
Inferring disease architecture and predictive ability with LDpred2-auto

Authors

Florian Privé, Clara Albiñana, Julyan Arbel,
Bogdan Pasaniuc, Bjarni J. Vilhjálmsson

Correspondence

florian.prive.21@gmail.com

This study extends the utility of LDpred2-auto, a widely used method for deriving predictive polygenic scores. It validates the use of the latest version of LDpred2-auto for inferring key genetic parameters that inform us about the genetic architecture of traits, for identifying causal genetic variants, and for estimating predictive performance.



Inferring disease architecture and predictive ability with LDpred2-auto

Florian Privé,^{1,*} Clara Albiñana,¹ Julyan Arbel,² Bogdan Pasaniuc,^{3,4,5,6} and Bjarni J. Vilhjálmsson^{1,7,8}

Summary

LDpred2 is a widely used Bayesian method for building polygenic scores (PGSs). LDpred2-auto can infer the two parameters from the LDpred model, the SNP heritability h^2 and polygenicity p , so that it does not require an additional validation dataset to choose best-performing parameters. The main aim of this paper is to properly validate the use of LDpred2-auto for inferring multiple genetic parameters. Here, we present a new version of LDpred2-auto that adds an optional third parameter α to its model, for modeling negative selection. We then validate the inference of these three parameters (or two, when using the previous model). We also show that LDpred2-auto provides per-variant probabilities of being causal that are well calibrated and can therefore be used for fine-mapping purposes. We also introduce a 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 we show that this new set of variants captures 12% more heritability and provides 6% more predictive performance, on average, in UK Biobank analyses.

Introduction

Most traits and diseases in humans are heritable. What differs is the genetic architecture of each trait, which 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 effect 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.^{1–4} Causal effects are larger when a trait is more heritable and smaller when the trait is more polygenic. As for the distribution of causal effects relative to their allele frequencies, it 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.⁵ In this model, the expected phenotypic variance explained by a genetic variant is proportional to $[f(1 - f)]^\alpha$, where f is the allele frequency of this variant. 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 (non-exhaustively) include GCTA⁶ (h^2), BOLT-REML⁷ (h^2 and p), LD Score regression⁸ (h^2), FINEMAP⁹ (per-variant p , also called posterior inclusion probabilities [PIPs], used for fine-mapping), HESS¹⁰ (local h^2), LDAK-SumHer^{11,12}

(h^2 and α), S-LD4M³ (p), GRM-MAF-LD¹³ (α), SuSiE¹⁴ (PIPs), SBayesS¹⁵ (h^2 , p , and a third parameter S , similar to α), and BEAVR¹⁶ (local p). Not all methods have the same modeling assumptions; for example, LDAK-SumHer assumes a different prior than SBayesS and LDpred2-auto and does not estimate all the same parameters. It can estimate the SNP heritability and α . However, it cannot estimate the polygenicity nor the per-variant probabilities of being causal (since it inherently assumes an infinitesimal model, i.e., $p = 1$). Moreover, since it does not sample effects, it cannot be used to estimate the predictive performance r^2 with the formula we propose in this paper.

As previously shown by Daetwyler et al.,¹⁷ h^2 and p can also be used to determine how well we can predict a phenotype using genetic variants alone, with $r_{\max}^2 = h^2 \cdot r_{gg}^2 = \frac{h^2}{1 + (1 - r_{\max}^2)} \frac{Mp}{Nh^2}$, where r_{\max}^2 is the maximum achievable squared correlation between the genetic predictor and the phenotype, r_{gg}^2 is the maximum achievable squared correlation between the genetic predictor and the genetic component, N is the sample size, M is the number of variants, and Mp is the number of causal variants. Such genetic predictors are called polygenic scores (PGSs), and they are getting closer to being included as part of existing clinical risk models for diseases.^{18–20} LDpred2 is a widely used PGS method that can directly build PGSs using resulting summary statistics from genome-wide associations studies (GWASs), making it highly applicable.^{21–23}

¹National Centre for Register-based Research, Aarhus University, Aarhus, Denmark; ²University Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, France; ³Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA, USA; ⁴Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; ⁵Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; ⁶Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; ⁷Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark; ⁸Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute, Cambridge, MA, USA

*Correspondence: florian.prive.21@gmail.com

<https://doi.org/10.1016/j.ajhg.2023.10.010>

© 2023 American Society of Human Genetics.



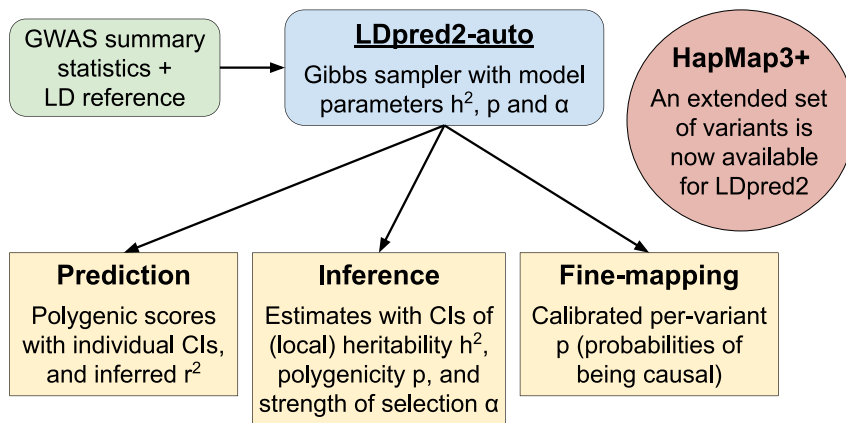


Figure 1. Overview of what LDpred2-auto can now provide

For individual CIs of polygenic scores, please refer to the work of Ding et al.^{24,25} HapMap3+ is the new extended set of 1,444,196 SNPs we introduce here (material and methods) and now recommend using for LDpred2 when power is sufficient (i.e., with a large sample size, large h^2 , and/or small p). CI means “confidence interval” (often called “credible interval” in a Bayesian setting).

LDpred2 is a Bayesian approach that uses the SNP heritability h^2 and polygenicity p as parameters of its model. LDpred2-auto, one variant of LDpred2, can directly estimate these two parameters from GWAS summary statistics, making it applicable even when no validation data are available for tuning these two model parameters.²¹

The main aim of this paper is to properly validate the use of LDpred2-auto for inferring multiple genetic parameters. Here we extend LDpred2-auto and show that it is a reliable method for estimating h^2 (global and local), p (also per-variant probabilities [PIPs] used for fine-mapping purposes), and α (by extending its model to also include this third parameter). So, on top of providing competitive PGSS, LDpred2-auto can also provide all these estimates of genetic architecture. Moreover, we show how it can now also reliably estimate the predictive ability r^2 of the PGSS it derives, allowing for direct assessment of the usefulness of the derived PGSS, without requiring an independent test set. An overview of what LDpred2-auto can now provide is presented in Figure 1. Finally, we extend the set of HapMap3 variants recommended for use with LDpred2, which enables us to capture around 12% more SNP heritability and achieve around 6% more predictive performance r^2 , on average, in UK Biobank (UKBB) analyses. We call this extended set of 1,444,196 SNPs “HapMap3+” and recommend using it when power is sufficient (i.e., with a large sample size, large h^2 , and/or small p).

Material and methods

We have extensively used R packages `bigstatsr` and `bigsnpr`²⁶ for analyzing large genetic data, packages from the future framework²⁷ for easy scheduling and parallelization of analyses on the high-performance computing cluster, and packages from the tidyverse suite²⁸ for shaping and visualizing results. We have extensively used the UKBB data,²⁹ which is available through a procedure described at <https://www.ukbiobank.ac.uk/using-the-resource/>. The UKBB received ethical approval from the NHS National Research Ethics Service North West (11/NW/0382). The present analyses were conducted under UKBB data application number 58024.

Data for simulations

For simulations, we use the UKBB imputed (BGEN) data, read as allele dosages with function `snpr_readBGEN` from R package `bigsnpr`.^{26,29} We use the set of 1,054,330 HapMap3 variants recommended to use for LDpred2.²¹ 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 UKBB (field 22020). These individuals are unrelated and have passed some quality control such as 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 Bycroft et al.²⁹). 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 4.5.³⁰ 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 UKBB analyses

We use the set of 1,054,330 HapMap3 variants recommended to use for LDpred2²¹ 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, Aschard, et al.³¹ About half of these consist of phecodes mapped from ICD10 and ICD9 codes using R package `PheWAS`.^{32,33} The other half consist of phenotypes defined in UKBB fields based on manual curation.³¹ As covariates, we first recompute PCs for the homogeneous subset of individuals previously defined using function `snpr_autoSVD` from R package `bigsnpr` and keep four PCs based on visual inspection.^{26,30} 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, Arbel, et al.³⁴ We design two other LD matrices: one using a smaller random subset of 2,000 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, Aschard, et al.³¹ (“hm3_altpop”). We apply the optimal algorithm developed in Privé³⁵ to obtain independent LD blocks, as recommended in Privé, Arbel, et al.³⁴ We finally define a fourth LD reference by extending the set of HapMap3 variants (see next 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 LDpred2²¹ 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., “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 UKBB with a minor allele frequency (MAF) larger than 0.005 in the whole data (i.e., the MAF from the MFI files). There are around 11.5 million such variants. Then we restrict to unrelated UKBB individuals that are *not* listed as White British (field 22006) and use these individuals of diverse ancestries^{29,36} to compute all pairwise correlations between variants within a 1 Mb distance, restricting to squared correlations larger than 0.3. Finally, we design an algorithm that aims at maximizing the tagging of all these variants read. We want to maximize $\sum \max_{k \in \text{HapMap3}+} r_{jk}^2$, where j spans the whole set of variants read and 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 that 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 LD matrix (within a 3 cM window) and apply the optimal algorithm developed in Privé (2021)³⁵ to obtain 431 independent LD blocks. Since we use individuals of diverse ancestries for computing the pairwise variant correlations used for constructing this extended set of variants, we expect this new set of variants to be beneficial across diverse ancestries. Indeed, from the 11.5 million variants we aimed at tagging, 82.1% (respectively, 69.1%), 80.0% (respectively, 66.7%), 79.3% (respectively, 66.1%), 75.4% (respectively, 65.3%), and 66.6% (respectively, 48.6%) are tagged at $r^2 > 50\%$ (respectively, $r^2 > 80\%$) by at least one HapMap3+ variant in Northwestern Europeans, Middle Easterns, South Asians, East Asians, and West Africans (results extrapolated from variants in chromosome 22).

Optional extended model and inference with LDpred2-auto

LDpred2 originally assumed the following model for effect sizes:

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

where p is the proportion of causal variants, M the number of variants, h^2 the (SNP) heritability, γ the effect sizes on the allele scale, S the standard deviations of the genotypes, and β the effects of the scaled genotypes.²¹ 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 (as in

LDpred2-grid). This makes LDpred2-auto a method free of hyper-parameters that can therefore be applied directly without requiring a validation dataset to choose best-performing hyper-parameters.²¹ h^2 is estimated by $\beta^T \mathbf{R} \beta$, where \mathbf{R} is the correlation matrix between variants and β is a vector of causal effect sizes (after scaling) from one iteration of the Gibbs sampler. As for p , it is sampled from $\text{Beta}(1+M_c, 1+M - M_c)$, where $M_c = \sum_j (\beta_j \neq 0)$. Note a small change: we now sample p from

$\text{Beta}(1+M_c/\bar{l}^2, 1+(M - M_c)/\bar{l}^2)$, where \bar{l}^2 is the average LD score, to add more variability in the sampling in order to account for a reduced effective number of (independent) variants.

Here we provide an extended model and sampling scheme for LDpred2-auto that can be optionally used by setting parameter `use_MLE = TRUE` (otherwise it is run as described in the previous paragraph when using `use_MLE = FALSE`). In this new option, we extend LDpred2-auto with a third parameter α that controls the relationship between MAFs (or, equivalently, standard deviations) of genotypes and expected effect sizes; the model becomes

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

Therefore, it was earlier assumed that $\alpha = -1$ and $\sigma_\beta^2 = h^2/(Mp)$ in Equation 1. This extended model in Equation 2 is similar to the model assumed by SBayesS, where α is denoted by S instead.¹⁵ In SBayesS, α and σ_β^2 are estimated by maximizing the likelihood of the normal distribution (over the causal variants from the Gibbs sampler). When using this 3-parameter model in LDpred2-auto, in order to add some sampling to these two parameters, we first sample causal variants with replacement before computing the maximum likelihood estimators. This maximum likelihood estimation (MLE) is implemented using R package `roptim` (see the web resources), 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. Note that we still estimate $h^2 = \beta^T \mathbf{R} \beta$, but that h^2 is not used in the variance of sampled effect sizes anymore (Equation 2). Note also that, to get local heritability estimates (e.g., for a single LD block), this h^2 estimation ($\beta^T \mathbf{R} \beta$) is simply restricted to the variants from this LD block. Finally, in both models and sampling schemes now implemented in LDpred2-auto, we now detect strong divergence when $\beta^T \hat{\beta} > 2\hat{\beta}^T \hat{\beta}$, where β is the vector of scaled effect sizes from one iteration of the Gibbs sampler and $\hat{\beta}$ is the marginal scaled effect sizes; corresponding chains are stopped, and missing values are returned for effect sizes and estimates of missing iterations.

Inference of predictive performance r^2

To infer the out-of-sample predictive performance r^2 (and its confidence interval [CI]) of the resulting PGS from LDpred2-auto, we use the distribution of $\beta_1^T \mathbf{R} \beta_2$, where β_1 and β_2 are two sampled vectors of causal effect sizes (after scaling) from two different chains of the Gibbs sampler. Intuitively, if the prediction is perfect, then $\beta_1 = \beta_2$ and $r^2 = h^2$; when power is very low, β_1 and β_2 are almost uncorrelated and $r^2 \approx 0$. Others have proposed to estimate r^2 from a reference genotype set³⁷ or from an additional set of external GWAS summary statistics and LD^{38,39}; here we only use the summary statistics that we input to LDpred2-auto. These previous works have shown that r^2 can be approximated by

$(\bar{\beta}^T \hat{\beta}_{\text{test}})^2 / (\bar{\beta}^T \mathbf{R} \bar{\beta})$ where $\bar{\beta}$ and $\hat{\beta}_{\text{test}}$ are, respectively, the predictive effects from the training set and the marginal effects from the test set (after scaling). Note that $E[\beta_1] = \bar{\beta}$, $E[\beta_1^T \mathbf{R} \beta_2] = \bar{\beta}^T \mathbf{R} \bar{\beta}$, and, when the test sample has the same genetic ancestry as the GWAS used for training (to get the summary statistics), $E[\mathbf{R} \beta_2] \approx \bar{\beta}_{\text{test}}$. Therefore, we propose to use $\beta_1^T \mathbf{R} \beta_2$ as a sample of r^2 . This can also be computed for a specific chain by taking two samples β_1 and β_2 that are far enough on the same chain to remove the possible autocorrelation. This is what we use for SBayesS here, and also as an alternative means for post-filtering chains for prediction (see next material and methods section). Note that Ding et al.²⁴ investigated autocorrelation in LDpred2(-grid) and showed that it decays very quickly. Therefore, picking two LDpred2-auto samples that are 100 iterations apart should be more than enough to ensure quasi-independence of these samples.

In this paper, we check this r^2 approximation 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 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). Corresponding 95% CIs are estimated through bootstrapping individual indices.

Post-filtering of chains in LDpred2-auto

Because a Gibbs sampler can be unstable, with so many variants and with possible mismatches between, e.g., the GWAS summary statistics and the LD reference used, we have always recommended running multiple chains in LDpred2-auto and post-filtering some of them as quality control.²¹ We originally proposed to filter chains by keeping the ones providing PGSs with the largest variances. Then we tested an almost equivalent and simpler alternative,³⁴ which is to keep only chains that provide top imputed marginal scaled effect sizes $\beta = R\bar{\beta}$, where R is the correlation matrix and $\bar{\beta}$ are the PGS scaled effect sizes. This is the default post-filtering of LDpred2-auto chains that we use here, for both prediction and inference.

Moreover, here we test two alternative filtering criteria in the first simulations based on continuous outcomes (and call this “LDpred2-auto_alfilter”). First, for prediction, we test filtering on the average of $\beta_1^T \mathbf{R} \beta_2$ within each chain, which is an estimate of r^2 (cf. previous material and methods section). Second, for inference, we filter on some convergence criterion. The split-Rhat statistic is a popular metric to test for good mixing and convergence of Markov chains.⁴⁰ However, we have found this statistic to perform poorly when, e.g., one parameter gets stuck and is constant; in this case, the chain does not mix well, but a perfect Rhat of 1.0 is obtained. Instead, we have found that a two-sample Cramer-von Mises statistic,⁴¹ by similarly using both parts of the chain after burn-in, is often highly correlated with the split-Rhat statistic while not suffering from the previous issue. We therefore chose to use this statistic and to average the three statistics computed for h^2 , p , and α for each chain. We use a threshold of 4, above which we filter out the chain, because we have found that a value of 4 for this statistic often corresponds to a value close to 1.05 for the split-Rhat.

Results

Validating the inference with simulations

For simulations, we use the UKBB imputed data.²⁹ We use 356,409 individuals of Northwestern European ancestry

and 322,805 SNPs across seven chromosomes (material and methods). We first simulate continuous phenotypes using function `snp_simuPheno` from R package `bigsnpr`,²⁶ 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 MAFs and expected effect sizes. This function first picks a proportion p of causal variants at random, samples effect sizes γ using the variance component parameterized by α , and then scales 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 final phenotype has a variance of 1. Then, we run a GWAS to obtain summary statistics using N individuals (either the 200,000 dedicated to this, or a random subset of 20,000), using fast linear regressions implemented in `big_univLinReg` from R package `bigstatsr`.²⁶ Finally, we run LDpred2-auto with and without the option `allow_jump_sign`, which was proposed in Privé, Arbel, et al.³⁴ for robustness (when disabled, it prevents effect sizes from changing sign without going through 0 first), and with and without the extended model including a third parameter α (using option `use_MLE`; material and methods). 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 some of these chains are filtered out for quality control (material and methods).

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, 3, 4, and S1–S6). 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 (Figures 2, S1, and S2). When power is low (e.g., $N = 20,000$ and $h^2 = 0.01$), LDpred2-auto_noMLE (with only two model parameters, as in previous versions of LDpred2-auto) and SBayesS are both over-confident in their estimate of h^2 (i.e., CIs are small and do not contain the true parameter value; Figure S7), and LDpred2-auto_jump overestimates the heritability (Figure S1). LDpred2-auto_nojump looks very reliable for estimating h^2 . When power is low, LDpred2-auto can overestimate the polygenicity when the true value is very small (e.g., $p = 0.0005$) and underestimate it when the polygenicity is large (e.g., $p = 0.1$; Figure S3). SBayesS, which uses a similar model with the same three parameters, often overestimates the polygenicity, especially when $p \leq 0.02$. The α estimate of LDpred2-auto (with the extended 3-parameter model) can become very imprecise when power is too low, which can be detected by a small number of chains kept from LDpred2-auto. Estimates of both p and α from SBayesS are often over-confident with small CIs that do not include the true simulated values (Figures S9 and S11). Generally, 95% CIs for all three parameters (h^2 , p , and α) obtained from LDpred2-auto_nojump contain the true simulated value (Figures S7–S12), therefore confirming the validity of these CIs; this is the preferred method we recommend

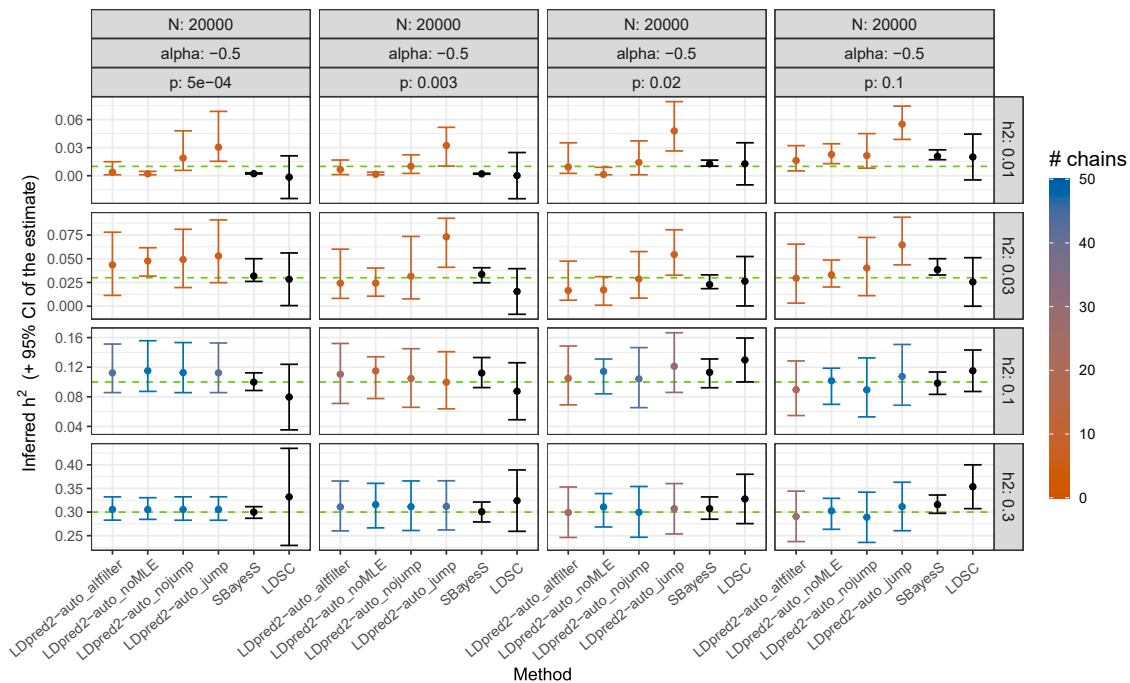


Figure 2. Inferred SNP heritability h^2 in simulations with continuous outcomes and $N = 20,000$

More results can be found in the supplemental information. Horizontal dashed lines represent the true simulated values. For LDpred2-auto, suffix “nojump”/“jump” refers to using `allow_jump_sign = FALSE/TRUE` (and `use_MLE = TRUE`), “noMLE” refers to using `use_MLE = FALSE` (and `allow_jump_sign = FALSE`), and “altfilter” is similar to “nojump” but uses a different post-filtering of chains (material and methods). Note that the recommended option is to use `allow_jump_sign = FALSE`.³⁴ The 95% CIs for the LDpred2-auto and SBayesS estimates are 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 (note that only one chain is used and kept in SBayesS). The 95% CI for the LD Score regression (LDSC) estimate is obtained as ± 1.96 of its standard error. Colors for LDpred2-auto models represent the number of chains kept (out of 50).

using. Finally, we also investigate alternative ways of post-filtering chains in LDpred2-auto, for both prediction and inference (LDpred2-auto_altfilter compared to LDpred2-auto_nojump; material and methods); results remain practically unchanged (Figures S1–S6).

Then, LDpred2-auto can also infer per-variant probabilities of being causal and local per-block heritability estimates, which are well calibrated (Figures S13 and S14). We recall that calibrated per-variant probabilities of being causal (also known as PIPs) can be used for fine-mapping purposes.¹⁴ LDpred2-auto provides PIPs that are more calibrated than with, e.g., SuSiE-RSS,⁴² which we run assuming there are 10 causal variants per LD block by default (Figure S15). 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 (material and methods), even when power is low, and we show it works with results from SBayesS’s Gibbs sampler as well (Figures S16 and S17). CIs for this r^2 estimate very often encompass the true simulated value, except for LDpred2-auto_jump when both h^2 and N are small (Figures S18 and S19).

We then run simulations with binary outcomes where the simulated continuous liabilities are transformed to binary outcomes using a threshold corresponding to the prevalence. Results are very similar to when using the continuous phenotypes above (Figures S20–S23), and

they are similar whether we use 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 must be transformed to the liability scale,⁴³ where $K_{\text{GWAS}} = 0.5$ should be used for transforming estimates of h^2 and r^2 when using N_{eff} in inference methods.⁴⁴

Genetic architectures of 248 phenotypes from the UKBB

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 eight covariates (material and methods). We first use the set of 1,054,330 HapMap3 variants recommended to use for LDpred2.²¹ Here, if not otherwise specified, we use options `use_MLE = TRUE` (i.e., the extended 3-parameter model and sampling scheme) and `allow_jump_sign = FALSE` (when disabled, this prevents effect sizes from changing sign without going through 0 first and has been proposed for extra robustness in Privé, Arbel, et al.³⁴).

Consistent with simulations, inferred SNP heritability h^2 estimates from LDpred2-auto closely match with those from LD Score regression, while generally being more

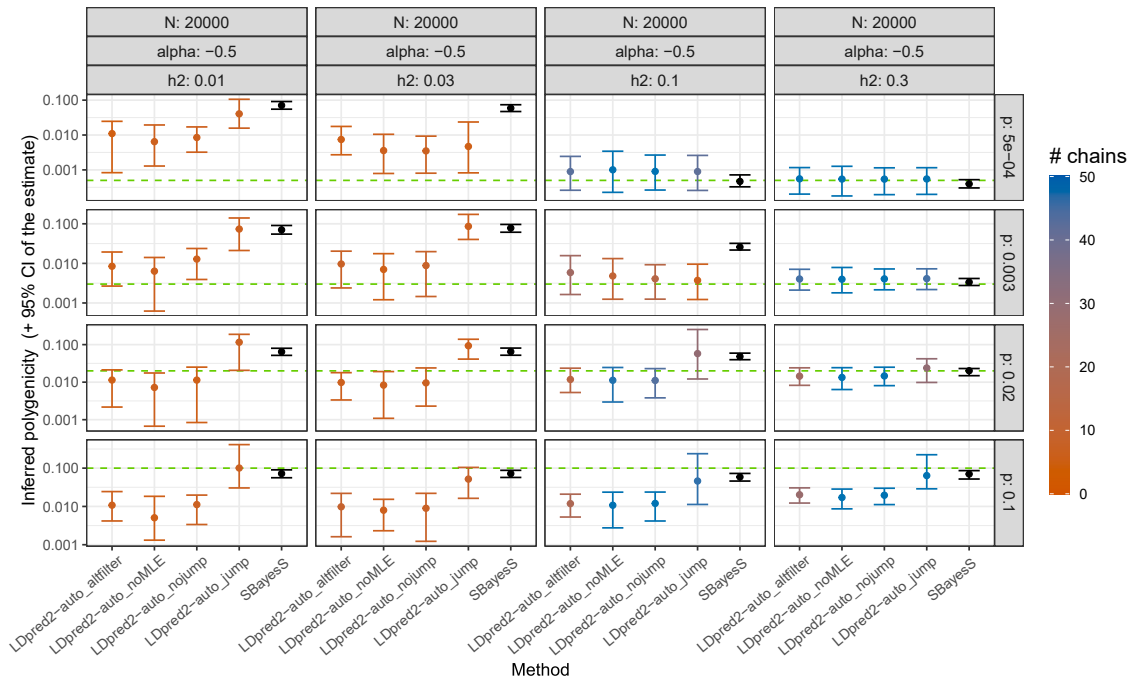


Figure 3. Inferred polygenicity p in simulations with continuous outcomes and $N = 20,000$

More results can be found in the supplemental information. Horizontal dashed lines represent the true simulated values. For LDpred2-auto, suffix “nojump”/“jump” refers to using `allow_jump_sign = FALSE/TRUE` (and `use_MLE = TRUE`), “noMLE” refers to using `use_MLE = FALSE` (and `allow_jump_sign = FALSE`), and “altfilter” is similar to “nojump” but uses a different post-filtering of chains (material and methods). Note that the recommended option is to use `allow_jump_sign = FALSE`.³⁴ The 95% CIs for the LDpred2-auto and SBayesS estimates are obtained from the 2.5% and 97.5% quantiles of all the p estimates from the iterations (after burn-in) of the chains kept (note that only one chain is used and kept in SBayesS). Colors for LDpred2-auto models represent the number of chains kept (out of 50).

precise, especially for phenotypes with a small polygenicity (Figure S24). 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 S25). Most phenotypes have an estimated α between -1.0 and -0.3 with a mode at -0.65 (Figure S26). As for the inferred predictive performance r^2 (from the Gibbs sampler of LDpred2-auto; material and methods), they are highly consistent with the predictive performance in the test set; only for standing height are they overestimated (Figures 5 and S27). Heritability estimates for height are probably overestimated as well since we use similar formulas for estimating h^2 and r^2 (material and 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](#)).

To investigate whether estimates from LDpred2-auto are robust to some misspecifications, we test using two alternative LD references (material and methods). Using a smaller number of individuals for computing the LD matrix results in a slightly overestimated p and h^2 (and r^2) with LDpred2-auto, 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 S28). When using an LD reference from an alternative population (South Europe instead of Northwest 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 of the LDpred2-auto chains is usable (distinct from the previous three; Figure S29). Interestingly, using the previous approach (`use_MLE = FALSE`) seems to provide more robust results, where we can always get some chains not to diverge (and therefore get non-zero predictive performance) for the seven (three and four) phenotypes mentioned before when using the previous alternative LD references (Figure S30).

Then, we investigate the effect of disabling the LDpred2-auto parameter `allow_jump_sign` on the estimates from LDpred2-auto; when disabled, this prevents effect sizes from changing sign without going through 0 first and has been proposed for extra robustness in Privé, Arbel, et al.³⁴ Consistent with simulations, p estimates from LDpred2-auto are conservatively lower than when allowing effects to “jump” sign (i.e., using a standard sampling; Figure S31). 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 (when using the Northwestern European LD reference) and a standard sampling can be used in this case (Figure S31).

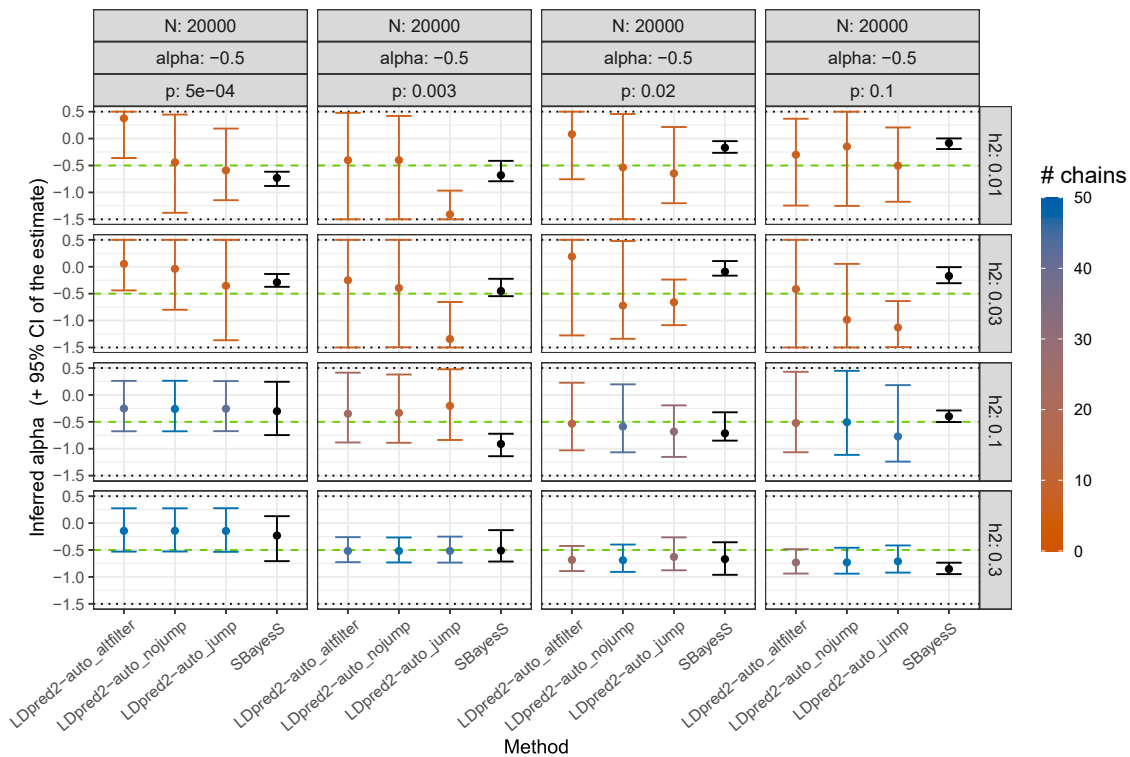


Figure 4. Inferred α in simulations with continuous outcomes and $N = 20,000$

More results can be found in the supplemental information. Horizontal dashed lines represent the true simulated values. Horizontal dotted lines represent boundaries imposed on the LDpred2-auto estimates. For LDpred2-auto, suffix “nojump”/“jump” refers to using `allow_jump_sign = FALSE/TRUE` (and `use_MLE = TRUE`), and “altfilter” is similar to “nojump” but uses a different post-filtering of chains (material and methods). Note that “LDpred2_noMLE” (`use_MLE = FALSE`) does not infer α . Note also that the recommended option is to use `allow_jump_sign = FALSE`.³⁴ The 95% CIs for the LDpred2-auto and SBayesS estimates are obtained from the 2.5% and 97.5% quantiles of all the α estimates from the iterations (after burn-in) of the chains kept (note that only one chain is used and kept in SBayesS). Colors for LDpred2-auto models represent the number of chains kept (out of 50).

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 and prediction from LDpred2-auto. We first compare to using raw (untransformed) phenotypes in Figure S32; 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; it probably makes sense to transform these phenotypes. We then compare to using the rank-based inverse normal (RIN) transformation in Figure S33; estimates for p and α are also highly consistent. Except for bilirubin and lipoprotein(a) concentration, higher h^2 estimates and predictive performance r^2 are generally obtained with the RIN transformation than the log-transformation.

More heritability and predictive accuracy with the new set of variants

Here we use the same individuals as in the previous section. We investigate using the extended set of HapMap3 variants proposed here, HapMap3+ (material and methods), which

includes $\sim 37\%$ more variants on top of the HapMap3 variants recommended to use for LDpred2 (i.e., 1,054,330 + 389,866 variants), to improve the genome coverage of this set. 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 6). 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 S34). However, running LDpred2 with this extended set of variants takes around 50% more time; yet we remind the reader that LDpred2 has been made much faster in Privé, Arbel, et al.³⁴ and now runs in less than 1 h for 50 chains parallelized over 13 cores (Figure S35), instead of 4–12 h before.

Local heritability and polygenicity

In this section, we use the extended set of variants constructed here, HapMap3+, for which we define 431

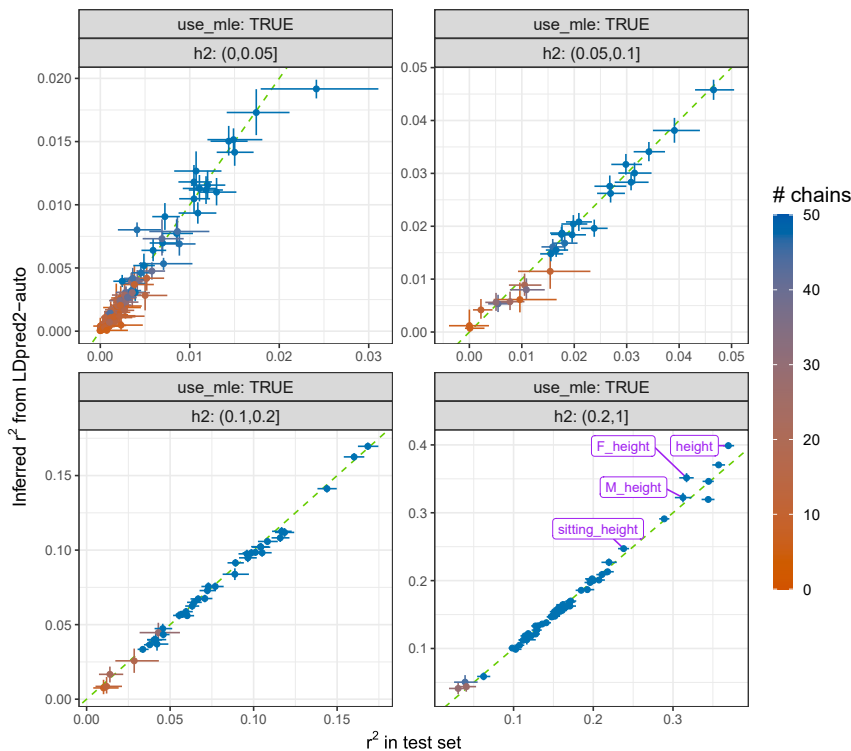


Figure 5. Inferred predictive performance r^2 in the UK Biobank

This compares the r^2 inferred from the Gibbs sampler of LDpred2-auto (material and methods) versus the ones obtained in the test set, for all 248 phenotypes defined from the UKBB. These are stratified by the SNP heritability estimated from LDpred2-auto. Green dashed lines represent the 1:1 line. The 95% CI 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% CI for r^2 in the test set is obtained from bootstrap. Colors represent the number of chains kept (out of 50). “F_height” and “M_height” use females and males only, respectively (in both GWAS and test sets).

the median probabilities of being causal have a correlation of 32.0% with the LD scores and of 26.4% with the log of LD scores. Therefore, the posterior probabilities of being causal obtained from LDpred2-auto tend to slightly increase with LD scores.

independent LD blocks (material and 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 S36. For lipoprotein(a) concentration, “red hair” and “disorders of iron metabolism” (phecode 275.1), almost all heritability comes from one LD block only. We also perform the same analysis with external GWAS summary statistics for 90 cardiovascular proteins⁴⁵; 22 (respectively, 8) of them have at least 50% (respectively, 80%) of their heritability explained by a single block (Figure S37).

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 S38). 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 S39. 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.⁴⁶ To verify that this is not driven by population structure, we compute pcadapt chi-squared statistics that quantify whether a variant is associated with population structure⁴⁷; the log-statistics have only a small correlation of 5.9% with the probabilities of being causal. To verify that this does not correspond to regions of low LD, we compute LD scores;

Finally, since the HapMap3+ variants still represent a rather small proportion of common variants, we showcase running LDpred2-auto using a more dense set of variants from a small region, as usually done for fine-mapping. We identify the most significant HapMap3+ variant for height, rs2871960 on chromosome 3, and read all the variants within a 500 kb distance that have a MAF larger than 0.005 and INFO score larger than 0.5 in the UKBB; there are 3,881 such variants. Then, we perform a GWAS for these variants using the same 305,338 training individuals as used before (Figure S40). Finally, using the resulting GWAS summary statistics and an LD reference from the same set of individuals, we run both SuSiE-RSS and LDpred2-auto (_nojump). We test two different values for L, the maximum number of causal variants in SuSiE-RSS (10 and 100), and three different values for the maximum value of the estimated p (1, 0.01, and 0.001). Per-variant posterior probabilities of being causal (also known as PIPs) are very similar with both methods, especially when restricting p in LDpred2-auto, which is similar to what SuSiE-RSS assumes (more conservative; Figure S41).

Application to height

Here we run three LDpred2-auto models for height, one from the same 305K training UKBB individuals used before (with available height, out of 306K), one based on 100K UKBB individuals (as a random subset of the previous 305K), and one from a large GWAS meta-analysis of 1.6 million individuals of European genetic ancestries.⁴⁸ We first infer the genetic ancestry proportions of individuals included in the meta-analysis using the method proposed in Privé³⁶ and find that 81.9% are from

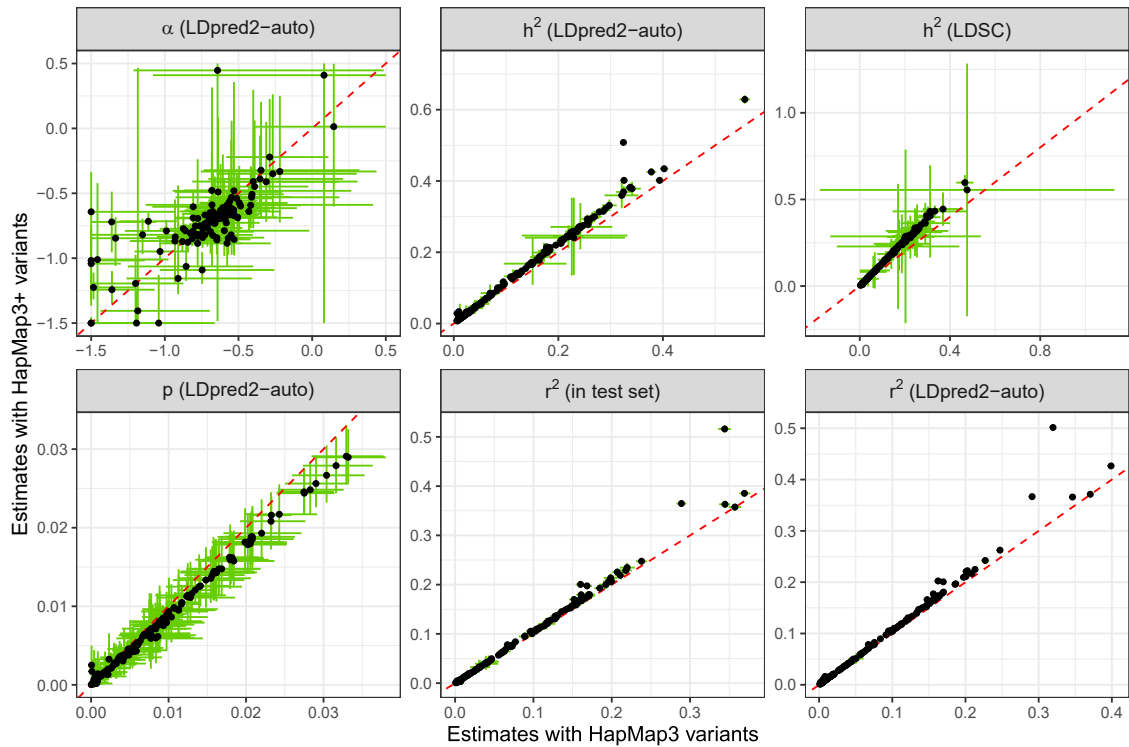


Figure 6. LDpred2-auto estimates for UKBB phenotypes using either the HapMap3 or HapMap3+ sets of variants. Only 154 phenotypes with more than 25 chains kept when using the HapMap3 variants are represented here. Red dashed lines represent the 1:1 line. The 95% CI for the LDpred2-auto estimate (in green) 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% CI for r^2 in the test set is obtained from bootstrap.

Northwestern Europe, 9.5% are from Eastern Europe, 6.5% are from Finland, 1.5% are of Ashkenazi genetic ancestry, 0.3% are from Southwest Europe, and 0.2% are from West Africa. For this set of external GWAS summary statistics, we therefore use the same Northwestern European LD matrix as used in UKBB analyses. Note that we use the HapMap3+ set of 1,444,196 SNPs here; however, for the GWAS meta-analysis, only 1,013,499 SNPs (out of 1,373,020) are overlapping with HapMap3+ and passing quality control.

As expected,⁴⁹ intercepts from LD Score regression increase with sample size: 1.02 (standard error [SE]: 0.008) with $N = 100K$, 1.11 (0.015) with $N = 305K$, and 2.31 (0.068) with $N = 1.6$ million. 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, estimated predictive performance r^2 (from the Gibbs sampler of LDpred2-auto) increase with sample size: 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 $\sim 49\%$, $\sim 67.5\%$, and $\sim 87\%$ of h^2 , respectively. Even though there are 1.6 million individuals in the meta-analysis, the predictive performance corresponds to around 87% of the SNP heritability only; therefore an even larger sample size is required to be able to better predict height. Polygenicity

estimates from LDpred2-auto increase with sample size—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 irrespective of whether `allow_jump_sign` is used or not, which is surprising to us. We also identify 1,753 SNPs with a greater than 95% probability of being causal (fine-mapping), which are spread over the entire genome (Figure S42). As for α estimates from LDpred2-auto, they remain consistent, at -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 from LDpred2-auto results to investigate functional enrichment. We perform this analysis using 50 non-flanking binary annotations from the baselineLD v.2.2 model.⁵⁰ Heritability enrichments that we obtain from LDpred2-auto results are rather modest, ranging from 0.7 to 2.5 with a GWAS sample size of $N = 305,000$, and of slightly smaller magnitude with $N = 100,000$, and of slightly larger magnitude with $N = 1.6$ million (Figure S43).

Application to other external GWAS summary statistics

A description of the eight external GWAS summary statistics used is provided in Table 1; these do not include UKBB individuals. Quality control of these summary statistics is performed as described in Privé, Arbel, et al.³⁴ We run LDpred2-auto using either the HapMap3 or HapMap3+

Table 1. Summary of external GWAS summary statistics used

Trait	GWAS citation	Effective GWAS sample size	# GWAS variants	# matched variants with HapMap3+
Asthma	Demenaïs et al. ⁵⁴	67,341.2	2,001,280	946,092
Breast cancer (BRCA)	Michailidou et al. ⁵⁵	254,862.6	11,792,542	1,411,710
Coronary artery disease (CAD)	Nikpay et al. ⁵⁶	129,014.3	9,455,778	1,325,052
Depression (MDD) (without UKBB)	Wray et al. ⁵⁷	168,040.2	9,874,289	1,314,499
Prostate cancer (PRCA)	Schumacher et al. ⁵⁸	135,316.1	20,370,946	1,411,710
Type 1 diabetes (T1D)	Censin et al. ⁵⁹	13,497.6	8,996,866	1,127,489
Type 2 diabetes (T2D)	Scott et al. ⁶⁰	72,143.0	12,056,346	1,408,283
Vitamin D	Jiang et al. ⁶¹	79,366	2,579,296	1,028,171

These do not include UKBB individuals. Note that some of them may contain a substantial amount of non-European genetic ancestry (e.g., >20% for CAD).

variants, with either the extended or previous model and sampling (via parameter `use_MLE`). Because of the increased mismatch between external GWAS summary statistics and the LD reference we use here (compared to in simulations and UKBB analyses), we also explore multiple values for parameter `coef_shrink`, which is a multiplicative coefficient for shrinking/regularizing off-diagonal elements of the LD matrix in LDpred2-auto. Note that to transform r^2 and h^2 estimates to the liability scale (except for vitamin D, which is a continuous trait), we use the prevalence in the UKBB as the population prevalence, which may be slightly biased.^{51,52} Results are presented in Figure S44. In terms of predictive performance, using the HapMap3+ variants provides equal or better predictive performance compared to using the HapMap3 variants, except for type 1 diabetes (T1D); therefore it seems more useful to use this new set of variants for larger sample sizes. Using the extended model and sampling scheme (with option `use_MLE`) provides equal or better predictive performance except for vitamin D, but proves to be less robust, especially when using `coef_shrink` close to 1 (low or no regularization of the LD matrix). The 3-parameter model also generally provides better predictive performance in the UKBB analyses, where power and robustness are often not an issue (Figure S45). Maximum r^2 are achieved at different LD regularization coefficients `coef_shrink` across phenotypes, reflecting possible substantial mismatches between the GWAS summary statistics and LD reference used. However, results are virtually unchanged when regularizing the LD matrix (“`hm3_plus_regul`”) by multiplying correlations between variants i and j by $\exp(-0.5 \cdot d_{ij})$, where d_{ij} is the distance in cM between the two variants (similar to Wen and Stephens⁵³), which is surprising to us. As for estimates of h^2 and r^2 (inferred from the Gibbs sampler), they tend to increase with smaller values of `coef_shrink`. This is also the case for estimates of p when using `use_MLE = TRUE`. Estimates of α are largely constant but can become very small (capped at -1.5) in the case of robustness issues when using, e.g., `use_MLE = TRUE` and almost no regularization on the LD matrix (i.e., `coef_shrink` close to 1).

Discussion

LDpred2-auto was originally developed for building polygenic scores.²¹ 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 and more locally), polygenicity (and per-variant probabilities of being causal, also known as PIPs, useful for fine-mapping), and the selection-related parameter α . We remind readers that LDpred2 can also be used to infer the uncertainty of individual polygenic scores.^{24,25} We also introduce a 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. Others have proposed to estimate r^2 from a reference genotype set³⁷ or from an additional set of external GWAS summary statistics and LD^{38,39}; here we only use the summary statistics that we input to LDpred2-auto. 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 (assuming an infinitesimal architecture, i.e., $p = 1$) generally providing lower predictive performance than LDpred2-grid (testing different values for parameter p) or LDpred2-auto (estimating p).²¹ We also obtain widespread signatures of negative selection, with most α estimates between -1.0 and -0.3 with a mode at -0.65 , consistent with previous findings from SBayesS, where they find a median S (same as α) of -0.578 (standard deviation [SD]: 0.096).¹⁵ In that paper, they found consistent results with BayesS, although the original BayesS publication reported weaker effects, with a median of -0.37 (SD: 0.11).⁶² In Schoech et al.,¹³ they found an average α of -0.38 (SE: 0.02) with GRM-MAF-LD, but they also noted that LDAK estimates were upward biased by 0.4 (originally centered at -0.25 ,¹² suggesting a likely value would be close to -0.65 , which is what we obtain here).

However, when looking at the heritability enrichments of several functional annotations for height, we obtain much smaller magnitudes with LDpred2-auto than with

stratified LD Score regression (S-LDSC).⁵⁰ For example, Yengo et al.⁴⁸ report fold enrichments of more than 10× for, e.g., coding and conserved variants, while we get less than 2×. This is partly due to LDpred2-auto estimates being more conservative, as they are shrunk toward no enrichment (the prior); however, we do use a very large sample size here, so the prior should not matter much. Another possible reason comes from using variants that capture the causal effects by LD, while these tagging variants may be annotated differently from the causal variants, which could cause functional enrichments to be diluted.⁶³ 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 for use with LDpred2, making it 37% larger to offer better coverage of the genome. Since we used individuals of diverse ancestries for computing the pairwise variant correlations used for constructing this extended set of variants, this new set of variants is beneficial across diverse ancestries (material and methods). Increasing the genome coverage 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% in UKBB analyses 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.⁶⁴ on the UKBB individual-level data while restricting to all variants within a 1 Mb 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. When using external GWAS summary statistics, using HapMap3+ instead of HapMap3 variants was more beneficial for larger sample sizes. Our intuition and conclusion is that using more variants is beneficial when power is sufficient; however, when power is low (e.g., small N , small h^2 , and/or large p), it may actually be detrimental.

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. Therefore, we recommend using the 3-parameter model (with α , when using `use_MLE = TRUE`), but only when power is sufficient and when robustness is not an issue (e.g., without substantial misspecifications such as an ancestry mismatch between the GWAS and LD panels). We also recommend using option `allow_jump_sign = FALSE` for robustness³⁴ and for getting accurate or conservative p estimates. When still obtaining a large p estimate ($> 1\%$) with this option, we recommend rerunning LDpred2-auto without this option

to get a less conservative estimate (cf. simulations). However, LDpred2-auto 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, which causes GWAS effects to be inflated,^{65,66} which causes our estimates to be inflated as well. Second, 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. Third, when using external GWAS summary statistics, it is often beneficial to regularize the LD matrix (via parameter `coef_shrink`, especially when using the extended model and sampling); however, it leads to an increased estimation for, e.g., r^2 and h^2 . Future work could focus on correcting these estimates and also on identifying the best shrinkage regularization coefficient to apply based on, e.g., some distance metric (mismatch) between GWAS summary statistics and the LD reference used. Fourth, because we use a limited set of variants as input for LDpred2, causal variants identified by LDpred2-auto are probably tagging variants that are highly correlated with unobserved causal variants close by. LDpred2-auto can also miss some causal variants when they are poorly tagged by the set of variants used. Future work will focus on scaling LDpred2-auto to using more variants to alleviate this limitation. LDpred2-auto runtime currently depends on the number of causal variants (or, equivalently, the polygenicity). When the polygenicity is close to 0, it runs linearly with the number of variants. When the polygenicity is close to 1, it runs quadratically. However, the polygenicity is almost always lower than 0.1 for human traits and diseases (cf. UKBB results here), which makes running LDpred2-auto efficient (we report runtimes always under 1 h, even with HapMap3+ variants). Currently, the main issue is the large size of the LD matrix that grows quadratically with the number of variants used. For now, to identify causal variants, one can concentrate on small regions of the genome, where LDpred2-auto can be rerun using a much more dense set of variants, as typically done in fine-mapping analyses (cf. [results section local heritability and polygenicity](#)).

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,^{67,68} for genetic correlation estimation,^{11,69–72} for multi-ancestry prediction and inference,^{73–76} as well as extending it to use more variants and to leverage functional annotations.^{63,77,78}

Data and code availability

- All code used for this paper is available at <https://github.com/privefl/paper-infer/tree/main/code>

- Descriptions of UK Biobank phenotypes used here can be found at <https://github.com/privefl/paper-infer/blob/main/phenotype-description.tsv>
- Simulation and UKBB results from this study can be found at <https://github.com/privefl/paper-infer/tree/main/results>

Supplemental information

Supplemental information can be found online at <https://doi.org/10.1016/j.ajhg.2023.10.010>.

Acknowledgments

The authors thank the members of the NCRR/QGG StatGen group and Marc-André Legault for helpful discussions, as well as the reviewers for their useful comments. The 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; the authors thank all the UK Biobank participants for contributing to such useful data for research. 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.

Received: June 14, 2023

Accepted: October 17, 2023

Published: November 8, 2023

Web resources

LD matrices for HapMap3+ variants computed from the Northwestern European UKBB data used in this paper, <https://doi.org/10.6084/m9.figshare.21305061>

Official tutorial on running LDpred2 with R package bigsnpr (for both prediction and inference), <https://privefl.github.io/bigsnpr/articles/LDpred2.html>

roptim: an R package for general purpose optimization with C++, <https://cran.r-project.org/package=roptim>

References

1. Sullivan, P.F., Kendler, K.S., and Neale, M.C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* *60*, 1187–1192.
2. 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. *Nat. Genet.* *42*, 565–569.
3. 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. *Am. J. Hum. Genet.* *105*, 456–476.
4. Trubetskoy, V., Pardiñas, A.F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T.B., Bryois, J., Chen, C.-Y., Dennison, C.A., Hall, L.S., et al. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* *604*, 502–508.
5. Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* *91*, 1011–1021.
6. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* *88*, 76–82.
7. Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., Schizophrenia Working Group of Psychiatric Genomics Consortium, de Candia, T.R., Lee, S.H., Wray, N.R., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* *47*, 1385–1392.
8. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* *47*, 291–295.
9. Benner, C., Spencer, C.C.A., 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*, 1493–1501.
10. Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. *Am. J. Hum. Genet.* *99*, 139–153.
11. Speed, D., and Balding, D.J. (2019). SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* *51*, 277–284.
12. Speed, D., Holmes, J., and Balding, D.J. (2020). Evaluating and improving heritability models using summary statistics. *Nat. Genet.* *52*, 458–462.
13. 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. *Nat. Commun.* *10*, 790–810.
14. 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. *J. R. Stat. Soc. Series B Stat. Methodol.* *82*, 1273–1300.
15. 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. *Nat. Commun.* *12*, 1164–1212.
16. 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 Comput. Biol.* *17*, e1009483.
17. 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*, e3395.

18. Torkamani, A., Wineinger, N.E., and Topol, E.J. (2018). The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* *19*, 581–590.
19. Lambert, S.A., Abraham, G., and Inouye, M. (2019). Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* *28*, R133–R142.
20. 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. *Hum. Genet.* *141*, 1697–1704.
21. Privé, F., Arbel, J., and Vilhjálmsón, B.J. (2020). LDpred2: better, faster, stronger. *Bioinformatics* *36*, 5424–5431.
22. Pain, O., Glanville, K.P., Hagenaaers, S.P., Selzam, S., Fürtjes, A.E., Gaspar, H.A., Coleman, J.R.I., Rimfeld, K., Breen, G., Plomin, R., et al. (2021). Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genet.* *17*, e1009021.
23. Kulm, S., Marderstein, A., Mezey, J., and Elemento, O. (2021). A systematic framework for assessing the clinical impact of polygenic risk scores. Preprint at medRxiv. <https://doi.org/10.1101/2020.04.06.20055574>.
24. Ding, Y., Hou, K., Burch, K.S., Lapinska, S., Privé, F., Vilhjálmsón, B., Sankaraman, S., and Pasaniuc, B. (2022). Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. *Nat. Genet.* *54*, 30–39.
25. Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., Privé, F., Vilhjálmsón, B.J., Olde Loohuis, L.M., and Pasaniuc, B. (2023). Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* *618*, 774–781.
26. 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*, 2781–2787.
27. Bengtsson, H. (2021). A unifying framework for parallel and distributed processing in R using futures. *The R Journal* *13*, 208–291.
28. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. *J. Open Source Softw.* *4*, 1686.
29. 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*, 203–209.
30. Privé, F., Luu, K., Blum, M.G.B., McGrath, J.J., and Vilhjálmsón, B.J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics* *36*, 4449–4457.
31. Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., O’Reilly, P.E., and Vilhjálmsón, B.J. (2022). Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* *109*, 12–23.
32. 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*, 2375–2376.
33. 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 Med. Inform.* *7*, e14325.
34. Privé, F., Arbel, J., Aschard, H., and Vilhjálmsón, B.J. (2022). Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores. *HGG Adv.* *3*, 100136.
35. Privé, F. (2021). Optimal linkage disequilibrium splitting. *Bioinformatics* *38*, 255–256.
36. 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*, 3477–3480.
37. 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. *Genet. Epidemiol.* *41*, 469–480.
38. Pattee, J., and Pan, W. (2020). Penalized regression and model selection methods for polygenic scores on summary statistics. *PLoS Comput. Biol.* *16*, e1008271.
39. Witteveen, M.J., Pedersen, E.M., Meijisen, J., Andersen, M.R., Privé, F., Speed, D., and Vilhjálmsón, B.J. (2022). Publicly available privacy-preserving benchmarks for polygenic prediction. Preprint at bioRxiv. <https://doi.org/10.1101/2022.10.10.510645>.
40. Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved rhat for assessing convergence of mcmc (with discussion). *Bayesian Analysis* *16*, 667–718.
41. Anderson, T.W. (1962). On the distribution of the two-sample Cramer-von Mises criterion. *Ann. Math. Statist.* *33*, 1148–1159.
42. Zou, Y., Carbonetto, P., Wang, G., and Stephens, M. (2022). Fine-mapping from summary data with the “Sum of Single Effect” model. *PLoS Genet.* *18*, e1010299.
43. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* *88*, 294–305.
44. Grotzinger, A.D., Fuente, J.d.I., Privé, F., Nivard, M.G., and Tucker-Drob, E.M. (2023). Pervasive downward bias in estimates of liability-scale heritability in genome-wide association study meta-analysis: a simple solution. *Biol. Psychiatry* *93*, 29–36.
45. 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. *Nat. Metab.* *2*, 1135–1148.
46. 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 Res.* *47*, D1005–D1012.
47. Privé, F., Luu, K., Vilhjálmsón, B.J., and Blum, M.G.B. (2020). Performing highly efficient genome scans for local adaptation with R package pcadapt version 4. *Mol. Biol. Evol.* *37*, 2153–2154.
48. Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliassen, A.U., Jiang, Y., Raghavan, S., et al. (2022). A saturated map of common genetic variants associated with human height. *Nature* *610*, 704–712.
49. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A.P., and Price, A.L. (2018). Mixed-model association for biobank-scale datasets. *Nat. Genet.* *50*, 906–908.
50. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gustav, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shores, N., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* *50*, 621–629.
51. Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N.E. (2017). Comparison of

- sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* *186*, 1026–1034.
52. van Alten, S., Domingue, B.W., Galama, T.J., and Marees, A.T. (2022). Reweighting the UK Biobank to Reflect its Underlying Sampling Population Substantially Reduces Pervasive Selection Bias Due to Volunteering. Preprint at medRxiv. <https://doi.org/10.1101/2022.05.16.22275048>.
 53. Wen, X., and Stephens, M. (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Ann. Appl. Stat.* *4*, 1158–1182.
 54. Demenais, F., Margaritte-Jeannin, P., Barnes, K.C., Cookson, W.O.C., Altmüller, J., Ang, W., Barr, R.G., Beaty, T.H., Becker, A.B., Beilby, J., et al. (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.* *50*, 42–53.
 55. 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*, 92–94.
 56. 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. *Nat. Genet.* *47*, 1121–1130.
 57. 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.F., et al. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* *50*, 668–681.
 58. 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. *Nat. Genet.* *50*, 928–936.
 59. Censin, J.C., 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 Med.* *14*, e1002362.
 60. Scott, R.A., Scott, L.J., Mägi, R., Marullo, L., Gaulton, K.J., Kaakinen, M., Pervjakova, N., Pers, T.H., Johnson, A.D., Eicher, J.D., et al. (2017). An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* *66*, 2888–2902.
 61. Jiang, X., O'Reilly, P.F., Aschard, H., Hsu, Y.-H., Richards, J.B., Dupuis, J., Ingelsson, E., Karasik, D., Pilz, S., Berry, D., et al. (2018). Genome-wide association study in 79,366 european-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin d levels. *Nat. Commun.* *9*, 260–312.
 62. Zeng, J., De Vlaming, R., Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., Yap, C.X., Xue, A., Sidorenko, J., McRae, A.F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* *50*, 746–753.
 63. Zheng, Z., Liu, S., Sidorenko, J., Yengo, L., Turley, P., Ani, A., Wang, R., Nolte, I.M., Snieder, H., Yang, J., et al. (2022). Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. Preprint at bioRxiv. <https://doi.org/10.1101/2022.10.12.510418>.
 64. Privé, F., Aschard, H., and Blum, M.G.B. (2019). Efficient implementation of penalized regression for genetic risk prediction. *Genetics* *212*, 65–74.
 65. 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. *Nat. Commun.* *13*, 660–710.
 66. Herzig, A.F., Noûs, C., Pierre, A.S., and Perdry, H. (2023). A model for co-occurrent assortative mating and vertical cultural transmission and its impact on measures of genetic associations. Preprint at bioRxiv. <https://doi.org/10.1101/2023.04.08.536101>.
 67. Rieger, S., McDaid, A., and Kutalik, Z. (2018). Evaluation and application of summary statistic imputation to discover new height-associated loci. *PLoS Genet.* *14*, e1007371.
 68. Julienne, H., Shi, H., Pasaniuc, B., and Aschard, H. (2019). RAISS: robust and accurate imputation from summary statistics. *Bioinformatics* *35*, 4837–4839.
 69. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., ReproGen Consortium; Psychiatric Genomics Consortium; and Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, and Duncan, L., et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* *47*, 1236–1241.
 70. Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* *101*, 737–751.
 71. 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. *Nat. Commun.* *10*, 2417–2511.
 72. Werme, J., van der Sluis, S., Posthuma, D., and de Leeuw, C.A. (2022). An integrated framework for local genetic correlation analysis. *Nat. Genet.* *54*, 274–282.
 73. Brown, B.C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C.J., Price, A.L., Zaitlen, N., et al. (2016). Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* *99*, 76–88.
 74. 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. *Am. J. Hum. Genet.* *106*, 805–817.
 75. Ruan, Y., Lin, Y.-F., Feng, Y.-C.A., Chen, C.-Y., Lam, M., Guo, Z., Stanley Global Asia Initiatives, He, L., Sawa, A., Martin, A.R., et al. (2022). Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* *54*, 573–580.
 76. 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. *Am. J. Hum. Genet.* *109*, 1388–1404.
 77. Zhang, Q., Privé, F., Vilhjálmsson, B., and Speed, D. (2021). Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.* *12*, 1–9.
 78. 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. *Nat. Commun.* *12*, 1–11.