

1     **Non-Parametric Genetic Prediction of Complex Traits**  
2             **with Latent Dirichlet Process Regression Models**

3

4     **Ping Zeng<sup>1, 2</sup>, Xiang Zhou<sup>2, 3\*</sup>**

5

6     <sup>1</sup>Department of Epidemiology and Biostatistics, Xuzhou Medical University,  
7     Xuzhou, Jiangsu 221004, China

8     <sup>2</sup>Department of Biostatistics, University of Michigan, Ann Arbor, Michigan  
9     48109, USA.

10    <sup>3</sup>Center for Statistical Genetics, University of Michigan, Ann Arbor,  
11    Michigan 48109, USA.

12

13    \* Correspondence and requests for materials should be addressed to XZ  
14    (email: xzhousph@umich.edu)

15

16

17

18

## 19 **Abstract**

20 Using genotype data to perform accurate genetic prediction of complex traits can  
21 facilitate genomic selection in animal and plant breeding programs, and can aid in the  
22 development of personalized medicine in humans. Because most complex traits have a  
23 polygenic architecture, accurate genetic prediction often requires modeling all genetic  
24 variants together via polygenic methods. Here, we develop such a polygenic method,  
25 which we refer to as the latent Dirichlet process regression model (DPR). DPR is non-  
26 parametric in nature, relies on the Dirichlet process to flexibly and adaptively model the  
27 effect size distribution, and thus enjoys robust prediction performance across a broad  
28 spectrum of genetic architectures. We compare DPR with several commonly used  
29 prediction methods with simulations. We further apply DPR to predict gene expressions,  
30 to conduct PrediXcan based gene set test, to perform genomic selection of four traits in  
31 two species, and to predict eight complex traits in a human cohort.

32

33

34

35

## 36 **Introduction**

37 Genome-wide association studies (GWASs) have identified thousands of genetic loci  
38 harboring associated single nucleotide polymorphisms (SNPs) for many complex traits  
39 and diseases, providing unprecedented insights into the genetic basis of phenotypic  
40 variation<sup>1-8</sup>. The accumulation of genetic data from existing association studies has led to  
41 a growing interest in predicting traits and diseases using genetic markers (in addition to  
42 using traditional environmental or clinical variables)<sup>9</sup>. In animals or plants, accurate  
43 phenotype prediction with genetic markers can assist the selection of individuals with  
44 desirable breeding values and can improve the effectiveness of breeding programs<sup>10</sup>. In  
45 humans, accurate phenotype prediction with genetic markers can facilitate disease  
46 prevention and intervention at early stages and can aid in the development of  
47 personalized medicine by using genotype information to customize the treatment and  
48 predict the outcome<sup>11</sup>. Phenotype prediction has also been proposed recently as a key  
49 step for integrating functional genomic sequencing studies with GWASs: we can  
50 construct more powerful and interpretable gene-set tests in GWASs by setting variant  
51 weights to be the coefficients inferred from predictive models in expression quantitative  
52 trait locus (eQTL) mapping studies<sup>12</sup>.

53 Progress towards accurate phenotype prediction requires the development of statistical  
54 methods that can model all SNPs jointly. Previous association studies have demonstrated  
55 that most complex traits and common diseases have a polygenic background and are each  
56 influenced by many genetic variants with small effects. For instance, it is estimated that  
57 thousands of causal variants influence human height<sup>13</sup>. Similarly, many animal or plant  
58 traits are contributed by hundreds of causal variants (e.g. maize-related traits, such as  
59 kernel oil and growing degree days<sup>14,15</sup>; and cattle-related traits, such as backfat thickness,  
60 milk yield and hot carcass weight<sup>16,17</sup>). Because most complex traits and common  
61 diseases have a polygenic architecture, a handful of identified associated SNPs often only  
62 capture a small proportion of the phenotypic variation and thus cannot be used to yield  
63 accurate phenotype and risk prediction. Instead, accurate phenotype prediction requires  
64 polygenic models that can make use of all genome-wide SNPs<sup>9,18-20</sup>. In the past decade,  
65 successful development and application of many polygenic models in the context of  
66 genomic selection has revolutionized many animal breeding programs<sup>16,21-23</sup>. More

67 recently, applications of polygenic models to human GWASs have also yielded fruitful  
68 results<sup>11,24-27</sup>.

69 Most existing polygenic models for prediction make an assumption on the effect size  
70 distribution and different methods differ mainly in such modeling assumption. For  
71 example, the commonly used linear mixed model (LMM), also known as the best linear  
72 unbiased predictor (BLUP), assumes that the effect sizes from all variants follow a  
73 normal distribution<sup>9,28</sup>. The Bayes alphabetic (e.g. BayesA and BayesB) methods assume  
74 that the variant effect sizes follow a t-distribution or its variation<sup>10,18,29</sup>. The Bayesian  
75 lasso assumes a double exponential/Laplace distribution<sup>30,31</sup>. NEG generalizes the  
76 Bayesian lasso by assuming a normal exponential gamma distribution<sup>32</sup>. BVSR and  
77 BayesC $\pi$  assume a point-normal distribution<sup>29,33</sup>. BSLMM assumes a mixture of two  
78 normal distributions<sup>34</sup> and is closely related to the early reversible jump Markov Chain  
79 Monte Carlo (rjMCMC) method<sup>20</sup>. BayesR<sup>35</sup> assumes a three-component normal mixture  
80 together with a point mass at zero. Given the large number of modeling choices, one  
81 naturally wonders which method to use for any given trait. Previous studies have  
82 suggested that accurate prediction requires choosing a prior effect size distribution that  
83 can closely match the shape of the true effect size distribution, such that the inferred  
84 posterior can approximate well the polygenic architecture of the given trait<sup>24,35,36</sup>.  
85 However, the effect size distribution for any given trait or disease is unknown *a priori*  
86 and varies for different diseases in terms of the number of causal variants, their minor  
87 allele frequencies (MAFs), and their individual effect sizes<sup>11</sup>. Therefore, to achieve  
88 robust performance, it is important to design prior distributions that are flexible enough  
89 to resemble the true effect distribution in many traits as close as possible<sup>34,35</sup>.

90 Up to now, almost all existing polygenic models are parametric in nature and use a  
91 prior effect size distribution that is characterized by a few parameters. From the  
92 information channel perspective<sup>37</sup>, the number of parameters in a parametric model  
93 determines model complexity and bounds the amount of information in data that can be  
94 captured by the model<sup>37-40</sup>. Therefore, using only a few parameters to characterize the  
95 effect size distribution can limit the flexibility of the model<sup>37,38</sup> and impede its robust  
96 performance across a range of genetic architectures. As an example, the commonly  
97 applied LMM uses a normal distribution with one variance component parameter to

98 characterize the effect size distribution. For highly polygenic traits, the assumed normal  
99 distribution can approximate the true effect size distribution well, and as a result, LMM  
100 can achieve good predictive performance<sup>34,35</sup>. However, for traits with large effect  
101 variants, the assumed normal distribution can no longer capture the true effect size  
102 distribution well and the performance of LMM decays<sup>34,35</sup>.

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

116

## 117 **Results**

118 **Method overview.** An overview of our method is provided in the [Methods](#) section with  
119 details provided in the [Supplementary Note](#). Briefly, we use a Dirichlet process to  
120 introduce a non-parametric effect size distribution that can robustly resemble a large  
121 classes of unimodal distributions. Indeed, our prior effect size distribution can be used to  
122 adaptively and accurately approximate a t-distribution, a point-t mixture distribution, a  
123 mixture of step functions, as well as the marginal effect sizes estimated from a real data  
124 set; whereas a normal distribution cannot ([Fig. 1](#)). Therefore, our prior distribution on the  
125 effect size can adaptively approximate a wide range of possible effect size distributions  
126 underlying complex traits. Since accurate modeling of the effect size distribution is a key  
127 to achieve accurate prediction performance<sup>24,34,36</sup>, we expect our non-parametric model to  
128 perform robustly well across a range of polygenic architectures. Our method is  
129 implemented in the DPR software, freely available at <http://www.xzlab.org/software.html>.

130 **Simulations.** We first compare the performance of DPR with several other commonly  
131 used prediction methods using simulations. A total of seven different methods are  
132 included for comparison: (1) BVSR<sup>29</sup>; (2) BayesR<sup>35</sup>; (3) LMM<sup>28</sup>; (4) MultiBLUP<sup>41</sup>; (5)  
133 rjMCMC<sup>20</sup>; (6) DPR.VB, the variational Bayesian (VB) version of DPR; and (7)  
134 DPR.MCMC, the Markov chain Monte Carlo (MCMC) version of DPR. Note that both  
135 BayesR and MultiBLUP have been recently demonstrated to outperform a range of  
136 existing prediction methods; thus, we do not include other prediction methods into  
137 comparison here.

138 To make our simulations as real as possible, we used genotypes from an existing  
139 cattle GWAS dataset<sup>17</sup> with 5,024 individuals and 42,551 SNPs and simulated  
140 phenotypes. To cover a range of possible genetic architectures, we consider eight  
141 simulation settings from four different simulation scenarios with the phenotypic variance  
142 explained (PVE) by all SNPs being either 0.2, 0.5, or 0.8 (details in [Methods](#)). In each  
143 setting for each PVE value, we performed 20 simulation replicates. In each replicate, we  
144 randomly split the data into a training data with 80% individuals and a test data with the

145 remaining 20% individuals. We then fitted different methods on the training data and  
146 evaluated their prediction performance on the test data (i.e. Monte Carlo cross validation).  
147 We evaluated prediction performance using either the squared correlation coefficient ( $R^2$ )  
148 or mean squared error (MSE). We contrasted the prediction performance of all other  
149 methods with that of DPR.MCMC by taking the difference of  $R^2$  or MSE between the  
150 other methods and DPR.MCMC. Therefore, an  $R^2$  difference below zero or an MSE  
151 difference above zero suggests worse performance than DPR.MCMC. Fig. 2 shows  $R^2$   
152 differences for different methods across 20 replicates in each of the eight simulation  
153 settings for PVE=0.5. Because Fig. 2 shows prediction performance difference, a large  
154 sample variance of a method in the figure only implies that the prediction performance of  
155 the method differs a lot from that of DPR.MCMC, but does not imply that the method  
156 itself has a large variation in predictive performance. Supplementary Table 1 shows the  
157 standard deviation of absolute  $R^2$  values across cross variation replicates; various  
158 methods display similar prediction variability. Supplementary Figs 1 and 2 show the  $R^2$   
159 differences for PVE=0.2 and PVE=0.8, respectively. The corresponding results for MSE  
160 differences are shown in Supplementary Figs 3-5. The  $R^2$  and MSE values of the baseline  
161 method, DPR.MCMC, are shown in the corresponding figure legend.

162 Overall, while each method works the best when their individual modeling  
163 assumption is satisfied, DPR.MCMC is robust and works well across all eight settings  
164 from four scenarios. For example, if we rank the methods based on their median  
165 performance across replicates, then when the total PVE is moderate (e.g. PVE=0.5, Fig. 2;  
166 note that for each PVE there are a total of eight simulation settings for the four scenarios),  
167 DPR.MCMC is the best or among the best in seven simulation settings (i.e. scenario I,  
168  $c=10, 100$  and  $1,000$  in scenario III, and normal,  $t$  and Laplace distributions in scenario  
169 IV; where “among the best” refers to the case when the difference between the given  
170 method and the best method is within  $\pm 0.001$ ) and is ranked as the second best in the rest  
171 one simulation setting (i.e. scenario II). Similarly, when the total PVE is high (e.g.  
172 PVE=0.8, Supplementary Fig. 2), DPR.MCMC is the best or among the best in seven  
173 simulation settings, and it is ranked as the second best in scenario IV when the effect size  
174 follows a normal distribution. Even when DPR.MCMC is ranked as the second best  
175 method, the difference between DPR.MCMC and the best method is often small. Among

176 the rest of the methods, LMM, MultiBLUP and rjMCMC all work well in polygenic  
177 settings (scenario I;  $c=1,000$  in scenario III; scenario IV) but can perform poorly in  
178 sparse settings (scenario II;  $c=10$  and  $c=100$  in scenario III). The performance of LMM,  
179 MultiBLUP and rjMCMC in polygenic vs sparse settings presumably stems from their  
180 polygenic assumptions on the effect size distribution. In contrast, because of the sparse  
181 assumption on the effect size distribution, both BayesR and BVSR have an advantage in  
182 sparse settings (scenario II;  $c=10$  or  $100$  in scenario III) but suffers in polygenic settings  
183 ( $c=1,000$  in scenario III; scenario IV). The performance of BVSR is also generally worse  
184 than BayesR in the challenging setting when PVE is either small or moderate,  
185 presumably because of the much simpler prior assumption employed in BVSR for the  
186 non-zero effects. Finally, the VB version of DPR (i.e. DPR.VB) performs considerably  
187 less well compared with the MCMC version of DPR (i.e. DPR.MCMC), especially when  
188 PVE is high ([Supplementary Fig. 2](#)). However, DPR.VB still compares favorably with  
189 the other methods when PVE is small or moderate ([Supplementary Fig. 1](#)).

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

194 Our first application is predicting gene expression levels using cis-SNPs in the  
195 GEUVADIS data<sup>42</sup>. The GEUVADIS data contains gene expression measurements on  
196 15,810 genes and 465 individuals after quality control ([Methods](#)). These individuals have  
197 their genotypes measured in the 1000 Genomes project<sup>43</sup>. In the data, we first identified  
198 cis-SNPs that are within 100 kb of each gene and obtained an average of 175 cis-SNPs  
199 per gene. Then, for each gene in turn, we applied different methods to predict gene  
200 expression levels using these cis-SNPs. To measure prediction performance, we carried  
201 out 20 Monte Carlo cross validation data splits as in simulations. In each data split, we  
202 fitted methods in a training set with 80% of randomly selected individuals and evaluated  
203 method performance using  $R^2$  in the test set with the remaining 20% of individuals. In  
204 addition to the seven methods used in the simulations (i.e. LMM, BVSR, MultiBLUP,



205 BayesR, rjMCMC, DPR.VB and DPR.MCMC), we also applied Elastic Net (ENET)<sup>44</sup>,  
206 which is the default method used in the original PrediXcan paper<sup>12</sup>. [Table 1](#) lists the  
207 number of genes with a predictive  $R^2$  above different thresholds for different methods.  
208 The predictive  $R^2$  obtained from DPR.MCMC versus various other methods across all  
209 genes is shown as scatter plots in [Supplementary Fig. 6](#), where each plot also lists the  
210 number of genes for which DPR.MCMC performs better and the number of genes for  
211 which DPR.MCMC performs worse.

212 The results are largely consistent with these in simulations. Overall, DPR.MCMC  
213 generally achieves better predictive performance than the other methods. For example,  
214 DPR.MCMC is able to achieve a higher predictive  $R^2 > 0.10$  in ~1,300 genes, which is  
215 ~100 more than that by the second best method at this threshold (i.e. LMM; [Table 1](#)).  
216 Similarly, compared with other methods, not only does DPR.MCMC achieve a higher  $R^2$   
217 for most genes; the  $R^2$  improvement from DPR.MCMC can be large for many genes  
218 ([Supplementary Fig. 6](#)). Among the rest of the methods, the performance of LMM,  
219 DPR.VB and ENET are comparable with each other and are ranked right behind  
220 DPR.MCMC. On the other hand, the two sparse models (i.e. BVSR and BayesR)  
221 perform poorly in this data, especially for some genes whose expression levels are highly  
222 predictive by the other methods ([Table 1](#), [Supplementary Fig. 6](#)).

223 The robust performance of DPR.MCMC in predicting gene expression levels also  
224 translates to a relatively high power in the subsequent PrediXcan gene set test. To  
225 perform PrediXcan gene set test, we consider the seven common diseases from WTCCC<sup>4</sup>  
226 as in Gamazon *et al.*<sup>12</sup>. For each disease and each gene in turn, we used the estimated cis-  
227 SNP effect sizes on expression levels from GEUVADIS as weights to construct gene set  
228 tests in WTCCC. Following Gamazon *et al.*<sup>12</sup>, we focused on a set of 4,343 genes that  
229 had a predictive  $R^2$  above 0.01 from all methods. The results are shown in [Table 2](#), which  
230 lists the number of significant genes identified by different methods for different diseases.  
231 In total, DPR.MCMC identified 38 genes associated with different phenotypes, more  
232 than that identified by any other methods. The performance of DPR.MCMC is followed  
233 by DPR.VB and subsequently LMM and rjMCMC. [Supplementary Table 2](#) lists the  
234 significant genes identified by DPR.MCMC, which are all consistent with previous  
235 studies. We also note that, in general, a higher gene expression predictive performance

236 leads to a higher power in the subsequent gene set analysis. In addition, consistent with  
237 their relatively poor gene expression prediction performance, the two sparse models  
238 (BayesR and BVSr) do not perform well in the gene set test as well.

239 Finally, we compare the performance of DPR with the other methods in predicting  
240 phenotypes in three GWAS data sets: (1) a cattle study<sup>17</sup>, where we focus on three  
241 phenotypes: milk fat percentage (MFP), milk yield (MY), as well as somatic cell score  
242 (SCS); (2) a maize study<sup>15</sup>, where we use growing degree day (GDD) as the phenotype;  
243 (3) the Framingham heart study (FHS) data<sup>45</sup>, where we focus on five plasma traits that  
244 include low-density lipoprotein (LDL) cholesterol, glucose (GLU), high-density  
245 lipoprotein (HDL) cholesterol, total cholesterol (TC) and triglycerides (TG), and three  
246 anthropometric traits that include height, weight and body mass index (BMI). As in  
247 simulations, for each phenotype, we performed 20 Monte Carlo cross validation data  
248 splits. In each data split, we fitted methods in a training set with 80% of randomly  
249 selected individuals and evaluated method performance using  $R^2$  or MSE in a test set  
250 with the remaining 20% of individuals. We again contrasted the performance of the other  
251 methods with that of DPR.MCMC by taking the  $R^2$  difference or MSE difference with  
252 respect to DPR.MCMC. The results are shown in [Fig. 3](#) ( $R^2$  difference) and  
253 [Supplementary Fig. 7](#) (MSE difference), with  $R^2$  and MSE of DPR.MCMC presented in  
254 the corresponding figure legend. [Supplementary Table 1](#) shows the standard deviation of  
255 absolute  $R^2$  values across cross variation replicates. [Supplementary Fig. 8](#) also displays  
256 trace plots of the log posterior likelihood from DPR.MCMC for all traits, suggesting  
257 reasonable convergence of our method.

258 Overall, consistent with simulations, DPR.MCMC shows robust performance across  
259 all traits and is ranked either as the best or the second best method. In the cattle data ([Fig.](#)  
260 [3A](#)), for MFP and MY, both DPR.MCMC and BayesR perform the best. For SCS,  
261 DPR.MCMC performs the best, followed by BayesR. rjMCMC performs well for MFP  
262 and MY but poorly for SCS; while LMM and MultiBLUP do not perform well for MFP  
263 and MY in the cattle data but their performance improves for SCS. The relative  
264 performance of BayesR vs LMM and MultiBLUP in the cattle data is consistent with the  
265 distinct genetic architectures that underlie the three complex traits<sup>17,46</sup>: while MFP and  
266 MY are affected by a few large or moderate effect SNPs together with many small effect

267 SNPs, SCS is a highly polygenic trait influenced by many SNPs with small effects.  
268 BVSR performs poorly for these three traits in the cattle data. In the maize data (Fig. 3A),  
269 DPR.MCMC performs the best, followed by BayesR, suggesting that GDD is influenced  
270 by a few SNPs with large effects<sup>15</sup>. In the Framingham heart study data (Fig. 3B and Fig.  
271 3C), DPR.MCMC performs the best or among the best for LDL, GLU, TC, Weight and  
272 BMI. Its performance is comparable to BayesR and rjMCMC for Height, and follows  
273 right behind BayesR for HDL and TG. Besides DPR and BayesR, rjMCMC also  
274 performs well in FHS and is often ranked as the third best method (e.g. for LDL, GLU,  
275 Height and Weight). In contrast to the relatively robust performance of DPR.MCMC,  
276 however, all other methods can perform poorly for certain phenotypes. In Fig. 3, for  
277 example, BayesR is the second worst method for predicting GLU; LMM is the second  
278 worst method for predicting LDL; MultiBLUP is the worst method for predicting Weight  
279 and BMI; DPR.VB performs among the worst for LDL and HDL; rjMCMC performs  
280 poorly for HDL; while BVSR performs the worst for almost all traits except for LDL,  
281 Height and Weight. The poor performance of BVSR presumably stems from its relatively  
282 simple and sparse assumption on the effect sizes.

283 Because the FHS is a family based study, we use this data to further examine the  
284 influence of individual relatedness on prediction performance. To do so, we divided the  
285 FHS data into two sub data sets (D1 and D2) with equal sample size but different levels  
286 of relatedness (details in Methods): individuals in D1 are more closely related to each  
287 other than those in D2. We then compared method performance by performing cross  
288 validation in each of the two data sets separately. While the difference between methods  
289 becomes smaller due to smaller sample size in the two sub data sets, the relative  
290 performance of most methods for the eight traits are largely unchanged in these two sub  
291 data sets as compared to that in the complete data (Fig. 9 vs Figs 3B and 3C). For  
292 example, DPR.MCMC is ranked as the best method or among the best methods for six  
293 traits in D1 and for four traits in D2. BayesR performs similarly and is ranked as the best  
294 or among the best for four traits in D1 and for five traits in D2. LMM ranks right after  
295 DPR.MCMC and BayesR, while the other methods do not perform well here. In addition,  
296 all methods generally perform better in D1 than in D2 (Supplementary Fig. 10),  
297 suggesting that relatedness improves prediction performance — a phenomenon that has

298 been well recognized by previous studies<sup>9,23,47-49</sup>. Besides cross validation within each  
299 data set separately, we also performed cross validation between D1 and D2 by predicting  
300 traits in one data with parameters inferred from another. The results are again largely  
301 consistent with the main results. In particular, DPR.MCMC is ranked as the best or  
302 among the best for six traits in D1 to D2 prediction and for eight traits in D2 to D1  
303 prediction. BayesR is ranked as the best or among the best for six traits in D1 to D2  
304 prediction and for eight traits in D2 to D1 prediction. rjMCMC also performs reasonably  
305 well and follows right behind DPR.MCMC and BayesR ([Supplementary Fig. 11](#)).

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

326

## 327 **Discussion**

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

336 We have focused on one application of DPR — genetic prediction of phenotypes.  
337 Like some other polygenic methods<sup>34,35,50</sup>, DPR can also be applied to many other  
338 polygenic applications. For example, DPR can be used to estimate the proportion of  
339 variance in phenotypes explained by all SNPs, a quantity that is commonly referred to as  
340 SNP heritability<sup>28,34</sup>. Because DPR assumes a flexible effect size distribution that is  
341 adaptive to the genetic architecture underlying a given trait, it has the potential to provide  
342 accurate estimation of SNP heritability. As another example, DPR can be applied to  
343 association mapping. There, we can view the normal component with the smallest  
344 variance as the polygenic background, and we can estimate the probability of a SNP  
345 being in any normal components other than the smallest one as the posterior inclusion  
346 probability (PIP). PIP computed in this way measures SNP marginal association strength  
347 in the presence of polygenic effects, and may represent a more powerful association  
348 indicator than standard single SNP association test statistics<sup>33,50</sup>. An important feature of  
349 using PIP in the context of Bayesian models is that PIP quantifies the uncertainty of  
350 association strength<sup>33,50</sup>, which is a desirable feature that is not easily achieved by  
351 penalized frequentist counterparts<sup>51</sup>.

352 Here, we have restricted ourselves to applying DPR to continuous phenotypes. For  
353 case control studies, we could follow previous approaches of treating binary phenotypes  
354 as continuous traits and apply DPR directly<sup>34,35,41</sup>. However, it would be desirable to  
355 extend DPR to accommodate case control data or other discrete data types in a principled  
356 way, by, for example, extending DPR into the generalized linear model framework. In

357 particular, we could use a probit or a logistic link to extend DPR to directly model case  
358 control data. We could use a Poisson or an over-dispersed Poisson distribution to extend  
359 DPR to model count data. Extending DPR to various discrete data types would likely  
360 lead to wider applications of DPR beyond GWASs. For instance, by modeling count data,  
361 DPR could be used to perform differential expression analysis<sup>52</sup> or expression QTL  
362 mapping in RNA sequencing studies<sup>53,54</sup>. Similarly, by modeling proportional data, DPR  
363 could be used to perform differential methylation analysis or methyl-QTL mapping in  
364 bisulfite sequencing studies<sup>55</sup>. Extending DPR to modeling discrete data types using the  
365 generalized linear model framework is thus an important avenue for future research.

366 In the present study, while we used unrelated individuals for GEUVADIS gene  
367 expression prediction and PrediXcan tests, we used related individuals for the other two  
368 real data applications. Related individuals not only share similar genetic background but  
369 also are likely influenced by a common set of environmental factors<sup>47,56</sup>. In addition,  
370 untyped causal SNPs in related individuals can be more easily tagged by neighboring  
371 typed SNPs than that in unrelated individuals, thanks to the relatively high linkage  
372 disequilibrium (LD) in related data. Because both the shared environmental factors and  
373 easy tagging of causal SNPs can facilitate prediction, cross validation using related  
374 individuals often results in better prediction performance than using unrelated  
375 individuals<sup>9,23,47-49</sup>. However, we caution that the prediction accuracy measured in the  
376 test data obtained with cross-validation in related individuals are likely inflated if our  
377 ultimate goal is to perform prediction in unrelated individuals instead of related ones. In  
378 addition, the predictive model inferred from related individuals may not generalize well  
379 to unrelated individuals who are not necessarily influenced by the same set of  
380 environmental factors and who do not share the same LD pattern near the causal SNPs.  
381 We have attempted to tease apart the influence of relatedness on prediction performance  
382 by splitting the FHS data into two parts with different levels of relatedness. Our results  
383 indeed show that, while the relative performance of various methods remains largely the  
384 same, the absolute performance of all methods do increase with individual relatedness.  
385 Additionally, while our method performs well relative to the other methods, we caution  
386 that DPR's prediction accuracy is still unlikely of practical use in human clinical setting.  
387 Studies on unrelated individuals or studies using a fully independent validation data are  
388 likely required to establish the practical utility of prediction methods, which often have  
389 unsatisfactory performance there<sup>9,47,57</sup>. Despite the practical importance of using

390 completely independent or cross-population studies for prediction performance validation,  
391 however, we also want to point out its potential caveat: using completely independent  
392 data for cross-validation may fail to correctly characterize the relative performance of  
393 different methods. In particular, a good method that properly captures the signal in the  
394 training data may suffer in the validation data due to different LD patterns between the  
395 two data sets. Similarly, a poor method that fails to capture the signal in the training data  
396 may perform well in the validation data where such signal is no longer relevant.  
397 Therefore, using training and validating data that are both representative of the study  
398 population is important to not only ensure a proper comparison among methods but also  
399 to ensure the clinical relevance and wide applicability of the prediction methods.  
400 Exploring the use of such data is an important direction for future research.

401 DPR is not without its limitations. Perhaps the biggest limitation is its computational  
402 cost. Like any other MCMC based approaches<sup>34,35,58</sup>, our Gibbs algorithm for fitting DPR  
403 is computationally slow and can only be applied to moderate-sized GWAS studies. To  
404 make DPR widely applicable, we have explored the use of variational Bayesian  
405 approximation for fitting DPR. Variational Bayesian approximation obtains an  
406 approximate posterior distribution through optimization<sup>59</sup> and represents a much faster  
407 alternative to MCMC sampling. Indeed, DPR.VB is orders of magnitude faster than  
408 DPR.MCMC. However, despite its faster computational speed, the VB algorithm is less  
409 accurate than MCMC when SNP heritability is large, sometimes by quite a large margin  
410 (e.g. PVE=0.80 in simulations). The loss of accuracy in VB is not unexpected because  
411 our VB assumes that the posterior distributions of the SNP effect sizes are independent  
412 from each other. Posterior independence is an unrealistic assumption given that SNP  
413 genotypes are correlated through LD. Therefore, it is important to explore alternative VB  
414 algorithms to incorporate the posterior correlation among effect sizes, by, for example,  
415 adapting algorithms developed elsewhere<sup>60,61</sup>. It would be ideal if we could develop  
416 algorithms that can achieve a high predictive performance as DPR.MCMC but incurs a  
417 small computational cost as DPR.VB. Certainly, besides developing alternative  
418 algorithms to MCMC, there is still room for improvement on our MCMC algorithm. For  
419 example, we could use all individuals to compute some quantities while use only a subset  
420 of individuals to compute other quantities, as in our previous MQS method<sup>62</sup>, in order to  
421 reduce the computational burden while maintaining the accuracy of the algorithm. In any

422 case, developing efficient and accurate algorithms likely represents a key step to adapt  
423 existing polygenic methods to association studies that are orders of magnitude larger.  
424



## 425 **Methods**

426 **Overview of DPR.** We provide a brief overview of DPR here. Detailed methods and  
427 algorithms are provided in the [Supplementary Note](#). To model the relationship between  
428 phenotypes and genotypes, we consider the following multiple regression model

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, \sigma_e^2 I_n), \quad (1)$$

429 where  $\mathbf{y}$  is an  $n$ -vector of phenotypes measured on  $n$  individuals;  $\mathbf{W}$  is an  $n$  by  $c$  matrix of  
430 covariates including a column of 1s for the intercept term;  $\boldsymbol{\alpha}$  is a  $c$ -vector of coefficients;  
431  $\mathbf{X}$  is an  $n$  by  $p$  matrix of genotypes;  $\boldsymbol{\beta}$  is the corresponding  $p$ -vector of effect sizes;  $\boldsymbol{\epsilon}$  is an  
432  $n$ -vector of residual errors where each element is assumed to be independently and  
433 identically distributed from a normal distribution with variance  $\sigma_e^2$ .

434 Like many previous methods<sup>9,19,28,34,41</sup>, we assume that the effect size of  $i$ th SNP,  $\beta_i$ ,  
435 follows a normal distribution with variance  $\sigma^2$ , i.e.  $\beta_i \sim N(0, \sigma^2)$ . Unlike previous  
436 methods, however, we specify a non-parameter prior on the hyper-parameter  $\sigma^2$  to  
437 induce a non-parametric prior on  $\beta_i$ . To motivate our prior choice for  $\sigma^2$ , it helps to  
438 provide a brief review of the previous polygenic prediction methods. Among the many  
439 polygenic prediction methods developed recently, a surprisingly large number of them  
440 assume *a priori* that the effect sizes follow a particular class of prior distribution – the  
441 scale mixture of normal distributions<sup>63</sup>. Specifically, these methods assume that each  
442 effect size  $\beta_i$  follows a normal distribution  $\beta_i \sim N(0, \sigma^2)$ , with the variance parameter (i.e.  
443 the scale parameter)  $\sigma^2$  following another distribution  $p(\sigma^2)$ . The prior distribution on  
444  $\sigma^2$ ,  $p(\sigma^2)$ , thus differentiates many different predictive methods. For example, LMM  
445 assumes a flat prior  $p(\sigma^2)$  that is proportional to a constant<sup>9,28</sup>. The Bayes alphabetic  
446 methods assume that  $\sigma^2$  follows an inverse gamma distribution to induce a t-prior on  
447  $\beta_i$ <sup>10,18,64</sup>. The Bayesian lasso assumes that  $\sigma^2$  follows a Rayleigh distribution to induce a  
448 double exponential distribution (a.k.a. Laplace distribution) on  $\beta_i$ <sup>30,58</sup>. NEG assumes an  
449 exponential gamma distribution on  $\sigma^2$  to induce an NEG prior on  $\beta_i$ <sup>32</sup>. BVSR and  
450 BayesC $\pi$  assume a mixture of a point mass at zero with another flat prior to induce a  
451 point-normal distribution on  $\beta_i$ <sup>29,33</sup>. BSLMM assumes a mixture of two point masses to  
452 induce a normal mixture distribution on  $\beta_i$ <sup>34</sup>. While BayesR assumes a three point

453 masses together with another point mass at zero on  $\sigma^2$  to also induce a normal mixture  
454 distribution on  $\beta_i$ <sup>35</sup>.

455 The scale mixture of normal distributions is flexible because different distributions on  
456 the scale parameter  $\sigma^2$  can be used to induce many smooth unimodal distributions on  $\beta_i$ .  
457 However, existing predictive methods explicitly make a parametric prior assumption on  
458  $\sigma^2$ , which necessarily relies on a limited number of parameters to characterize the  
459 distribution on  $\sigma^2$ . Consequently, the induced effect size distribution on  $\beta_i$  from a  
460 parametric prior on  $\sigma^2$  can be restrictive and may sometimes fail to resemble closely the  
461 unknown truth effect size distribution underlying complex traits. Motivated by the  
462 potential drawback of parametric priors on  $\sigma^2$ , we instead develop a non-parametric prior  
463 distribution on  $\sigma^2$  to induce a more flexible distribution on  $\beta_i$ . Because a non-parametric  
464 distribution is characterized by effectively infinitely many parameters, our induced effect  
465 size distribution on  $\beta_i$  has the potential to resemble a wide range of genetic architectures  
466 and achieve robust predictive performance across a variety of traits.

467 Technically, we assume  $\sigma^2$  follows a Dirichlet process (DP) prior<sup>37-40</sup>

$$\sigma^2 \sim G, G \sim DP(H, \lambda), \quad (2)$$

468 where  $H$  is the base distribution, and  $\lambda$  is the concentration parameter that describes how  
469 the distribution on  $\sigma^2$ ,  $G$ , deviates from the base distribution. Here, we use an inverse  
470 gamma distribution as the base distribution and set the two parameters in the inverse  
471 gamma distribution to small values to keep the prior relatively uninformative. We treat  
472 the concentration parameter  $\lambda$  as an unknown hyper-parameter and intend to infer it from  
473 the data at hand. Because we use the Dirichlet process as a prior for the latent variance  
474 parameter  $\sigma^2$  we refer to our regression model based on equations (1)-(2) as the latent  
475 Dirichlet Process Regression, or DPR. The induced marginal distribution on  $\beta_i$  (after  
476 integrating out  $\sigma^2$ ) is also non-parametric and can robustly resemble a large classes of  
477 unimodal distributions. Indeed, the distribution on  $\beta_i$  can be used to adaptively and  
478 accurately approximate a t-distribution, a point-t mixture distribution, a mixture of step  
479 functions, as well as the marginal effect sizes estimated from a real data set; whereas a  
480 normal distribution cannot (Fig. 1). Therefore, our prior distribution on the effect size can  
481 adaptively approximate a wide range of possible effect size distributions underlying

482 complex traits. Since accurate modeling of the effect size distribution is a key to achieve  
483 accurate prediction performance<sup>24,34,36</sup>, we expect our non-parametric model to perform  
484 robustly well across a range of polygenic architectures.

485 It is important to point out that our modeling assumption on the effect sizes  $\beta_i$  is also  
486 mathematically equivalent to a Dirichlet process normal mixture, which is a mixture of  
487 normal distributions with infinitely many normal components. Specifically, using the  
488 stick-breaking constructive representation of the Dirichlet process<sup>59</sup>, we can re-write our  
489 modeling assumption on  $\beta_i$  in an equivalent form as

$$\beta_i \sim \sum_{k=1}^{+\infty} \pi_k N(0, \sigma_k^2), \quad (3)$$

$$\pi_k = v_k \prod_{l=1}^{k-1} (1 - v_l), v_k \sim \text{Beta}(1, \lambda),$$

490 where  $\lambda$  is the same concentration parameter as in equation (2), and determines the  
491 number of normal components in the model and subsequently the model complexity<sup>59</sup>.  
492 Each  $\sigma_k^2$  in the above equation follows the base distribution  $H$ . From the normal mixture  
493 equivalence aspect, our method effectively generalizes many previous methods<sup>18,34,35</sup> that  
494 use a fixed, often small, number of normal components to using infinitely many normal  
495 components *a priori*. Although the prior number of normal components in our model is  
496 infinite, the posterior number of components for any given data set will be finite, and can  
497 be automatically inferred based on the data at hand. Therefore, our model has the  
498 potential to adjust the model complexity according to the data complexity, and has the  
499 potential to adapt to a wide range of polygenic architectures.

500 To fit our model, we develop two complementary algorithms: one is based on the  
501 MCMC algorithm, and the other is based on the variational Bayesian (VB)  
502 approximation. The MCMC sampling algorithm, which we refer to as DPR.MCMC, is  
503 accurate but computationally slow. The variational Bayesian algorithm, which we refer  
504 to as DPR.VB, is computationally fast, but, as we will show in the results, is often less  
505 accurate. The two algorithms provide users the choice of speed vs accuracy. The details  
506 of the two algorithms are provided in [Supplementary Note](#).

507 **Simulations.** we used genotypes from an existing cattle GWAS dataset<sup>17</sup> with 5,024  
508 individuals and 42,551 SNPs and simulated phenotypes. To cover a range of possible  
509 genetic architectures, we consider *eight* simulation settings from four different simulation  
510 scenarios to cover a range of possible genetic architectures:

511 (1) Scenario I satisfies the DPR modeling assumption, where all SNPs are causal and  
512 SNPs in different effect-size groups have different effects. Specifically, we  
513 randomly selected 10 group-one SNPs, 100 group-two SNPs, 1,000 group-three  
514 SNPs, and set the remaining SNPs as group-four SNPs. We simulated SNP effect  
515 sizes all from a standard normal distribution but scaled their effects in each group  
516 separately so that the proportion of genetic variance explained by the four groups  
517 are 0.05, 0.15, 0.20, and 0.60, respectively. We set the total proportion of  
518 phenotypic variance (PVE; i.e. SNP heritability) to be either 0.2, 0.5 or 0.8,  
519 representing low, moderate, and high heritability, respectively. This simulation  
520 scenario consists of one simulation setting for each PVE.

521 (2) Scenario II satisfies the BayesR modeling assumption, where a small proportion  
522 of SNPs are causal. These causal SNPs come from three effect-size groups. The  
523 simulations were similar to scenario I with the only exception that the group-four  
524 SNPs have zero effects. Here, the proportion of PVE by the three groups are 0.1,  
525 0.2, and 0.7, respectively. Again, we set the total PVE to be either 0.2, 0.5 or 0.8.  
526 This simulation scenario consists of one simulation setting for each PVE.

527 (3) Scenario III is similar to Scenario I except that SNPs come from two effect-size  
528 groups, thus representing a simpler scenario than I. In particular, we selected  
529 either  $c=10$ , 100 or 1,000 SNPs as group-one SNPs and set the remaining SNPs  
530 as group-two SNPs. We simulated their effect sizes from a standard normal  
531 distribution and scaled their effects in each group separately so that the proportion  
532 of PVE by the two groups are 0.2 and 0.8, respectively. Again, we set the total  
533 PVE to be either 0.2, 0.5 or 0.8. This simulation scenario consists of three  
534 simulation settings for each PVE ( $c=10$ , 100 or 1,000).

535 (4) Scenario IV is related to the assumption made in LMM and MultiBLUP. Here, all  
536 SNPs have non-zero effects and their effect sizes come from either a normal  
537 distribution, a t-distribution with four degrees of freedom, or a Laplace

538 distribution. We scaled their effect sizes further so that the total PVE equals 0.2,  
539 0.5, or 0.8. This simulation scenario consists of three simulation settings for each  
540 PVE (normal, t, or Laplace).

541 In each setting, we performed 20 simulation replicates. In each replicate, we randomly  
542 split the data into a training data with 80% individuals and a test data with the remaining  
543 20% individuals. We then fitted different methods on the training data and evaluated  
544 their prediction performance on the test data (i.e. Monte Carlo cross validation).

545 **GEUVADIS data.** The GEUVADIS data<sup>42</sup> contains gene expression measurements for  
546 465 individuals from five different populations: CEPH (CEU), Finns (FIN), British  
547 (GBR), Toscani (TSI) and Yoruba (YRI). Following previous studies<sup>65</sup>, we focused only  
548 on protein coding genes and lincRNAs that are annotated from GENCODE<sup>66</sup> (release 12).  
549 We removed lowly expressed genes that have zero counts in at least half of the  
550 individuals and obtained a final set of 15,810 genes. Afterwards, following previous  
551 studies<sup>67</sup>, we performed PEER normalization to remove confounding effects and  
552 unwanted variations. In order to remove potential population stratification, we quantile  
553 normalized the gene expression measurements across individuals in each population to a  
554 standard normal distribution, and then quantile normalized the gene expression  
555 measurements to a standard normal distribution across individuals from all five  
556 populations. In addition to the gene expression data, all individuals in GEUVADIS also  
557 have their genotypes sequenced in the 1000 Genomes Project. Among the sequenced  
558 genotypes, we filtered out SNPs that have a Hardy-Weinberg equilibrium (HWE) p-value  
559  $< 10^{-4}$ , a genotype call rate  $< 95\%$ , or an MAF  $< 0.01$ . We retained a total of 7,072,917  
560 SNPs for analysis. We intersected these SNPs with imputed SNPs from WTCCC data<sup>4</sup>  
561 (see below; for the purpose of performing gene set tests) and kept a final set of 2,793,818  
562 overlapping SNPs for analysis. Then, for each gene in turn, we obtained its cis-SNPs that  
563 are located within either 100 kb upstream of the transcription start site (TSS) or 100 kb  
564 downstream of the transcription end site (TES), resulting in an average of 175 cis-SNPs  
565 per gene.

566 **WTCCC data.** The Wellcome Trust Case Control Consortium<sup>4</sup> (WTCCC) 1 data  
567 consists of about 14,000 cases from seven common diseases and 2,938 shared controls.

568 The cases include 1,963 individuals with type 1 diabetes (T1D), 1,748 individuals with  
569 Crohn's disease (CD), 1,860 individuals with rheumatoid arthritis (RA), 1,868  
570 individuals with bipolar disorder (BD), 1,924 individuals with type 2 diabetes (T2D),  
571 1,926 individuals with coronary artery disease (CAD), and 1,952 individuals with  
572 hypertension (HT). We obtained quality controlled genotypes from WTCCC and imputed  
573 missing genotypes using BIMBAM<sup>68</sup>. We obtained a total of 458,868 SNPs shared  
574 across all individuals. We then further imputed SNPs using the 1000 Genomes as the  
575 reference panel with SHAPEIT and IMPUTE2<sup>69</sup>. We filtered out SNPs that have a HWE  
576 p-value  $< 10^{-4}$ , a genotype call rate  $< 95\%$ , or an MAF  $< 0.01$  to obtain a total of  
577 2,793,818 imputed SNPs. For PrediXcan analysis<sup>12</sup>, as in the GEUVADIS data (see  
578 above), we focused on the same 15,810 genes. As in<sup>12</sup>, we further restricted our  
579 association analysis on a set of 4,343 genes that have a predictive  $R^2$  above 0.01 by all  
580 predictive methods.

581 **Cattle data.** The cattle data<sup>17</sup> consists of 5,024 samples and 42,551 SNPs after removing  
582 SNPs that have a HWE p-value  $< 10^{-4}$ , a genotype call rate  $< 95\%$ , or an MAF  $< 0.01$ .  
583 For the remaining SNPs, we imputed missing genotypes with the estimated mean  
584 genotype of that SNP. We analyzed three traits: milk fat percentage (MFP), milk yield  
585 (MY), and somatic cell score (SCS). All phenotypes were quantile normalized to a  
586 standard normal distribution before analysis.

587 **Maize data.** The maize data<sup>15</sup> consists of 2,267 inbred accessions and 98,385 SNPs after  
588 removing SNPs that have a HWE p-value  $< 10^{-4}$ , a genotype call rate  $< 95\%$ , or an MAF  
589  $< 0.01$ . For the remaining SNPs, we imputed missing genotypes with the estimated mean  
590 genotype of that SNP. We used the growing degree days (GDD) to silking as the  
591 phenotype in genomic selection. GDD was calculated using climate data from weather  
592 stations located near the farms<sup>15</sup>, and was quantile normalized to a standard normal  
593 distribution before analysis.

594 **Framingham heart study data.** The Framingham heart study (FHS) data contains  
595 genotype data on 6,950 individuals and 394,174 SNPs. We filtered out SNPs that have a  
596 HWE p-value  $< 10^{-4}$ , a genotype call rate  $< 95\%$ , or an MAF  $< 0.01$  to obtain a final set  
597 of 387,741 SNPs. For these SNPs, we imputed missing genotypes with the estimated

598 mean genotype of that SNP. We performed analysis on eight traits: five commonly used  
599 plasma traits that include low-density lipoprotein (LDL) cholesterol, glucose (GLU),  
600 high-density lipoprotein (HDL) cholesterol, total cholesterol (TC), and triglycerides (TG);  
601 and three anthropometric traits that include height, weight, and body mass index (BMI).  
602 Each trait was quantile normalized to a standard normal distribution before analysis.  
603 Note that the FHS data is a family-based study where individuals are genetically related.  
604 To tease apart the influence of individual relatedness on prediction performance among  
605 methods, we also divided the samples in FHS into two separate data sets with different  
606 levels of relatedness. Specifically, we first used genotypes to compute the genome-wide  
607 genetic relatedness matrix (GRM). We then ordered individual pairs based on their  
608 genetic relatedness values. From top to bottom of the ordered individual pair list, we  
609 selected individuals from individual pairs with high levels of relatedness into a data set  
610 D1, and continued this process until the sample size of D1 was half of the full data. We  
611 then kept the remaining individuals from individual pairs with low levels of relatedness  
612 into a data set D2. The relatedness threshold for separating individual pairs between the  
613 two data sets was 0.151. Nevertheless, the majority pairs in D1 and D2 have genetic  
614 relatedness values close to zero: 99.6% of pairs in D1 and 99.9% of pairs in D2 have a  
615 genetic relatedness value between +/-0.01. As another way of measuring relatedness, we  
616 also computed the effective number of chromosome segments ( $M_e$ )<sup>49</sup> in the two data.  $M_e$   
617 is a crucial parameter that measures the effective number of independent SNPs and is  
618 also closely related to the effective number of independent individuals:  $M_e$  is small in a  
619 data with related individuals and is large in a data with unrelated individuals. A small  
620 value of  $M_e$  often correlates to high prediction accuracy<sup>48,49,70</sup>. With 20 cross-validation  
621 replicates, we estimated  $M_e$  in D1 and D2 sub data to be 34541.39 (sd=140.87) and  
622 81786.01 (sd=651.52), respectively.

623 **Other methods.** We compared the performance of DPR.MCMC and DPR.VB mainly  
624 with five existing methods: (1) LMM<sup>28</sup> as implemented in the GEMMA software  
625 (version 0.95alpha); (2) BVSR<sup>29</sup> as implemented in the GEMMA software (version  
626 0.95alpha); (3) MultiBLUP<sup>41</sup> as implemented in the LDAK software (version 4.9); (4)  
627 BayesR<sup>35</sup> as implemented in the bayesR software; (5) rjMCMC<sup>20</sup> as implemented in the  
628 gwas\_rjmc1.163 software. We used default settings to fit all these methods. For rjMCMC,

629 because it requires us to provide a variance component parameter, we used LMM to  
630 estimate the variance component first in all analyses. In addition, rjMCMC does not  
631 output parameter estimates. Therefore, for the PrediXcan analysis, we first merged the  
632 GEUVADIS and WTCCC files for every gene, labeled WTCCC individuals as having  
633 missing phenotypes, and then ran rjMCMC on these files to obtain predicted gene  
634 expression values using the WTCCC genotype data. The same strategy was also applied  
635 to perform cross-validation prediction between D1 and D2 sub data sets in FHS. For gene  
636 expression prediction and PrediXcan analysis, following the original PrediXcan paper<sup>12</sup>,  
637 we also used Elastic Net (ENET)<sup>44</sup>, which is implemented in the R package glmnet  
638 (version 1.9-5). For ENET, following<sup>12</sup>, we set one penalty parameter (i.e.  $\alpha$ ) to be 0.5  
639 and selected the other one using 100-fold cross validation in the training data.

640 **Data availability.** No data were generated in the present study. The GEUVADIS gene  
641 expression data is publicly available at <http://www.geuvadis.org>. The genotype data from  
642 the 1000 genomes project is publicly available at <http://www.internationalgenome.org>.  
643 The WTCCC genotype and phenotype data is publicly available at  
644 <https://www.wtccc.org.uk>. The genotype and phenotype data from the cattle and maize  
645 studies are available from the authors upon reasonable request and with permission of  
646 Prof. Xiaolei Liu at the HuaZhong Agriculture University ([xiaoleiliu@mail.hzau.edu.cn](mailto:xiaoleiliu@mail.hzau.edu.cn)).  
647 The Framingham heart study genotype and phenotype data is available in dbGaP  
648 (<https://www.ncbi.nlm.nih.gov/gap>) with accession number phs000007.

649 **Software availability.** Our method is implemented in the DPR software, freely available  
650 at <http://www.xzlab.org/software.html>.



## 651 **References**

- 652 1. Fritsche, L. G. *et al.* A large genome-wide association study of age-related  
653 macular degeneration highlights contributions of rare and common variants. *Nat.*  
654 *Genet.* **48**, 134-143 (2016).
- 655 2. Mancuso, N. *et al.* The contribution of rare variation to prostate cancer  
656 heritability. *Nat. Genet.* **48**, 30-35 (2016).
- 657 3. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41-  
658 47 (2016).
- 659 4. The Wellcome Trust Case Control Consortium. Genome-wide association study  
660 of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**,  
661 661-678 (2007).
- 662 5. Global Lipids Genetics Consortium. Discovery and refinement of loci associated  
663 with lipid levels. *Nat. Genet.* **45**, 1274-1283 (2013).
- 664 6. Afshari, N. A. *et al.* Genome-wide association study identifies three novel loci in  
665 Fuchs endothelial corneal dystrophy. *Nat. Commun.* **8**, 14898 (2017).
- 666 7. Hoffmann, T. J. *et al.* Genome-wide association study of prostate-specific antigen  
667 levels identifies novel loci independent of prostate cancer. *Nat. Commun.* **8**,  
668 14248 (2017).
- 669 8. Warren, H. R. *et al.* Genome-wide association analysis identifies novel blood  
670 pressure loci and offers biological insights into cardiovascular risk. *Nat. Genet.*  
671 **49**, 403-415 (2017).
- 672 9. Makowsky, R. *et al.* Beyond Missing Heritability: Prediction of Complex Traits.  
673 *PLoS Genet.* **7**, e1002051 (2011).
- 674 10. Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., & Goddard, M. E.  
675 Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction:  
676 Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting  
677 Model Traits. *PLoS Genet.* **6**, e1001139 (2010).
- 678 11. Chatterjee, N., Shi, J., & Garcia-Closas, M. Developing and evaluating polygenic  
679 risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392-  
680 406 (2016).
- 681 12. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using  
682 reference transcriptome data. *Nat. Genet.* **47**, 1091-1098 (2015).
- 683 13. Allen, H. L. *et al.* Hundreds of variants clustered in genomic loci and biological  
684 pathways affect human height. *Nature* **467**, 832-838 (2010).
- 685 14. Li, H. *et al.* Genome-wide association study dissects the genetic architecture of  
686 oil biosynthesis in maize kernels. *Nat. Genet.* **45**, 43-50 (2013).
- 687 15. Romay, M. C. *et al.* Comprehensive genotyping of the USA national maize  
688 inbred seed bank. *Genome Biol.* **14**, R55 (2013).
- 689 16. Fernandes Júnior, G. A. *et al.* Genomic prediction of breeding values for carcass  
690 traits in Nellore cattle. *Genet. Sel. Evol.* **48**, 7 (2016).
- 691 17. Zhang, Z. *et al.* Accuracy of whole-genome prediction using a genetic  
692 architecture-enhanced variance-covariance matrix. *G3: Genes | Genomes |*  
693 *Genetics* **5**, 615-627 (2015).
- 694 18. Meuwissen, T., Hayes, B., & Goddard, M. Prediction of total genetic value using  
695 genome-wide dense marker maps. *Genetics* **157**, 1819-1829 (2001).

- 696 19. de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., & Sorensen,  
697 D. Prediction of complex human traits using the genomic best linear unbiased  
698 predictor. *PLoS Genet.* **9**, e1003608 (2013).
- 699 20. Lee, S. H., van der Werf, J. H. J., Hayes, B. J., Goddard, M. E., & Visscher, P. M.  
700 Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP  
701 Data. *PLoS Genet.* **4**, e1000231 (2008).
- 702 21. Hayes, B., Bowman, P., Chamberlain, A., & Goddard, M. Genomic selection in  
703 dairy cattle: Progress and challenges. *J. Dairy Sci.* **92**, 433-443 (2009).
- 704 22. Goddard, M. E., & Hayes, B. Genomic selection. *J. Anim. Breed. Genet.* **124**,  
705 323-330 (2007).
- 706 23. Meuwissen, T., Hayes, B., & Goddard, M. Accelerating improvement of livestock  
707 with genomic selection. *Annu. Rev. Anim. Biosci.* **1**, 221-237 (2013).
- 708 24. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on  
709 polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400-405  
710 (2013).
- 711 25. Shah, S. *et al.* Improving Phenotypic Prediction by Combining Genetic and  
712 Epigenetic Associations. *Am. J. Hum. Genet.* **97**, 75-85 (2015).
- 713 26. Maier, R. *et al.* Joint analysis of psychiatric disorders increases accuracy of risk  
714 prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am.*  
715 *J. Hum. Genet.* **96**, 283-294 (2015).
- 716 27. Weissbrod, O., Geiger, D., & Rosset, S. Multikernel: linear mixed models for  
717 complex phenotype prediction. *Genome Res.* **26**, 969-979 (2016).
- 718 28. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for  
719 human height. *Nat. Genet.* **42**, 565-569 (2010).
- 720 29. Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. Extension of the  
721 bayesian alphabet for genomic selection. *BMC Bioinformatics* **12**, 186 (2011).
- 722 30. Park, T., & Casella, G. The Bayesian Lasso. *J. Am. Stat. Assoc.* **103**, 681-686  
723 (2008).
- 724 31. Yi, N., & Xu, S. Bayesian LASSO for Quantitative Trait Loci Mapping. *Genetics*  
725 **179**, 1045-1055 (2008).
- 726 32. Hoggart, C. J., Whittaker, J. C., De Iorio, M., & Balding, D. J. Simultaneous  
727 analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS*  
728 *Genet.* **4**, e1000130 (2008).
- 729 33. Guan, Y., & Stephens, M. Bayesian variable selection regression for genome-  
730 wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5**, 1780-  
731 1815 (2011).
- 732 34. Zhou, X., Carbonetto, P., & Stephens, M. Polygenic modeling with Bayesian  
733 sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).
- 734 35. Moser, G. *et al.* Simultaneous Discovery, Estimation and Prediction Analysis of  
735 Complex Traits Using a Bayesian Mixture Model. *PLoS Genet.* **11**, e1004969  
736 (2015).
- 737 36. Goddard, M. Genomic selection: prediction of accuracy and maximisation of long  
738 term response. *Genetica* **136**, 245-257 (2009).
- 739 37. Ghahramani, Z. Bayesian non-parametrics and the probabilistic approach to  
740 modelling. *Philos. T. R. Soc. A* **371**, 20110553 (2013).

- 741 38. Müller, P., & Mitra, R. Bayesian Nonparametric Inference—Why and How.  
742 *Bayesian. Anal.* **8**, 269-302 (2013).
- 743 39. Gershman, S. J., & Blei, D. M. A tutorial on Bayesian nonparametric models. *J.*  
744 *Math. Psychol.* **56**, 1-12 (2012).
- 745 40. Müller, P., & Quintana, F. A. Nonparametric Bayesian Data Analysis. *Stat. Sci.*  
746 **19**, 95-110 (2004).
- 747 41. Speed, D., & Balding, D. J. MultiBLUP: improved SNP-based prediction for  
748 complex traits. *Genome Res.* **24**, 1550-1557 (2014).
- 749 42. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional  
750 variation in humans. *Nature* **501**, 506-511 (2013).
- 751 43. 1000 Genomes Project Consortium. An integrated map of genetic variation from  
752 1,092 human genomes. *Nature* **491**, 56-65 (2012).
- 753 44. Zou, H., & Hastie, T. Regularization and variable selection via the Elastic Net. *J.*  
754 *R. Stat. Soc. Ser. B.* **67**, 301-320 (2005).
- 755 45. Splansky, G. L. *et al.* The third generation cohort of the National Heart, Lung,  
756 and Blood Institute's Framingham Heart Study: design, recruitment, and initial  
757 examination. *Am. J. Epidemiol.* **165**, 1328-1335 (2007).
- 758 46. Hu, Z. L., Park, C. A., Wu, X. L., & Reecy, J. M. Animal QTLdb: an improved  
759 database tool for livestock animal QTL/association data dissemination in the post-  
760 genome era. *Nucleic Acids Res.* **41**, D871-D879 (2013).
- 761 47. Spiliopoulou, A. *et al.* Genomic prediction of complex human traits: relatedness,  
762 trait architecture and predictive meta-models. *Hum. Mol. Genet.* **24**, 4167-4182  
763 (2015).
- 764 48. Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. E. Using the genomic  
765 relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed.*  
766 *Genet.* **128**, 409-421 (2011).
- 767 49. Lee, S. H., Weerasinghe, W. M. S. P., Wray, N. R., Goddard, M. E., & van der  
768 Werf, J. H. J. Using information of relatives in genomic prediction to apply  
769 effective stratified medicine. *Scientific Reports* **7**, 42091 (2017).
- 770 50. Carbonetto, P., & Stephens, M. Scalable variational inference for Bayesian  
771 variable selection in regression, and its accuracy in genetic association studies.  
772 *Bayesian. Anal.* **7**, 73-108 (2012).
- 773 51. Yi, H., Breheny, P., Imam, N., Liu, Y., & Hoeschele, I. Penalized multimarker vs.  
774 single-marker regression methods for genome-wide association studies of  
775 quantitative traits. *Genetics* **199**, 205-222 (2015).
- 776 52. Sun, S. *et al.* Differential expression analysis for RNAseq using Poisson mixed  
777 models. *Nucleic Acids Res.* gkx204. doi: 210.1093/nar/gkx1204 (2017).
- 778 53. Tung, J., Zhou, X., Alberts, S. C., Stephens, M., & Gilad, Y. The genetic  
779 architecture of gene expression levels in wild baboons. *Elife* **4**, e04729 (2015).
- 780 54. Zhou, X. *et al.* Epigenetic modifications are associated with inter-species gene  
781 expression variation in primates. *Genome Biol.* **15**, 1 (2014).
- 782 55. Lea, A. J., Tung, J., & Zhou, X. A Flexible, Efficient Binomial Mixed Model for  
783 Identifying Differential DNA Methylation in Bisulfite Sequencing Data. *PLoS*  
784 *Genet.* **11**, e1005650 (2015).
- 785 56. Manolio, T. *et al.* Finding the missing heritability of complex diseases. *Nature*  
786 **461**, 747-753 (2009).

- 787 57. Shi, J. *et al.* Winner's Curse Correction and Variable Thresholding Improve  
788 Performance of Polygenic Risk Modeling Based on Genome-Wide Association  
789 Study Summary-Level Data. *PLoS Genet.* **12**, e1006493 (2016).
- 790 58. Li, J., Das, K., Fu, G., Li, R., & Wu, R. The Bayesian lasso for genome-wide  
791 association studies. *Bioinformatics* **27**, 516-523 (2011).
- 792 59. Blei, D. M., & Jordan, M. I. Variational inference for Dirichlet process mixtures.  
793 *Bayesian. Anal.* **1**, 121-143 (2006).
- 794 60. Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. Variational inference: A review  
795 for statisticians. *J. Am. Stat. Assoc.* (*in press*), *Preprint at*  
796 *<https://arxiv.org/abs/1601.00670>* (2017).
- 797 61. Ranganath, R., Tran, D., & Blei, D. M. (2016). *Hierarchical variational models*.  
798 Paper presented at the International Conference on Machine Learning.
- 799 62. Zhou, X. A Unified Framework for Variance Component Estimation with  
800 Summary Statistics in Genome-wide Association Studies. *Ann. Appl. Stat.* (*in*  
801 *press*), *Preprint at <http://biorxiv.org/content/early/2016/03/08/042846>* (2017).
- 802 63. Andrews, D. F., & Mallows, C. L. Scale mixtures of normal distributions. *J. R.*  
803 *Stat. Soc. Ser. B.* **36**, 99-102 (1974).
- 804 64. Verbyla, K. L., Hayes, B. J., Bowman, P. J., & Goddard, M. E. Accuracy of  
805 genomic selection using stochastic search variable selection in Australian  
806 Holstein Friesian dairy cattle. *Genet. Res.* **91**, 307-311 (2009).
- 807 65. Wen, X., Luca, F., & Pique-Regi, R. Cross-Population Joint Analysis of eQTLs:  
808 Fine Mapping and Functional Annotation. *PLoS Genet.* **11**, e1005176 (2015).
- 809 66. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The  
810 ENCODE Project. *Genome Res.* **22**, 1760-1774 (2012).
- 811 67. Stegle, O., Parts, L., Piipari, M., Winn, J., & Durbin, R. Using probabilistic  
812 estimation of expression residuals (PEER) to obtain increased power and  
813 interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500-507 (2012).
- 814 68. Guan, Y., & Stephens, M. Practical Issues in Imputation-Based Association  
815 Mapping. *PLoS Genet.* **4**, e1000279 (2008).
- 816 69. Howie, B. N., Donnelly, P., & Marchini, J. A Flexible and Accurate Genotype  
817 Imputation Method for the Next Generation of Genome-Wide Association Studies.  
818 *PLoS Genet.* **5**, e1000529 (2009).
- 819 70. Lee, S. H., Clark, S., & van der Werf, J. Estimation Of Genomic Prediction  
820 Accuracy From Reference Populations With Varying Degrees Of Relationship.  
821 *bioRxiv*, *Preprint at <http://biorxiv.org/content/early/2017/03/22/119164>* (2017).  
822  
823  
824  
825

## 826 **Acknowledgments**

827 This research is supported by National Institutes of Health grant R01HG009124. XZ is  
828 also supported by NIH Grants R01HL117626 (PI Abecasis), R21ES024834 (PI Pierce),  
829 R01HL133221 (PI Smith), and a grant from the Foundation for the National Institutes of  
830 Health through the Accelerating Medicines Partnership (BOEH15AMP, co-PIs Boehnke  
831 and Abecasis). We thank Xiaolei Liu from the HuaZhong Agriculture University  
832 ([xiaoleiliu@mail.hzau.edu.cn](mailto:xiaoleiliu@mail.hzau.edu.cn)) for providing us with the cattle and maize data. We thank  
833 Dr. Shiquan Sun for help with the development of DPR. We thank Dr. Doug Speed for  
834 help with LDAK and Dr. Sang Hong Lee for help with gwas\_rjmc1.163 software. This  
835 study also makes use of data generated by the Wellcome Trust Case Control Consortium  
836 (WTCCC). A full list of the investigators who contributed to the generation of the data is  
837 available from <http://www.wtccc.org.uk>. Funding for the WTCCC project was provided  
838 by the Wellcome Trust under award 076113 and 085475. This research was conducted in  
839 part using data and resources from the Framingham Heart Study of the NHLBI and  
840 Boston University School of Medicine, which was partially supported by the NHLBI  
841 Framingham Heart Study (Contract No. N01-HC-25195) and its contract with  
842 Affymetrix, Inc for genotyping services (Contract No. N02-HL-6-4278). We thank all  
843 participants and staff from the Framingham Heart Study. We thank the three anonymous  
844 reviewers for their constructive comments which greatly improved the quality of the  
845 paper.

## 846 **Author contributions**

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

850 **Competing financial interests:** The authors declare no competing financial  
851 interest.

852

## 853 **Additional information**

### 854 **Supplementary Information**

855 **Supplementary Figure 1. Comparison of prediction performance of several methods**  
856 **with DPR.MCMC in simulations when PVE=0.2.** Performance is measured by  $R^2$   
857 difference with respect to DPR.MCMC, where a negative value (i.e. values below the red  
858 horizontal line) indicates worse performance than DPR.MCMC. The sample  $R^2$   
859 differences are obtained from 20 replicates in each scenario. Methods for comparison  
860 include BVSR (cyan), BayesR (chocolate), LMM (purple), MultiBLUP (green), DPR.VB  
861 (red), rjMCMC (black blue) and DPR.MCMC. Simulation scenarios include: (A)  
862 Scenario I, which satisfies the DPR modeling assumption; (B) Scenario II, which  
863 satisfies the BayesR modeling assumption; (C) Scenario III, where the number of SNPs  
864 in the large effect group is 10, 100, or 1,000; and (D) Scenario IV, where the effect sizes  
865 are generated from either a normal distribution, a t-distribution or a Laplace distribution.  
866 For each box plot, the bottom and top of the box are the first and third quartiles, while the  
867 ends of whiskers represent either the lowest datum within 1.5 interquartile range of the  
868 lower quartile or the highest datum within 1.5 interquartile range of the upper quartile.  
869 For DPR.MCMC, the mean predictive  $R^2$  in the test set and the standard deviation for the  
870 eight settings are respectively 0.074 (0.020), 0.081 (0.016), 0.076 (0.018), 0.072 (0.019),  
871 0.064 (0.016), 0.083 (0.023), 0.077 (0.016) and 0.077 (0.017).

872 **Supplementary Figure 2. Comparison of prediction performance of several methods**  
873 **with DPR.MCMC in simulations when PVE=0.8.** Performance is measured by  $R^2$   
874 difference with respect to DPR.MCMC, where a negative value (i.e. values below the red  
875 horizontal line) indicates worse performance than DPR.MCMC. The sample  $R^2$   
876 differences are obtained from 20 replicates in each scenario. Methods for comparison  
877 include BVSR (cyan), BayesR (chocolate), LMM (purple), MultiBLUP (green), DPR.VB  
878 (red), rjMCMC (black blue) and DPR.MCMC. Simulation scenarios include: (A)  
879 Scenario I, which satisfies the DPR modeling assumption; (B) Scenario II, which  
880 satisfies the BayesR modeling assumption; (C) Scenario III, where the number of SNPs  
881 in the large effect group is 10, 100, or 1,000; and (D) Scenario IV, where the effect sizes

882 are generated from either a normal distribution, a t-distribution or a Laplace distribution.  
883 For each box plot, the bottom and top of the box are the first and third quartiles, while the  
884 ends of whiskers represent either the lowest datum within 1.5 interquartile range of the  
885 lower quartile or the highest datum within 1.5 interquartile range of the upper quartile.  
886 For DPR.MCMC, the mean predictive  $R^2$  in the test set and the standard deviation for the  
887 eight settings are respectively 0.554 (0.028), 0.622 (0.022), 0.569 (0.023), 0.548 (0.027),  
888 0.537 (0.030), 0.543 (0.028), 0.546 (0.027) and 0.539 (0.022).

889 **Supplementary Figure 3. Comparison of prediction performance of several methods**  
890 **with DPR.MCMC in simulations when PVE=0.2.** Performance is measured by MSE  
891 difference with respect to DPR.MCMC, where a positive value (i.e. values above the red  
892 horizontal line) indicates worse performance than DPR.MCMC. The sample MSE  
893 differences are obtained from 20 replicates in each scenario. Methods for comparison  
894 include BVSR (cyan), BayesR (chocolate), LMM (purple), MultiBLUP (green), DPR.VB  
895 (red), rjMCMC (black blue) and DPR.MCMC. Simulation scenarios include: (A)  
896 Scenario I, which satisfies the DPR modeling assumption; (B) Scenario II, which  
897 satisfies the BayesR modeling assumption; (C) Scenario III, where the number of SNPs  
898 in the large effect group is 10, 100, or 1,000; and (D) Scenario IV, where the effect sizes  
899 are generated from either a normal distribution, a t-distribution or a Laplace distribution.  
900 For each box plot, the bottom and top of the box are the first and third quartiles, while the  
901 ends of whiskers represent either the lowest datum within 1.5 interquartile range of the  
902 lower quartile or the highest datum within 1.5 interquartile range of the upper quartile.  
903 For DPR.MCMC, the mean predictive MSE in the test set and the standard deviation for  
904 the eight settings are respectively 0.919 (0.044), 0.910 (0.038), 0.929 (0.036), 0.944  
905 (0.053), 0.923 (0.038), 0.925 (0.033), 0.924 (0.037) and 0.918 (0.037).

906 **Supplementary Figure 4. Comparison of prediction performance of several methods**  
907 **with DPR.MCMC in simulations when PVE=0.5.** Performance is measured by MSE  
908 difference with respect to DPR.MCMC, where a positive value (i.e. values above the red  
909 horizontal line) indicates worse performance than DPR.MCMC. The sample MSE  
910 differences are obtained from 20 replicates in each scenario. Methods for comparison

911 include BVSR (cyan), BayesR (chocolate), LMM (purple), MultiBLUP (green), DPR.VB  
912 (red), rjMCMC (black blue) and DPR.MCMC. Simulation scenarios include: (A)  
913 Scenario I, which satisfies the DPR modeling assumption; (B) Scenario II, which  
914 satisfies the BayesR modeling assumption; (C) Scenario III, where the number of SNPs  
915 in the large effect group is 10, 100, or 1,000; and (D) Scenario IV, where the effect sizes  
916 are generated from either a normal distribution, a t-distribution or a Laplace distribution.  
917 For each box plot, the bottom and top of the box are the first and third quartiles, while the  
918 ends of whiskers represent either the lowest datum within 1.5 interquartile range of the  
919 lower quartile or the highest datum within 1.5 interquartile range of the upper quartile.  
920 For DPR.MCMC, the mean predictive MSE in the test set and the standard deviation for  
921 the eight settings are respectively 0.722 (0.043), 0.701 (0.028), 0.707 (0.034), 0.717  
922 (0.037), 0.727 (0.034), 0.734 (0.040), 0.721 (0.032) and 0.720 (0.028).

923 **Supplementary Figure 5. Comparison of prediction performance of several methods**  
924 **with DPR.MCMC in simulations when PVE=0.8.** Performance is measured by MSE  
925 difference with respect to DPR.MCMC, where a positive value (i.e. values above the red  
926 horizontal line) indicates worse performance than DPR.MCMC. The sample MSE  
927 differences are obtained from 20 replicates in each scenario. Methods for comparison  
928 include BVSR (cyan), BayesR (chocolate), LMM (purple), MultiBLUP (green), DPR.VB  
929 (red), rjMCMC (black blue) and DPR.MCMC. Simulation scenarios include: (A)  
930 Scenario I, which satisfies the DPR modeling assumption; (B) Scenario II, which  
931 satisfies the BayesR modeling assumption; (C) Scenario III, where the number of SNPs  
932 in the large effect group is 10, 100, or 1,000; and (D) Scenario IV, where the effect sizes  
933 are generated from either a normal distribution, a t-distribution or a Laplace distribution.  
934 For each box plot, the bottom and top of the box are the first and third quartiles, while the  
935 ends of whiskers represent either the lowest datum within 1.5 interquartile range of the  
936 lower quartile or the highest datum within 1.5 interquartile range of the upper quartile.  
937 For DPR.MCMC, the mean predictive MSE in the test set and the standard deviation for  
938 the eight settings are respectively 0.443 (0.032), 0.379 (0.016), 0.429 (0.024), 0.454  
939 (0.023), 0.464 (0.030), 0.465 (0.027), 0.454 (0.032) and 0.457 (0.022).



940 **Supplementary Figure 6. Comparison of predictive  $R^2$  from DPR.MCMC with the**  
941 **other six methods for predicting gene expression levels in the GEUVADIS data.**  
942 Scatter plots show (A) predictive  $R^2$  in the test data obtained by DPR.MCMC vs that  
943 obtained by BVSR for all genes; (B) DPR.MCMC vs ENET; (C) DPR.MCMC vs  
944 BayesR; (D) DPR.MCMC vs LMM; (E) DPR.MCMC vs MultiBLUP; (F) DPR.MCMC  
945 vs DPR.VB; (G) DPR.MCMC vs rjMCMC. Each panel also lists the number of genes  
946 where DPR.MCMC performs better (first number) and the number of genes where  
947 DPR.MCMC performs worse (second number).

948 **Supplementary Figure 7. Comparison of prediction performance of several methods**  
949 **with DPR.MCMC for twelve traits from three data sets.** Performance is measured by  
950 MSE difference with respect to DPR.MCMC, where a positive value (i.e. values above  
951 the red horizontal line) indicates worse performance than DPR.MCMC. Methods for  
952 comparison include BVSR (cyan), BayesR (chocolate), LMM (purple), MultiBLUP  
953 (green), DPR.VB (red), rjMCMC (black blue) and DPR.MCMC. The sample MSE  
954 differences are obtained from 20 replicates of Monte Carlo cross validation for each trait.  
955 For each box plot, the bottom and top of the box are the first and third quartiles, while the  
956 ends of whiskers represent either the lowest datum within 1.5 interquartile range of the  
957 lower quartile or the highest datum within 1.5 interquartile range of the upper quartile.  
958 For DPR.MCMC, the mean predictive MSE in the test set and the standard deviation are  
959 0.246 (0.011) for MFP, 0.371 (0.019) for MY, 0.446 (0.028) for SCS, 0.170 (0.012) for  
960 GDD, 0.928 (0.029) for LDL, 0.954 (0.034) for GLU, 0.833 (0.063) for HDL, 0.970  
961 (0.044) for TC, 0.960 (0.035) for TG, 0.519 (0.050) for height, 0.834 (0.065) for weight  
962 and 0.868 (0.074) for BMI. The SNP heritability estimates are 0.912 (0.007) for MFP,  
963 0.810 (0.012) for MY, 0.801 (0.012) for SCS, 0.880 (0.013) for GDD, 0.397 (0.024) for  
964 LDL, 0.357 (0.036) for GLU, 0.418 (0.024) for HDL, 0.402 (0.036) for TC, 0.334 (0.034)  
965 for TG, 0.905 (0.013) for Height, 0.548 (0.022) for Weight and 0.483 (0.023) for BMI.

966 **Supplementary Figure 8. Trace plots of the log posterior likelihood of DPR.MCMC**  
967 **in real data applications.** For each of the twelve traits in the three GWAS data sets, we  
968 plot the log posterior likelihood versus the first 10,000 iterations (i.e. burn-in period)

969 using the first cross-validation data. In each panel, the log posterior likelihood values  
970 were centered to have a median value of zero.

971 **Supplementary Figure 9. Comparison of prediction performance of several methods**  
972 **with DPR.MCMC for eight traits in each of the two sub data sets of FHS.** The two  
973 sub data sets D1 and D2 have the same sample size but different levels of relatedness  
974 (individuals in D1 are more related to each other than those in D2). (A) The  $R^2$  difference  
975 of five plasma traits (LDL, GLU, HDL, TC and TG) with respect to DPR.MCMC in the  
976 D1 and D2 sub data of FHS; (B) The  $R^2$  difference of three anthropometric traits (Height,  
977 Weight and BMI) with respect to DPR.MCMC in the D1 and D2 sub data of FHS. For  
978 each box plot, the bottom and top of the box are the first and third quartiles, while the  
979 ends of whiskers represent either the lowest datum within 1.5 interquartile range of the  
980 lower quartile or the highest datum within 1.5 interquartile range of the upper quartile.  
981 FHS: Framingham heart study.

982 **Supplementary Figure 10. Prediction performance of various methods are higher in**  
983 **a data with more related individuals (D1) than in a data with less related**  
984 **individuals (D2).** The two data sets D1 and D2 from FHS have the same sample size but  
985 different levels of relatedness (individuals in D1 are more related to each other than those  
986 in D2). For each trait in the FHS data (x-axis), we first computed the median predictive  
987  $R^2$  across 20 replicates in D1 and D2 separately, and then contrast the difference between  
988 the two averaged predictive  $R^2$  values in the two data sets (D1 minus D2; y-axis).  
989 Positive averaged predictive  $R^2$  differences suggest that all methods have higher  
990 predictive performance in D1 versus D2. FHS: Framingham heart study.

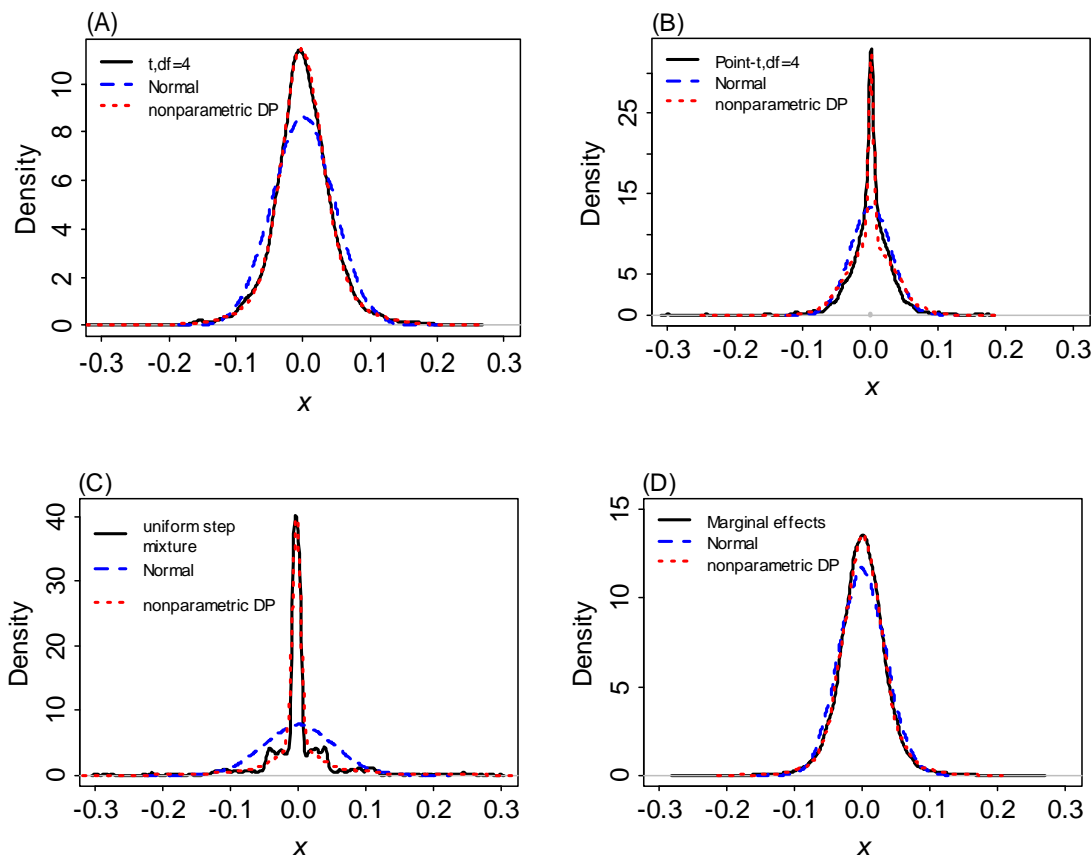
991 **Supplementary Figure 11. Comparison of prediction performance of several**  
992 **methods with DPR.MCMC using cross-validation between the two sub data sets of**  
993 **FHS.** The two sub data sets D1 and D2 have the same sample size but different levels of  
994 relatedness (individuals in D1 are more related to each other than those in D2). (A)  
995 Predictive  $R^2$  difference of different methods in D1 using parameters inferred in D2. For  
996 DPR.MCMC, the  $R^2$  is 0.024 for LDL, 0.012 for GLU, 0.021 for HDL, 0.022 for TC,

997 0.016 for TG, 0.131 for Height, 0.061 for Weight and 0.041 for BMI. (B) Predictive  $R^2$   
998 difference of different methods in D2 using parameters inferred in D1; For DPR.MCMC,  
999 the  $R^2$  is 0.043 for LDL, 0.009 for GLU, 0.033 for HDL, 0.021 for TC, 0.015 for TG,  
1000 0.226 for Height, 0.083 for Weight and 0.058 for BMI. FHS: Framingham heart study.

1001 **Supplementary Table 1. Sampling variation of  $R^2$  measured by standard deviation**  
1002 **across Monte Carlo cross validation replicates for various methods in simulations**  
1003 **and real data analysis.**

1004 **Supplementary Table 2. Significant genes identified by DPR.MCMC for different**  
1005 **diseases in the PrediXcan gene set analysis of WTCCC.**

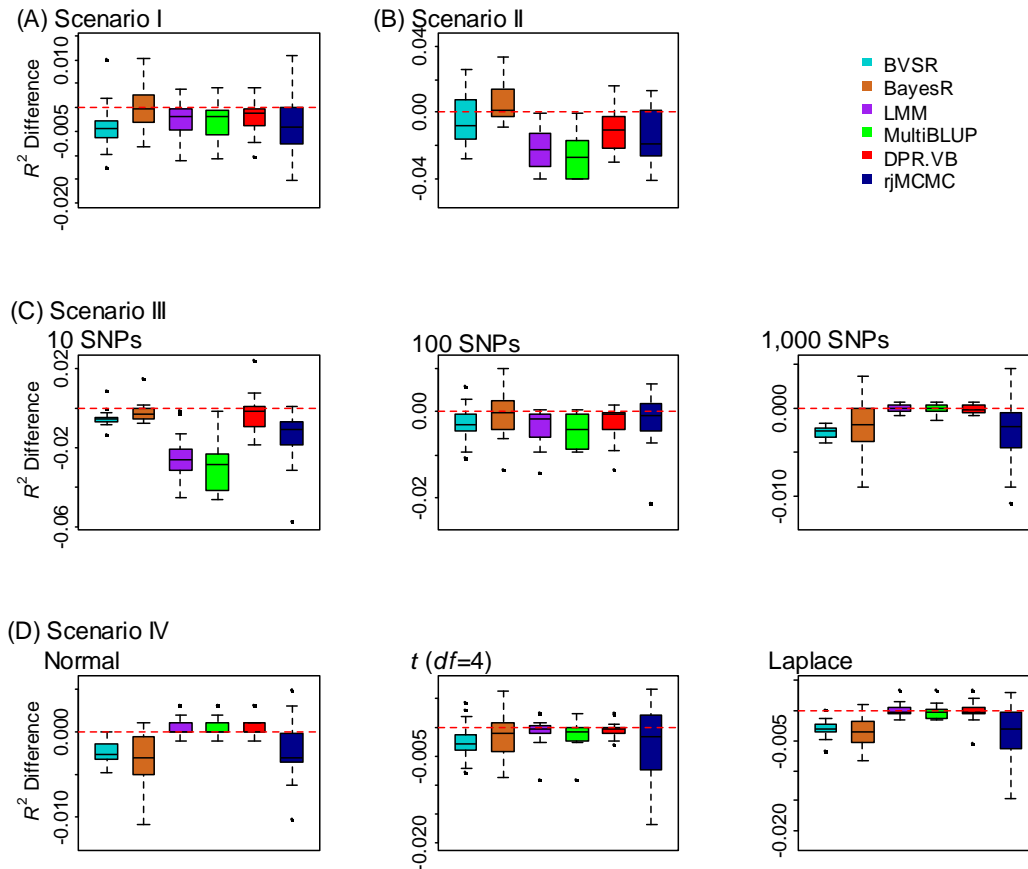
1006 **Supplementary Note. Model and Algorithm Details for DPR**  
1007



1008

1009

1010 **Figure 1. The induced non-parametric Dirichlet process (nonparametric DP)**  
 1011 **normal mixture prior on the effect sizes can be used to approximate a large number**  
 1012 **of unimodal distributions.** We either simulated 2,000 values from (A) a standard t-  
 1013 t-distribution with  $df=4$ ; (B) a point-t mixture distribution with the zero proportion being  
 1014 0.2, or equivalently,  $0.8 \times t(df=4) + 0.2 \times \delta_0$ , where  $\delta_0$  denotes a point mass at zero; (C) a  
 1015 four-component uniform step mixture distribution  $0.50 \times U(-0.05, 0.05) + 0.25 \times U(-$   
 1016  $0.3, 0.3) + 0.15 \times U(-0.8, 0.8) + 0.05 \times U(-2, 2)$ , where  $U$  denotes a uniform distribution; or  
 1017 obtained (D) the estimated marginal effect sizes from a linear mixed model in the cattle  
 1018 data with SCS (somatic cell score) as the phenotype. To make the first three data  
 1019 comparable with the last data in (D), we centered and scaled the values from the first  
 1020 three data for them to have a mean of zero and within the range of  $(-0.3, 0.3)$ . We then fit  
 1021 each data with either our non-parametric distribution (red) or a normal distribution (blue),  
 1022 and displayed the fitted curves on top of the sample distribution (black). Clearly, the non-  
 1023 parametric Dirichlet process normal mixture can approximate all these distributions well,  
 1024 while a simple normal distribution cannot.



1025

1026 **Figure 2. Comparison of prediction performance of several methods with**  
 1027 **DPR.MCMC in simulations when PVE=0.5.** Performance is measured by  $R^2$  difference  
 1028 with respect to DPR.MCMC, where a negative value (i.e. values below the red horizontal  
 1029 line) indicates worse performance than DPR.MCMC. The sample  $R^2$  differences are  
 1030 obtained from 20 replicates in each scenario. Methods for comparison include BVSr  
 1031 (cyan), BayesR (chocolate), LMM (purple), MultiBLUP (green), DPR.VB (red),  
 1032 rjMCMC (black blue) and DPR.MCMC. Simulation scenarios include: (A) Scenario I,  
 1033 which satisfies the DPR modeling assumption; (B) Scenario II, which satisfies the  
 1034 BayesR modeling assumption; (C) Scenario III, where the number of SNPs in the large  
 1035 effect group is 10, 100, or 1,000; and (D) Scenario IV, where the effect sizes are  
 1036 generated from either a normal distribution, a t-distribution or a Laplace distribution. For  
 1037 each box plot, the bottom and top of the box are the first and third quartiles, while the  
 1038 ends of whiskers represent either the lowest datum within 1.5 interquartile range of the  
 1039 lower quartile or the highest datum within 1.5 interquartile range of the upper quartile.

1040 For DPR.MCMC, the mean predictive  $R^2$  in the test set and the standard deviation for the  
1041 eight settings are respectively 0.272 (0.031), 0.299 (0.026), 0.295 (0.026), 0.281 (0.030),  
1042 0.277 (0.035), 0.278 (0.030), 0.282 (0.025) and 0.273 (0.022).

1043

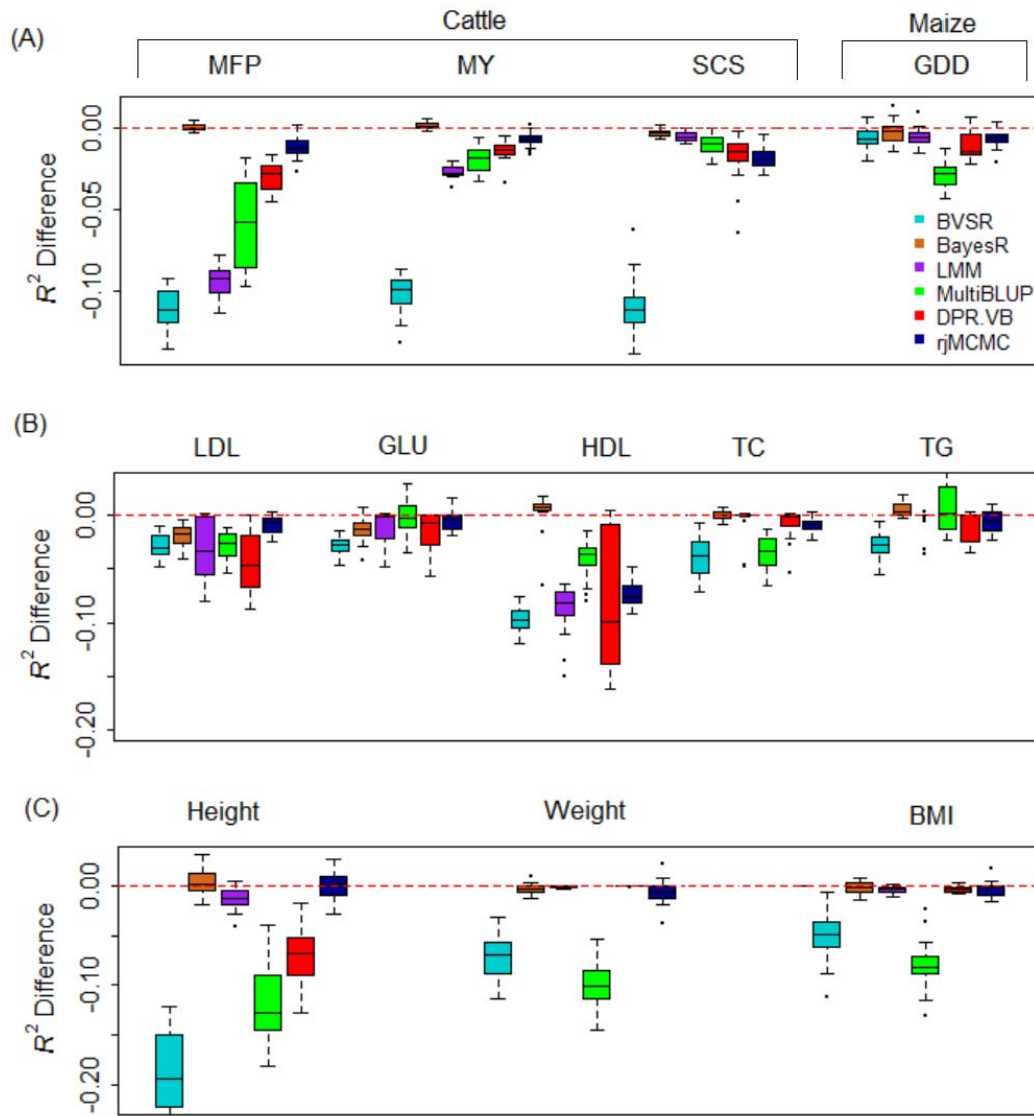
1044

1045

1046

1047

1048



1049

1050 **Figure 3. Comparison of prediction performance of several methods with**

1051 **DPR.MCMC for twelve traits from three data sets.** Performance is measured by  $R^2$

1052 difference with respect to DPR.MCMC, where a negative value (i.e. values below the red

1053 horizontal line) indicates worse performance than DPR.MCMC. Methods for comparison

1054 include BVSR (cyan), BayesR (chocolate), LMM (purple), MultiBLUP (green), DPR.VB

1055 (red), rjMCMC (black blue) and DPR.MCMC. For each box plot, the bottom and top of

1056 the box are the first and third quartiles, while the ends of whiskers represent either the

1057 lowest datum within 1.5 interquartile range of the lower quartile or the highest datum

1058 within 1.5 interquartile range of the upper quartile. The sample  $R^2$  differences are

1059 obtained from 20 replicates of Monte Carlo cross validation for each trait. For

1060 DPR.MCMC, the mean predictive  $R^2$  in the test set and the standard deviation across  
1061 replicates are 0.751 (0.011) for MFP, 0.624 (0.012) for MY, 0.551 (0.017) for SCS and  
1062 0.828 (0.012) for GDD, 0.081 (0.033) for LDL, 0.047 (0.017) for GLU, 0.153 (0.044) for  
1063 HDL, 0.050 (0.020) for TC, 0.044 (0.015) for TG, 0.478 (0.031) for height, 0.169 (0.038)  
1064 for weight and 0.137 (0.037) for BMI. The SNP heritability estimates are 0.912 (0.007)  
1065 for MFP, 0.810 (0.012) for MY, 0.801 (0.012) for SCS, 0.880 (0.013) for GDD, 0.397  
1066 (0.024) for LDL, 0.357 (0.036) for GLU, 0.418 (0.024) for HDL, 0.402 (0.036) for TC,  
1067 0.334 (0.034) for TG, 0.905 (0.013) for Height, 0.548 (0.022) for Weight and 0.483  
1068 (0.023) for BMI.

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092



1093

1094

1095

1096 **Table 1. Comparison of seven different methods in predicting gene expression levels**

1097 **in the GEUVADIS data.**

Threshold	ENET	BayesR	BVSR	LMM	MultiBLUP	rjMCMC	DPR	
							VB	MCMC
0.10	1061	809	486	<b>1195</b>	1098	1013	1163	<b>1280</b>
0.20	<b>449</b>	338	142	403	299	321	389	<b>467</b>
0.30	<b>182</b>	170	48	162	110	123	155	<b>194</b>
0.40	78	<b>84</b>	24	76	46	47	70	<b>86</b>
0.50	<b>37</b>	35	10	33	16	19	32	<b>38</b>
0.60	<b>15</b>	14	4	14	5	9	12	<b>17</b>
0.70	2	<b>3</b>	1	<b>3</b>	1	2	2	<b>3</b>

1098 To compare prediction performance, we counted the number of genes whose median  $R^2$

1099 across 20 replicates in the test set is above a given  $R^2$  threshold. A larger number thus

1100 indicates better performance. For each given threshold, we colored the best method with

1101 red and the second best method with blue.

1102

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121 **Table 2. Comparison of seven different methods in the PrediXcan gene set test in**  
1122 **the WTCCC data.**

Disease	ENET	BayesR	BVSR	LMM	MultiBLUP	rjMCMC	DPR	
							VB	MCMC
T1D	21	22	16	23	22	24	<b>26</b>	<b>25</b>
CD	<b>6</b>	0	1	4	4	<b>5</b>	3	<b>6</b>
RA	7	1	5	<b>9</b>	<b>8</b>	7	<b>8</b>	7
BD	0	0	0	0	0	0	0	0
CAD	0	0	0	0	0	0	0	0
HT	0	0	0	0	0	0	0	0
T2D	0	0	0	0	0	0	0	0
Total	34	23	22	36	34	36	<b>37</b>	<b>38</b>

1123 The table lists the number of genes passing the genome-wide significance threshold via  
1124 Bonferroni correction ( $\alpha=1.15\times 10^{-5}$ ) in each of the seven common diseases. For each  
1125 disease, we colored the best method with red and the second best method with blue.

1126

1127

1128

1129

1130

1131

1132

1133

1134

1135

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147 **Table 3. Mean computational time of the seven methods in the model fitting stage**  
 1148 **for twelve traits across three data sets.**

Data	Traits	BVSR	rjMCMC	BayesR	LMM	MultiBLUP	DPR	
							VB	MCMC
Cattle	MFP	2.26(0.49)	3.04(0.24)	5.01(0.75)	0.27(0.05)	0.40(0.12)	0.22(0.11)	6.29(3.07)
	MY	2.51(0.52)	2.95(0.31)	5.95(1.04)	0.27(0.08)	0.46(0.07)	0.21(0.09)	4.01(0.55)
	SCS	4.56(0.78)	3.15(0.27)	6.17(1.05)	0.24(0.04)	0.27(0.06)	0.20(0.08)	5.23(2.38)
Maize	GDD	2.38(0.72)	1.08(0.11)	7.86(1.57)	0.19(0.05)	0.03(0.01)	0.08(0.01)	4.53(1.29)
	LDL	1.02(0.17)	1.78(0.15)	78.56(27.78)	1.76(1.15)	1.71(0.33)	1.24(0.79)	85.76(18.22)
FHS	GLU	0.25(0.14)	1.86(0.18)	47.87(17.86)	1.06(0.52)	1.63(0.13)	0.43(0.12)	61.16(23.46)
	HDL	0.49(0.16)	1.83(0.14)	80.45(38.23)	3.39(1.26)	1.74(0.11)	1.28(0.56)	84.38(10.61)
	TC	0.24(0.13)	1.92(0.12)	51.17(16.72)	1.05(0.48)	1.62(0.37)	0.42(0.11)	51.69(11.77)
	TG	0.25(0.17)	1.98(0.15)	59.41(17.72)	0.99(0.35)	1.91(0.46)	0.45(0.13)	50.78(10.72)
	Height	0.68(0.16)	1.75(0.16)	71.14(13.80)	2.27(1.12)	4.13(1.18)	1.56(0.18)	71.62(11.89)
	Weight	0.59(0.13)	1.61(0.15)	72.66(12.15)	2.28(1.11)	1.95(0.34)	1.61(0.10)	79.67(15.04)
	BMI	0.47(0.10)	1.71(0.13)	76.08(15.28)	2.31(1.13)	2.35(0.27)	1.57(0.17)	75.15(14.91)

1149 The computational time is in hours. Values in parentheses are standard deviations. Mean  
 1150 and standard deviation are calculated based on 20 replicates. For MCMC based methods  
 1151 (rjMCMC, BVSR, BayesR and DPR.MCMC), the computational time is based on 50,000  
 1152 iterations of Metropolis Hastings steps for BVSR, reversible jump steps for rjMCMC,  
 1153 and Gibbs steps for BayesR and DPR.MCMC.

1154