Zeng *et al. BMC Genomics* (2017) 18:368
DOI 10.1186/s12864-017-3759-6

BMC Genomics

RESEARCH ARTICLE Open Access

# Prediction of gene expression with cis-SNPs using mixed models and regularization methods

Ping Zeng[1,2*], Xiang Zhou[2] and Shuiping Huang[1*]

## Abstract

**Background:** It has been shown that gene expression in human tissues is heritable, thus predicting gene expression using only SNPs becomes possible. The prediction of gene expression can offer important implications on the genetic architecture of individual functional associated SNPs and further interpretations of the molecular basis underlying human diseases.

**Methods:** We compared three types of methods for predicting gene expression using only cis-SNPs, including the polygenic model, i.e. linear mixed model (LMM), two sparse models, i.e. Lasso and elastic net (ENET), and the hybrid of LMM and sparse model, i.e. Bayesian sparse linear mixed model (BSLMM). The three kinds of prediction methods have very different assumptions of underlying genetic architectures. These methods were evaluated using simulations under various scenarios, and were applied to the Geuvadis gene expression data.

**Results:** The simulations showed that these four prediction methods (i.e. Lasso, ENET, LMM and BSLMM) behaved best when their respective modeling assumptions were satisfied, but BSLMM had a robust performance across a range of scenarios. According to $R^2$ of these models in the Geuvadis data, the four methods performed quite similarly. We did not observe any clustering or enrichment of predictive genes (defined as genes with $R^2 \geq 0.05$) across the chromosomes, and also did not see there was any clear relationship between the proportion of the predictive genes and the proportion of genes in each chromosome. However, an interesting finding in the Geuvadis data was that highly predictive genes (e.g. $R^2 \geq 0.30$) may have sparse genetic architectures since Lasso, ENET and BSLMM outperformed LMM for these genes; and this observation was validated in another gene expression data. We further showed that the predictive genes were enriched in approximately independent LD blocks.

**Conclusions:** Gene expression can be predicted with only cis-SNPs using well-developed prediction models and these predictive genes were enriched in some approximately independent LD blocks. The prediction of gene expression can shed some light on the functional interpretation for identified SNPs in GWASs.

**Keywords:** Gene expression, Cis-SNPs, Prediction model, Linear mixed model, Lasso, Elastic net, Bayesian sparse linear mixed model

* Correspondence: zpstat@xzhmu.edu.cn; hsp@xzhmu.edu.cn
[1]Department of Epidemiology and Biostatistics, Xuzhou Medical University, 209 Tongshan Rd, Xuzhou, Jiangsu 221004, China
Full list of author information is available at the end of the article

## Background

In the last decade tens of thousands of SNPs have been identified by genome wide association studies (GWASs) for many complex human diseases and traits [1–3], such as type I and II diabetes [4–7], lung cancer [8–11], Crohn's disease [12, 13], rheumatoid arthritis [13–18], blood pressure and hypertension [19–21], prostate cancer [22–26], height [27, 28], schizophrenia and bipolar disorder [29], and many others. These successes offer unprecedented insights into the genetic architectures of human diseases and traits, and may lead to clinically promising preventions and treatments for diseases in the future [30, 31]. However, the majority of identified SNPs in GWASs are located outside the protein-coding regions and their causal genetic mechanisms remain largely unknown. One way to explain this is that the identified SNPs are associated with molecular-level traits, such as methylation levels and gene expression levels, which are thought to mediate the effects of SNPs on many complex traits and diseases, and hold the key to understand the genetic basis of disease susceptibility and phenotypic variation. Recently, molecular QTL mapping have gained increasing attention [32–37], and have revealed that many cis-regulatory SNPs are not only related to diseases but also have influences on molecular phenotypes [37–39], e.g. gene expression levels which are quantitative molecular traits and can be influenced by cis-regulatory variants.

It has been found that gene expression in human tissues is heritable [38, 40, 41], meaning that predicting gene expression using only genetic variants is feasible. Gene expression levels can be effectively incorporated into models in a direct manner or in a mediated manner [42, 43], leading to a higher power for association and prediction. Additionally, accurate prediction of gene expression is a crucial step for transcriptome-wide association studies [34, 44] which attempt to construct a more biologically meaningful relationship between genes and diseases. Therefore, in addition to being significant interest in its own right for examining the relationship between SNPs and gene expression levels, knowledge of genetic variations in gene expression is also useful and important for association studies as well as phenotypic prediction [45]; integrative analysis of these information can result in a more accurate and powerful risk prediction and makes an advance towards to the precision medicine and personalized treatment of diseases. Most recently, it has been shown that, based on effective predicted values of gene expression, more powerful and interpretable gene-set tests in GWASs can be constructed [34]. Therefore, investigation of gene expression measurements can offer important implications on the genetic architecture of individual functional associated SNPs and provide further interpretations of the molecular basis underlying human diseases [32, 35, 37, 38].

Predicting complex phenotypes using genome-wide SNPs simultaneously has been increasingly used for human diseases and traits as well as animal and plant breeding [46–51], whereas predicting gene expression using SNPs is currently little studied. Based on regularized models it has recently been demonstrated [52, 53] that for some genes their expression measurements can be successfully predicted using only cis-SNPs, which are defined as SNPs located nearby a gene. In this paper we explore to predict gene expression with only cis-SNPs by borrowing two risk prediction models that are well studied and widely used in GWASs, i.e. linear mixed model (LMM) [46, 54–57] and Bayesian sparse linear mixed model (BSLMM) [58]. We evaluate the prediction performance of LMM and BSLMM with gene expression levels as phenotypes and compare them with the regularized models (i.e. Lasso and elastic net) previously employed in [34, 52, 53]. We use the Geuvadis gene expression data as an illustrative example.

## Methods

### Overview of Lasso, Elastic Net, LIMM and BSLMM

We first give a brief overview of the four methods (i.e. Lasso, elastic net, LMM and BSLMM) for predicting gene expression using only cis-SNPs. These methods are widely employed in phenotypic prediction of human complex traits and genomic selection in plant and animal breeding [51, 54, 56–68]. Compared with other methods, such as polygenic scores [29] and stepwise models, the four methods mentioned above have many advantages, e.g. they are numerically stable [69], can analyze all variants jointly while avoiding model overfitting, and incorporate the information of linkage disequilibrium (LD); thus they have the potential to improving prediction accuracy.

Let $\mathbf{y}$ be an $n$-vector of gene expression measured on $n$ individuals and assume it is centered; $\mathbf{X}$ is an $n$ by $p$ matrix of genotypes for $p$ cis-SNPs. Lasso [70] and elastic net (ENET) [71] are both popular regularization regressions, which select important cis-SNPs and estimate their effects simultaneously by imposing a penalty [34, 52, 63] on the cis-SNPs effect sizes. Specifically, Lasso and ENET fit the following linear model

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \frac{1}{n}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{'}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \sum_{j=1}^{p} P_{\lambda}\left(\left|\beta_j\right|\right), \\ \text{Lasso}: &\quad P_{\lambda} = \lambda|\beta|, \\ \text{ENET}: &\quad P_{\lambda} = \lambda\left(\alpha|\beta| + (1-\alpha)\beta^2\right), \end{aligned} \tag{1}$$

where $P_{\lambda}$ is the penalty function, $\lambda$ is the turning parameter controlling the extent of shrinkage, and $\alpha$ provides a mix between ridge regression and Lasso [70–73]. We ignore the intercept term in the model due to the fact that $\mathbf{y}$ is centered. The coordinate descent algorithm [74, 75] is employed to efficiently fit Lasso and ENET, and $\lambda$ is typically selected via *k*-fold cross validation [72]. Due to $P_\lambda$, small effects will be exactly estimated to be zero with reasonably selected turning parameter. Therefore, in this sense Lasso and ENET are sparse models. In contrast, LMM [46, 54–57] assumes every cis-SNP influences the gene expression measurements, with the effects sizes following a normal distribution

$$\begin{aligned} &\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \ \boldsymbol{\varepsilon}\sim N\left(0, \sigma_e^2\mathbf{I}_n\right), \\ &\beta_j\sim N\left(0, \sigma_b^2\sigma_e^2/p\right). \end{aligned} \tag{2}$$

Again we ignore the intercept term here. In model (2) $\sigma_e^2$ is the residual variance, $\mathbf{I}_n$ is an $n$-dimensional identity matrix, and $\sigma_b^2$ is the genetic variance scaled by $\sigma_e^2$. Note that the narrow-sense heritability $h^2$ can be defined as $\sigma_b^2/(\sigma_b^2 + 1)$ [55]. Because of assuming all variants have nonzero impacts on gene expression, LMM is thus a polygenic model [58, 76, 77]. We adopt the restricted maximum likelihood method to fit LMM using the efficient GEMMA algorithm [58, 78]. In GWASs, a few variants have displayed much larger effects than other SNPs. For example, the markers in major histocompatibility complex (MHC) region [79] in chromosome 6 show strong effects on some autoimmune diseases [13], e.g. type I diabetes, Crohn's disease and rheumatoid arthritis. To consider this, BSLMM [58] extends LMM by additionally incorporating SNPs with stronger effect sizes into the model. That is, BSLMM models the gene expression using

$$\begin{aligned} &\mathbf{y} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{u} + \boldsymbol{\varepsilon},\ \ \boldsymbol{\varepsilon}\sim N\left(0, \sigma_e^2\mathbf{I}_n\right), \\ &\mathbf{u}\sim N\left(0, \mathbf{K}\sigma_b^2\sigma_e^2\right), \\ &\tilde{\beta}\sim\pi N\left(0, \sigma_a^2\sigma_e^2\right) + (1-\pi)\delta_0, \end{aligned} \tag{3}$$

where $\mathbf{K}$ is the relatedness matrix, $\tilde{\beta}$ is the large SNP effect size, $\pi$ is probability that SNPs have large effect sizes, $\mathbf{u}$ can be viewed as the collection of small effects sizes, $\sigma_a^2$ is the corresponding variance, and $\delta_0$ is a point mass at zero. BSLMM is essentially a hybrid of LMM and sparse model via a spike and slab prior on affect sizes rather than imposing a penalty. In the special case of $\mathbf{K} = \mathbf{X}\mathbf{X}^T/p$, we can decompose the small effects sizes as $\mathbf{u} = \mathbf{X}\boldsymbol{\beta}$ with $\beta_j\sim N\left(0, \sigma_b^2\sigma_e^2/p\right)$. Based on re-parameterization [58], BSLMM is efficiently fit using Monte Carlo Markov Chain (MCMC) sampling. As BSLMM includes both LMM and sparse model as special cases, thus it is expected to enjoy both the advantages of LMM and sparse model.

## Simulations

We compared the performance of Lasso, ENET, LMM and BSLMM using simulations. To make our simulations much close to the real data, we used genotypes of gene *TPRG1L* from the Geuvadis program [80]. Briefly, there were a total of 465 individuals and 5,818 SNPs (minor allele frequency, or MAF, ≥0.05) in *TPRG1L*. We simulated gene expression $\mathbf{y}$ under three scenarios: (I) In addition to including all 5,818 SNPs into model as causal markers (the polygenic part), we also selected either 5 or 15 SNPs randomly with relatively large effect sizes (the sparse part). We simulated the effect sizes of the two parts from standard normal distributions and scaled the effects in each part separately so that the proportion of variance of gene expression explained (PVE) [58] by the two parts was 0.60 and 0.40, respectively. This scenario corresponded to the BSLMM modeling assumption. (II) We only modeled the polygenic part, i.e. all the SNPs were contained in the model with effect sizes following a standard normal distribution, corresponding to the LMM modeling assumption. (III) We only modeled the sparse part, i.e. again only either 5 or 15 SNPs with relatively large effect sizes were contained in the model, corresponding to the sparse modeling assumption in Lasso and ENET. In all the three scenarios the total PVE was set to 0.10, 0.30 or 0.50. In each scenario, we performed 20 simulation replicates. In each replicate, we randomly split the simulated data into a training data with 80% individuals and a test data with the rest 20% individuals. We then fit Lasso, ENET, LMM and BSLMM on the training data and assessed their performance in the test data. The performance was measured by the squared correlation coefficient ($R^2$) between the predicted values and the observed values in the test data. Both Lasso and ENET were implemented via the R package glmnet (version 2.0–5) [75], the penalty parameters in Lasso and ENET were selected using 100-fold cross validation. Additionally, we set $\alpha = 0.5$ in ENET as done in [34]. LMM and BSLMM were implemented via the GEMMA software (version 0.94) [58, 78]. For BSLMM we set both burn-in and MCMC sampling sizes to 10,000.

## Application to the Geuvadis data

The Geuvadis project [80] performed mRNA and small RNA sequencing on 465 Epstein-Barr-virus-transformed lymphoblastoid cell line samples from five populations. The genotype data was from the 1000 Genomes project [81]. Since the original gene expression measurements were read counts, the PEER normalization [82–84] was employed to remove technical variations and batch effects. We quantile-normalized every gene expression to a standard normal distribution separately in the five populations and then quantile-normalized together. According to GENCODE (release 12) [85], in the Geuvadis data we selected 15,810 genes that were expressed in at least half of the individuals. For each gene we only included common cis-SNPs (MAF ≥ 0.05) that were located within the gene or in the 1 Mb upstream and downstream regions near that gene, resulting in an average of about 580 SNPs per gene. Note that here only cis-SNPs are used due to the following reasons. First, it has been found that most expression quantitative trait loci (eQTL) are near the regulated gene and only a few eQTLs are trans-acting [33, 86]. Second, the effects of trans-SNPs are usually too weak to be detected with a reasonably high power [87]. Third, incorporating trans-SNPs into the model (e.g. using a two-variance-component model [88]) may improve the predictive accuracy, but with limited sample sizes the model fitting will become difficult and may lead to numerical issues. We randomly split each gene expression in the Geuvadis data into a training data with 80% individuals and a test data with the rest 20% individuals. We then fit Lasso, ENET, LMM and BSLMM on the training data and assessed their performance in the test data. Lasso and ENET were conducted using the R package glmnet (version 2.0–5) [75]. The penalty parameters of Lasso and ENET were selected via 100-fold cross validation. LMM and BSLMM were implemented via the GEMMA software (version 0.94) [58, 78]. For BSLMM we set burn-in to 2,000 and MCMC sampling size to 10,000.

## Results

The simulations show that these four prediction methods behave best when their individual modeling assumptions are satisfied. (The patterns are very similar for the two cases that there were 5 or 15 causal SNPs with relatively large effect sizes in scenarios I and III, so only results for 15 are displayed) For example, in scenario I where the BSLMM modeling assumptions were satisfied (Fig. 1a), BSLMM outperforms the other methods, whereas in scenarios II and III, as expected, the best methods are LMM and Lasso (or ENET), respectively. When the underlying model assumptions are not satisfied, LMM and Lasso (or ENET) are subject to reductions of prediction accuracy; for example, LMM in scenario II (Fig. 1b) and Lasso (or ENET) in scenario I or II (Fig. 1b and c). In contrast, BSLMM is very robust across various scenarios and has a compatible performance with the best method in scenarios II and III. For instance, BSLMM is only slightly worse than LMM in scenario II (Fig. 1b) where only polygenetic effect sizes were simulated, and behaves similarly to Lasso (or ENET) in scenario III (Fig. 1c) where only sparse effect sizes were included.

To compare the speed of these methods, we selected seven genes with various numbers of cis-SNPs. In terms of the computation time (Table 1), all the four methods are very fast, but LMM is more efficient than other methods. The computation speeds of Lasso, ENET and

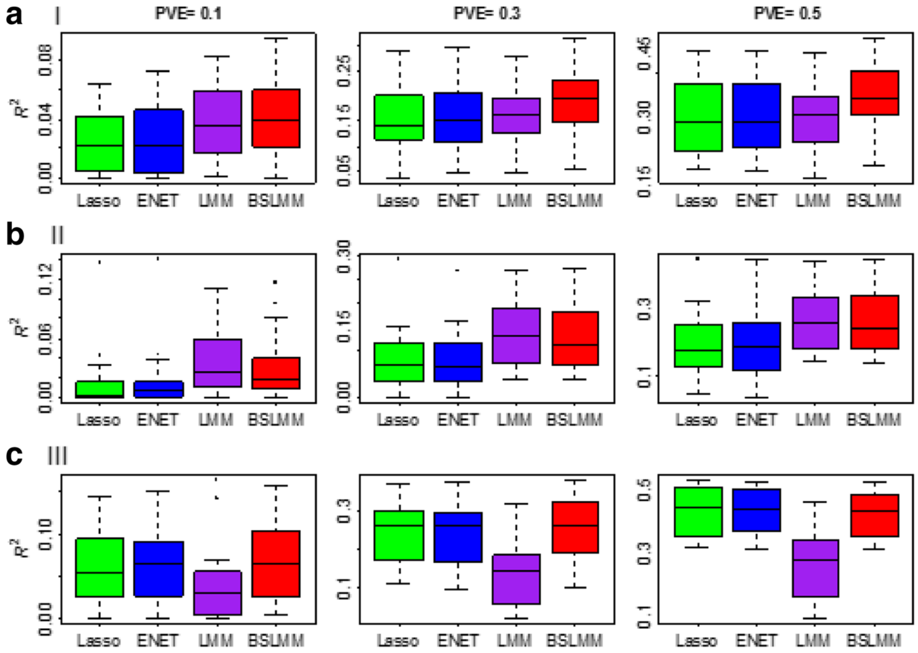

**Fig. 1** Comparison of the four methods (i.e. Lasso, ENET, LMM and BSLMM) for predicting gene expression in scenarios I-III. **a** The results of scenario I where the BSLMM modeling assumption is satisfied and 15 causal SNPs are included in the sparse part. **b** The results of scenario II where the LMM modeling assumption is satisfied. **c** The results of scenario III where the sparse modeling assumption is satisfied and there are only 15 causal SNPs and the rest are all neutral. The performance is measured by $R^2$. In each panel from left to right it corresponds to PVE = 0.1, 0.3 or 0.5 respectively

**Table 1** Computational time (in second) for the four models for predicting gene expression measurements

| #SNP | PVE | Lasso | ENET | LMM | BSLMM |
|---|---|---|---|---|---|
| 510 | 0.118 | 4.117 (0.203) | 2.937 (0.073) | 0.159 (0.148) | 1.780 (1.789) |
| 1375 | 0.002 | 6.594 (0.273) | 5.345 (0.110) | 0.560 (0.021) | 3.895 (0.917) |
| 2011 | 0.000 | 5.805 (0.172) | 5.134 (0.100) | 0.727 (0.076) | 1.502 (0.841) |
| 3045 | 0.357 | 8.623 (0.177) | 7.992 (0.234) | 1.097 (0.011) | 8.286 (8.159) |
| 4120 | 0.046 | 8.649 (0.282) | 8.385 (0.227) | 1.412 (0.073) | 16.129 (8.792) |
| 4953 | 0.523 | 10.019 (0.248) | 9.772 (0.285) | 1.621 (0.182) | 7.626 (3.406) |
| 5818 | 0.124 | 13.492 (0.199) | 13.077 (0.237) | 1.957 (0.057) | 2.269 (0.854) |

#SNP denotes the number of cis-SNPs included in this gene; PVE is the proportion of variance of gene expression explained by cis-SNPs; the tuning parameters of LASSO ENET are selected using 100-fold cross validation; BSLMM uses 10,000 Monte Carlo samplings after 2,000 burn-in samplings. The times are averaged across 20 replicates, and values in parentheses are the standard deviations

BSLMM are comparable and can vary with the number of cross validation or the burn-in and MCMC sampling sizes.

We now turn to the real application of the Geuvadis data. The predictive $R^2$ obtained from BSLMM versus other methods across all genes is presented in Fig. 2, where each panel also shows the number of genes for which BSLMM performs better and the number of genes for which BSLMM performs worse. In the top panel of Fig. 2a–c, these numbers are computed across all the genes, and in the bottom panel of Fig. 2d–f these numbers are computed across only the genes with predictive $R^2$ in the test data larger than 0.05. Table 2 lists the number of genes with a predictive $R^2$ above certain thresholds (from 0.05 to 0.60) for different methods. The four methods perform quite similarly to each other (Fig. 2 and Table 2). For example, the correlation coefficients of $R^2$ between BSLMM and other three methods are all above 0.970, and the correlation coefficient of $R^2$ between ENET and Lasso is even 0.999. Nevertheless, we can observe that BSLMM has a slightly higher predictive accuracy than other three methods. For instance, for these genes with $R^2 \geq 0.05$ (Fig. 2d–f), the difference of $R^2$ between BSLMM and LMM, BSLMM and Lasso, and BSLMM and ENET has a mean of $8.49 \times 10^{-3}$(standard deviation, or sd, $=3.33 \times 10^{-4}$), $7.67 \times 10^{-3}$ (sd $= 3.51 \times 10^{-4}$), and $7.53 \times 10^{-3}$ (sd $= 3.46 \times 10^{-4}$), respectively.

More interestingly, it is observed from Fig. 2 and Table 2 that in the Geuvadis data there is little predictive difference among Lasso, ENET and BSLMM for highly predictive genes (e.g. with $R^2 \geq 0.30$); whereas for these genes ($R^2 \geq 0.30$) LMM achieves a smaller $R^2$. We further validate this finding using another gene expression data from GenoExp [52]. The GenoExp data was obtained from the HapMap Phase II data set [89], include 210

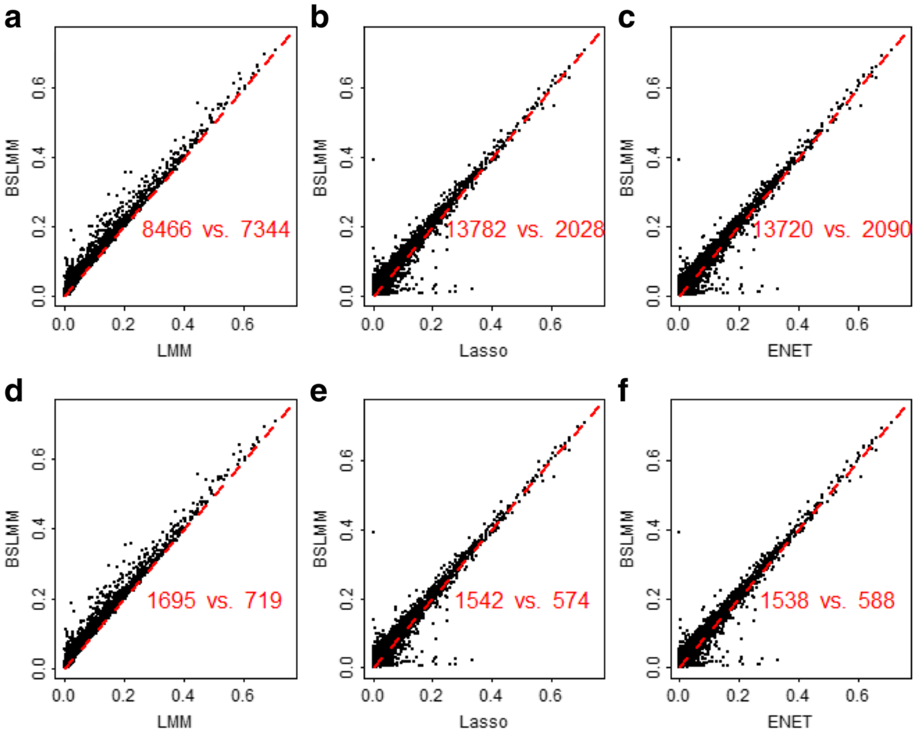

**Fig. 2** Comparison of the prediction performance of the four methods (i.e. Lasso, ENET, LMM and BSLMM) for the Geuvadis data. In each panel it lists the number of genes where BSLMM performs better and the number of genes where BSLMM performs worse; in the top (**a**)-(**c**), these numbers are computed across all the genes, and in the bottom (**d**)-(**f**) these numbers are computed across only the genes with $R^2 \geq 0.05$

**Table 2** Number of predictive genes passing the given $R^2$ threshold in the Geuvadis data and GenoExp data

| threshold | Geuvadis data | | | | GenoExp data | | | |
|---|---|---|---|---|---|---|---|---|
| | Lasso | ENET | LMM | BSLMM | Lasso | ENET | LMM | BSLMM |
| 0.05 | 2252 | 2262 | 2447 | 2567 | 1785 | 1414 | 1560 | 1758 |
| 0.10 | 1144 | 1145 | 1145 | 1266 | 831 | 788 | 734 | 826 |
| 0.20 | 420 | 422 | 383 | 466 | 315 | 309 | 276 | 323 |
| 0.30 | 161 | 162 | 152 | 178 | 156 | 148 | 124 | 160 |
| 0.40 | 75 | 75 | 65 | 76 | 70 | 70 | 56 | 70 |
| 0.50 | 33 | 33 | 25 | 32 | 36 | 32 | 27 | 37 |
| 0.60 | 14 | 14 | 12 | 14 | 25 | 21 | 20 | 24 |

There are 15,810 and 15,427 genes in the Geuvadis data and GenoExp data, respectively. It can be seen that in both data sets when the given $R^2$ threshold is large (e.g. ≥0.30) the number of predictive genes passing that value in LMM is less than that of LASSO, ENET or BSLMM, implying that these highly predictive genes may have a sparse genetic architecture

unrelated Epstein-Barr-virus-transformed lymphoblastoid cell line samples and 15,427 genes (with an average of about 304 cis-SNPs per gene). As before, for each gene the expression levels were quantile normalized to a standard normal distribution using the same procedure as in the Geuvadis data and were randomly divided into a training data with 80% individuals and a test data with the rest 20% individuals. We then fit Lasso, ENET, LMM and BSLMM on the training data and assessed their performance in the test data. For highly predictive genes ($R^2 \geq 0.30$) in the GenoExp data, it can be also seen (Table 2) that LMM have a smaller $R^2$ compared with Lasso, ENET and BSLMM, which validates our previous finding and, together with the result of the Geuvadis data, supports the recent finding that these highly predictive genes may be influenced by a few of cis-SNPs with relatively large effect sizes [90]; in other words, these highly predictive genes may have sparse genetic architectures.

To further see whether the predictive genes show special pattern across the genome, we display four plots in Fig. 3. However, we do not observe any obvious clustering or enrichment of $R^2$ across the chromosomes (Fig. 3a and b), and we also do not see there is any clear relationship between the proportion of the predictive genes ($R^2 \geq 0.05$) and the proportion of genes in each chromosome (Fig. 3c). The predictive genes are defined the genes with $R^2 \geq 0.05$, which means that about 5% variation of gene expression is explained by only cis-SNPs

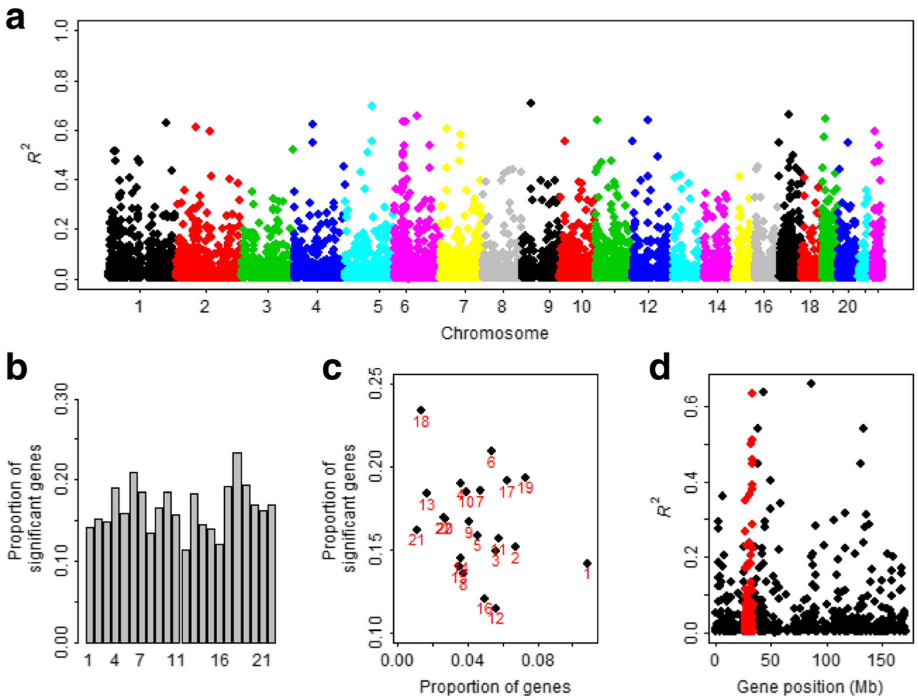

**Fig. 3** Distribution of $R^2$ of BSLMM for the Geuvadis data. **a** A Manhattan-type plot shows $R^2$ and gene positions across chromosomes, in which the y-axis is $R^2$ for each gene, the x-axis is the gene position and the various colors represent different chromosomes. **b** The barplot shows the proportion of predictive genes ($R^2 \geq 0.05$) for each chromosome. **c** The scatter of the proportion of the predictive genes against the proportion of gene in each chromosome. **d** The $R^2$ pattern for the MHC region (chr6: 26-34 Mb); there are a total of 179 genes with $R^2 \geq 0.05$ in chromosome 6, among which 45 are located on the MHC region (in red). The total length of chromosome 6 is about 171 Mb, and the length of the MHC region is 8 Mb. Then the enrichment-fold is 5.37, which is computed as the ratio of the proportion of predictive genes (i.e. 0.25 = 45/179) and the proportion of the length of MHC (i.e. 0.05 = 8/171), and is significantly higher ($P = 1.79 \times 10^{-3}$) than the average enrichment-fold (the median is 1.70) of other regions in chromosome 6

and is selected arbitrarily to some extent; although other larger values can be also used and may lead to different results, the conclusions can not be changed. However, we indeed find enrichments of predictive genes in some special genetic regions. For example, for the MHC region of chromosome 6 (Fig. 3d), there are a total of 179 genes with $R^2 \geq 0.05$ in chromosome 6, among which 45 are located in the MHC region. The total length of chromosome 6 is about 171 Mb, and the length of the MHC region is about 8 Mb (from 26 Mb to 34 Mb [91]). Then the enrichment-fold is 5.37, which is computed as the ratio of the proportion of predictive genes (i.e. 0.25 = 45/179) and the proportion of the length of MHC (i.e. 0.05 = 8/171), and is significantly higher than the average enrichment-fold (the median is 1.70) of other regions in chromosome 6 ($P = 1.79 \times 10^{-3}$ based on an approximate z test [92]). For the Geuvadis data we obtained 1,324 approximately independent blocks (1.6 Mb on average) (Fig. 4) of LD [51], with the median enrichment-fold being 1.49. Among these, there are 17 LD blocks with enrichment-fold ≥ 20 (Table 3), within which it has been identified by previous GWASs [93] that many SNPs are related to a lot of complex diseases and traits, including type 2 diabetes, aging-related traits, blood pressure, body mass index, bipolar disorder, Crohn's disease, lung cancer, obesity, schizophrenia and coronary heart disease. Therefore, the enrichment of predictive genes in these LD blocks may provide important implications for the underlying functional basis of identified SNPs in GWASs.

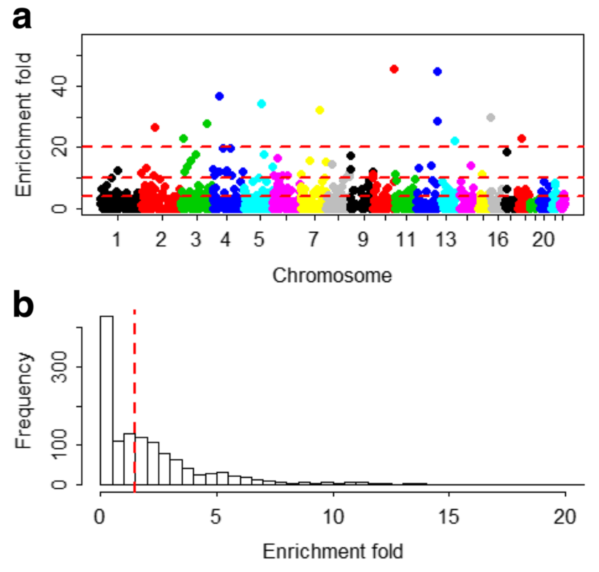

**Fig. 4** Enrichment-fold in 1,324 approximately independent LD blocks. **a** The enrichment-fold distributed across the chromosomes; the reference lines are 4, 10 and 20, respectively; (**b**) The histogram of enrichment-fold in 1,324 independent LD blocks; the median is 1.49 (indicating with red reference line) and the maximum is 299.82. The enrichment-fold is computed as the ratio of the proportion of predictive genes (i.e. $R^2 \geq 0.05$) and the proportion of the length of that LD block

**Table 3** Enrichment-fold (≥20) of independent LD blocks in the Geuvadis data

| Enrichment fold | #Identified SNPs | Chromosome | LD block | |
|---|---|---|---|---|
| | | | lower | upper |
| 26.58 | 16 | 2 | 84,687,169 | 84,743,579 |
| 53.20 | 11 | 2 | 152,118,393 | 152,146,571 |
| 22.76 | 17 | 3 | 19,988,517 | 20,053,822 |
| 184.35 | 8 | 3 | 75,713,481 | 75,721,542 |
| 27.75 | 17 | 3 | 161,090,668 | 161,144,215 |
| 36.70 | 15 | 4 | 44,680,444 | 44,728,612 |
| 81.95 | 8 | 4 | 47,465,736 | 47,487,305 |
| 34.18 | 13 | 5 | 107,006,596 | 107,052,542 |
| 32.41 | 25 | 7 | 120,965,421 | 121,036,418 |
| 299.82 | 29 | 10 | 18,940,551 | 18,948,334 |
| 45.61 | 12 | 10 | 131,909,081 | 131,934,663 |
| 28.64 | 8 | 12 | 127,210,816 | 127,256,957 |
| 44.88 | 23 | 12 | 129,308,528 | 129,337,972 |
| 22.28 | 17 | 13 | 101,241,782 | 101,327,347 |
| 29.80 | 13 | 16 | 5,084,142 | 5,147,789 |
| 23.06 | 22 | 18 | 23,671,164 | 23,806,409 |
| 151.19 | 16 | 18 | 61,616,535 | 61,637,159 |

We obtained a total of 18,896 complete records (mainly including the information of disease/trait, chromosome id and position) of identified SNPs by GWASs from https://www.genome.gov/gwastudies/. We counted the number (given in the second column) of related SNPs within 1 Mb upstream and downstream regions near each LD block. These identified SNPs are extensively related to about 130 different types of complex diseases and traits. For example, in the first LD block (Chr2: 84,687,169-84,743,579), previous GWASs have discovered 16 associated SNPs, which, in terms of the catalog of published GWASs, are related to aging traits, protein quantitative trait loci, pulmonary function decline, IgG glycosylation, RR interval heart rate, the response to antipsychotic therapy, coronary artery calcification, prostate cancer, response to cytadine analogues cytosine arabinoside, bilirubin levels, orthostatic hypotension, breast cancer and conduct disorder

## Discussion and conclusions

In this paper we have explored to predict gene expression using only cis-SNPs and compared four prediction methods (i.e. Lasso, ENET, LMM and BSLMM). The four methods represent three types of prediction approaches that are widely used for genetic data in which the number for predictors (i.e. SNPs) is typically larger than the sample size [57, 62, 66, 94, 95]. Lasso and ENET assume the underlying model is sparse and only include important cis-SNPs into the model by regularization. In contrast to the sparsity, LMM assumes all cis-SNPs have impacts on the gene expression and thus is an explicit polygenic model. BSLMM combines the sparse model and LMM, and can have the benefits of both the models. Therefore, as shown in simulations the sparse model and LMM work well under individual model assumption, but become worse when their model assumptions are not met. On the other hand, BSLMM has a robust performance across different scenarios and is the best model or performs comparably with the best model.

Note that there are other risk prediction methods that are not considered here. For example, the Bayes-alphabet models [58, 59], which use slight different mixture priors from BSLMM and thus should have similar performance. BayesR [96] and Multi-BLUP [97] are more recently developed risk prediction methods, but they typically require more dense SNPs to achieve a better prediction accuracy, thus may improve little compared with BSLMM in the context of gene expression prediction. Besides single-trait prediction methods, multi-trait prediction approaches have also attracted significant recent attention. It has been shown that by leveraging shared genetic basis underlying correlated phenotypes multi-trait prediction approaches are typically more powerful than single-trait prediction methods [98–100]. Since multiple gene expression levels in an independent LD block may be highly correlated and have common genetic basis, analyzing a set of gene expression levels jointly using multi-trait approaches is expected to offer a potential to further increasing prediction accuracy. We will investigate this interesting problem in our further work.

In the application of the Geuvadis gene expression data, the four methods behave similarly; but it is very interesting that BSLMM and the two sparse models (i.e. Lasso and ENET) have a better performance for some genes that have high $R^2$ (e.g. ≥0.30), more importantly, this finding is further validated in an external data set, suggesting that these highly predictive genes may have sparse genetic architectures [90]. In the Geuvadis data, we also find that the predictive genes are enriched in some approximately independent LD blocks, meaning that for some special genome regions (e.g. MHC) in human [79] the gene expression values are more predictive relative to other regions, and thus can provide further useful insights for revealing the biological function of regulatory variants.

According to the computational efficiency, LMM is the fastest method; BSLMM, Lasso and ENET are computationally comparable. As we use the R package glmnet [75] to conduct Lasso and ENET, which may limit their utility for larger data set; but this limitation seems to not be a problem in the context of gene expression prediction using cis-SNPs, since currently the sample size of the gene expression data is relatively small. On the other hand, LMM and BSLMM are performed using the GEMMA software [58, 78], which can be applicable to large scale data set. Note that the computation time is dependent not only on implementational environment, computer language, the number of cis-SNPs and the sample sizes but also on other factors, for instance, the number of the cross-validation used in Lasso and ENET, and the burn-in steps and the posterior sampling steps in BSLMM.

Finally, we need to emphasize that like in [52] the prediction accuracies of these models are still low for most genes, although we discover some gene expression levels can be effectively predicted by cis-SNPs in the Geuvadis data. There may be other factors that are also responsible for gene expression, such as trans-SNPs and environmental factors. In summary, in this paper we have demonstrated that gene expression can be predicted with only cis-SNPs using well-developed prediction models that are commonly-used in GWASs and the prediction of gene expression can shed some light on the functional interpretation for these identified SNPs in GWASs.

### Abbreviations

BSLMM: Bayesian sparse linear mixed model; ENET: Elastic net; GWASs: Genome wide association studies; LD: Linkage disequilibrium; LMM: Linear mixed model; MCMC: Monte Carlo Markov chain; MHC: Major histocompatibility complex; PVE: Proportion of variance of explained

### Acknowledgements

We are grateful to the Geuvadis consortium for making the Geuvadis gene expression data available to us. We thank the Editor and two referees for their useful comments which have helped to improve the paper.

### Funding

This work was supported by National Natural Science Foundation of China (no. 81402765), Statistical Science Research Project from National Bureau of Statistics of China (no. 2014LY112), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

### Availability of data and materials

The Geuvadis data is publicly available from http://www.geuvadis.org/web/geuvadis. The GEMMA software is available from http://www.xzlab.org/software.html, the glmnet package is available from https://cran.r-project.org/web/packages/glmnet/index.html.

### Authors' contributions

PZ, XZ and SH conceived and designed the experiment, PZ analyzed and interpreted the data, PZ and SH wrote the manuscript. All authors read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interests.

### Consent for publication

Not applicable.

### Ethics approval and consent to participate

Not applicable.

## Publisher's Note

### Author details

[1]Department of Epidemiology and Biostatistics, Xuzhou Medical University, 209 Tongshan Rd, Xuzhou, Jiangsu 221004, China. [2]Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48104, USA.

Received: 20 October 2016 Accepted: 3 May 2017
Published online: 11 May 2017

## References

1. Visscher P, Brown M, McCarthy M, Yang J. Five Years of GWAS Discovery. Am J Hum Genet. 2012;90(1):7–24.
2. Hindorff L, Sethupathy P, Junkins H, Ramos E, Mehta J, Collins F, Manolio T. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106(23):9362–7.
3. Zeng P, Zhao Y, Qian C, Zhang L, Zhang R, Gou J, Liu J, Liu L, Chen F. Statistical analysis for genome-wide association study. J Biomed Res. 2015;29(4):285–97.
4. Hakonarson H, Grant SF, Bradfield JP, Marchand L, Kim CE, Glessner JT, Grabs R, Casalunovo T, Taback SP, Frackelton EC. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature. 2007;448(7153):591–4.
5. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445(7130):881–5.
6. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, Ma C, Fontanillas P, Moutsianas L, McCarthy DJ, et al. The genetic architecture of type 2 diabetes. Nature. 2016;536(7614):41–7.
7. Steinthorsdottir V, Thorleifsson G, Sulem P, Helgason H, Grarup N, Sigurdsson A, Helgadottir HT, Johannsdottir H, Magnusson OT, Gudjonsson SA, et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat Genet. 2014;46(3):294–8.
8. Dong J, Hu Z, Wu C, Guo H, Zhou B, Lv J, Lu D, Chen K, Shi Y, Chu M, et al. Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. Nat Genet. 2012;44(8):895–9.
9. Hu Z, Wu C, Shi Y, Guo H, Zhao X, Yin Z, Yang L, Dai J, Hu L, Tan W, et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in han chinese. Nat Genet. 2011;43(8):792–6.
10. Lan Q, Hsiung CA, Matsuo K, Hong YC, Seow A, Wang Z, Hosgood HD, Chen K, Wang J-C, Chatterjee N, et al. Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. Nat Genet. 2012;44(12):1330–5.
11. Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, Zong X, Laplana M, Wei Y, Han Y, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet. 2014;46(7):736–41.
12. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010;42(12):1118–25.
13. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78.
14. Jiang L, Yin J, Ye L, Yang J, Hemani G, Liu A, Zou H, He D, Sun L, Zeng X. Novel risk loci for rheumatoid arthritis in Han Chinese and congruence with risk variants in Europeans. Arthritis Rheumatol. 2014;66(5):1121–32.
15. Orozco G, Viatte S, Bowes J, Martin P, Wilson AG, Morgan AW, Steer S, Wordsworth P, Hocking LJ, Barton A. Novel Rheumatoid Arthritis Susceptibility Locus at 22q12 Identified in an Extended UK Genome-Wide Association Study. Arthritis Rheumatol. 2014;66(1):24–30.
16. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, Yoshida S. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014;506(7488):376–81.
17. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FA, Zhernakova A, Hinks A. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet. 2010;42(6):508–14.
18. Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, Kraft P, Chen R, Kallberg HJ, Kurreeman FAS, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet. 2012;44(5):483–9.
19. Ehret GB, Ferreira T, Chasman DI, Jackson AU, Schmidt EM, Johnson T, Thorleifsson G, Luan J, Donnelly LA, Kanoni S, et al. The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. Nat Genet. 2016;48(10):1171–84.
20. Liu C, Kraja AT, Smith JA, Brody JA, Franceschini N, Bis JC, Rice K, Morrison AC, Lu Y, Weiss S, et al. Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. Nat Genet. 2016;48(10):1162–70.
21. Surendran P, Drenos F, Young R, Warren H, Cook JP, Manning AK, Grarup N, Sim X, Barnes DR, Witkowska K, et al. Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension. Nat Genet. 2016;48(10):1151–61.
22. Mancuso N, Rohland N, Rand KA, Tandon A, Allen A, Quinque D, Mallick S, Li H, Stram A, Sheng X, et al. The contribution of rare variation to prostate cancer heritability. Nat Genet. 2016;48(1):30–5.
23. Eeles RA, Kote-Jarai Z, Al Olama AA, Giles GG, Guy M, Severi G, Muir K, Hopper JL, Henderson BE, Haiman CA. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. Nat Genet. 2009;41(10):1116–21.
24. Kote-Jarai Z, Al Olama AA, Giles GG, Severi G, Schleutker J, Weischer M, Campa D, Riboli E, Key T, Gronberg H. Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. Nat Genet. 2011;43(8):785–91.
25. Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, Benediktsdottir KR, Sigurdsson A, Magnusson OT, Gudjonsson SA, Magnusdottir DN, et al. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nat Genet. 2012;44(12):1326–9.
26. Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, Yu K, Chatterjee N, Welch R, Hutchinson A, et al. Multiple loci identified in a genome-wide association study of prostate cancer. Nat Genet. 2008;40(3):310–5.
27. Allen HL, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467(7317):832–8.
28. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46(11):1173–86.
29. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P, Ruderfer DM, McQuillin A, Morris DW. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748–52.
30. Florez JC. Leveraging Genetics to Advance Type 2 Diabetes Prevention. PLoS Med. 2016;13(7):e1002102.
31. Chatterjee N, Shi J, Garcia-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet. 2016;17(7):392–406.
32. Wen X, Lee Y, Luca F, Pique-Regi R. Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. Am J Hum Genet. 2016;98(6):1114–29.
33. Wen X, Luca F, Pique-Regi R. Cross-Population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation. PLoS Genet. 2015;11(4):e1005176.
34. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Consortium GT, Nicolae DL, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091–8.
35. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJH, Shishkin AA, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518(7539):337–43.
36. Kichaev G, Yang W-Y, Lindstrom S, Hormozdiari F, Eskin E, Price AL, Kraft P, Pasaniuc B. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. 2014;10(10):e1004722.
37. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014;24(1):14–24.
38. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, et al. Population genomics of human gene expression. Nat Genet. 2007;39(10):1217–24.
39. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms

underlying human gene expression variation with RNA sequencing. Nature. 2010;464(7289):768–72.

40. Cheung V, Spielman R, Ewens K, Weber T, Morley M, Burdick J. Mapping determinants of human gene expression by regional and genome-wide association. Nature. 2005;437(7063):1365–9.
41. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G. Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003;422(6929):297–302.
42. Huang Y-T, VanderWeele TJ, Lin X. Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. Ann Appl Stat. 2014;8(1):352–76.
43. Huang Y-T. Integrative modeling of multiple genomic data from different types of genetic association studies. Biostatistics. 2014;15(4):587–602.
44. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, Jansen R, de Geus EJC, Boomsma DI, Wright FA, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48(3):245–52.
45. Fay JC, McCullough HL, Sniegowski PD, Eisen MB. Population genetic variation in gene expression is associated with phenotypic variation in Saccharomyces cerevisiae. Genome Biol. 2004;5(4):R26.
46. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics. 2013;193(2):327–45.
47. Crossa J, Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, et al. Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics. 2010;186(2):713–24.
48. Jannink JL, Lorenz AJ, Iwata H. Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics. 2010;9(2):166–77.
49. Hayes BJ, Pryce J, Chamberlain AJ, Bowman PJ, Goddard ME. Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. PLoS Genet. 2010;6(9):e1001139.
50. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009;10(6): 381–91.
51. Heslot N, Yang HP, Sorrells ME, Jannink JL. Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci. 2012;52(1):146–60.
52. Manor O, Segal E. Robust prediction of expression differences among human individuals using only genotype information. PLoS Genet. 2013;9(3): e1003396.
53. Manor O, Segal E. GenoExp: a web tool for predicting gene expression levels from single nucleotide polymorphisms. Bioinformatics. 2015;31(11): 1848–50.
54. de los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet. 2010; 11(12):880–6.
55. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565–9.
56. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de Los Campos G. Beyond Missing Heritability: Prediction of Complex Traits. PLoS Genet. 2011;7(4):e1002051.
57. de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 2013;9(7):e1003608.
58. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2013;9(2):e1003264.
59. Meuwissen T, Hayes B, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–29.
60. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 2007;17(10): 1520–8.
61. Park T, Casella G. The Bayesian Lasso. J Am Stat Assoc. 2008;103(482):681–6.
62. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk of complex disease. Curr Opin Genet Dev. 2008;18(3):257–63.
63. Kooperberg C, LeBlanc M, Obenchain V. Risk prediction using genome-wide association studies. Genet Epidemiol. 2010;34(7):643–52.
64. Li J, Das K, Fu G, Li R, Wu R. The Bayesian lasso for genome-wide association studies. Bioinformatics. 2011;27(4):516–23.
65. Erbe M, Hayes B, Matukumalli L, Goswami S, Bowman P, Reich C, Mason B, Goddard M. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95(7):4114–29.
66. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet. 2013;45(4):400–5.
67. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013;14(7):507–15.
68. Golan D, Rosset S. Effective Genetic-Risk Prediction Using Mixed Models. Am J Hum Genet. 2014;95(4):383–93.
69. Zeng P, Wei Y, Zhao Y, Liu J, Liu L, Zhang R, Gou J, Huang S, Chen F. Variable selection approach for zero-inflated count data via adaptive lasso. J Appl Stat. 2014;41(4):879–94.
70. Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat Soc Ser B. 1996;58(1):267–88.
71. Zou H, Hastie T. Regularization and variable selection via the Elastic Net. J R Stat Soc Ser B. 2005;67(2):301–20.
72. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical learning: Data Mining, Inference, and Prediction, 2 nd edn. New York: Springer; 2009.
73. Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. New York: CRC Press; 2015.
74. Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise coordinate optimization. Ann Appl Stat. 2007;1(2):302–32.
75. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1–22.
76. Ott J. Polygenic Models for Risk Prediction in Human Genetics. Hum Hered. 2016;80(4):162–4.
77. Dandine-Roulland C, Perdry H. The Use of the Linear Mixed Model in Human Genetics. Hum Hered. 2016;80(4):196–206.
78. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44(7):821–4.
79. Beck S, Geraghty D, Inoko H, Rowen L. Complete sequence and gene map of a human major histocompatibility complex. Nature. 1999; 401(6756):921–3.
80. Lappalainen T, Sammeth M, Friedländer MR, AC't Hoen P, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501(7468):506–11.
81. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.
82. Stegle O, Parts L, Durbin R, Winn J. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. PLoS Comput Biol. 2010;6(5):e1000770.
83. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012;7(3):500–7.
84. AC't Hoen P, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, Laros JF, Buermans HP, Karlberg O, Brännvall M. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol. 2013;31(11):1015–22.
85. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760–74.
86. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 2008;24(8):408–15.
87. Bryois J, Buil A, Evans DM, Kemp JP, Montgomery SB, Conrad DF, Ho KM, Ring S, Hurles M, Deloukas P, et al. Cis and Trans Effects of Human Genomic Variants on Gene Expression. PLoS Genet. 2014;10(7):e1004461.
88. Tucker G, Loh P-R, MacLeod IM, Hayes BJ, Goddard ME, Berger B, Price AL. Two-variance-component model improves genetic prediction in family datasets. Am J Hum Genet. 2015;97(5):677–90.
89. Frazer K, Ballinger D, Cox D, Hinds D, Stuve L. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851–61.
90. Wheeler HE, Shah KP, Brenner J, Garcia T, Aquino-Michaels K, Cox NJ, Nicolae DL, Im HK, Consortium G. Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. PLoS Genet. 2016;12(11):e1006423.
91. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, Zang C, Ripke S, Bulik-Sullivan B, Stahl E. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. 2014; 95(5):535–52.

92. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, Anttila V, Xu H, Zang C, Farh K. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47(11):1228–35.
93. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42(D1):D1001–6.
94. Abraham G, Inouye M. Genomic risk prediction of complex human disease and its clinical application. Curr Opin Genet Dev. 2015;33:10–6.
95. Weissbrod O, Geiger D, Rosset S. Multikernel linear mixed models for complex phenotype prediction. Genome Res. 2016;26(7):969–79.
96. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. PLoS Genet. 2015;11(4):e1004969.
97. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 2014;24(9):1550–7.
98. Maier R, Moser G, Chen G-B, Ripke S, Coryell W, Potash JB, Scheftner WA, Shi J, Weissman MM, Hultman CM. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am J Hum Genet. 2015;96(2):283–94.
99. Li C, Yang C, Gelernter J, Zhao H. Improving genetic risk prediction by leveraging pleiotropy. Hum Genet. 2014;133(5):639–50.
100. Guo G, Zhao F, Wang Y, Zhang Y, Du L, Su G. Comparison of single-trait and multiple-trait genomic prediction models. BMC Genet. 2014;15(1):30.