BIOINFORMATICS DISCOVERY NOTE

Vol. 27 no. 6 2011, pages 745–748
doi:10.1093/bioinformatics/btr024

Data and text mining

Advance Access publication January 31, 2011

Quantitative trait prediction based on genetic marker-array data, a simulation study

Wai-ki Yip[1,*] and Christoph Lange[1,2,3]

[1]Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA, [2]Institute for Genomic Mathematics, University of Bonn and [3]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Using simulation studies for quantitative trait loci (QTL), we evaluate the prediction quality of regression models that include as covariates single-nucleotide polymorphism (SNP) genetic markers which did not achieve genome-wide significance in the original genome-wide association study, but were among the SNPs with the smallest P-value for the selected association test. We compare the results of such regression models to the standard approach which is to include only SNPs that achieve genome-wide significance. Using mean square prediction error as the model metric, our simulation results suggest that by using the coefficient of determination (R^2) value as a guideline to increase or reduce the number of SNPs included in the regression model, we can achieve better prediction quality than the standard approach. However, important parameters such as trait heritability, the approximate number of QTLs, etc. have to be determined from previous studies or have to be estimated accurately.

Contact: wkyip@hsph.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 6, 2010; revised on January 4, 2011; accepted on January 9, 2011

1 INTRODUCTION

Height is one of the classic human traits studied by many statisticians. It is easy to measure and data are readily available. Galton and Fisher studied the problem extensively. Using a model of polygenes—a large number of independently segregating genes, each contributes a small, equal and direct effect on the observed continuous phenotype, Fisher illustrated that Mendelian approach and the Biometric approach to population genetics can be bridged using height as an example (Fisher, 1918). There are many subsequent studies on the genetic variations of height in different population. Height is found to be highly heritable. Its heritability is ∼0.8 for European populations and ∼0.65 for Asian populations (Visschler, 2008; Slventoinen *et al*., 2003). With the advent of genome-wide SNP-chip technology, it is now feasible to search for genetic associations with height at a genome-wide level. Several genome-wide association studies (GWASs) for height have been conducted by Lettre *et al*. (2008), Sanna *et al*. (2008), Weedon *et al*. (2007) and Weedon *et al*. (2008) and have been combined using meta-analysis approaches. Based on a total sample size of ∼63 000, 54 loci are validated as associated with height (Visschler, 2008). However, each locus explains only a very small proportion of the phenotypic variance (∼0.3 to ∼0.5%) adding to a total of ∼20% of explained phenotypic variance of height. A recent paper by Yang *et al*. (2010) illustrates that if all the SNPs of a GWAS panel are used in a regression model, we can account for up to 45% of the heritability of the phenotype height. The weak linkage disequilibrium (LD) among SNPs could be the remaining heritability up to 85% (Yang *et al*., 2010).

While this article is not specific to height, it illustrates the difficulties in identifying potential causal genes [quantitative trait loci (QTL), disease susceptible loci (DSL) if it is associated with diseases] associated with height even when there is an abundance of data. With the type I error rate of 5×10^{-7} set for a GWAS, one would need a sample size ($\sim 10000-20000$) to detect a variant with the given effect size (0.3–0.5% of variance explained). It is worthwhile to mention two more issues. While height is a commonly available phenotype, a sample size of 40 000 is not realistic for most clinical phenotypes. Secondly, given the sample size of about 63 000 in the meta-analysis for height, heterogeneity induced by combining different studies in GWAS meta-analyses is likely to reduce the ability to further discover QTLs.

While the identification of QTLs is important as it can direct scientists to understand the underlying biological mechanism for the disease, it is not the only goal. Predicting the phenotypic value based on genetic information, i.e. SNPs, can be an important epidemiological tool to assess the predisposition of a population to certain diseases and conditions.

Traditionally, predictive model in genetics have only included genes with established association. But, in case of the quantitative traits, the majority of the genetic association remains undiscovered. By using just the identified loci in a model, the prediction will not be optimal. Our goal here is not to find the missing heritability or to identify causal genes, but to find better ways to predict the phenotypic value. In experiments with agriculture and animal studies, a similar technique known as genome-wide selection, which calculates the genomic estimated breeding value based on information from all markers is used regularly for breeding as described in Schaeffer (2006).

2 METHODS

The purpose of this study is to evaluate the ability to predict quantitative traits in humans based on GWAS data. This is in contrast to standard application of GWAS which so far have mostly been used to discover QTLs/DSLs.

*To whom correspondence should be addressed.

Table 1. Average number of significant SNPs identified by association tests after Bonferroni correction as a function of heritability and number of QTLs

|QTL–h^2	0.1	0.3	0.5	0.7	0.9
200	4.59 (105)	60.18 (195)	110.44 (195)	139.68 (198)	155.70 (198)
500	0.33 (90)	7.24 (195)	19.25 (225)	31.76 (255)	42.84 (270)
1000	0.10 (90)	1.21 (165)	3.04 (186)	4.99 (210)	7.27 (255)

The number of QTLs used for prediction in the 300 predictor set are shown in parentheses.

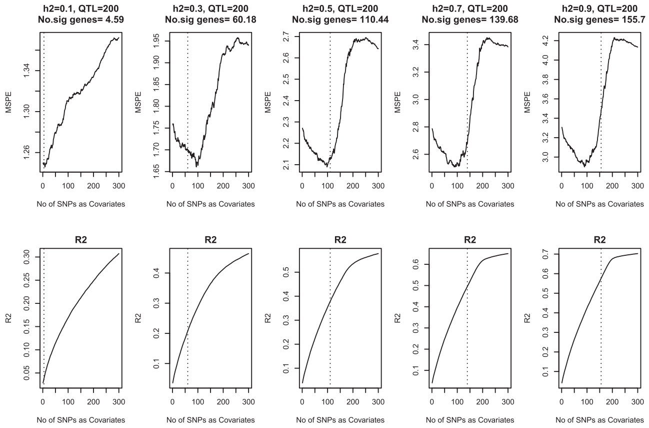

Fig. 1. Relation between MSPE and number of SNPs with 200 QTLs for the additive uniform model. The bottom panel shows where the R^2 from the training population is.

By examining different regression models, we assess the prediction quality of GWAS data and provide some guidelines for efficient model building in scenarios where the emphasis is on the prediction of the outcome variable.

We simulate a quantitative trait with different total heritability (e.g. 0.1, 0.3, 0.5, 0.7 and 0.9) from a theoretical GWAS sample of 10 000 subjects. A total of 100 000 markers are examined. To simplify the simulation, we assume that the markers are biallelic SNPs and the causal QTLs are included in these marker set. The minor allele frequency (MAF) of each marker is generated, using a β distribution. Assuming Hardy–Weinberg equilibrium, the genotypes of the markers in each study subject are then generated based on the generated MAFs. The subject's phenotype value is simulated based on the genotypes of QTLs. The details of the model are presented in the Section 2. The sample dataset is used for training/model building—creating a prediction model based on linear regression of phenotypic value on a selected set of markers. The corresponding z-score/P-value of each marker is calculated.

We then create a validation sample of 1000 subjects. The phenotypic values and the marker values are regenerated for this validation population. The prediction is then applied to the validation sample. The mean square prediction error (MSPE) for the result from each prediction model is computed for comparison of the different approaches. This procedure, i.e. creation of a training set, model building and validation step is then repeated 100 times. The reported MSPE is the average MSPE over all the runs.

3 RESULTS

We examine scenarios with different number of QTLs (200, 500 and 1000) and different values of the overall heritability ($h^2 = 0.1, 0.3, 0.5, 0.7, 0.9$). In terms of the genetic effect sizes of each QTL, we evaluate two models: an additive uniform genetic model where each QTL's contribution is equal (α, α, α,...) and an additive geometric where each QTL's contribution follows a geometric sequence (α, α^2, α^3,...) to the effect of the trait (Lande and Thompson, 1990)

To assess the performance of the standard approach, i.e. the inclusion of SNPs in the regression model that achieve genome-wide significance, we assess the statistical power to discover the QTLs in the training dataset. For the additive uniform genetic model, our simulation results are summarized in Table 1. As the number of QTLs increases, the signal gets weaker and so the number of significant SNPs detected by GWAS association tests is smaller. On the other hand, as heritability goes up, the signal gets stronger. More association tests for the QTLs become significant. The result is summarized in the following table.

The simulation results for the prediction quality of the different regression models are presented in Figure 1. The plots are organized into two panels. The top panel shows the plots of the MSPE against the number of markers used in the prediction of the phenotypic value; and the bottom panel shows the plots of R^2 values, the classical statistical measurement of variability explained by the model, based on the training sample. The vertical dotted line in every plot indicates the number of SNPs that achieve genome-wide significance based on Bonferroni correction in the regression model. Consequently, the line visualizes the performance/properties of the standard approach.

The optimal MSPE is hard to predict. As the number of predictive SNPs in the regression model increases, the number of non-QTL also increases and eventually, the quality of the prediction is affected. The association tests find most of the QTLs if the heritability is high. For traits with lower heritability, more non-QTLs will be included by chance. However, as expected, the value for R^2 continues to increase as the number of predictive SNPs increases and seems to be a good indicator of the number of QTLs found. Note that the R^2 value we used is computed directly from the training set. Theoretically, it is always bounded between 0 and 1. It is different from the R^2 value from the validation dataset.

If R^2 is small (<0.2), the signal is too weak to detect the QTLs. Too many non-QTLs are included in the predictive set; hence, the prediction outcome will be poor regardless. However, if R^2 is sufficiently large (>0.5), it indicates that most of the QTLs have already been included. The prediction outcome will be reasonable if we just use or reduce the number of significant SNPs for the prediction model. For values of R^2 between 0.2 and 0.5, we can safely increase the number of SNPs in the prediction model if we want to improve the quality of prediction. If a reference/validation dataset is available, the actual number of SNPs to be included in the prediction model can be set to where the minimum MSPE is.

For the additive geometric genetic model, the association signal is concentrated on a very small number of QTLs. The effects from subsequent QTLs beyond the first few are too weak to be detected in the training set and also too small to contribute substantially to the prediction model. The results are summarized in the plots of Figure 2. Since this scenario for the QTL distribution is not supported by current GWASs, it is not further discussed in this manuscript.

Besides uniform and geometric series, a recent paper by Park *et al.* (2010) proposes a method to estimate the effect size distribution based on GWAS. We have not assessed the different prediction qualities based on those distribution estimates in this report.

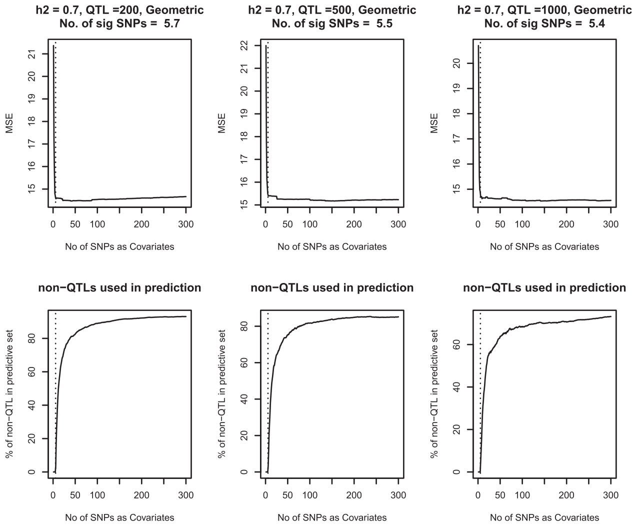

Fig. 2. Relation between MSPE and number of SNPs for the additive geometric model for $h^2 = 0.7$ showing the signal concentrated on the first few QTLs.

4 IMPLEMENTATION OF THE COMPUTER SIMULATION MODEL

The following steps describe how the simulation model works:

Step 1: Generate the MAF for each allele.

The distribution of alleles for SNP data on commercially available chips follows a β distribution. We pick $\alpha = 2.0$ and $\beta = 8.0$ for this simulation study. The SNP set is fixed at 100 000 representing a common number of SNPs used in GWASs. We use a cutoff of 0.05 to simulate quality control process where SNPs with MAF < 0.05 are usually discarded. The MAF is used in both the training and the validation phase but is regenerated for each of the 100 training–validation cycles.

Step 2: Generate the genotype of the training sample.

The training sample size consists of 10 000 simulating 10 000 individuals. Based on the MAF generated in Step 1, we generate the genotype (aa, aA or AA) for each SNP for these 10 000 individuals. No ascertainment condition is assumed.

Step 3: Generate the phenotype based on input to the model.

There are a number of parameters being input to the model: heritability and the number of QTLs. The QTLs are randomly picked from the SNP set of 100 000 and the theoretical phenotypic value is generated based on the genotype of the QTLs using an additive model:

$$\text{Phenotypic value from each QTL} = \text{number of minor allele for the QTL} \times \alpha$$

where α is calculated using the standard formula. It is a function of heritability, MAF and estimated variance $= 0.01$.

For the additive uniform model, heritability, h^2, is the total heritability divided by the number of QTL specified in the model; for the additive geometric model, heritability, h^2, is divided according to the geometric sequence (α, α^2, $\alpha^3, \ldots$). The phenotypic value for the individual is the sum of all phenotypic value for all the QTLs based on the genotype of that individual.

Step 4: Perform a single marker analysis.

Based on the generated data, we perform a single marker analysis for each of the 100 000 markers for the 10 000 individuals based on the generated phenotypic values from Step 3. It is a simple univariate regression of phenotype on the single marker.

The z-score for the i-th SNP is calculated using the standard formula:

$$z_i = \frac{y_i - \bar{y}}{\text{standard error}}$$

which has a $N(0, 1)$ distribution. A large z-score corresponds to a low P-value. Since we are doing simultaneous testing of 100 000 markers, we are bound to find significance due to chance. Bonferroni correction is applied to maintain the overall type I error rate of 0.05. The adjusted two-sided P-value threshold of 0.05×10^{-5} is needed. For a marker to be significant, its z-score must be ≥ 5.025.

Step 5: Formulate the prediction model.

We identify the significant markers based on Bonferroni corrected z-score. A linear regression model is constructed based on these significant markers. We call this the standard model.

Other regression model can be used. For simplicity, a linear regression model is used at this point. Ignoring the significance, we also created linear regression models based on $1, 2, \ldots,$ up to 300 of the top significant markers as predictors. The 'standard' model, which uses only the number of significant markers in the prediction model, is represented by the dotted vertical line in each plot. We can now compare the efficacy of the standard model with the other models.

For the training dataset, we calculate R^2, the coefficient of determination, which is used to explain the variance of the model with the total variance.

Step 6: Generate validation data.

We create a validation population of 1000 individuals. The MAF from Step 1 is used to generate the genome for these 1000 individuals. The corresponding phenotypic values are generated based on the genome of the QTLs.

Step 7: Evaluate the prediction model.

We can now apply the prediction models to the validation population. MSPE is used as the metric for comparison and is calculated for all the models. The following is the formula used to calculate MSPE for each model.

$$\text{MSPE} = \frac{\sum_i (y_i - \widehat{y_i})^2}{n}$$

where y_i is the true value, $\hat{y}_i$ is the predicted value from the model and n is the total number of sample in the validation dataset.

The training and validation cycle is repeated 100 times to ensure estimates of MSPE and R^2. The average MSPE is calculated, reported and plotted as in previous section. The average number of significant markers for each case is also reported. The average percentage of non-QTL used in the prediction model is also tracked so that we can report the amount of noise in prediction.

5 DISCUSSION

As we have demonstrated by the simple linear regression technique, it is very difficult to determine the optimal number of markers that should be used for prediction. The minimum MSPE (and probably other metrics as well) depends on other model parameters. The standard model, which uses only the significant SNPs, was compared with regression models that include either fewer number or larger number of SNPs as covariates. The number of SNPs for which the optimal MSPEs is obtained can be greater, equal to or less than the number of significant SNPs. However, by using the R^2 value of regression as a guideline, we can include more markers or remove some markers to improve the quality of prediction. In this simple simulation model, we have shown, for the given sample size, that for an R^2 value that is <0.2, there is little we can do to improve the prediction quality because, among the non-significant SNPs, there are too many non-QTLs which will be included in the prediction set. However, for R^2 values >0.5, we may consider just using or removing some of the SNPs in the prediction set. Then, for R^2 values between 0.2 and 0.5, we can certainly improve on the quality of prediction by increasing the number of SNPs in the prediction set.

There remains the question of how to pick the actual number of SNPs to be in the prediction set to achieve the optimal prediction result. The number depends on (i) heritability of this particular trait, (ii) the actual number of QTLs, (iii) the distribution of the gene and (iv) how much noise is in the system.

Heritability can be estimated by other studies. For example, the heritability for height is well studied in many populations—0.8 for European populations and 0.65 for Asian populations. Similar studies can be done to ascertain the heritability of a particular trait. The distribution of the QTLs is assumed to be uniform in this simulation. The recent paper by Park *et al.* (2010) proposes a method to estimate the effect size distribution based on GWAS. In addition, we have used causal SNPs as markers. In reality, it is highly unlikely. Most causal SNPs will just be in LD with the marker SNPs. However, that should not affect the validity of this study as weak LD can be translated into high heritability in our stimulation.

How much noise is in the system? For noise, we mean all other non-genetic factors, e.g. environment. We can estimate the amount of noise for a well-known studied trait such as height. We should be able to account for the noise in the simulation model.

By calibrating the simulation model to these parameters, we should be able to estimate through simulation the optimal number of prediction markers using R^2 as a guideline. That would provide a better prediction model than the regular significant marker approach. That is the next step for this research to quantify the parameters (heritability, number of genes and noise) that allow us to pick the optimal number of genes to use for prediction by calibrating against an existing dataset. Besides linear regression, different subset selection and regression techniques such as LASSO (Tibshirani, 1996) can be used. We can compare these techniques to see which one produces the best outcome as our next step.

In our simulation study, we assume that the heritability is due to genetic factors. If some proportion of the heritability estimate should be attributable to shared environment, the h^2 value in the simulation has to be reduced. This means that the simulation results would not change, but one would have to be able to identify the shared environmental component.

6 CONTACT SUPPLEMENTARY INFORMATION

The supplementary information contains data generated by the computer model for heritability (0.1, 0.3, 0.5, 0.7 and 0.9) for QTLs (200, 500 and 1000). The naming convention is r200R2-1 for simulation results with 200 QTLs and heritability of 0.1. The corresponding *R* script to generate the plots for each heritability is also provided. Figure 1 in this article is produced by the *R* script plotR2.R.

The computer simulation model is developed and writing in C/C++ on a Linux64 system (fedora11) and is only available on that platform. For source code, executable and any additional information, please contact the author, Wai-ki Yip, at wkyip@hsph.harvard.edu.

ACKNOWLEDGEMENTS

W.-K.Y. would like to acknowledge the generous support and encouragement from the Biostatistics Department, Harvard School of Public Health.

Funding: The project described was supported by Award Number (R01MH081862, R01MH087590) from the National Institute of Mental Health and Award Number (HL089856-4) from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Heart, Lung, and Blood Institute.

Conflict of Interest: none declared.

REFERENCES

Christoph,L. and John,C.W. (2001) On prediction of genetic values in marker-assisted selection. *Genetics*, **159**, 1375–1381.

Fisher,R.A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.*, **52**, 399–433.

Lande,R. and Thompson,R. (1990) Efficiency of marker-assisted Selection in the improvement of quantitative traits. *Genetics*, **124**, 743–756.

Lettre,G. *et al.* (2008) Identification of ten loci associated with height highlights new biological pathway in human growth. *Nat. Genet.*, **40**, 584–591.

Park,J.-H. *et al.* (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.*, **42**, 570–575.

Sanna,S. *et al.* (2008) Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat. Genet.*, **40**, 198–203.

Schaeffer,L.R. (2006) Strategy for applying genome-wide selection in dairy cattle. *J. Anim. Breed. Genet.*, **123**, 218–223.

Silventoinen,K. *et al.* (2003) Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res.*, **6**, 399–408.

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

Visschler,P.M. (2008) Sizing up human height variation. *Nat. Genet.*, **40**, 489–490.

Weedon,M.N. *et al.* (2007) A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nat. Genet.*, **39**, 1245–1250.

Weedon,M.N. *et al.* (2008) Genome-wide analysis identifies 20 loci that influence adult height. *Nat. Genet.*, **40**, 575–583.

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