# Effect of shrinkage on prediction accuracy of methionine and proline fruit contentsin a broad-based tomato population

Janejira Duangjit, Mathilde M. Causse, Christopher Sauvage

# Effect of shrinkage on prediction accuracy of methionine and proline fruit contentsin a broad-based tomato population

**Janejira Duangjit[1]*, Mathilde Causse[2], Christopher Sauvage[2]**

*[1]Departments of Horticulture, Faculty of Agriculture, Kasetsart University, Bangkok 10900, Thailand*
*[2]INRA, UR1052 GAFL, Génétique et Amélioration des Fruits et Légumes, 67 allée des chênes, CS60094, 84143 Montfavet cedex, France*

**Corresponding author: fagrjrd@ku.ac.th*

## ABSTRACT

Genomic selection is a promising marker assisted selection based application for quantitative traits improvement. In this study, the genomic estimated breeding values (GEBVs) were estimated by the ridge-regression best linear unbiased prediction (rrBLUP) statistical model with a training set of 122 accessions (122/163 - 75%) from an available dataset representing a broad-based tomato population. Methionine and proline contents were randomly picked as representative of low heritability ($h^2$<0.5) and high heritability ($h^2$>0.7) metabolomics traits, respectively. The goals are to minimize mean-squared error (MSE) and to see the potential of using low marker density (250≤m≤1500). Results showed that shrinkage intensity was affected by the number of markers used in the model; and it ranged from 1% to 7% (the lower number of markers, the higher shrinkage intensity). In this tomato population, accuracies predicted by shrinked approach revealed benefit of shrinkage over non-shrinked approach, in methionine but not proline fruit contents, when used small number of marker (m=250). This suggests that shrinkage should be applied to low heritability traits to improve prediction accuracy in a board-based tomato population.

**Keywords**: genomic selection; predicted breeding value; shrinkage intensity

## INTRODUCTION

Genomic selection (GS) is an approach utilizing information from molecular markers covering the whole genome of organisms. The aim of this technique is to shorten the selection by using genotypic information to predict the performance of individuals in the form of genomic estimated breeding value (GEBV), and finally to select for the elite individuals to be used in the breeding program (Borém and Fritsche-Neto, 2014).

In the past decade, GS was conducted in many species of animals [e.g. cattle (Hayes *et al*., 2009), chicken (Wang *et al*., 2013), and pig (Tribout *et al*., 2012)] and plants [e.g. maize (Owens *et al*., 2014), wheat (Storlie *et al*., 2013), apple (Kumar *et al*., 2012)]. With the availability of high-throughput sequencing technology, it is practicable for researchers and breeders to use information from dense genetic markers to benefit from genetic and breeding studies.

The effectiveness of many factors in GS have been tested (Asoro *et al*., 2011; Würschum *et al*., 2013), mostly aim to improve accuracy in prediction. Besides, attempt to use less markers in order to reduce genotyping cost was made (Wang *et al*., 2013). It has been found that prediction accuracy dropped when fewer markers were used in GS. However, increasing the number of markers was not always improving the prediction (i.e. accuracy reached a plateau when 3000 markers were fitted in the model) (Duangjit *et al.,* under review). In addition, when minimizing markers number, it is possible that the number of individuals exceed the number of markers. This causes high mean-squared error (MSE) which can be fixed by performing shrinkage (Stein, 1956). In the rrBLUP R package, shrinkage was implemented (Endelman, 2011; R Development Core Team, 2014). It has been previously shown that shrinkage estimation improved accuracy of GEBVs using rrBLUP (Endelman and Jannink, 2012; Zhao *et al.*, 2013).

In this study, we want to test the effect of shrinkage of methionine and proline fruit contents from publicly available data of tomato. Here, we assessed shrinkage intensity that minimizes the expected MSE when different ranges of low marker density used in the rrBLUP model. We also evaluated the effect of shrinkage on the prediction accuracy.

## MATERIALS AND METHODS

### Statistical Model in Genomic Selection

Genotypic and phenotypic data used in this study is available publically, from a broad-based population of tomato accessions (Sauvage *et al.*, 2014). Previous cross-validation study of genomic selection with different sizes of training set has revealed the potential of genomic selection using ridge-regression best linear unbiased prediction (rrBLUP) statistical model (Asoro *et al*., 2011). The study showed that the high number of individuals in training step resulted in high accuracy of prediction. Therefore, from the population of 163 tomato accessions, 122 (75%) accessions were used to fit the model, and 41 (25%) accessions were used as a criterion in validation step in this study. The analyses were performed in rrBLUP package in R software published by Endelman (2011) (see also: https://cran.r-project.org/web/packages/rrBLUP/index.html).

Mixed model for prediction of breeding values were implemented with the model $\mathrm{Y} = 1\mu + Xg + e$ where $\mathrm{Y}$ is vector of observed phenotypic value, $\mu$ is overall mean of the training set fitted as fixed effect, $g$ is a vector of random SNP effect, $X$ is an incidence matrix $g$, constructed from covariates based on the genotype, and $e$ is residual effect. Genomic estimated breeding value (GEBV) was predicted using genotypic information, and SNP effects obtaining from *mixed.solve* function in rrBLUP (Endelman, 2011).

### Shrinkage Estimation in rrBLUP

Marker matrix (X) was created by decoding bi-allelic loci to be X ∈ {-1, 0 ,1}. Shrinkage estimation was performed with *A.mat* function in the rrBLUP package. Estimations were repeated with different number of markers, which were randomly selected from a set of 5995 SNP markers (Sim *et al.*, 2012). Shrinkage intensities when 250 markers were randomly picked and fitted into the rrBLUP model were estimated from the software. The averages of intensities which lead to minimal mean square errors were calculated from 1000 iterations. Calculations were made for different sets of markers (i.e. 500, 1000, and 1500).

### Breeding Value Prediction

Population accuracy was predicted using 75% (122 accessions) of tomato population. The shrinkage intensity used in this analysis was ranged based on results from the first part (i.e. at 1.0, 2.0, 3.6, and 7.0%), and also at zero as a non-shrinked estimation. In order to assess effects of shrinkage in genomic selection, accuracies from shrinked and non-shrinked estimations were compared.

Finally, prediction accuracies were evaluated by correlating values between predicted breeding values (GEBVs) and measured phenotypic data of methionine and proline fruit contents as reported in Sauvage *et al*. (2014). Average values and standard deviations were calculated from 1000 iterations. Statistical significant difference between i) shrinked and non-shrinked methods and ii) shrinkage intensities were evaluated using t-tests.

## RESULTS AND DISCUSSION

### Estimated Optimal Shrinkage

Shrinkage intensity values under the criteria to minimize the expected MSE were obtained from this study. The optimal shrinkage intensity increased when smaller number of markers were randomly selected from a set of 5995 markers. The minimum ($0.01±6.938x10^{-18}$) was obtained from using 1500 markers, while the highest intensity was found when a set of random 250 markers was used (0.07±0.007) (Figure 1).

The results were estimated from random set of 250, 500, 1000, and 1500 out of a total set of 5995 markers. As the size of marker set is bigger, the shrinkage intensity is closer to zero. Difference in shrinkage intensity between 250 and 500 markers was significantly greater than that between 1000 and 1500 markers (P<0.05). A similar trend was also found in rice, maize, barley, and pig populations as studied by Endelman and Jannick (2012) where approximate 10% shrinkage was found, when a set of 96 markers was used. However, very low number of markers (less than 250 markers) was not tested here as it was assumed to cause poor prediction ability.

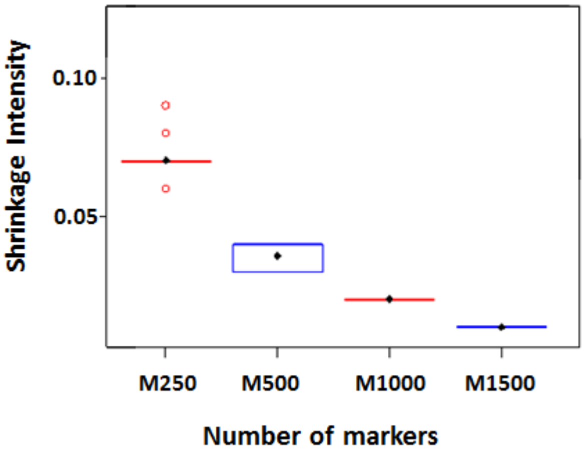

**Figure 1** Optimal shrinkage intensity predicted by rrBLUP. Boxplot of shrinkage intensity estimated by using 122 accessions in training step with different sets of markers. Midbar and dot indicate average and median from 1000 imputations. M250, M500, M1000, and M1500 indicate number of markers used in statistical model training step.

**Effects of Shrinkage on Prediction Accuracy**

The accuracies from shrinked and non-shrinked approaches were compared in the two traits. When shrinkage was performed, in methionine fruit content, accuracy was 0.136±0.150 when 250 markers were included in the model. Accuracy increased to 0.143±0.143 and 0.147±0.148 when 500 and 1000 markers were used, respectively. Although, the accuracy slightly dropped when used 1500 markers, they were not significantly different from accuracy calculated from 1000 markers (Table 1). Similar trends were observed in proline fruit content; the larger the number of markers, the higher the accuracy of predicting breeding value. Accuracy increased from 0.337±0.104 (used 250 markers) to 0.369±0.096 (used 1500 markers). This can be explained by the fact that markers with higher effect were fitted into the model.

When compared to the accuracies obtained from non-shrinked approach, for both traits, in most cases, accuracy values were higher when shrinkage was performed (Table 1 and Figure 2). When 250 markers were included in the model, accuracy was 0.129±0.148 without shrinkage. When shrinked with intensity of 7% (average shrinkage intensity when used 250 markers), accuracy increased to 0.136±0.150. The difference was significant at P<0.05. This trend was not observed when more markers were added, and similar result was not found in proline fruit content.

**Table 1** Prediction ability in methionine and proline traits obtained from genomic selection with rrBLUP model.

| Traits | | Methionine | Proline |
|---|---|---|---|
| Heritability | | 0.427 | 0.773 |
| Shrinked<br>(mean±SD) | 250 | 0.136±0.150* | 0.337±0.104 |
| | 500 | 0.143±0.143 | 0.355±0.098 |
| | 1000 | 0.147±0.148 | 0.355±0.103 |
| | 1500 | 0.144±0.145 | 0.369±0.096 |
| Non-shrinked<br>(mean±SD) | 250 | 0.129±0.148* | 0.343±0.104 |
| | 500 | 0.134±0.149 | 0.354±0.098 |
| | 1000 | 0.147±0.144 | 0.357±0.103 |
| | 1500 | 0.141±0.145 | 0.358±0.096 |

Mean and standard deviation (SD) are shown. Mean accuracies are the average correlation values of GEBVs and measured phenotypic values. Prediction was performed using 122 accessions in training step with different numbers of markers (1000 imputations). Heritability of each trait is from Sauvage *et al.* (2014). Significant difference assessed by t-test is indicated by asterisk (*).

Although with a low number of markers used (m≤1500), proline fruit content with high heritability ($h^2$=0.773) had higher accuracy than methionine fruit content ($h^2$=0.427), as shown above. This result agreed with previous study of traits with different heritability (Wimmer *et al.*, 2013) supporting that trait with high heritability is the trait underlined with big proportion of genetic component. The results showed that significant increase of accuracy of GEBV, for most cases, cannot be obtained from shrinkage in this structured tomato population, particularly for the methionine and proline fruit content, which agreed with results from Endelman and Jannick (2012) where shrinkage can significantly increase GEBV only in unstructured population. However, the significant accuracy gain was found when shrinked at 7% only in a low heritability trait (methionine). This revealed that, in trait with high heritability, shrinkage does not help improving accuracy, as reported by Endelman and Jannick (2012).

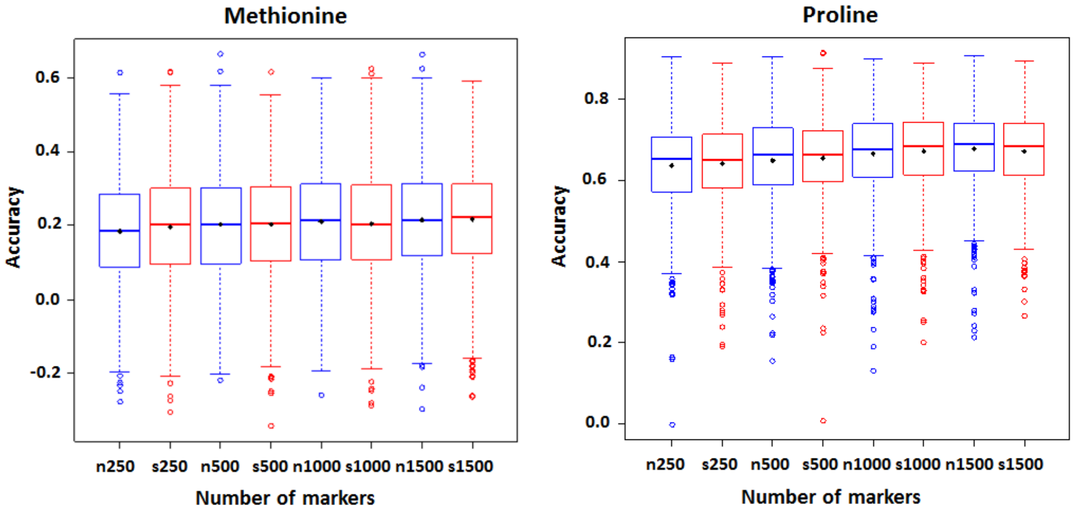

**Figure 2** Impact of shrinkage on prediction accuracy. Boxplot of prediction accuracy estimated by rrBLUP model using 122 accessions in training step with different sets of markers. Midbar and dot indicate average and median from 1000 imputations. n and s represent non-shrinked and shrinked calculation; numbers (i.e. 250, 500, 1000, and 1500) indicate number of markers used in statistical model training step.

Overall, as using methionine and proline fruit contents as case studies, factors such as numbers of markers and shrinkage affect genetic prediction using the rrBLUP method. Moreover, genetic background (i.e. heritability) of each trait also plays important role. This study showed that shrinkage benefits GS in trait with low heritability when the marker number is low. In tomato, the low-cost (due to small number of markers use) and the most efficient number of markers need to be revealed.

## ACKNOWLEDGEMENTS

This work was supported by the European Union (European Solanaceae grant no. PL016214–2). We would like to thank to Jeffery Endelman (University of Wisconsin-Madison, USA) for his advises on our work.

## REFERENCES

Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink JL (2011) Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. Plant Gen 4: 132–144.

Borém A, Fritsche-Neto R (2014) Biotechnology and Plant Breeding - Applications and Approaches for Developing Improved Cultivars . 1st ed. Scientific Publishing Services, LTD, CA.

Duangjit J, Causse M, Sauvage C, Efficiency of Genomic Selection for Tomato Fruit Quality. Submitted.

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Gen 4: 250–255.

Endelman JB, Jannink JL (2012) Shrinkage estimation of the realized relationship matrix. G3 2: 1405–1413.

Hayes B, Bowman P, Chamberlain A, Goddard M (2009) Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci 92: 433–443.

Kumar S, Chagné D, Bink MCAM, Volz RK, Whitworth C, Carlisle C (2012) Genomic selection for fruit quality traits in apple (*Malus×domestica* Borkh.) PLoS ONE 7: 1–10.

Owens B, Lipka A, Magallanes-Lundback M, Tiede T, Diepenbrock C, Kandianis C, Kim E, Cepela J, Mateos-Hernandez M, Buell C, *et al.* (2014) A foundation for provitamin a biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels. Genetics 198: 1699–1716.

R Development Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012, ISBN 3-900051-07-0.

Sauvage C, Segura V, Bauchet G, Stevens R, Phuc Thi D, Nikoloski Z, Fernie AR, Causse M (2014) Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. Plant Physiol 165: 1120–1132.

Sim SC, Durstewitz G, Plieske J, Wieseke R, Ganal MW, Van Deynze A, Hamilton JP, Buell CR, Causse M, Wijeratne S, et al (2012) Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. PLoS ONE 7: e40563

Stein C (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, Proceedings of the Third Berkeley symposium on mathematical statistics and probability. pp. 197–206.

Storlie E, Charmet G (2013) Genomic selection accuracy using historical data generated in a wheat breeding program. Plant Gen 6: 1–9.

Tribout T, Larzul C, Phocas F (2012) Efficiency of genomic selection in a purebred pig male line. J Anim Sci 90: 4164–4176.

Wang C, Habier D, Wolc A, Garrick DJ, Fernando RL, Lamont SJ, Dekkers J, Kranis A, Watson KA (2013) Application of genomic selection using an evenly spaced low-density marker panel in broiler chickens. Animal Industry Report 659: 58.

Wimmer V, Lehermeier C, Albrecht T, Auinger H, Wang Y, Schön C (2013) Genome-wide prediction of traits with different genetic architechture through efficient variable selection. Genetics 195: 573–587.

Würschum T, Reif J, Kraft T, Janssen G, Zhao Y (2013) Genomic selection in sugar beet breeding populations. BMC Genet 14: 1–8.

Zhao Y, Gowda M, Liu W, Würschum T, Maurer HP, Longin FH, Ranc N, Piepho HP, Reif J (2013) Choice of shrinkage parameter and prediction of genomic breeding values in elite maize breeding populations. Plant Breed 132: 99–106.