Efficiency of Variable Selection in Genome-Wide Prediction for Traits of Different Genetic Architecture

***C.C. Schön*[1] , V. Wimmer[1,2] and C. Lehermeier[1]**
[1]Technische Universität München, [2]current address: KWS SAAT AG

ABSTRACT: The choice of statistical method to obtain maximum prediction accuracy in genome-based prediction is still under debate. For traits influenced by a small number of quantitative trait loci, predictions should benefit from methods performing variable selection compared to methods distributing effects across the genome. However, assumptions underlying successful variable selection are frequently violated in experimental data. Based on computer simulations and experimental data sets from different species we investigated the breakdown behavior of different statistical methods with respect to recovering true non-zero predictors in the underlying genetic model. The efficiency of variable selection was strongly influenced by the level of determinedness of the data, the heritability of the trait, and the extent of linkage disequilibrium in the population. Based on our results, upper bounds for the number of causal mutations which can be identified by a variable selection method can be inferred.

Keywords: variable selection; genomic prediction

Motivation

A model predicting complex phenotypes is assumed to perform well, if prior assumptions about the factors contributing to phenotypic trait expression match the pattern in the data. However, the true genetic architecture of quantitative traits is unknown, making model selection a complex task in practice. Therefore, key questions in genomic prediction are whether or not selection of a subset of markers that tag quantitative trait loci (QTL) can enhance prediction performance and which methods should be chosen to perform this task.

In simulation studies, it was shown that statistical methods employing variable selection such as BayesB (Meuwissen et al. 2001) or LASSO (Tibshirani et al. 1996) were superior over ridge regression best linear unbiased prediction (RR-BLUP; Hoerl and Kennard, 1970) or GBLUP (Habier et al. 2007) even for traits of considerable complexity (Zhong et al. 2009; Daetwyler et al. 2010). However, most experimental studies conducted on livestock and plant populations revealed only small differences between methods employing variable selection and those distributing effects across the genome (e.g. Heslot et al. 2012). Only few examples exist where variable selection consistently improved prediction accuracy in real life data. Thus, the potential of variable selection in high-dimensional data sets still warrants further investigation.

Here, we report results from several studies. Wimmer et al. (2013) investigated the performance of different prediction methods in simulated data and inferred the breakdown behavior of these methods with respect to recovering true non-zero predictors in the underlying genetic model. They also investigated the prediction performance of variable selection methods for traits of different genetic architecture and heritability in experimental data sets from different species. With experimental studies in maize we addressed the question if variable selection methods can outperform GBLUP in data sets with large allelic diversity as hypothesized by Lorenz et al. (2011). Furthermore, we investigated the predictive power of SNP markers that had been selected based on genome annotation and on their assignment to one of the two subgenomes of maize (Schnable et al. 2011).

Computer Simulations

A simulation study investigating the assumptions underlying successful variable selection was presented by Wimmer et al. (2013). Replicated data sets were generated for each of 400 scenarios. In each scenario, p=2000 independent biallelic single nucleotide polymorphism (SNP) markers segregating for n individuals were simulated. Varying the number of phenotypic observations from 100 to 2000 resulted in 20 different levels of determinedness (n/p). The level of model complexity (number of true non-zero coefficients p_0/n) was varied from 0.05 to 1.0 in increments of 0.05. For all scenarios linkage equilibrium was assumed and four different trait heritabilities (0.25, 0.50, 0.75, 1.0) were simulated. Performance of two methods employing variable selection (LASSO and BayesB) was compared to the performance of RR-BLUP by assessing the normalized L_2 error

$$L_2\left(\hat{\boldsymbol{\beta}},\boldsymbol{\beta_0}\right)=\frac{\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta_0}\right\|_2}{\left\|\boldsymbol{\beta_0}\right\|_2},$$

with $\boldsymbol{\beta_0}$ denoting the true, $\hat{\boldsymbol{\beta}}$ the estimated marker effects, and $\|\cdot\|_2$ the L_2 norm of a vector.

The magnitude of the L_2 error for RR-BLUP was mainly a function of the level of determinedness, while performance of LASSO and BayesB was also strongly affected by model complexity. For smaller population sizes, i.e. lower levels of determinedness (n/p < 0.5, h^2=0.75), the variable selection methods performed better than RR-BLUP if model complexity was also low (p_0/n < 0.4, h^2=0.75). Thus, if the true model is sparse, variable selection methods can estimate marker effects with higher precision than RR-BLUP. However, as the genetic trait architecture becomes more complex, variable selection methods will not be successful in identifying the true model and precision of marker estimates will be low.

When increasing the number of predictors in the model for a given sample size, the level of determinedness will decrease leading to an increase in L_2 error. Consequently, increasing marker density with high-density SNP arrays or whole-genome sequencing will not improve prediction accuracy unless effective sample sizes increase accordingly.

Trait heritability also had a strong effect on the breakdown behavior of LASSO and BayesB. For the 0.5 level of determinedness (n=1000, p=2000) and h^2=1.0, variable selection methods performed better than RR-BLUP up to a complexity level of approximately 600 true non-zero effects. However, with h^2<0.5 the ability to identify the true non-zero predictors in the model disappeared already for much lower complexity levels and variable selection methods could not improve prediction over RR-BLUP when the number of true non-zero effects exceeded 300. As in simulation studies a number of simplifying assumptions have to be made (e.g. absence of dominance and epistasis, linkage equilibrium between markers) these numbers can be considered to be rather optimistic. Thus, we conclude that when trait heritability is low variable selection methods cannot be expected to recover the true model and show superior performance to RR-BLUP unless sample size is much higher than the number of segregating QTL.

Experimental Studies

We investigated prediction performance of the statistical methods LASSO, BayesB, and RR-BLUP for 13 traits of different genetic architecture in four experimental data sets from wheat, rice, Arabidopsis, and maize. One aim of the study was to identify experimental settings in which variable selection methods were consistently superior to RR-BLUP. Results are in part presented in Wimmer et al. (2013).

The wheat (n=254, number of SNP markers m=2,056) and the maize data sets (n=698, m=11,646) represent typical breeding populations with familial substructure (Poland et al. 2012; Lehermeier et al. 2013). The rice data set (n=413, m=36,901) comprises a global collection of highly diverse rice lines derived from six distinct subpopulations of different geographic origin generating high long-range linkage disequilibrium (LD) when calculated for the entire population (Zhao et al. 2011). The Arabidopsis data set consists of 199 accessions genotyped with 215,908 SNP markers (Atwell et al. 2010) with no obvious population structure and substantially less LD than the other data sets. Phenotypic traits were selected based on results from genome-wide association studies indicating different genetic trait architectures. Of the 13 traits analyzed, three were assumed to be controlled by few major QTL. Performance of the three statistical methods was assessed using fivefold cross-validation with random assignment of genotypes to estimation and test sets and repeated sampling.

For all traits and all data sets, BayesB had the same prediction performance as RR-BLUP. In the Arabidopsis data set, LASSO was superior to RR-BLUP and BayesB for two traits, with the largest difference exhibited for the trait that was assumed to have a sparse genetic architecture (FRIGIDA gene expression).

In the wheat and rice data set prediction with LASSO never outperformed the other two methods and was significantly decreased for most traits. The results indicated that the decrease in predictive power of LASSO for all traits, irrespective of their genetic architecture, was the result of high long-range LD prevalent in the rice and wheat data sets. In the presence of LD, a loss in prediction performance can be expected for LASSO in comparison to RR-BLUP as LASSO randomly selects one predictor variable from a group of correlated variables while RR-BLUP distributes effects across several SNPs.

These findings were corroborated by simulation studies. Imposing the correlation structure of the three experimental data sets on the simulated data described in the previous section showed an increase in normalized L_2 error in the presence of LD as compared to simulations with independent SNP markers. Simulations showed that for LASSO the number of phenotypic observations required to

achieve the same average normalized L_2 error as in scenarios with independent markers was at least doubled when the LD structure of the Arabidopsis data set was superimposed.

In the literature it has been hypothesized that Bayesian and variable selection models rely more on information from LD whereas GBLUP mainly uses information from relatedness (Habier *et al.* 2007). Consequently, we studied if variable selection methods could improve prediction performance when employed in experimental populations with distinct familial substructure and extensive allelic diversity. We investigated this question in a maize data set (n=841, m=32,801) comprising ten biparental families representing a large spectrum of the allelic diversity of European dent maize germplasm (Bauer et al. 2013; Lehermeier et al. in review). Five traits were recorded: biomass yield, dry matter content, male and female flowering, and plant height. From QTL analyses prior knowledge was available that large effect QTL were segregating for all traits in this population. We chose BayesCπ (Habier et al. 2011) as variable selection method and compared its prediction performance to GBLUP using cross-validation. Prediction performance across families was assessed by predicting phenotypic values of all individuals from one biparental family based on a model trained on genotypic and phenotypic data from the other nine families.

Prediction performance achieved with BayesCπ was not improved compared to GBLUP despite the diverse material under study and QTL with sizeable effects segregating. Within and across families predictive abilities differed only marginally between the two methods giving no indication that the variable selection method could increase prediction by capturing LD between SNP markers and QTL.

Towards whole-genome sequences

The promise of whole-genome sequencing data is that causal variants will be included in the data with high probability (Meuwissen and Goddard 2010). On the other hand, the number of predictor variables is vast relative to the number of individuals for which sequencing and phenotyping data will be available. Even if sequencing technologies will allow the analysis of thousands of individuals in the near future, for many plant and livestock species precision phenotyping will remain a severe bottleneck leading to highly underdetermined models.

As we have seen that the performance of prediction methods is mainly dominated by dimensionality (n, p, p_0), pre-screening of sequencing data to extract meaningful predictors might be a viable strategy. It has been shown that integrating knowledge on marker-trait associations from functional or QTL detection studies into whole genome-based prediction can lead to an increase in prediction performance (de los Campos et al. 2013; Zhang et al. 2014). In a maize genome-wide association study Li et al. (2012) demonstrated that trait-marker associations were enriched in specific genomic regions. We therefore studied if prediction accuracies could be improved by classifying SNPs according to bioinformatic information. In the Arabidopsis and the maize data set described by Lehermeier et al. (2013) we used Software SnpEff (Cingolani et al. 2012) in conjunction with the respective gene models to classify SNPs into different bioinformatic categories (e.g. genic, non-genic, synonymous, non-synonymous). For each class of SNPs, prediction performance was assessed using cross-validation and compared to the prediction performance that could be obtained with the same number of randomly chosen SNPs.

For all traits in both data sets, predictive abilities obtained with SNPs in a specific genomic region were in the range of predictive abilities expected when the same number of SNPs was selected at random from the entire genome. Slightly different results were obtained by Morota et al. (2014) for prediction of three complex traits in chicken. They found small but significant differences in prediction accuracy for some genomic regions although these results were not consistent across traits and none of the predictions based on SNPs in a specific genomic region substantially outperformed prediction with all markers.

When incorporating recent findings on genome evolution of maize in our prediction models, we did find substantial differences in predictive abilities between genomic regions. It is conjectured that the ten maize chromosomes trace back to a tetraploid ancestor and Schnable et al. (2011) separated the genome of modern maize into two subgenomes representing the two duplicated genomes present in its tetraploid ancestor. Based on the maize data set described by Bauer et al. (2013) we assessed predictive abilities for five traits separately for the two subgenomes using GBLUP. Approximately two thirds of segregating SNPs were assigned to subgenome 1, one third to subgenome 2. Differences in size between the two subgenomes were accounted for by randomly sampling fragments from subgenome 1 to achieve the same genome coverage as subgenome 2. For all traits, a higher predictive ability was achieved for genomic regions assigned to subgenome 1 as compared to subgenome 2 indicating that in the maize genome specific regions exhibit higher predictive ability than

others. Further research will be needed to investigate the underlying mechanisms of these findings.

Conclusion

Our findings from computer simulations and experimental data clearly show that variable selection methods can outperform methods retaining all predictors in the model. Crucial assumptions for variable selection to perform successfully are that in model training the effective sample size needs to scale with the number of causal mutations affecting the trait of interest and the total number of predictors in the model. High trait heritability and low LD increase the ability of the statistical methods to successfully perform variable selection and accurately estimate SNP effects. More profound knowledge on the role of specific genomic regions contributing to trait expression might enhance predictive power.

Literature Cited

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J. et al. (2010). Nature 465: 627–631.

Bauer, E., Falque, M., Walter, H. et al. (2013). Genome Biol. 14: R103.

Cingolani, P., Platts, A., Wang, L. L. et al. (2012). Fly 6:80-92.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B. et al. (2010). Genetics 185:1021–1031.

de los Campos, G., Vazquez, A. I., Fernando, R. et al. (2013). PLoS Genet 9: e1003608.

Habier, D., Fernando, R. L., and Dekkers, J. C. (2007). Genetics 177:2389–2397.

Habier, D., Fernando, R. L., Kizilkaya, K. et al. (2011). BMC Bioinformatics 12: 186.

Heslot, N., Yang, H., Sorrells, M. E. et al. (2012) Crop Sci. 52:146–160.

Hoerl, A. E., and Kennard, R. W. (1970) Technometrics 12:55–67.

Lehermeier, C., Wimmer, V., Albrecht, T. et al. (2013). Stat. Appl. Genet. Mol. Biol. 12: 375–391.

Li, X., Zhu, C., Yeh, C.T. et al. (2012). Genome Res. 22:2436-2444.

Lorenz, A. J., Chao, S., Asoro, F. G. et al. (2011). Advances in Agronomy 110:77–123.

Meuwissen, T., and Goddard, M. (2010). Genetics 185:623–631.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Genetics 157:1819 –1829.

Morota, G., Abdollahi-Arpanahi, R., Kranis, A. et al. (2014). BMC Genomics 15:109.

Poland, J., Endelman, J., Dawson, J. et al. (2012). Plant Genome 5:103.

Schnable, J. C., Springer, N. M., and Freeling, M. (2011). Proc. Natl. Acad. Sci. 108:4069–4074.

Tibshirani, R. (1996). J. Roy. Stat. Soc. B 58:267–288.

Wimmer, V., Lehermeier, C., Albrecht, T. et al. (2013). Genetics 195: 573–587.

Zhang, Z., Ober, U., Erbe, M. et al. (2014). PLoS ONE 9(3): e93017.

Zhao, K., Tung, C.W., Eizenga, G. C. et al. (2011). Nat. Commun. 2:467.

Zhong, S., Dekkers, J. C., Fernando R. L. et al. (2009). Genetics 182:355–364.