Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models

Ping Zeng¹,² & Xiang Zhou²,³

Using genotype data to perform accurate genetic prediction of complex traits can facilitate genomic selection in animal and plant breeding programs, and can aid in the development of personalized medicine in humans. Because most complex traits have a polygenic architecture, accurate genetic prediction often requires modeling all genetic variants together via polygenic methods. Here, we develop such a polygenic method, which we refer to as the latent Dirichlet process regression model. Dirichlet process regression is non-parametric in nature, relies on the Dirichlet process to flexibly and adaptively model the effect size distribution, and thus enjoys robust prediction performance across a broad spectrum of genetic architectures. We compare Dirichlet process regression with several commonly used prediction methods with simulations. We further apply Dirichlet process regression to predict gene expressions, to conduct PrediXcan based gene set test, to perform genomic selection of four traits in two species, and to predict eight complex traits in a human cohort.
Genome-wide association studies (GWASs) have identified thousands of genetic loci harboring associated single-nucleotide polymorphisms (SNPs) for many complex traits and diseases, providing unprecedented insights into the genetic basis of phenotypic variation\cite{1,2,3,4}. The accumulation of genetic data from existing association studies has led to a growing interest in predicting traits and diseases using genetic markers (in addition to using traditional environmental or clinical variables)\cite{5}. In animals or plants, accurate phenotype prediction with genetic markers can assist the selection of individuals with desirable breeding values and can improve the effectiveness of breeding programs\cite{6,7}. In humans, accurate phenotype prediction with genetic markers can facilitate disease prevention and intervention at early stages and can aid in the development of personalized medicine by using genotype information to customize the treatment and predict the outcome\cite{8,9}.

Phenotype prediction has also been proposed recently as a key step for integrating functional genomic sequencing studies with GWASs: we can construct more powerful and interpretable gene-set tests in GWASs by setting variant weights to be the coefficients inferred from predictive models in expression quantitative trait locus mapping studies\cite{10}.

Progress toward accurate phenotype prediction requires the development of statistical methods that can model all SNPs jointly. Previous association studies have demonstrated that most complex traits and common diseases have a polygenic background and are each influenced by many genetic variants with small effects. For instance, it is estimated that thousands of causal variants influence human height\cite{11}. Similarly, many animal or plant traits are contributed by hundreds of causal variants (e.g., maize-related traits, such as kernel oil and growing degree days (GDDs)\cite{12,13}; and cattle-related traits, such as backfat thickness, milk yield (MY) and hot carcass weight)\cite{14,15}. Because most complex traits and common diseases have a polygenic architecture, a handful of identified associated SNPs often only capture a small proportion of the phenotypic variation and thus cannot be used to yield accurate phenotype and risk prediction. Instead, accurate phenotype prediction requires polygenic models that can make use of all genome-wide SNPs\cite{16,17,18,19,20}. In the past decade, successful development and application of many polygenic models in the context of genomic selection has revolutionized many animal breeding programs\cite{21,22,23,24,25}. More recently, applications of polygenic models to human GWASs have also yielded fruitful results\cite{26,27,28}.

Most existing polygenic models for prediction make an assumption on the effect size distribution and different methods differ mainly in such modeling assumption. For example, the commonly used linear mixed model (LMM), also known as the best linear unbiased predictor (BLUP), assumes that the effect sizes from all variants follow a normal distribution\cite{29,30}. The Bayes alphabetic (e.g., BayesA and BayesB) methods assume that the variant effect sizes follow a t-distribution or its variation\cite{31,32,33}. The Bayesian lasso assumes a double exponential/Laplace distribution\cite{34,35}. NEG generalizes the Bayesian lasso by assuming a normal exponential gamma distribution\cite{36}. BVSR and BayesC\pi assume a point-normal distribution\cite{37,38}. BSLMM assumes a mixture of two normal distributions\cite{39} and is closely related to the early reversible jump Markov Chain Monte Carlo (rjMCMC) method\cite{40}. BayesR\pi assumes a three-component normal mixture together with a point mass at zero. Given the large number of modeling choices, one naturally wonders which method to use for any given trait. Previous studies have suggested that accurate prediction requires choosing a prior effect size distribution that can closely match the shape of the true effect size distribution.

**Fig. 1** The induced non-parametric Dirichlet process (nonparametric DP) normal mixture prior on the effect sizes can be used to approximate a large number of unimodal distributions. We either simulated 2000 values from a standard t-distribution with df = 4; b a point-t mixture distribution with the zero proportion being 0.2, or equivalently, 0.8 × t(df = 4) + 0.2 × δ0, where δ0 denotes a point mass at zero; c a four-component uniform step mixture distribution 0.50 × U(−0.05,0.05) + 0.25 × U(−0.3,0.3) + 0.15 × U(−0.8,0.8) + 0.05 × U(−2,2), where U denotes a uniform distribution; or obtained d the estimated marginal effect sizes from a linear mixed model in the cattle data with SCS (somatic cell score) as the phenotype. To make the first three data comparable with the last data in d, we centered and scaled the values from the first three data for them to have a mean of zero and within the range of (−0.3,0.3). We then fit each data with either our non-parametric distribution (red) or a normal distribution (blue), and displayed the fitted curves on top of the sample distribution (black). Clearly, the non-parametric Dirichlet process normal mixture can approximate all these distributions well, while a simple normal distribution cannot.
such that the inferred posterior can approximate well the polygenic architecture of the given trait\textsuperscript{24, 35, 36}. However, the effect size distribution for any given trait or disease is unknown a priori and varies for different diseases in terms of the number of causal variants, their minor allele frequencies (MAFs), and their individual effect sizes\textsuperscript{31}. Therefore, to achieve robust performance, it is important to design prior distributions that are flexible enough to resemble the true effect size distribution in many traits as close as possible\textsuperscript{34, 35}.

Up to now, almost all existing polygenic models are parametric in nature and use a prior effect size distribution that is characterized by a few parameters. From the information channel perspective\textsuperscript{37}, the number of parameters in a parametric model determines model complexity and bounds the amount of information in data that can be captured by the model\textsuperscript{37–40}. Therefore, using only a few parameters to characterize the effect size distribution can limit the flexibility of the model\textsuperscript{37, 38} and impede its robust performance across a range of genetic architectures. As an example, the commonly applied LMM uses a normal distribution with one variance component parameter to characterize the effect size distribution. For highly polygenic traits, the assumed normal distribution can approximate the true effect size distribution well, and as a result, LMM can achieve good predictive performance\textsuperscript{34, 35}. However, for traits with large effect variants, the assumed normal distribution can no longer capture the true effect size distribution well and the performance of LMM decays\textsuperscript{34, 35}.

To allow for greater flexibility on the a priori effect size distribution and to enable robust phenotype prediction performance across a range of phenotypes, we develop a Bayesian non-parametric model, which we refer to as the latent Dirichlet process regression (DPR). DPR does not use any fixed parametric distribution as the prior choice for the effect size distribution. Instead, DPR relies on the Dirichlet process to assign a prior on the effect size distribution itself and is thus capable of inferring an effect size distribution from the data at hand. Effectively, DPR uses infinitely many parameters a priori to characterize the effect size distribution, and with such a flexible modeling assumption, DPR is capable of adapting to a broad spectrum of genetic architectures and achieves robust predictive performance across a wide range of complex traits. We illustrate the benefits of DPR with simulations and real data applications for gene expression prediction, gene-based test via PrediXcan, genomic selection for four traits in two species, as well as genetic prediction of eight complex traits in a human cohort.

**Results**

**Method overview.** An overview of our method is provided in the Methods section with details provided in the Supplementary Note. Briefly, we use a Dirichlet process to introduce a non-parametric effect size distribution that can robustly resemble a large classes of unimodal distributions. Indeed, our prior effect size distribution can be used to adaptively and accurately approximate a t-distribution, a point-t mixture distribution, a mixture of step functions, as well as the marginal effect sizes estimated from a real data set; whereas a normal distribution cannot (Fig. 1). Therefore, our prior distribution on the effect size can adaptively approximate a wide range of possible effect size distributions underlying complex traits. Since accurate modeling

---

**Fig. 2** Comparison of prediction performance of several methods with DPR.MCMC in simulations when PVE = 0.5. Performance is measured by $R^2$ difference with respect to DPR.MCMC, where a negative value (i.e., values below the red horizontal line) indicates worse performance than DPR.MCMC. The sample $R^2$ differences are obtained from 20 replicates in each scenario. Methods for comparison include BVSR (cyan), BayesR (chocolate), LMM (purple), MultiBLUP (green), DPR.VB (red), rjMCMC (black blue), and DPR.MCMC. Simulation scenarios include: a Scenario I, which satisfies the DPR modeling assumption; b Scenario II, which satisfies the BayesR modeling assumption; c Scenario III, where the number of SNPs in the large effect group is 10, 100, or 1000; and d Scenario IV, where the effect sizes are generated from either a normal distribution, a t-distribution or a Laplace distribution. For each box plot, the bottom and top of the box are the first and third quartiles, while the ends of whiskers represent either the lowest datum within 1.5 interquartile range of the lower quartile or the highest datum within 1.5 interquartile range of the upper quartile. For DPR.MCMC, the mean predictive $R^2$ in the test set and the standard deviation for the eight settings are, respectively, 0.272 (0.031), 0.299 (0.026), 0.295 (0.026), 0.281 (0.030), 0.277 (0.035), 0.278 (0.030), 0.282 (0.025), and 0.273 (0.022)
of the effect size distribution is a key to achieve accurate prediction performance\(^2^4, 34, 36\), we expect our non-parametric model to perform robustly well across a range of polygenic architectures. Our method is implemented in the DPR software, freely available at http://www.xzlab.org/software.html.

**Simulations.** We first compare the performance of DPR with several other commonly used prediction methods using simulations. A total of seven different methods are included for comparison: (1) BVSR\(^2^9\); (2) BayesR\(^3^5\); (3) LMM\(^2^8\); (4) MultiBLUP\(^4^1\); (5) rjMCMC\(^2^0\); (6) DPR.VB, the variational Bayesian (VB) version of DPR; and (7) DPR.MCMC, the Markov chain Monte Carlo (MCMC) version of DPR. Note that both BayesR and MultiBLUP have been recently demonstrated to outperform a range of existing prediction methods; thus, we do not include other prediction methods into comparison here.

To make our simulations as real as possible, we used genotypes from an existing cattle GWAS data set\(^1^7\) with 5024 individuals and 42,551 SNPs and simulated phenotypes. To cover a range of possible genetic architectures, we consider eight simulation settings from four different simulation scenarios with the phenotypic variance explained (PVE) by all SNPs being either 0.2, 0.5, or 0.8 (details in Methods). In each setting for each PVE value, we performed 20 simulation replicates. In each replicate, we randomly split the data into a training data with 80% individuals and a test data with the remaining 20% individuals. We then fitted different methods on the training data and evaluated their prediction performance on the test data (i.e., Monte Carlo cross validation). We evaluated prediction performance using either the squared correlation coefficient ($R^2$) or mean squared error (MSE). We contrasted the prediction performance of all other methods with that of DPR.MCMC by taking the difference of $R^2$ or MSE between the other methods and DPR.MCMC. Therefore, an $R^2$ difference below zero or an MSE difference above zero suggests worse performance than DPR.MCMC. Figure 2 shows $R^2$ differences for different methods across 20 replicates in each of the eight simulation settings for PVE = 0.5. Because Fig. 2 shows prediction performance difference, a large sample variance of a method in the figure only implies that the prediction performance of the method differs a lot from that of DPR.MCMC, but does not imply that the method itself has a large variation in predictive performance. Supplementary Table 1 shows the standard deviation of absolute $R^2$ values across cross validation replicates; various methods display similar prediction variability. Supplementary Figs. 1 and 2 show the $R^2$ differences for PVE = 0.2 and PVE = 0.8, respectively. The corresponding results for MSE differences are shown in Supplementary Figs. 3–5. The $R^2$ and MSE values of the baseline method, DPR.MCMC, are shown in the corresponding figure legend.

Overall, while each method works the best when their individual modeling assumption is satisfied, DPR.MCMC is robust and works well across all eight settings from four scenarios. For example, if we rank the methods based on their median performance across replicates, then when the total PVE is moderate (e.g., PVE = 0.5, Fig. 2; note that for each PVE there are a total of eight simulation settings for the four scenarios), DPR.MCMC is the best or among the best in seven simulation settings (i.e., scenario I, $c = 10$, 100, and 1000 in scenario III, and normal, t and Laplace distributions in scenario IV; where "among the best" refers to the case when the difference between the given method and the best method is within $\pm 0.001$) and is ranked as the second best in the rest one simulation setting (i.e., scenario II). Similarly, when the total PVE is high (e.g., PVE = 0.8, Supplementary Fig. 2), DPR.MCMC is the best or among the best in seven simulation settings, and it is ranked as the second best in scenario IV when the effect size follows a normal distribution. Even when DPR.MCMC is ranked as the second best method, the difference between DPR.MCMC and the best method is often small. Among the rest of the methods, LMM, MultiBLUP, and rjMCMC all work well in polygenic settings (scenario I; $c = 1000$ in scenario III; scenario IV) but can perform poorly in sparse settings (scenario II; $c = 10$ and $c = 100$ in scenario III). The performance of LMM, MultiBLUP, and rjMCMC in polygenic vs. sparse settings presumably stems from their polygenic assumptions on the effect size distribution. In contrast, because of the sparse assumption on the effect size distribution, both BayesR and BVSR have an advantage in sparse settings (scenario II; $c = 10$ or 100 in scenario III) but suffers in polygenic settings ($c = 1000$ in scenario III; scenario IV). The performance of BVSR is also generally worse than BayesR in the challenging setting when PVE is either small or moderate, presumably because of the much simpler prior assumption employed in BVSR for the non-zero effects. Finally, the VB version of DPR (i.e., DPR.VB) performs considerably less well compared with the MCMC version of DPR (i.e., DPR.MCMC), especially when PVE is high (Supplementary Fig. 2). However, DPR.VB still compares favorably with the other methods when PVE is small or moderate (Supplementary Fig. 1).

**Real data applications.** To gain further insights, we compare the performance of DPR with the other methods in several real data sets to (1) predict gene expression levels using cis-SNPs; (2) conduct subsequent PrediXcan based gene set test; (3) perform genomic selection in animal studies; and (4) predict complex traits in humans.

Our first application is predicting gene expression levels using cis-SNPs in the GEUVADIS data\(^3^2\). The GEUVADIS data contains gene expression measurements on 15,810 genes and 465 individuals after quality control (Methods). These individuals

---

**Table 1** Comparison of seven different methods in predicting gene expression levels in the GEUVADIS data

| Threshold | ENET | BayesR | BVSR | LMM | MultiBLUP | rjMCMC | DPR |
|----------|------|--------|------|-----|-----------|--------|-----|
| 0.10     | 1061 | 809    | 486  | 1195| 1098      | 1013   | 1163|
| 0.20     | 449  | 338    | 142  | 403 | 299       | 321    | 389 |
| 0.30     | 182  | 170    | 49   | 162 | 110       | 123    | 195 |
| 0.40     | 78   | 84     | 24   | 76  | 46        | 47     | 70  |
| 0.50     | 37   | 35     | 23   | 26  | 16        | 19     | 32  |
| 0.60     | 15   | 14     | 4    | 19  | 5         | 9      | 12  |
| 0.70     | 2    | 3      | 1    | 3   | 1         | 2      | 2   |

To compare prediction performance, we counted the number of genes whose median $R^2$ across 20 replicates in the test set is above a given $R^2$ threshold. A larger number thus indicates better performance. For each given threshold, we colored the best method with red and the second best method with blue.
have their genotypes measured in the 1000 Genomes project. In the data, we first identified cis-SNPs that are within 100 kb of each gene and obtained an average of 175 cis-SNPs per gene. Then, for each gene in turn, we applied different methods to predict gene expression levels using these cis-SNPs. To measure prediction performance, we carried out 20 Monte Carlo cross validation data splits as in simulations. In each data split, we fitted methods in a training set with 80% of randomly selected individuals and evaluated method performance using $R^2$ in the test set with the remaining 20% of individuals. In addition to the seven methods used in the simulations (i.e., LMM, BVSR, MultiBLUP, BayesR, rjMCMC, DPR.VB, and DPR.MCMC), we also applied Elastic Net (ENET), which is the default method used in the original PrediXcan paper. Table 1 lists the number of genes with a predictive $R^2$ above different thresholds for different methods. The predictive $R^2$ obtained from DPR.MCMC vs. various other methods across all genes is shown as scatter plots in Supplementary Fig. 6, where each plot also lists the number of genes for which DPR.MCMC performs better and the number of genes for which DPR.MCMC performs worse.

The results are largely consistent with these in simulations. Overall, DPR.MCMC generally achieves better predictive performance than the other methods. For example, DPR.MCMC is able to achieve a higher predictive $R^2 > 0.10$ in ~1300 genes, which is ~100 more than that by the second best method at this threshold (i.e., LMM; Table 1). Similarly, compared with other methods, not only does DPR.MCMC achieve a higher $R^2$ for most genes; the $R^2$ improvement from DPR.MCMC can be large for many genes (Supplementary Fig. 6). Among the rest of the methods, the performance of LMM, DPR.VB, and ENET are comparable with each other and are ranked right behind DPR.MCMC. On the other hand, the two sparse models (i.e., BVSR and BayesR) perform poorly in this data, especially for some genes whose expression levels are highly predictive by the other methods (Table 1, Supplementary Fig. 6).

The robust performance of DPR.MCMC in predicting gene expression levels also translates to a relatively high power in the subsequent PrediXcan gene set test. To perform PrediXcan gene set test, we consider the seven common diseases from Wellcome Trust Case Control Consortium (WTCCC) as in Gamazon et al. For each disease and each gene in turn, we used the estimated cis-SNP effect sizes on expression levels from GEUVADIS as weights to construct gene set tests in WTCCC. Following Gamazon et al., we focused on a set of 4343 genes that had a predictive $R^2$ above 0.01 from all methods. The results are shown in Table 2, which lists the number of significant genes identified by different methods for different diseases. In total, DPR.MCMC identified 38 genes associated with different phenotypes, more than that identified by any other methods.

The performance of DPR.MCMC is followed by DPR.VB and subsequently LMM and rjMCMC. Supplementary Table 2 lists the significant genes identified by DPR.MCMC, which are all consistent with previous studies. We also note that, in general, a higher gene expression predictive performance leads to a higher power in the subsequent gene set analysis. In addition, consistent with their relatively poor gene expression prediction performance, the two sparse models (BayesR and BVSR) do not perform well in the gene set test as well.

Finally, we compare the performance of DPR with the other methods in predicting phenotypes in three GWAS data sets: (1) a cattle study, where we focus on three phenotypes: milk fat percentage (MFP), MY, as well as somatic cell score (SCS); (2) a maize study, where we use GDD as the phenotype; (3) the Framingham heart study (FHS) data, where we focus on five plasma traits that include low-density lipoprotein (LDL) cholesterol, glucose (GLU), high-density lipoprotein (HDL) cholesterol, total cholesterol (TC) and triglycerides (TGs), and three anthropometric traits that include height, weight and body mass index (BMI). As in simulations, for each phenotype, we performed 20 Monte Carlo cross validation data splits. In each data split, we fitted methods in a training set with 80% of randomly selected individuals and evaluated method performance using $R^2$ or MSE in a test set with the remaining 20% of individuals. We again contrasted the performance of the other methods with that of DPR.MCMC by taking the $R^2$ difference or MSE difference with respect to DPR.MCMC. The results are shown in Fig. 3 ($R^2$ difference) and Supplementary Fig. 7 (MSE difference), with $R^2$ and MSE of DPR.MCMC presented in the corresponding figure legend. Supplementary Table 1 shows the standard deviation of absolute $R^2$ values across cross variation replicates. Supplementary Fig. 8 also displays trace plots of the log posterior likelihood from DPR.MCMC for all traits, suggesting reasonable convergence of our method.

Overall, consistent with simulations, DPR.MCMC shows robust performance across all traits and is ranked either as the best or the second best method. In the cattle data (Fig. 3a), for MFP and MY, both DPR.MCMC and BayesR perform the best. For SCS, DPR.MCMC performs the best, followed by BayesR. rjMCMC performs well for MFP and MY but poorly for SCS; while LMM and MultiBLUP do not perform well for MFP and MY in the cattle data but their performance improves for SCS. The relative performance of BayesR vs. LMM and MultiBLUP in the cattle data is consistent with the distinct genetic architectures that underlie the three complex traits: while MFP and MY are affected by a few large or moderate effect SNPs together with many small effect SNPs, SCS is a highly polygenic trait influenced by many SNPs with small effects. BVSR performs poorly for these three traits in the cattle data. In the maize data (Fig. 3a), DPR.MCMC performs the best,
SNPs with large effects
In the FHS data (Fig. 3b, c), DPR.MCMC method for predicting Weight and BMI; DPR.VB performs poorly for certain phenotypes. In Fig. 3, for example, performance of DPR.MCMC, however, all other methods can perform the best or among the best for LDL, Height, and Weight. The poor performance of BVSR presumably stems from its relatively simple and sparse assumption on the effect sizes.

Because the FHS is a family based study, we use this data to further examine the influence of individual relatedness on prediction performance. To do so, we divided the FHS data into two sub data sets (D1 and D2) with equal sample size but different levels of relatedness (details in Methods): individuals in D1 are more closely related to each other than those in D2. We then compared method performance by performing cross validation in each of the two data sets separately. While the difference between methods becomes smaller due to smaller sample size in the two sub data sets as compared to that in the complete data (Supplementary Fig. 9 vs. Fig. 3b, c). For example, DPR.MCMC is ranked as the best method or among the best methods for six traits in D1 and for four traits in D2. BayesR performs similarly and is ranked as the best or among the best for four traits in D1 and for five traits in D2. LMM ranks right after DPR.MCMC and BayesR, while the other methods do not perform well here. In addition, all methods generally perform better in D1 than in D2 (Supplementary Fig. 10), suggesting that relatedness improves prediction performance—a phenomenon that has been well recognized by previous studies. Besides cross validation within each data set separately, we also performed cross validation between D1 and D2 by predicting traits in one data with parameters inferred from another. The results are again largely consistent with the main results. In particular, DPR.MCMC is ranked as the best or among the best for six traits in D1 to D2 prediction and for eight traits in D2 to D1 prediction. BayesR is ranked as the best or among the best for six traits in D1 to D2 prediction and for eight traits in D2 to D1 prediction. rjMCMC also performs reasonably well and follows right behind DPR.MCMC and BayesR (Supplementary Fig. 11).

Computational time. Finally, we list the computational time of the seven methods for the 12 traits in Table 3. Note that some differences in computational time among methods may reflect implementation issues, including the language environment in which the methods are implemented, rather than fundamental differences between algorithms. In addition, we only list in the table the computation time in the fitting stage. Computation time spent in the prediction stage by plugging in estimated coefficients in the new data is almost ignorable and is thus not listed. For sampling based methods (BVSR, rjMCMC, DPR.MCMC, and BayesR), we measure the computational time based on a fixed number of iterations. However, due to different convergence properties of different algorithms (e.g., BVSR uses a Metropolis-Hastings algorithm, rjMCMC uses a reversible jump MCMC algorithm, while both DPR.MCMC and BayesR use a Gibbs sampling algorithm), a fixed number of iterations in different methods may correspond to different mixing performance. Nevertheless, we can see that DPR.MCMC has a similar computational cost as the other Gibbs based approach (e.g., BayesR), though in the human data both these Gibbs based approaches (DPR.MCMC and BayesR) can be slower than the Metropolis-Hastings approach (BVSR) and the reversible jump MCMC algorithm (rjMCMC) that effectively update only a small subset of significant SNPs in each iteration. In contrast, DPR.VB is orders of magnitude faster than its MCMC counterpart, and is computationally as efficient as the other two non-MCMC based approaches (LMM and MultiBLUP).
Discussion

We have presented a novel statistical method, DPR, for genetic prediction of complex traits. DPR uses an infinitely many parameters a priori to flexibly model the effect size distribution, and represents the first non-parametric method developed for modeling polygenic traits in genetic association studies. By flexibly modeling the effect size distribution, DPR is capable of adapting to the polygenic architecture underlying many complex traits and enjoys robust performance across a range of phenotypes. With simulations and applications to four real data sets, we have illustrated the benefits of DPR.

We have focused on one application of DPR—genetic prediction of phenotypes. Like some other polygenic methods34,35,50 DPR can also be applied to many other polygenic applications. For example, DPR can be used to estimate the proportion of variance in phenotypes explained by all SNPs, a quantity that is commonly referred to as SNP heritability. Because DPR assumes a flexible effect size distribution that is adaptive to the genetic architecture underlying a given trait, it has the potential to provide accurate estimation of SNP heritability. As another example, DPR can be applied to association mapping. There, we can view the normal component with the smallest variance as the polygenic background, and we can estimate the probability of a SNP being in any normal components other than the smallest one as the posterior inclusion probability (PIP). PIP computed in this way measures SNP marginal association strength in the presence of polygenic effects, and may represent a more powerful association indicator than standard single SNP association test statistics33,50. An important feature of using PIP in the context of Bayesian models is that PIP quantifies the uncertainty of association strength33,50, which is a desirable feature that is not easily achieved by penalized frequentist counterparts.

Here, we have restricted ourselves to applying DPR to continuous phenotypes. For case control studies, we could follow previous approaches of treating binary phenotypes as continuous traits and apply DPR directly34,35,41. However, it would be desirable to extend DPR to accommodate case control data or other discrete data types in a principled way, by, for example, extending DPR into the generalized linear model framework. In particular, we could use a probit or a logistic link to extend DPR to directly model case control data. We could use a Poisson or an over-dispersed Poisson distribution to extend DPR to model count data. Extending DPR to various discrete data types would likely lead to wider applications of DPR beyond GWASs. For instance, by modeling count data, DPR could be used to perform differential expression analysis52 or expression QTL mapping in RNA sequencing studies53,54. Similarly, by modeling proportional data, DPR could be used to perform differential methylation analysis or methyl-QTL mapping in bisulfite sequencing studies55. Extending DPR to modeling discrete data types using the generalized linear model framework is thus an important avenue for future research.

In the present study, while we used unrelated individuals for GEUVADIS gene expression prediction and PrediXcan tests, we used related individuals for the other two real data applications. Related individuals not only share similar genetic background but also are likely influenced by a common set of environmental factors. In addition, untyped causal SNPs in related individuals can be more easily tagged by neighboring typed SNPs than in unrelated individuals, thanks to the relatively high linkage disequilibrium (LD) in related data. Because both the shared environmental factors and ease tagging of causal SNPs can facilitate prediction, cross validation using related individuals often results in better prediction performance than using unrelated individuals9,23,47,49. However, we caution that the prediction accuracy measured in the test data obtained with cross-validation in related individuals are likely inflated if our ultimate goal is to perform prediction in unrelated individuals instead of related ones. In addition, the predictive model inferred from related individuals may not generalize well to unrelated individuals who are not necessarily influenced by the same set of environmental factors and who do not share the same LD pattern near the causal SNPs. We have attempted to tease apart the influence of relatedness on prediction performance by splitting the FHS data into two parts with different levels of relatedness. Our results indeed show that, while the relative performance of various methods remains largely the same, the absolute performance of all methods do increase with individual relatedness. Additionally, while our method performs well relative to the other methods, we caution that DPR's prediction accuracy is still unlikely of practical use in human clinical setting. Studies on unrelated individuals or studies using a fully independent validation data are likely required to establish the practical utility of prediction methods, which often have unsatisfactory performance there9,47,57. Despite the practical importance of using completely independent or cross-population studies for prediction performance validation, however, we also want to point out its potential caveat: using completely independent data for cross-validation may fail to correctly characterize the relative performance of different methods. In particular, a good method that properly captures the signal in the training data may suffer in the validation data due to different LD patterns between the two populations.
data sets. Similarly, a poor method that fails to capture the signal in the training data may perform well in the validation data where such signal is no longer relevant. Therefore, using training and validating data that are both representative of the study population is important to not only ensure a proper comparison among methods but also to ensure the clinical relevance and wide applicability of the prediction methods. Exploring the use of such data is an important direction for future research.

DPR is not without its limitations. Perhaps the biggest limitation is its computational cost. Like any other MCMC-based approaches34, 35, 58, our Gibbs algorithm for fitting DPR is computationally slow and can only be applied to moderate-sized GWAS studies. To make DPR widely applicable, we have explored the use of VB approximation for fitting DPR. VB approximation obtains an approximate posterior distribution through optimization59 and represents a much faster alternative to MCMC sampling. Indeed, DPR.VB is orders of magnitude faster than DPR.MCMC. However, despite its faster computational speed, the VB algorithm is less accurate than MCMC when SNP heritability is large, sometimes by quite a large margin (e.g., PVE = 0.80 in simulations). The loss of accuracy in VB is not unexpected because our VB assumes that the posterior distributions of the SNP effect sizes are independent from each other. Posterior independence is an unrealistic assumption given that SNP genotypes are correlated through LD. Therefore, it is important to explore alternative VB algorithms to incorporate the posterior correlation among effect sizes, by, for example, adapting algorithms developed elsewhere60, 61. It would be ideal if we could develop algorithms that can achieve a high predictive performance as DPR.MCMC but incurs a small computational cost as DPR.VB. Certainly, besides developing alternative algorithms to MCMC, there is still room for improvement on our MCMC algorithm. For example, we could use all individuals to compute some quantities while use only a subset of individuals to compute other quantities, as in our previous MQS method62, in order to reduce the computational burden while maintaining the accuracy of the algorithm. In any case, developing efficient and accurate algorithms likely represents a key step to adapt existing polygenic methods to association studies that are orders of magnitude larger.

Methods

Overview of DPR. We provide a brief overview of DPR here. Detailed methods and algorithms are provided in the Supplementary Note. To model the relationship between phenotypes and genotypes, we consider the following multiple regression model:

\[ y = W \alpha + X \beta + e \sim N(0, \sigma^2_\epsilon I_n), \]

(1)

where \( y \) is an \( n \)-vector of phenotypes measured on \( n \) individuals; \( W \) is an \( n \) by \( c \) matrix of covariates including a column of 1's for the intercept term; \( \alpha \) is a \( c \)-vector of coefficients; \( X \) is an \( n \) by \( p \) matrix of genotypes; \( \beta \) is the corresponding \( p \)-vector of effect sizes; \( e \) is an \( n \)-vector of residual errors where each element is assumed to be independently and identically distributed from a normal distribution with variance \( \sigma^2_\epsilon \).

Like many previous methods19, 28, 34, 41, we assume that the effect size of each SNP, \( \beta_i \), follows a normal distribution with variance \( \sigma^2_\beta \), i.e., \( \beta_i \sim N(0, \sigma^2_\beta) \). Unlike previous methods, however, we specify a non-parameter prior on the hyper-parameter \( \sigma^2_\beta \) to induce a non-parametric prior on \( \beta \). To motivate our prior choice for \( \sigma^2_\beta \), it helps to provide a brief review of the previous polygenic prediction methods. Among the many polygenic prediction methods developed recently, a surprisingly large number of them assume a priori that the effect sizes follow a particular class of prior distribution—the scale mixture of normal distributions24, 36. Specifically, these methods assume that each effect size \( \beta_i \) follows a normal distribution \( \beta_i \sim N(0, \sigma^2_\beta) \), with the variance parameter (i.e., the scale parameter) \( \sigma^2_\beta \) following another distribution \( p(\sigma^2_\beta) \). The prior distribution on \( \sigma^2_\beta, p(\sigma^2_\beta) \), thus differentiates many different predictive methods. For example, LMM assumes a flat prior \( p(\sigma^2_\beta) \) that is proportional to a constant.24 The Bayes lasso methods assume that \( \sigma^2_\beta \) follows an inverse gamma distribution to induce a\( p(\sigma^2_\beta) \) that is proportional to \( \sigma^2_\beta^{-Q} \) for any given data set will be finite, and can automatically infer the scale parameter \( \sigma^2_\beta \) based on the data at hand. Therefore, our model has the potential to adjust the model complexity according to the data complexity, and has the potential to adapt to a wide range of polygenic architectures.

It is important to point out that our modeling assumption on the effect sizes \( \beta_i \) is also mathematically equivalent to a Dirichlet process normal mixture, which is a mixture of normal distributions with infinitely many normal components. Specifically, using the stick-breaking constructive representation of the Dirichlet process63, we can rewrite our modeling assumption on \( \beta_i \) in an equivalent form as

\[ \beta_i \sim \sum_{\ell=1}^{\infty} N(0, \sigma^2_\beta) |_{\lambda \sim G, G \sim DP(H, \lambda)} \]

(2)

where \( H \) is the base distribution, and \( \lambda \) is the concentration parameter that describes how the distribution on \( \sigma^2_\beta, G \), deviates from the base distribution. Here, we use an inverse gamma distribution as the base distribution and set the two parameters in the inverse gamma distribution to small values to keep the prior relatively uninformative. We treat the concentration parameter \( \lambda \) as an unknown hyper-parameter and intend to infer it from the data at hand. Because we use the Dirichlet process as a prior for the latent variance parameter \( \sigma^2_\beta \) we refer to our regression model based on Eqs. (1) and (2) as the latent Dirichlet Process Regression, or DPR. The individual marginal distribution on \( \beta_i \) (after integrating out \( \sigma^2_\beta \)) is also non-parametric and can robustly resemble a large class of unimodal distributions. Indeed, the distribution on \( \beta_i \) can be used to adaptively and accurately approximate a t-distribution, a point-t mixture distribution, a mixture of step functions, as well as the marginal effect sizes estimated from a real data set; whereas a normal distribution cannot (Fig. 1). Therefore, our prior distribution on \( \sigma^2_\beta \) can adaptively approximate a wide range of possible effect size distributions underlying complex traits. Since accurate modeling of the effect size distribution is a key to achieve accurate prediction performance24, 34, 36, we expect our non-parametric model to perform robustly well across a range of polygenic architectures.

Simulations. We used genotypes from an existing cattle GWAS data set17 with 5024 individuals and 42,551 SNPs and simulated phenotypes. To cover a range of possible genetic architectures, we consider eight simulation settings from four different simulation scenarios to cover a range of possible genetic architectures: (1) a simple regression scenario; (2) an adding assumption that causal SNPs and non-causal SNPs in different effect-size groups have different effects. Specifically, we randomly selected 10 group-one SNPs, 100 group-two SNPs, 1000 group-three
SNPs, and set the remaining SNPs as group-four SNPs. We simulated SNP effect sizes all from a standard normal distribution but scaled their effects in each group separately so that the proportion of genetic variance explained by the four groups are 0.05, 0.15, 0.20, and 0.60, respectively. We set the total proportion of phenotypic variance (PVE; i.e., SNP heritability) to be either 0.2, 0.5, or 0.8, representing low, moderate, and high heritability, respectively. This simulation scenario consists of one simulation setting for each PVE.

Scenario II satisfies the BayesR modeling assumption, where a small proportion of SNPs are causal. These causal SNPs come from three effect-size groups. The simulations were similar to scenario I with the only exception that the group-four SNPs have zero effects. Here, the proportion of PVE by the three groups are 0.1, 0.2, and 0.7, respectively. Again, we set the total PVE to be either 0.2, 0.5, or 0.8. This simulation scenario consists of one simulation setting for each PVE.

Scenario III is similar to Scenario I except that SNPs come from two effect-size groups, thus representing a simpler scenario than I. In particular, we selected either ε = 10 or ε = 1000 SNPs as group-one SNPs and set the remaining SNPs as group-two SNPs. We simulated their effect sizes from a standard normal distribution and scaled their effects in each group separately so that the proportion of PVE by the two groups are 0.2 and 0.8, respectively. Again, we set the total PVE to be either 0.2, 0.5, or 0.8. This simulation scenario consists of three simulation settings for each PVE (ε = 10, 100, or 1000).

Scenario IV is related to the assumption made in LMM and MultiBILUP. Here, all SNPs have non-zero effects and their effect sizes come from either a normal distribution, a t-distribution with four degrees of freedom, or a Laplace distribution. We scaled their effect sizes further so that the total PVE equals 0.2, 0.5, or 0.8. This simulation scenario consists of three simulation settings for each PVE (normal, t, or Laplace).

In each setting, we performed 20 simulation replicates. In each replicate, we randomly split the data into a training data with 80% individuals and a test data with the remaining 20% individuals. We then fitted different methods on the training data and evaluated their prediction performance on the test data (i.e., Monte Carlo cross-validation).

GEUVADIS data. The GEUVADIS data contains gene expression measurements for 465 individuals from five different populations: CEPH (CEU), Finns (FIN), British (GBR), and Yoruban (YRI). Following previous studies, we focused only on protein coding genes and lincRNAs that are annotated from GENCODEv18. Following previous studies, we released lowly expressed genes that have zero counts in at least half of the individuals and obtained a final set of 15,810 genes. Afterwards, following previous studies, we performed PEER normalization to remove confounding effects and unwanted variations. In order to remove potential population stratification, we quantile normalized the gene expression measurements across individuals in each population to a standard normal distribution, and then quantile normalized the gene expression measurements to a standard normal distribution across individuals from all five populations. In addition to the gene expression data, all individuals in GEUVADIS also have their genotypes sequenced in the 1000 Genomes Project. Among the sequenced genotypes, we filtered out SNPs that have a Hardy-Weinberg equilibrium (HWE) p-value < 10^{-4}, a genotype call rate < 95%, or an MAF < 0.01. We retained a total of 7,072,917 SNPs for analysis. We intersected these SNPs with imputed SNPs from WTCCC data (see below). We further selected the group-one SNPs and group-two SNPs based on the estimated proportion of genetic variance by each group.

WTCCC data. The WTCCC data consists of about 14,000 cases from seven common diseases and 2938 shared controls. The cases include 1963 individuals with type 1 diabetes (T1D), 1748 individuals with Crohn’s disease (CD), 1860 individuals with rheumatoid arthritis (RA), 1868 individuals with bipolar disorder (BD), 1924 individuals with type 2 diabetes (T2D), 1926 individuals with coronary artery disease (CAD), and 1952 individuals with hypertension (HT). We obtained quality controlled genotypes from WTCCC and imputed missing genotypes using BIMBAM. We obtained a total of 458,688 SNPs shared across all individuals. We then further imputed SNPs using the 1000 Genomes as the reference panel with SHAPEIT and IMPUTE2. We filtered out SNPs that have a HWE p-value < 10^{-4}, a genotype call rate < 95%, or an MAF < 0.01 to obtain a total of 2,793,818 imputed SNPs for PrediXcan analysis, as in the GEUVADIS data (see above), we focused on the same 15,810 genes. As in ref. 17, we further restricted our association analysis on a set of 4343 genes that have a predictive R^2 above 0.1 by all predictive methods.

Cattle data. The cattle data consists of 5024 samples and 42,551 SNPs after removing SNPs that have a HWE p-value < 10^{-4}, a genotype call rate < 95%, or an MAF < 0.01. For the remaining SNPs, we imputed missing genotypes with the estimated proportion of genetic variance by each group. We analyzed three traits: MTP, MY, and SCS. All phenotypes were quantile normalized to a standard normal distribution before analysis.

Maize data. The maize data consists of 2287 inbred accessions and 98,385 SNPs after removing SNPs that have a HWE p-value < 10^{-4}, a genotype call rate < 95%, or an MAF < 0.01. For the remaining SNPs, we imputed missing genotypes with the estimated proportion of genetic variance by each group. We performed analysis on eight traits: five commonly used plasma traits that include LDL cholesterol, HDL cholesterol, TC, and TG, and three anthropometric traits that include height, weight, and BMI. Each trait was quantile normalized to a standard normal distribution before analysis. Note that the maize data is a family-based study where individuals are genetically related. To tease apart the influence of individual relatedness on prediction performance among methods, we also divided the samples in FHS into two separate data sets with different levels of relatedness. Specifically, we first used genotypes to compute the genome-wide genetic relatedness matrix. We then ordered individuals based on their genetic relatedness values. From top to bottom of the ordered individual pair list, we selected individuals from individual pairs with high levels of relatedness into a data set D1, and continued this process until the sample size of D1 was half of the full data. We then kept the remaining individuals from individual pairs with low levels of relatedness into a data set D2. The relatedness threshold for separating individual pairs into the two data sets was 0.151. Nevertheless, the majority pairs in D1 and D2 have genetic relatedness values close to zero: 99.6% of pairs in D1 and 99.0% of pairs in D2 have a genetic relatedness value between -0.01 and 0.01. As another way of measuring relatedness, we also computed the effective number of chromosome segments (M_e) in the two data. M_e is a crucial parameter that measures the effective number of independent SNPs and is also closely related to the effective number of independent individuals: M_e is small in a data with related individuals and is large in a data with unrelated individuals. A small value of M_e often correlates to high prediction accuracy48, 49, 70. With 20 cross-validation replicates, we estimated M_e in D1 and D2 sub data to be 3454.39 (sd = 140.87) and 8178.601 (sd = 651.52), respectively.

Other methods. We compared the performance of DPR.MCMC and DPR.VB mainly with five existing methods: (1) LMM28 as implemented in the GEMMA software (version 0.95alpha); (2) BVS46 as implemented in the GWASctools software (version 0.95alpha); (3) MultiBILUP47 as implemented in the LDAK software (version 4.9); (4) BayesR55 as implemented in the bayesR software; (5) rjMCMC20 as implemented in the gwas_rjmc1.163 software. We used default settings to fit all these methods. For rMCMC, because it requires us to provide a variance component parameter, we used LMM to estimate the variance component first in all analyses. In addition, rjMCMC does not output parameter estimates. Therefore, for the PrediXcan analysis, we first merged the GEUVADIS and WTCCC files for every gene, labeled WTCCC individuals as having missing phenotypes, and then ran rjMCMC on these files to obtained predicted gene expression values using the WTCCC genotype data. The same strategy was also applied to perform cross-validation prediction between D1 and D2 sub data sets in FHS. For gene expression prediction and PrediXcan analysis, following the original PrediXcan paper17, we also used ENET22, which is implemented in the R package glmnet (version 1.9-5). For ENET, following23, we set one penalty parameter (i.e., α) to be 0.5 and selected the other one using 10-fold cross validation in the training data.

Code availability. Our method is implemented in the DPR software, freely available at http://www.xzlab.org/software.html.

Data availability. No data were generated in the present study. The GEUVADIS gene expression data is publicly available at http://www.geuvadis.org. The genotype data from the 1000 genomes project is publicly available at http://www.internationalgenome.org. The WTCCC genotype and phenotype data is publicly available at https://www.wtccc.uk. The genotype and phenotype data from the cattle and maize studies are available from the authors upon reasonable request and with permission of Prof. Xiaolei Liu at the HuaZhong Agriculture University (xiaoleliu@mail.hzau.edu.cn). The FHS genotype and phenotype data is available in dbGaP (https://www.ncbi.nlm.nih.gov/gap) with accession number phs000007.

Received: 26 November 2016 Accepted: 30 June 2017 Published online: 06 September 2017

References
1. Friguet, L. G. et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat. Genet. 48, 134–143 (2016).
5. Global Lipids Genetics Consortium. Discovery and re...tudes in Framingham heart study: design, recruitment, and initial examination. Am. J. Epidemiol. 165, 1328–1335 (2007).

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

1829 (2001).

421 (2011).

222 (2015).

475 (2009).

753 (2009).

415 (2017).

1058 (2013).

511 (2013).

371 (20110553) (2013).

20110553 (2013).

1098 (2015).

49, 403–415 (2017).

e1000251 (2011).

19, 95–110 (2004).

233 (2009).

311 (2009).

257 (2009).

511 (2013).

210.1093/nar/gkx1204 (2017).

723 (20110555). 1092 human genomes.

112 (2017).

877 (2017).

875 (2011).

569 (2010).

260, 747 (2009).

275 (2009).

310, 320–321 (2005).

42091 (2017).

128, 409–421 (2011).

4167–4182 (2015).

59, 506–511 (2013).

1901–1998 (2015).

210.1093/nar/gkx204 (2017).

12, 753–768 (2014).

280 (2011).

7, 536–541 (2014).

112 (2017).

105 (2013).

205, 222 (2015).

511 (2013).

511 (2013).

91 (2017).

1557 (2013).

113 (2016).

104 (2013).

128, 409–421 (2011).
Acknowledgements
This research is supported by National Institutes of Health grant R01HG009124. We thank Xiaolei Liu from the HuaZhong Agriculture University (xiaoleiliu@mail.hzau.edu.cn) for providing us with the cattle and maize data. We thank Dr Shiquan Sun for help with the development of DPR. We thank Dr Doug Speed for help with LDAR and Dr Sang Hong Lee for help with gwas_rjmc1.163 software. This study also makes use of data generated by the Wellcome Trust Case Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113 and 085475. This research was conducted in part using data and resources from the Framingham Heart Study of the NHLBI and Boston University School of Medicine, which was partially supported by the NHBI Framingham Heart Study (Contract No. N01-HC-25195) and its contract with Affymetrix, Inc for genotyping services (Contract No. N02-HL-6-4278). We thank all participants and staff from the Framingham Heart Study.

Author contributions
X.Z. and P.Z. conceived and designed the experiments, developed the algorithm, and implemented the software used in analysis. P.Z. performed the experiments, analyzed the data. X.Z. and P.Z. wrote the paper.

Additional information
Supplementary Information accompanies this paper at doi:10.1038/s41467-017-00470-2.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2017