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 Aarhus Institute of Advanced Studies (AIAS), Aarhus University, Denmark.
4 Quantitative Genetics and Genomics (QGG), Aarhus University, Denmark.
*Correspondence: doug@qgg.au.dk

At present, most tools for constructing genetic prediction models begin with the assumption that all genetic variants contribute equally towards the phenotype. However, this represents a sub-optimal model for how heritability is distributed across the genome. Here we construct prediction models for 14 phenotypes from the UK Biobank (200,000 individuals per phenotype) using four of the most popular prediction tools: lasso, ridge regression, Bolt-LMM and BayesR. When we improve the assumed heritability model, prediction accuracy always improves (i.e., for all four tools and for all 14 phenotypes). When we construct prediction models using individual-level data, the best-performing tool is Bolt-LMM; if we replace its default heritability model with the most realistic model currently available, the average proportion of phenotypic variance explained increases by 19% (s.d. 2), equivalent to increasing the sample size by about a quarter. When we construct prediction models using summary statistics, the best tool depends on the phenotype. Therefore, we develop MegaPRS, a summary statistic prediction tool for constructing lasso, ridge regression, Bolt-LMM and BayesR prediction models, that allows the user to specify the heritability model.

There is a great demand for accurate genetic prediction models of complex traits. In particular, for precision medicine to become a reality, we require models that can reliably predict how likely individuals are to develop different diseases, and how well they will respond to different treatments.[1,2] Many complex traits have substantial SNP heritability,[3,4] indicating that it is theoretically possible to construct clinically-useful linear SNP-based prediction models (polygenic risk scores).[5,6] It is already understood that the accuracy of a polygenic risk score depends on the available sample size.[7,8] Here we show that it also depends on the realism of the assumed heritability model.

The heritability model describes how $E[h^2_j]$, the expected heritability contributed by each SNP, varies across the genome.[9] In human genetics, it is common to 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.[3] In particular, the GCTA Model is assumed by any multi-SNP prediction tool that assigns the same penalty or prior distribution to standardized SNP effect sizes.[4,10] Recently, we provided a method for comparing different heritability models using summary statistics from genome-wide association studies.[11] 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 prediction models for a variety of complex traits using four of the most widely-used prediction tools: lasso, ridge regression, Bolt-LMM and BayesR.[10,12–14] Existing versions of these tools almost exclusively assume the GCTA Model (see Discussion).[13–17] We instead develop versions that allow the user to specify the heritability model. We show that when we switch from the GCTA to the BLD-LDAK Model, prediction accuracy always improves. When individual-level genotype and phenotype data are available, we recommend using our new tool Bolt-Predict to construct Bolt-LMM models. With access only to summary statistics and a reference panel, we recommend using our new tool MegaPRS, which constructs lasso, ridge regression, Bolt-LMM and BayesR models, then selects the most accurate one. Both Bolt-Predict and MegaPRS are available in our software package LDAK (www.ldak.org).

Results

For our main analysis, we construct prediction models for 14 phenotypes from the UK Biobank:[18,19] eight continuous (body mass index, forced vital capacity, height, impedance, neuroticism score, pulse rate, reaction time and systolic blood pressure), four binary (college education, ever smoked, hypertension and snorer) and two 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 prediction models, 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 prediction model via R^2, the squared correlation between observed and predicted phenotypes across the 20,000 test samples (we estimate the s.d. of R^2 via jackknifing, and when summarizing across phenotypes, compute the inverse-variance-weighted average). 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).

We consider three different heritability models: the GCTA Model ($E[h^2_j]$ assumed to be constant), the LDAK-Thin Model ($E[h^2_j]$ depends only on the MAF of SNP j), and the BLD-LDAK Model ($E[h^2_j]$ depends on the MAF of SNP j, local levels of linkage disequilibrium and functional annotations).[11] Our previous work compared heritability models based on how well they fit real data.[11] Specifically, we measured their performance via the Akaike Information Criterion (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). Of 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).

We use four tools to construct prediction models: lasso, ridge regression, Bolt-LMM and BayesR.[10,12–14] These tools differ according to their prior distributions on SNP effect sizes (Table 1). For each tool, we develop two versions, one that uses individual-level data and one that uses summary statistics. All eight versions allow the user to specify the heritability model (i.e., provide $E[h^2_j]$ for each SNP). In Methods we provide algorithmic details, including how each version selects prior parameters via cross-validation (using only the training samples). For the analyses below, we always use our versions of each tool. Extended Data Fig. 1 confirms that when we run our versions assuming the GCTA Model, they perform at least as as well, both in terms of prediction accuracy and computational efficiency, as existing versions (namely the original versions of Bolt-LMM and BayesR,[13,14] both of which use individual-level data, and the summary statistic tools lassosum, sBLUP, LDPred, AnnoPred and SBayesR[15,17,20–23]). Additionally, we develop a new summary statistics tool, MegaPRS, which constructs lasso, ridge regression, Bolt-LMM and BayesR models, then selects the best-performing tool via cross-validation (it does this at the same time as it selects prior parameters for each tool).

Tool	**Prior distribution for SNP effect sizes**
Lasso	$DE(\lambda/\sigma)$
Ridge regression	$N(0,\sigma^2)$
Bolt-LMM	$p\ N(0,\sigma^2_{Big}) + (1-p)\ N(0,\sigma^2_{Small})$
BayesR	$\pi_1\ N(0,\sigma^2) + \pi_2\ N(0,\sigma^2/10) + \pi_3\ N(0,\sigma^2/100) + (1-\pi_1-\pi_2-\pi_3)\ \delta_0$

Table 1 - Prediction Tools | Each tool uses a different prior distribution for effect sizes. DE(a) denotes a double exponential distribution with rate a, N(b,c) denotes a normal distribution with mean b and variance c, while δ_0 denotes a point mass at zero. Note that the variance terms (σ^2, σ^2_{Big} and σ^2_{Small}) are determined by the choice of heritability model, and therefore generally vary across SNPs.

Improving the heritability model improves prediction accuracy.

For Figure 1, we construct prediction models using individual-level data from all 200,000 training samples. Using a single processor, this takes approximately 4 hours (ridge regression), 20 hours (Bolt-LMM) or 50 hours (lasso or BayesR), and requires 35Gb memory (for lasso, Bolt-LMM and BayesR, the runtimes can be reduced substantially by using multiple CPUs). Figure 1a shows that for all four tools and for all 14 phenotypes, R^2 always increases when we replace the GCTA Model with either the LDAK-Thin or BLD-LDAK Model. Supplementary Table 3 provides numerical values; replacing the GCTA Model with the BLD-LDAK model increases R^2 of Lasso models by 1-56% (mean 22%, median 13%), increases R^2 of ridge regression models by 13-39% (mean 22%, median 23%), increases R^2 of Bolt-LMM models by 5-29% (mean 14%, median 12%) and increases R^2 of BayesR models by 0-25% (mean 15%, median 17%). Of the four prediction tools, Bolt-LMM almost always performs best (in particular, it produces the most accurate prediction model for each of the 14 phenotypes). Figure 1b shows that for Bolt-LMM, replacing the GCTA Model with the LDAK-Thin Model increases average R^2 by 13% (SD 2), while replacing the GCTA model with the BLD-LDAK Model increases average R^2 by 19% (SD 2).

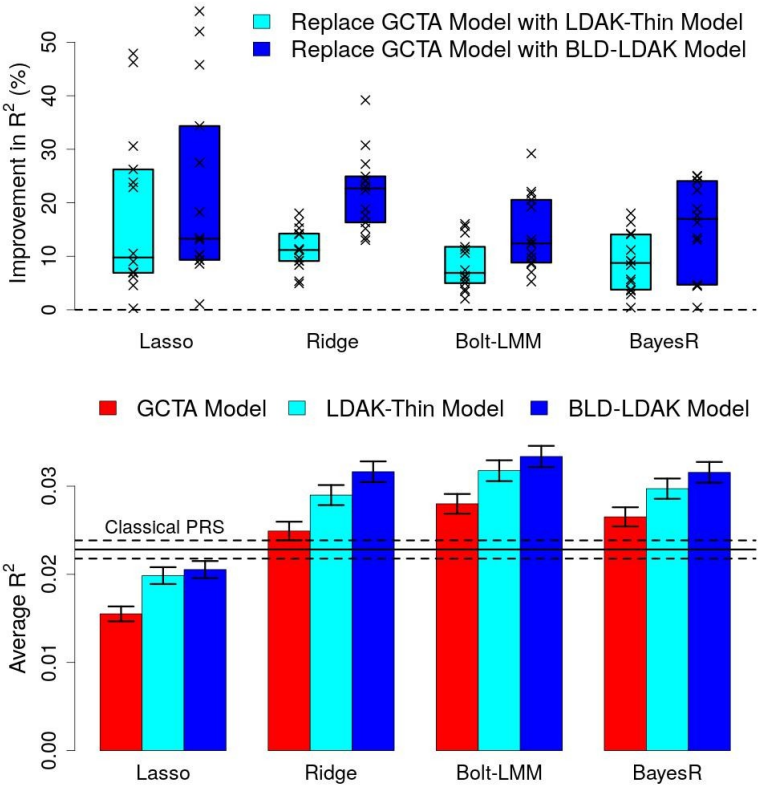


Fig. 1 | Impact of the heritability model when using individual-level data. We construct lasso, ridge regression, Bolt-LMM and BayesR prediction models using individual-level data from all 200,000 training samples, then measure their accuracy via R^2, the squared correlation between observed and predicted phenotypes across 20,000 test samples. **a,** 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,** 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 polygenic risk scores and a 95% confidence interval.

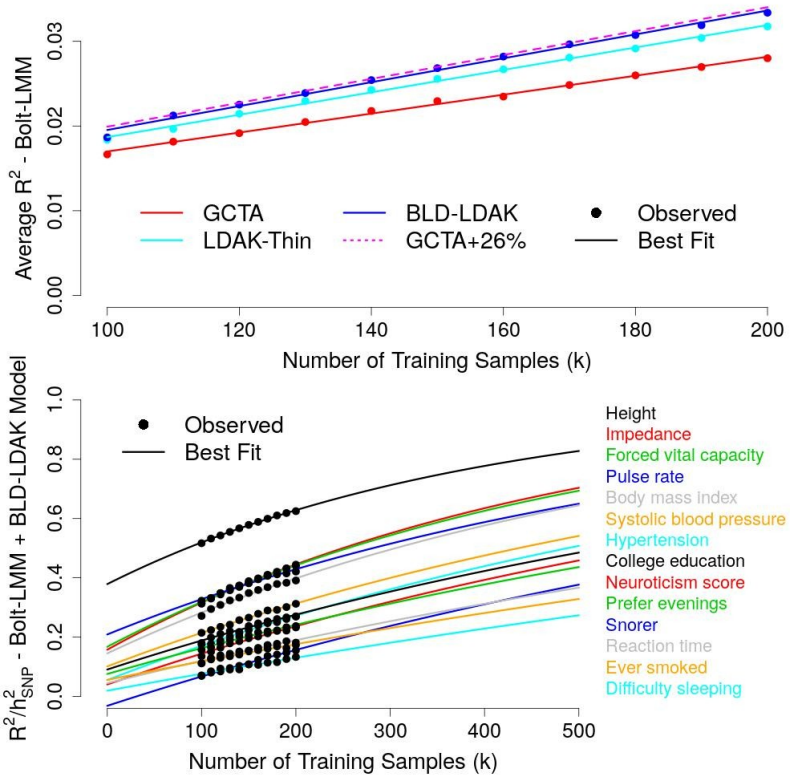


Fig. 2 | Dependency of Bolt-LMM prediction accuracy on sample size. We construct Bolt-LMM prediction models using individual-level data from between 100,000 and 200,000 training samples, then measure their accuracy via R^2, the squared correlation between observed and predicted phenotypes across 20,000 test samples. **a,** 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, where n is the number of training samples; for the GCTA Model, we use the best fit line to predict average R^2 if the sample size was 26% higher than specified (dashed line). **b,** Points report R^2/h^2_{SNP} for individual phenotypes, 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 + b n), where n is the number of training samples.

For Figure 2, we focus on the best-performing tool, Bolt-LMM, and vary the number of training samples. Figure 2a shows that to increase average R^2 by 19% (the increase we observed above when we replaced the GCTA with the BLD-LDAK Model), it is necessary to increase the number of training samples by about 26%. Figures 2b and Supplementary Table 1 report estimates of R^2/h^2_{SNP} for individual phenotypes, which indicate the accuracy of each prediction model relative to the maximum possible. Using 200,000 training samples, the models 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 prediction models would explain between 23% and 78% of SNP heritability.

For Figure 3, we construct prediction models using summary statistics computed from all 200,000 training samples. Using a single processor, this takes under two hours (regardless of which tool we use), and requires less than 10Gb memory. Figure 3a shows that, the same as when using individual-level data, R^2 always increases when we replace the GCTA Model with either the LDAK-Thin or BLD-LDAK Model. Supplementary Table 4 provides numerical values; we now find that the best-performing tool depends on the phenotype, and thus it is advantageous to instead use MegaPRS (which constructs models using all four tools, then selects the best one via cross-validation). Figure 3b shows that for MegaPRS, replacing the GCTA Model with the BLD-LDAK Model increases average R^2 by 17% (SD 2). Extended Data Fig. 2 shows that this improvement is equivalent to increasing the number of training samples by about 25%.

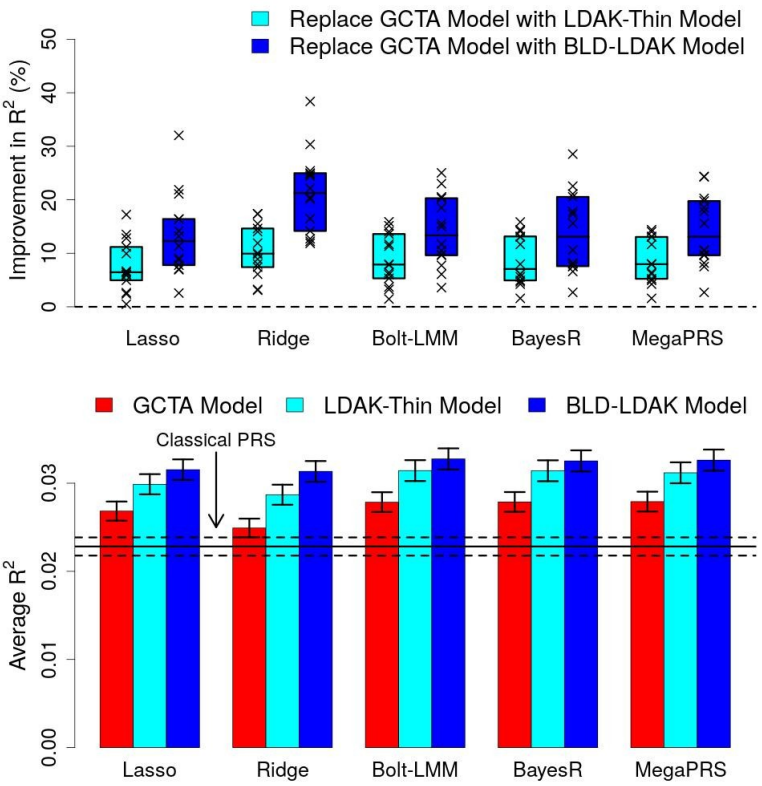


Fig. 3 | Impact of the heritability model when using summary statistics. We construct lasso, ridge regression, Bolt-LMM, BayesR and MegaPRS prediction models using summary statistics from all 200,000 training samples, then measure their accuracy via R^2, the squared correlation between observed and predicted phenotypes across 20,000 test samples. **a,** 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,** 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 polygenic risk scores and a 95% confidence interval.

Extended Data Figs. 3-5 show that improving the heritability model improves prediction performance if we instead measure accuracy using mean absolute error or (for the binary phenotypes) area under the curve, and when we increase the number of SNPs from 629,000 to 7.5M by including imputed genotypes. For Extended Data Fig. 6 and Supplementary Table 5, we consider five additional phenotypes: asthma, breast cancer, prostate cancer, type 2 diabetes and rheumatoid arthritis. For these diseases, there are relatively few cases in UK Biobank (average 7,000, range 1,000-21,000), so we instead train prediction models using summary statistics from published studies[24–28] (average sample size 139,000 individuals, range 58,000-215,000). Again, we see that prediction accuracy improves when we replace the GCTA Model with the LDAK-Thin or BLD-LDAK Model.

Discussion

Most prediction tools start with the assumption that each SNP contributes equal heritability.[4] We have shown that the accuracy of four of the most widely-used prediction tools substantially improves when we assume a more realistic heritability model. As explained in Methods, changing the assumed heritability model typically requires at most a small increase in computational demands (in our analyses, switching to the LDAK-Thin Model required no additional computation, while switching to the BLD-LDAK Model increased the total run time by less than an hour).

A strength of our study is that we have considered a variety of complex traits. These include continuous, binary and categorical phenotypes, that have have low, medium and high SNP heritability, and are both closely and distantly related to diseases. Therefore, the fact that we observed improvement for all the phenotypes we analyzed, makes us confident that similar improvements will be observed for many more complex traits.

When performing heritability analysis, we previously recommended choosing the model with lowest AIC.[11] We now recommend the same when constructing prediction models. Based on average AIC, the BLD-LDAK, LDAK-Thin and GCTA models rank first, second and third, respectively, which matches the ranking of models based on average prediction accuracy. For Extended Data Fig. 7, we additionally consider the GCTA-LDMS-I and Baseline LD models, those currently recommended by the authors of GCTA and LDSC, respectively.[29,30] Based on AIC, these two models rank between the LDAK-Thin and BLD-LDAK Models (Supplementary Table 2), which similarly matches their ranking based on prediction accuracy.

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 summary statistic prediction tools where the user can specify the heritability model, AnnoPred and LDPred-funct.[22,23] AnnoPred is similar to Bolt-LMM. It assumes that SNP effect sizes have the prior distribution $p_0\ N(0,\sigma^2) + (1\text{-}p_0)\ \delta_0$ (this matches the Bolt-LMM prior distribution when $\sigma^2_{Small}=0$), then incorporates the chosen heritability model by allowing either σ^2 or p_0 to vary across SNPs.[22] Extended Data Fig. 1 shows that AnnoPred is outperformed by our summary statistic version of Bolt-LMM, 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 ridge regression. It first estimates effect sizes assuming the the prior distribution $N(0,\sigma^2)$, where σ^2 varies according to the chosen heritability model, then regularizes these estimates via cross-validation.[23] Extended Data Fig. 1 shows that LDPred-funct is outperformed by our summary statistic version of ridge regression, regardless of whether we assume the BLD-LDAK Model (our recommended model) or the Baseline LD Model (recommended by the authors of LDPred-funct).

Aside from demonstrating the advantage of using a more realistic heritability model, we note three additional conclusions from our analyses (evident from Figs. 1 & 3). It is generally better to estimate SNP effect sizes jointly rather than individually (ridge regression, Bolt-LMM and BayesR tend to outperform classical polygenic risk scores). It is generally better to use a multi-distribution prior for effect sizes rather than a single distribution (Bolt-LMM and BayesR tend to outperform ridge regression). It is generally better to create polygenic prediction models rather than sparse ones (this is the reason why our individual-level data version of lasso, which produces models where the majority of effect sizes are zero, tends to perform poorly, and is generally outperformed by our summary statistic version of lasso, whose models have more non-zero effect sizes[15]). Furthermore, Extended Data Figs. 2, 5 show that when feasible, it is generally better to construct models using individual-level data, rather than summary statistics, and to include imputed genotypes, rather than restrict to directly-genotyped SNPs.

Finally, we feel that our analyses provide optimism regarding the prospects of precision medicine. With the advent of population-based biobanks (e.g., Japan Biobank, China Kadoorie Biobank, deCODE and the Estonian Genome Project[31–34]), and the creation of global collaborations for many complex traits and diseases (e.g., the GIANT Consortium and the Psychiatric Genomics Consoritum[35,36]), sample sizes over 100,000 are now relatively common. We have shown that with 200,000 samples we can construct prediction models explaining a substantial proportion of SNP heritability (typically between 25 and 50%). However, our work shows that to speed up the arrival of precision medicine, we should not only continue to increase the sample size, but also strive to create more realistic heritability models.

Methods

Suppose there are n individuals and m SNPs. Let X denote the matrix of genotypes (size n x m, where column X_j contains the genotypes for SNP j), and Y denote the vector of phenotypes (length n). For convenience, we assume the X_j and Y are standardized, so that Mean(X_j) = Mean(Y) = 0 and Var(X_j) = Var(Y) = 1. We assume that the $X^2(1)$ test statistic for SNP j from single-SNP analysis is $S_j = n\, r_j^2 / (1 - r_j^2)$, where $r_j = X_j Y/n$ 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[37]). Each of lasso, ridge regression, Bolt-LMM and BayesR assumes the linear model

$$E[Y] = X_j \beta_1 + X_2 \beta_2 + \ldots + X_m \beta_m = X \beta \qquad (1)$$

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

Heritability models. The heritability model takes the form[11]

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

where the a_{jk} are pre-specified SNP annotations, while the parameters τ_k are estimated from the data.[37] 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;[3] 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;[11,30] 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.[11,29] The GCTA Model is the most common in statistical genetics.[4] The GCTA-LDMS and Baseline LD Models are the recommended models of the authors of GCTA and LDSC, respectively. The BLD-LDAK Model is our preferred model, however, we recommend the LDAK-Thin Model for applications that demand a simple heritability model.[11]

For a given phenotype, we estimate the τ_k in Equation (2) using our software SumHer, which uses summary statistics from single-SNP analysis and a reference panel.[37] SumHer has two steps: first it uses the reference panel to calculate a tagging file (this file contains $E[S_j]$, the expected value of S_j, given the heritability model), then it regresses the summary statistics onto the tagging file (i.e., regresses S_j onto $E[S_j]$). The computational demands of SumHer depend on the complexity of the heritability model; for our 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.

Prediction tools. All four tools assume the error terms in Equation (1) are normally distributed: $Y \sim N(X\beta, \sigma^2_e)$, where σ^2_e is the residual variance. The tools differ in their prior distributions for SNP effect sizes. Lasso[10] uses a double exponential distribution $\beta_j \sim DE(\lambda / E[h^2_j]^{0.5})$. Ridge regression[12] uses a single Gaussian distribution, $\beta_j \sim N(0, E[h^2_j])$. Bolt-LMM[13] uses a mixture of two Gaussian distributions

$$\beta_j \sim p\, N(0, (1\text{-}f_2)/p\, E[h^2_j]) + (1\text{-}p)\, N(0, f_2/(1\text{-}p)\, E[h^2_j])$$

BayesR[14] uses a mixture of three Gaussian distributions and a point mass at zero.

$$\beta_j \sim \pi_1\, N(0, sE[h^2_j]) + \pi_2\, N(0, sE[h^2_j]/10) + \pi_3\, N(0, sE[h^2_j]/100) + (1\text{-}\pi_1\text{-}\pi_2\text{-}\pi_3)\, \delta_0,$$

where $s = 1/(\pi_1 + \pi_2/10 + \pi_3/100)$. 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 priors parameters (λ, p, f_2, π_1, π_2 and π_3) are decided using cross-validation, as explained below.

Model fitting using individual-level data. The function big_spLinReg (available within our R package bigstatsr[16]) fits lasso models using coordinate descent.[38,39] 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 effect[39]) 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 effect. big_spLinReg uses ten-fold cross-validation to decide when to stop reducing λ.

The functions Ridge-Predict, Bolt-Predict and BayesR-Predict (available within our software package LDAK[9]) use variational Bayes. Bolt-Predict uses the same algorithm for estimating β_j and deciding parameter values as the original Bolt-LMM software.[13] Given values for p and f_2, 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. 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 values for (p,f_2). First Bolt-Predict estimates effect sizes for each of the 18 pairs, using data from 90% of training samples. Then it identifies which pair results in the best fitting model (measured as the mean squared difference between observed and predicted phenotypes for the remaining 10% of training samples). Finally, for the best-fitting pair, it re-estimates effect sizes using data from all training samples. Note that the original Bolt-LMM software begins by using REML[40] to estimate the $E[h^2_j]$; Bolt-Predict does not require this step because it instead uses estimates from SumHer. Extended Data Fig. 1 shows that the results from Bolt-Predict, when run assuming the GCTA Model, are very similar to those from the original Bolt-LMM software (both have average R^2 0.028, s.d. 0.0006).

The prior distribution used by ridge regression is equivalent to that used by Bolt-LMM when p=0.5 and $f_2=0.5$. Therefore, Ridge-Predict uses the same algorithm as Bolt-Predict, except it is no longer necessary to perform the cross-validation step. Extended Data Fig. 1 shows that results from Ridge-Predict are very similar to those from the original Bolt-LMM software when the latter is run with p=0.5 and $f_2=0.5$ (both have average R^2 0.00025, s.d. 0.0005). Extended Data Fig. 1 also shows that results from Ridge-Predict are very similar to those from Best Linear Unbiased Prediction[41] (BLUP), to be expected considering BLUP uses the same prior distribution on effect sizes (note that BLUP is much more computationally demanding, because it must compute and eigen-decompose a genome-wide kinship matrix).

The original version of BayesR estimates parameters using Markov Chain Monte Carlo (MCMC).[14] However, we do not have sufficient resources to apply this version to the full UK Biobank data (we estimate that this would require approximately 900Gb and weeks of CPU time). Therefore, BayesR-Predict instead uses variational Bayes. The algorithm is the same as for Bolt-LMM, except it is now necessary to select suitable values for π_1, π_2 and π_3. In total, we consider 35 different triplets: the first is the ridge regression model $(\pi_1, \pi_2, \pi_3)=(0,0,1)$; the remaining 34 are obtained by allowing five values (0, 0.001, 0.005, 0.01, 0.02) for each fraction, with the restrictions $\pi_3 \geq \pi_2 \geq \pi_1$ and $\pi_1 + \pi_2 + \pi_3 > 0$. We investigated omitting the restriction $\pi_3 \geq \pi_2 \geq \pi_1$, 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 Extended Data Fig. 1, we compare our implementation of BayesR to the original version (for computational reasons, we analyze only 20,000 individuals and 122,000 SNPs); the accuracy of our implementation is consistent with that of the original version (average R^2 0.00034 vs 0.00033, SD 0.0001), yet our implementation 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, Ridge-Predict, Bolt-Predict and BayesR-Predict, respectively) correspond to using a single CPU. However, for big_spLinReg, Bolt-Predict and BayesR-Predict, we also provide parallel versions. For Bolt-Predict and BayesR-Predict, the parallel versions use 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 Ridge-Predict, Bolt-

Predict and 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. Our tool MegaPRS (available within our software package LDAK[9]) has three steps. Suppose we have full summary statistics (computed using all training samples), partial summary statistics (computed using, say, 90% of training samples), and a reference panel. MegaPRS begins by using the reference panel to calculate SNP-SNP correlations. Next it constructs a variety of lasso, ridge regression, Bolt-LMM and BayesR models, first using the partial summary statistics (we refer to these as the "partial models"), then using the full summary statistics (the "full models"). Finally, MegaPRS identifies the most accurate of the partials models, based on how well each predicts phenotypes for the individuals excluded when calculating the partial summary statistics, then reports effect sizes for the corresponding full model. We provide full details for each step below. For our analyses, each step took less than 30 minutes, and required less than 10Gb memory. Note that our other summary statistic tools use the same algorithm as MegaPRS, except that in the second step they consider only one type of model (e.g., our summary statistic version of lasso constructs only lasso models). In Extended Data Fig. 1, we confirm that our summary versions of lasso, ridge regression, Bolt-LMM and BayesR perform at least as well as existing summary statistic software (specifically, we compare our lasso with lassosum,[15] our ridge regression with sBLUP, LDPred-inf and LDPred-funct[20,21,23], our Bolt-LMM with LDPred2 and AnnoPred,[22,42] and our BayesR with SBayesR[17]).

MegaPRS exploits that, in the absence of individual-level data, X_jY 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 $r_j = X_jY/n$), while X_jX_k can be estimated from the reference panel (specifically, MegaPRS uses $X_jX_k = n\, c_{jk}^2$, where c_{jk} is the observed correlation between SNPs j and k in the reference panel). In the first step, MegaPRS searches the reference panel for local pairs of SNPs with significant c_{jk} (by default, we consider pairs within 3cM and define significant as P<0.01). MegaPRS saves the 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 significant pairs (on average, 413 per SNP), and so the corresponding binary file had size 1.9Gb.

In the second step, MegaPRS estimates effect sizes for 325 models (100 lasso models, 11 ridge regression models, 131 Bolt-LMM models and 83 BayesR models; see Supplementary Table 8 for full details). Like our individual-level data tools, MegaPRS uses either coordinate descent (lasso models), or variational Bayes (ridge regression, Bolt-LMM and BayesR models). This is possible because for all four prior distributions, the posterior distribution for β_j can be expressed in terms of X_jY and X_jX_k. For example, when constructing Bolt-LMM Models, the conditional posterior distribution of β_j is $p'N(\mu_{Big},v_{Big}) + (1-p')N(\mu_{Small},v_{Small})$, where

$$\mu_{Big} = X_j^T (Y - X\beta + X_j \beta_j) / (X_j^T X_j\, \sigma^2_e/\sigma^2_{Big}+1) \qquad v_{Big} = \sigma^2_e / (X_j^T X_j\, \sigma^2_e/\sigma^2_{Big}+1)$$

$$\mu_{Small} = X_j^T (Y - X\beta + X_j \beta_j) / (X_j^T X_j\, \sigma^2_e/\sigma^2_{Small}+1) \qquad v_{Small} = \sigma^2_e / (X_j^T X_j\, \sigma^2_e/\sigma^2_{Small}+1)$$

$$p' = [1 + (1-p)/p\ \sigma^2_{Big} / \sigma^2_{Small} \exp([\mu^2_{Small/}v_{Small} - \mu^2_{Big/}v_{Big}]/2)]^{-1}$$

When performing coordinate descent or 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 X_jX_k 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 $X_jX_k=0$ for pairs of SNPs that are either distant or not significantly correlated). These differences accumulate over the genome, resulting in poor estimates of $X_j^TX\beta = \Sigma_k\, X_j^T\, X_k\beta_k$, and therefore poor estimates of the conditional posterior distribution of β_j. To avoid these problems, MegaPRS uses sliding windows (see Extended Data 8 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 0.00001 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 $\sigma^2_e = 1$ (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 very rarely. For example, our main analysis constructed 13,650 models (14 phenotypes x three heritability models x 325 prior distributions), and not once did a region fail to converge.

In the third step, MegaPRS measures the accuracy of the partial models via R, the correlation between observed and predicted phenotypes for the individuals excluded when computing the partial summary statistics. If X_E and Y_E denote the standardized genotypes and phenotypes of the n_E excluded individuals, respectively, then $R = \beta'^T X_E^T Y_E / (n_E \beta'^T X_E^T X_E \beta')^{1/2}$, where β' is the vector of estimated effect sizes. Note that if we do not have access to individual-level data for the excluded individuals, we can instead recover $X_E^T Y_E$ from summary statistics (calculated across the excluded individuals) and estimate $X_E^T X_E$ from a reference panel.

When analyzing the UK Biobank phenotypes, it is straightforward to compute partial summary statistics (because we have access to individual-level data, we are able to repeat the single-SNP analysis using only 90% of training samples). This is not the case when analyzing asthma, breast cancer, prostate cancer, rheumatoid arthritis and type 2 diabetes, for which we use summary statistics from published studies. Therefore, we instead create "pseudo" partial summary statistics.[43] Let $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_m)^T$ denote the vector of true SNP effect sizes from single-SNP analysis (note that γ_j will usually differ from β_j, because β_j reflects how much SNP j contributes directly to the phenotype, whereas γ_j reflects how much contribution it tags). Given $X^T Y/n$, the estimate of γ from all n training samples, our aim is to generate $X_A^T Y_A/n_A$ an estimate of γ from n_A samples (where $n_A = n - n_E$). Our approach is similar to that of Zhao et al.,[43] who propose sampling $X_A^T Y_A/n_A$ from $N(X^T Y/n, n_E/n_A\, V/n)$, where V is the variance of $X^T Y$. However, while Zhao et al. restrict to independent SNPs, and subsequently derive $V = I + X^T Y Y^T X/n^2$, where I is an identity matrix, we instead use $V = X^T X$, as proposed by Zhu and Stephens.[44] If the matrix X' denotes the genotypes of the reference panel (size n' x m), we can achieve this sampling by setting $X_A^T Y_A/n_A = X^T Y/n + (n_E/n_A)^{1/2}\, X'^T/n'^{1/2}\, G$, where G is a vector of length n' with elements drawn from a standard Gaussian distribution. Additionally, we calculate $(X_E^T Y_E)/n_E = (X^T Y - X_A^T Y_A)/n_E$, the complementary estimate of γ, which we use when measuring the accuracy of the partial models (as explained above).

In Extended Data Fig. 9 we show that for the UK Biobank phenotypes, the estimated accuracy of the partial models is similar whether we construct the models using actual partial summary statistics (then calculate R directly from X_E and Y_E), or construct them using pseudo partial summary statistics (then calculate R by recovering $X_E^T Y_E$ from the complementary estimate of γ and estimating $X_E^T X_E$ from a reference panel). We note two caveats. Firstly, we observed that the estimate of R can be unreliable when the reference panel used to estimate $X_E^T X_E$ is used also to create the pseudo partial summary statistics or by MegaPRS to estimate SNP-SNP correlations. Therefore, when running MegaPRS using pseudo partial summary statistics, we recommend using three independent reference panels (either by sourcing two extra reference panels, or by dividing the original reference panel into three). Secondly, we found that estimates of R can be unreliable when there are very strong effect loci within regions of long-range linkage disequilibrium (this was only an issue for rheumatoid arthritis, where a single SNP within the major histocompatibility complex explains 2% of phenotypic variation). Therefore, when estimating R, we recommend excluding a region of long-range linkage disequilibrium if it contains a SNP that explains at least 1% of phenotypic variation (see URLs for lists of regions) .

Data. We accessed UK Biobank data via Project 21432. The 14 phenotypes we analyze are the same as for our previous study: 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 487k 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), 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 which we picked 200,000 and 20,000 to use as training and test samples, respectively.

The imputed data contains 97M SNPs, but in general we used only the 628,694 autosomal SNPs with info score >0.9, MAF >0.01 and present on the UK Biobank Axiom Array (the exception is for Extended Data Fig. 5, 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 for these analyses, we used a hard-call-threshold of 0.5. For all analyses, we used adjusted phenotypes, obtained by regressing the original phenotypic values on 13 covariates: age (data field 21022), sex (31), Townsend Deprivation Index (189) and ten principal components. To obtain summary statistics, we performed single-SNP analysis using linear regression (regardless of whether the phenotype was continuous, binary or categorical). When we required a reference panel, we used genotype data from 20,000 UK Biobank individuals, randomly picked from the 86k common to the training samples of each phenotype (when analyzing asthma, breast cancer, prostate cancer, rheumatoid arthritis and type 2 diabetes, we constructed two extra reference panels, each containing an additional 20,000 UK Biobank individuals).

We downloaded summary statistics for asthma,[26] breast cancer,[28] prostate cancer,[27] rheumatoid arthritis[24] and type 2 diabetes[25] from the websites of the corresponding studies. We chose these diseases as they were the ones for which we could find at least 1000 cases in the UK Biobank and summary statistics from a genome-wide association study of at least 50,000 samples (that did not use UK Biobank data). We excluded SNPs that had ambiguous alleles (A&T or C&G) or were not present in our UK Biobank dataset, after which on average 470,000 SNPs remained (range 191,000 to 559,000). For more details, see Supplementary Table 5.

Sensitivity of MegaPRS to setting choices. In Extended Data Fig. 10, 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 fairly small. It is largest when we replace the UK Biobank reference panel (20,000 individuals) with genotypes of European individuals from the 1000 Genome Project[45] (489 individuals). In this case, average R^2 reduces by approximately 3% (about two thirds of this is due to reducing the number of individuals, one third due to replacing UK Biobank genotypes with 1000 Genome Project genotypes).

Other tools. When running Bolt-LMM, BayesR, lassosum, sBLUP, LDPred-funct, LDPred2, AnnoPred and SBayesR (Extended Data Fig. 1), we generally used the recommended settings of each software (see the Supplementary Note for explicit scripts). For lassosum, LDPred2 and AnnoPred, we selected prior parameters via cross-validation, using the same approach as when running our summary statistic tools (i.e., by constructing partial and full models, as described above). For sBLUP, we found that average R^2 improved if we repeated the analyses excluding high linkage disequilibrium regions.[17,23] For AnnoPred, we found it was necessary to exclude SNPs from the major histocompatibility region (otherwise, the software would fail to complete). Figures 1 & 3 include results from Classical polygenic risk scores. For these, we used the estimates of β_j from single-SNP analysis; we considered six p-value thresholds ($P \leq 5e\text{-}8$, $P \leq .0001$, $P \leq 0.001$, $P \leq 0.01$, $P \leq 0.1$, all SNPs) and four clumping thresholds ($s^2_{jk} \leq 0.2$, $s^2_{jk} \leq 0.5$, $s^2_{jk} \leq 0.8$, and no clumping), reporting results from the pair of thresholds that resulted in highest R^2.

URLs

LDAK, http://www.ldak.org; big_spLinReg and LDPred2, https://privefl.github.io/bigsnpr; Bolt-LMM, https://data.broadinstitute.org/alkesgroup/BOLT-LMM; BayesR, sBLUP and SBayesR, https://cnsgenomics.com/software/gctb; LDPred-funct, https://github.com/carlaml/Ldpred-funct; AnnoPred, https://github.com/yiminghu/AnnoPred; UK Biobank, https://www.ukbiobank.ac.uk. High-LD regions, https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD)

Acknowledgements

D.S. is funded 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 and B.V. wrote the paper.

Competing interests
The authors declare no competing interests.

Data availability
UK Biobank data can be applied for from the UK Biobank website (see URLs).

Code availability
We provide step-by-step scripts for constructing prediction models in the Supplementary Note.

References

1. Hodson, R. Precision medicine. *Nature* **537**, S49–S49 (2016).

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

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

4. 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).

5. Wray, N., Yang, J., Goddard, M. & Visscher, P. The genetic interpretation of area under the ROC curve in genetic profiling. *PLoS Genet.* **6**, e1000864 (2010).

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

7. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9**, e03348 (2013).

8. Palla, L. & Dudbridge, F. A Fast Method that Uses Polygenic Scores to Estimate the Variance Explained by Genome-wide Marker Panels and the Proportion of Variants Affecting a Trait. *Am. J. Hum. Genet.* **97**, 250–259 (2015).

9. Speed, D., Hemani, G., Johnson, M. & Balding, D. Improved heritability estimation from genome-wide SNP data. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).

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

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

12. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*. (Springer, 2001).

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

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

15. 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).

16. 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).

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

18. Sudlow, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *PLoS Med.* **12**, e1001779 (2015).

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

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

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

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

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

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

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

26. 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).

27. 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).

28. 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).

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

30. 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).

31. Hakonarson, H., Gulcher, J. R. & Stefansson, K. deCODE genetics, Inc. *Pharmacogenomics* **4**, 209–215 (2003).

32. Chen, Z. *et al.* China Kadoorie Biobank of 0 . 5 million people : survey methods , baseline characteristics and long-term follow-up. 1652–1666 (2011) doi:10.1093/ije/dyr120.

33. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2014).

34. Nagai, A., Hirata, M., Kamatani, Y., Muto, K. & Matsuda, K. Overview of the BioBank Japan Project : Study design and pro fi le. **27**, 2–8 (2017).

35. Sullivan, P. F. The Psychiatric GWAS Consortium: Big Science Comes to Psychiatry. *Neuron* **68**, 182–186 (2010).

36. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~ 700 000 individuals of European ancestry. **27**, 3641–3649 (2018).

37. Speed, D. & Balding, D. Better estimation of SNP heritability from summary statistics provides a new understanding of the genetic architecture of complex traits. *Nat. Genet.* **51**, 277–284 (2019).

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

39. 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).

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

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

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

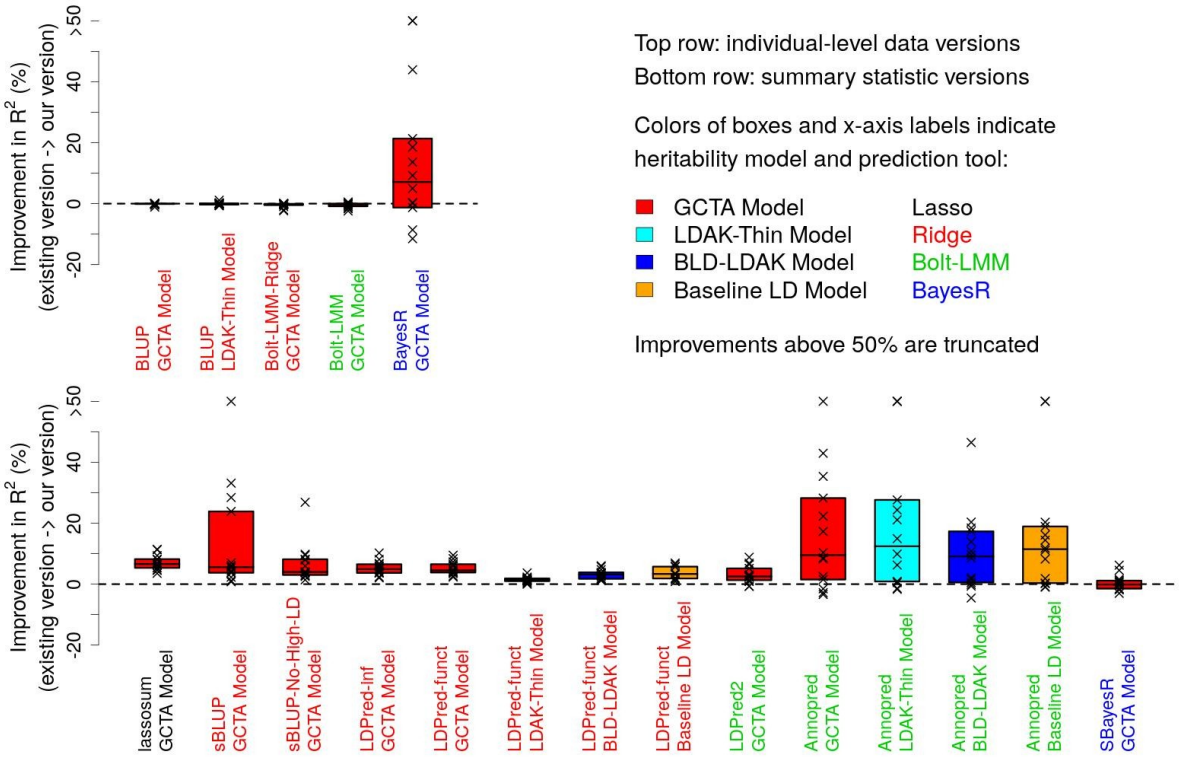
43. Zhao, Z. *et al.* Fine-tuning Polygenic Risk Scores with GWAS Summary Statistics. *bioRxiv* 810713 (2019) doi:10.1101/810713.

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

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

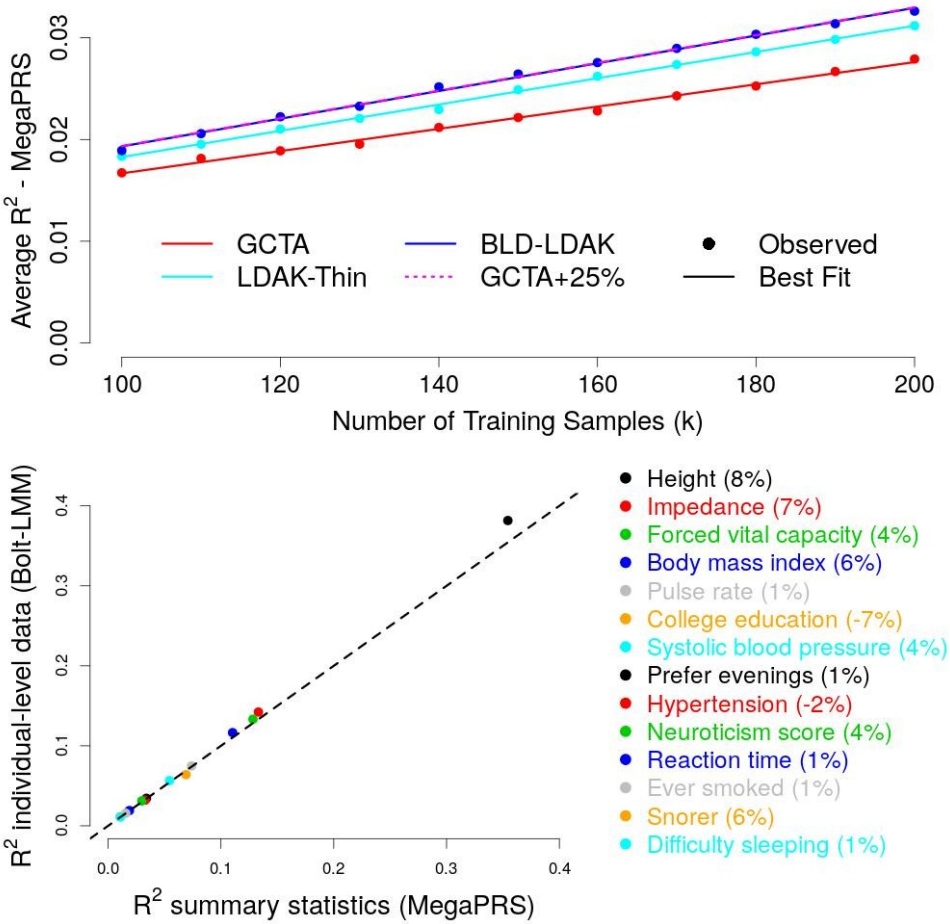
46. Bulik-Sullivan, B. *et al.* LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. *Nat. Genet.* **47**, 291–295 (2015).

Extended Data Figures

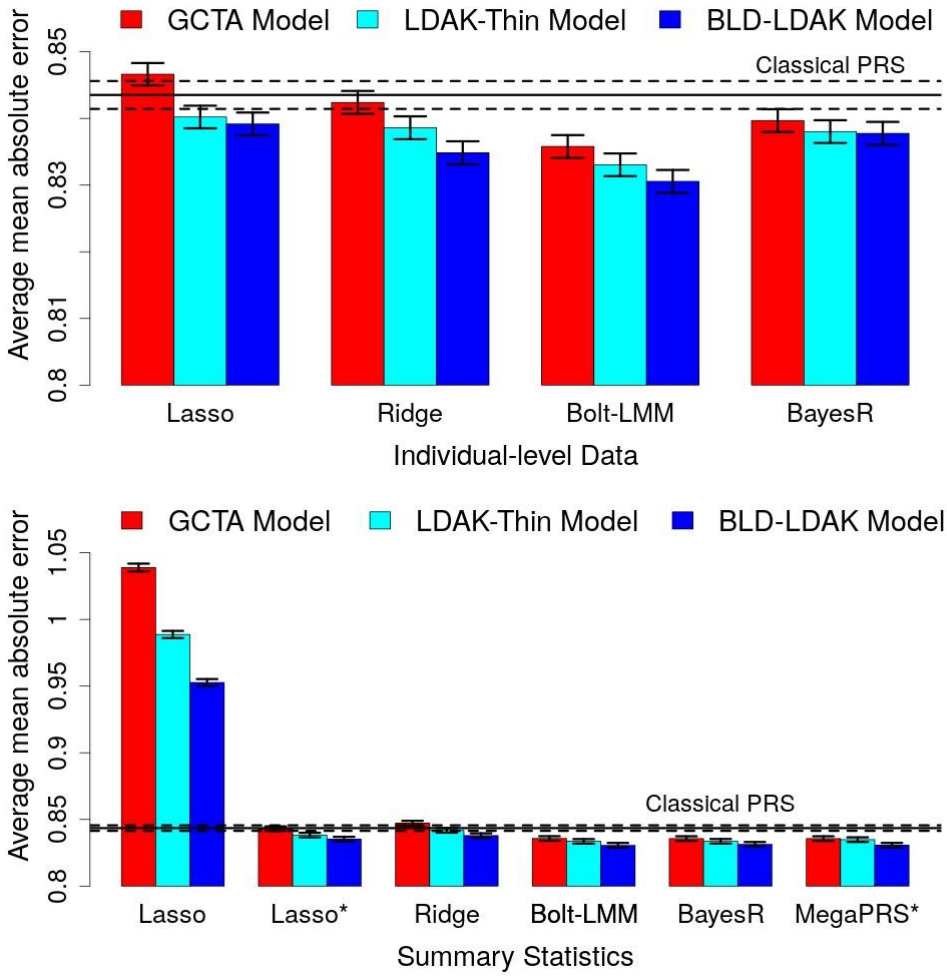


Extended Data Fig. 1 | Comparison with existing software. Points report the improvement in prediction accuracy for each phenotype when we switch from using existing versions of each tool to our versions. Accuracy is measured via R^2, the squared correlation between observed and predicted phenotypes across 20,000 test samples; improvements above 50% are truncated. Boxes mark the median and inter-quartile range across the 14 phenotypes. The top row considers versions of tools that use individual-level data, the bottom row considers versions that use summary statistics. The colors of boxes indicate the assumed heritability model (GCTA, LDAK-Thin, BLD-LDAK or Baseline LD Model), while the colors of the x-axis labels indicates the type of prediction tool (lasso, ridge regression, Bolt-LMM or BayesR).

Here we summarize the different analyses; for more details see Methods, while for scripts see the Supplementary Note. In general, we trained models using the full training data for each phenotype (200,000 individuals and 628,694 SNPs). However, this was not computationally feasible for BLUP (best linear unbiased prediction) and the original BayesR software, so for these we instead restricted to 50,000 individuals and 628,694 SNPs, and to 20,000 individuals and 99,852 SNPs (Chromosomes 1 & 2), respectively. Further, when comparing with AnnoPred, it was necessary to exclude the major histocompatibility complex (Chr6:25-34Mb), as otherwise AnnoPred would often fail to complete. For Bolt-LMM-Ridge, we run the original Bolt-LMM software with the options "--pEst .5 --varFrac2Est .5" (i.e., forcing the ridge regression model). For sBLUP-No-High-LD, we run sBLUP with SNPs in high-LD regions excluded.[23] We compared results from LDPred-funct to those from our version of ridge regression. Strictly, this comparison is not fair (our version is disadvantaged), because LDPred-funct uses a generalized version of the ridge regression model (having estimated effect sizes, these are then regularized via cross-validation). However, we observed that this regularization made little difference to accuracy (we found that results from LDPred-funct were very similar to those from LDPred-inf, which omits the regularization).

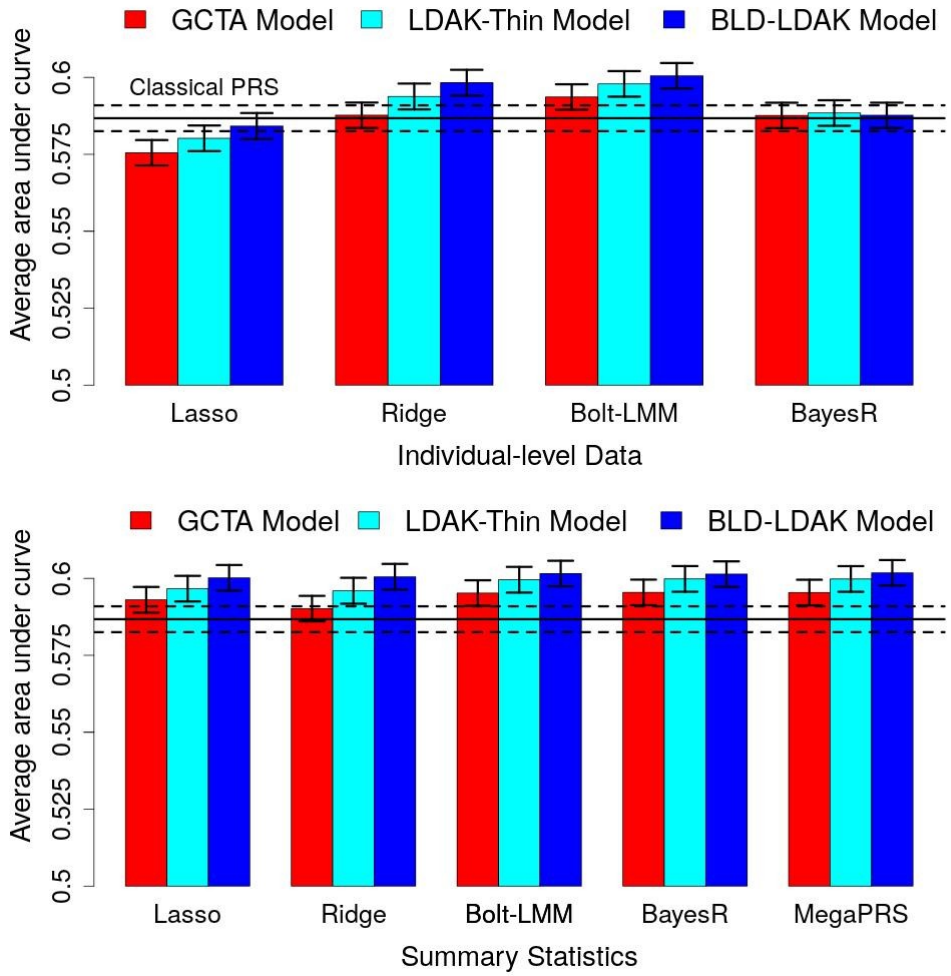


Extended Data Fig. 2 | Prediction accuracy of MegaPRS. We construct MegaPRS prediction models using summary statistics from between 100,000 and 200,000 training samples, then measure their accuracy via R^2, the squared correlation between observed and predicted phenotypes across 20,000 test samples. **a,** 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, where n is the number of training samples; for the GCTA Model, we use the best fit line to predict average R^2 if the sample size was 25% higher than specified (dashed line). **b,** Points compare the accuracy of MegaPRS models constructed using summary statistics (x-axis) to the accuracy of Bolt-LMM models constructed using individual-level data (y-axis); colors indicate the phenotype (numbers in brackets indicate the percentage improvement in R^2 when switching from summary statistics to individual-level data).

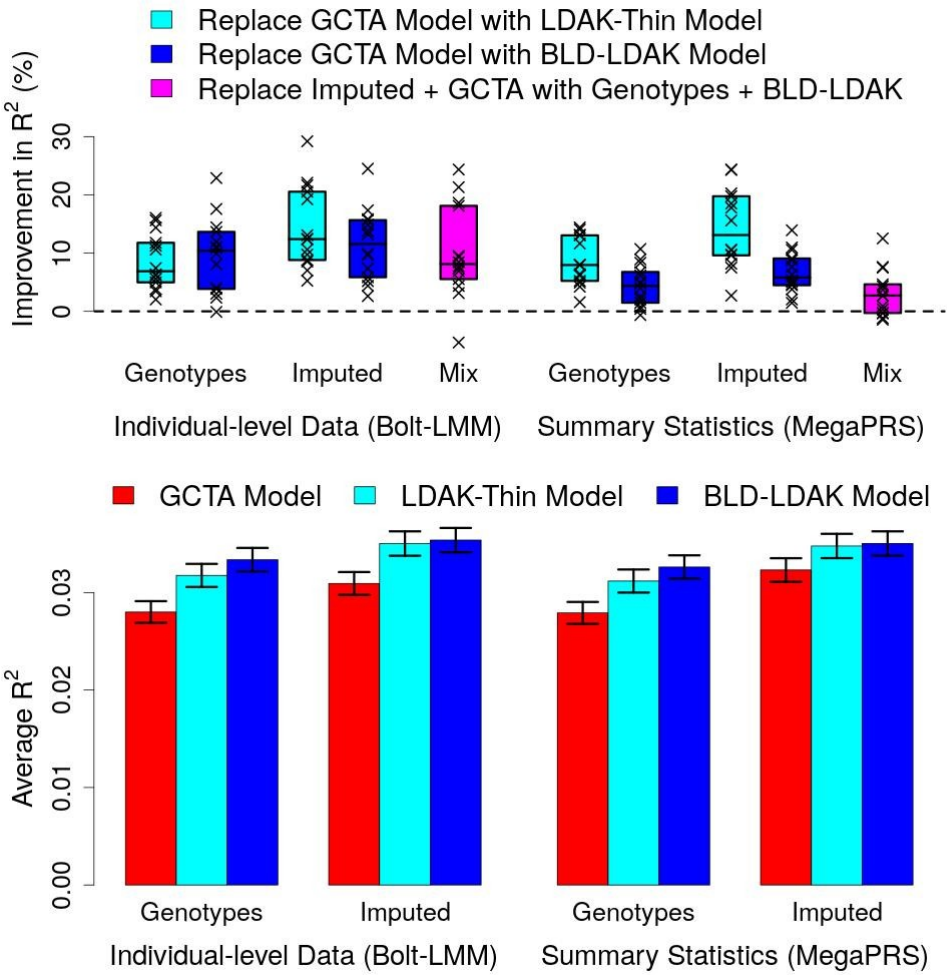


Extended Data Fig. 3 | Measuring prediction accuracy via mean absolute error. We construct lasso, ridge regression, Bolt-LMM, BayesR and MegaPRS prediction models using all 200,000 training samples, then measure their accuracy based on the mean absolute error between observed and predicted phenotypes across 20,000 test samples. Bars report mean absolute error 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 mean absolute error for classical classical polygenic risk scores and a 95% confidence interval. **a,** The models are constructed using individual-level data. **b,** The models are constructed using summary statistics (note that when running MegaPRS, we do not consider lasso models, for the reason explained below).

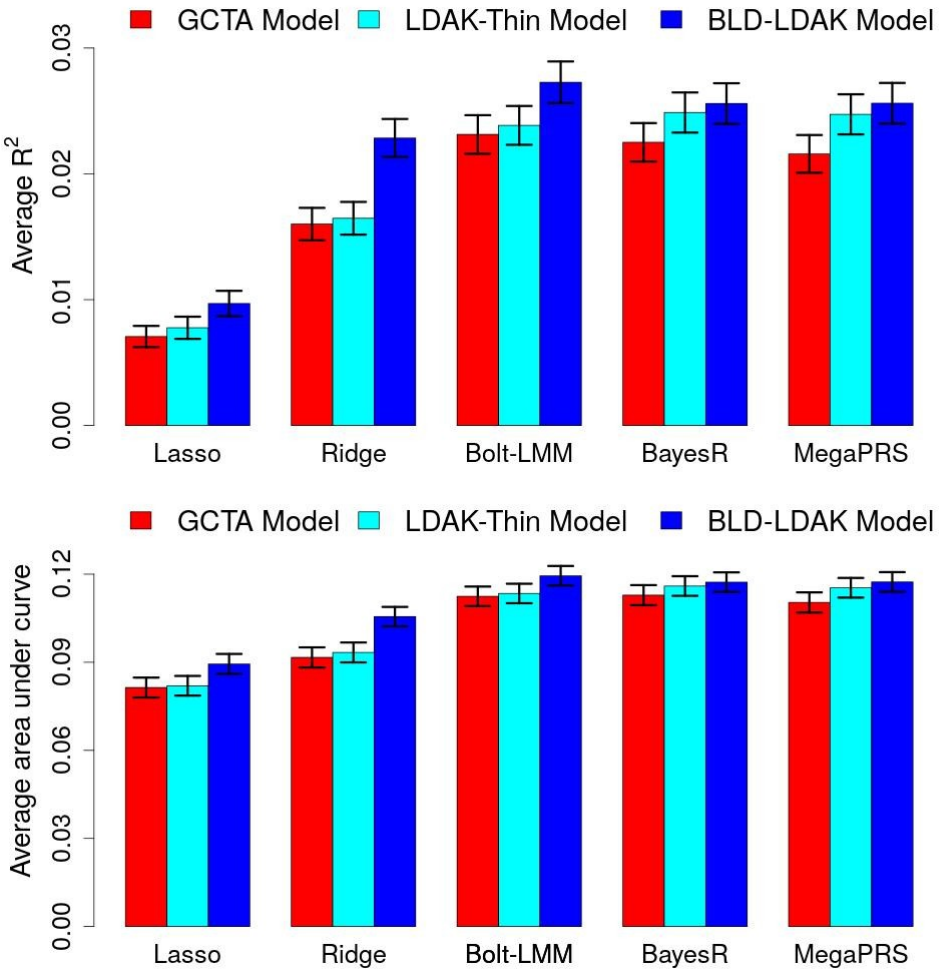
We see that for all tools, improving the heritability model (i.e., replacing the GCTA Model with either the LDAK-Thin or BLD-LDAK Model) improves average prediction accuracy (reduces average mean absolute error). However, we note that relative to the other tools, our standard summary statistics version of lasso performs poorly. This is because when choosing the smoothing parameter λ, the value that maximizes R^2 is in general not the value that minimizes mean absolute error. Therefore, if the aim is to construct prediction models from summary statistics that minimize mean absolute error, we recommend using MegaPRS*, which replaces the standard lasso solver with lasso*, a non-sparse version (specifically, lasso* uses posterior mean estimates of effect sizes, instead of posterior modes, which ensures all effect sizes are non-zero).



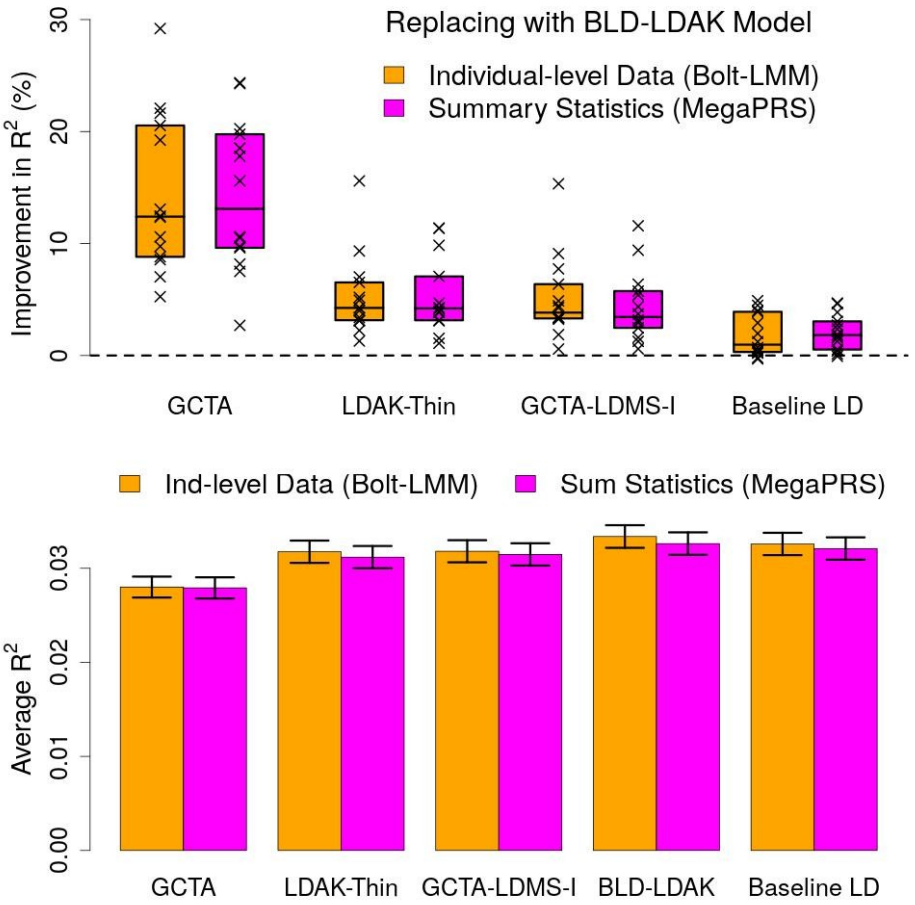
Extended Data Fig. 4 | Measuring prediction accuracy via area under curve. We construct lasso, ridge regression, Bolt-LMM, BayesR and MegaPRS prediction models using all 200,000 training samples, then for the four binary phenotypes (college education, ever smoked, hypertension and snorer) measure their accuracy based on area under the receiver operating curve for the 20,000 test samples. Bars report area under curve averaged across the four phenotypes (vertical segments mark 95% confidence intervals); colors indicate the assumed heritability model, while blocks indicate the prediction tool. The horizontal lines mark average area under curve for classical classical polygenic risk scores and a 95% confidence interval. **a,** The models are constructed using individual-level data. **b,** The models are constructed using summary statistics.



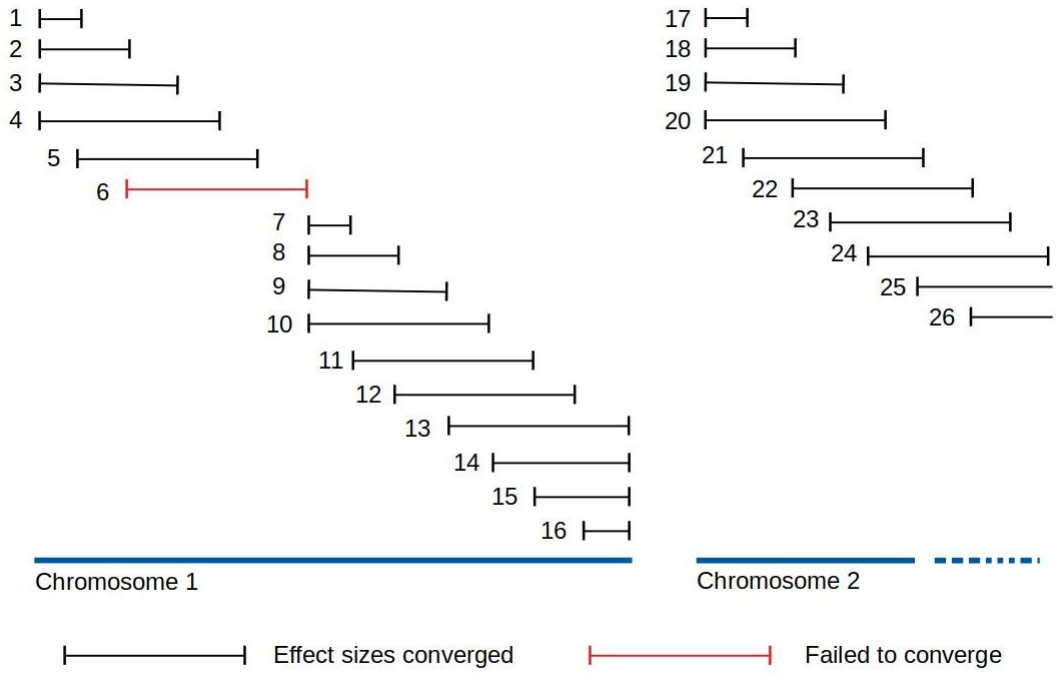
Extended Data Fig. 5 | Including imputed SNP genotypes. We construct prediction models using all 200,000 training samples, then measure their accuracy via R^2, the squared correlation between observed and predicted phenotypes across 20,000 test samples. First we restrict to 629,000 directly-genotyped SNPs (the same as for our main analysis), then we increase the number of SNPs to 7.5M by including imputed genotypes. When using individual-level data, we construct Bolt-LMM models; when using summary statistics, we use MegaPRS. Note that when including imputed genotypes, it was not computationally feasible to analyze all SNPs together, so we instead analyzed each chromosome separately. We then merged effect sizes across chromosomes before performing cross-validation, so that we continued to select prior parameters based on genome-wide data (rather than separately for each chromosome). **a,** 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). The dark blue boxes show that when using imputed data, R^2 generally increases, similar to when using directly-genotyped data (light blue boxes). The purple boxes show that the improvement in accuracy by switching from the GCTA Model to the BLD-LDAK Model is generally larger than the improvement in accuracy by including imputed SNPs. **b,** Bars report R^2 averaged across the 14 phenotypes (vertical segments mark 95% confidence intervals); colors indicate the assumed heritability model.



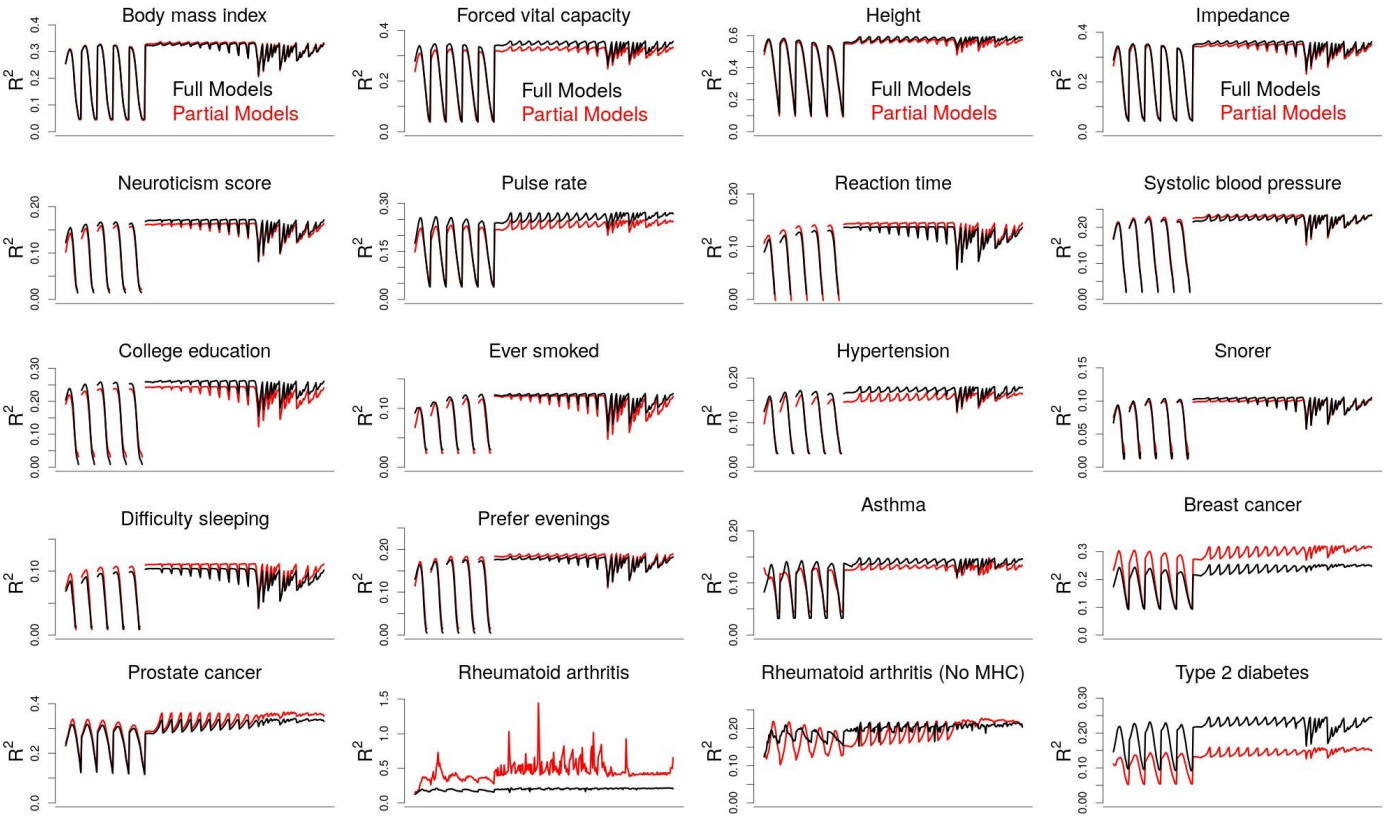
Extended Data Fig. 6 | Asthma, breast cancer, prostate cancer, rheumatoid arthritis and type 2 diabetes. For each disease, we use summary statistics from published studies to construct lasso, ridge regression, Bolt-LMM, BayesR and MegaPRS prediction models, then test their accuracy based on how well they predict for UK Biobank individuals. **a,** Bars report R^2, the squared correlation between observed and predicted phenotypes across UK Biobank individuals, averaged across the five diseases (vertical segments mark 95% confidence intervals); colors indicate the assumed heritability model. **b,** Bars report area under the receiver operating curve for the UK Biobank individuals, averaged across the five diseases (vertical segments mark 95% confidence intervals); colors indicate the assumed heritability model.



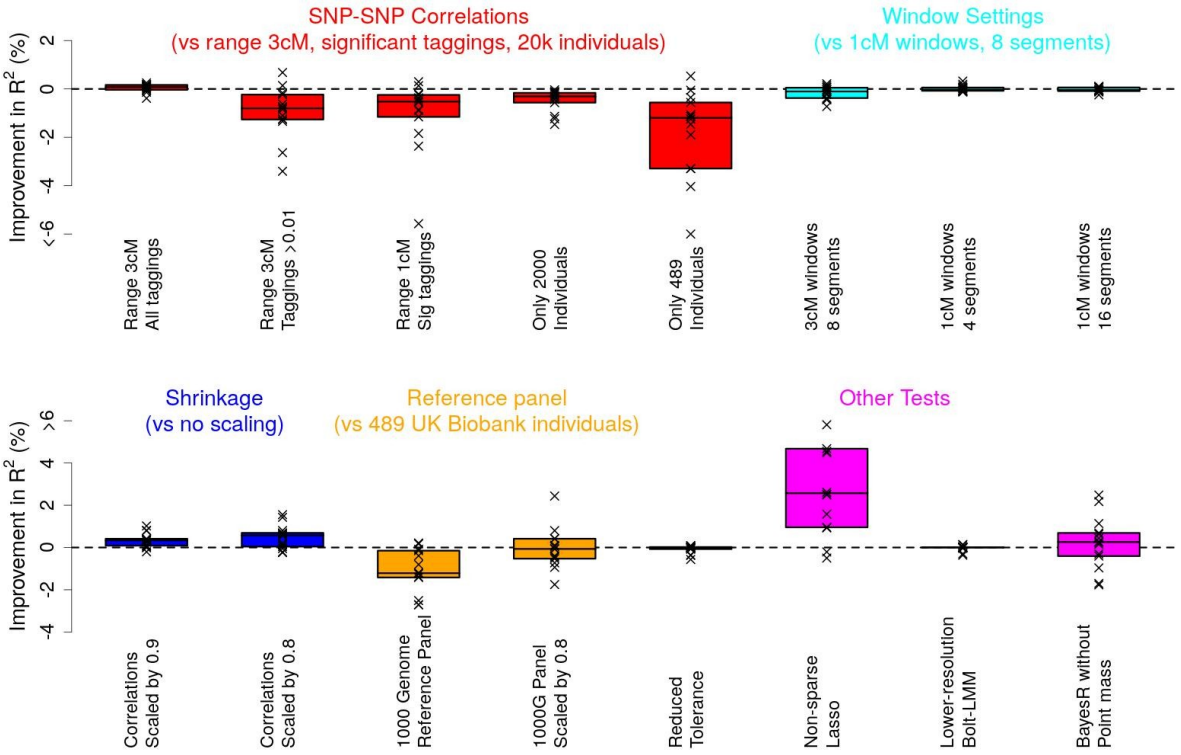
Extended Data Fig. 7 | Alternative heritability models. We construct prediction models using all 200,000 training samples, then measure their accuracy via R^2, the squared correlation between observed and predicted phenotypes across 20,000 test samples. When constructing models using individual-level data, we use Bolt-LMM; when constructing models from summary statistics, we use MegaPRS. In addition to the GCTA, LDAK-Thin and BLD-LDAK Models, we also consider the GCTA-LDMS-I[30] and Baseline LD Model,[29] the models recommended by the authors of GCTA[3] and LD Score Regression[46], respectively. **a,** Points report the percentage increase in R^2 for individual phenotypes when we switch from assuming the GCTA, LDAK-Thin, GCTA-LDMS-I or Baseline LD Model to the BLD-LDAK Model (boxes mark the median and inter-quartile range across the 14 phenotypes). **b,** For each heritability model, bars report R^2 averaged across the 14 phenotypes (vertical segments mark 95% confidence intervals).



Extended Data Fig. 8 | A sliding window approach for estimating effect sizes using summary statistics. When performing estimating effect sizes using summary statistics (for which we used either coordinate descent or variational Bayes), we found it was not feasible to iterate over all SNPs in the genome. Therefore, we instead used sliding windows, illustrated above (numbers indicate the order in which windows are processed). By default, we iteratively estimate effect sizes for all SNPs in a 1cM window, stopping when the estimated proportion of variance explained by these SNPs converges. Then we move 1/8 cM along the genome, and repeat for the next 1cM window. If a window fails to converge within 50 iterations, we reset the effect sizes to their values prior to that window (note that this happens very rarely, and not once for our main analysis).



Extended Data Fig. 9 | Cross-validation using pseudo summary statistics. We select prior parameters for each tool via cross-validation. When constructing prediction models using summary statistic, this cross validation requires (independent) training and test summary statistics (for example, one constructed using 90% of training samples, the other constructed using the remaining 10%). For asthma, breast cancer, prostate cancer, rheumatoid arthritis and type 2 diabetes, we are unable to compute these directly, as we do not have access to the data used to generate the summary statistics. Therefore, we instead create pseudo partial summary statistics (for details, see Methods). This figure demonstrates that this approach is effective. Each plot reports R, the correlation between observed and predicted phenotypes across test samples, for each of the 325 models considered by MegaPRS (100 lasso models, 11 ridge regression models, 131 Bolt-LMM models and 83 BayesR models). For the black lines, the models are constructed using full summary statistics, and R is calculated using the test datasets (this can be viewed as the "true R"); for the red lines, the models are constructed using the pseudo training summary statistics, and R is estimated using the pseudo test summary statistics (the "estimated R"). The first 14 boxes correspond to the UK Biobank phenotypes; the last six to asthma, breast cancer, prostrate cancer, rheumatoid arthritis (twice) and type 2 diabetes. Note that in practice, there is no need to create pseudo summary statistics for the UK Biobank phenotypes, as with access to individual-level data, we can compute the actual summary statistics; we do so here only for demonstration. We see that in general, the black and red lines mirror each other, indicating that we can reliably use pseudo summary statistics to select model parameters. The exception is for rheumatoid arthritis, a consequence of there being very strong effect loci within the major histocompatibility complex (MHC), a region of long-range linkage disequilibrium. However, in this case, we find it suffices to exclude the MHC region when using the pseudo summary statistics (we continue to include the region when using the full summary statistics).



Extended Data Fig. 10 | Sensitivity analysis of MegaPRS. Points report the percentage increase in R^2, the squared correlation between observed and predicted phenotypes across 20,000 test samples, for individual phenotypes when we change settings from their default values (boxes mark the median and inter-quartile range across the 14 phenotypes). For the five red boxes, we change settings when calculating SNP-SNP correlations: by default, we record correlations within 3cM that are significant (P<0.99, which corresponds to those with magnitude greater than $\sqrt{0.0003}$), using a reference panel of 20,000 individuals; here we instead record correlations within 1cM, or record all correlations, or record those whose magnitude is greater than $\sqrt{0.01}$, or reduce the reference panel to 2000 individuals, or reduce the reference panel to 489 individuals. For the three light-blue boxes, we change settings when estimating effect sizes: by default, we estimate effect sizes for a 1cM window, then move 1/8th of a window along the genome and repeat; here we instead use a 3cM window, or move 1/4th of a window along the genome, or move 1/16th of a window along the genome. For the two dark-blue boxes, we scale the estimates of SNP-SNP correlations by 0.9 or 0.8. For the first orange box, we replace the UK Biobank reference panel with 489 Europeans individuals from the 1000 Genome Project (note that here we compare results with those obtained using a reference panel of only 489 UK Biobank individual); for the second orange box, we do the same except scale estimates of SNP-SNP correlations by 0.8. For the first purple box, we reduce the convergence tolerance from 0.00001 to 0.000001. For the second purple box, we only construct lasso models, replacing the sparse solver (effect sizes are conditional posterior modes) with a non-sparse solver (effect sizes are conditional posterior means). For the third purple box, we only construct Bolt-LMM models, reducing the number of possible pairs for (p,f2) from 132 to 18. For the fourth purple box, we construct only BayesR models, using a shrinkage version of the prior distribution (we replace the point mass at zero with a Gaussian distribution with variance $\sigma^2/1000$).

Overall, we find that the performance of MegaPRS is fairly robust to the changes of settings. The largest impact is if we replace the UK Biobank reference panel with a 1000 Genomes panel. In this case, R^2 reduces on average by about 2% due to reducing the number of individuals from 20,000 to 489 (fifth box on the top row) and on average a further 1% due to replacing UK Biobank genotypes with 1000 Genome genotypes (third box of the bottom row). We note that there is a small advantage using shrunk estimates of SNP-SNP correlations (first two boxes on bottom row), and that by shrinking, we can offset some of the reduction due to substituting in the 1000 Genome reference panel (fourth box on bottom row). Lastly, we note that the non-sparse version of the lasso tends to out-perform the sparse version (sixth box of the bottom row), reflecting once more that polygenic prediction models generally outperform sparse models.