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. When we improve the heritability model, the proportion of phenotypic variance explained increases by on average 14%, which is equivalent to increasing the sample size by a quarter.
here 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\(^\text{1-5}\). 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\(^\text{3}\).

Many complex traits have high SNP heritability, which justifies the use of genome-wide, linear, SNP-based prediction models\(^\text{8,9}\). The resulting predictions are called polygenic risk scores (PRS). They take the form

\[ \text{PRS} = \sum_{j=1}^{m} \beta_j X_j, \]

where \( X_j \) denote the genotypes and \( \beta_j \) the estimated effect size for SNP \( j \). The variance of the prior distribution to 5\( e^{-7} \). The resulting predictions are called polygenic risk scores (PRS). They take the form

\[ \text{PRS} = \sum_{j=1}^{m} \beta_j X_j, \]

where \( X_j \) denote the genotypes and \( \beta_j \) the estimated effect size for SNP \( j \). 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\(^\text{10-14}\).

Advanced prediction tools start by making prior assumptions regarding how SNPs contribute toward 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\(^\text{15}\). Almost all existing advanced prediction tools automatically assume that \( E[h^2] \) is constant. We refer to this as the GCTA Model, because it is a core assumption of the software GCTA (Genome-wide Complex Trait Analysis)\(^\text{8}\).

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\(^\text{9,16}\). However, the GCTA Model is suboptimal. Recently, we provided a method for comparing different heritability models using summary statistics from genome-wide association studies\(^\text{17}\). 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\(^\text{18}\)). 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\(^\text{14}\)). Both tools are available in our software LDAK\(^\text{15}\) (www.ldak.org).

### 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 (least absolute shrinkage and selection operator)\(^\text{16}\), BLUP (best linear unbiased prediction)\(^\text{19}\), Bolt-LMM\(^\text{18}\) and BayesR\(^\text{11}\), respectively. Our four

![Fig. 1 Prior distributions for SNP effect sizes.](image)

**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).
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\(^{15}\), sBLUP\(^{20}\), LDpred\(^{15}\) 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 distanty-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).

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\(^{21}\) (AIC), equal to 2 \(K - 2\log\ell\), where \(K\) is the number of parameters in the heritability model and \(\log\ell\) 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 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, thePRS 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, thePRS 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 \(4\) h (LDAK-Ridge-Predict), \(20\) h (LDAK-Bolt-Predict) or \(50\) h (big_spLinReg and LDAK-BayesR-Predict), and require \(35\) Gb memory (note that for big_spLinReg, LDAK-Bolt-Predict and LDAK-BayesR-Predict, the runtimes can be reduced substantially by using multiple CPUs).

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-BayesR-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 are obtained by regressing $R^2/h^2_{SNP}$ on $1-\exp(a+bn)$. Source data are provided within the Source Data file.

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 2 h (regardless of which tool we use) and requires <10 Gb 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.

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 ($P < 1 \times 10^{-16}$ from a one-sided Wald Test) if we consider only the 106 phenotypes with $R^2 < 0.1$, or 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

Figure 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 $a + bn + \sigma^2$; 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 $1-\exp(a+bn)$. Source data are provided within the Source Data file.

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. Source data are provided within the Source Data file.
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 7.5 M 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 arthri-
tis, 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 data24–31, 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).

Discussion

Most existing prediction tools start with the assumption that each
SNP contributes equal heritability9. 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 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 dis-
tribution 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.

We are aware of two existing summary statistic prediction tools
where the user can specify the heritability model, AnnoPred32
and LDpred-funct33. AnnoPred is similar to LDAK-Bolt-SS. It
assumes that SNP effect sizes have the prior distribution $p_0 (0, \sigma^2) + (1-p_0) \delta_0$, then incorporates the chosen heritability model
by allowing either $\sigma^2$ or $p_0$ to vary across SNPs23. 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 prior dis-
tribution $N(0,\sigma^2)$, where $\sigma^2$ varies across SNPs according to
the chosen heritability model, then regularizes these estimates via
cross-validation33. 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-
func).

When performing heritability analysis, we previously recom-
mended choosing the heritability model with lowest AIC17. 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 corre-
sponding PRS. We additionally construct PRS assuming the
GCTA-LDMS-f34 and Baseline LD Models35, those currently
recommended by the authors of GCTA8 and LDSC36, respec-
tively. 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
models37,38 involve PRS with small or modest $R^2$. For example in
psychiatric research, a PRS with $R^2 \approx 0.05$ was used to show that
impulsivity is an endophenotype for attention deficit hyper-
activity disorder39, a PRS with $R^2 \approx 0.07$ was used to show that
individuals with chronic schizophrenia had higher-than-average
genetic liability to schizophrenia40, a PRS with $R^2 \approx 0.02$ was used
to identify clinically-defined subtypes of autism that have sig-
nificantly different genetic liabilities41, 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 parenting42, 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 disorder43.
Away from psychiatric research, Khera et al.3 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 predic-
tion 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 X denote the matrix of genotypes (size n x m, where column Xj contains the genotypes for SNP j), and Y denote the vector of phenotypes (length n). For convenience, we standardize X and Y so that Mean(X) = Mean(Y) = 0 and Var(X) = Var(Y) = 1. We assume that the chi-squared (one degree of freedom) test statistic for SNP j from single-SNP analysis is $S_j = \frac{n}{\sigma^2_{p}} (1 - \tau)^2$, where $\tau = X_j Y_j$ is the correlation between the phenotype (which 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

$$E[Y] = X_1 \beta_1 + X_2 \beta_2 + \ldots + X_m \beta_m = X \beta$$

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

Heritability models.

The heritability model takes the form

$$E[h^2_j] = q_1 \tau_1 + q_2 \tau_2 + \ldots + q_k \tau_k$$

where the $q_k$ are pre-specified SNP annotations, while the $\tau_k$ are estimated from the data44. In total, we consider five heritability models (see Supplementary Tables 6 and 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 LDAK-LDMCs Model allows $E[h^2_j]$ to vary based on MAF and local Levels of linkage disequilibrium43,45. 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 annotations47-55. The GCTA Model is the most used model in statistical genetics9. The GCTA-LDMS and Baseline LD Models are the recommended models of the authors of GCTA1 and LDSC49, respectively. The BLD-LDAK Model is our preferred model, as we find the GCTA-LDThin and BLD-LDAK Models in Supplementary Fig. 10 to be the best performers. For a given phenotype, we estimate the $\tau_k$ in Eq. (2) using SumHer (an existing tool within the LDAK software), which requires summary statistics from single-SNP analysis and a reference panel44. SumHer first calculates the expected value of $S_j$ given the heritability model

$$E[S_j] = 1 + 2 \pi \sigma_p^2 \sum q_k \tau_k$$

where $\sigma_p^2$ is the squared correlation between SNPs j and k in the reference panel, while the summation is across SNPs near SNP j (e.g., within 1 cM). Then SumHer estimates the $\tau_k$ by regressing the $S_j$ on the $E[S_j]$. For further details see our earlier publications47,44. Note that while SumHer can allow for confounding bias (by adding an extra parameter to Eq. (3) designed to capture inflation of test statistics due to population structure and familial relatedness), we do not consider this feature, nor use it when constructing prediction models45. The computational demands of SumHer depend on the complexity of the heritability model; for our analyses, it took ~20 min when assuming the GCTA or LDAK-Thin Model, and about 1 h when assuming the BLD-LDAK Model (each time requiring <10 GB memory). As well as estimating $\tau_k$, SumHer also reports $\epsilon_k$, the estimate of $E[h^2_j]$ obtained by replacing the $\tau_k$ in Eq. (2) with their estimated values.

New prediction tools.

Each of our new tools assumes that the error terms in Eq. (1) are normally distributed, so that $Y = \mathcal{N}(X \beta, \sigma^2_\epsilon)$, where $\sigma^2_\epsilon$ is the residual variance. They differ in their prior forms of SNP effect size for SNPs with non-zero effects49 and are used in an iterative process (starting from zero) until they converge. Within each iteration, each $\beta_j$ 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 mean. $\lambda$ 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.snpLinReg uses ten-fold cross-validation to decide when to stop reducing $\lambda$.

LDAK-Bolt-Predict uses the same algorithm for estimating the $\beta_j$ and deciding values for $p$ and $f_j$ as the existing tool Bolt-LMM36,37. In summary, LDAK-Bolt-Predict uses variational Bayes to estimate the $\beta_j$, given values for $p$ and $f_j$. LDAK-Bolt-Predict estimates $E[h^2_j]$ 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 the existing tool BLUP38 (Best Linear Unbiased Prediction).

The existing tool BayesR estimates all parameters using Markov Chain Monte Carlo (MCMC)31. However, we do not have sufficient resources to apply BayesR to the full UK Biobank data (we estimate that this would require ~900 GB and weeks of computational resources). Therefore, for the LDAK-BayesR-Predict tool, we use the heritability model and cross-validation. The algorithm is the same as for LDAK-Bolt-Predict, except that it is now necessary to perform the cross-validation step. Supplementary Fig. 1 shows that the results from LDAK-BayesR-Predict, when run assuming the GCTA Model, are very similar to those from the existing tool BLUP38 (Best Linear Unbiased Prediction).

For each tool, we set $E[h^2_j]$ on the basis of estimates from SumHer (see above). The resulting $E[h^2_j]$ is the correlation between SNP j and the phenotype (as explained above, we assume $S_j = \frac{n}{\sigma^2_p} (1 - \tau)^2$, where $\tau$ is the correlation between the phenotype and $E[h^2_j]$ for SNP j). Then SumHer provides the expected value of $S_j$ given the heritability model.

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 (computed using all available training and test sets): full 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, $X$ and $Y$ can be recovered from the results of single-SNP regression (as explained above, we assume $S_j = \frac{n}{\sigma^2_p} (1 - \tau)^2$, where $\tau$ is the correlation between the phenotype and $E[h^2_j]$ for SNP j). Then SumHer provides the expected value of $S_j$ given the heritability model.

Model fitting using individual-level data.

big.snpLinReg is a function within our R package bigSNP47. 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 $\beta_j \sim \mathcal{N}(0, E[h^2_j])$ to $\beta_j \sim \mathcal{N}(0, E[h^2_j]/\lambda)$. In summary, big.snpLinReg estimates the $\beta_j$ using coordinate descent with warm starts48,49. Given a value for $\lambda$, the $\beta_j$ are updated iteratively (starting from zero) until they converge. Within each iteration, each $\beta_j$ 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 mean. $\lambda$ 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.snpLinReg uses ten-fold cross-validation to decide when to stop reducing $\lambda$. LDAK-Bolt-Predict uses the same algorithm for estimating the $\beta_j$ and deciding values for $p$ and $f_j$ as the existing tool Bolt-LMM36,37. In summary, LDAK-Bolt-Predict uses variational Bayes to estimate the $\beta_j$, given values for $p$ and $f_j$. LDAK-Bolt-Predict estimates $E[h^2_j]$ 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 the existing tool BLUP38 (Best Linear Unbiased Prediction).

The existing tool BayesR estimates all parameters using Markov Chain Monte Carlo (MCMC)31. However, we do not have sufficient resources to apply BayesR to the full UK Biobank data (we estimate that this would require ~900 GB and weeks of computational resources). Therefore, for the LDAK-BayesR-Predict tool, we use the heritability model and cross-validation. The algorithm is the same as for LDAK-Bolt-Predict, except that it is now necessary to perform the cross-validation step. Supplementary Fig. 1 shows that the results from LDAK-BayesR-Predict, when run assuming the GCTA Model, are very similar to those from the existing tool BLUP38 (Best Linear Unbiased Prediction).
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 <30 min and required <10 GB memory.

In Step 1, MegaPRS searches the reference panel for local pairs of SNPs with significant $r_j$ (by default, we define local as within 3 cM and significant as $P < 0.01$ from a two-sided conditional test that $r_j$ = 0). If 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 260 M local significant pairs (on average, 413 per SNP, and so sparse that we used in Supplementary Fig. 1 and Supplementary Table 1). We do not use the raw summary statistics to identify the most accurate of the training models, because we found that, in general, the estimates of $R$ for the training models (measured using pseudo 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 and 2. Secondly, we find 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.lodist.org/high-id-regions).

Data. When using our individual-level tools, we constructed prediction models for 14 phenotypes from UK Biobank21,22, 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 (23106), neuroticism score (2127), 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)2, 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 500,000 individuals to use as 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), sex23, Townsend Deprivation Index (189) and ten principal components (five from the UK Biobank imputed data, five derived from the 1000 Genomes data). In 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 $\text{I}^2 > 0.01$ (using any tool and any heritability model). 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 Neale Lab 2018 supplementary information). For each phenotype, we began 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 we can use our pseudo summary statistics to construct 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 Step 1 we only kept SNPs with $\text{I}^2 > 0.01$. It is 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
When we required a reference panel, we used genotypes of 20,000 individuals (available from www.ldak.org) and bigstatsr (https://priveleged.github.io/bigstatsr), bgsnpr (https://priveleged.github.io/bgsnpr), LDpred-funct (https://github.com/carlamli/LDpred-funct) and AnnoPred (https://github.com/yiminghu/AnnoPred).

Received: 24 March 2021; Accepted: 17 June 2021; Published online: 07 July 2021

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) https://doi.org/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).
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 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. Vilhjálmsdóttir, 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, 5086 (2019).
15. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. 91, 1011–1021 (2012).
16. Tishkarian, 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. Nature https://doi.org/10.1038/s41586-018-0579-z (2018).
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 1561 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. Liu, Y. et al. Leveraging functional annotations in genetic risk prediction for human complex diseases. Plos Comput. Biol. 13, 1–16 (2017).

33. Eva, M. et al. LDpr6-funct: incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. bioRxiv https://www.biorxiv.org/content/10.1101/375337v3 (2020).

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, pages 737–745 (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. EBioMed. 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. 27, 277–284 (2019).

45. Holmes, J., Speed, D. & Balding, D. Summary statistic analyses can mistake confounding bias for heritability. Genet Epidemiol. 43:930–940 (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. Bioinformat 34, 2781–2787 (2018).

47. Privé, F., Aschard, H. & Blum, M. G. B. Efficient implementation of penalized regression for genetic risk prediction. Genetics 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 Method.) 74, 245–266 (2010).

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. & Villjamsson, B. J. LDpred2: better, faster, stronger. bioRxiv 2020.04.28.066720 (2020) https://doi.org/10.1101/2020.04.28.066720.

52. Zhao, Z. et al. Fine-tuning polygenic risk scores with GWAS summary statistics. bioRxiv 810713 (2019) https://doi.org/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).

Acknowledgements
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 John McGrath) and the Lundbeck Foundation Initiative for Integrative Psychiatric Research, JPSYCHI (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 Sklodowska-Curie grant agreement no. 754513, by Aarhus University Research Foundation (AUFF), by the Independent Research Fund Denmark under Project no. 7625-00094B, 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.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-24485-y.

Correspondence and requests for materials should be addressed to D.S.

Peer review information Nature Communications thanks Xia Shen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2021