bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Improved genetic prediction of complex traits from individual-level data or summary statistics

Qianqian Zhang[1], Florian Privé[2], Bjarni Vilhjálmsson[1,2] and Doug Speed[1,3,4,*]

1 Bioinformatics Research Centre (BiRC), Aarhus University, Denmark.
2 National Center for Register-based Research (NCRR), Department of Economics and Business Economics, Aarhus University, Denmark.
3 Quantitative Genetics and Genomics (QGG), Aarhus University, Denmark.
4 Aarhus Institute of Advanced Studies (AIAS), Aarhus University, Denmark.
*Correspondence: doug@qgg.au.dk

ABSTRACT - ***Most existing tools for constructing genetic prediction models begin with the assumption that all genetic variants contribute equally towards the phenotype. However, this represents a suboptimal model for how heritability is distributed across the genome. Therefore, we develop prediction tools that allow the user to specify the heritability model. We compare individual-level data prediction tools using 14 UK Biobank phenotypes; our new tool LDAK-Bolt-Predict outperforms the existing tools Lasso, BLUP, Bolt-LMM and BayesR for all 14 phenotypes. We compare summary statistic prediction tools using 225 UK Biobank phenotypes; our new tool LDAK-BayesR-SS outperforms the existing tools lassosum, sBLUP, LDpred and SBayesR for 223 of the 225 phenotypes. The increase in prediction accuracy from improving the heritability model tends to be substantial. For example, when using LDAK-Bolt-Predict, the proportion of phenotypic variance explained increased by on average 14% (range 5-29%), equivalent to increasing the sample size by a quarter.***

INTRODUCTION

There is a great demand for more accurate genetic prediction models of complex traits. Better models will, for example, improve our ability to investigate genetic architecture, detect genetic overlap between traits and search for gene-environment interactions.[1,2] They will also enable more widespread use of precision medicine, for example, by enabling us to better identify subgroups of individuals with elevated risk of developing a particular disease, or those with lowest chance of responding to a particular medication.[3–7]

Many complex traits have high SNP heritability, which justifies the use of genome-wide, linear, SNP-based prediction models.[8,9] The resulting predictions are called polygenic risk scores (PRS). They take the form $\mathbf{P} = \mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \ldots + \mathbf{X}_m \beta_m$, where m is the total number of SNPs, while $\mathbf{X}_j$ and β_j denote, respectively, the genotypes and estimated effect size for SNP j. Tools for constructing PRS differ in how they estimate the SNP effect sizes. The simplest way to construct a PRS is using effect size estimates from single-predictor regression (classical PRS). However, it is generally better to use an advanced prediction tool that estimates effect sizes using a multi-SNP regression model.[10–14]

Advanced prediction tools start by making prior assumptions regarding how SNPs contribute towards the phenotype. These assumptions include specifying a heritability model, which describes how $E[h^2_j]$, the expected heritability contributed by each SNP, varies across the genome.[15] Almost all existing advanced prediction tools automatically assume that $E[h^2_j]$ is constant. We refer to this as the GCTA Model, because it a core assumption of the software GCTA.[8] In particular, the GCTA Model is assumed by any prediction tool that uses a multi-SNP regression model and assigns the same penalty or prior distribution to standardized SNP effect sizes.[9,16] However, the GCTA Model is suboptimal. Recently, we provided a method for comparing different heritability models using summary statistics from genome-wide association studies.[17] Across tens of complex traits, the model that fit real data best was the BLD-LDAK Model, in which $E[h^2_j]$ depends on minor allele frequency (MAF), local levels of linkage disequilibrium and functional annotations.

In this paper, we construct PRS for a variety of complex traits using eight new prediction tools. The main difference between these and existing tools is that they allow the user to specify the heritability model. We show that for all eight tools, the accuracy of the PRS improves when we switch from the GCTA Model to the BLD-LDAK Model. When individual-level genotype and phenotype data are available, we recommend using our new tool LDAK-Bolt-Predict (a generalized version of the prediction tool contained within the existing software Bolt-LMM[18]). With access only to summary statistics and a reference panel, we recommend using our new tool LDAK-BayesR-SS (a generalized version of the existing prediction tool SBayesR[14]). Both tools are available in our software LDAK[15] (www.ldak.org).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

RESULTS

Overview of Methods.

Figure 1a classifies our eight new prediction tools based on the form of the prior distribution they assign to SNP effect sizes. Our four individual-level tools, big_spLinReg, LDAK-Ridge-Predict, LDAK-Bolt-Predict and LDAK-BayesR-Predict, use the same prior distribution forms as the existing individual-level data tools Lasso,[16] BLUP,[19] Bolt-LMM[18] and BayesR,[11] respectively. Our four new summary statistic tools, LDAK-Lasso-SS, LDAK-Ridge-SS, LDAK-Bolt-SS and LDAK-BayesR-SS, use the same prior distribution forms as the existing summary statistic tools lassosum,[13] sBLUP,[20] LDpred[12] and SBayesR,[14] respectively. Figure 1b illustrates how our new tools incorporate alternative heritability models by allowing the parameters of the effect size prior distribution to vary across SNPs. We provide full details of our new tools in Methods, and scripts for repeating our analyses in Supplementary Note 1.

In total, we construct PRS for 225 phenotypes from the UK Biobank[21,22] (Supplementary Data 1). When using individual-level prediction tools, we restrict to the 14 phenotypes for which we have access to individual-level data. Of these, eight are continuous (body mass index, forced vital capacity, height, impedance, neuroticism score, pulse rate, reaction time and systolic blood pressure), four are binary (college education, ever smoked, hypertension and snorer), and two are ordinal (difficulty falling asleep and preference for evenings). For each phenotype, we have 220,000 distantly-related (pairwise allelic correlations <0.03125), white British individuals, recorded for 628,694 high-quality (information score >0.9), common (MAF >0.01), autosomal, directly-genotyped SNPs. When constructing PRS, we use 200,000 individuals as training samples, and the remaining 20,000 individuals as test samples. When we require a reference panel, we use the genotypes of 20,000 individuals picked at random from the 200,000 training samples. We measure the accuracy of a PRS via R^2, the squared correlation between observed and predicted phenotypes across the 20,000 test samples, and estimate the s.d. of R^2 via jackknifing. For a given phenotype, R^2 is upper-bounded by h^2_{SNP}, the SNP heritability, estimates of which range from 0.07 to 0.61 (Supplementary Table 1). When using summary statistic prediction tools, we construct PRS for all 225 phenotypes, using results released by the Neale Lab. These results come from association studies with average sample size 285k (range 35-361k), and the average h^2_{SNP} is 0.22 (range 0.07-0.63).

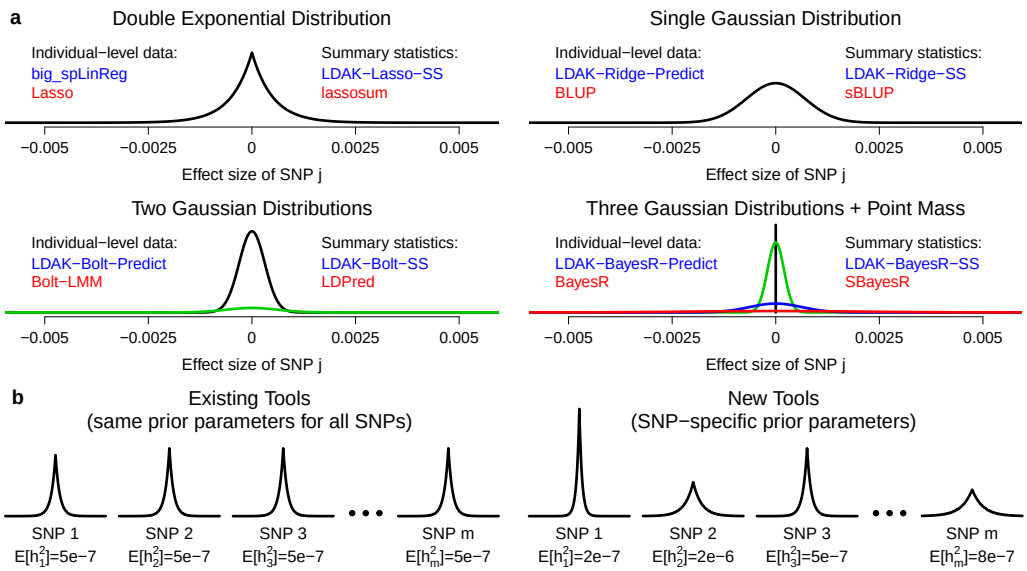

Fig. 1 *Prior distributions for SNP effect sizes.* ***a*** *We divide prediction tools based on the form of the prior distribution they assign to SNP effect sizes, and whether they use individual-level data or summary statistics. For each of our eight new tools (names in blue), there is an existing tool that uses the same prior distribution form (names in red).* ***b*** *Having selected the form of the effect size prior distribution, most existing prediction tools use the same parameters for each SNP. Our new tools, by contrast, use SNP-specific prior distribution parameters. To illustrate this difference, we consider lasso-based prediction tools that assign a double exponential prior distribution to standardized SNP effect sizes. While existing tools might, for example, set the variance of the prior distribution to 5e-7 (so that $E[h^2_j]$=5e-7 for all SNPs), our new tools instead let the variance vary across the genome (allowing $E[h^2_j]$ to be set according to the chosen heritability model).*

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We consider three different heritability models: the GCTA Model assumes $E[h^2_j]$ is constant, the LDAK-Thin Model allows $E[h^2_j]$ to vary based on the MAF of SNP j, while the BLD-LDAK Model allows $E[h^2_j]$ to vary based on the MAF of SNP j, local levels of linkage disequilibrium and functional annotations.[17] Our previous work compared heritability models based on how well they fit real data.[17] Specifically, we measured their performance via the Akaike Information Criterion[23] (AIC), equal to $2K$ - 2logl, where K is the number of parameters in the heritability model and logl is the approximate log likelihood (lower AIC is better). Across the 12 models we considered, AIC was lowest for the BLD-LDAK Model, highest for the the GCTA Model, and intermediate for the LDAK-Thin Model (we reproduce these results in Supplementary Table 2).

Supplementary Fig. 1 shows that when run assuming the GCTA Model, each of our new prediction tools performs at least as well as the corresponding existing tool. For some pairs of tools, the results are almost identical. For example, the PRS constructed using LDAK-Bolt-Predict and LDAK-BayesR-SS assuming the GCTA Model have similar accuracy to those constructed using Bolt-LMM and SBayesR, respectively. However, for other pairs, our tools are superior. For example, the PRS constructed using LDAK-Lasso-SS and LDAK-Ridge-SS assuming the GCTA Model tend to be more accurate than those from lassosum and sBLUP, respectively. We explain the algorithmic innovations that lead to these improvements in Supplementary Note 2. As the aim of this paper is to demonstrate the impact on prediction accuracy of improving the heritability model (not due to algorithmic innovations), for the analyses below, we always use our new tools.

Performance of individual-level data prediction tools.

First we use our four new individual-level data tools to construct PRS for the first 14 UK Biobank phenotypes. When using all 200,000 training samples, the tools take approximately 4 hours (LDAK-Ridge-Predict), 20 hours (LDAK-Bolt-Predict) or 50 hours (big_spLinReg and LDAK-BayesR-Predict), and require 35Gb memory (note that for big_spLinReg, LDAK-Bolt-Predict and LDAK-BayesR-Predict, the runtimes can be reduced substantially by using multiple CPUs).

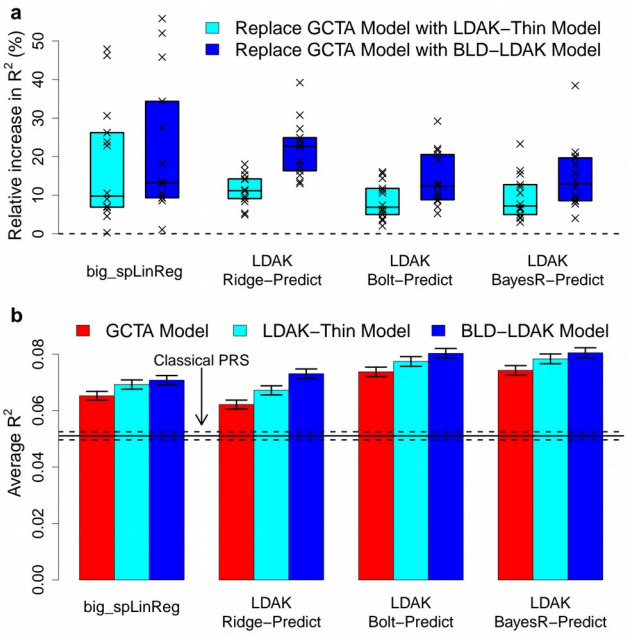

Fig. 2 *Impact of changing the heritability model when using individual-level data.* ***a*** *We use our four new individual-level data prediction tools to construct PRS for the first 14 UK Biobank phenotypes (using all 200,000 training samples). We measure the accuracy of each PRS via R^2, the squared correlation between observed and predicted phenotypes across 20,000 test samples. Points report the percentage increase in R^2 for individual phenotypes when each tool is switched from assuming the GCTA Model to either the LDAK-Thin or BLD-LDAK Model (boxes mark the median and inter-quartile range across the 14 phenotypes).* ***b*** *For the same analysis as* ***a****, bars report R^2 averaged across the 14 phenotypes (vertical segments mark 95% confidence intervals). Colors indicate the assumed heritability model, while blocks indicate the prediction tool. The horizontal lines mark average R^2 for classical PRS and a 95% confidence interval.*

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2 and Supplementary Table 3 show that the accuracy of PRS always increases when we replace the GCTA Model with either the LDAK-Thin or BLD-LDAK Model (i.e., for all four tools and for all 14 phenotypes). For our recommended tool, LDAK-Bolt-Predict, replacing the GCTA Model with the LDAK-Thin Model increases R^2 by on average 9% (s.d. 2%), while replacing the GCTA Model with the BLD-LDAK Model increases R^2 by on average 14% (s.d. 2%). Moreover, when run assuming the BLD-LDAK Model, LDAK-Bolt-Predict outperforms our implementations of the existing tools Lasso, BLUP, Bolt-LMM and BayesR for all 14 phenotypes. We note that the performances of LDAK-Bolt-Predict and LDAK-Bayes-Predict are very similar. For example, when run assuming the BLD-LDAK Model, the tools have average R^2 0.080 and 0.081, respectively (s.d.s 0.001), and each tool produces the most accurate PRS for seven of the 14 phenotypes. Therefore, our decision to recommend LDAK-Bolt-Predict simply reflects its faster runtime.

Figure 3 and Supplementary Table 1 show how the accuracy of PRS constructed using LDAK-Bolt-Predict varies with the number of training samples. We find that the increase we observed when we switched from the GCTA Model to the BLD-LDAK Model is equivalent to increasing the number of training samples by about 24%. The ratio R^2/h^2_{SNP} indicates the accuracy of a PRS relative to the maximum possible accuracy. When we use 200,000 training samples, the PRS achieve between 13% (difficulty falling asleep) and 62% (height) of their potential. The lines of best fit suggest that if we had individual-level data for 400,000 samples, the PRS would explain between 23% and 78% of SNP heritability.

Performance of summary statistic prediction tools

Now we use our four new summary statistic tools to construct PRS for all 225 UK Biobank phenotypes. To construct each PRS takes under two hours (regardless of which tool we use) and requires less than 10Gb memory. Supplementary Fig. 2 and Supplementary Data 1 show that switching from the GCTA Model to the LDAK-Thin Model increases R^2 for between 217 and 225 phenotypes (depending on tool), while switching from the GCTA Model to the BLD-LDAK Model increases R^2 for between 223 and 225 phenotypes. LDAK-BayesR-SS has the highest average R^2 of the four prediction tools, and produces the most accurate PRS for 137 of the 225 phenotypes.

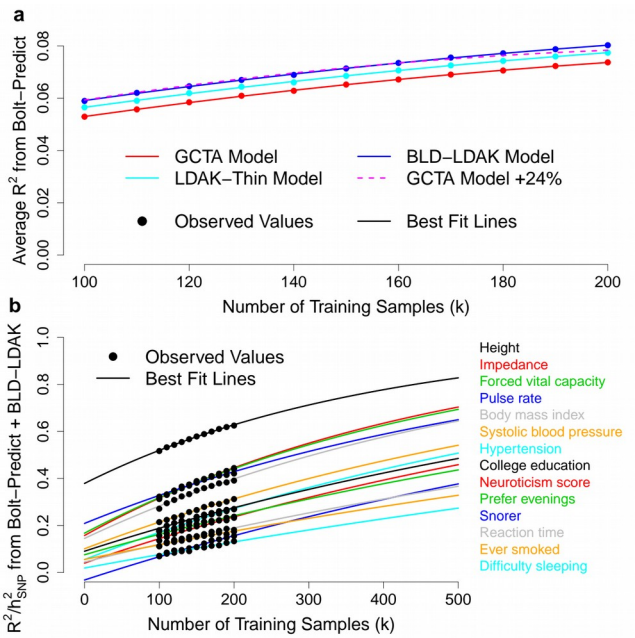

Fig. 3 *Dependency of prediction accuracy on sample size.* ***a*** *We use LDAK-Bolt-Predict to construct PRS for the first 14 phenotypes, varying n, the number of training samples, between 100,000 and 200,000. We measure the accuracy of each PRS via R^2, the squared correlation between observed and predicted phenotypes across 20,000 test samples. Points report R^2 averaged across the 14 phenotypes; colors indicate the assumed heritability model. The lines of best fit are obtained by regressing average R^2 on $a+bn+cn^2$; for the GCTA Model, we use the best fit line to predict average R^2 if the sample size was 24% higher than specified (dashed line).* ***b*** *For the same analysis as* ***a****, points report R^2/h^2_{SNP} for PRS constructed assuming the BLD-LDAK Model, where h^2_{SNP} is the estimated SNP heritability (the maximum possible R^2). The lines of best fit are obtained by regressing R^2/h^2_{SNP} on $1-\exp(a+bn)$.*

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

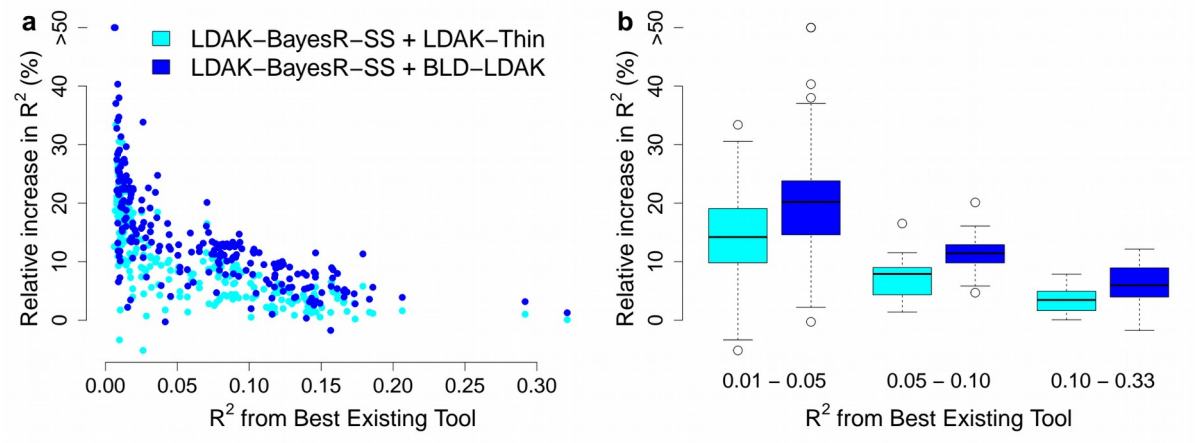

Fig. 4 *Impact of changing the heritability model when using summary statistics.* ***a*** *For each of the 225 phenotypes, we compare PRS constructed using LDAK-BayesR-SS assuming either the LDAK-Thin or BLD-LDAK Model, to PRS constructed using our implementations of the existing tools lassosum, sBLUP, LDpred and SBayesR. We measure the accuracy of PRS via R^2, the squared correlation between observed and predicted phenotypes. The x-axis reports highest R^2 across the four existing tools, while the y-axis reports the percentage increase in R^2 if instead of using the existing tool with highest R^2, we use LDAK-BayesR-SS assuming either the LDAK-Thin or BLD-LDAK Model (improvements above 50% are truncated).* ***b*** *The same as* ***a****, except phenotypes are grouped based on highest R^2 across the four existing tools: 0.01-0.05 (106 phenotypes), 0.05-0.10 (51 phenotypes) or 0.10-0.33 (68 phenotypes). Boxes mark the median increase in R^2 and the inter-quartile range.*

Figure 4 shows that when run assuming the BLD-LDAK Model, LDAK-BayesR-SS outperforms our implementations of the existing tools lassosum, sBLUP, LDpred and SBayesR for 223 of the 225 phenotypes. Compared to the best existing tool, the average increase in R^2 is 14% (s.d. 1%). Consistent with simulations (Supplementary Figs. 3 & 4), we find that the increase tends to be higher for phenotypes with lower R^2. Nonetheless, the average increase remains substantial and significant (P<1e-16 from a one-sided Wald Test) if we consider only the 106 phenotypes with R^2<0.05, only the 51 phenotypes with 0.05<R^2<0.1, or only the 68 phenotypes with R^2>0.1.

Additional Analyses.

For our main analyses, we measured the accuracy of PRS using R^2. Supplementary Fig. 5 shows that improving the heritability model improves accuracy if we instead measure mean absolute error or (for the binary phenotypes) area under the curve. For our main analyses, we used only directly-genotyped SNPs. Supplementary Fig. 6 shows that improving the heritability model also improves the accuracy of PRS when we increase the number of SNPs from 629,000 to 7.5M by including imputed genotypes.

For Supplementary Fig. 7 and Supplementary Table 4, we consider eight diseases: asthma, atrial fibrillation, breast cancer, inflammatory bowel disease, prostate cancer, rheumatoid arthritis, schizophrenia and type 2 diabetes. For each disease, we construct PRS using summary statistics from published studies (average sample size 117,000, range 35,000-215,000) that did not include UK Biobank data,[24–31] then test them using UK Biobank data. Again, we find that for all phenotypes, the accuracy of PRS improves when we replace the GCTA Model with the LDAK-Thin or BLD-LDAK Model. This indicates that the improvements we observed in the main analyses are not an artifact of genotyping errors (as were this the case, we would expect the improvements to disappear when using training and test individuals that have been genotyped independently).

For our main analyses, we used white British individuals from the UK Biobank both to train and test the PRS. For Supplementary Fig. 8 and Supplementary Table 5, we instead test the PRS using UK Biobank individuals of South Asian, African and East Asian ancestry. While absolute accuracy is substantially lower, it remains that PRS constructed assuming the LDAK-Thin or BLD-LDAK Models are more accurate than those constructed assuming the GCTA Model. This indicates that the improvements we observed in the main analyses are not due to population structure (as were this the case, we would expect prediction models constructed assuming the LDAK-Thin or BLD-LDAK Models to perform worse across populations than those constructed assuming the GCTA Model).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DISCUSSION

Most existing prediction tools start with the assumption that each SNP contributes equal heritability.[9] We have instead developed tools that allow the user to specify more realistic heritability models, and shown how these enable the creation of substantially more accurate PRS. Of our eight new tools, we recommend using LDAK-Bolt-Predict when analyzing individual-level data, and LDAK-BayesR-SS when analyzing summary statistics (in both cases, we advise using the tools assuming the BLD-LDAK Model).

When using LDAK-Bolt-Predict, the average increase in R^2 due to changing from the GCTA Model to the BLD-LDAK Model was 14% (s.d. 2%). We showed that this increase is equivalent to increasing the sample size by about a quarter. To provide further perspective, consider that the average increase when switching from using LDAK-Bolt-SS to LDAK-Bolt-Predict (i.e., changing from using summary statistics to individual-level data) was 2% (s.d. 2%), the average increase when switching from using directly-genotyped SNPs to imputed genotypes was 7% (s.d. 2%), the average increase when switching from using LDAK-Ridge-Predict to LDAK-Bolt-Predict (i.e., changing from a single prior distribution for effect sizes to a mixture prior) was 16% (s.d. 2%), while the average increase when switching from classical PRS to LDAK-Ridge-Predict (i.e., changing from classical PRS to the worst-performing advanced prediction tool) was 17% (s.d. 3%).

A strength of our study is that we have considered a variety of complex traits. These include continuous, binary and ordinal phenotypes, that have low, medium and high SNP heritability, and that are both closely and distantly related to diseases. Therefore, the fact that we increased prediction accuracy for almost all of the 225 phenotypes we analyzed, makes us confident that improvements will be observed for many more complex traits. Similarly, our new prediction tools have varying forms of prior distribution for SNP effect sizes. Therefore, the fact that prediction accuracy increased for all tools, indicates that if a new tool is developed with a superior prior distribution form, it is likely that this tool could also be made more accurate by improving the heritability model.

Except for ours, we are not aware of any individual-level data prediction tools that can both analyze biobank-sized datasets (say, over 50,000 samples) and allow the user to specify the heritability model. We are aware of two summary statistic prediction tools where the user can specify the heritability model, AnnoPred[32] and LDpred-funct.[33] AnnoPred is similar to LDAK-Bolt-SS. It assumes that SNP effect sizes have the prior distribution p_0 N(0,σ^2) + (1-p_0) δ_0, then incorporates the chosen heritability model by allowing either σ^2 or p_0 to vary across SNPs.[32] Supplementary Fig. 1 shows that AnnoPred is outperformed by LDAK-Bolt-SS, regardless of whether we assume the BLD-LDAK Model (our recommended model) or the Baseline LD Model (recommended by the authors of AnnoPred). LDpred-funct is similar to LDAK-Ridge-SS. It first estimates effect sizes assuming the the prior distribution N(0,σ^2), where σ^2 varies across SNPs according to the chosen heritability model, then regularizes these estimates via cross-validation.[33] Supplementary Fig. 1 shows that LDpred-funct is outperformed by LDAK-Ridge-SS, regardless of whether we assume the BLD-LDAK Model (our recommended model) or the Baseline LD Model (recommended by the authors of LDpred-funct).

When performing heritability analysis, we previously recommended choosing the heritability model with lowest AIC.[17] We now recommend the same when constructing PRS. Based on average AIC, the BLD-LDAK, LDAK-Thin and GCTA models rank first, second and third, respectively, which matches their order when ranked based on the average accuracy of the corresponding PRS. We additionally construct PRS assuming the GCTA-LDMS-I[34] and Baseline LD Models,[35] those currently recommended by the authors of GCTA[8] and LDSC,[36] respectively. Based on average AIC, these two models rank between the LDAK-Thin and BLD-LDAK Models (Supplementary Table 2), which similarly matches their order when ranked based on the average accuracy of the corresponding PRS (Supplementary Fig. 9).

Although we observed improvement for almost all of the 225 UK Biobank phenotypes, we found that the relative advantage of our new prediction tools was largest for phenotypes with small and modest R^2 (e.g., those with R^2<0.1). This is relevant because, at present, most successful applications of genetic prediction models[37,38] involve PRS with small or modest R^2. For example in psychiatric research, a PRS with R^2≈0.05 was used to show that impulsivity is an endophenotype for attention deficit hyperactivity disorder,[39] a PRS with R^2≈0.07 was used to show that individuals with chronic schizophrenia had higher-than-average genetic liability to schizophrenia,[40] a PRS with R^2≈0.02 was used to identify clinically-defined subtypes of autism that have significantly different genetic liabilities,[41] a PRS with R^2<0.05 was used to demonstrate that risk of developing emotional problems is moderated by an interaction between environmental sensitivity and type of

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

parenting,[42] and a PRS with $R^2 \approx 0.01$ was used to demonstrate that stressful life events and childhood trauma are risk factors for the development of major depressive disorder.[43] Away from psychiatric research, Khera et al.[5] demonstrated the utility of genetic risk prediction for atrial fibrillation, breast cancer, coronary artery disease, inflammatory bowel diseases and type 2 diabetes using PRS with R^2 between 0.02 and 0.04.

We finish by noting that the performance of our new prediction tools will increase as more realistic heritability models are developed. To date, most of the improvement in PRS accuracy has come from increasing sample size, algorithmic innovations or developing more effective forms of prior distribution for SNP effect sizes. Our work indicates that in future, more focus should be placed on improving the heritability model.

METHODS

We begin by explaining our new prediction tools. Note that before running each tool, it is necessary to estimate the expected heritability contributed by each SNP, given the heritability model. Our prediction tools then use these estimates to set the parameters of the effect size prior distribution for each SNP.

Suppose there are n individuals and m SNPs. Let $\mathbf{X}$ denote the matrix of genotypes (size n x m, where column $\mathbf{X}_j$ contains the genotypes for SNP j), and $\mathbf{Y}$ denote the vector of phenotypes (length n). For convenience, the $\mathbf{X}_j$ and $\mathbf{Y}$ are standardized, so that Mean($\mathbf{X}_j$) = Mean($\mathbf{Y}$) = 0 and Var($\mathbf{X}_j$) = Var($\mathbf{Y}$) = 1. We assume that the chi-squared (one degree of freedom) test statistic for SNP j from single-SNP analysis is $S_j = n\, r_j^2 / (1 - r_j^2)$, where $r_j = \mathbf{X}_j\mathbf{Y}/n$ is the correlation between SNP j and the phenotype (this assumes the analysis performed linear regression, but remains a good approximation for S_j computed using logistic regression[44]). We consider prediction tools that use the linear model

$$E[\mathbf{Y}] = \mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \ldots + \mathbf{X}_m \beta_m = \mathbf{X}\boldsymbol{\beta} \qquad (1)$$

where β_j is the effect size for SNP j, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)^T$. Because $\mathbf{X}_j$ and $\mathbf{Y}$ are standardized, the heritability contributed by SNP j is $h^2_j = \beta_j^2$.

Heritability models.

The heritability model takes the form[17]

$$E[h^2_j] = a_{j1}\, \tau_1 + a_{j2}\, \tau_2 + \ldots + a_{jK}\, \tau_K \qquad (2)$$

where the a_{jk} are pre-specified SNP annotations, while the parameters τ_k are estimated from the data.[44] In total, we consider five heritability models (see Supplementary Tables 6 & 7 for formal definitions): the one-parameter GCTA Model assumes $E[h^2_j]$ is constant;[8] the one-parameter LDAK-Thin and 20-parameter GCTA-LDMS-I Model allow $E[h^2_j]$ to vary based on MAF and local levels of linkage disequilibrium;[34,35] the 66-parameter BLD-LDAK and 75-parameter Baseline LD Models allow $E[h^2_j]$ to vary based on MAF, linkage disequilibrium and functional annotations.[17,35] The GCTA Model is the most used model in statistical genetics.[9] The GCTA-LDMS and Baseline LD Models are the recommended models of the authors of GCTA[8] and LDSC,[36] respectively. The BLD-LDAK Model is our preferred model, however, we recommend the LDAK-Thin Model for applications that demand a simple heritability model.[17] We explain the biological intuition behind the GCTA, LDAK-Thin and BLD-LDAK Models in Supplementary Fig. 10.

For a given phenotype, we estimate the τ_k in Equation (2) using SumHer (an existing tool within the LDAK software), which requires summary statistics from single-SNP analysis and a reference panel.[44] SumHer first calculates the expected value of S_j given the heritability model

$$E[S_j] \approx 1 + n\, \Sigma_l\, c^2_{jl}\, (a_{l1}\, \tau_1 + a_{l2}\, \tau_2 + \ldots + a_{lK}\, \tau_K) \qquad (3)$$

where c^2_{jl} is the squared correlation between SNPs j and l in the reference panel, while the summation is across SNPs near SNP j (e.g., within 1cM). Then SumHer estimates the τ_k by regressing the S_j on the $E[S_j]$. For further details see our earlier publications.[17,44] Note that while SumHer can allow for confounding bias (by adding an extra parameter to Equation (3) designed to capture inflation of test statistics due to population structure and familial relatedness), we no longer recommend this feature, nor use it when constructing prediction models.[45] The computational demands of SumHer depend on the complexity of the heritability model; for our

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

analyses, it took approximately 20 minutes when assuming the GCTA or LDAK-Thin Model, and about one hour when assuming the BLD-LDAK Model (each time requiring less than 10Gb memory). As well as estimating τ_k, SumHer also reports e_j, the estimate of $E[h^2_j]$ obtained by replacing the τ_k in Equation (2) with their estimated values.

New prediction tools.

Each of our new tools assumes that the error terms in Equation (1) are normally distributed, so that $\mathbf{Y} \sim N(\mathbf{X\beta}, \sigma^2_e)$, where σ^2_e is the residual variance. They differ in their forms of prior distributions for SNP effect sizes (Fig. 1a). big_spLinReg and LDAK-Lasso-SS use a double exponential distribution, $\beta_j \sim DE(\lambda / E[h^2_j]^{0.5})$. LDAK-Ridge-Predict and LDAK-Ridge-SS use single Gaussian distributions, $\beta_j \sim N(0, E[h^2_j])$ and $\beta_j \sim N(0, vE[h^2_j])$, respectively. LDAK-Bolt-Predict and LDAK-Bolt-SS use a mixture of two Gaussian distributions, $\beta_j \sim p\, N(0, (1-f_2)/p\, E[h^2_j]) + (1-p)\, N(0, f_2/(1-p)\, E[h^2_j])$. LDAK-BayesR-Predict and LDAK-BayesR-SS use a mixture of a point mass at zero and three Gaussian distributions, $\beta_j \sim \pi_1\, \delta_0 + \pi_2\, N(0, sE[h^2_j]/100) + \pi_3\, N(0, sE[h^2_j]/10) + \pi_4\, N(0, sE[h^2_j])$, where $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$ and $s = (\pi_2/100 + \pi_3/10 + \pi_4)^{-1}$. The biological intuition behind the different prior distribution forms is explained in Supplementary Fig. 11. For each tool, we set $E[h^2_j]=e_j$ (the estimate from SumHer), and $\sigma^2_e = 1-\Sigma e_j$. The remaining prior parameters (λ, v, p, f_2, π_1, π_2, π_3 and π_4) are decided using cross-validation, as explained below.

Model fitting using individual-level data.

big_spLinReg is a function within our R package bigstatsr.[46] The original version of the function is described in Prive et al.;[47] the most recent version is the same, except that it allows the user to provide penalty factors that transform the prior from $\beta_j \sim DE(\lambda)$ to $\beta_j \sim DE(\lambda / E[h^2_j]^{0.5})$. In summary, big_spLinReg estimates the β_j using coordinate descent with warm starts.[48,49] Given a value for λ, the β_j are updated iteratively (starting from zero) until they converge. Within each iteration, each β_j within the strong set (the subset of predictors determined most likely to have non-zero effects[49]) is updated once, by replacing its current value with its conditional posterior mode. λ starts at a value sufficiently high that $\beta_j = 0$ for all SNPs, then is gradually lowered to allow an increasing number of SNPs to have non-zero effects. big_spLinReg uses ten-fold cross-validation to decide when to stop reducing λ.

LDAK-Bolt-Predict uses the same algorithm for estimating the β_j and deciding values for p and f_2 as the existing tool Bolt-LMM.[18] In summary, LDAK-Bolt-Predict uses variational Bayes to estimate the β_j. Given values for p and f_2, LDAK-Bolt-Predict updates the β_j iteratively (starting from zero), until the approximate log likelihood converges. Within each iteration, each β_j is updated once, by replacing its current value with its conditional posterior mean. LDAK-Bolt-Predict considers 6 values for p (0.01, 0.02, 0.05, 0.1, 0.2 and 0.5) and three values for f_2 (0.1, 0.3 and 0.5), resulting in 18 possible pairs for p and f_2. First LDAK-Bolt-Predict estimates effect sizes for each of the 18 pairs using data from 90% of samples. Then it identifies which pair results in the best fitting model (based on the mean squared difference between observed and predicted phenotypes for the remaining 10% of samples). Finally, for the best-fitting pair, it re-estimates effect sizes using data from all samples. Note that whereas Bolt-LMM begins by using REML[50] to estimate h^2_{SNP}, then sets $E[h^2_j]=h^2_{SNP}/m$, LDAK-Bolt-Predict does not require this step because it instead sets $E[h^2_j]$ based on estimates from SumHer (see above). Supplementary Fig. 1 shows that the results from LDAK-Bolt-Predict, when run assuming the GCTA Model, are very similar to those from Bolt-LMM.

The prior distribution used by LDAK-Ridge-Predict matches that used by LDAK-Bolt-Predict when $p = f_2 = 0.5$. Therefore, LDAK-Ridge-Predict uses the same algorithm as LDAK-Bolt-Predict, except that it fixes $p = f_2 = 0.5$ and it is no longer necessary to perform the cross-validation step. Supplementary Fig. 1 shows that the results from LDAK-Ridge-Predict, when run assuming the GCTA Model, are very similar to those from the existing tool BLUP[19] (Best Linear Unbiased Prediction).

The existing tool BayesR estimates all parameters using Markov Chain Monte Carlo (MCMC).[11] However, we do not have sufficient resources to apply BayesR to the full UK Biobank data (we estimate that this would require approximately 900Gb and weeks of CPU time). Therefore, LDAK-BayesR-Predict instead uses variational Bayes and cross-validation. The algorithm is the same as for LDAK-Bolt-Predict, except that it is now necessary to select suitable values for π_1, π_2, π_3 and π_4. In total, we consider 35 different combinations: the first is the ridge regression model ($\pi_1, \pi_2, \pi_3, \pi_4$) = (0, 0, 0, 1); the remaining 34 are obtained by allowing each of π_2, π_3 and π_4 to take five values (0, 0.01, 0.05, 0.1, 0.2), with the restrictions $\pi_4 \leq \pi_3 \leq \pi_2$ and $\pi_2 + \pi_3 + \pi_4 > 0$. We

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

investigated omitting the restriction $\pi_4 \leq \pi_3 \leq \pi_2$, in which case there are 125 different triplets, however, we found that while this takes approximately four times longer to run, it did not significantly improve prediction accuracy. In Supplementary Fig. 1, we compare LDAK-BayesR-Predict to BayesR (for computational reasons, we analyze only 20,000 individuals and 99,852 SNPs); the accuracy of LDAK-BayesR-Predict is consistent with that of BayesR, yet our tool is approximately 60 times faster (takes under 20 minutes, compared to 20 hours) and requires 10 times less memory (2Gb instead of 20Gb).

The runtimes reported in the main text (approximately 50, 4, 20 and 50 hours for big_spLinReg, LDAK-Ridge-Predict, LDAK-Bolt-Predict and LDAK-BayesR-Predict, respectively) correspond to using a single CPU. However, for big_spLinReg, LDAK-Bolt-Predict and LDAK-BayesR-Predict, we also provide parallel versions. For LDAK-Bolt-Predict and LDAK-BayesR-Predict, the parallel versions utilize the fact that models corresponding to different parameter choices can be generated independently (i.e., on different CPUs). For big_spLinReg, this is not possible (because the final β_j for one value of λ are used as the starting β_j when λ is reduced), but instead, each of the ten cross-validation runs can be performed independently. Additionally, for the functions LDAK-Ridge-Predict, LDAK-Bolt-Predict and LDAK-BayesR-Predict, LDAK automatically creates a save-point every 10 iterations, so that the job can be restarted if it fails to complete within the allocated time.

Model fitting using summary statistics.

LDAK-Lasso-SS, LDAK-Ridge-SS, LDAK-Bolt-SS and LDAK-BayesR-SS are all contained within a new tool called MegaPRS. To run MegaPRS requires a reference panel and three sets of summary statistics: full summary statistics (computed using all samples), training summary statistics (computed using, say, 90% of samples) and test summary statistics (computed using the remaining samples). In some cases, you will already have (or be able to construct) training and test summary statistics. However, most likely, you will only have full summary statistics, in which case you should first generate pseudo training and test summary statistics (see below).

MegaPRS exploits that, in the absence of individual-level data, $\mathbf{X}_j\mathbf{Y}$ can be recovered from the results of single-SNP regression (as explained above, we assume $S_j = n\, r_j^2 / (1 - r_j^2)$, where n is the sample size and $r_j = \mathbf{X}_j\mathbf{Y}/n$), while $\mathbf{X}_j\mathbf{X}_l$ can be estimated from the reference panel (specifically, MegaPRS uses $\mathbf{X}_j\mathbf{X}_l = n\, c_{jl}$, where c_{jl} is the observed correlation between SNPs j and l in the reference panel). Note that in the equations below, $\mathbf{X}_j$, $\mathbf{Y}$ and n vary depending on context. When using full summary statistics, $\mathbf{X}_j$ and $\mathbf{Y}$ contain genotypes and phenotypes for all samples, and n is the total number of samples. When using training (test) summary statistics, $\mathbf{X}_j$ and $\mathbf{Y}$ contain genotypes and phenotypes for only training (test) samples, and n is the number of training (test) samples.

MegaPRS has three steps. In Step 1, it uses the reference panel to estimate SNP-SNP correlations. In Step 2, it constructs pairs of prediction models, first using the training summary statistics (we refer to these as the "training models"), then using the full summary statistics (the "full models"). In Step 3, it uses the test summary statistics to identify the most accurate of the training models, then reports effect sizes for the corresponding full model. For our analyses, each step took less than 30 minutes and required less than 10Gb memory.

In Step 1, MegaPRS searches the reference panel for local pairs of SNPs with significant c_{jl} (by default, we define local as within 3cM and significant as $P<0.01$ from a two-sided likelihood ratio test that $c_{jl} = 0$). MegaPRS saves the local, significant pairs in a binary file, which requires 8 bytes for each pair (one integer to save the index of the second SNP, one float to save the correlation). For the UK Biobank data, there were 260M local, significant pairs (on average, 413 per SNP), and so the corresponding binary file had size 1.9Gb.

In Step 2, MegaPRS uses the training and full summary statistics to estimates effect sizes for training and full prediction models, respectively. Pairs of training and test models correspond to different prior distribution parameters. MegaPRS constructs 11 pairs of models if running LDAK-Lasso-SS or LDAK-Ridge-SS, 132 pairs of models if running LDAK-Bolt-SS and 84 pairs of models if running LDAK-BayesR-SS (Supplementary Table 8 lists the prior parameters to which these correspond). MegaPRS estimates effect sizes iteratively using variational Bayes. As explained above (in the description of LDAK-Bolt-Predict), the variational Bayes algorithm replaces each current estimate of β_j with its conditional posterior mean. This is possible because for all four tools, the posterior distribution for β_j can be expressed in terms of $\mathbf{X}_j\mathbf{Y}$ and $\mathbf{X}_j\mathbf{X}_l$. For example, if we

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

write the prior distribution for LDAK-Bolt-SS in the form $\beta_j \sim p\ \mathrm{N}(0, \sigma^2_{\mathrm{Big}}) + (1\text{-}p)\ \mathrm{N}(0, \sigma^2_{\mathrm{Small}})$, then the conditional posterior distribution of β_j has the form $p'\mathrm{N}(\mu_{\mathrm{Big}},v_{\mathrm{Big}}) + (1\text{-}p')\mathrm{N}(\mu_{\mathrm{Small}},v_{\mathrm{Small}})$, where

$$
\begin{aligned}
\mu_{\mathrm{Big}} &= \mathbf{X}_j^{\mathrm{T}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_j \beta_j) / (\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j + \sigma^2_{\mathrm{e}}/\sigma^2_{\mathrm{Big}}) \\
v_{\mathrm{Big}} &= \sigma^2_{\mathrm{e}} / (\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j + \sigma^2_{\mathrm{e}}/\sigma^2_{\mathrm{Big}}) \\
\mu_{\mathrm{Small}} &= \mathbf{X}_j^{\mathrm{T}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_j \beta_j) / (\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j + \sigma^2_{\mathrm{e}}/\sigma^2_{\mathrm{Small}}) \qquad (4) \\
v_{\mathrm{Small}} &= \sigma^2_{\mathrm{e}} / (\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j + \sigma^2_{\mathrm{e}}/\sigma^2_{\mathrm{Small}}) \\
p' &= [1 + (1\text{-}p)/p\ (v_{\mathrm{Small}}/v_{\mathrm{Big}})\ (\sigma_{\mathrm{Big}}/\sigma_{\mathrm{Small}}) \exp([\mu^2_{\mathrm{Small}}/v_{\mathrm{Small}} - \mu^2_{\mathrm{Big}}/v_{\mathrm{Big}}]/2)]^{-1}
\end{aligned}
$$

When performing variational Bayes using summary statistics, we found it was not feasible to iterate over all SNPs in the genome. This was due to differences between estimates of $\mathbf{X}_j\mathbf{X}_l$ from the reference panel and their true values (a consequence of the fact that individuals in the reference panel are different to those used in the original association analysis, and because we assume $\mathbf{X}_j\mathbf{X}_l = 0$ for pairs of SNPs that are either distant or not significantly correlated). These differences accumulate over the genome, resulting in poor estimates of $\mathbf{X}_j^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \Sigma_l\ \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_l\beta_l$, and therefore poor estimates of the conditional posterior distribution of β_j. To avoid these problems, MegaPRS uses sliding windows (see Supplementary Fig. 12 for an illustration). By default, MegaPRS iteratively estimates effect sizes for all SNPs in a 1cM window, stopping when the estimated proportion of variance explained by these SNPs converges (changes by less than 1e-5 between iterations). At this point, MegaPRS moves 1/8 cM along the genome, and repeats for the next 1cM window. Within each window, MegaPRS assumes $\sigma^2_{\mathrm{e}} = 1$ (this approximation is reasonable because the expected heritability contributed by a single window will be close to zero). If a window fails to converge within 50 iterations, MegaPRS resets the β_j to their values prior to that window. We found this happened rarely. For example, our main analyses constructed 160,650 full models (across 225 phenotypes, four tools and three heritability models), and for only 990 of these (0.6%) did any of the ~12,000 regions fail to converge.

In Step 3, MegaPRS uses the test summary statistics to measure the accuracy of the training models. If $\boldsymbol{\beta}'$ denotes the vector of estimated effect sizes for a model, MegaPRS calculates $R = \boldsymbol{\beta}'^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} / (n\,\boldsymbol{\beta}'^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}')^{1/2}$, an estimate of the correlation between observed and predicted phenotypes for the individuals used to compute the test summary statistics. MegaPRS then constructs the final model by extracting effect sizes from the full model corresponding to the training model with highest R.

In Supplementary Fig. 1, we compare LDAK-Lasso-SS with lassosum,[13] LDAK-Ridge-SS with sBLUP,[20] LDpred-inf[12] and LDpred-funct,[33] LDAK-Bolt-SS with LDpred2[51] and AnnoPred,[32] and LDAK-BayesR-SS with SBayesR.[14] When we run our new tools assuming the GCTA model, they perform at least as well as the corresponding existing tools.

Pseudo summary statistics.

Given results from single-SNP analysis using n samples, we wish to generate two sets of results, mimicking those we would obtain from first analyzing $n_{\mathrm{A}} < n$ samples, then analyzing the remaining $n_{\mathrm{B}} = n - n_{\mathrm{A}}$ samples. We can reword the task as follows. Let $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_m)^{\mathrm{T}}$ denote the vector of true SNP effect sizes from single-SNP analysis (note that γ_j differs from β_j, because β_j reflects how much SNP j contributes directly to the phenotype, whereas γ_j reflects how much contribution the SNP tags). Given $\mathbf{X}^{\mathrm{T}}\mathbf{Y}/n$, the estimate of $\boldsymbol{\gamma}$ from all n samples, our aim is to generate $\mathbf{X}_{\mathrm{A}}^{\mathrm{T}}\mathbf{Y}_{\mathrm{A}}/n_{\mathrm{A}}$ and $\mathbf{X}_{\mathrm{B}}^{\mathrm{T}}\mathbf{Y}_{\mathrm{B}}/n_{\mathrm{B}}$, estimates of $\boldsymbol{\gamma}$ from n_{A} and n_{B} samples, respectively.

The method we use to generate $\mathbf{X}_{\mathrm{A}}^{\mathrm{T}}\mathbf{Y}_{\mathrm{A}}/n_{\mathrm{A}}$ and $\mathbf{X}_{\mathrm{B}}^{\mathrm{T}}\mathbf{Y}_{\mathrm{B}}/n_{\mathrm{B}}$ is a modified version of that proposed by Zhao et al.[52] First we sample $\mathbf{X}_{\mathrm{A}}^{\mathrm{T}}\mathbf{Y}_{\mathrm{A}}/n_{\mathrm{A}}$ from $\mathrm{N}(\mathbf{X}^{\mathrm{T}}\mathbf{Y}/n,\ n_{\mathrm{B}}/n_{\mathrm{A}}\ \mathbf{V}/n)$, where $\mathbf{V}$ is the variance of $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, then we set $\mathbf{X}_{\mathrm{B}}^{\mathrm{T}}\mathbf{Y}_{\mathrm{B}}/n_{\mathrm{B}} = (\mathbf{X}^{\mathrm{T}}\mathbf{Y} - \mathbf{X}_{\mathrm{A}}^{\mathrm{T}}\mathbf{Y}_{\mathrm{A}})/n_{\mathrm{B}}$. In their method, Zhao et al. restrict to independent SNPs, leading them to derive $\mathbf{V} = \mathbf{I} + \mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}/n^2$, where $\mathbf{I}$ is an identity matrix. However, as we wish to accommodate SNPs in linkage disequilibrium, we instead use $\mathbf{V} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$, as proposed by Zhu and Stephens.[53] If $\mathbf{X}'$ denotes the (standardized) genotypes of the reference panel (size n' x m), then an estimate of $\mathbf{V}$ is $\mathbf{X}'^{\mathrm{T}}\mathbf{X}'n/n'$, and therefore we achieve the desired sampling by setting $\mathbf{X}_{\mathrm{A}}^{\mathrm{T}}\mathbf{Y}_{\mathrm{A}}/n_{\mathrm{A}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}/n + (n_{\mathrm{B}}/n_{\mathrm{A}})^{1/2}\ \mathbf{X}'^{\mathrm{T}}/n'^{1/2}\ \mathbf{g}$, where $\mathbf{g}$ is a vector of length n' with elements drawn from a standard Gaussian distribution.

As explained above, our primary use of pseudo summary statistics is to construct and test training prediction models, in order to decide parameters of the effect size prior distribution. Supplementary Fig. 13 investigates this use of pseudo summary statistics for the first 14 UK Biobank phenotypes and the eight additional diseases (those we used in Supplementary Fig. 7 and Supplementary Table 4). We see that, in general, the estimates of R

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for the training models (measured using pseudo test summary statistics) mirror the estimates of *R* for the corresponding full models (measured using the independent test data), indicating that it is valid to use pseudo summary statistics to decide prior distribution parameters. However, we note two caveats. Firstly, we observe that estimates of *R* can be unreliable when calculated using a reference panel that was also used to create the pseudo summary statistics and/or to construct the prediction models. Therefore, when running MegaPRS using pseudo partial summary statistics, we ensure that the reference panel used in Step 3 is distinct to the reference panel used in Steps 1 & 2. Secondly, we found that estimates of *R* can be unreliable when there are strong effect loci within regions of long-range linkage disequilibrium (e.g, this was an issue for rheumatoid arthritis, where a single SNP within the major histocompatibility complex explains 2% of phenotypic variation). Therefore, when estimating *R* using pseudo summary statistics, we recommend excluding regions of long-range linkage disequilibrium (a list of these are provided at www.ldak.org/high-ld-regions).

Data.

When using our individual-level tools, we constructed prediction models for 14 phenotypes from UK Biobank,[21,22] for which we have access to phenotype and genotype data via Application 21432. These phenotypes are: body mass index (data field 21001), forced vital capacity (3062), height (50), impedance (23106), neuroticism score (20127), pulse rate (102), reaction time (20023), systolic blood pressure (4080), college education (6138), ever smoked (20160), hypertension (20002), snorer (1210), difficulty falling asleep (1200) and preference for evenings (1180). Starting with all 487k UK Biobank individuals, we first filtered based on ancestry (we only kept individuals who were both recorded and inferred through principal component analysis to be white British),[17] then filtered so that no pair remained with allelic correlation >0.0325 (that expected for second cousins). Depending on phenotype, there were between 220,399 and 253,314 individuals (in total, 392,214 unique). From these, we picked 200,000 and 20,000 individuals to use for training and testing prediction models, respectively. For all analyses, we used adjusted phenotypes, obtained by regressing the original phenotypic values on 13 covariates (across all 220,000 training and test individuals). These covariates are age (data field 21022), sex (31), Townsend Deprivation Index (189) and ten principal components (five from the UK Biobank data, five derived from the 1000 Genomes Project[54] data). Supplementary Table 1 reports the estimated proportion of phenotypic variation explained by cryptic relatedness (population structure and familial relatedness); across the 14 phenotypes, it is at most 0.001, and never significant (all P>0.7 from a one-sided likelihood ratio test).

The UK Biobank provides imputed genotype data, but in general we restricted to the 628,694 autosomal SNPs with information score >0.9, MAF >0.01 and present on the UK Biobank Axiom Array (the exception is for Supplementary Fig. 6, where we did not require that SNPs were present on the Axiom Array). We converted dosages to genotypes using a hard-call-threshold of 0.1 (i.e., dosages were rounded to the nearest integer, unless they were between 0.1 and 0.9 or between 1.1 and 1.9, in which case the corresponding genotype was considered missing). After this conversion, on average, 0.1% of genotypes were missing. Note that big_spLinReg does not allow missing values, so when using this tool, we used a hard-call-threshold of 0.5.

When using our summary statistic tools, we constructed prediction models for 225 phenotypes from UK Biobank (Supplementary Data 1), using the August 2018 results from the Neale Lab. In total, the Neale Lab analyzed 4,203 UK Biobank phenotypes, using up to 361,194 British individuals. We downloaded results for the 283 phenotypes that were computed using both sexes and had estimated SNP heritability >0.05 (using details provided in the file ukb31063_h2_topline.02Oct2019.tsv.gz). For each phenotype, we begun by generating pseudo training and test summary statistics corresponding to 90% and 10% of samples, respectively. We subsequently used the pseudo training summary statistics to construct prediction models, and the pseudo test summary statistics to measure their accuracy. Supplementary Fig. 2 confirms that it is possible to both construct and (fairly) measure the accuracy of different prediction models using a single set of summary statistics. Specifically, it shows that for the first 14 phenotypes, estimates of R^2 are similar whether we use data from our own UK Biobank application (for which we have independent training and test data) or use summary statistics from the Neale Lab. Although we downloaded results for 283 phenotypes, in the main text, we restrict to the 225 phenotypes for which it was possible to generate a PRS with R^2>0.01 (using any tool and any heritability model). We made this choice because it is difficult to reliably compare the performance of tools using PRS with very low, and often non-significant, R^2. However, we note that for the 58 phenotypes we rejected, it remained that improving the heritability model increased R^2 on average 93% of time, and that the average improvement in R^2 was 19% (s.d. 3%) when we switched from the best existing tool to LDAK-BayesR-SS assuming the BLD-LDAK Model.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

When we required a reference panel, we used genotypes of 20,000 individuals from the UK Biobank (when requiring multiple reference panels, we ensured 20,000 different individuals were used for each). Note that when using individual-level data prediction tools, for which we use a reference panel to estimate E[h^2_j] given the heritability model, we always picked the 20,000 individuals from the 200,000 training samples, to ensure that there was no overlap between the data used to train and test prediction models.

For the analysis in Supplementary Fig. 7 and Supplementary Table 4, we used results from published studies to construct PRS for eight diseases: asthma,[29] atrial fibrillation,[28] breast cancer,[31] inflammatory bowel disease,[26] prostate cancer,[30] rheumatoid arthritis,[24] schizophrenia[25] and type 2 diabetes.[27] We chose these diseases as they were the ones for which we could find both cases in the UK Biobank and publicly-available summary statistics from a genome-wide association study that did not use UK Biobank data. We excluded SNPs with ambiguous alleles (A&T or C&G) or that were not present in our UK Biobank dataset, after which on average 470,000 SNPs remained (range 191,000 to 559,000).

For the analysis in Supplementary Fig. 8 and Supplementary Table 5, we measured how well prediction models for the first 14 UK Biobank phenotypes (constructed using data from white British individuals) performed for individuals of non-European ancestry. For this, we used principal component analysis to identify 7,057, 2,717 and 1,331 individuals from the UK Biobank whose ancestries were consistent with individuals reported to be South Asian (Indian or Pakistani), African and East Asian (Chinese), respectively.

Sensitivity of MegaPRS to setting choices.

In Supplementary Fig. 14, we test the impact on prediction accuracy of changing the definitions of local and significant when calculating SNP-SNP correlations, the window settings and convergence threshold used when estimating effect sizes, and the choice of reference panel. In general, the impact on accuracy is small. It is largest when we replace the UK Biobank reference panel (20,000 individuals) with genotypes of 489 European individuals from the 1000 Genome Project.[54] In this case, average R^2 reduces by approximately 3% (about two thirds of this is due to reducing the number of individuals, and one third due to replacing UK Biobank genotypes with 1000 Genome Project genotypes).

Other prediction tools.

Supplementary Fig. 1 compares the performance of our new tools with nine existing tools, using the first 14 UK Biobank phenotypes. Note that when using summary statistic tools, we use our own UK Biobank data, rather than summary statistics from the Neale Lab (so there was no need to use pseudo summary statistics). When running BLUP,[19] Bolt-LMM,[18] BayesR,[11] lassosum,[13] sBLUP,[20] LDpred-funct,[33] LDpred2,[51] AnnoPred[32] and SBayesR,[14] we used the default settings of each software (see Supplementary Note 1 for scripts). When it was necessary to select prior parameters (this was the case for lassosum, LDpred2 and AnnoPred, as well as for our four summary statistic tools), we used cross-validation. Similar to above, we constructed models corresponding to different parameter choices using 90% of the training samples (180,000 individuals), then tested these using the remaining 10% of the training samples (20,000 individuals). Having identified the best-performing parameters, we then used these to make the final model (using all training samples). Note that for sBLUP, we found that average R^2 improved if we repeated the analyses excluding regions of long-range linkage disequilibrium,[14,33] while for AnnoPred, we found it was necessary to exclude SNPs from the major histocompatibility complex (otherwise, the software would fail to complete).

When constructing Classical PRS, we used estimates of β_j from single-SNP analysis. We considered six p-value thresholds (P≤5e-8, P≤.0001, P≤0.001, P≤0.01, P≤0.1, all SNPs) and four clumping thresholds (c^2_{jk}≤0.2, c^2_{jk}≤0.5, c^2_{jk}≤0.8, and no clumping). We decided the most-suitable pair of thresholds via cross-validation. Similar to above, we first constructed 18 models using 90% of training samples, then tested these models using the remaining 10% of training samples. We then used the best-performing p-value and clumping thresholds to construct the final model (using summary statistics from all training samples).

DATA AVAILABILITY

Individual-level UK Biobank data can be applied for from www.ukbiobank.ac.uk. Neale Lab summary statistics can be downloaded from www.nealelab.is/uk-biobank. Summary statistics for the eight additional diseases can

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

be downloaded from the websites of the corresponding studies: asthma (www.ebi.ac.uk/gwas/studies/GCST006862), atrial fibrillation (www.ebi.ac.uk/gwas/studies/GCST004296), breast cancer (bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray), inflammatory bowel disease (www.ibdgenetics.org/downloads.html), prostate cancer (http://practical.icr.ac.uk/blog/?page_id=8164), rheumatoid arthritis (plaza.umin.ac.jp/~yokada/datasource/software.htm), schizophrenia (www.med.unc.edu/pgc/download-results/) and type 2 diabetes (diagram-consortium.org/downloads.html).

CODE AVAILABILITY

We provide step-by-step scripts for constructing prediction models in Supplementary Note 1. Our eight new prediction tools are provided within our software packages LDAK (available from www.ldak.org) and bigstatsr (privefl.github.io/bigstatsr). When comparing against existing prediction tools, we additionally used the software packages Bolt-LMM (data.broadinstitute.org/alkesgroup/BOLT-LMM), gctb (cnsgenomics.com/software/gctb), lassosum (github.com/tshmak/lassosum), bigsnpr (privefl.github.io/bigsnpr), LDpred-funct, (github.com/carlaml/Ldpred-funct) and AnnoPred, (github.com/yiminghu/AnnoPred).

ACKNOWLEDGMENTS

The authors thank Dr Veera Rajagopal for testing the LDAK software. F.P. and B.J.V. are supported by the Danish National Research Foundation (Niels Bohr Professorship to Prof. John McGrath) and the Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH (R102-A9118, R155-2014-1724 and R248-2017-2003). B.J.V. is also supported by a Lundbeck Foundation Fellowship (R335-2019-2339). D.S. is supported by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement no. 754513, by Aarhus University Research Foundation (AUFF), by the Independent Research Fund Denmark under Project no. 7025-00094B, and by a Lundbeck Foundation Experiment Grant.

AUTHOR CONTRIBUTIONS

D.S., Q.Z. and F.P. performed the analyses, D.S, F.P. and B.V. wrote the paper.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

1. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. Nat. Protoc. **15**, 2759–2772 (2020).

2. Murray, G. K. et al. Could polygenic risk scores be useful in psychiatry? A review. JAMA Psychiatry 1–10 (2020) doi:10.1001/jamapsychiatry.2020.3042.

3. Speed, D. et al. Describing the genetic architecture of epilepsy through heritability analysis. Brain **137**, 2680–2689 (2014).

4. Niemi, M. E. K. et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. Nature **562**, 268–271 (2018).

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

6. Gibson, G. Going to the negative: genomics for optimized medical prescription. Nat. Rev. Genet. **20**, 1–2 (2019).

7. Mars, N. et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat. Med. **26**, 549–557 (2020).

8. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. **42**, 565–569 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

9. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. Nat. Genet. **49**, 986–992 (2017).

10. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. Gen. Res. **24**, 1550–1557 (2014).

11. Moser, G. et al. Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. PLoS Genet. **11**, e1004969 (2015).

12. Vilhjálmsson, B. et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am. J. Hum. Genet **97**, 576–592 (2015).

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

14. Lloyd-Jones, L. R. et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun. **10**, (2019).

15. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. **91**, 1011–21 (2012).

16. Tibshirani, R. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B **58**, 267–288 (1996).

17. Speed, D., Holmes, J. & Balding, D. J. Evaluating and improving heritability models using summary statistics. Nat. Genet. **52**, 458–462 (2020).

18. Loh, P. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. **47**, 284–290 (2015).

19. Henderson, C. Estimation of genetic parameters. Ann. Math. Stat. **21**, 309–310 (1950).

20. Robinson, M. R. et al. Genetic evidence of assortative mating in humans. Nat. Hum. Behav. **1**, 1–13 (2017).

21. Sudlow, C. et al. UK Biobank : An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. **12**, e1001779 (2015).

22. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. (2018) doi:10.1038/s41586-018-0579-z.

23. Akaike, H. A new look at the statistical model identification. Trans. Autom. Contr **19**, 716–723 (1974).

24. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature **506**, 376–381 (2014).

25. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature **511**, 421–427 (2014).

26. Liu, J. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. **47**, 979–986 (2015).

27. Scott, R. et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes **66**, 2888–2902 (2017).

28. Christophersen, I. E. et al. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. Nat. Genet. **49**, 946–952 (2017).

29. Demenais, F. et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nat. Genet. **50**, 42–50 (2018).

30. Schumacher, F. R. et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat. Genet. **50**, 928–936 (2018).

31. Zhang, H. et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nat. Genet. **52**, 572–581 (2020).

32. Hu, Y. et al. Leveraging functional annotations in genetic risk prediction for human complex diseases. 1–16 (2017).

33. Carla, M. et al. LDpred-funct : incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. bioRxiv 1–32 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265280; this version posted May 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

34. Evans, L. M. et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. Nat. Genet. **50**, (2018).

35. Gazal, S. et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nat. Genet. **49**, 1421–1427 (2017).

36. Bulik-Sullivan, B. et al. LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. Nat. Genet. **47**, 291–295 (2015).

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

38. Gibson, G. On the utilization of polygenic risk scores for therapeutic targeting. PLoS Genet. **15**, 1–14 (2019).

39. Hari Dass, S. A. et al. A biologically-informed polygenic score identifies endophenotypes and clinical conditions associated with the insulin receptor function on specific brain regions. EBioMedicine **42**, 188–202 (2019).

40. Meier, S. M. et al. High loading of polygenic risk in cases with chronic schizophrenia. Mol. Psychiatry **21**, 969–974 (2016).

41. Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. Nat. Genet. **51**, 431–444 (2019).

42. Keers, R. et al. A Genome-Wide Test of the Differential Susceptibility Hypothesis Reveals a Genetic Predictor of Differential Response to Psychological Treatments for Child Anxiety Disorders. Psychother. Psychosom. **85**, 146–158 (2016).

43. Musliner, K. L. et al. Association of Polygenic Liabilities for Major Depression, Bipolar Disorder, and Schizophrenia with Risk for Depression in the Danish Population. JAMA Psychiatry **76**, 516–525 (2019).

44. Speed, D. & Balding, D. SumHer better estimates the SNP heritability of complex traits from summary statistics. Nat. Genet. **51**, 277–284 (2019).

45. Holmes, J., Speed, D. & Balding, D. Summary statistic analyses can mistake confounding bias for heritability. (2019).

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

47. Privé, F., Aschard, H. & Blum, M. G. B. Efficient Implementation of Penalized Regression for Genetic Risk Prediction. **212**, 65–74 (2019).

48. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. **33**, 1–22 (2010).

49. Tibshirani, R. et al. Strong rules for discarding predictors in lasso-type problems. J. R. Stat. Soc. Ser. B (Statistical Methodol. **74**, 245–266 (2012).

50. Corbeil, R. R. & Searle, S. R. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. Technometrics **18**, 31–38 (1976).

51. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. bioRxiv 2020.04.28.066720 (2020) doi:10.1101/2020.04.28.066720.

52. Zhao, Z. et al. Fine-tuning Polygenic Risk Scores with GWAS Summary Statistics. bioRxiv 810713 (2019) doi:10.1101/810713.

53. Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. Ann. Appl. Stat. **11**, 1561–1592 (2017).

54. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature **467**, 1061–1073 (2010).