and determined the number and characteristics of additional variants that were detected. For an objective assessment of performance on real data, where the true associations are unknown, we divided the data into test and training sets and predicted the phenotype in the test set. The improved performance on the critical task of complex phenotype prediction illustrates the utility of the PGS as a means of accounting for background genetic variance.

Results

We simulated data to evaluate the impact of including the PGS as a fixed effect in GWAS. The simulations consisted initially of a normally-distributed continuous trait in 100,000 individuals. The trait had a narrow-sense heritability (h^2) of 0.5 and there were 1,000 causal SNPs with normally-distributed effects on the trait (see Methods for details). We incorporated the LOCO PGS as a fixed effect in a linear mixed model using GCTA fastGWA[23] (we refer to this as fastGWA-PGS) and found a substantial improvement in power to detect the known causal SNPs (Fig. 1). In 100 simulations we found that fastGWA-PGS recovered, on average, 76 additional causal variants below the conventional P-value threshold of $5x10^{-8}$ compared to fastGWA, corresponding to a relative increase in power of 17%. The contribution to phenotype variance of off-target SNPs can also be modelled as a random effect in a linear mixed model. This approach is applied by Bolt-LMM, which uses a normal mixture random effect, with a component corresponding to SNPs with large effects; however, adding the PGS as a fixed effect to Bolt-LMM (which we refer to here as Bolt-LMM-PGS) led to an improvement in power (Fig. 1). In 100 simulations Bolt-LMM-PGS recovered, on average, 52 more causal variants than Bolt-LMM without the PGS fixed effect, corresponding to a 10% relative improvement in power. Bolt-LMM achieved higher power than fastGWA (Fig 2b), at a substantial cost in terms of computational speed, as has been shown previously[23]. However, the power of fastGWA-PGS exceeded that of Bolt-LMM and was intermediate between Bolt-LMM and Bolt-LMM-PGS. These results suggest that first calculating a PGS from an initial GWAS and including this as a fixed effect in a linear mixed model can provide an improved means to control for off-target genetic effects.

We calculated receiver operator characteristic (ROC) curves to investigate whether the increased number of causal variants recovered when we included the LOCO PGS as a fixed effect reflected a reduction in P-values across the board or also an improvement in the ordering of the variants, when the variants are ordered by the evidence of an association with the phenotype. Over 100 simulations we found that the area under the ROC curve (AUC) was always higher for fastGWA-PGS than for fastGWA without the LOCO PGS fixed effect. This was also the case for 99 of the 100 simulations when we added the PGS fixed effect to Bolt-LMM. The difference in sensitivity as a function of specificity (Fig. 2) showed that the sensitivity was consistently higher at a given specificity when the LOCO PGS was included as a fixed effect, indicating an improvement in the ordering of the SNPs. The increase in mean sensitivity was up to 0.052 in the case of fastGWA, corresponding to a relative increase of 8.2% (at a specificity of 0.9992). The addition of the PGS fixed effect led to a smaller but still consistent increase in sensitivity for Bolt-LMM (Fig. 2b). In this case, the greatest increase in the mean sensitivity was 0.029, corresponding to a 4.4% relative increase in sensitivity (at a sensitivity of 0.9994). We also compared the ROC curves from the simulated data

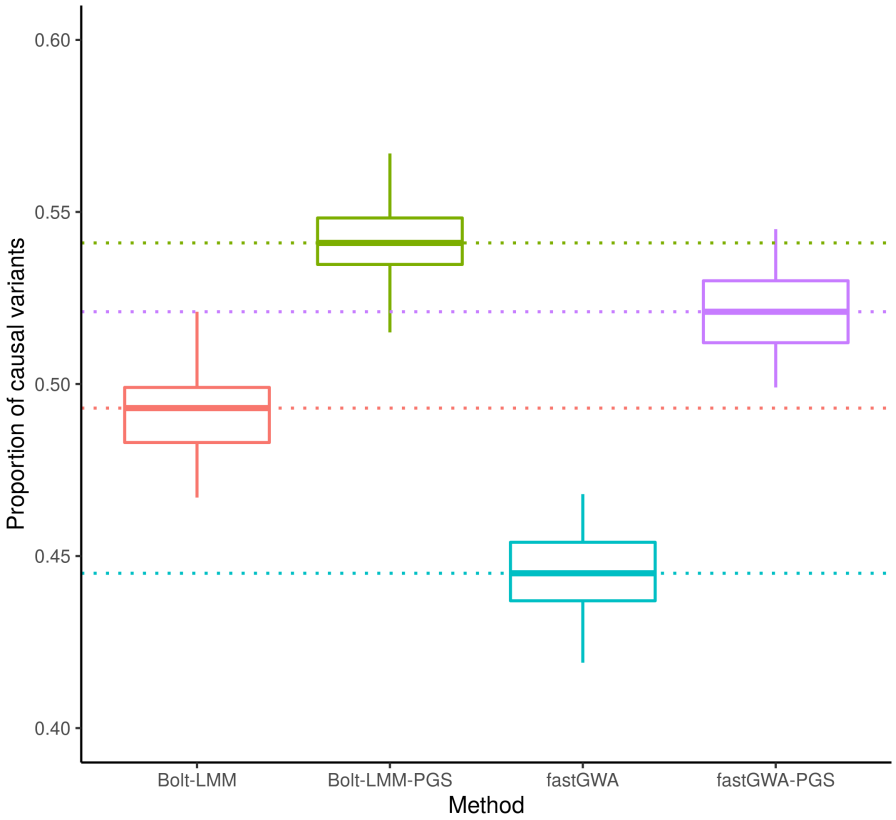

Figure 1. Proportion of causal variants identified over 100 simualtions by Bolt-LMM, Bolt-LMM-PGS, fastGWA & fastGWA-PGS.

between fastGWA-PGS and Bolt-LMM. Over the 100 simulations the mean sensitivity of fastGWA-PGS was consistently higher than Bolt-LMM, at a given specificity (the largest increase in mean sensitivity was 0.021, which corresponded to a relative increase in sensitivity of 3.2% at a specificity of 0.9991; Fig. S1, Additional File 1).

Dependence on trait heritability and number of causal variants

We simulated data over a range of values of h^2 and of the number of causal SNPs to investigate the impact of including the LOCO PGS as a fixed effect on GWAS over the range of these simulation parameters. Due to the computational intensity of Bolt-LMM-PGS this investigation is restricted to comparison of fastGWA-PGS to fastGWA and Bolt-LMM. As we would expect, for very low values of h^2 including the PGS fixed effect offers very little improvement (although the proportion of causal variants detected below a p-value threshold of $5x10^{-8}$ was always greater for fastGWA-PGS than for fastGWA). From a h^2 value of around 0.4 the proportion of the causal variants recovered tended to be significantly greater for fastGWA-PGS than for fastGWA or Bolt-LMM (Fig. 3a). Similarly, when there was a small number of causal variants (with a fixed value of h^2 = 0.5), including the polygenic score as a covariate provided little advantage; however, again, the number of causal variants recovered when the PGS was included as a covariate was always greater than or equal to the number recovered when the PGS covariate was omitted. For simulations with at least 80 causal variants the proportion of causal variants correctly identified with fastGWA-PGS always exceeded that obtained with Bolt-LMM and fastGWA (Fig. 3b).

Results of simulations with variable numbers of GWAS participants

Including the PGS as fixed effect resulted in a higher proportion of causal variants recovered across the full range of sample sizes investigated (Fig. 4). Even at N=10,000 more causal variants were recovered with fastGWA-PGS than with fastGWA or Bolt-LMM. This was somewhat surprising given that it is assumed that large sample sizes are required for accurate phenotype prediction from PGS[24]. As the sample size increased, the number of variants recovered increased sharply for all methods (fastGWA, fastGWA-PGS and Bolt-LMM), before beginning to level out at around 100,000 participants.

Application to UK Biobank height data

We assessed performance of fastGWA-PGS on real data using standing height of individuals of British ancestry (429,359 individuals) from the UK Biobank. As fastGWA is far more computationally efficient than Bolt-LMM, fastGWA-PGS was used for the UK Biobank data. The distribution of P-values obtained from fastGWA-PGS was lower than that obtained using fastGWA (See Fig. S2, Additional File 1). At a genome-wide significance level of $5x10^{-8}$ there were 127,306 and 152,626 variants that achieved significance using fastGWA and fastGWA-PGS, respectively. A total of 31,332 variants were identified by fastGWA-PGS that did not pass the significance threshold using fastGWA, termed from here as fastGWA-PGS-only variants (See Fig.S3, Additional File 1). These variants were significantly enriched for low ($< 1\%$) minor allele frequency (OR=2.44; CI=2.2-2.7; P = $6x10^{-65}$, using Fisher's Exact test; Fig. S4 Table S1, Additional File 1). We performed a gene-property analysis for tissue specificity of the fastGWA-PGS-only variants across 53 GTEx (v8) tissue types. Muscle and skeletal tissue had the strongest, albeit modest, association (P = 0.02; Fig. S5, Table S2, Additional File 1).

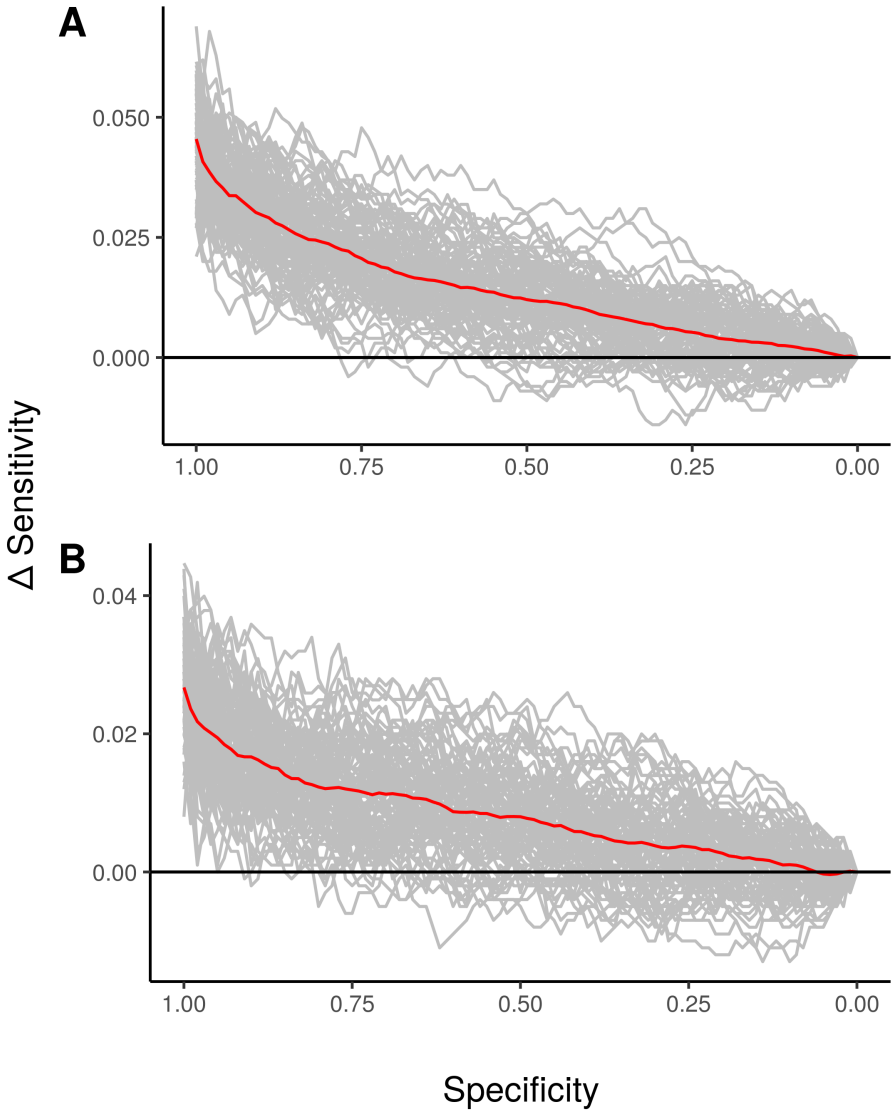

Figure 2. (A) Difference in sensitivity (between fastGWA-PGS and fastGWA) as a function of specificity. The red line shows the mean difference over all simulations. (B) Difference in sensitivity (between Bolt-LMM and Bolt-LMM-PGS) as a function of specificity.

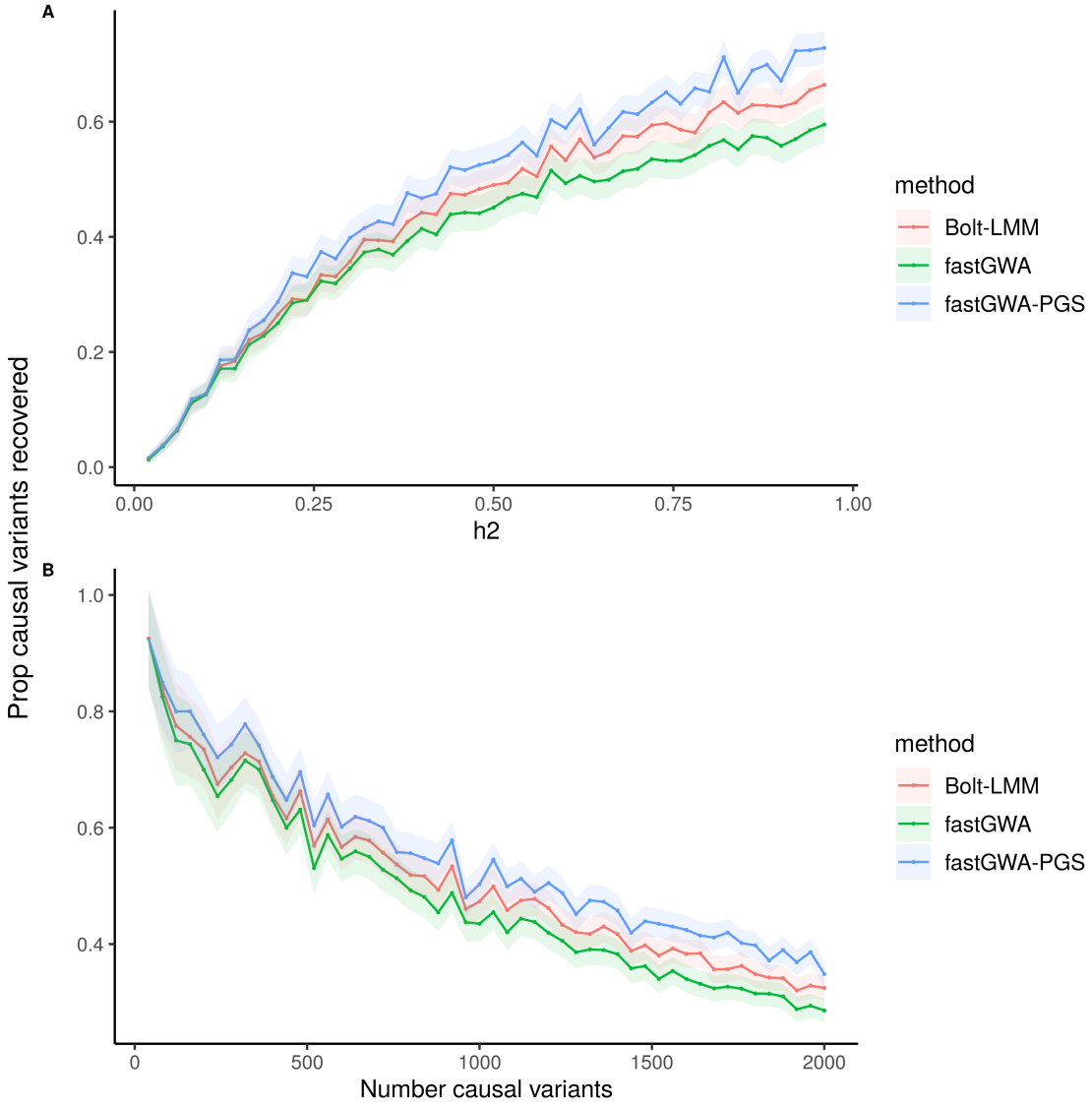

Figure 3. Proportion of causal variants identified as a function of (A) the narrow-sense heritability of the simulated trait and (B) the number of causal variants. The envelope shows ± twice the standard error.

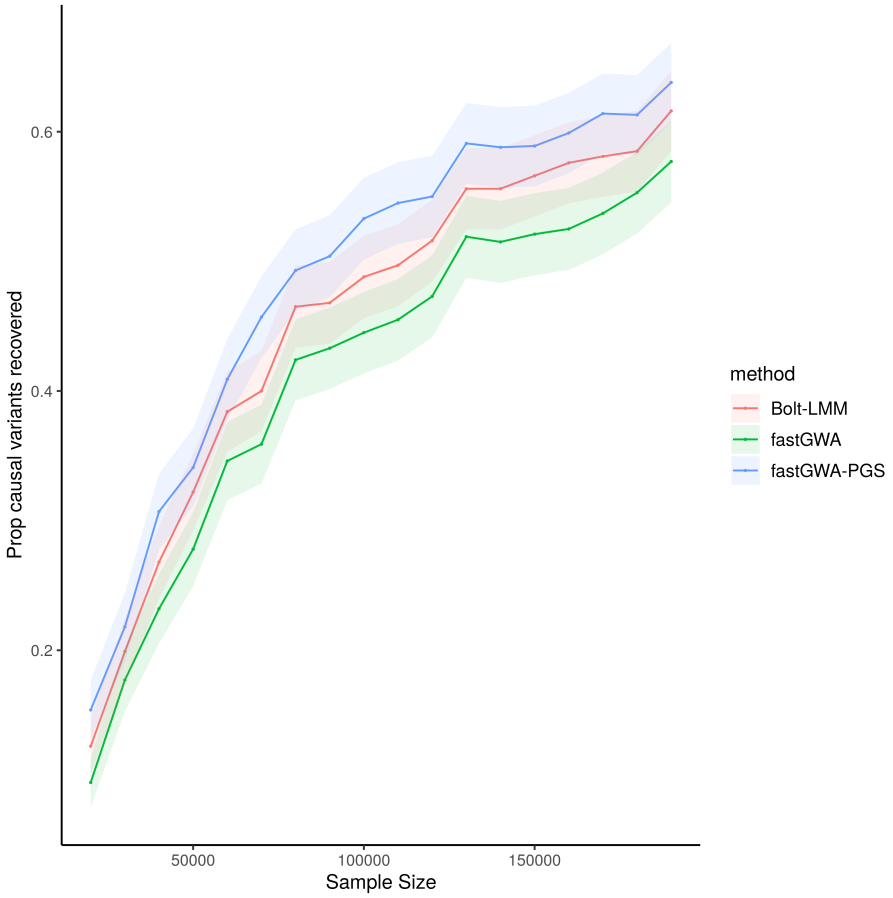

Figure 4. Proportion of causal variants identified as a function of sample size. The envelope shows $\pm$ twice the standard error.

Application to phenotype prediction

One way to determine objectively whether fastGWA-PGS outperforms fastGWA on real data is to apply both methods on the key task of phenotype prediction. We partitioned the UKB height data into an 80%/20% training and testing datasets and calculated the PGS scores using a partitioning and thresholding (P+T) method as well as a more powerful Bayesian method to estimate the phenotype in the test dataset (see Methods for details). In both cases fastGWA-PGS consistently outperformed fastGWA across all inclusion thresholds (Posterior inclusion probability (PIP) and P-value). The maximum R^2 using P+T was achieved at a P value threshold of 0.05 for both fastGWA ($R^2 = 0.129$) and fastGWA-PGS ($R^2 = 0.138$), corresponding to a relative improvement in the proportion of the phenotype explained of 7% (Fig. 5a). The Bayesian model provided substantially better prediction of the phenotype than the P+T method when variants below 0.4 PIP were included in the model and, again, this performance was further enhanced when the SNP effects were estimated using fastGWA-PGS rather than fastGWA (Fig. 5b). The maximum R^2 for fastGWA-PGS was 0.194, compared to 0.184 with fastGWA (corresponding to a relative increase of 5.4%). Although smaller, the improvement in performance on the prediction task remained consistent when we repeated this analysis with a 20%/80% training-test split (Fig. S6, Additional File 1). There are three possible (non-mutually exclusive) reasons for the improvement in prediction obtained using fastGWA-PGS. It could result from a larger number of SNPs passing a given threshold being used in prediction, from improvement in the ordering of SNPs, so that a better set of SNPs are being used or, lastly, from better estimates of the SNP effect sizes. To distinguish between these possibilities we repeated the prediction using the same number of SNPs for fastGWA-PGS and fastGWA and then for the same set of SNPs for both (see Methods for details). The performance gain was maintained when we used the same number of SNPs, but was much smaller when we used exactly the same set of SNPs (Table S3, Additional File 1).

Discussion

Omitting covariates that are associated with a response and independent of an effect of interest can result in a reduction in the efficiency of the estimation of the effect of interest[25]. Complex traits are associated with the genotype of many loci across the genome, but the effects of the off-target SNPs are frequently not explicitly modelled in GWAS and, instead tests of association are performed for one target SNP at a time. We evaluated a simple two-stage approach to accounting for off-target genetic effects that consists of performing an initial GWAS and using the summary statistics to calculate a polygenic score and then including the polygenic score, derived from SNPs not on the same chromosome as the target SNP, as a fixed effect in a second round of association testing. We found that this led to a substantial improvement in GWAS power using simulated data. A popular approach to account for the effects of off-target SNPs is to make use of a mixture distribution for the random effect in a linear mixed model, with a component with large variance corresponding to off-target SNPs with large effect sizes. This approach is implemented in Bolt-LMM[6]. However, we found that incorporating the PGS as a fixed effect in the linear model offers a better alternative for accounting for off-target genetic effects. This led to improvements in power for both fastGWA and Bolt-LMM, two commonly used implementations of linear mixed models for GWAS. In the case of Bolt-LMM-PGS,

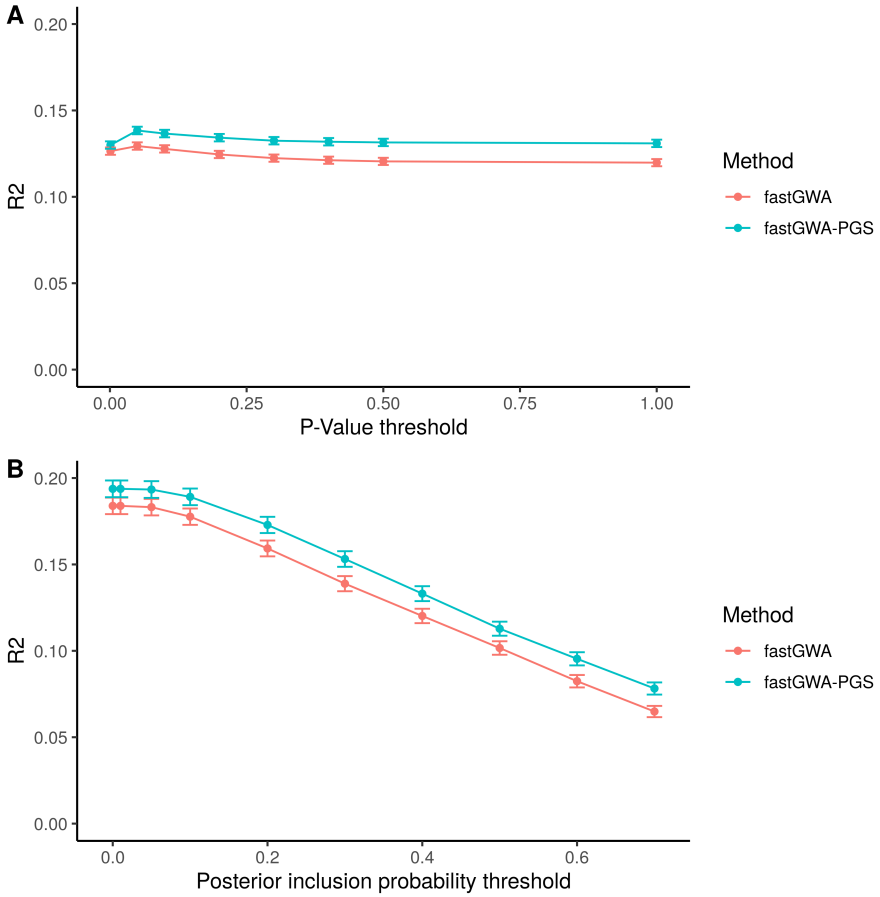

Figure 5. P+T prediction (A) across the range of P-value thresholds for each method. Bayesian prediction (B) across PIP scores. Summary statistics are calculated on the 80% training data

the effects of off-target SNPs are modelled with both a fixed and a random effect. Bolt-LMM-PGS had higher power than fastGWA-PGS (Fig. 1); however, this improvement came at a substantial cost, as fastGWA is much more computationally efficient than Bolt-LMM[23].

The increase in power using the PGS fixed effect was largest for simulated phenotypes with high heritability and a large number of causal variants (Fig. 3b). In these cases the many off-target SNPs collectively explain a substantial proportion of the phenotypic variance and summarizing the contribution of these off-target SNPs to the phenotype via the LOCO PGS is likely to result in a better estimate of the effect of the target SNP and its standard error. The boost in performance derived from including the LOCO PGS as a fixed effect was evident even for relatively small study sizes. Across all the simulation parameters we investigated, the performance of the fastGWA-PGS was never worse than fastGWA without the LOCO PGS. We also note that we calculated the LOCO PGS using SNPs that were selected based on a fixed P-value threshold. Further increases in power may be possible by optimizing the SNPs that are used to calculate the PGS separately for each omitted chromosome.

We also applied the method to real data (standing height in individuals of British ancestry in the UK Biobank). Height was chosen as there have been over 3,000 near independent variants associated with height, explaining 0.483 of the phenotypic variance. Heritability is often estimated to be as high as 80%. Although GWAS can only give us insight into h^2, the additive component of heritability, estimates of h^2 for height range from 0.3 for high quality common variants up to 0.88 in monozygotic twins for all SNPs[26–28]. Consistent with the simulation results, we found nearly 20% more associated variants with fastGWA-PGS than with fastGWA, including a large number of variants (31,332) that were significant using fastGWA-PGS, but not using fastGWA (a much smaller set of 6,012 SNPs were identified as significant by fastGWA but not fastGWA-PGS; Fig. S3, Additional File 1). The newly identified variants are enriched for rare variants while common variants were depleted. The additional variants identified using fastGWA-PGS were associated with expression in skeletal and muscle tissue, consistent with fastGWA-PGS identifying additional true causal variants.

The h^2 estimate in the simulations corresponds to the combined additive contribution to the variance of the causal SNPs; however, when we apply GWAS methods to the simulated data we do not recover all of the associated SNPs and, consequently, the estimated SNP heritabilities tends to be lower than the simulated values. For example, the simulations that were the basis of Figure 1 used $h^2 = 0.5$, but the mean estimated h^2 for these simulations was 0.41. This has implications for the interpretation of Figure 3a, which shows the relationship between simulated h^2 and power, rather than the estimated h^2. Within UK Biobank, standing height has the highest estimated SNP h^2, at 0.52[23]. fastGWA-PGS offers improvement over fastGWA at higher values of h^2, so we can expect it to yield more independent association signals for other traits such as weight, heel bone mineral density, basal metabolic rate and platelet count that all have estimated $h^2 > 0.2$ in UK Biobank[23,29]. These are traits with high heritabilities that can be easily measured in large population-based cohorts, which has enabled GWAS to already detect many association signals. Some diseases have equally high heritability, e.g., type 1 diabetes, schizophrenia and Alzheimer's disease, but do not yet have sample sizes on the same scale as height. As this changes and statistical power of GWAS of these diseases increases, inclusion of the LOCO PGS as a fixed effect in GWAS may prove useful for identifying additional risk loci for them

over and above conventional GWAS analysis.

The use of polygenic scores to contribute to phenotype prediction from genotype is an increasingly important application of the results of GWAS[30]. Recent work has shown that PGS can identify sizeable proportions of a population with substantially greater risk for CAD, type II diabetes (T2D), arterial fibrillation (AF), infammatory bowel disease (IBD) and breast cancer[16]. A separate study using the FinnGen dataset compared the overall lifetime risk of coronary heart disease (CHD), T2D, AF, breast cancer and prostate cancer between individuals having an average compared to a high PGS score. A high PGS score was associated with a 21% to 38% increase in overall lifetime risk with age of onset 4 to 9 years earlier than in the average group[31]. Prophylactic interventions can reduce both the physical and economic stress on health systems by entering individuals on measures to prevent and manage disease earlier. PGS when used to target individuals with high efficacy of treatment can also reduce the numbers of individuals to treat[32]. Although having a high PGS for a disease is not akin to a diagnosis, it can guide individuals to limit their exposure to other risk factors. Performing GWAS on a subset of samples and predicting on the remainder, we observed a consistent improvement in accuracy of the prediction of standing height in UK Biobank when we included the PGS as a fixed effect. Although there are many caveats surrounding the prediction of complex phenotypes, particularly clinical phenotypes[22], accurate prediction of disease susceptibility can have profound implications for disease management and prevention. Our results suggest that incorporating PGS into the GWAS before recalculating PGS or otherwise predicting the phenotype or disease risk leads to improved performance on this crucial task.

Conclusion

The tasks of detecting trait-associated variants and predicting the trait in a new sample from the summary statistics of these variants are closely intertwined. Improved performance on the trait-association task can result in more associated variants and better estimates of their effect sizes, resulting in improvement on the prediction task. On the other hand, improved methods for phenotype prediction can help to control for off-target effects in methods that identify the trait-associated variants and their effects. The method that we have explored here consists of incorporating a LOCO PGS as a fixed-effect covariate to control for these off-target effects; however, any method for phenotype prediction could play this role, once its application is restricted to variants that are not linked to the target SNP. We show here that incorporating the PGS as a fixed-effect covariate results in increased power to detect trait-associated variants in GWAS. The resulting trait-associated variants and effect size estimates lead to an improvement in the PGS, as illustrated by improved performance in the task of predicting the phenotype in a test dataset.

Methods

Genotype data

The simulations and application to real data were based on autosomal genotype data from the UK Biobank. To limit the effects of population stratification only individuals reporting white British ancestry (data field 21000; code 1001) were included in all

analyses. The genotype data for the simulation analysis was based on directly genotyped variants with minor allele frequency (MAF) greater than 0.005%. Variants with genotype missingness greater than 2% or that failed a test for Hardy-Weinberg equilibrium (HWE) at $\alpha = 0.0001$ were excluded. To compute the genetic relationship matrix (GRM), a set of variants was extracted with MAF greater than 0.25%, a genotype missingness rate of less than 2% and passing a HWE test at $\alpha = 0.0005$. The GRM set was further reduced by elimination of variants with an R^2 greater than 0.1 in a 200 variant sliding window of size 500. All genotype QC was implemented in plink2[33]. The sparse GRM required by fastGWA was created by setting entries corresponding to sample pairs with an estimated relatedness of less than 0.05 to 0. To account for population structure in the association studies, principal component analysis (PCA) was performed on the set of GRM variants using plink2.

Simulations

Based on the above genotype data, we simulated a continuous phenotype using the GCTA software suite[34]. The initial simulation consisted of 100,000 individuals, 1,000 causal variants and $h^2 = 0.5$ (Fig. S7, Additional File 1). This simulation was repeated 100 times with 165,037 variants remaining after variant filtering for the GRM calculation. Power was calculated as the proportion of the causal variants recovered. To calculate false positive rates we first removed all SNPs within 1 Mb of the causal SNPs. Further simulations were carried out to investigate the effects of varying the number of causal SNPs, h^2 and the sample size on method performance. In each case all parameters other than the one being varied were the same as the initial simulation, and one simulation was performed per set of parameter values. The pROC R package was used to generate receiver operating characteristic (ROC) curves[35].

Association tests

Association testing was performed using fastGWA and using Bolt-LMM with 10 PCs, genotype batch and assessment centre as fixed covariates. For the PGS method we first performed GWAS (using fastGWA or Bolt-LMM) and calculated PGS scores on a Leave One Chromosome Out (LOCO) basis. This resulted in 22 sets of PGS values (one for each autosomal chromosome, calculated from the summary statistics of variants on all other autosomal chromosomes). The PGS were calculated using PRSice2 (version 2.2.12 (2020-02-20))[36]. To decrease computation time and reduce the likelihood of overfitting a p-value threshold of 5e-05 was chosen a priori for the LOCO PGS calculation. Association testing was then performed using fastGWA in a chromosome-wise manner with the corresponding LOCO PGS included as a fixed effect.

Application to standing height from the UK biobank

Autosomal genotype data was cleaned using MAF of 0.1%, SNP-missingness of 2% and sample-missingness cut-off of 10% and imputation quality thresholds of 0.8. Variants that failed HWE testing at $\alpha = 5x10^{-7}$ were excluded from subsequent analysis. The total number of variants remaining after filtering was 9,156,362. The GRM was calculated using a set of LD-pruned variants with a MAF of at least 0.01, window size of 1000, R^2 of 0.9 and window shift of 100 variants. This left a total of 1,326,927 to be used in the calculation of the sparse GRM using the same methods as described for the simulation analysis. Batch number, sex, age, assessment centre and 10 PCs were included as covariates in all association tests.

Gene-property analysis

To test whether the fastGWA-PGS-only variants had an association with tissue specific expression data from GTEx, gene-property analysis was performed using the SNP2GENE functionality of the FUMA web platform[37]. Briefly, FUMA using MAGMA[38], maps variants to genes assigning a P-value based gene score. The Z-transformed gene scores are then modelled along with the log2 RPKM average expression in the tissue of interest, log2 RPKM of average expression across all tissues and technical confounders. To test for a positive relationship a one-sided test is then performed.

Prediction

To test the performance of fastGWA-PGS on the task of predicting standing height, the UK Biobank data was partitioned into 80/20 and an independent 20/80 training and test datasets. We used two PGS strategies, pruning and thresholding (P+T) and a Bayesian method (SBayesR)[36,39]. As the Bayesian method requires LD matrices we used a pre-computed set of matrices built on 2.8 million common pruned UK Biobank variants. The precomputed LD matrices for 2.8 million high-confidence UK Biobank variants were downloaded from the Zenodo public repository (10.5281/zenodo.3375373). Both PRSice and SBayesR predictions were performed using this 2.8 million UK Biobank set of genotypes.

For the P+T method, the same fastGWA-PGS methodology was used as in the simulation study and PGS scores were calculated in the test set using P-value thresholds (0.001, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1). The Bayesian method replaces the P+T method in the PGS-LOCO step with SBayesR. To limit the effect of over-fitting, we excluded variants with a posterior inclusion probability (PIP) score of less than 0.2 in the PGS-LOCO calculation. As SBayesR uses Bayesian inference to estimate the true effect sizes, Plink2 was used to calculate the PGS-LOCO score from the genotype and SBayesR effect size data before association testing. SBayesR was then applied once more on the fastGWA-PGS summary statistics over a range of PIP scores. The model fit was assessed for each method by fitting a linear model to the values of the phenotype in the test set as a function of their predicted values.

We repeated the phenotype prediction using the same number of SNPs (74,000) with fastGWA-PGS and fastGWA. We chose 74,000 as it fell in the range of the number of top SNPs that maximise the R^2 value for fastGWA-PGS and fastGWA. We also performed the prediction using the 37,514 SNPs that were in common among the top 74,000 SNPs from fastGWA and fastGWA-PGS.

Data Availability

Code to implement the fastGWA-PGS method described in this work is available under MIT license from the github repository, https://github.com/declan93/PGS-LMM/.

References

1. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. genetics* **38**, 203–208 (2006).

2. Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).

3. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. genetics* **42**, 348 (2010).

4. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. genetics* **44**, 821–824 (2012).

5. Listgarten, J., Lippert, C. & Heckerman, D. Fast-lmm-select for addressing confounding from spatial structure and rare variants. *Nat. genetics* **45**, 470–471 (2013).

6. Loh, P.-R. *et al.* Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. genetics* **47**, 284 (2015).

7. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. genetics* **50**, 906–908 (2018).

8. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).

9. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. genetics* **46**, 100 (2014).

10. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nat. methods* **9**, 525 (2012).

11. Lippert, C. *et al.* The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Sci. reports* **3**, 1815 (2013).

12. Tucker, G., Price, A. L. & Berger, B. Improving the power of gwas and avoiding confounding from population stratification with pc-select. *Genetics* **197**, 1045–1049 (2014).

13. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in uk biobank. *Nat. genetics* **50**, 1593–1599 (2018).

14. Kadie, C. & Heckerman, D. Ludicrous speed linear mixed models for genome-wide association studies. *bioRxiv* DOI: 10.1101/154682 (2019). https://www.biorxiv.org/content/early/2019/12/07/154682.full.pdf.

15. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).

16. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. genetics* **50**, 1219–1224 (2018).

17. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581 (2018).

18. Yanes, T., Young, M.-A., Meiser, B. & James, P. A. Clinical applications of polygenic breast cancer risk: a critical review and perspectives of an emerging field. *Breast Cancer Res.* **22**, 1–10 (2020).

19. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92 (2017).

20. Bahcall, O. Common variation and heritability estimates for breast, ovarian and prostate cancers. *Nat Genet.* **10** (2013).

21. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. molecular genetics* **28**, R133–R142 (2019).

22. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. communications* **10**, 1–9 (2019).

23. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. Tech. Rep., Nature Publishing Group (2019).

24. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS genetics* **9** (2013).

25. Neuhaus, J. M. Estimation efficiency with omitted covariates in generalized linear models. *J. Am. Stat. Assoc.* **93**, 1124–1129 (1998).

26. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry. *Hum. molecular genetics* **27**, 3641–3649 (2018).

27. Hou, K. *et al.* Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture. *Nat. genetics* 1 (2019).

28. Nolte, I. M. *et al.* A comparison of heritability estimates by classical twin modeling and based on genome-wide genetic relatedness for cardiac conduction traits. *Twin Res. Hum. Genet.* **20**, 489–498 (2017).

29. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. genetics* **51**, 1339–1348 (2019).

30. Martin, A. R., Daly, M. J., Robinson, E. B., Hyman, S. E. & Neale, B. M. Predicting polygenic risk of psychiatric disorders. *Biol. psychiatry* **86**, 97–109 (2019).

31. Mars, N. *et al.* Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat. Medicine* 1–9 (2020).

32. Gibson, G. On the utilization of polygenic risk scores for therapeutic targeting. *PLoS genetics* **15** (2019).

33. Chang, C. C. *et al.* Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742–015 (2015).

34. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. Gcta: a tool for genome-wide complex trait analysis. *The Am. J. Hum. Genet.* **88**, 76–82 (2011).

35. Robin, X. *et al.* proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics* **12**, 77 (2011).

36. Choi, S. W. & O'Reilly, P. F. Prsice-2: Polygenic risk score software for biobank-scale data. *Gigascience* **8**, giz082 (2019).

37. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with fuma. *Nat. communications* **8**, 1–11 (2017).

38. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. Magma: generalized gene-set analysis of gwas data. *PLoS computational biology* **11** (2015).

39. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nat. communications* **10**, 1–11 (2019).

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 23739. This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 16/IA/4612.

Author contributions statement

CS conceived and supervised the project and performed analyses. DB implemented the pipeline and performed analyses. DB and CS wrote the manuscript, with input from DM. DM advised on application of the method to human phenotypes.

Additional information

Competing interests

The authors declare that they have no competing interests.