Identifying and correcting multiple sources of misspecification in GWAS summary statistics for polygenic scores

Florian Privé,[1,*] Julyan Arbel,[2] Hugues Aschard,[3,4] and Bjarni J. Vilhjálmsson[1,5]

[1]National Centre for Register-Based Research, Aarhus University, Aarhus, 8210, Denmark.
[2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, 38000, France.
[3]Department of Computational Biology, Institut Pasteur, Paris, 75015, France.
[4]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, 02115, USA.
[5]Bioinformatics Research Centre, Aarhus University, Aarhus, 8000, Denmark.
[*]To whom correspondence should be addressed.

Contact: florian.prive.21@gmail.com

Abstract

Are genome-wide association studies (GWAS) summary statistics of good enough quality for performing follow-up analyses? Can we detect possible misspecifications in GWAS summary statistics and correct for them in order to improve predictive performance of polygenic scores?

1 Introduction

Contrary to individual-level genotypes and phenotypes, summary statistics resulting from genome-wide association studies (GWAS) are widely available. Therefore, they have been extensively used e.g. to derive polygenic scores (PGS). However, GWAS summary statistics come with uneven qualities, such as the imputation qualities of each variant reported. There is also some heterogeneity in the methods used for performing the individual GWAS and their meta-analyses, as well as the information reported in the resulting summary statistics. Moreover, many PGS methods such as PRS-CS, SBayesR, and LDpred2 (Ge *et al.* 2019; Lloyd-Jones *et al.* 2019; Privé *et al.* 2020b) use Bayesian models, which can be sensitive to model misspecifications (Walker 2013; Miller and Dunson 2018). Previously, to make sure that input parameters passed to LDpred2 were consistent with its modeling assumptions, we proposed a quality control (QC) based on comparing standard deviations inferred from GWAS summary statistics with the ones computed from a reference panel (Privé *et al.* 2020b). This was particularly important for LDpred2-auto, which directly estimates key model parameters from the data.

Here, we investigate some of the possible misspecifications that come with GWAS summary statistics and propose corrections in order to improve the predictive performance of polygenic scores. We approach this from three different angles. First, based on additional summary information such as the imputation INFO scores and allele frequencies from the GWAS summary statistics, we refine our previously proposed QC Privé *et al.* (2020b). For instance, we show that standard deviations of imputed genotypes (allele dosages) are lower than the expected values under Hardy-Weinberg equilibrium. Second, we investigate possible corrections to apply to the input parameters of polygenic methods, namely the reference LD (linkage disequilibrium) matrix, the GWAS effect sizes, their standard errors and the corresponding sample sizes. For example, we show that GWAS effect sizes computed from imputed dosages are larger in magnitude compared to if true genotypes were available. Third, we introduce two new optional parameters in LDpred2-auto to make it more robust to these types of misspecification. We focus our investigations on LDpred2 and lassosum for two reasons. First, multiple studies have shown that LDpred2 and lassosum are consistently ranking best among methods for polygenic prediction (Mak *et al.* 2017; Privé *et al.* 2020b; Pain *et al.* 2021; Kulm *et al.* 2021). Second, we reimplement and use a new version of lassosum, called lassosum2, that uses the exact same input parameters as LDpred2, which makes it easy for us to test the QCs and corrections presented here. We perform our investigations using simulations first, then we use public summary statistics while restricting to the widely-used HapMap3 variants, finally we investigate two alternative sets with more variants.

2 Results

2.1 Misspecification of GWAS sample sizes

We design simulations where variants have different GWAS sample sizes, which is often the case when meta-analyzing GWAS from multiple cohorts without the same genome coverage. Using 40,000 variants from chromosome 22 (Methods), we simulate quantitative phenotypes with a heritability of 20% and 2000

causal variants. We then divide the 40,000 variants into three groups: for half of the variants, we use 100% of 300,000 individuals for GWAS, but use only 80% for one quarter of variants and 60% for the remaining quarter. We then run C+T, lassosum, lassosum2 (Methods), LDpred2-inf, LDpred2(-grid), and LDpred2-auto (Privé *et al.* 2019; Mak *et al.* 2017; Privé *et al.* 2020b) by using either the true per-variant GWAS sample sizes, the total sample size, or imputed sample sizes. Note that we initially included PRS-CS and SBayesR in our comparison (Ge *et al.* 2019; Lloyd-Jones *et al.* 2019). However, results for SBayesR always diverged and the overlap with the LD reference provided for PRS-CS was too small. Averaged over 10 simulations, when providing true per-variant GWAS sample sizes, squared correlations between the polygenic scores and the simulated phenotypes are of 0.123 for C+T, 0.161 for lassosum, 0.169 for lassosum2, 0.159 for LDpred2(-grid), 0.140 for LDpred2-auto, and 0.141 for LDpred2-inf (Figure 1). Results when using imputed (instead of true) sample sizes are quite similar. Note that C+T does not use this sample size information. When using the total GWAS sample size instead of the per-variant sample sizes, predictive performance slightly decreases to 0.157 for lassosum and to 0.163 for lassosum2, but dramatically decreases for LDpred2 with new values of 0.134 for LDpred2-grid, 0.119 for LDpred2-auto, and 0.123 for LDpred2-inf (Figure 1). This extreme simulation scenario shows that LDpred2 can be sensitive to GWAS sample size misspecification, whereas lassosum (and lassosum2) seems little affected by this.

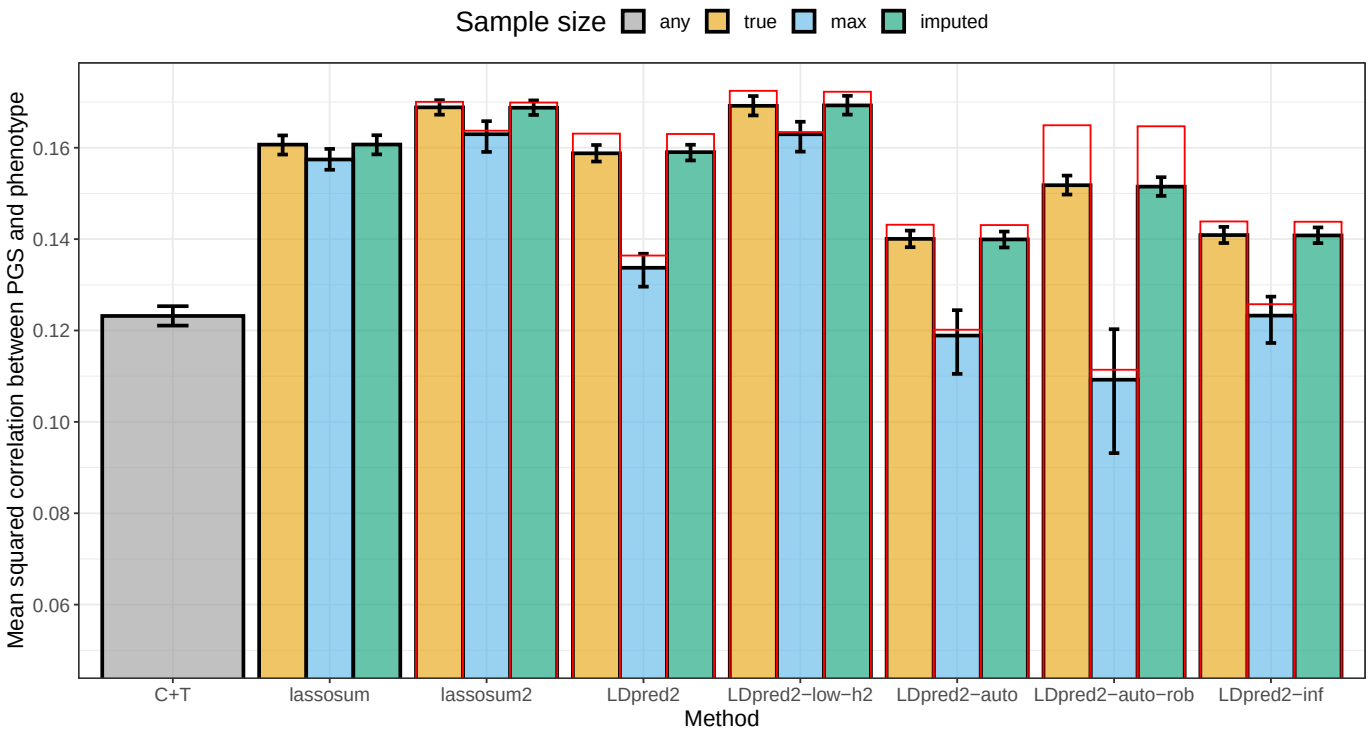

Figure 1: Results for the simulations with sample size misspecification, averaged over 10 simulations for each scenario. Reported 95% confidence intervals are computed from 10,000 non-parametric bootstrap replicates of the mean. The GWAS sample size is "true" when providing the true per-variant sample size, "max" when providing instead the maximum sample size as a unique value to be used for all variants, "imputed" (Methods), or "any" when the method does not use this information (the case for C+T). Red bars correspond to using the LD with independent blocks (Methods).

We conduct further investigations to explain results of figure 1. First, the results for LDpred2-auto are the same as with LDpred2-inf because it always converges to an infinitesimal model ($p = 1$) in these simulations. To solve this limitation, we introduce two new parameters to make LDpred2-auto more robust, referred to as "LDpred2-auto-rob" here (Methods). Second, for lassosum2, results for a grid of parameters (over λ and δ) are quite smooth compared to LDpred2 (Figures S2 and S3). In these simulations with misspecified sample sizes, it seems highly beneficial to use a small value for the SNP heritability hyper-parameter h^2 in LDpred2, e.g. a value of 0.02 or even 0.002 when the true value is 0.2 (Figure S3). Indeed, using a small value for this hyper-parameter induces a larger regularization (shrinkage) on the effect sizes. When running LDpred2(-grid) with a grid of hyper-parameters including these low values for h^2, we refer to this as "LDpred2-low-h2" here. Results with LDpred2-low-h2 improves from 0.159 to 0.169 when using true sample sizes and from 0.134 to 0.163 when using the maximum sample size. Finally, we introduce a last change for robustness here: we form independent LD blocks in the LD matrix to prevent small errors in the Gibbs sampler to propagate to too many variants (Methods). This change seems to solve convergence issues of LDpred2 in these simulations (Figure S3) and further improves predictive performance for all LDpred2 methods (Figure 1).

2.2 When using allele dosages from imputation

Marchini and Howie (2010) showed that the IMPUTE INFO measure is highly concordant with the MACH measure $\hat{r}^2_j = \frac{\text{var}(G_j)}{2\hat{\theta}_j(1-\hat{\theta}_j)}$, where $\hat{\theta}_j$ is the estimated allele frequency of G_j, the genotypes for variant j. Therefore, when using the expected genotypes from imputation (dosages), their standard deviations are often lower than the expected value under Hardy-Weinberg equilibrium, because $\text{INFO}_j \approx \frac{\text{var}(G_j)}{2\hat{\theta}_j(1-\hat{\theta}_j)}$. In simulations (cf. Methods section "Data for simulations"), we verify that $\text{sd}(G_j)^{\text{true}} \approx \text{sd}(G_j)^{\text{imp}} / \sqrt{\text{INFO}_j}$ (Figure S6). We also show that GWAS effect sizes $\hat{\gamma}$ computed from imputed dosages are over-estimated: $\hat{\gamma}_j^{\text{true}} \approx \hat{\gamma}_j^{\text{imp}} \cdot \sqrt{\text{INFO}_j}$ and $\text{se}(\hat{\gamma}_j)^{\text{true}} \approx \text{se}(\hat{\gamma}_j)^{\text{imp}} \cdot \sqrt{\text{INFO}_j}$ (Figures S8 and S7). This is the first correction of summary statistics we consider in the simulations below. Note that we recompute INFO scores for the subset of 362,307 European individuals used in this paper as they can differ substantially from the ones reported by the UK Biobank for the whole data (Figure S5). As a second option, instead of using dosages to compute the GWAS summary statistics, it has been argued that using multiple imputation (MI) would be more appropriate (Palmer and Pe'er 2016). In simulations, we show that $\hat{\gamma}_j^{\text{MI}} \approx \hat{\gamma}_j^{\text{imp}} \cdot \text{INFO}_j$ and $Z_j^{\text{MI}} \approx Z_j^{\text{imp}} \cdot \text{INFO}_j$, where $Z = \hat{\gamma}/\text{se}(\hat{\gamma})$ (Figure S9). This is the second correction of summary statistics we implement in the simulations below, along with $n_j = N \cdot \text{INFO}_j$. Finally, we consider an in-between solution as a third correction, using $\hat{\gamma}_j = \hat{\gamma}_j^{\text{imp}} \cdot \text{INFO}_j$, $\text{se}(\hat{\gamma}_j) = \text{se}(\hat{\gamma}_j)^{\text{imp}} \cdot \sqrt{\text{INFO}_j}$, and $n_j = N \cdot \text{INFO}_j$.

Using 40,000 variants from chromosome 22, we simulate quantitative phenotypes assuming a heritability of 20% and 2000 causal variants using the "true" dataset (cf. Methods section "Data for simulations"). We compute GWAS summary statistics from the dosage dataset and use these summary statistics to run LDpred2 and lassosum2 with either no correction of the summary statistics, or with one of the three corrections described above. The LD reference used by LDpred2 and lassosum2 is computed from the validation set using the dataset with the "true" genotypes. For lassosum2 and LDpred2(-grid) which tune parameters using the validation set, correcting for imputation quality slightly improves predictive performance in

these simulations (Figure 2). However, correcting for imputation quality can dramatically improve predictive performance for LDpred2-auto. Moreover, new additions for robustness introduced before, namely LDpred2-low-h2, LDpred2-auto-rob, and forming independent blocks in the LD matrix also improve predictive performance for all corrections (Figure 2).

In the real data applications hereinafter, we choose to use the first correction, "sqrt_info", which is simple because it is equivalent to post-processing PGS effects by multiplying them by $\sqrt{\text{INFO}}$.

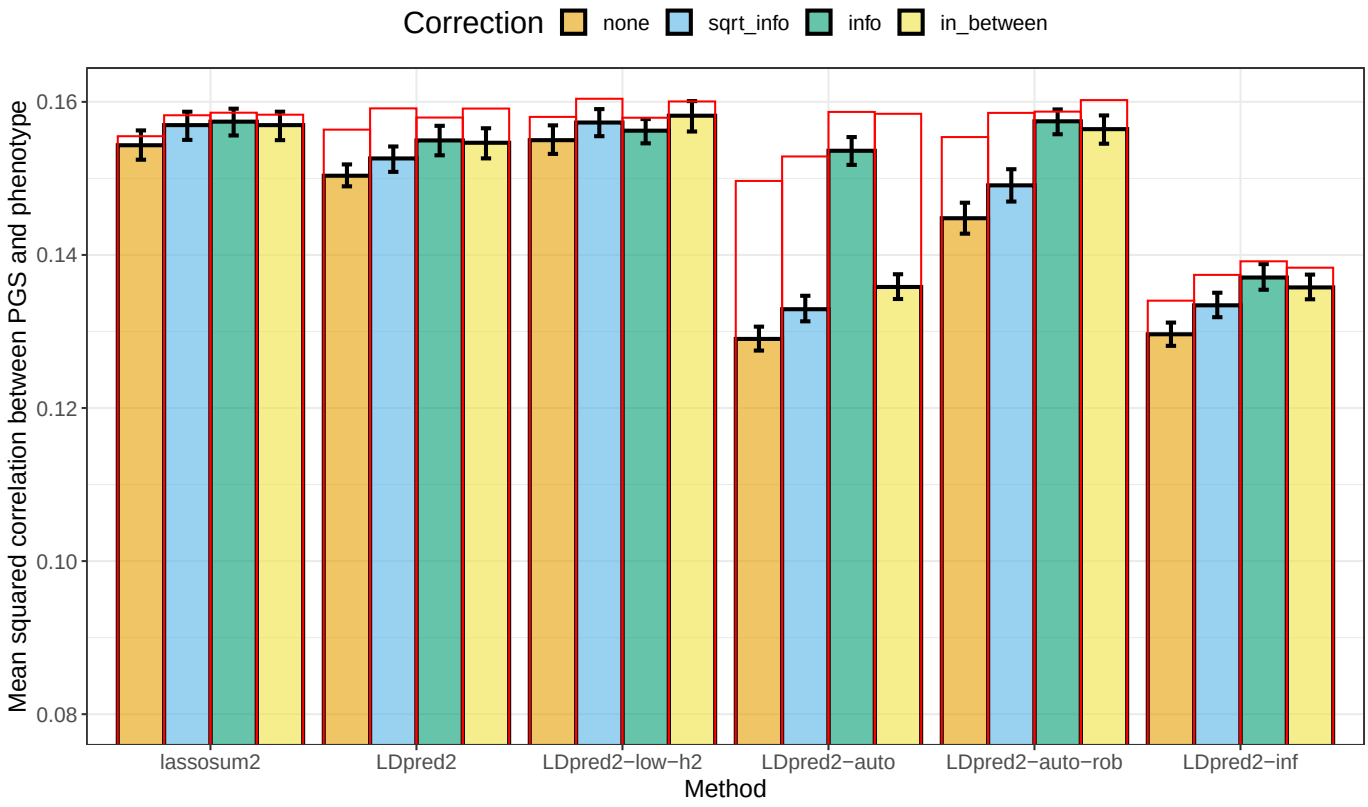

Figure 2: Results of predictive performance for the simulations using GWAS summary statistics from imputed dosage data, averaged over 10 simulations for each scenario. Reported 95% confidence intervals are computed from 10,000 non-parametric bootstrap replicates of the mean. Correction "sqrt_info" corresponds to using $\hat{\gamma}_j^{\text{imp}} \cdot \sqrt{\text{INFO}_j}$ and $\text{se}(\hat{\gamma}_j)^{\text{imp}} \cdot \sqrt{\text{INFO}_j}$. Correction "info" corresponds to using $\hat{\gamma}_j^{\text{imp}} \cdot \text{INFO}_j$ and $N \cdot \text{INFO}_j$. Correction "in_between" corresponds to using $\hat{\gamma}_j^{\text{imp}} \cdot \text{INFO}_j$, $\text{se}(\hat{\gamma}_j)^{\text{imp}} \cdot \sqrt{\text{INFO}_j}$, and $N \cdot \text{INFO}_j$. Red bars correspond to using the LD with independent blocks (Methods).

2.3 Application to breast cancer summary statistics

Breast cancer summary statistics are interesting because they include results from two mega analyses (Michailidou *et al.* 2013, 2015, 2017), which means there is some larger precision in the parameters reported, such as the INFO scores and the sample sizes. Imputation INFO scores for the OncoArray summary statistics (after having restricted to HapMap3 variants) are generally very good (Figure S11) and better than the ones from iCOGS (Figure S10), probably because the chip used included around 200K variants only, compared to more than 500K variants for the OncoArray. For both summary statistics, we first compare the standard

deviations inferred from the reported allele frequencies (i.e. $\sqrt{2f(1-f)}$ where f is the allele frequency, and denoted as sd_af) versus the ones inferred from the GWAS summary statistics (Equation (2), and denoted as sd_ss). When coloring by INFO scores, we see a clear trend with sd_ss being lower than sd_af as INFO decreases; indeed, using sd_ss$/\sqrt{\text{INFO}}$ provides a very good fit for sd_af, except for some variants of chromosome 6 and 8 for the OncoArray summary statistics (Figures 3 and S14). Most of these outlier variants are in regions 25-33 Mbp of chromosome 6 and 8-12 Mbp of chromosome 8 (Figure S12), which are two known long-range LD regions (Price *et al.* 2008). We hypothesize that this is due to using principal components (PCs) as covariates in GWAS that capture LD structure instead of population structure (Privé *et al.* 2020a). To validate this hypothesis, we simulate a phenotype using HapMap3 variants of chromosome 6 for 10,000 individuals from the UK Biobank, then we run GWAS with or without PC19 as covariate. PC19 from the UK Biobank was previously reported to capture LD structure in region 70-91 Mbp of chromsome 6 (Privé *et al.* 2020a). In these simulations, the same bias as in figure 3B is observed for the variants in this region (Figure S13), confirming our hypothesis.

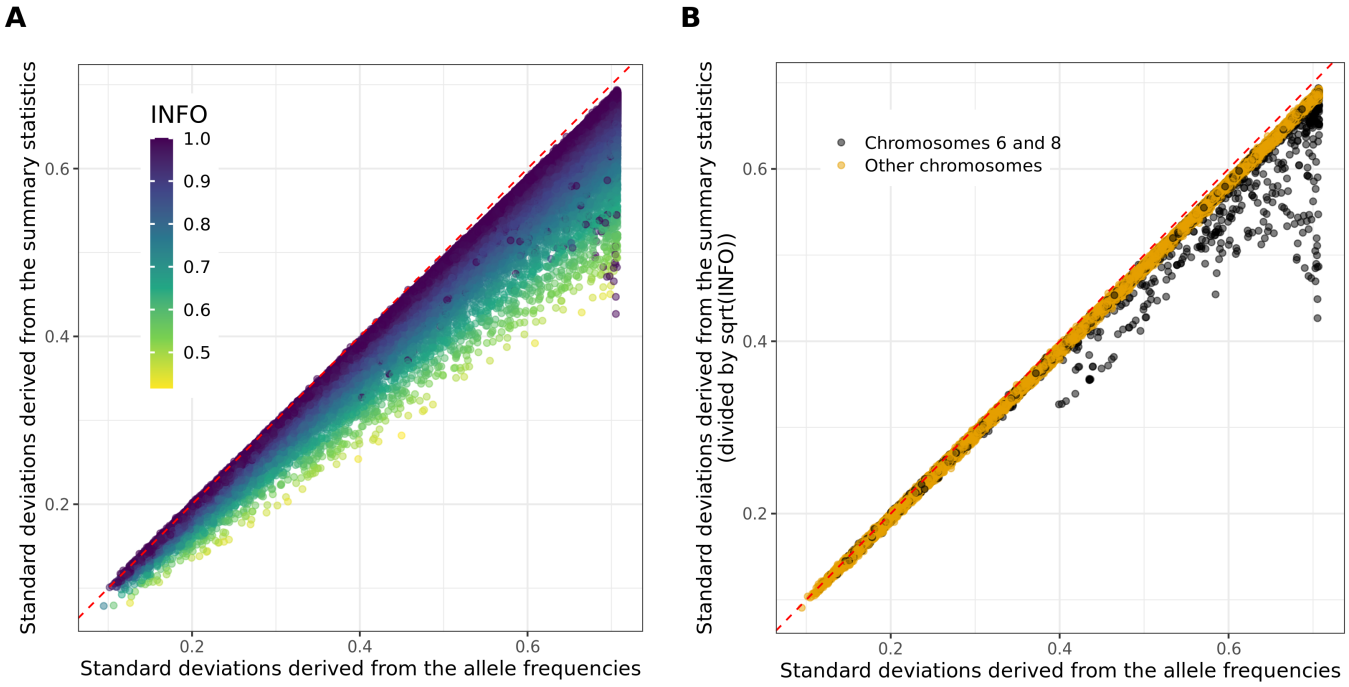

Figure 3: Standard deviations inferred from the OncoArray breast cancer GWAS summary statistics (Equation 2) versus the ones inferred from the reported GWAS allele frequencies ($\sqrt{2f(1-f)}$). Only 100,000 HapMap3 variants are represented, at random.

Therefore, providing an accurate imputation INFO score is important for two reasons. First, it allows for correcting for a reduced standard deviation when using imputed data in the QC step we propose, in order to better uncover problems with the summary statistics. Second, using one of the proposed corrections may lead to an improved prediction when deriving polygenic scores. We apply this correction to the two breast cancer summary statistics. We first compare the standard QC proposed in Privé *et al.* (2020b) (which ends up filtering on MAF here, which we call "qc1"). We then also filter out the two long-range LD regions of

chromosome 6 and 8 for the OncoArray summary statistics and remove around 500 variants when filtering on differences of MAFs between summary statistics and the validation dataset ("qc2"). As for the correction for the INFO scores, we use the first correction, "sqrt_info", which is simple because it is equivalent to post-processing PGS effects by multiplying them by $\sqrt{\text{INFO}}$. Although results for both QC used are very similar, correcting for the INFO score slightly improves prediction when deriving polygenic scores based on iCOGS summary statistics (Figure 4). All other improvements introduced before have little or no effect here, probably because misspecifications are much smaller than in the simulations.

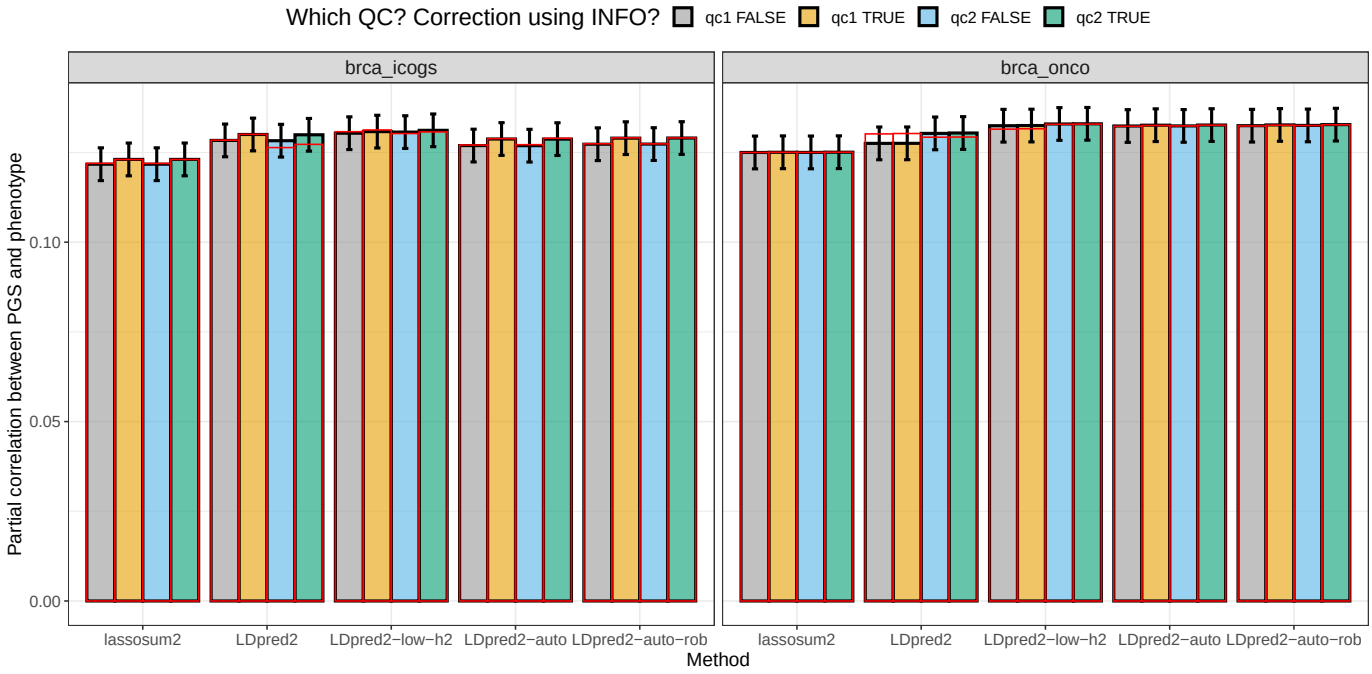

Figure 4: Raw partial correlations (for *all* models) for predicting breast cancer in the UK Biobank when using either the OncoArray or the iCOGS summary statistics. These are computed using function `pcor` of R package bigstatsr where 95% confidence intervals are obtained through Fisher's Z-transformation.

2.4 Other phenotypes and larger sets of variants

We use external summary statistics for which INFO scores are reported (Table 1); they all have a very high mean INFO score (larger than 0.94), except for BRCA-iCOGS (0.841) and T1D-affy (0.885). QC plots comparing standard deviations usually show little deviation from $x = y$ (after the INFO score correction), except for coronary artery disease (CAD) summary statistics (Figures S14-S19). We show previous results as figure 4 for other phenotypes in figures S20-S23; most changes introduced before have little or no impact on the predictions, except for the second quality control performed for CAD when using LDpred2-auto (Figure S20). We then introduce two new sets of variants as possible replacement for HapMap3 variants (Methods). These sets include more variants (more than 2M), therefore possibly of lower quality on average. We present similar results with these two new sets of variants in figures S24-S28. Using the "maxtag" set generally provides larger predictive performance than using HapMap3 variants, especially when predicting prostate cancer (Figure S27).

3 Discussion

We have investigated misspecifications in GWAS summary statistics, focusing on the impact of sample size heterogeneity and imputation quality. Previously, we proposed a quality control (QC) based on comparing standard deviations inferred from GWAS summary statistics with the ones computed from a reference panel (Privé *et al.* 2020b). Here we show that we can refine this quality control by deriving the latter directly from the reported allele frequencies in the GWAS summary statistics, and by correcting the former using the GWAS imputation INFO scores (e.g. see figure 3). Using this refined QC, we are able to identify a potential issue with how principal components were derived in a set of breast cancer summary statistics. Fortunately, this has practically no effect on the predictive performance of the derived polygenic scores. Additional QC can also be performed, e.g. comparing reported GWAS allele frequencies with the ones from the LD reference panel, e.g. to detect genotyping or allelic errors. We do perform this additional QC as part of "qc2" here, and also when designing the large set of variants from the UK Biobank (Methods). One can also run other QC tools such as DENTIST (Chen *et al.* 2021).

Note that, in this study, we use summary statistics that include extended information (e.g. INFO scores and allele frequencies), yet most GWAS summary statistics do not. We acknowledge that, in the case of a meta-analysis from multiple studies, providing a single INFO score per variant may not be possible; would a weighted averaged INFO score work? Nevertheless, this quality control could be performed within each study before pooling results, to make sure that summary statistics have the best possible quality for follow-up analyses such as deriving polygenic scores. Another information, the effective sample size per variant, is often missing from GWAS summary statistics. Sometimes, it can even be challenging to recover the total effective sample size from large meta-analyses. We recall that if some studies of a meta-analysis have an imbalanced number of cases and controls, the global effective sample size should not be computed from the total numbers of cases and controls overall, but instead from the sum of the effective sample sizes of each study. Indeed, take the extreme example of meta-analyzing two studies, one with 1000 cases and 0 controls, and another one with 0 cases and 1000 controls, then the effective sample size of the meta-analysis is 0, not 2000. Fortunately, an overestimated sample size can be detected from the QC plot we propose, where the slope is then less than 1 for case-control studies using logistic regression; otherwise the standard deviation of the phenotype is also needed (Equation 1), but can be estimated (Privé *et al.* 2020b).

We have assessed the impact of these misspecifications in GWAS summary statistics on the predictive performance of some polygenic score methods. Using both the Bayesian LDpred2 models (Privé *et al.* 2020b) and our reimplementation of the frequentist lassosum model (Mak *et al.* 2017) for deriving polygenic scores, we have introduced and investigated some changes to possibly make these models more robust to misspecifications. Overall, these changes have provided large improvements of predictive performance in simulations with large misspecifications. However, they have almost no effect when using the real GWAS summary statistics we chose, except for breast cancer and coronary artery disease summary statistics. Although these results are somewhat unfortunate, they are reassuring because it means that the summary statistics we use in this study are usually of good quality for follow-up analyses such as deriving polygenic scores.

4 Materials and Methods

4.1 Data for simulations

We use the UK Biobank imputed (BGEN) data (Bycroft *et al.* 2018). We restrict individuals to the ones used for computing the principal components (PCs) in the UK Biobank (Field 22020). These individuals are unrelated and have passed some quality control including removing samples with a missing rate on autosomes larger than 0.02, having a mismatch between inferred sex and self-reported sex, and outliers based on heterozygosity (more details can be found in section S3 of Bycroft *et al.* (2018)). To get a set of genetically homogeneous individuals, we compute a robust Mahalanobis distance based on the first 16 PCs and further restrict individuals to those within a log-distance of 5 (Privé *et al.* 2020a). This results in 362,307 individuals. We sample 300,000 individuals to form a training set (e.g. to run GWAS), 10,000 individuals to form a validation set (to tune hyper-parameters), and use the remaining 52,307 individuals to form a test set (to evaluate final predictive models).

Among genetic variants on chromosome 22 and with a minor allele frequency larger than 0.01 and an imputation INFO score larger than 0.4 (as reported by the UK Biobank), we sample 40,000 of them according to the inverse of the INFO score density so that they have varying levels of imputation accuracy (Figure S4). We read the UK Biobank data into two different datasets using function `snp_readBGEN` from R package bigsnpr (Privé *et al.* 2018), one by reading the BGEN data at random according to imputation probabilities, and another one reading it as dosages (i.e. expected values according to imputation probabilities). The first dataset is used as what could be the real genotype calls and the second dataset as what would be its imputed version; this design technique was used in Privé *et al.* (2019).

4.2 Data for real analyses

We also use the UK Biobank data, and use the same individuals as described in the previous section. We sample 10,000 individuals to form a validation set and use the remaining 352,307 individuals as test set. We restrict to the genetic variants to the 1,054,315 HapMap3 variants used in the LD reference provided in Privé *et al.* (2020b). We also try two new sets of variants (see next section).

To define phenotypes in the UK Biobank, we first map ICD10 and ICD9 codes (UKBB fields 40001, 40002, 40006, 40013, 41202, 41270 and 41271) to phecodes using R package PheWAS (Carroll *et al.* 2014; Wu *et al.* 2019). We use published GWAS summary statistics listed in table 1 to derive polygenic scores.

4.3 Two new sets of variants

We also design two larger sets of imputed variants to compare against using only HapMap3 variants for prediction. Following Privé *et al.* (2021), we first restrict to UKBB variants with MAF > 0.01 and INFO > 0.3. We then compile frequencies and imputation INFO scores from other datasets, iPSYCH and summary statistics for breast cancer, prostate cancer, coronary artery disease, type-1 diabetes and depression (Bybjerg-Grauholm *et al.* 2020; Michailidou *et al.* 2017; Schumacher *et al.* 2018; Nikpay *et al.* 2015; Censin *et al.* 2017; Wray *et al.* 2018). We restrict to variants with a mean INFO > 0.3 in these other datasets, and

Trait	GWAS citation	Effective GWAS sample size	# GWAS variants	# matched variants with INFO > 0.4	Mean INFO
Breast cancer (BRCA) [iCOGS]	Michailidou *et al.* (2017)	87,037	11,792,542	1,051,242	0.841
Breast cancer (BRCA) [OncoArray]	Michailidou *et al.* (2017)	104,442	11,792,542	1,054,233	0.968
Type 1 diabetes (T1D) [Affymetrix]	Censin *et al.* (2017)	5516	8,996,866	934,712	0.885
Type 1 diabetes (T1D) [Illumina]	Censin *et al.* (2017)	7982	8,996,866	949,334	0.942
Prostate cancer (PRCA)	Schumacher *et al.* (2018)	135,316	20,370,946	818,400	0.969
Depression (MDD) [without UKBB]	Wray *et al.* (2018)	110,464	9,874,289	1,049,455	0.968
Coronary artery disease (CAD)	Nikpay *et al.* (2015)	129,014	9,455,778	1,052,200	0.982

Table 1: Summary of external GWAS summary statistics used. PRCA summary statistics have many variants with a missing INFO score, which we discard.

also compute the median frequency per variant. To exclude potential mismappings in the genotyped data (Kunert-Graf *et al.* 2020) that might have propagated to the imputed data, we compare median frequencies in the external data to the ones in the UK Biobank. As we expect these potential errors to be localized around errors in the genotype data, we apply a moving-average smoothing on the frequency differences to increase power to detect these errors and also reduce false positives. We define the threshold (of 0.03) on these smoothed differences based on visual inspection of their histogram. This results in an initial set of 9,394,361 variants.

We then define the two sets from this large set of variants. One is based on clumping, using a threshold $r^2 = 0.9$ over a radius of 100 Kbp and prioritizing HapMap3 variants and larger INFO scores. This results in a set "clump" of 2,465,478 variants, among which there are 554,655 of the initial HapMap3 variants. For the second set, we aim at maximizing the tagging of all the initial 9,394,361 variants, i.e. $\sum_{i \in \text{all}} \max_{j \in \text{set}} |R_{i,j}|$, where $R_{i,j}$ is the correlation between variants i and j (inspired from the alternative sensitivity of Agier *et al.* (2016)). We design a greedy algorithm that first selects all HapMap3 variants, then adds one variant at a time, the one that maximizes the addition to this sum, until no variant can add more than 0.2. This results in a set "maxtag" of 2,029,086 variants.

4.4 GWAS sample size imputation

In this paper, we extensively use the following formula

$$\text{sd}(G_j) \approx \frac{\text{sd}(y)}{\sqrt{n_j \, \text{se}(\hat{\gamma}_j)^2 + \hat{\gamma}_j^2}}, \tag{1}$$

where $\hat{\gamma}_j$ is the marginal (GWAS) effect of variant j, n_j is the GWAS sample size associated with variant j, y is the vector of phenotypes and G_j is the vector of genotypes for variant j. This formula is used in LDpred2 Privé *et al.* (2020b, 2021). Note that, for a binary trait for which logistic regression is used, we have instead

$$\text{sd}(G_j) \approx \frac{2}{\sqrt{n_j^{\text{eff}} \, \text{se}(\hat{\gamma}_j)^2 + \hat{\gamma}_j^2}}, \tag{2}$$

where $n_j^{\text{eff}} = \frac{4}{1/n_j^{\text{cases}} + 1/n_j^{\text{controls}}}$.

We can then impute n_j from equation (1) using

$$n_j \approx \frac{\text{var}(y)/\text{var}(G_j) - \hat{\gamma}_j^2}{\text{se}(\hat{\gamma}_j)^2} \,, \tag{3}$$

and impute n_j^{eff} from equation (2) using

$$n_j^{\text{eff}} \approx \frac{4/\text{var}(G_j) - \hat{\gamma}_j^2}{\text{se}(\hat{\gamma}_j)^2} \,. \tag{4}$$

In practice, we also bound this estimate to be between $0.5 \cdot N$ and $1.1 \cdot N$, where N is the global sample size.

4.5 New implementation of lassosum in bigsnpr

Instead of using a regularized version of the correlation matrix R parameterized by s, $R_s = (1-s)R + sI$ (where $0 < s \leq 1$), we use $R_\delta = R + \delta I$ (where $\delta > 0$), which makes it clearer that lassosum is also using L2-regularization (therefore elastic-net). Then, from Mak *et al.* (2017), the solution from lassosum can be obtained by iteratively updating

$$\beta_j^{(t)} = \begin{cases} \text{sign}\left(u_j^{(t)}\right)\left(\left|u_j^{(t)}\right| - \lambda\right) / \left(\widetilde{X}_j^T \widetilde{X}_j + \delta\right) & \text{if } \left|u_j^{(t)}\right| > \lambda \,, \\ 0 & \text{otherwise.} \end{cases}$$

where

$$u_j^{(t)} = r_j - \widetilde{X}_j^T \left(\widetilde{X}\beta^{(t-1)} - \widetilde{X}_j \beta_j^{(t-1)}\right) \,.$$

Following the notations from Privé *et al.* (2020b) and denoting $\widetilde{X} = \frac{1}{\sqrt{n-1}} C_n G S^{-1}$, where G is the genotype matrix, C_n is the centering matrix and S is the diagonal matrix of standard deviations of the columns of G. Then $\widetilde{X}_j^T \widetilde{X} = R_{j,\cdot} = R_{\cdot,j}^T$ and $\widetilde{X}_j^T \widetilde{X}_j = 1$. Moreover, using the notations from Privé *et al.* (2020b), $u_j^{(t)} = \widehat{\beta}_j - R_{\cdot,j}^T \beta^{(t-1)} + \beta_j^{(t-1)}$, where $r_j = \widehat{\beta}_j = \dfrac{\widehat{\gamma}_j}{\sqrt{n_j \, \text{se}(\widehat{\gamma}_j)^2 + \widehat{\gamma}_j^2}}$ and $\widehat{\gamma}_j$ is the GWAS effect of variant j and n is the GWAS sample size (Mak *et al.* 2017; Privé *et al.* 2021). Then the most time-consuming part is computing $R_{\cdot,j}^T \beta^{(t-1)}$. To make this faster, instead of computing $R_{\cdot,j}^T \beta^{(t-1)}$ at each iteration (j and t), we can start with an initial vector of 0s only (for all j) since $\beta^{(0)} \equiv 0$, and then updating this vector when $\beta_j^{(t)} \neq \beta_j^{(t-1)}$ only. Note that only positions k for which $R_{k,j} \neq 0$ must be updated in this vector $R_{\cdot,j}^T \beta^{(t-1)}$.

In this new implementation of the lassosum model, the input parameters are the correlation matrix R, the GWAS summary statistics ($\widehat{\gamma}_j$, $\text{se}(\widehat{\gamma}_j)$ and n_j), and the two hyper-parameters λ and δ. Therefore, except for the hyper-parameters, this is the exact same input as for LDpred2 (Privé *et al.* 2020b). We try $\delta \in \{0.001, 0.005, 0.02, 0.1, 0.6, 3\}$ by default in lassosum2, instead of $s \in \{0.2, 0.5, 0.8, 1.0\}$ in lassosum. For λ, the default in lassosum uses a sequence of 20 values equally spaced on a log scale between 0.1 and 0.001. We use instead a sequence between λ_0 and $\lambda_0/100$ by default in lassosum2, where $\lambda_0 = \max_j \left|\widehat{\beta}_j\right|$ is the minimum λ for which no variable enters the model because the L1-regularization is too strong. Note that we do not provide an "auto" version using pseudo-validation (as in Mak *et al.* (2017)) as we have not found it to

be very robust (Figure S29). Also note that, as in LDpred2, we run lassosum2 genome-wide using a sparse correlation matrix which assumes that variants further away than 3 cM are not correlated, and therefore we do not require splitting the genome into independent LD blocks anymore (as done in lassosum).

4.6 New LD reference

We make three changes to the LD reference. First, when using imputed data, we multiply the correlation between variants j and k by $\sqrt{\text{INFO}_j \cdot \text{INFO}_k}$ (for $j \neq k$) since it approximates well the correlation from non-imputed data (Figure S30). Second, we also define nearly independent LD blocks using the optimal algorithm developed in Privé (2021). For different numbers of blocks and maximum number of variants in each block, we use the split with the minimum cost within the ones reducing the original number of non-zero values to less than 60% (70% for chromosome 6). Having a correlation matrix with independent blocks prevents the small errors in the algorithm (e.g. the Gibbs sampler in LDpred2) from propagating to too many variants. It also makes running LDpred2 (and lassosum2) faster, taking about 60% of the initial time (since only 60% of the initial non-zero values of the correlation matrix are kept). Finally, we have developed a new "compact" format for the SFBMs (sparse matrices on disk). Instead of using something similar to the standard "compressed sparse column" format which stores all $\{i,\ x(i,j)\}$ for a given column j, we only store the first index i_0 and all the contiguous values $\{x(i_0,j),\ x(i_0+1,j),\ \dots\}$ up to the last non-zero value for this column j. This makes this format about twice as efficient for both LDpred2 and lassosum2.

4.7 LDpred2-low-h2 and LDpred2-auto-rob

Here we introduce the small changes made to LDpred2 (-grid and -auto) in order to make them more robust. First, LDpred2-low-h2 simply consists in running LDpred2-grid by testing h^2 within $\{0.3, 0.7, 1, 1.4\} \cdot h^2_{\text{LDSC}}$ (note the added 0.3 compared to Privé *et al.* (2020b)), where h^2_{LDSC} is the heritability estimate from LD score regression. Indeed, we show in simulations here that using lower values for h^2 may provide higher predictive performance in the case of misspecifications (thanks to more shrinkage of the effects). In simulations, because of the large misspecifications, we use a larger grid over $\{0.01, 0.1, 0.3, 0.7, 1, 1.4\} \cdot h^2_{\text{LDSC}}$.

For LDpred2-auto, we introduce two new parameters. The first one, `shrink_corr`, allows for shrinking off-diagonal elements of the correlation matrix. This is similar to parameter 's' in lassosum, and act as a regularization. We use a value of 0.9 in simulations and 0.95 in real data when running "LDpred2-auto-rob" (and the default value of 1, without any effect, when running "LDpred2-auto"). The second new parameter, `allow_jump_sign`, controls whether variants can change sign over two consecutive iterations of the Gibbs sampler. When setting this parameter to false (in the method we name "LDpred2-auto-rob" here), this forces the effects to go through 0 before changing sign. This is useful to prevent instability (oscillation and ultimately divergence) of the Gibbs sampler under large misspecifications, and is also useful for accelerating convergence of chains with a large initial value for p, the proportion of causal variants.

Code availability

All code used for this paper is available at `https://github.com/privefl/paper-misspec/tree/master/code`. The latest version of R package bigsnpr can be installed from GitHub. Two tutorials on running LDpred2 and lassosum2 using R package bigsnpr are available at `https://privefl.github.io/bigsnpr/articles/LDpred2.html` and `https://privefl.github.io/bigsnpr-extdoc/polygenic-scores-pgs.html`. We have extensively used R packages bigstatsr and bigsnpr (Privé *et al.* 2018) for analyzing large genetic data, packages from the future framework (Bengtsson 2021) for easy scheduling and parallelization of analyses on the HPC cluster, and packages from the tidyverse suite (Wickham *et al.* 2019) for shaping and visualizing results.

Acknowledgements

Authors thank Timothy Shin Heng Mak, Shing Wan Choi, and Matthew Stephens for helpful discussions. Authors also thank GenomeDK and Aarhus University for providing computational resources and support that contributed to these research results. This research has been conducted using the UK Biobank Resource under Application Number 58024.

Funding

F.P. and B.J.V. are supported by the Danish National Research Foundation (Niels Bohr Professorship to Prof. John McGrath). B.J.V. is also supported by a Lundbeck Foundation Fellowship (R335-2019-2339).

Declaration of Interests

The authors declare no competing interests.

References

Agier, L., Portengen, L., Chadeau-Hyam, M., Basagaña, X., Giorgis-Allemand, L., Siroux, V., Robinson, O., Vlaanderen, J., González, J. R., Nieuwenhuijsen, M. J., *et al.* (2016). A systematic comparison of linear regression–based statistical methods to assess exposome-health associations. *Environmental Health Perspectives*, **124**(12), 1848–1856.

Bengtsson, H. (2021). A Unifying Framework for Parallel and Distributed Processing in R using Futures. *The R Journal*.

Bybjerg-Grauholm, J., Pedersen, C. B., Baekvad-Hansen, M., Pedersen, M. G., Adamsen, D., Hansen, C. S., Agerbo, E., Grove, J., Als, T. D., Schork, A. J., *et al.* (2020). The ipsych2015 case-cohort sample: updated directions for unravelling genetic and environmental architectures of severe mental disorders. *medRxiv*.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.* (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**(7726), 203–209.

Carroll, R. J., Bastarache, L., and Denny, J. C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*, **30**(16), 2375–2376.

Censin, J., Nowak, C., Cooper, N., Bergsten, P., Todd, J. A., and Fall, T. (2017). Childhood adiposity and risk of type 1 diabetes: A mendelian randomization study. *PLoS Medicine*, **14**(8), e1002362.

Chen, W., Wu, Y., Zheng, Z., Qi, T., Visscher, P. M., Zhu, Z., and Yang, J. (2021). Improved analyses of gwas summary statistics by reducing data heterogeneity and errors. *bioRxiv*, pages 2020–07.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, **10**(1), 1776.

Kulm, S., Marderstein, A., Mezey, J., and Elemento, O. (2021). A systematic framework for assessing the clinical impact of polygenic risk scores. *medRxiv*, pages 2020–04.

Kunert-Graf, J. M., Sakhanenko, N. M., and Galas, D. J. (2020). Allele frequency mismatches and apparent mismappings in UK Biobank SNP data. *bioRxiv*.

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., *et al.* (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nature Communications*, **10**(1), 1–11.

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, **41**(6), 469–480.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, **11**(7), 499–511.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., Schmidt, M. K., Chang-Claude, J., Bojesen, S. E., Bolla, M. K., *et al.* (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genetics*, **45**(4), 353–361.

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., Maranian, M. J., Bolla, M. K., Wang, Q., Shah, M., *et al.* (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature Genetics*, **47**(4), 373–380.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., *et al.* (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, **551**(7678), 92–94.

Miller, J. W. and Dunson, D. B. (2018). Robust bayesian inference via coarsening. *Journal of the American Statistical Association*.

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., Hopewell, J. C., *et al.* (2015). A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, **47**(10), 1121.

Pain, O., Glanville, K. P., Hagenaars, S. P., Selzam, S., Fürtjes, A. E., Gaspar, H. A., Coleman, J. R., Rimfeld, K., Breen, G., Plomin, R., *et al.* (2021). Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genetics*, **17**(5), e1009021.

Palmer, C. and Pe'er, I. (2016). Bias Characterization in Probabilistic Genotype Data and Improved Signal Detection with Multiple Imputation. *PLoS Genetics*, **12**(6), e1006091.

Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K. D., *et al.* (2008). Long-range LD can confound genome scans in admixed populations. *The American Journal of Human Genetics*, **83**(1), 132–135.

Privé, F. (2021). Optimal linkage disequilibrium splitting. *Bioinformatics (in press)*.

Privé, F., Aschard, H., Ziyatdinov, A., and Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*, **34**(16), 2781–2787.

Privé, F., Vilhjálmsson, B. J., Aschard, H., and Blum, M. G. B. (2019). Making the most of clumping and thresholding for polygenic scores. *The American Journal of Human Genetics*, **105**(6), 1213–1221.

Privé, F., Luu, K., Blum, M. G., McGrath, J. J., and Vilhjálmsson, B. J. (2020a). Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics*, **36**(16), 4449–4457.

Privé, F., Arbel, J., and Vilhjálmsson, B. J. (2020b). LDpred2: better, faster, stronger. *Bioinformatics*, **36**(22-23), 5424–5431.

Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., O'Reilly, P. F., and Vilhjálmsson, B. J. (2021). High-resolution portability of 245 polygenic scores when derived and applied in the same cohort. *medRxiv*.

Schumacher, F. R., Al Olama, A. A., Berndt, S. I., Benlloch, S., Ahmed, M., Saunders, E. J., Dadaev, T., Leongamornlert, D., Anokian, E., Cieza-Borrella, C., *et al.* (2018). Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nature Genetics*, **50**(7), 928.

Walker, S. G. (2013). Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, **143**(10), 1621–1633.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., *et al.* (2019). Welcome to the tidyverse. *Journal of Open Source Software*, **4**(43), 1686.

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J., Agerbo, E., Air, T. M., Andlauer, T. M., *et al.* (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, **50**(5), 668.

Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., Denny, J. C., *et al.* (2019). Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. *JMIR Medical Informatics*, **7**(4), e14325.