

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

Janejira Duangjit, Mathilde Causse, Christopher Sauvage

► **To cite this version:**

Janejira Duangjit, Mathilde Causse, Christopher Sauvage. Effect of shrinkage on prediction accuracy of methionine and proline fruit contents in a broad-based tomato population. Genomics, Bioinformatics and System Biology Conference (GBSBC 2015), Sep 2015, Bangkok, Thailand. hal-02739886

HAL Id: hal-02739886

<https://hal.inrae.fr/hal-02739886>

Submitted on 2 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Effect of shrinkage on prediction accuracy of methionine and proline fruit**
2 **contents in a broad-based tomato population**

3 **Janejira Duangjit^{1*}, Mathilde Causse², Christopher Sauvage²**

4
5 ¹*Departments of Horticulture, Faculty of Agriculture, Kasetsart University, Bangkok 10900, Thailand*

6 ²*INRA, UR1052 GAFL, Génétique et Amélioration des Fruits et Légumes, 67 allée des chênes, CS60094, 84143*
7 *Montfavet cedex, France*

8
9 **Corresponding author: fagrjrd@ku.ac.th*

10
11 **ABSTRACT**

12 Genomic selection is a promising marker assisted selection based application for quantitative traits
13 improvement. In this study, the genomic estimated breeding values (GEBVs) were estimated by the ridge-
14 regression best linear unbiased prediction (rrBLUP) statistical model with a training set of 122 accessions
15 (122/163 - 75%) from an available dataset representing a broad-based tomato population. Methionine and
16 proline contents were randomly picked as representative of low heritability ($h^2 < 0.5$) and high heritability
17 ($h^2 > 0.7$) metabolomics traits, respectively. The goals are to minimize mean-squared error (MSE) and to see
18 the potential of using low marker density ($250 \leq m \leq 1500$). Results showed that shrinkage intensity was
19 affected by the number of markers used in the model; and it ranged from 1% to 7% (the lower number of
20 markers, the higher shrinkage intensity). In this tomato population, accuracies predicted by shranked approach
21 revealed benefit of shrinkage over non-shranked approach, in methionine but not proline fruit contents, when
22 used small number of marker ($m=250$). This suggests that shrinkage should be applied to low heritability
23 traits to improve prediction accuracy in a board-based tomato population.

24
25 **Keywords:** genomic selection; predicted breeding value; shrinkage intensity

26
27 **INTRODUCTION**

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

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

1 The effectiveness of many factors in GS have been tested (Asoro *et al.*, 2011; Würschum *et al.*,
2 2013), mostly aim to improve accuracy in prediction. Besides, attempt to use less markers in order to reduce
3 genotyping cost was made (Wang *et al.*, 2013). It has been found that prediction accuracy dropped when
4 fewer markers were used in GS. However, increasing the number of markers was not always improving the
5 prediction (i.e. accuracy reached a plateau when 3000 markers were fitted in the model) (Duangjit *et al.*,
6 under review). In addition, when minimizing markers number, it is possible that the number of individuals
7 exceed the number of markers. This causes high mean-squared error (MSE) which can be fixed by
8 performing shrinkage (Stein, 1956). In the rrBLUP R package, shrinkage was implemented (Endelman, 2011;
9 R Development Core Team, 2014). It has been previously shown that shrinkage estimation improved
10 accuracy of GEBVs using rrBLUP (Endelman and Jannink, 2012; Zhao *et al.*, 2013).

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

16 MATERIALS AND METHODS

17 Statistical Model in Genomic Selection

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

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

31 Shrinkage Estimation in rrBLUP

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

1 **Breeding Value Prediction**

2 Population accuracy was predicted using 75% (122 accessions) of tomato population. The shrinkage
3 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%),
4 and also at zero as a non-shrunked estimation. In order to assess effects of shrinkage in genomic selection,
5 accuracies from shrunked and non-shrunked estimations were compared.

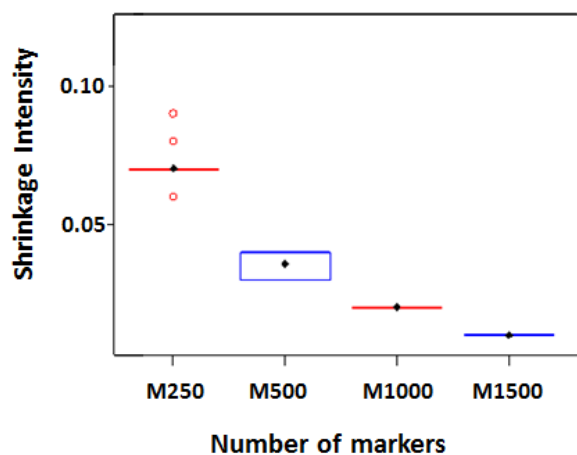
6 Finally, prediction accuracies were evaluated by correlating values between predicted breeding
7 values (GEBVs) and measured phenotypic data of methionine and proline fruit contents as reported in
8 Sauvage *et al.* (2014). Average values and standard deviations were calculated from 1000 iterations.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

11
12 **RESULTS AND DISCUSSION**

13 **Estimated Optimal Shrinkage**

14 Shrinkage intensity values under the criteria to minimize the expected MSE were obtained from this
15 study. The optimal shrinkage intensity increased when smaller number of markers were randomly selected
16 from a set of 5995 markers. The minimum ($0.01 \pm 6.938 \times 10^{-18}$) was obtained from using 1500 markers, while
17 the highest intensity was found when a set of random 250 markers was used (0.07 ± 0.007) (Figure 1).

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



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

4

5 **Effects of Shrinkage on Prediction Accuracy**

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

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

20

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

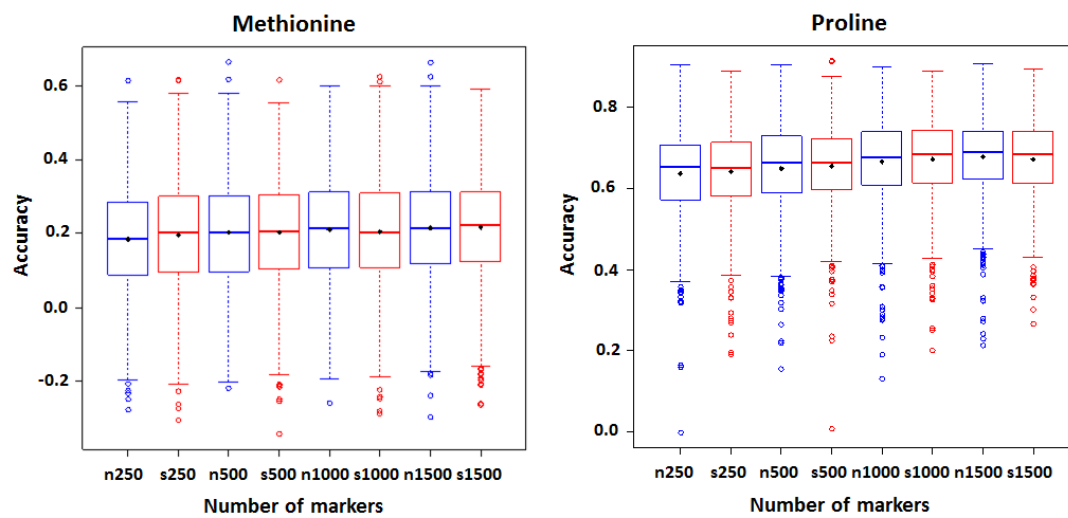
22

Traits		Methionine	Proline
Heritability		0.427	0.773
Shrinked (mean±SD)	250	$0.136\pm 0.150^*$	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 (mean±SD)	250	$0.129\pm 0.148^*$	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

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

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

11



12 **Figure 2** Impact of shrinkage on prediction accuracy. Boxplot of prediction accuracy estimated by rrBLUP
 13 model using 122 accessions in training step with different sets of markers. Midbar and dot indicate average
 14 and median from 1000 imputations. n and s represent non-shrunked and shrunked calculation; numbers (i.e.
 15 250, 500, 1000, and 1500) indicate number of markers used in statistical model training step.

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

21 **ACKNOWLEDGEMENTS**

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

24 **REFERENCES**

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

- 1 Borém A, Fritsche-Neto R (2014) *Biotechnology and Plant Breeding - Applications and Approaches for*
2 *Developing Improved Cultivars* . 1st ed. Scientific Publishing Services, LTD, CA.
- 3 Duangjit J, Causse M, Sauvage C, Efficiency of Genomic Selection for Tomato Fruit Quality. Submitted.
- 4 Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant*
5 *Gen* 4: 250–255.
- 6 Endelman JB, Jannink JL (2012) Shrinkage estimation of the realized relationship matrix. *G3* 2: 1405–1413.
- 7 Hayes B, Bowman P, Chamberlain A, Goddard M (2009) Genomic selection in dairy cattle: Progress and
8 challenges. *J Dairy Sci* 92: 433–443.
- 9 Kumar S, Chagné D, Bink MCAM, Volz RK, Whitworth C, Carlisle C (2012) Genomic selection for fruit quality
10 traits in apple (*Malus×domestica* Borkh.) *PLoS ONE* 7: 1–10.
- 11 Owens B, Lipka A, Magallanes-Lundback M, Tiede T, Diepenbrock C, Kandianis C, Kim E, Cepela J, Mateos-
12 Hernandez M, Buell C, *et al.* (2014) A foundation for provitamin a biofortification of maize: genome-
13 wide association and genomic prediction models of carotenoid levels. *Genetics* 198: 1699–1716.
- 14 R Development Core Team (2014) *R: A language and environment for statistical computing*. R Foundation for
15 *Statistical Computing*, Vienna, Austria, 2012, ISBN 3-900051-07-0.
- 16 Sauvage C, Segura V, Bauchet G, Stevens R, Phuc Thi D, Nikoloski Z, Fernie AR, Causse M (2014)
17 Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiol*
18 165: 1120–1132.
- 19 Sim SC, Durstewitz G, Plieske J, Wieseke R, Ganai MW, Van Deynze A, Hamilton JP, Buell CR, Causse M,
20 Wijeratne S, *et al* (2012) Development of a large SNP genotyping array and generation of high-
21 density genetic maps in tomato. *PLoS ONE* 7: e40563
- 22 Stein C (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution,
23 *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*. pp. 197–
24 206.
- 25 Storlie E, Charmet G (2013) Genomic selection accuracy using historical data generated in a wheat breeding
26 program. *Plant Gen* 6: 1–9.
- 27 Tributou T, Larzul C, Phocas F (2012) Efficiency of genomic selection in a purebred pig male line. *J Anim Sci*
28 90: 4164–4176.
- 29 Wang C, Habier D, Wolc A, Garrick DJ, Fernando RL, Lamont SJ, Dekkers J, Kranis A, Watson KA (2013)
30 Application of genomic selection using an evenly spaced low-density marker panel in broiler
31 chickens. *Animal Industry Report* 659: 58.
- 32 Wimmer V, Lehermeier C, Albrecht T, Auinger H, Wang Y, Schön C (2013) Genome-wide prediction of traits
33 with different genetic architecture through efficient variable selection. *Genetics* 195: 573–587.
- 34 Würschum T, Reif J, Kraft T, Janssen G, Zhao Y (2013) Genomic selection in sugar beet breeding
35 populations. *BMC Genet* 14: 1–8.
- 36 Zhao Y, Gowda M, Liu W, Würschum T, Maurer HP, Longin FH, Ranc N, Piepho HP, Reif J (2013) Choice of
37 shrinkage parameter and prediction of genomic breeding values in elite maize breeding populations.
38 *Plant Breed* 132: 99–106.