Efficiency of genomic selection for tomato fruit quality

Janejira Duangjit, Mathilde Causse, Christopher Sauvage

Mol Breeding (2016)36:29
DOI 10.1007/s11032-016-0453-3

Efficiency of genomic selection for tomato fruit quality

Janejira Duangjit · Mathilde Causse · Christopher Sauvage

Received: 20 July 2015 / Accepted: 18 February 2016

Abstract Fruit quality is polygenic; each component has variable heritability and is difficult to assess. Genomic selection, which allows the prediction of phenotypes based on the whole-genome genotype, could vastly help to improve fruit quality. The goal of this study is to evaluate the accuracy of genomic selection for several metabolomic and quality traits by cross-validation and to estimate the impact of different factors on its accuracy. We analyzed data from 45 phenotypic traits and genotypic data obtained from a previous study of genetic association on a collection of 163 tomato accessions. We tested the influence of (1) the size of training population, (2) the number and density of molecular markers and (3) individual relatedness on the accuracy of prediction. The prediction accuracy of phenotypic values was largely related to the heritability of the traits. The size of training population increased the accuracy of predictions. Using 122 accessions and 5995 single nucleotide polymorphisms (SNPs) was the optimal condition. The density of markers and their numbers also affected the accuracy of the prediction. Using 2313 SNP markers distributed 0.1 cM or more apart from each other reduced the accuracy of prediction, and no gain in prediction accuracy was found when more markers were used in the model. Additionally, the more accessions were related, the more accurate were the predictions. Finally, the structure of the population negatively affected the prediction accuracy. In conclusion, the results obtained by cross-validation illustrated the effect of several parameters on the accuracy of prediction and revealed the potential of genomic selection in tomato breeding programs.

Keywords Genomic selection · Cross-validation · SNP · Tomato · Metabolomics · Fruit quality

Introduction

Until recently, selection to improve traits of economic interest in crops was achieved through a conventional pedigree-based approach. Thus, individuals of superior phenotypes were selected to mate and produce a progeny showing better performance. With the advent of molecular biology and techniques, it became feasible to tag genes or quantitative trait loci (QTLs) underlying the genetic architecture of traits of interest using molecular markers and linkage mapping. These

Electronic supplementary material The online version of this article (doi:10.1007/s11032-016-0453-3) contains supplementary material, which is available to authorized users.

J. Duangjit
Department of Horticulture, Faculty of Agriculture, Kasetsart University, Bangkok 10900, Thailand

M. Causse · C. Sauvage (✉)
INRA, UR1052 GAFL, Génétique et Amélioration des Fruits et Légumes, 67 allée des chênes, CS60094, 84143 Montfavet Cedex, France
e-mail: christopher.sauvage@avignon.inra.fr

Published online: 02 March 2016

markers could then be used in breeding programs following marker-assisted gene introgression or marker-assisted selection for the selection of several QTLs. However, a small number of recombination events that occurred during the construction of a mapping population limited the accuracy of QTLs and thus marker-assisted selection, especially for polygenic traits (Price 2006; Zhu et al. 2008). This approach rapidly revealed its shortcomings with long selection cycles and low efficiency for traits relying on the effect of multiple genes of minor effect (Xu et al. 2012) and/or a rapid reduction of the genetic variability following the fixation of main QTLs (Blanc et al. 2008). In the last 10 years, whole-genome sequencing of major crops and model plant species was initiated following the advent of high-throughput sequencing technologies. A decade after the first plant genome of *Arabidopsis thaliana* was made publicly available, genome sequences of many plants have been continuously published with more than 100 different fully sequenced genomes (Michael and VanBuren 2015; see also CoGePedia 2015), including four Solanaceae species, eggplant, tomato, pepper and potato (Xu et al. 2011; Sato et al. 2012; Qin et al. 2014; Hirakawa et al. 2014). Whole-genome and expressed sequence tag sequencing efforts also provided thousands of single nucleotide polymorphisms (SNPs).

Thus, it became feasible to study the genetic architecture of complex traits using the combination of dense genetic marker information and genome-wide association (GWA) approaches and to decipher more accurately the loci responsible for variation in traits. However, while the sequencing technologies produce large amounts of genomic data, accurate phenotyping is still limited, especially in large populations. Moreover, both QTL and GWA approaches are still unable to capture loci of minor effect, mainly due to the stringent statistical threshold used to avoid the detection of false-positive effects (Manolio et al. 2009).

Genomic selection (GS) has been proposed as a new tool for selecting livestock and crops by predicting the performance of individuals for traits of interest following statistical models (Meuwissen et al. 2001). In theory, GS is able to capture the Mendelian component of trait variation without any record of phenotypes if molecular markers tag all the loci underlying this trait (using linkage disequilibrium). GS relies on a dense genome-wide coverage to produce genomic estimated breeding values (GEBV) from all the marker effects. Many simulated and empirical GS experiments were conducted in the late 2000s (Calus et al. 2008; Solberg et al. 2008). Some plant populations were generated from biparents (Wang et al. 2014; Beyene et al. 2015) or multiple parents (Kumar et al. 2012), but in most cases the populations studied were broad-based (Asoro et al. 2011), where cross-validation experiments were commonly performed (Resende et al. 2012; Würschum et al. 2013; Wang et al. 2014; Zhang et al. 2015) to assess GS feasibility. In practice, four main steps are required in cross-validation: (1) assigning individuals to training and validation sets; (2) estimating marker effects based on genotypes and phenotypes in the training set; (3) calculating GEBV by using genotype information from the validation set fitted in a predictive model resulting from the previous step and (4) estimating accuracy of prediction by correlation between predicted and observed breeding values in the validation dataset (Gianola et al. 2011; González-Camacho et al. 2012; Crossa et al. 2013). This approach uses information from all markers across the genome regardless of their effect sizes, providing reliable prediction when compared to conventional approaches where only significant markers are included in calculations (Meuwissen 2007). Therefore, GS has opened the door for a new generation of breeding programs, as reviewed by Meuwissen et al. (2013). This led to studies of GS application in different crops, for example oat (Asoro et al. 2011), maize (Owens et al. 2014), wheat (Storlie and Charmet 2013), sugarcane (Gouy et al. 2013), apple (Kumar et al. 2012), sugar beet (Würschum et al. 2013) and oil palm (Cros et al. 2014). These cases have shown the potential of applying GEBV for selection in future breeding programs. Many statistical models (viz., parametric, semi-parametric and nonparametric) with different assumptions about the marker effect distribution have been created. It has been shown that they were able to predict breeding values in plants (Heslot et al. 2012). Several statistical models were provided and implemented in free R packages. For instance, the rrBLUP and BGLR packages provide a broad range of methods such as GBLUP, BayesB, BayesA, BayesC, Bayesian Lasso, Bayesian Ridge Regression and RKHS (Pérez and de los Campos 2014; Endelman 2011). These methods differ mainly by the theoretical distribution of the SNP effects.

Cultivated tomato is one of the major horticultural crops, with both economic and scientific importance. For many years, linkage maps have been used to characterize genetic resources and map QTLs, as well as increase selection efficiency in breeding programs (Tanksley et al. 1992; Tanksley and Fulton 2007; Sim et al. 2012a; Pascual et al. 2014). Detailed genetic maps have permitted map-based cloning of several economically important genes such as disease resistance genes (Martin et al. 1993; Jones et al. 1994) and fruit weight (FW) and sugar content (SC) QTLs (Frary et al. 2000; Fridman et al. 2000; Causse et al. 2004). Marker-assisted selection has been shown to be feasible but limited by the low polymorphism at the intraspecific level (Lecomte et al. 2004). The tomato genome has been fully sequenced (Sato et al. 2012). Additionally, a transcriptome-based high-throughput genotyping platform of 7720 SNPs, specific to the cultivated tomato genome, has been developed (Sim et al. 2012a). These genomic tools help in understanding the genetic background of tomato and, meanwhile, provide challenges for their application in tomato breeding (Blanca et al. 2012; Sim et al. 2012b; Blanca et al. 2015). Recently, a GWA study was reported to be efficient in tomato (Ranc et al. 2012; Xu et al. 2013; Ruggieri et al. 2014; Sacco et al. 2015), and the study of a large panel of tomato accessions using this 8K SNP genotyping array identified 44 loci controlling fruit metabolic traits (Sauvage et al. 2014). Selection for tomato fruit quality has been requested by consumers for years but remains difficult due to the number of traits concerned and their complex inheritance (Causse et al. 2002; Lecomte et al. 2004; Pascual et al. 2014). GS could thus be a good alternative for increasing selection efficiency for such complex objectives.

In this context, we have tested GS approaches in fresh market tomato for fruit quality traits, using a collection of tomato accessions and the high-throughput genotyping data available. The goals of this study were to (1) evaluate the accuracy of GS for several metabolomic and quality traits by cross-validation, and (2) assess the impact of factors including training population size, number and density of markers, and relatedness on the accuracy of GS. The long-term goal of this research is to optimize GS approaches and use it in tomato breeding programs to improve fruit quality.

Materials and methods

Plant material and SNP genotyping

We used the molecular data described in Xu et al. (2013), Sauvage et al. (2014) and Pascual et al. (2016). The panel was composed of 163 tomato accessions conserved in INRA-Montfavet, France (see Sauvage et al. 2014). Based on the classification of Blanca et al. (2015), the collection contains three (sub)species of tomato: 29 *Solanum lycopersicum* L. (SL), 118 *Solanum lycopersicum* var. *cerasiforme* (SLC) and 16 *Solanum pimpinellifolium* (SP) (Supplemental Table 1). The accessions were genotyped using the 7720 SNP array on the Infinium assay (Illumina), developed by the Solanaceae Coordinate Agricultural Project (SolCAP) (Hamilton and Buell 2012; Sim et al. 2012a) and the protocols as recommend by the manufacturer. The probe sequences and SNP information are available at SolCAP Web site (http://solcap.msu.edu). Well-supported SNPs were identified at 90 % rate threshold per locus. Filtering step was set in the range of minor allele frequency (MAF) from 0.037 to 0.45 as reported in Sauvage et al. (2014). The SNPs and accessions that were unmatched for all conditions were removed via option *–maf* in PLINK! version 1.9 (Purcell et al. 2007). A set of 5995 reliable SNP markers was used to support our work.

Genetic relationship and population structure assessment

The matrix of the identical-by-state (IBS) distance for all pairs of accessions was calculated in PLINK! to illustrate genetic kinship of the tomato population (Purcell et al. 2007). The IBS distance matrix was visualized in R computer software (R Development Core Team 2008). To assess the population stratification, the most likely number of clusters K in all simulations were assumed to be in the range of $K = 1$ to $K = 10$. Ten replicates were conducted in Structure software (Falush et al. 2003) for each K with a burn-in period of 1×10^6, followed by 5×10^6 Markov chain Monte Carlo (MCMC) steps. The Evanno correction was then applied (Evanno et al. 2005).

Phenotype data

The population of 163 accessions was evaluated in 2007 and 2008 in plastic tunnels and phenotyped as described in Xu et al. (2013) and Sauvage et al. (2014) for a total of 35 metabolic traits (ASA, asparagine, aspartate, beta-alanine, citrate, DHA, erythritol, fructose, fucose, GABA, galacturonate, glucuronate, glutamate, glutamine, 2-oxoglutarate, glycerol-3P, inositol-1P, lysine, malate, maltitol, maltose, methionine, nicotinate, phenylalanine, proline, putrescine, rhamnose, saccharate, serine, sucrose, threonate, threonine, tocopherol, tyramine and xylose) and ten quality traits [pH, firmness, FW, soluble solid content (SSC), SC, titratable acid (TA), locule number (LCN) and three color components: lightness (L), color from red to green (a*) and color from yellow to blue (*b) assessed with a Konica Minolta CR-300 chromameter]. For all the traits, correlation of the phenotypes over the 2 years higher than 0.6 and the average over the 2 years was used. Non-normally distributed traits were $\log_{10}$-transformed.

Statistical model in genomic prediction

Previous empirical studies that investigated the effect of predictive parametric models on the accuracy of the phenotype prediction showed that most of the models performed almost identically, especially for polygenic traits (Daetwyler et al. 2013; Howard et al. 2014). On this basis, to obtain accuracy, the ridge regression best linear unbiased prediction (rrBLUP) statistical model was used (Endelman 2011). This model is equivalent to best linear unbiased prediction (BLUP), and it runs quickly, as it uses an algorithm for mixed models with a single variance component besides the residual error (Kang et al. 2008) and is implemented as an R package (see http://cran.r-project.org/web/packages/rrBLUP/index.html).

Thus, the genotype information on the training set was fitted in a model $Y = 1\mu + Xg + e$, where Y is a vector of observed phenotypic value, μ is the overall mean of the training set fitted as covariates based on the genotype, and e is a residual effect. Once the SNP effects were calculated using the mixed-model solver function (mixed.solve), the GEBV of the validation set were predicted based on genotype information and SNP effects. Graphical representations were obtained using R v3.0.2. The heritability of each trait was estimated through genetic variance components computed at step 0 of the multi-locus mixed model in GWA analyzed as described in Sauvage et al. (2014).

Training and validation set design

The power of genomic prediction was assessed by comparing assessed GEBV with the actual phenotyping values. The impact of several factors was studied:

(i) *The size of training set* From the population of 163 tomato accessions, different numbers of accessions were used to fit the model, with the criterion of having at least 40 accessions (25 %) in the training and validation sets. Estimations were repeated at 5 and 10 % increments (i.e., 30, 40, 50, 60 and 70) up to the maximum of 75 % of the population (122 individuals). The remaining accessions were used as a validation set, with all 5995 SNP markers. Prediction accuracies were tested by the correlation between predicted breeding values and evaluated phenotype (true breeding values) for 1000 iterations.

(ii) *Marker density* To assess the impact of the marker density on prediction ability, two sets of markers were used to fit the model. The first was a set of 5995 SNP markers as described in (i). The second was a set of 2313 markers published by Blanca et al. (2015), derived from the same genotyping platform but gathering markers distributed no less than 0.1 cM apart across the tomato genome.

(iii) *Number of markers* To determine the effect of marker numbers in prediction accuracy, sets of 500, 1000, 1500 and 2000 were randomly selected and used for comparison with the 2313 SNP marker set, while sets of 500, 1000, 2000, 3000, 4000 and 5000 SNPs were randomly selected and used for comparison with the 5995 SNP marker set. To avoid the effect of unequal training sets and differences in a genetic relationship, across all traits, a fixed training set of 122 accessions (75 %) was randomly chosen for all 1000 replicates. Accuracies were determined as described above in (i).

(iv) *Relatedness* Based on the results of the IBS distance for all pairs of accessions, two

clusters were created. Subpopulation I contained 16 SP accessions, and subpopulation II was a combined group of SLC and SL (147 accessions). A random sample of 60 % (98 accessions) from subpopulation II was used as a training set, while validation sets were (i) remaining accessions from subpopulation II and (ii) all accessions from subpopulation I. These analyses were performed with all the 5995 markers.

Results

Population structure and trait variation

The studied population was composed of 163 accessions representative of the variation present in cultivated, cherry and wild-related tomato (Ranc et al. 2012). This population showed a large range of genetic diversity and different degrees of relatedness. Following the Evanno correction, the analysis of the population structure revealed two ancestral populations composed of (1) the SL and the SLC accessions ($n = 147$) and (2) a cluster of 16 SP accessions. These two groups were also identified on the kinship matrix (Supplemental Figure 1). Phenotypes were measured on the 163 accessions over 2 years in a row, and only the 45 traits (ten fruit quality traits and 35 metabolic composition traits) highly correlated ($R^2 > 0.6$) over these 2 years were analyzed to circumvent any genotype × environment interaction that may have affected our results. Briefly, as described in Sauvage et al. (2014) for the metabolomics traits and in Xu et al. (2013) for the ten quality traits, significant differences were assessed between genetic groups for most of the traits, especially between SP and SL-SLC. The phenotypic heritability ranged from 0.124 to 0.940, and previous GWA studies revealed 35 associations for 18 metabolic traits (for 17 traits, no association was detected) and 37 for ten quality traits (Sauvage et al. 2014; Pascual et al. 2016; Table 1).

Prediction accuracy and heritability

The prediction accuracy using 75 % of the population (122 accessions) with 5995 markers in the training step showed large differences across traits (Table 1, reporting the median and standard deviation of the correlation values between the observed and the predicted phenotypes). The maximum accuracy was found for FW (0.814 ± 0.068), and the lowest accuracy was obtained for threonate (0.052 ± 0.115), with heritability of the two traits very high (0.88) and very low (0.168), respectively. The correlation between mean accuracy and heritability was high ($r = 0.69$), but some traits showed discrepancy (Fig. 1), particularly for traits with extreme heritability.

Prediction with different sizes of training populations

Maximum accuracies were found for the training population composed of 75 % of the accessions, except for six out of 45 traits (13.3 %) (tocopherol, threonate, glucuronate, galacturonate, fucose and ASA), which showed a slight decrease in average accuracy (ranging from −0.006 to −0.001) (Fig. 2; Supplemental Table 2). Decreasing the size of the training set from 75 % (122 accessions) to 25 % (41) of the population resulted in a reduction of accuracy with a maximum increment of 0.341 of average accuracy for pH and minimum increment of 0.022 of average accuracy for threonate content (Supplemental Table 2). However, high prediction accuracy was obtained even with a small-sized training population for several traits. For instance, in soluble solid content, an accuracy of 0.679 ± 0.047 was obtained when the model was trained by 40 accessions (25 %). The highest accuracies using optimal conditions were obtained for FW, SSC and proline (Fig. 2; Supplemental Table 2). The standard deviation of accuracy also increased with the size of the training set.

Prediction with different sets of markers

Increasing the number of markers in the statistical model raised the prediction accuracy. Not all the traits responded similarly to the number of markers (Fig. 3). For example, prediction of FW did not vary much when using different numbers of markers from the whole set of 5995 markers, ranging from 0.777 ± 0.03 to 0.780 ± 0.006 when using 500 and 5000 markers, respectively (Fig. 3a; Supplemental Table 3). Higher accuracies were obtained with a higher number of markers for every trait, except for serine and firmness (−0.002). Color component b*

Table 1 Prediction ability obtained with rrBLUP for 45 traits

Trait	Mean accuracy	Median	SD	Heritability	PVE	No. of associations
Metabolome traits						
ASA	0.601	0.611	0.086	0.553	0.561	5
Asparagine	0.449	0.461	0.130	0.417	0.220	1
Aspartate	0.126	0.133	0.130	0.284	0.162	1
Beta-alanine	0.413	0.419	0.096	0.624	NA	0
Citrate	0.246	0.251	0.111	0.423	0.181	1
DHA	0.658	0.665	0.082	0.595	0.743	2
Erythritol	0.621	0.628	0.083	0.534	0.358	2
Fructose	0.594	0.602	0.090	0.565	0.386	2
Fucose	0.438	0.445	0.111	0.415	0.237	1
GABA	0.639	0.644	0.088	0.415	0.237	1
Galacturonate	0.439	0.446	0.109	0.235	NA	0
Glucuronate	0.363	0.365	0.126	0.183	NA	0
Glutamate	0.522	0.533	0.098	0.706	NA	0
Glutamine	0.478	0.488	0.119	0.377	NA	0
Glycerol-3P	0.444	0.445	0.091	0.471	NA	0
Inositol-1P	0.432	0.436	0.116	0.276	NA	0
Lysine	0.210	0.211	0.117	0.322	NA	0
Malate	0.519	0.525	0.100	0.642	0.390	2
Maltitol	0.674	0.679	0.078	0.853	NA	0
Maltose	0.471	0.477	0.107	0.619	NA	0
Methionine	0.239	0.241	0.114	0.427	NA	0
Nicotinate	0.705	0.712	0.066	0.595	0.279	1
2-Oxoglutarate-	0.509	0.516	0.101	0.541	NA	0
Phenylalanine	0.352	0.360	0.109	0.391	NA	0
Proline	0.709	0.719	0.074	0.773	0.461	2
Putrescine	0.473	0.477	0.091	0.241	NA	0
Rhamnose	0.529	0.538	0.102	0.579	0.504	4
Saccharate	0.309	0.316	0.117	0.293	NA	0
Serine	0.229	0.235	0.113	0.124	NA	0
Sucrose	0.571	0.574	0.089	0.585	0.439	3
Threonate	0.052	0.041	0.115	0.168	0.170	1
Threonine	0.363	0.370	0.120	0.348	0.187	1
Tocopherol	0.451	0.460	0.119	0.306	0.224	1
Tyramine	0.604	0.609	0.081	0.612	0.472	3
Xylose	0.376	0.384	0.110	0.325	NA	0
Quality traits						
a*	0.567	0.591	0.151	0.645	0.481	1
b*	0.581	0.598	0.127	0.750	0.430	1
Firmness	0.614	0.624	0.096	0.416	0.431	1
FW	0.814	0.824	0.068	0.880	0.800	9
L	0.617	0.632	0.114	0.710	0.385	3
LN	0.423	0.433	0.096	0.941	0.894	2
pH	0.506	0.516	0.102	0.540	0.240	2
Soluble solid	0.714	0.724	0.074	0.600	0.550	9
Sugar content	0.649	0.653	0.081	0.611	0.489	3
TA	0.619	0.630	0.091	0.670	0.470	6

Mean accuracies are the average of correlation values of evaluated phenotypes and GEBV predicted by rrBLUP using 122 accessions (75 %) in training step with 5995 markers (1000 imputations). Median and standard deviation (SD) are indicated. Trait heritability, the number of associations and percentage of variation explained are from Sauvage et al. (2014) for metabolomic traits and Pascual et al. (2016) for quality traits. *No. of associations* shows the total of loci that were significantly associated with the particular trait

PVE represents percentage variance explained by the identified loci

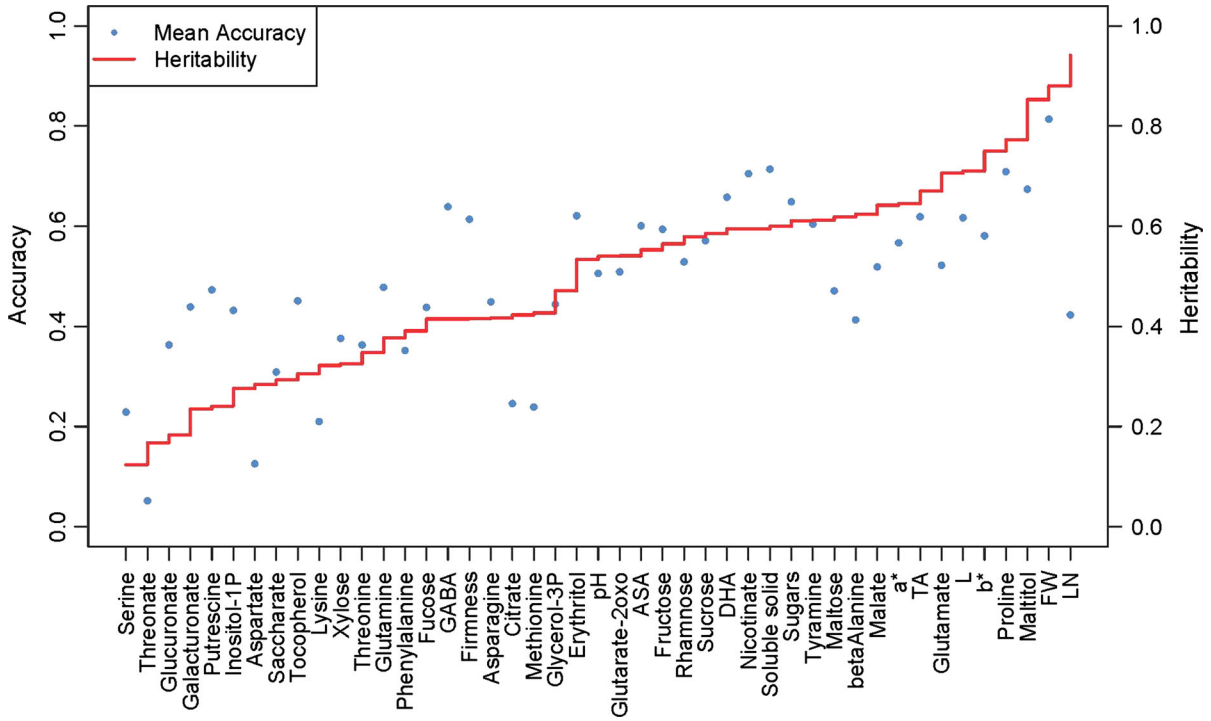

Fig. 1 Mean accuracy and heritability for 45 fruit quality traits. Mean accuracy is based on rrBLUP predictive ability (analyzed using 5995 SNP markers and 122 accessions in the training panel). Heritability adapted from Sauvage et al. (2014) and Pascual et al. (2016). Each *blue dot* represents mean accuracy and a *red line* represents heritability estimate

had the highest increase (0.096) in accuracy when marker numbers increased from 500 to 5000. In every case, the standard deviation of accuracy increased when the number of markers in the model decreased (Supplemental Table 3).

When using a set of 2313 markers distributed 0.1 cM or more apart from each other (Supplemental Figure 2), as previously, greater accuracies were observed when more markers were used and the greatest average accuracy remained for FW, with increases ranging from 0.604 ± 0.049 (500 markers) to 0.661 ± 0.013 (2000 markers), which was significantly different ($P < 0.05$). The highest increase in accuracy (+0.150) concerned asparagine content (0.326 ± 0.153 to 0.476 ± 0.015). However, for most traits, using markers from the set of 5995 markers gave higher accuracies. For example, for malate content, accuracies were estimated to be 0.485 ± 0.072 and 0.184 ± 0.057 with the 5995- and 2313-marker set, respectively (Fig. 3b; Supplemental Table 3). Accuracies were not significantly different for seven traits (15.6 %) (aspartate, glycerol-3P, inositol-1P, saccharate, tyramine, b* and L) when compared to the set of 2313 markers (Supplemental Table 3). With a subset of 2000 SNP markers, the prediction was much more variable, as standard deviations calculated from rrBLUP ranged from 0.004 to 0.156 (when the training populations contained 122 accessions).

Prediction with training populations from different subpopulations

To examine the effect of the genetic relatedness between accessions on the accuracy of GEBV, a random set of 98 accessions from subpopulation II (composed of 147 accessions of SL and SLC) was trained and validated by the accessions within the group and from the other group (SP). Outperformance was found if related subpopulations were used in the validation step. For most of the traits, prediction ability was reduced when validated with members of the out-group (SP) (Supplemental Table 4). For GABA content, for example, the average accuracy was 0.408 ± 0.218 when validating the model with

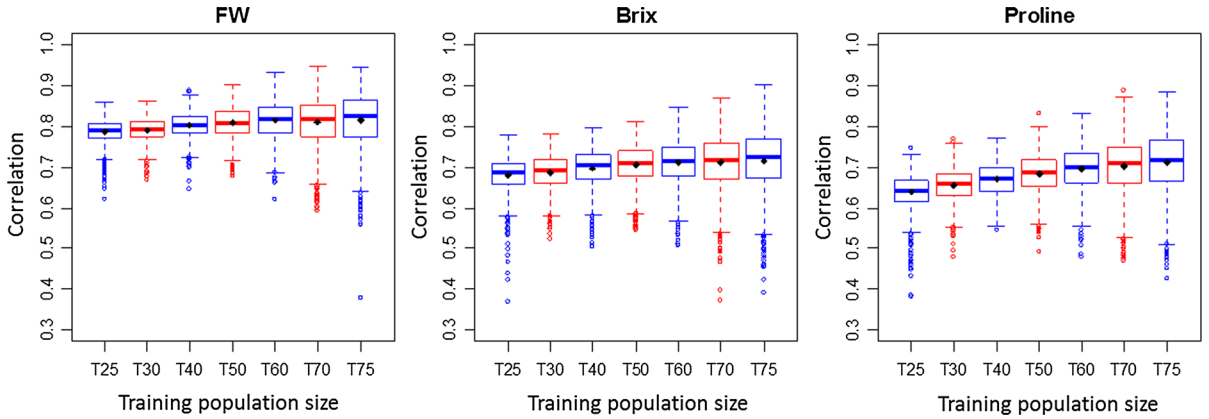

Fig. 2 *Boxplots* representing the impact of the population size in the training set on prediction accuracy. *Boxplot* of correlation values of evaluated phenotypes and GEBV predicted by rrBLUP with a set of 5995 markers for fruit weight (FW), soluble solid content (SSC or Brix) and proline content. *Midbar* and *dot* indicate average and median from 1000 imputations. *T25*, *T30*, *T40*, *T50*, *T60*, *T70* and *T75* indicate the percentages of accessions in the population used in the statistical model training step

SLC and SL members and it dropped to a negative value (-0.264 ± 0.065) when validated with SP accessions (Fig. 4a). Similarly, one of the most significant drops was found for ASA content. The average accuracies were 0.561 ± 0.174 when using the in-group (SLC and SL). Conversely, in the out-group validation, the accuracy dropped to 0.179 ± 0.093 (Fig. 4b).

Discussion

Fruit quality is a multi-component trait, with many components of variable heritabilities and difficult to assess; GS could thus vastly help its improvement. In the present study, we successfully evaluated the accuracy of prediction of the breeding value for a large range of quality traits and investigated the impact of parameters such as population size, relatedness between individuals, marker number and density on the prediction ability (in terms of accuracy) of phenotype prediction. We used the rrBLUP model for a set of 45 traits covering a large range of heritability and genetic architecture (as shown by the GWA results) in a broad-based population of tomato accessions. To our knowledge, these results are the first ones reported in tomato, an economically and scientifically important species, paving the way for further investigations into the potential of GS in this species, especially for quality traits.

Accuracy of GS using rrBLUP model

We tested GS with phenotypic data on 45 traits and 163 tomato accessions, selected to represent a large range of genetic variation. Accessions were characterized in greenhouse trials for 2 years under similar growth conditions. This set of accessions was previously studied for GWA, and zero (for 17 traits) to nine significant associations (for FW and SSC) could be detected, explaining up to 89 % (for LN) of the phenotypic variation (Sauvage et al. 2014; Pascual et al. 2016) and illustrating the variety of genetic architecture underlying quality traits. We used rrBLUP as it was less computationally demanding than other approaches (Würschum et al. 2013), and it provided very similar results to other methods (preliminary results not presented). The results indicated that GS could be applied to and was encouraging to be used in tomato breeding programs. The highest accuracy was obtained when 122 accessions were used to train the model and 5995 markers were employed.

Overall, using rrBLUP, we obtained accuracies ranging from 0.052 ± 0.115 (threonate content) to 0.814 ± 0.068 (FW), which represents a very large variation from almost unpredictable traits to highly predictable traits. In addition, we obtained better higher accuracies than similar studies using rrBLUP. For example, in spruce and rice, where wood attributes and grain yield traits were predicted, the highest

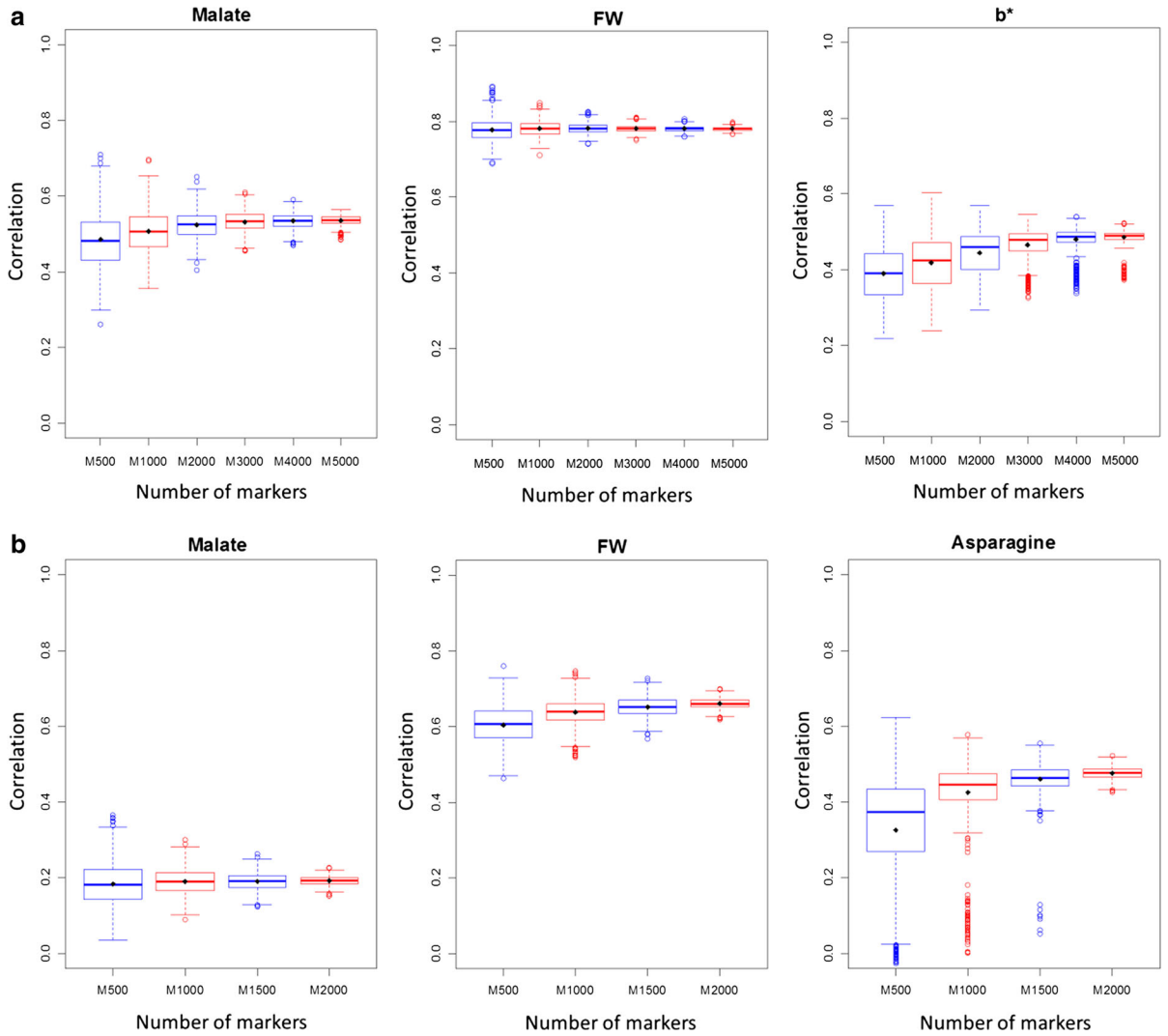

Fig. 3 The impact of the number and density of markers on prediction accuracy. *Boxplot* of correlation values of evaluated phenotypes and GEBV predicted by rrBLUP with a training set of 122 accessions (75 %) for some predicted traits of tomato. *Midbar* and *dot* indicate average and median from 1000 imputations. M500, M1000 and so on indicate the number of markers used in the statistical model training step. **a** Plots on *top row* from analyses with a set of 5995 markers; **b** plots from *bottom row* from analyses with a set of 2313 markers

accuracies reached 0.58 (diameter) and 0.63 (flowering time), respectively (see El-Dien et al. 2015; Spindel et al. 2015). In these two studies, rrBLUP also demonstrated its higher power over other models (e.g., GRR) to predict with a better accuracy on average, supporting our results. However, in apple (Kumar et al. 2012), rrBLUP was also more accurate than the conventional BLUP but on average high to very high accuracies were reported (ranging from 0.68 to 0.89), reflecting the effect of the size of the training population.

Characteristics of the traits affect prediction accuracy

Preferably, breeders want to apply GS to all the traits, whatever their heritability and genetic architecture. Differences in trait heritability and genetic architecture were found based on the GWA approach (Sauvage et al. 2014; Pascual et al. 2016). Under optimal conditions (122 accessions in the training set and 5995 markers), traits with high heritability ($h^2 > 0.6$) had higher prediction ability than those with a lower

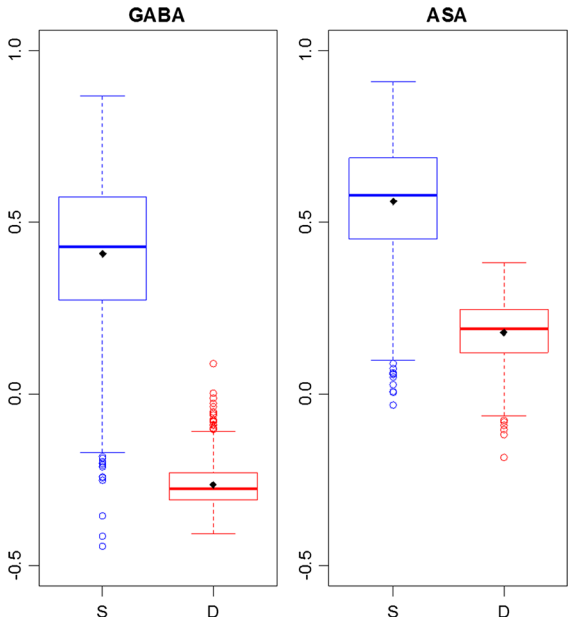

Fig. 4 *Boxplot* of correlation values of evaluated phenotypes and GEBV predicted by rrBLUP using a set of 5995 markers and a training set of 98 accessions (60 %) for **a** GABA content and **b** ASA content. *Midbar* and *dot* indicate average and median from 1000 imputations. *S* and *D* indicate validation populations from same and different subpopulations, respectively

heritability. This correlation was consistent with previous as previously studies as expected (Hayes et al. 2009; Lorenz et al. 2012; Wimmer et al. 2013). This can be explained by the fact that a trait with high heritability was underlined by a strong genetic component; consequently, it will be more predictable using GS (Resende et al. 2012). Twenty-three traits (50 %) showed high prediction ability (accuracy >0.5). From this set, the majority of traits, 21, were highly heritable ($h^2 > 0.5$) and firmness and GABA content had heritability around 0.41. An inconsistent trend was found for LN, with high heritability (0.94) but low accuracy (0.422). Such an observation is supported by the results of Pascual et al. (2016), where two QTLs were responsible for most of the trait variation (89 %). For such a trait, taking the QTL information into consideration, as a cofactor of the model, may be important. In a general way, taking into account below-threshold QTL effects by GS has been shown to improve the selection efficiency of low-heritability traits (Calus et al. 2008). However, increasing the number of markers does not always improve the prediction. Especially for a trait controlled by only a few loci (e.g., LN), over-parameterization effects occurred when fitting too many markers into rrBLUP (Resende et al. 2012). This can be fixed by performing shrinkage as suggested by Maltecca et al. (2012). Furthermore, the use of GS for traits controlled by a limited set of loci can be debated, as marker-assisted selection could be more powerful for improving the traits.

Several factors (marker density, the number of markers, training population size and relatedness) have an effect on prediction accuracy estimated by the rrBLUP method. Exploiting only a small number of markers (500 SNPs) was enough to obtain high accuracy (>0.50) for 18 traits (40 %), within the range 0.509–0.777, for a training population size of 122 accessions. With the largest number of markers (5995 SNP markers), 21 traits (46.6 %) showed good accuracy (>0.50), in the range 0.505–0.780, when the same training population size was used (Table 1). Based on these results, we can expect that high accuracy may be obtained when applying GS to tomato breeding. Several factors could be improved: (1) The number of accessions in a population; we used 122 accessions as the largest training population size. This is relatively small compared to other studies in horticultural crops with around 1000 accessions used (Kumar et al. 2012; Würschum et al. 2013). (2) Efficiency depends on the trait and prediction accuracy could be improved if we consider the population structure effect in the model. (3) There will be less genetic variation in a real breeding program compared to the broad-based population tested in this study.

An increase in marker numbers in the rrBLUP model increased the accuracy of prediction and decreased its standard deviation, but for most of the traits a plateau was reached when 3000 markers were incorporated into the model. This indicated that only one SNP per gene provided enough information, as the 5995 markers were originated from approximately 3000 genes (Sim et al. 2012a). When considering the increment of accuracy by adding markers into the model (500–5000), a plateau was reached for some traits but increase continued for others (Fig. 3). Therefore, including more markers for the traits controlled by many low-effect loci may contribute to increasing the chance of fitting more markers with low effect, which explain more variation in the model. The hypothesis that genetic architecture has an influence on the performance of prediction is also proved by a simulation study (Wimmer et al. 2013).

Effect of SNP distribution across the genome

The accuracy was lower when using markers relatively evenly distributed (every 0.1 cM) across the genome than using the same number of markers regardless of their location (Fig. 3; Supplemental Table 3). This can be explained by the structure of the tomato genome. Although tomato has a gene density of 6.7 kb/gene in euchromatin with a total of 34,727 genes, the gene density is highly heterogeneous (Wang et al. 2006; Sato et al. 2012). Furthermore, some chromosomes carry more QTLs than others; for example, many QTLs of important traits were detected on chromosome 2 (e.g., for FW, fruit morphology and LCN) (Frary et al. 2000; Causse et al. 2002; Muños et al. 2011). Using evenly distributed markers decreased the chance of fitting markers underlying the traits, especially since the patterns of recombination (and linkage disequilibrium) are not linear along tomato chromosomes. Given a specified number of markers, it is thus important to use markers in gene-rich regions to increase the probability of tagging QTLs and thus accuracy. The regions to enrich in markers can be also defined when the genetic architecture of traits of interested has been revealed. High-throughput sequencing approaches, such as genotyping-by-sequencing or exon capture, will help to tag most of the 'haploblocks' in the tomato genome. The recent release of hundreds of tomato genome sequences (Aflitos et al. 2014; Lin et al. 2014) is also a valuable resource to achieve this objective.

The most effective number of SNPs required is still under debate. It has to be assessed to implement the GS strategy efficiently. This number depends on parameters such as individual relatedness, population structure and linkage disequilibrium. Thus, simulation of genomic and phenotypic datasets can contribute to the success of a GS and its application for breeders to develop tools, such as SNP arrays. Towards this objective, simulation packages have been developed (e.g., AlphaDrop; see Hickey and Gorjanc 2012, but see Daetwyler et al. 2013 for a complete review) and have to be implemented in any GS cross-validation experiment to explore larger sets of parameters such as SNP density and number. In tomato, the use of simulated datasets represents the next steps in our approach for the development of next-generation arrays, for example.

Prediction of breeding values in different subpopulations

In order to assess genetic prediction ability, we separated accessions based on species criteria to study the effect of genetic relatedness and population structure. However, with the limited number of accessions in SL, we divided the population into two groups based on IBS analysis and tested the effect of training and validating within or across groups: (1) model trained and validated within SLC + SL group or (2) trained with SL + SLC group and validated by SP accessions (the other way around could not be done due to the limited number of accessions in SP). As expected, on average, higher accuracy was obtained when training and validating within the group. For most of the cases, the accuracy dropped dramatically when the model was trained by information from SL + SLC and validated by information from SP (Supplemental Figure 1). The antagonistic trend was obvious for the traits with significant differences in the group means. Genetic differences between SL and SLC were less than those between SL or SLC and SP, which supports our results (Bauchet et al. 2014). Therefore, the effect of genetic relationship on genomic prediction confirmed the results obtained by Habier et al. (2007, 2010) in cattle. In addition, for some traits, high prediction accuracy was obtained even in a small size of training population (Fig. 2). This can be explained by the fact that a large training population size tends to include more accessions from different groups and, therefore, reduces prediction accuracy. This observation is supported by a study in maize where subpopulation structure affected accuracy (Windhausen et al. 2012).

Similarly to the GWA approach, the success of GS is strongly linked to the linkage disequilibrium decay and the number of markers that tag the genome of the target species. In this case, the ideal situation is to have very short LD decay (over a few kb) and sufficient number of markers to tag every LD block. However, since LD decay is shaped by natural and artificial forces, such as the strength of natural selection, reproduction regimes (e.g., selfing) or domestication and breeding, variable degrees of linkage are expected depending on variable genomic distances. In tomato, many studies investigated the consequences of modern breeding for genome-wide LD patterns (Sim et al. 2012b; Sato et al. 2012; Ranc et al. 2012; Xu et al.

2013; Sauvage et al. 2014). The population used in the present study was previously investigated by Sauvage et al. (2014), who showed that LD patterns were different between the three tomato groups (admixed type or SLC, cultivated type or SL, and wild type or SP). LD in market and cherry types (SL and SLC) was high but lower in the wild type (SP). Such LD influence has been observed in a GS experiment in rice (Wimmer et al. 2013), suggesting that accuracy of marker effect estimation decreases as the training population size increases. Thus, the advent of molecular techniques such genotyping-by-sequencing or very dense SNP arrays will improve genomic prediction in tomato by providing an extensive representation of LD patterns in mapping populations.

Implications for tomato breeding program

We have evaluated the accuracy of GS using rrBLUP models for 45 traits from a broad-based population of tomatoes from the INRA core collection. The accessions covered the breadth of alleles present in tomato populations. The large representation of cherry tomatoes is due to (1) their admixed position between SL and SP (with allele frequency more equilibrated than SL) and (2) their interest as a source of favorable alleles for fruit quality. Nowadays, sequencing and genotyping are cheaper and more effective, while phenotyping is a limiting step, particularly for improving fruit quality and composition. Recently, hundreds of genome sequences were released in the cultivated tomato clade by whole-genome sequencing (Aflitos et al. 2014; Lin et al. 2014), which constitutes a large basis for further GS analyses. The millions of SNPs discovered represent a unique opportunity to assess whether or not increasing marker density, up to hundreds of thousands of markers, makes any difference in phenotype prediction. Size and composition of training population are also important factors influencing GS performance. In our study, the accuracy was improved when increasing population size, but not when the composition of the training population was heterogeneous. Such a trend was also found in other GS experiments such as in sugar beet, oat or white spruce (Asoro et al. 2011; Beaulieu et al. 2014; Würschum et al. 2013). These factors, together with other important factors discussed above (viz., number and density of markers), can be classed as operative factors, which can be optimized to increase prediction accuracy. Other factors are genetic factors, which cannot be modified. They involve the genetic background of the population (i.e., population structure, relatedness of individuals), heritability and genetic architecture of the traits. The optimization of the training populations to improve accuracy has received much interest recently. Up to five sampling algorithms were assessed in Isidro et al. (2015), demonstrating the benefit of this approach to increasing the accuracy, especially if the population is structured (Rincent et al. 2012). Applying such an optimization would have notably required a larger training population, especially in the subpopulation II composed only of 16 SP accessions. However, with the increase in genomic and phenotypic data in tomato, the optimization of training panels will have to be included in the GS strategy.

Over the long term, GS together with recurrent selection should help improve fruit quality components. GS could be an effective way to increase genetic gain by genotyping and phenotyping individuals in the training panel and then predicting GEBV of other tomato lines. Sequentially, selection in subsequent cycles can be based on GEBV. For some traits, the correlation between experimental breeding value and GEBV reached a high level of accuracy (maximum = 0.81). This could reduce the time and cost to recombine the best individuals and create populations with improved quality. This will shorten the time and reduce cost in the selection, as previously proposed (Heffner et al. 2009; Windhausen et al. 2012). Based on our results, prediction within SL and SLC accessions is strongly encouraging while prediction across populations is not favorable. Additionally, enlarging the training population will promote accuracy.

Overall, the present study, in addition to the ones published to date in crops, contributes to answering the question raised by Jonas and de Koning (2013): 'Does GS have a future in plant breeding?' Yes, the availability of an increasing number of crop genomes, the reports of high accuracies for traits of interest as well as methodological refinements (such as the implementation of genotype-by-environment interaction in prediction models) are notably contributing to paving the way for a successful implementation of GS in plant breeding.

Acknowledgments We gratefully acknowledge the support for JD from the Embassy of France in Thailand in Junior Research Fellowship Program 2014. The INRA SelGen Méta-programme also funded this work (http://www.selgen.inra.fr/en). The authors are indebted to Yolande Carretero and Renaud Duboscq from INRA UR1052 who provided the biological material and performed the molecular work required. Finally, we would like to thank David Cros (CIRAD, Montpellier, France) for his help and advice on our work, as well as Marie Christine Le Paslier and Dominque Brunel from INRA-EPGV for their help with the SNP genotyping.

Authors' contributions MC and CS designed the study. JD and CS analyzed the data, and JD, MC, and CS wrote the manuscript. All authors read and approved the final version.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Aflitos S, Schijlen E, de Jong H, de Ridder D, Smit S, Finkers R, Wang J, Zhang G, Li N, Mao L, Bakker F, Dirks R, Breit T, Gravendeel B, Huits H, Struss D, Swanson-Wagner R, van Leeuwen H, van Ham R, Fito L, Guignier L, Sevilla M, Ellul P, Ganko E, Kapur A, Reclus E, de Geus B, van de Geest H, Te Lintel Hekkert B, van Haarst J, Smits L, Koops A, Sanchez-Perez G, van Heusden A, Visser R, Quan Z, Min J, Liao L, Wang X, Wang G, Yue Z, Yang X, Xu N, Schranz E, Smets E, Vos R, Rauwerda J, Ursem R, Schuit C, Kerns M, vanden Berg J, Vriezen W, Janssen A, Datema E, Jahrman T, Moquet F, Bonnet J, Peters S (2014) Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. Plant J 80:136–148

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 Genome 4:132–144

Bauchet G, Munos S, Sauvage C, Bonnet J, Grivet L, Causse M (2014) Genes involved in floral meristem in tomato exhibit drastically reduced genetic diversity and signature of selection. BMC Plant Biol 14:279. doi:10.1186/s12870-014-0279-2

Beaulieu J, Doerksen T, Clément S, MacKay J, Bousquet J (2014) Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity 113:343–352

Beyene Y, Semagn K, Mugo S, Tarekegne A, Babu B, Meisel B, Sehabiague P, Makumbi D, Magorokosho C, Oikeh S, Gakunga J, Vargas M, Olsen M, Prasanna B, Marianne Crossa J (2015) Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. Crop Sci 55:154–163

Blanc G, Charcosset A, Veyrieras JB, Gallais A, Moreau L (2008) Marker-assisted selection efficiency in multiple connected populations: a simulation study based on the results of a QTL detection experiment in maize. Euphytica 161:71–84

Blanca J, Cañizares J, Cordero L, Pascual L, Diez M, Nuez F (2012) Variation revealed by SNP genotyping and morphology provides insight into the origin of the tomato. PLoS One 7:e48198. doi:10.1371/journal.pone.0048198

Blanca J, Montero-Pau J, Sauvage C, Bauchet G, Illa E, Díez MJ, Francis D, Causse M, van der Knaap E, Cañizares J (2015) Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. BMC Genomics 16:257. doi:10.1186/s12864-015-1444-1

Calus M, De Roos A, Veerkamp R (2008) Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553–561

Causse M, Saliba-Colombani V, Lecomte L, Duffé P, Rousselle P, Buret M (2002) QTL analysis of fruit quality in fresh market tomato: a few chromosome regions control the variation of sensory and instrumental traits. J Exp Bot 53:2089–2098

Causse M, Duffe P, Gomez MC, Buret M, Damidaux R, Zamir D, Gur A, Chevalier C, Lemaire-Chamley M, Rothan C (2004) A genetic map of candidate genes and QTLs involved in tomato fruit size and composition. J Exp Bot 55:1671–1685

CoGePedia (2015) https://genomevolution.org/wiki/index.php/Sequenced_plant_genomes

Cros D, Denis M, Sánchez L, Cochard B, Flori A, Durand-Gasselin T, Nouy B, Omoré A, Pomiès V, Riou V, Suryana E, Bouvet JM (2014) Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). Theor Appl Genet 128:397–410

Crossa J, Beyene Y, Kassa S, Perez P, Hickey JM, Chen C, de los Campos G, Burgueno J, Windhausen VS, Buckler E, Jannink JL, Lopez Cruz MA, Babu R (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 G3(3):1903–1926

Daetwyler HD, Calus MPL, Pong-Wong R, de los Campos G, Hickey JM (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347–365

El-Dien OG, Ratcliffe B, Klapste J, Chen C, Porth I, El-KAssaby Y (2015) Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC Genomics 1–16(13):370

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

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

Falush D, Stephens M, Pritchard J (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Frary A, Nesbitt T, Grandillo S, Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert K, Tanksley S (2000) fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289:85–88

Fridman E, Pleban T, Zamir D (2000) A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. Proc Natl Acad Sci USA 97:4718–4723

Gianola D, Okut H, Weigel K, Rosa G (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet 12:87. doi:10.1186/1471-2156-12-87

González-Camacho J, de los Campos G, Pérez P, Gianola D, Cairns J, Mahuku G, Raman B, Crossa J (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theor Appl Genet 125:759–771

Gouy M, Rousselle Y, Bastianelli D, Lecomte P, Bonnal L, Roques D, Efile JC, Rocher S, Daugrois L, Toubi L, Nabeneza S, Hervouet C, Telismart H, Denis M, Thong-Chane A, Glaszmann JC, Hoarau JY, Nibouche S, Costet L (2013) Experimental assessment of the accuracy of genomic selection in sugarcane. Theor Appl Genet 126:2575–2586

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397

Habier D, Tetens J, Seefried F, Lichtner P, Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 42:5. doi:10.1186/1297-9686-42-5

Hamilton J, Buell C (2012) Advances in plant genome sequencing. Plant J 70:177–190

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

Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1–12

Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146–160

Hickey JM, Gorjanc G (2012) Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. G3 2:425–427. doi:10.1534/g3.111.001297/-/DC1

Hirakawa H, Shirasawa K, Miyatake K, Nunome T, Negoro S, Ohyama A, Yamaguchi H, Sato S, Isobe S, Tabata S, Fukuoka H (2014) Draft genome sequence of eggplant (*Solanum melongena* L.): the representative *Solanum* species indigenous to the old world. DNA Res 21(6):649–660

Howard R, Carriquiry AL, Beavis WD (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 4:1027–1046

Isidro J, Jannink J, Akdemir D, Poland J, Heslot N, Sorrells M (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet 128:145–158

Jonas E, de Koning D-J (2013) Does genomic selection have a future in plant breeding? Trends Biotechnol 31:497–504

Jones D, Thomas C, Hammondkosack K, Balintkurti P, Jones J (1994) Isolation of the tomato *Cf-9* gene for resistance to *Cladosporium fulvum* by transposon tagging. Science 266:789–793

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723

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:e36674. doi:10.1371/journal.pone.0036674

Lecomte L, Duffé P, Buret M, Servin B, Hospital F, Causse M (2004) Marker-assisted introgression of 5 QTLs controlling fruit quality traits into three tomato lines revealed interactions between QTLs and genetic backgrounds. Theor Appl Genet 109:658–668

Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, Huang Z, Li J, Zhang C, Wang T, Zhang Y, Wang A, Zhang Y, Lin K, Li C, Xiong G, Xue Y, Mazzucato A, Causse M, Fei Z, Giovannoni JJ, Chetelat RT, Zamir D, Stadler T, Li J, Ye Z, Du Y, Huang S (2014) Genomic analyses provide insights into the history of tomato breeding. Nat Genet 46:1220–1226

Lorenz A, Smith K, Jannink J (2012) Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Sci 52:1609–1621

Maltecca C, Kristen ParkerKL, Cassady JP (2012) Application of multiple shrinkage methods to genomic predictions. J Anim Sci 90:1777–1787

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Martin G, Brommonschenkel S, Chunwongse J, Frary A, Ganal M, Spivey R, Wu T, Earle E, Tanksley S (1993) Map-based cloning of a protein kinase gene conferring disease resistance in tomato. Science 262:1432–1436

Meuwissen T (2007) Genomic selection: marker assisted selection on a genome wide scale. J Anim Breed Genet 124:321–322

Meuwissen T, Hayes B, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Meuwissen T, Hayes B, Goddard M (2013) Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci 1:221–237

Michael T, VanBuren R (2015) Progress, challenges and the future of crop genomes. Curr Opin Plant Biol 24:71–81

Muños S, Ranc N, Botton E, Bérard A, Rolland S, Duffé P, Carretero Y, Le Paslier M, Delalande C, Bouzayen M, Brunel D, Causse M (2011) Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near *WUSCHEL*. Plant Physiol 156:2244–2254

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

Pascual L, Desplat N, Huang BE, Desgroux A, Bruguier L, Bouchet J-P, Le QH, Chauchard B, Verschave P, Causse M (2014) Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect

causal variants in the resequencing era. Plant Biotechnol J. doi:10.1111/pbi.12282

Pascual L, Albert E, Sauvage C, Duangjit J, Bouchet J, Bitton F, Desplat N, Brunel D, Paslier M, Ranc N, Bruguier L, Chauchard B, Verschave P, Causse M (2016) Dissecting quantitative trait variation in the resequencing era: complementarity of bi-parental, multi-parental and association panels. Plant Sci 242:120–130

Pérez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198(2):483–495

Price AH (2006) Believe it or not, QTLs are accurate! Trends Plant Sci 11:213–216

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, de Bakker P, Daly M, Sham P (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81:559–575

Qin C et al (2014) Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. Proc Natl Acad Sci USA 111:5135–5140

Ranc N, Munos S, Xu J, LePaslier MC, Chauveau A, Bounon R, Rolland S, Bouchet JP, Brunel D, Causse M (2012) Genome-wide association mapping in tomato (*Solanum lycopersicum*) is possible using genome admixture of *Solanum lycopersicum* var. *cerasiforme*. G3 2:853–864

R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. ISBN: 3-900051-07-0, http://www.R-project.org

Resende JM, Muñoz P, Resende M, Garrick D, Fernando R, Davis J, Jokela E, Martin T, Peter G, Kirst M (2012) Accuracy of genomic selection methods in a standard data set of Loblolly Pine (*Pinus taeda* L.). Genetics 190:1503–1510

Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodríguez V, Moreno-Gonzalez J, Melchinger A, Bauer E, Schoen C, Meyer N, Giauffret C, Bauland C, Jamin P, Laborde J, Monod H, Flament P, Charcosset A, Moreau L (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). Genetics 192:715–728

Ruggieri V, Francese G, Sacco A, D'Alessandro A, Rigano MM, Parisi M, Milone M, Cardi T, Mennella G, Barone A (2014) An association mapping approach to identify favourable alleles for tomato fruit quality breeding. BMC Plant Biol 14:337. doi:10.1186/s12870-014-0337-9

Sacco A, Ruggieri V, Parisi M, Festa G, Rigano MM, Picarella ME, Mazzucato A, Barone A (2015) Exploring a tomato landraces collection for fruit-related traits by the aid of a high-throughput genomic platform. PLoS One. doi:10.1371/journal.pone.0137139

Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki K, Egholm M, Knight J, Bogden R, Li C, Shuang Y, Xu X, Pan S, Cheng S, Liu X, Ren Y, Wang J, Albiero A, Dal Pero F, Todesco S, Van Eck J, Buels RM, Bombarely A, Gosselin JR, Huang M, Leto JA, Menda N, Strickler S, Mao L, Gao S, Tecle IY, York T, Zheng Y, Vrebalov JT, Lee J, Zhong S, Mueller LA, Stiekema WJ, Ribeca P, Alioto T, Yang W, Huang S, Du Y, Zhang Z, Gao J, Guo Y, Wang X, Li Y, He J, Li C, Cheng Z, Zuo J, Ren J, Zhao J, Yan L, Jiang H, Wang B, Li H, Li Z, Fu F, Chen B, Han B, Feng Q, Fan D, Wang Y, Ling H, Xue Y, Ware D, McCombie WR, Lippman ZB, Chia JM, Jiang K, Pasternak S, Gelley L, Kramer M, Anderson LK, Chang SB, Royer SM, Shearer LA, Stack SM, Rose JKC, Xu Y, Eannetta N, Matas AJ, McQuinn R, Tanksley SD, Camara F, Guigo R, Rombauts S, Fawcett J, Van de Peer Y, Zamir D, Liang C, Spannagl M, Gundlach H, Bruggmann R et al (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641

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, Francis DM (2012a) Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. PLoS One 7:e40563. doi:10.1371/journal.pone.0040563

Sim SC, Van DeynzeA, Stoffel K, Douches DS, Zarka D, Ganal MW, Chetelat RT, Hutton SF, Scott JW, Gardner RG, Panthee DR, Mutschler M, Myers JR, Francis DM (2012b) High-density SNP genotyping of tomato (*Solanum lycopersicum* L.) reveals patterns of genetic variation due to breeding. PLoS One 7:e45520. doi:10.1371/journal.pone.0045520

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen TH (2008) Genomic selection using different marker types and densities. J Anim Sci 86:2447–2454

Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redona E, Atlin G, Jannink JL, McCouch S (2015) Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet 11:e1004982

Storlie E, Charmet G (2013) Genomic selection accuracy using historical data generated in a wheat breeding program. Plant Genome. doi:10.3835/plantgenome2013.01.0001

Tanksley SD, Fulton TM (2007) Dissecting quantitative trait variation-examples from the tomato. Euphytica 154:365–370

Tanksley S, Ganal M, Prince J, de-Vicente M, Bonierbale M, Broun P, Fulton T, Giovannoni J, Grandillo S, Martin G, Messeguer R, Miller J, Miller L, Paterson A, Pineda O, Roder M, Wing R, Wu W, Young N (1992) High density molecular linkage maps of the tomato and potato genomes. Genetics 132:1141–1160

Wang Y, Tang X, Cheng Z, Mueller L, Giovannoni J, Tanksley S (2006) Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome. Genetics 172:2529–2540

Wang Y, Mette M, Miedaner T, Gottwald M, Wilde P, Reif J, Zhao Y (2014) The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test

years. BMC Genomics 15:556. doi:10.1186/1471-2164-15-556

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

Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink JL, Sorrells ME, Raman B, Cairns JE, Tarekegne A, Semagn K, Beyene Y, Grudloyma P, Technow F, Riedelsheimer C, Melchinger AE (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 2:1427–1436. doi:10.1534/g3.112.003699

Würschum T, Reif J, Kraft T, Janssen G, Zhao Y (2013) Genomic selection in sugar beet breeding populations. BMC Genet 14:1–8. doi:10.1186/1471-2156-14-85

Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, Orjeda G, Guzman F, Torres M, Lozano R, Ponce O, Martinez D, De la Cruz G, Chakrabarti SK, Patil VU, Skryabin KG, Kuznetsov BB, Ravin NV, Kolganova TV, Beletsky AV, Mardanov AV, Di Genova A, Bolser DM, Martin DMA, Li G, Yang Y, Kuang H, Hu Q, Xiong X, Bishop GJ, Sagredo B, Mejia N, Zagorski W, Gromadka R, Gawor J, Szczesny P, Huang S, Zhang Z, Liang C, He J, Li Y, He Y, Xu J, Zhang Y, Xie B, Du Y, Qu D, Bonierbale M, Ghislain M, del Rosario HerreraM, Giuliano G, Pietrella M, Perrotta G, Facella P, O'Brien K, Feingold SE, Barreiro LE, Massa GA, Diambra L, Whitty BR, Vaillancourt B, Lin H, Massa A, Geoffroy M, Lundback S, DellaPenna D, Buell CR, Sharma SK, Marshall DF, Waugh R, Bryan GJ, Destefanis M, Nagy I, Milbourne D, Thomson SJ, Fiers M, Jacobs JME, Nielsen KL, Sonderkaer M, Iovene M, Torres GA, Jiang J, Veilleux RE, Bachem CWB, de Boer J, Borm T, Kloosterman B, van Eck H, Datema E, Hekkert BTL, Goverse A, van Ham RCHJ, Visser RGF, Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189–195. doi:10.1038/nature10158

Xu Y, Lu Y, Xie C, Gao S, Wan J, Prasanna B (2012) Whole-genome strategies for marker-assisted plant breeding. Mol Breed 29:833–854

Xu J, Ranc N, Munos S, Rolland S, Bouchet JP, Desplat N, Le Paslier MC, Liang Y, Brunel D, Causse M (2013) Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. Theor Appl Genet 126:567–581

Zhang X, Pérez-Rodríguez P, Semagn K, Beyene Y, Babu R, López-Cruz M, Vicente F, Olsen M, Buckler E, Jannink J, Prasanna B, Crossa J (2015) Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. Heredity 114:291–299

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. Plant Genome 1:5–20