Inferring disease architecture and predictive ability with LDpred2-auto

Florian Privé,[1,*] Clara Albiñana,[1] Bogdan Pasaniuc,[2,3,4,5] and Bjarni J. Vilhjálmsson[1,6,7]

[1]National Centre for Register-based Research, Aarhus University, Aarhus, Denmark.
[2]Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA, USA.
[3]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA.
[4]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA, USA.
[5]Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA, USA.
[6]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark.
[7]Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute, Cambridge, MA, USA.
[*]To whom correspondence should be addressed.

Contact: florian.prive.21@gmail.com

Abstract

LDpred2 is a widely used Bayesian method for building polygenic scores (PGS). LDpred2-auto can infer the two parameters from the LDpred model, h^2 and p, so that it does not require an additional validation dataset to choose best-performing parameters. Here, we present a new version of LDpred2-auto, which adds a third parameter α to its model for modeling negative selection. Additional changes are also made to provide better sampling of these parameters. We then validate the inference of these three parameters. LDpred2-auto also provides per-variant probabilities of being causal that are well calibrated, and can therefore be used for fine-mapping purposes. We also derive a new formula to infer the out-of-sample predictive performance r^2 of the resulting PGS directly from the Gibbs sampler of LDpred2-auto. Finally, we extend the set of HapMap3 variants recommended to use with LDpred2 with 37% more variants to improve the coverage of this set, and show that this new set of variants captures 12% more heritability and provides 6% more predictive performance, on average.

Introduction

Most traits and diseases in humans are heritable. What differs is the genetic architecture of each trait that can be parameterized by three key terms: the heritability (i.e. the proportion of phenotypic variation explained by genetics), the polygenicity (i.e. the fraction of genomic variants that have a non-zero impact on the trait), and the causal effect distribution (i.e. how the effect size distribution varies across causal variants). Some phenotypes, such as schizophrenia or height, are highly heritable and highly polygenic (Sullivan *et al.*, 2003; Yang *et al.*, 2010; O'Connor *et al.*, 2019). Causal effects are larger when a trait is more heritable, and smaller when it is more polygenic. The distribution of causal effects relative to their allele frequencies is often investigated through a single parameter, usually called α or S, to model the effect of negative selection on complex traits whereby variants with lower frequencies are expected to have higher causal effect sizes (Speed *et al.*, 2012). Many methods have been developed to estimate the SNP heritability (referred to as h^2 for brevity) and polygenicity (p), either globally for the whole genome or locally for specific regions of the genome, as well as α. These methods include GCTA (h^2, Yang *et al.* (2011)), BOLT-REML (h^2 and p, Loh *et al.* (2015)), LD Score regression (h^2, Bulik-Sullivan *et al.* (2015b)), FINEMAP (per-variant p used for fine-mapping, Benner *et al.* (2016)), HESS (local h^2, Shi *et al.* (2016)), LDAK-SumHer (h^2, Speed and Balding (2019)), S-LD4M (p, O'Connor *et al.* (2019)), GRM-MAF-LD (α, Schoech *et al.* (2019)), SuSiE (per-variant p used for fine-mapping, Wang *et al.* (2020)), SBayesS (h^2, p, and a third parameter S, similar to α, Zeng *et al.* (2021)), and BEAVR (local p, Johnson *et al.* (2021)).

As previously shown by Daetwyler *et al.* (2008), h^2 and p can also be used to determine how well we can predict a phenotype from using genetic variants alone. Such genetic predictors are called polygenic scores (PGS), and are getting closer to being included as part of existing clinical risk models for diseases (Torkamani *et al.*, 2018; Lambert *et al.*, 2019; Kumuthini *et al.*, 2022). LDpred2 is a widely used polygenic score method that can directly build PGS using summary statistics results from genome-wide associations studies (GWAS), making it highly applicable (Privé *et al.*, 2020b; Pain *et al.*, 2021; Kulm *et al.*, 2021). LDpred2 is a Bayesian approach that uses the SNP heritability h^2 and polygenicity p as parameters of its model. In LDpred2-auto, it can directly estimate these parameters from the data, making it applicable even when no validation data is available for tuning these two model hyper-parameters (Privé *et al.*, 2020b).

Here we extend LDpred2-auto to make it a highly reliable method for estimating h^2 (global and local), p (also per-variant probabilities for fine-mapping purposes), and α (by extending its model to also include this third parameter). So, on top of providing competitive PGS, LDpred2-auto can now provide all these estimates of genetic architecture. Moreover, we show how it can now also reliably estimate the predictive ability r^2 of PGS it derives, allowing for directly assessing the usefulness of the derived PGS, without needing an independent test set. An overview of what LDpred2-auto can provide is presented in Figure 1. Finally, we extend the set of HapMap3 variants recommended to use with LDpred2, which enables us to capture around 12% more SNP heritability and achieve around 6% more predictive performance r^2.

We call 'HapMap3+' this new preferred set of 1,444,196 SNPs.

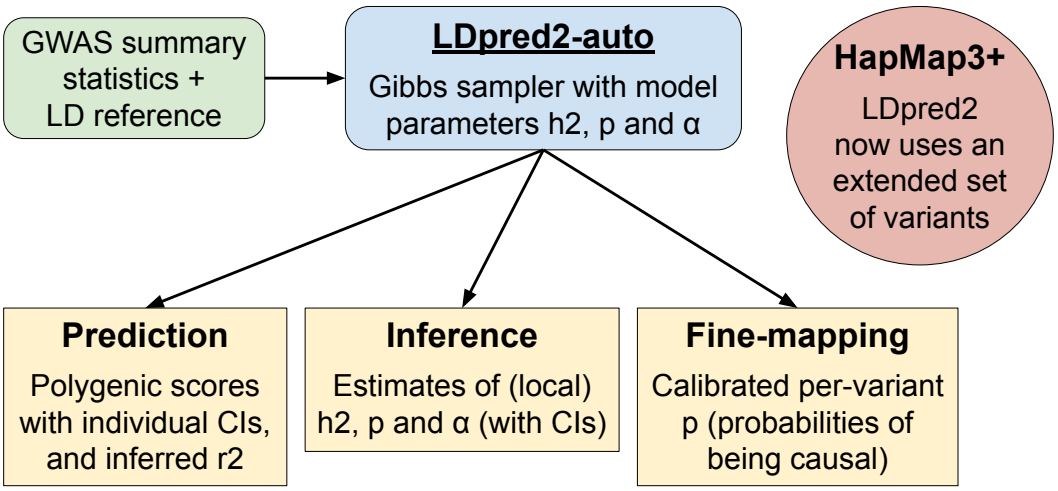

Figure 1: Overview of what LDpred2-auto can now provide. For individual CIs of polygenic scores, please refer to the work of Ding *et al.* (2022a,b). HapMap3+ is the new extended set of 1,444,196 SNPs we introduce here, and now recommend to use for LDpred2 (Methods). CI means "confidence interval" (often called "credible interval" in a Bayesian setting).

Results

Validating the inference with simulations

For simulations, we use the UK Biobank imputed data (Bycroft *et al.*, 2018). We use 356,409 individuals of Northwestern European ancestry and 322,805 SNPs across seven chromosomes (Methods). We first simulate continuous phenotypes using function `snp_simuPheno` from R package bigsnpr (Privé *et al.*, 2018), varying three parameters: the SNP-heritability h^2, the polygenicity p (i.e. the proportion of causal variants), and the parameter α in equation (2) that controls the relationship between minor allele frequencies and expected effect sizes. This function first picks a proportion p of causal variants at random, sample effect sizes γ using the variance component parameterized by α and then scale the effect sizes so that the genetic component $G\gamma$ has a variance h^2, where G is the genotype matrix. Finally, some Gaussian noise is added so that the full phenotype has a variance of 1. Then, a GWAS is run to obtain summary statistics using N individuals (either the 200,000 dedicated to this, or a subset of 20,000) using a simple linear regression implemented in `big_univLinReg` from R package bigstatsr (Privé *et al.*, 2018). Finally, we run the new LDpred2-auto model (Methods) with and without the option `allow_jump_sign`, which was proposed in Privé *et al.* (2022a) for robustness (when disabled, in which case it prevents effect sizes from changing sign without going through 0 first). LDpred2-auto is run with 50 Gibbs sampler chains with different starting values for p (from 0.0001 to 0.2, equally spaced on a log scale). Then only chains that provide top correlations $R\beta$ are kept.

First, LDpred2-auto generally reliably infers the three parameters from its model, i.e. the SNP heritability h^2, polygenicity p, and α (Figures 2, S1 and S2). Compared to LD Score regression, heritability estimates are as precise when power is low, and much more precise when power is large, especially for small polygenicity values (Figure 2). When power is low, the polygenicity can be overestimated when the true value is very small (e.g. $p = 0.0005$), and underestimated when the polygenicity is large (e.g. $p = 0.1$, Figure S1). The α estimate can become unreliable when power is too low, which can be detected by a small number of chains kept from LDpred2-auto (Figure S2). Then, LDpred2-auto can also infer per-variant probabilities of being causal and local per-block heritability estimates, which are well calibrated (Figures S3 and S4). We recall that calibrated per-variant probabilities of being causal can be used for fine-mapping purposes (Wang *et al.*, 2020). Finally, LDpred2-auto can also be used to reliably infer the predictive performance r^2 of its resulting polygenic score, directly from within the Gibbs sampler, even when power is low (Figure S5).

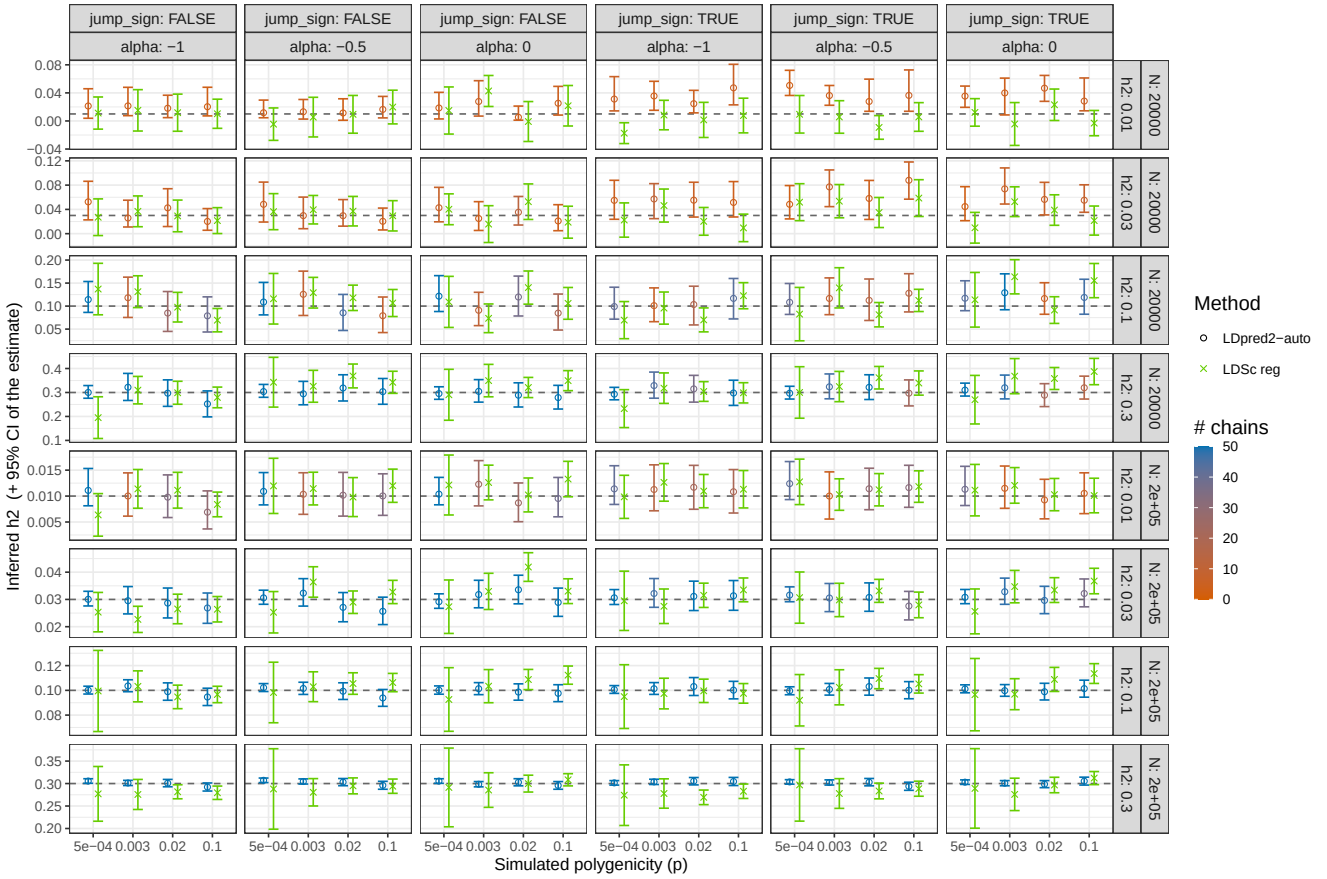

Figure 2: Inferred SNP heritability h^2 in simulations with continuous outcomes. Horizontal dashed lines represent the true simulated values. The 95% confidence interval for the LDpred2-auto estimate is obtained from the 2.5% and 97.5% quantiles of all the h^2 estimates from the iterations (after burn-in) of the chains kept. The 95% confidence interval for the LD Score regression estimate is obtained from ± 1.96 times its standard error. Note that the recommended option is to use `allow_jump_sign = FALSE` (Privé *et al.*, 2022a).

We then run simulations with binary outcomes where the simulated continuous liabilities are trans-

formed to binary outcomes using a threshold corresponding to the prevalence. Results are very similar as with the continuous phenotypes above (Figures S6–S9), and are similar whether we use either a linear regression GWAS and the total sample size N, or a logistic regression GWAS and the effective sample size (i.e. $N_{\text{eff}} = 4/(1/N_{\text{case}} + 1/N_{\text{control}})$). The main difference is that the h^2 and r^2 estimates need to be transformed to the liability scale (Lee *et al.*, 2011), where $K_{\text{GWAS}} = 0.5$ should be used when using N_{eff} in LDpred2-auto or LD Score regression (Grotzinger *et al.*, 2022).

More heritability and predictive accuracy with new set of variants

We use the same 356,409 unrelated individuals of Northwestern European ancestry as in the simulations. To form the test set, we randomly select 50,000 of these, while the other 306,409 are used to run a GWAS using linear regression (with function `big_univLinReg` from R package bigstatsr) for each of all 248 phenotypes and using 8 covariates (Methods). Then the new implementation of LDpred2-auto is used (Methods).

We investigate using the extended set of HapMap3 variants proposed here, HapMap3+, which includes ~37% more variants on top of HapMap3 variants recommended to use for LDpred2 (i.e. 1,054,330 + 389,866 variants), to improve the coverage of this set (Methods). As expected, compared to HapMap3, higher h^2 (average increase of 12.3% [95% CI: 10.8, 13.7]) and lower p (decrease of 11.5% [10.7, 12.3]) estimates are obtained with this extended set HapMap3+ (Figure 3). This is consistent with higher predictive performance r^2 in the test set (increase of 6.1% [4.1, 8.2]). In particular, a much larger h^2 estimate is obtained for lipoprotein(a) concentration (0.508 [0.502, 0.514] instead of 0.324 [0.320, 0.329]), which is also reflected in a larger predictive performance (r^2 in the test set of 0.516 [0.508, 0.524] instead of 0.344 [0.335, 0.353]). Interestingly, when using this extended set of HapMap3 variants, more chains are kept on average, which is a sign of better convergence of the models (Figure S10). However, running LDpred2 with this extended set of variants takes around 50% more time; yet, we recall that LDpred2 has been made much faster in Privé *et al.* (2022a), and now runs in less than one hour for 50 chains parallelized over 13 cores (Figure S11), instead of 4–12 hours before.

Genetic architectures of 248 phenotypes from the UK Biobank

Here we use the same individuals as in the previous section. We first use the set of 1,054,330 HapMap3 variants recommended to use for LDpred2 (Privé *et al.*, 2020b). Consistent with simulations, inferred SNP heritability h^2 estimates from LDpred2-auto closely match with those from LD Score regression, while generally being more precise, especially for phenotypes with a smaller polygenicity (Figure S12). Note that these h^2 estimates (and later the r^2 estimates) have not been transformed to the liability scale (i.e. are on the observed scale). Most phenotypes have an estimated polygenicity p between 0.001 and 0.04; these have therefore a very polygenic architecture, but not an infinitesimal one (Figure S13). Most phenotypes have an estimated α between -1.0 and -0.4 with a mode at -0.65 (Figure S14), which is

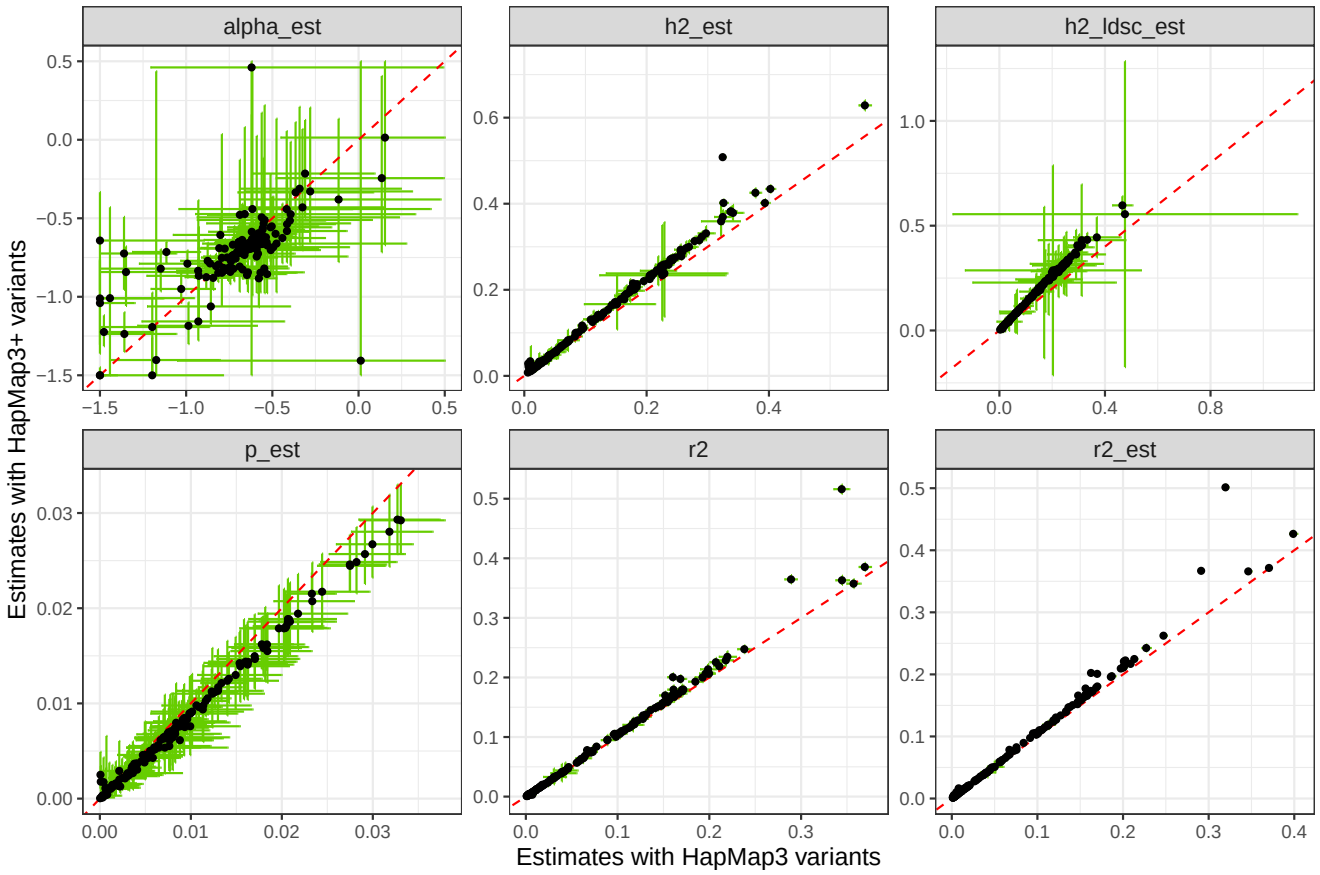

Figure 3: LDpred2-auto estimates for UKBB phenotypes using either the HapMap3 or HapMap3+ sets of variants. Only 154 phenotypes with more than 25 chains kept when using the HapMap3 variants are represented here. Red dashed lines represent the 1:1 line. The 95% confidence interval for the LDpred2-auto estimate (in green) is obtained from the 2.5% and 97.5% quantiles of all the estimates from the iterations (after burn-in) of the chains kept. The 95% confidence interval for r^2 in the test set is obtained from bootstrap.

consistent with widespread negative selection. As for the inferred predictive performance r^2, they are highly consistent with the ones derived from the test set; only for standing height are they overestimated (Figure 4). Heritability estimates for height are probably overestimated as well since we use similar formulas for estimating h^2 and r^2 (Methods), and because the SNP heritability estimate h^2 for standing height is higher than values reported in the literature (also see Section "Application to height").

Then, to investigate whether estimates from LDpred2-auto are robust to misspecifications, we test using two alternative LD references (Methods). Using a smaller number of individuals for computing the LD results in a slightly overestimated p and h^2 (and r^2), while the α estimate remains consistent, and the predictive performance in the test set remains mostly similar, except for three phenotypes for which none of the LDpred2-auto chains is usable (Figure S15). When using an LD reference from an alternative population (South Europe instead of North-West Europe), p, h^2, and r^2 are slightly overestimated as well, and a few phenotypes have lower predictive performance while there are four phenotypes for which none

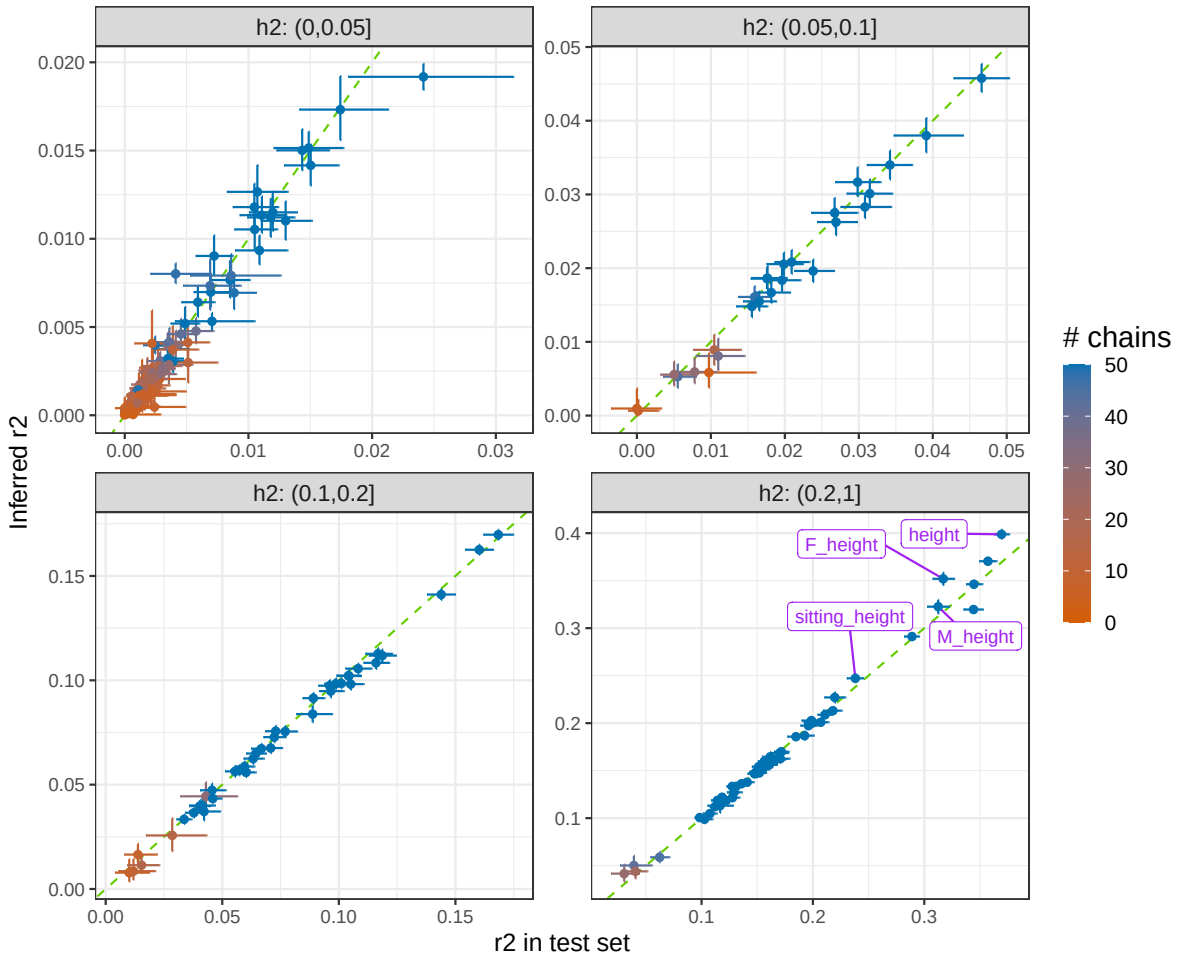

Figure 4: Inferred predictive performance r^2 from the Gibbs sampler of LDpred2-auto versus the ones obtained in the test set, for all 248 phenotypes defined from the UK Biobank. These are stratified by the polygenicity estimated from LDpred2-auto. Green dashed lines represent the 1:1 line. The 95% confidence interval for the LDpred2-auto estimate is obtained from the 2.5% and 97.5% quantiles of all the r^2 estimates from the iterations (after burn-in) of the chains kept. The 95% confidence interval for r^2 in the test set is obtained from bootstrap. "F_height" and "M_height" use females and males only, respectively.

of the LDpred2-auto chains is usable (distinct from the previous three, Figure S16).

Then, we investigate the effect of disabling "allow_jump_sign" (for extra robustness as shown in Privé *et al.* (2022a)) on the estimates from LDpred2-auto. Note that we now use the HapMap3+ set of variants here. Consistent with simulations, p estimates from LDpred2-auto are conservatively lower than when allowing effects to "jump" sign (i.e. normal sampling, Figure S17). h^2 estimates can also be slightly lower, while α estimates are broadly consistent. As for predictive performance r^2 (on the test set), they are similar, suggesting there is no problem of robustness here and normal sampling can be used (Figure S17).

Finally, we investigate different transformations to apply to some continuous phenotypes used here. Indeed, 49 of the phenotypes used here seem log-normally distributed or heavy-tailed (when visualizing

their histogram); we therefore log-transform them. However, we do investigate alternative transformations here to decide which one should be preferred and to check how this impacts the inference from LDpred2-auto. Note that we use the HapMap3+ set of variants here. We first compare to using raw (untransformed) phenotypes in Figure S18; estimates of p and α are highly consistent. However, h^2 estimates and predictive performance r^2 (in the test set) are generally larger with the log-transformation, hinting that it probably makes sense to transform these phenotypes. We then compare to using the rank-based inverse normal (RIN) transformation in Figure S19; estimates for p and α are also highly consistent. Except for bilirubin and lipoprotein(a) concentration, generally higher h^2 estimates and predictive performance r^2 are obtained with the RIN-transformation than the log-transformation.

Local heritability and poligenicity

In this section, we use the extended set of variants constructed here, HapMap3+, for which we define 431 independent LD blocks (Methods). We compute local per-block h^2 estimates, and report the UKBB phenotypes for which one block contributes to at least 10% of the total heritability of all blocks in Figure S20. For lipoprotein(a) concentration, "red hair" and "disorders of iron metabolism" (phecode 275.1), almost all heritability comes from one block only. We also perform the same analysis with external GWAS summary statistics for 90 cardiovascular proteins (Folkersen *et al.*, 2020); 22 (resp. 8) of them have at least 50% (resp. 80%) of their heritability explained by a single block (Figure S21).

Across 169 UKBB phenotypes with more than 25 chains kept, we compute the median heritability per block, and compare it to the number of variants in these blocks; the median heritability explained by a block of variants is largely proportional to the number of variants in this block (Figure S22). The outlier block explaining a much larger heritability contains the HLA region. Across the same phenotypes, we then compute per-variant median probabilities of being causal, and report them in a Manhattan-like plot in Figure S23. Some variants in multiple small regions across the genome have a larger probability of being causal across many phenotypes; interestingly, these are mapped to genes that are known to be associated with many different traits (up to more than 300) in the GWAS Catalog (Buniello *et al.*, 2019). To verify that this is not driven by population structure, we compute pcadapt chi-square statistics that quantify whether a variant is associated with population structure (Privé *et al.*, 2020c); the log-statistics have a negative correlation of -5.5% with the probabilities of being causal. To verify that this does not correspond to regions of low LD, we compute LD scores; the median probabilities of being causal have a correlation of 11.6% with the LD scores and of -6.8% with the log of LD scores.

Application to height

Here we run three LDpred2-auto models for height, one based on 100K UKBB individuals (as a subset of the 356K used before), one from the same 305K UKBB individuals used before, and one from a large GWAS meta-analysis of 1.6M individuals of European genetic ancestries across 1,373,020 variants

(Yengo *et al.*, 2022). We first infer the genetic ancestry proportions of individuals included in the meta-analysis using the method proposed in Privé (2022), and find that 81.9% are from N.W. Europe, 9.5% from E. Europe, 6.5% from Finland, 1.5% of Ashkenazi genetic ancestry, 0.3% from S.W. Europe, and 0.2% from W. Africa. We therefore use the same N.W. European LD matrix with these GWAS summary statistics. Note that we use the HapMap3+ set of 1,444,196 SNPs here, however, for the GWAS meta-analysis, only 1,013,499 SNPs are overlapping and passing quality control.

Intercepts from LD Score regression are increasing with sample size: 1.02 (SE: 0.008) with N=100K, 1.11 (0.015) with N=305K, and 2.31 (0.068) with N=1.6M, as expected (Loh *et al.*, 2018). SNP heritability estimates are 64.6% (SE: 2.7), 59.7% (2.2), and 39.2% (1.7) with LD Score regression, respectively, and 60.2% [95% CI: 57.2, 63.2], 63.2% [62.0, 64.4], and 54.2% [53.9, 54.5] with LDpred2-auto. As expected, predictive performance r^2 (estimated from the Gibbs sampler) are increasing with sample size, with 29.6% [28.7, 30.5], 42.7% [42.2, 43.1], and 47.0% [46.8, 47.1], respectively. Note that these r^2 estimates are probably overestimated by the same margin as the (SNP heritability) h^2 estimates, and correspond to ~49%, ~67.5%, and ~87% of h^2, respectively. We emphasize that, even though there are 1.6M individuals in the meta-analysis, the predictive performance corresponds to around 87% of the SNP heritability only, therefore requiring an even larger sample size to be able to better predict height. Polygenicity estimates from LDpred2-auto are increasing with sample size with 1.2% [1.0, 1.5], 2.3% [2.0, 2.5], and 5.9% [5.6, 6.3], consistent with results of simulations with a large polygenicity (p=10%). Therefore, we estimate that height has at least 50,000 causal variants. These results are similar independent of whether "allow_jump_sign" is used or not, which is surprising to us. We also identify 1753 SNPs with a probability of being causal larger than 95% (fine-mapping), which are spread over the entire genome (Figure S24). As for α estimates from LDpred2-auto, they remain consistent, with -0.71 [-0.75, -0.67], -0.74 [-0.76, -0.72], and -0.78 [-0.82, -0.76], respectively. Finally, we compute per-annotation heritability estimates to investigate functional enrichment. We perform this analysis using 50 non-flanking binary annotations from the baselineLD v2.2 model (Finucane *et al.*, 2018). Heritability enrichments are rather modest, ranging from 0.7 to 2.5 with a GWAS sample size of N=305K, and of slightly smaller magnitude with N=100K and slightly larger magnitude with N=1.6M (Figure S25).

Discussion

LDpred2-auto was originally developed for building polygenic scores (Privé *et al.*, 2020b). Here we have extended the LDpred2-auto model and shown that it can be used to reliably infer genetic architecture parameters such as the SNP-heritability (both genome-wide, local, and for annotations), polygenicity (and per-variant probabilities of being causal), and the selection-related parameter α. We remind readers that it can also be used to infer the uncertainty of individual polygenic scores (Ding *et al.*, 2022a). We also present a new way to infer the out-of-sample predictive performance r^2 of the resulting PGS, assuming the target sample has the same genetic ancestry as the GWAS used for training. Results across 248

phenotypes demonstrate that most of these phenotypes are very polygenic, yet do not have an infinitesimal architecture (i.e. not all variants are causal); this is consistent with LDpred2-inf generally providing lower predictive performance than LDpred2-grid or LDpred2-auto (Privé *et al.*, 2020b). We also obtain widespread signatures of negative selection with most α estimates between -1.0 and -0.4, consistent with previous findings (Zeng *et al.*, 2021). However, when looking at the heritability enrichment of several functional annotations for height, we obtain much smaller magnitudes than stratified LD Score regression (S-LDSC, Finucane *et al.* (2018)). For example, Yengo *et al.* (2022) report fold enrichments of more than 10x for e.g. coding and conserved variants, while we get less than 2x. This is partly due to LDpred2-auto estimates being more conservative as they are shrunk towards no enrichment (the prior), however we do use a very large sample size here so that the prior should not matter much. We also note that this heritability partitioning is performed after running LDpred2-auto for each annotation independently, therefore, unlike S-LDSC, the LDpred2-auto heritability partitioning does not depend on the set of annotations used.

Here we have also extended the set of HapMap3 variants recommended to use with LDpred2, making it 37% larger to offer a better coverage of the genome. This enables us to capture more of the heritability of phenotypes and therefore reduce the missing heritability (i.e. the difference between the family-based heritability and the SNP-based heritability). Using the new HapMap3+ set also improves predictive performance by an average of 6.1% here, and particularly for lipoprotein(a) concentration with an r^2 of 0.516 instead of 0.344. However, we note that we are able to achieve an r^2 of 0.677 [0.671, 0.682] when using the penalized regression implementation of Privé *et al.* (2019) on the UKBB individual-level data while restricting to all variants within a 1Mb window of the *LPA* gene. This means that this extended SNP set is still not tagging all variants perfectly, and that it might be preferable to use a more localized set of variants for phenotypes for which most of the heritability is contained in a single region of the genome.

Our proposed method has limitations. First, when power is low (i.e. when Nh^2/p is low), estimates of α and p become less reliable. However, estimates of h^2 and r^2 seem always reliable, except for height for which they are probably overestimated. We think this is likely due to assortative mating (Border *et al.*, 2022; Yengo *et al.*, 2022). Moreover, the h^2 from LDpred2-auto is also slightly overestimated when using a small LD reference panel or when the reference panel does not closely match with the ancestry of the GWAS summary statistics. Future work could focus on correcting these issues.

Nevertheless, LDpred2-auto users can now get much from running a single method. The reliable estimates provided by LDpred2-auto are very encouraging to further extend LDpred2-auto in multiple directions. As future research directions, we are interested in using LDpred2-auto for GWAS summary statistics imputation (Rüeger *et al.*, 2018; Julienne *et al.*, 2019), for genetic correlation estimation (Bulik-Sullivan *et al.*, 2015a; Shi *et al.*, 2017; Speed and Balding, 2019; Frei *et al.*, 2019; Werme *et al.*, 2022), multi-ancestry prediction and inference (Brown *et al.*, 2016; Shi *et al.*, 2020; Ruan *et al.*, 2022; Lu *et al.*, 2022), as well as extending it to learn from functional annotations (Zhang *et al.*, 2021; Márquez-Luna *et al.*, 2021).

Materials and Methods

Data for simulations

For simulations, we use the UK Biobank imputed (BGEN) data, read as allele dosages with function `snp_readBGEN` from R package bigsnpr (Bycroft *et al.*, 2018; Privé *et al.*, 2018). We use the set of 1,054,330 HapMap3 variants recommended to use for LDpred2 (Privé *et al.*, 2020b). Since we run lots of different models, we restrict the simulations to chromosomes 3, 6, 9, 12, 15, 18 and 21, resulting in a set of 322,805 SNPs. 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 (field 22009) and further restrict individuals to those within a log-distance of 4.5 (Privé *et al.*, 2020a). This results in 356,409 individuals of Northwestern European ancestry. We randomly sample 200,000 individuals to form a training set (to run the GWAS), and use the remaining individuals to form a test set (to evaluate the predictive models).

Data for the UK Biobank analyses

We use the set of 1,054,330 HapMap3 variants recommended to use for LDpred2 (Privé *et al.*, 2020b), and the same 356,409 individuals of Northwestern European ancestry as in the simulations. We randomly sample 50,000 individuals to form a test set (to evaluate the predictive models), and use the remaining individuals to form a training set (to run the GWAS).

We construct and use the same phenotypes as in Privé *et al.* (2022b). About half of these consists of phecodes mapped from ICD10 and ICD9 codes using R package PheWAS (Carroll *et al.*, 2014; Wu *et al.*, 2019). The other half consists of phenotypes defined in UKBB fields based on manual curation (Privé *et al.*, 2022b). As covariates, for the subset of individuals previously defined, we first recompute PCs using function `snp_autoSVD` from R package bigsnpr and keep four PCs based on visual inspection (Privé *et al.*, 2018, 2020a). We also use sex (field 22001), age (field 21022), birth date (combining fields 34 and 52) and deprivation index (field 189) as additional covariates (to a total of eight).

We use the LD matrix with independent LD blocks computed in Privé *et al.* (2022a). We design two other LD matrices: one using a smaller subset of 2000 individuals from the previously selected ones (which we call "hm3_small"), and one based on 10,000 individuals from around South Europe by using the "Italy" center defined in Privé *et al.* (2022b) ("hm3_altpop"). We apply the optimal algorithm developed in Privé (2021) to obtain independent LD blocks, as recommended in Privé *et al.* (2022a). We finally define a fourth LD reference by extending the set of HapMap3 variants (see next Methods section) and using 20,000 individuals from the previously selected ones ("hm3_plus").

Extending the set of HapMap3 variants used

The HapMap3 variants generally provide a good coverage of the whole genome. We recall that the set of 1,054,330 HapMap3 variants recommended to use for LDpred2 (Privé *et al.*, 2020b) is a subset of the original set of HapMap3 variants, which does not include duplicated positions (e.g. multi-allelic variants), nor ambiguous variants (e.g. both 'A' and 'T' or 'C' and 'G'), and which includes SNPs only (e.g. no indel). Here we propose to extend this set of 1,054,330 HapMap3 variants to make sure many genetic variants are well tagged by the extended set. To design this new set, we first read all variants from the UK Biobank (UKBB) with a minor allele frequency (MAF) larger than 0.005 in the whole data (i.e. the MAF from the MFI files). We then compute all pairwise correlations between variants within a 1 Mb distance, restricting to squared correlations larger than 0.3, and using all unrelated UKBB individuals excluding all White British (field 22006) to have a set of individuals from diverse ancestries. Finally, we design an algorithm which aims at maximizing the tagging of all these variants read. We want to maximize $\sum_j \max_{k \in \text{HapMap3+}} r^2_{j,k}$, where j spans the whole set of variants read, while k spans the variants kept in the new set, which we call HapMap3+. We start by including all previously used HapMap3 variants. Then, for the sake of simplicity, we use a greedy approach, where we repeatedly include the variant which increases this sum most, until no variant improves it by more than 2. Note that we only allow non-ambiguous SNPs to be included. This results in an extended set of 1,444,196 SNPs, of which we compute the correlation between variants (within a 3 cM window) and apply the optimal algorithm developed in Privé (2021) to obtain 431 independent LD blocks.

New model and inference with LDpred2-auto

LDpred2 originally assumed the following model for effect sizes,

$$\beta_j = S_j \gamma_j \sim \begin{cases} \mathcal{N}\left(0, \dfrac{h^2}{Mp}\right) & \text{with probability } p, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where p is the proportion of causal variants, M the number of variants, h^2 the (SNP) heritability, $\boldsymbol{\gamma}$ the effect sizes on the allele scale, $\boldsymbol{S}$ the standard deviations of the genotypes, and $\boldsymbol{\beta}$ the effects of the scaled genotypes (Privé *et al.*, 2020b). In LDpred2-auto, p and h^2 are directly estimated within the Gibbs sampler, as opposed to testing several values of p and h^2 from a grid of hyper-parameters. This makes LDpred2-auto a method free of hyper-parameters which can therefore be applied directly without requiring a validation dataset to choose best-performing hyper-parameters (Privé *et al.*, 2020b). Previously, p was sampled from $\text{Beta}(1 + M_c, 1 + M - M_c)$, where $M_c = \sum_j (\beta_j \neq 0)$.

Here we introduce a few changes to LDpred2-auto, which makes it better at inferring these important parameters. First, we extend LDpred2-auto with a third parameter α that controls the relationship between minor allele frequencies (or equivalently, standard deviations) of genotypes and expected effect

sizes; the model becomes

$$\beta_j = S_j \gamma_j \sim \begin{cases} \mathcal{N}\big(0,\ \sigma_\beta^2 \cdot (S_j^2)^{(\alpha+1)}\big) & \text{with probability } p, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Therefore, it was earlier assumed that $\alpha = -1$ and $\sigma_\beta^2 = h^2/(Mp)$ in equation (1). This new model in equation (2) is similar to the model assumed by SBayesS, where α is called S (Zeng *et al.*, 2021). In SBayesS, they estimate α and σ_β^2 by maximizing the likelihood of the normal distribution (over the causal variants from the Gibbs sampler). In the new LDpred2-auto, we first sample causal variants with replacement (bootstrap) before computing the maximum likelihood estimators, such that we add some proper sampling to these two parameters. This maximum likelihood estimation is implemented using R package roptim (Pan and Pan, 2022), and we bound the estimate of α to be between -1.5 and 0.5 (the default, but can be modified), and the estimate of σ_β^2 to be between 0.5 and 2 times the one from the previous iteration of the Gibbs sampler. We now sample p from Beta$(1 + M_c/\bar{l^2}, 1 + (M - M_c)/\bar{l^2})$, where $\bar{l^2}$ is the average LD score, to add more sampling by properly accounting for the reduced effective number of variants. As for h^2, we still estimate it by $h^2 = \boldsymbol{\beta}^T \boldsymbol{R} \boldsymbol{\beta}$, where $\boldsymbol{R}$ is the correlation matrix between variants and $\boldsymbol{\beta}$ is a vector of causal effect sizes (after scaling) from one iteration of the Gibbs sampler. We constrain this estimate to be at least 0.001 to prevent the Gibbs sampler from being trapped in very small heritability estimates. Note that this h^2 estimate can be restricted to e.g. variants from a single LD block to get estimates of local heritability.

Inference of predictive performance r^2

To infer the out-of-sample predictive performance r^2 (and CI) of the resulting PGS from LDpred2-auto, we use the distribution of $\boldsymbol{\beta_1}^T \boldsymbol{R} \boldsymbol{\beta_2}$, where $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are two sampled vectors of causal effect sizes (after scaling) from two different chains of the Gibbs sampler. Intuitively, if prediction is perfect then $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are the same and $r^2 = h^2$; when power is very low, these two are uncorrelated and $r^2 \approx 0$. Note that this assumes the target sample has the same genetic ancestry as the GWAS used for training (to get the summary statistics). Although we would have liked to, we do not provide any theoretical justification for this equation. Instead, we do check it using extensive simulations (across many genetic architectures) and real data analyses (across 248 different phenotypes). These are compared to the partial-r^2 (on individual-level data from a separate test set). The partial correlation (and 95% CI) is computed with function `pcor` from R package bigstatsr, adjusting for the same eight covariates as in the GWAS, then squared (while keeping the sign).

Acknowledgements

Authors thank members of the NCRR/QGG StatGen group and Marc-André Legault 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; authors thank all the UK Biobank participants for contributing to such useful data for Research.

Funding

F.P., C.A. and B.J.V. are supported by the Danish National Research Foundation (Niels Bohr Professorship to Prof. John McGrath), the Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH (R102-A9118, R155-2014-1724, R248-2017-2003), and a Lundbeck Foundation Fellowship (R335-2019-2339 to B.J.V.).

Declaration of Interests

B.J.V. is on Allelica's international advisory board. The other authors have no competing interests to declare.

Code and data availability

The UK Biobank data is available through a procedure described at `https://www.ukbiobank.ac.uk/using-the-resource/`. Descriptions of UK Biobank phenotypes used here are available at `https://github.com/privefl/paper-infer/blob/main/phenotype-description.tsv`. LD matrices for HapMap3+ variants computed from the N.W. European UKBB data used in this paper are available at `https://doi.org/10.6084/m9.figshare.21305061`. All code used for this paper is available at `https://github.com/privefl/paper-infer/tree/master/code`. 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. The latest version of R package bigsnpr can be installed from GitHub.

References

Bengtsson, H. (2021). A unifying framework for parallel and distributed processing in R using futures. *The R Journal*, **13**(2), 273–291.

Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, **32**(10), 1493–1501.

Border, R., O'Rourke, S., de Candia, T., Goddard, M. E., Visscher, P. M., Yengo, L., Jones, M., and Keller, M. C. (2022). Assortative mating biases marker-based heritability estimators. *Nature Communications*, **13**(1), 1–10.

Brown, B. C., Ye, C. J., Price, A. L., Zaitlen, N., Consortium, A. G. E. N. T. . D., *et al.* (2016). Transethnic genetic-correlation estimates from summary statistics. *The American Journal of Human Genetics*, **99**(1), 76–88.

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R., Patterson, N., Robinson, E. B., *et al.* (2015a). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, **47**(11), 1236–1241.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015b). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, **47**(3), 291–295.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., *et al.* (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics. *Nucleic Acids Research*, **47**(D1), D1005–D1012.

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.

Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*, **3**(10), e3395.

Ding, Y., Hou, K., Burch, K. S., Lapinska, S., Privé, F., Vilhjálmsson, B., Sankararaman, S., and Pasaniuc, B. (2022a). Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. *Nature Genetics*, **54**(1), 30–39.

Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., Privé, F., Vilhjálmsson, B. J., Loohuis, L. O., and Pasaniuc, B. (2022b). Polygenic scoring accuracy varies across the genetic ancestry continuum in all human populations. *bioRxiv*.

Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shoresh, N., *et al.* (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics*, **50**(4), 621–629.

Folkersen, L., Gustafsson, S., Wang, Q., Hansen, D. H., Hedman, Å. K., Schork, A., Page, K., Zhernakova, D. V., Wu, Y., Peters, J., *et al.* (2020). Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nature Metabolism*, **2**(10), 1135–1148.

Frei, O., Holland, D., Smeland, O. B., Shadrin, A. A., Fan, C. C., Maeland, S., O'Connell, K. S., Wang, Y., Djurovic, S., Thompson, W. K., *et al.* (2019). Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. *Nature Communications*, **10**(1), 1–11.

Grotzinger, A. D., de la Fuente, J., Privé, F., Nivard, M. G., and Tucker-Drob, E. M. (2022). Pervasive downward bias in estimates of liability-scale heritability in GWAS meta-analysis: A simple solution. *Biological Psychiatry*.

Johnson, R., Burch, K. S., Hou, K., Paciuc, M., Pasaniuc, B., and Sankararaman, S. (2021). Estimation of regional polygenicity from gwas provides insights into the genetic architecture of complex traits. *PLoS Computational Biology*, **17**(10), e1009483.

Julienne, H., Shi, H., Pasaniuc, B., and Aschard, H. (2019). RAISS: robust and accurate imputation from summary statistics. *Bioinformatics*, **35**(22), 4837–4839.

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.

Kumuthini, J., Zick, B., Balasopoulou, A., Chalikiopoulou, C., Dandara, C., El-Kamah, G., Findley, L., Katsila, T., Li, R., Maceda, E. B., *et al.* (2022). The clinical utility of polygenic risk scores in genomic medicine practices: a systematic review. *Human Genetics*, pages 1–8.

Lambert, S. A., Abraham, G., and Inouye, M. (2019). Towards clinical utility of polygenic risk scores. *Human Molecular Genetics*, **28**(R2), R133–R142.

Lee, S. H., Wray, N. R., Goddard, M. E., and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, **88**(3), 294–305.

Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., de Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., *et al.* (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics*, **47**(12), 1385–1392.

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

Lu, Z., Gopalan, S., Yuan, D., Conti, D. V., Pasaniuc, B., Gusev, A., and Mancuso, N. (2022). Multi-ancestry fine-mapping improves precision to identify causal genes in transcriptome-wide association studies. *The American Journal of Human Genetics*, **109**(8), 1388–1404.

Márquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S. S., Furlotte, N., Auton, A., and Price, A. L. (2021). Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *Nature Communications*, **12**(1), 1–11.

O'Connor, L. J., Schoech, A. P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A. L. (2019). Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics*, **105**(3), 456–476.

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.

Pan, Y. and Pan, J. (2022). roptim: An R Package for General Purpose Optimization with C++. *R package version 0.1.6*.

Privé, F. (2021). Optimal linkage disequilibrium splitting. *Bioinformatics*, **38**(1), 255–256.

Privé, F. (2022). Using the UK Biobank as a global reference of worldwide populations: application to measuring ancestry diversity from GWAS summary statistics. *Bioinformatics*, **38**(13), 3477–3480.

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., Aschard, H., and Blum, M. G. (2019). Efficient implementation of penalized regression for genetic risk prediction. *Genetics*, **212**(1), 65–74.

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., Luu, K., Vilhjálmsson, B. J., and Blum, M. G. (2020c). Performing highly efficient genome scans for local adaptation with R package pcadapt version 4. *Molecular Biology and Evolution*, **37**(7), 2153–2154.

Privé, F., Arbel, J., Aschard, H., and Vilhjálmsson, B. J. (2022a). Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores. *Human Genetics and Genomics Advances*.

Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., O'Reilly, P. F., and Vilhjálmsson, B. J. (2022b). Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics*, **109**(1), 12–23.

Ruan, Y., Lin, Y.-F., Feng, Y.-C. A., Chen, C.-Y., Lam, M., Guo, Z., He, L., Sawa, A., Martin, A. R., Qin, S., *et al.* (2022). Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics*, **54**(5), 573–580.

Rüeger, S., McDaid, A., and Kutalik, Z. (2018). Evaluation and application of summary statistic imputation to discover new height-associated loci. *PLoS Genetics*, **14**(5), e1007371.

Schoech, A. P., Jordan, D. M., Loh, P.-R., Gazal, S., O'Connor, L. J., Balick, D. J., Palamara, P. F., Finucane, H. K., Sunyaev, S. R., and Price, A. L. (2019). Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nature Communications*, **10**(1), 1–10.

Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, **99**(1), 139–153.

Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. *The American Journal of Human Genetics*, **101**(5), 737–751.

Shi, H., Burch, K. S., Johnson, R., Freund, M. K., Kichaev, G., Mancuso, N., Manuel, A. M., Dong, N., and Pasaniuc, B. (2020). Localizing components of shared transethnic genetic architecture of complex traits from GWAS summary data. *The American Journal of Human Genetics*, **106**(6), 805–817.

Speed, D. and Balding, D. J. (2019). SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature Genetics*, **51**(2), 277–284.

Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics*, **91**(6), 1011–1021.

Sullivan, P. F., Kendler, K. S., and Neale, M. C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of General Psychiatry*, **60**(12), 1187–1192.

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, **19**(9), 581–590.

Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society*, **82**(5), 1273–1300.

Werme, J., van der Sluis, S., Posthuma, D., and de Leeuw, C. A. (2022). An integrated framework for local genetic correlation analysis. *Nature Genetics*, **54**(3), 274–282.

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.

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.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., *et al.* (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, **42**(7), 565–569.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, **88**(1), 76–82.

Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A. U., Jiang, Y., Raghavan, S., *et al.* (2022). A saturated map of common genetic variants associated with human height from 5.4 million individuals of diverse ancestries. *bioRxiv*.

Zeng, J., Xue, A., Jiang, L., Lloyd-Jones, L. R., Wu, Y., Wang, H., Zheng, Z., Yengo, L., Kemper, K. E., Goddard, M. E., *et al.* (2021). Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nature Communications*, **12**(1), 1–12.

Zhang, Q., Privé, F., Vilhjálmsson, B., and Speed, D. (2021). Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nature Communications*, **12**(1), 1–9.