Heredity (2017) 118, 202–209

www.nature.com/hdy

ORIGINAL ARTICLE

# Efficiency of genomic selection for breeding population design and phenotype prediction in tomato

E Yamamoto[1], H Matsunaga[1], A Onogi[2], A Ohyama[3], K Miyatake[1], H Yamaguchi[1], T Nunome[1], H Iwata[2] and H Fukuoka[1]

Genomic selection (GS), which uses estimated genetic potential based on genome-wide genotype data for a breeding selection, is now widely accepted as an efficient method to improve genetically complex traits. We assessed the potential of GS for increasing soluble solids content and total fruit weight of tomato. A collection of big-fruited $F_1$ varieties was used to construct the GS models, and the progeny from crosses was used to validate the models. The present study includes two experiments: a prediction of a parental combination that generates superior progeny and the prediction of progeny phenotypes. The GS models successfully predicted a better parent even if the phenotypic value did not vary substantially between candidates. The GS models also predicted phenotypes of progeny, although their efficiency varied depending on the parental cross combinations and the selected traits. Although further analyses are required to apply GS in an actual breeding situation, our results indicated that GS is a promising strategy for future tomato breeding design.
*Heredity* (2017) **118,** 202–209; doi:10.1038/hdy.2016.84; published online 14 September 2016

## INTRODUCTION

Genomic selection (GS) is now widely accepted as an efficient method for improving genetically complex traits (Desta and Ortiz, 2014). In GS, a training population, which has been phenotyped and genotyped, is used to construct a model that predicts genetic potential (genomic estimated breeding value, GEBV) of unphenotyped individuals by using genome-wide genotype data (Meuwissen *et al.*, 2001). In simulations, GS provided superior efficiency in terms of genetic gain per year and total cost per genetic gain by saving time, cost and the effort required for phenotypic observations (Bernardo and Yu, 2007; Heffner *et al.*, 2010, 2011). GS has already been implemented and shown robust success in animal breeding (Hayes *et al.*, 2009). In plant species, several applied studies have used GS in maize and wheat. In maize, Massman *et al.* (2013) compared the genetic gain in grain yield and stover-quality traits between GS and conventional marker-assisted recurrent selection, and showed that the genetic gains were larger with GS. Combs and Bernardo (2013) performed five cycles of GS and observed that the realized genetic gains in maize grain yield generally agreed with the predicted level, although the gains after the first cycle were unstable. By using GS, Beyene *et al.* (2015) achieved higher genetic gain in maize grain yield compared with conventional pedigree breeding. In wheat, Rutkoski *et al.* (2015b) reported that GS showed an almost equivalent genetic gain in disease resistance on a per-unit-time basis compared with phenotypic selection. Bassi *et al.* (2016) have reviewed additional studies of GS in plants.

GS efficiency is affected by the genetic architecture of the target trait, marker density, the statistical method for model construction and training population composition. Because GS models use linkage disequilibrium (LD) between a marker and quantitative trait loci to estimate the marker effect, high-density markers are preferable (Meuwissen and Goddard, 2010). However, this may not be cost-effective when the breeding population size is large and the economic benefit per selection is low. Habier *et al.* (2009) indicated that the loss of predictability in low-density markers was small when the markers were evenly spaced; subsequent studies have confirmed this in various crops and traits (Heffner *et al.*, 2011; Spindel *et al.*, 2015). On the other hand, Cleveland *et al.* (2010) reported that markers selected on the basis of an additive effect size showed higher predictability compared with evenly spaced markers when the target trait was oligogenic. In practical GS, a statistical method should be selected based on empirically estimated accuracy, which is usually calculated through cross-validation. Genomic best linear unbiased prediction (GBLUP; VanRaden, 2008) is a popular and computationally feasible method that has performed well in many case studies (de los Campos *et al.*, 2013). However, Habier *et al.* (2007, 2013) demonstrated that Bayesian methods are more suitable when a training population and a breeding population are genetically distant. Onogi *et al.* (2015) used simulated data to indicate that the choice of statistical method is especially important when the size of the training population is small, and when the target trait includes nonadditive genetic factors, non-linear methods are more advantageous than linear methods.

[1]Vegetable Breeding and Genome Division, Institute of Vegetable and Tea Science, National Agriculture and Food Research Organization, Tsu, Mie, Japan; [2]Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo-Ku, Tokyo, Japan and [3]Vegetable Production Technology Division, Institute of Vegetable and Tea Science, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, Japan
Correspondence: Dr E Yamamoto, Vegetable Breeding and Genome Division, Institute of Vegetable and Tea Science, National Agriculture and Food Research Organization, 360 Kusawa, Ano, Tsu, Mie 514-2392, Japan.
E-mail: yame@affrc.go.jp
Received 26 April 2016; revised 21 July 2016; accepted 28 July 2016; published online 14 September 2016

Nevertheless, in most empirical studies, differences between methods were often small (de los Campos *et al.*, 2013). The design of the training population may be the most important factor affecting GS efficiency. A training population that is genetically close to the breeding population, ideally full-sibling, increases the predictability of GS models and the selection efficiency (Habier *et al.*, 2007). Riedelsheimer *et al.* (2013) indicated that half-siblings provided more GS efficiency than more distantly related populations. In general, development of a new training population requires additional cost, time and effort. Therefore, historical populations such as breeding lines are more suitable for GS implementation in an actual breeding setting. Rutkoski *et al.* (2015a) empirically demonstrated the utility of historical wheat lines as a training population, although a training population of closer relatives increased predictability.

GS efficiency cannot be reliably predicted in advance without any empirical data (Bassi *et al.*, 2016). Therefore, GS efficiency must be carefully assessed in a practical experiment before it is applied in an actual breeding situation. Tomato (*Solanum lycopersicum*) is one of the world's most important crops and represents the largest production among all vegetable crops (164 million tons in 2013, FAO statistics; http://faostat.fao.org/). In 2012, the tomato genome was released (Tomato Genome Consortium, 2012). In addition, several high-density single-nucleotide polymorphism (SNP) marker sets have been developed for the species (Sim *et al.*, 2012; Shirasawa *et al.*, 2013; Yamamoto *et al.*, 2016). These developments facilitated the use of DNA markers for tomato breeding. In tomato, Yamamoto *et al.* (2016) assessed the potential of GS for yield-related traits in big-fruited varieties. Duangjit *et al.* (2016) analyzed the potential of GS for fruit quality-related traits by using a collection of tomato accessions that consisted of cultivated, cherry and wild-related tomato. These studies analyzed prediction accuracy of the GS models using cross-validation.

In the present study, we assessed the potential of GS to improve soluble solids content and total fruit weight in tomato. We used 96 big-fruited $F_1$ tomato varieties as a training population, and evaluated the predictability of the GS models in their progeny populations. Breeding strategies of tomatoes vary by breeding objective, breeding sectors and breeder preference. However, because crossing among elite lines or varieties is a common strategy in modern plant breeding programs, our choice of plant materials was suitable for our purposes. We had a small population size with few phenotypic observations because it was not feasible to observe phenotypes in hundreds of hydroponically grown lines. In addition, we used a low-density marker set of a few hundred markers for the GS model construction, which is preferable for an actual breeding program. Our study is the first to empirically assess the efficiency of GS in tomato on a realistic experimental scale.

## MATERIALS AND METHODS

### Plant materials and growth conditions

We used a collection of 96 big-fruited $F_1$ tomato varieties that is detailed in Yamamoto *et al.* (2016). One plant of each variety was grown in 2011 and 2014. In addition, we developed four progeny populations derived from selected crosses between these varieties: SL10 × SL65 ($n = 21$), SL10 × SL75 ($n = 23$), SL65 × SL88 ($n = 23$) and SL75 × SL88 ($n = 23$). Progeny populations were grown in 2015. All plants were grown hydroponically with a high-wire system in a greenhouse at the National Agriculture and Food Research Organization, Institute of Vegetable and Tea Science, Tsu, Japan. Plant growth started in the first week of February and finished in July. Tomato seeds were sown on a granular soil (Nippi Engei Baido 1; Nihon Hiryo Co., Tokyo, Japan), and 20 days later, seedlings were transplanted onto rock wool slabs. A mixture of Otsuka-A nutrient solution and Otsuka-5 nutrient solution (Otsuka AgriTechno, Tokyo, Japan) was provided to the plants. The electrical conductivity level was adjusted to 0.80, 1.20, 1.60, 2.00 and 2.40 dS $m^{-1}$ in accordance with plant growth. The plants received 300 ml of water each time they were watered (six times a day, in accordance with plant growth and climate conditions). To promote fruiting, Tomato-tone (including 0.15% 4-chlorophenoxy acetic acid; Ishihara Biosciences, Tokyo, Japan) was diluted 100-fold and sprayed on each truss when the second to fifth flowers were open. Each truss was limited to six flowers to promote uniform fruit size. The plants were pinched above the fourth truss. The plants were phenotyped for soluble solids content and total fruit weight per plant (Supplementary Table S1). Soluble solids content, which indicates fruit sugar content, was measured in degrees Brix and obtained from the average of four fruits per plant. The phenotypic values were averaged per variety over the 2 years.

### Genotyping

Total genomic DNA was isolated from the leaves of a single plant from each variety using a DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). The plant material was genotyped for 337 SNP markers (Supplementary Table S2). These markers were selected from 16 782 previously developed markers (Yamamoto *et al.*, 2016) based on the following criteria: (1) the selected markers were polymorphic between the four parental varieties (i.e., SL10, SL65, SL75 and SL88), and (2) the markers were distributed throughout the genome and as evenly spaced as possible. SNP genotyping was performed using a 96.96 Dynamic Array on the BioMark platform (Fluidigm, South San Francisco, CA, USA), according to the manufacturer's protocol. The data were analyzed using Fluidigm SNP genotyping analysis version 4.1.2 to obtain genotype calls. Samples were classified into three genotypes, based on SNP type normalization, using a *k*-means clustering algorithm. The SNP genotype data were assessed with BEAGLE version 3.3.2 (Browning and Browning, 2007) to impute missing data and estimate the most likely linkage phases of the individuals (Supplementary Table S3).

### Heritability, population structure and LD

To calculate narrow-sense heritability $\left(\hat{h}^2\right)$ of the traits, we estimated the genetic and error variance components ($\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$) with a restricted maximum likelihood approach (Kang *et al.*, 2008), because the replication of phenotypic observations was insufficient to calculate these components using the standard method. For the restricted maximum likelihood approach, the phenotypic variance **V** was defined by the following equation:

$$\mathbf{V} = \mathbf{A}\hat{\sigma}_g^2 + \mathbf{I}\hat{\sigma}_e^2 \quad (1)$$

where **A** is a genetic relationship matrix between individuals and **I** is an $n \times n$ identity matrix. In the genetic relationship matrix, the element $A_{jk}$ was defined as

$$A_{jk} = \sum_{i=1}^{m}(x_{ij} - 2p_i)(x_{ik} - 2p_i)/2p_i(1 - p_i) \quad (2)$$

where $x_{ij}$ (coded as 0, 1, 2) is the number of copies of the reference allele for the *i*th SNP of the *j*th individual, $p_i$ is the minor allele frequency for the *i*th SNP, and *m* is the total number of markers. The restricted maximum likelihood solution of Equation (1) was obtained by using the function 'mixed.solve' in the R package *rrBLUP* version 4.4 (Endelman, 2011). The estimated variance components were used to calculate heritability with the following equation:

$$\hat{h}^2 = \hat{\sigma}_g^2/(\hat{\sigma}_g^2 + \hat{\sigma}_e^2) \quad (3)$$

To explore genetic population structure in the varieties and the progeny population, we conducted a principal component analysis (PCA) with the R function 'prcomp'. LD between pairs of markers was evaluated using squared Pearson's correlation coefficient ($r^2$) calculated with the function 'LD' in the R package *genetics* version 1.3.8.1. The relationship between the degree of LD and the linkage map distance was analyzed. The linkage map positions of SNP markers were estimated from their physical positions via local polynomial regression, using the relationships reported in Shirasawa *et al.* (2010). The local polynomial regression was conducted with the R function 'loess' with the default parameter settings. When the estimated distance between two successive markers became negative, it was replaced with $1.0^{-6}$. The relation between LD values ($r^2$) and the linkage map distance between the corresponding markers was modeled by fitting local polynomials with the function 'locpoly' in the R package *KernSmooth* version 2.23.

### GS models

To construct GS models based on the genotype and the phenotype data from the varieties, we tested six statistical methods. GBLUP (VanRaden, 2008), Bayesian Lasso (BL; Park and Casella, 2008), weighted Bayesian shrinkage regression (wBSR; Hayashi and Iwata, 2010) and Bayes C (Habier *et al.*, 2011) are linear methods, whereas reproducing Kernel Hilbert spaces regression (RKHS; Gianola and van Kaam, 2008) and random forest (RF; Breiman, 2001) are nonlinear methods. GBLUP and RKHS assume the following model:

$$y = 1_n\mu + Zg + \varepsilon \tag{4}$$

where $y$ is a vector of phenotypes, $1_n$ is a vector of ones, $\mu$ is a mean, $Z$ is a design matrix allocating records to genetic values and $g$ is a vector of the additive genetic effects of markers. In GBLUP, the variance of $g$ is $\mathbf{G}\sigma_g^2$, where $\mathbf{G}$ is the realized additive genetic relationship matrix calculated using genotypes of SNP markers, and $\sigma_g^2$ is the genetic variance for this model. In RKHS, the realized genetic relationship matrix $\mathbf{G}$ was replaced with the Gaussian kernel matrix. GBLUP and RKHS were performed using the function 'kinship.BLUP' in the R package *rrBLUP* version 4.4 (Endelman, 2011), with K.method equal to 'RR' and 'GAUSS' for GBULP and RKHS, respectively.

In the present study, BL, wBSR and Bayes C assumed the following linear regression model:

$$y_i = \sum_{p=1}^{P} \gamma_p x_{ip}\beta_p + \varepsilon_i \tag{5}$$

where $y_i$ is the phenotypic value, $P$ is the number of markers, $\gamma_p$ is the indicator variable that takes 0 or 1, $x_{ip}$ is the genotype of marker $p$, $\beta_p$ is the effect of marker $p$ and $\varepsilon_i$ is the residual. The indicator variable $\gamma_p$ is fixed at 1 except for wBSR. $\varepsilon_i$ is assumed to follow a normal distribution with mean = 0 and variance $= 1/\tau_0^2$. Distributions of priors for certain variables differed depending on the method. In BL, $\beta_p$ was assumed to follow

$$\beta_p \sim N(0, 1/\tau_0^2\tau_p^2) \tag{6}$$

where $\tau_p^2$ determines the magnitude of shrinkage for $\beta_p$, and $1/\tau_0^2$ is the residual variance. $\tau_p^2$ was assumed to follow

$$\tau_p^2 \sim Inv - G\left(1, \lambda^2/2\right) \tag{7}$$

where *Inv–G* indicates the inverse gamma distribution and $\lambda^2$ is a regularization parameter that defines the distribution of $\tau_p^2$, and assumed to follow

$$\lambda^2 \sim G(\varphi, \omega) \tag{8}$$

In BL, $\varphi$ and $\omega$ are the hyperparameters. In the present study, $\varphi$ is fixed at 1, whereas five values of $\omega$ were tested: 0.001, 0.01, 0.1, 1 and 5. In wBSR, $\gamma_p$ assumed to follow

$$\gamma_p \sim \text{Bernoulli}(\pi) \tag{9}$$

If $\gamma_p = 1$, $\beta_p$ was assumed to follow

$$\beta_p \sim N(0, \sigma_p^2) \tag{10}$$

then the prior was

$$\sigma_p^2 \sim \chi^{-2}\left(\nu, S^2\right) \tag{11}$$

where $\chi^{-2}$ indicates a scaled inverse chi-square distribution, $\nu$ is the degree-of-freedom and $S^2$ is a scale parameter. In wBSR, $\nu$, $S^2$ and $\pi$ are the hyperparameters. In the present study, $\nu$ and $S^2$ are fixed at 1, whereas five values of $\pi$ were tested: 0.01, 0.1, 0.2, 0.5 and 1. In Bayes C, $\beta_p$ was assumed to follow

$$\beta_p = \begin{cases} 0 & \text{if } p_p = 0 \\ \sim N(0, \sigma^2) & \text{if } p_p = 1 \end{cases} \tag{12}$$

where $p_p$ is the indicator variable that determines whether the marker effect is included in the regression model ($p_p = 1$) or not ($p_p = 0$), with the prior distribution

$$p_p \sim \text{Bernoulli}(\pi) \tag{13}$$

Unlike BL and wBSR, all SNP effects have a common variance $\sigma^2$ in Bayes C. A prior distribution of $\sigma^2$ was as follows:

$$\sigma^2 \sim \chi^{-2}\left(\nu, S^2\right) \tag{14}$$

The sets of hyperparameters tested were the same as in wBSR. In BL, wBSR and Bayes C, a nested fivefold cross-validation was performed to determine the optimal hyperparameter value that showed the least mean square. We used VIGoR, which is based on variational Bayesian algorithms (Onogi and Iwata, 2016). RF is an ensemble learning method that uses a combination of decision trees, each generated from a subset of SNP markers selected by bootstrap. RF was performed using the function 'randomForest' in the R package *randomForest* version 4.6 with default parameter settings, namely, the number of variables tried at each split $m_{try} = p/3$, number of trees = 500 and minimum node size = 5. We performed 10-fold cross-validation to evaluate predictive accuracy of the GS models. We conducted 100 replicates for each trait and the same fold was used for each statistical method. The predictive accuracy was measured as Pearson's correlation coefficient between the predicted and actual phenotypic values using the R function 'cor.test'. The accuracies among the predictive methods were compared using Tukey's test with the R functions 'aov' and 'TukeyHSD' ($P < 0.05$).

### Simulation of trait segregation

We used the simulation method of Yamamoto *et al.* (2016). The tomato genome in the simulation was represented by the linkage map information from Shirasawa *et al.* (2010), with a bin size of 0.1 cM. The number of recombinations on each chromosome was determined using a random variable drawn from a Poisson distribution. For each chromosome, the lambda parameter of the Poisson distribution (i.e., the expected value of the random variable) was set as the length of the linkage map (in Morgans). The position of each recombination in a chromosome was sampled from a uniform distribution. To construct the genotype data of the simulated genome, the genotype of each marker was determined from the haplotype of the nearest bin in the simulated genome. To predict the trait segregation in the four progeny populations, 1000 simulated genomes were produced for each population. All simulation analyses were written and conducted in R (https://www.r-project.org/, Supplementary Method). The GEBVs and the observed phenotypic values among the populations were compared with Tukey's test, using the R functions 'aov' and 'TukeyHSD' ($P < 0.05$).

## RESULTS

### $F_1$ tomato variety characterization

The 96 big-fruited $F_1$ tomato varieties were phenotyped for soluble solids content and total fruit weight (Figure 1a). The estimated trait heritability was 0.626 and 0.248 for soluble solids content and total fruit weight, respectively. Strong population structure can result in unstable predictability of GS models (Riedelsheimer *et al.*, 2013). To analyze genetic population structure in the varieties, we performed a PCA with 337 SNP markers and found no strong population structure (Figure 1b). The extent of LD also affects predictability because GS models are designed to capture quantitative trait loci effects by using LD between markers and quantitative trait loci (Habier *et al.*, 2007, 2013). In the varieties, the estimated LD size based on 337 SNP markers was 16 cM (Figure 1c).

### GS model construction

The genotype and the phenotype data of the $F_1$ varieties were used to construct GS models (Figures 1a and b). We tested six statistical methods and evaluated predictability by using 10-fold cross-validation (Figure 2). The predictability was higher for soluble solids content than total fruit weight, which corresponded well to the higher estimated heritability for soluble solids content. For soluble solids content, GBLUP and BL showed significantly higher predictability ($P < 0.05$) compared with other methods for this trait, whereas RF showed the lowest predictability. For total fruit weight, nonlinear methods (i.e., RKHS and RF) showed significantly higher predictability ($P < 0.05$) than linear methods, suggesting that nonadditive genetic factors contribute to the trait.

**Trait segregation prediction**

Parental selection for a breeding population is a critical step in a breeding design. Iwata *et al.* (2013) proposed a method that predicted trait segregation by using GS models and a computationally simulated breeding population, and demonstrated its efficiency with empirical data from Japanese pear. To validate the efficiency of this method in

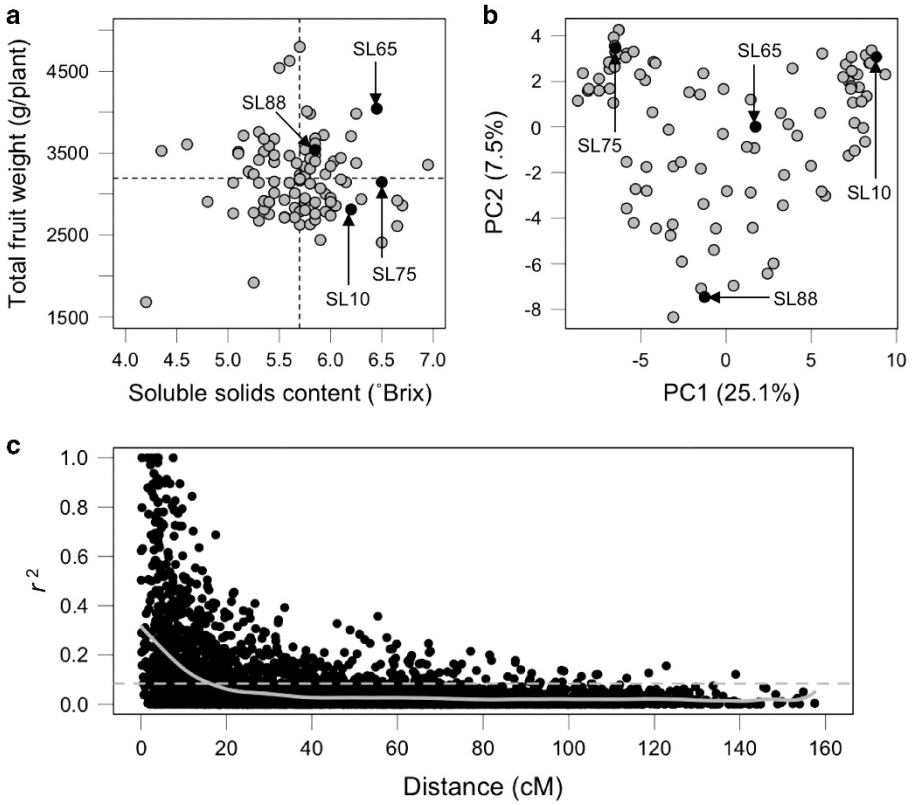


**Figure 1** The 96 big-fruited $F_1$ tomato varieties used in the present study. (**a**) Phenotypic distribution for soluble solids content and total fruit weight. The black dots indicate the varieties used to develop progeny populations. The values are means from 2 years of phenotypic observation. Dashed lines indicate the mean value for each trait. (**b**) The principal component analysis of the varieties based on 337 SNP genotypes. The black dots indicate the varieties used to develop progeny populations. (**c**) Plot of linkage disequilibrium values ($r^2$) against linkage map distance. The horizontal dashed line represents the base line $r^2$ values based on the 95th percentile of the distribution of $r^2$ values between pairs of unlinked markers. The curve shows local polynomial fitting using kernel smoothing regression.

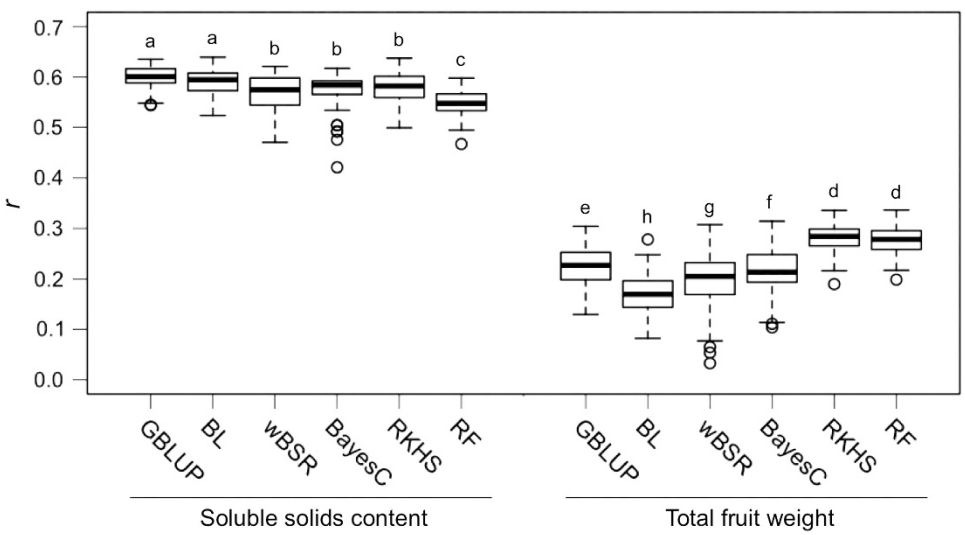


**Figure 2** Result of 10-fold cross-validation for the GS models using the 96 big-fruited $F_1$ tomato varieties as the training population. Boxplot for Pearson's correlation coefficients between phenotypic values and genomic estimated breeding values. GBLUP, genomic best linear unbiased prediction; BL, Bayesian Lasso; wBSR, weighted Bayesian shrinkage regression; RKHS, reproducing Kernel Hilbert spaces regression; RF, random forest. Different lowercase letters indicate significant differences with Tukey's test ($P<0.05$).

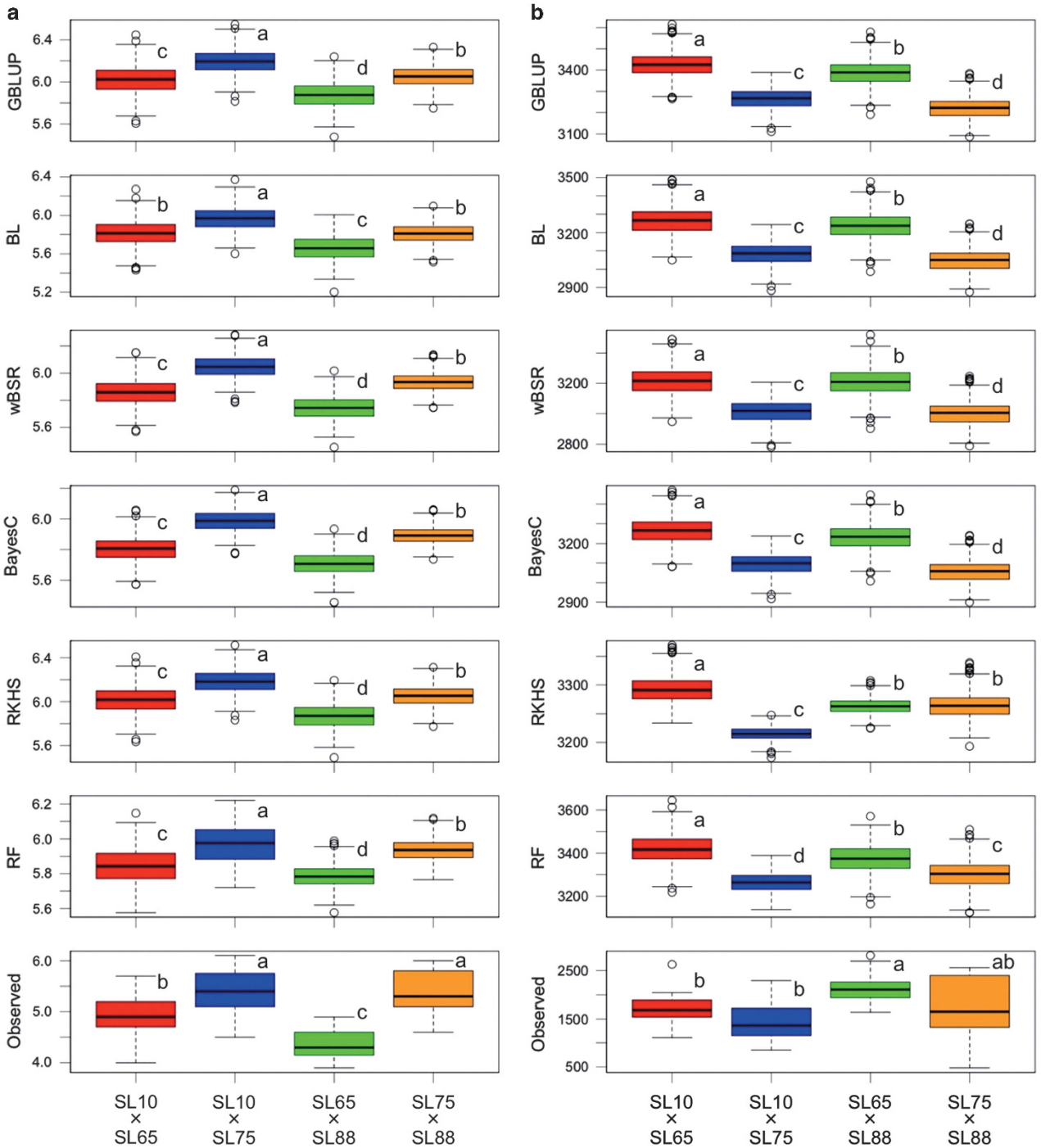


**Figure 3** Comparison between predicted and observed trait segregation in the progeny populations. Boxplots for trait segregation in soluble solids content (**a**) and total fruit weight (**b**). Predicted trait segregations were based on the genomic estimated breeding values (GEBVs) of 1000 simulated genomes. The GEBVs were calculated by using the genomic selection (GS) models based on the 96 big-fruited $F_1$ tomato varieties. The number of individuals for the observed trait segregation in each progeny population was 21, 23, 23 and 23 for SL10×SL65, SL10×SL75, SL65×SL88 and SL75×SL88, respectively. Labels on the *y*-axis of each panel indicate statistical method used for the GS model construction except for 'Observed,' which indicates the observed phenotypic distribution. GBLUP, genomic best linear unbiased prediction; BL, Bayesian Lasso; wBSR, weighted Bayesian shrinkage regression; RKHS, reproducing Kernel Hilbert spaces regression; RF, random forest. Different lowercase letters indicate significant differences with Tukey's test ($P<0.05$).

tomato, we designed four progeny populations to predict trait segregation: SL10×SL65, SL10×SL75, SL65×SL88 and SL75×SL88. The parental varieties were selected using the following criteria: the phenotypic values were more than the average values for at least one trait (Figure 1a), and the selected varieties were genetically distant from each other in the PCA results (Figure 1b). We simulated 1000 individuals for each population and calculated the GEBVs using the GS models (Figure 2). The predicted results were compared with the observed trait segregations ('Observed' in Figure 3). The observed phenotypic values in the progeny population were lower than in the

varieties (the training population) for both traits (Figures 1a and 3). This may reflect differences in growing conditions between the varieties (i.e., 2011 and 2014) and the progeny population (i.e., 2015). Nevertheless, the predicted and observed trait segregations were consistent with respect to the order of mean phenotypic values in the progeny populations, with an exception (see below).

SL65 and SL75 had very similar phenotypic values of soluble solids content (6.45 and 6.50 for SL65 and SL75, respectively; Figure 1a). However, the simulation predicted that SL75 would generate progeny with a significantly higher mean phenotypic value ($P<0.05$) than SL65 would when crossed with SL10 or SL88 (Figure 3a). The observed trait segregation supported this conclusion ('Observed' in Figure 3a). For total fruit weight, linear models (i.e., GBLUP, BL, wBSR and Bayes C) predicted that SL65 × SL88 would generate progeny with a considerably higher phenotypic value compared with SL75 × SL88, whereas nonlinear models (i.e., RKHS and RF) predicted a small or insignificant difference (Figure 3b). The observed trait segregation was similar to the result predicted by the nonlinear models ('Observed' in Figure 3b). The estimated predictability of the GS models in cross-validation was significantly higher for the nonlinear methods than the linear methods (Figure 2). Thus, the total fruit weight result indicated that cross-validation can effectively evaluate GS model predictability. However, for total fruit weight, the predictions were strongly inconsistent with the observations. The mean value of SL10 × SL65 was significantly higher than the other populations in the prediction, whereas it was significantly lower than SL65 × SL88 and not significantly different from other populations in the observations (Figure 3b).

### Progeny phenotype prediction

We investigated whether the GS models were efficient for predicting phenotypes in the progeny populations. All individuals in the four progeny populations were genotyped with the same 337 SNP markers used to construct the GS model. To summarize genetic relationships between the varieties and their progeny, we performed a PCA and found that the progeny were genetically distinct from each other but intermediate between their parents (Figure 4). The GEBV was calculated for all progeny by using the GS models constructed with the original 96 varieties (Figure 2). The correlation coefficients between the GEBVs and the phenotypic values were comparable to the estimated predictability in the cross-validation (see Figure 2 and 'All' in Table 1). This indicated that cross-validation could accurately assess the estimated predictability of GS models. In addition, the correlation coefficients were comparable to the estimated heritability of the traits in the progeny population (0.599 for soluble solids content and 0.443 for total fruit weight). Thus, we confirmed that the GS models could efficiently predict phenotypes ('All' in Table 1 and Figure 5). We surveyed the GS model predictability for each progeny population (Supplementary Figures S1 and S2). The genetic variation within each population was clearly lower than within the training population (i.e., all 96 varieties; Figure 4). Therefore, this was a challenging approach. GS model predictability differed by the parental combination and the trait (Table 1 and Supplementary Figures S1 and S2). For SL10 × SL75, the GS models showed predictability for both traits, whereas they were not efficient for SL65 × SL88. For SL10 × SL65, the GS models for soluble solids content showed high predictability, whereas the predictability for total fruit weight was negative. For SL75 × SL88, the GS models showed predictability for total fruit weight but not for soluble solids content (except for BL). Thus, although the GS models based on the varieties were useful for predicting progeny phenotypes, the efficiency strongly varied depending on the cross combinations and the traits.

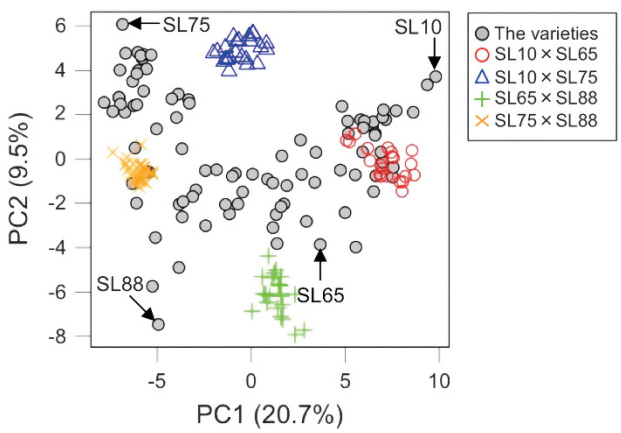


**Figure 4** The principal component analysis of the 96 big-fruited $F_1$ tomato varieties and the four progeny populations, based on 337 SNP markers. The arrows indicate the parental varieties of the progeny populations.

## DISCUSSION

In GS, a full-sibling training population that is genetically close to a breeding population is ideal (Habier *et al.*, 2007; Riedelsheimer *et al.*, 2013; Rutkoski *et al.*, 2015a). However, such a population is seldom available. Developing a new population for GS model construction is not usually feasible, so existing populations, such as breeding lines, are used instead for breeding projects. Therefore, we used 96 big-fruited $F_1$ tomato varieties, which had been characterized in a previous study (Yamamoto *et al.*, 2016), as a training population (Figure 1). It has long been recognized that there is a negative correlation between soluble solids content and total fruit weight in tomato. However, we did not observe a negative correlation in the studied varieties ($r=0.02$). This may be due to low heritability of total fruit weight in the present study (see above), or to breeding selection for both soluble solids content and yield performance in the varieties we used, as suggested by Higashide *et al.* (2012). Yamamoto *et al.* (2016) identified 16 782 DNA markers, 337 of which were used in the present study (Supplementary Table S2). In an actual breeding program, using fewer markers is more cost-effective. For example, the genotyping cost for the 96 varieties in the present study ($n=337$) was about one-quarter of that in the previous study ($n=16\,782$; Yamamoto *et al.*, 2016). Use of fewer markers could lead to a biased evaluation of a population's genetic architecture (Heslot *et al.*, 2013). However, in the present study, the PCA and LD results (Figures 1b and c) were very similar to those of the previous study (Yamamoto *et al.*, 2016). This indicated that the 337 selected markers showed no discernible ascertainment bias. Another difference between the present and the previous study was the cropping season. Tomato growth in the warm season (present study) is less stable compared with that in the cool season (previous study). Nevertheless, the GS models in the present study showed good predictability (Figures 2 and 5 and Table 1).

In previous studies, potential assessment of the GS in tomato breeding was performed based on cross-validation using training data, and demonstrated that could be applicable to breeding of agronomically important traits related to fruit quality (Duangjit *et al.*, 2016) and yield performance (Yamamoto *et al.*, 2016). In the present study, we assessed the potential of GS using the progenies derived from crosses between the varieties in the training

**Table 1 Correlation coefficients between the genomic estimated breeding values (GEBVs) and the phenotypic values in the progeny populations**

| *Trait* | *Methods* | *Population* | | | | |
|---|---|---|---|---|---|---|
| | | *SL10×SL65* | *SL10×SL75* | *SL65×SL88* | *SL75×SL88* | *All* |
| SSC | GBLUP | 0.518* | 0.305 | −0.044 | 0.136 | 0.633*** |
| | BL | 0.567** | 0.341 | 0.060 | 0.295 | 0.639*** |
| | wBSR | 0.518* | 0.308 | −0.091 | 0.041 | 0.674*** |
| | Bayes C | 0.520* | 0.325 | −0.082 | 0.038 | 0.689*** |
| | RKHS | 0.510* | 0.296 | −0.054 | 0.121 | 0.650*** |
| | RF | 0.468* | 0.135 | −0.059 | −0.221 | 0.567*** |
| TFW | GBLUP | −0.218 | 0.524** | 0.100 | 0.282 | 0.223* |
| | BL | −0.290 | 0.519** | 0.104 | 0.261 | 0.235* |
| | wBSR | −0.264 | 0.569** | 0.100 | 0.269 | 0.273** |
| | Bayes C | −0.263 | 0.530** | 0.111 | 0.277 | 0.230* |
| | RKHS | −0.208 | 0.548** | 0.180 | 0.351 | 0.256* |
| | RF | −0.025 | 0.363 | −0.011 | 0.247 | 0.275** |

Abbreviations: SSC, soluble solids content; TFW, total fruit weight. Asterisks indicate significance with Pearson's product–moment correlation test (*$P<0.05$, **$P<0.01$, ***$P<0.005$).

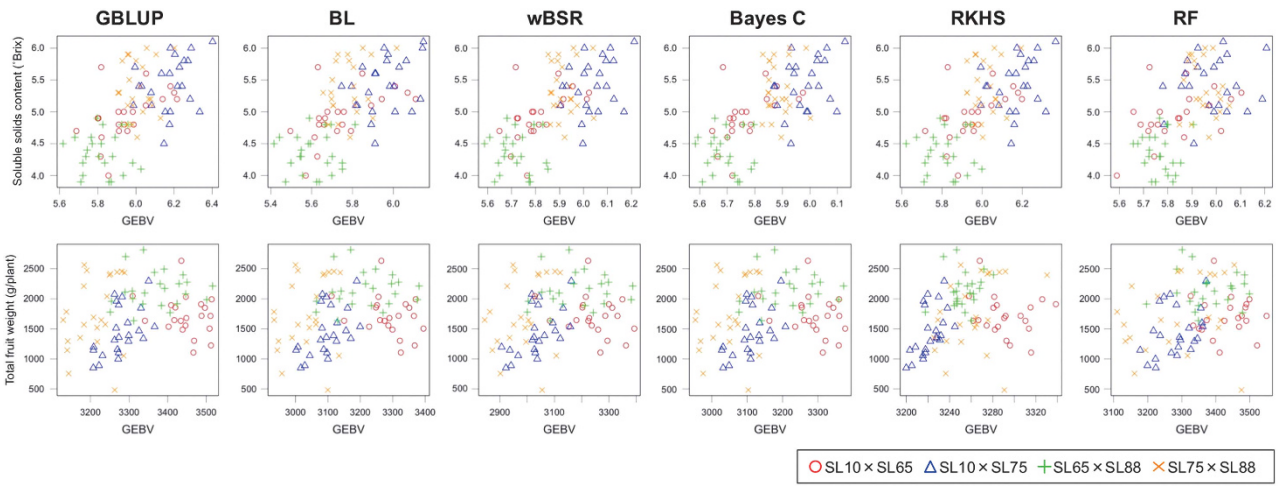


**Figure 5** Comparison between the phenotypic values and the genomic estimated breeding values (GEBV) in the progeny populations. GEBVs were calculated by using genomic selection (GS) models based on the 96 big-fruited $F_1$ tomato varieties. The title of each panel indicates the statistical method used for the GS model construction. GBLUP, genomic best linear unbiased prediction; BL, Bayesian Lasso; wBSR, weighted Bayesian shrinkage regression; RKHS, reproducing Kernel Hilbert spaces regression; RF, random forest.

population, and demonstrated that the GS models could predict phenotypes (Table 1 and Figure 5) and parental combinations that generated superior progeny by using the method of Iwata *et al.* (2013) in tomato (Figure 3). While these results indicated that GS is useful for designing a tomato breeding program, the GS models were not always efficient. For example, the GS models did not accurately predict soluble solids content of SL65 × SL88 (Table 1 and Supplementary Figure S1). For total fruit weight of SL10 × SL65, the correlations between the GEBVs and the phenotypic values were negative (Table 1 and Supplementary Figure S2), perhaps owing to a population-specific genotype-by-environment interaction during the different years of the experiment. If this hypothesis is true, the predicted and actual observations will converge as more individuals are assessed or phenotyped. However, a more logical explanation is that quantitative trait loci segregated differently between the training and the progeny populations, which has been reported in both theoretical and empirical studies (Riedelsheimer *et al.*, 2013; Rutkoski *et al.*, 2015a; Duangjit *et al.*, 2016). Therefore, training populations and the GS models should be updated when the genetic relationship between the training and breeding populations is dramatically changed (Rutkoski *et al.*, 2015a; Bassi *et al.*, 2016). We confirmed that although the GS models were useful, careful attention is required for their use in a long-term selection process.

In the present study, it is especially notable that GS model efficiency was confirmed even though the experiment was conducted at a small scale. In horticultural crops such as tomato, growing a large number of lines requires an enormous facility and prohibitive costs. Our study indicated that GS could reduce the expenditures required for tomato breeding. Although more studies are needed to test GS in actual breeding programs, our results highlight GS as a promising strategy for future tomato breeding.

## DATA ARCHIVING

There are no data to deposit.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## ACKNOWLEDGEMENTS

We thank Mr K Kamiya, Ms S Negoro and Ms N Fukushima for their technical assistance. This work was supported by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Genomics-based Technology for Agricultural Improvement, NGB-1004, 2005 and 2010).

Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J (2016). Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). *Plant Sci* **242**: 23–36.

Bernardo R, Yu J (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Sci* **47**: 1082–1090.

Beyene Y, Semagn K, Mugo S, Tarekegne A, Babu R, Meisel B *et al.* (2015). Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Sci* **55**: 154–163.

Breiman L (2001). Random forests. *Mach Learn* **45**: 5–32.

Browning SR, Browning BL (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**: 1084–1097.

Cleveland MA, Forni S, Deeb N, Maltecca C (2010). Genomic breeding value prediction using three Bayesian methods and application to reduced density marker panels. *BMC Proc* **4**: S6.

Combs E, Bernardo R (2013). Genomewide selection to introgress semidwarf maize germplasm into US corn belt inbreds. *Crop Sci* **53**: 1427–1436.

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**: 327–345.

Desta ZA, Ortiz R (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci* **19**: 592–601.

Duangjit J, Causse M, Sauvage C (2016). Efficiency of genomic selection for tomato fruit quality. *Mol Breed* **36**: 29.

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

Gianola D, van Kaam JB (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* **178**: 2289–2303.

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

Habier D, Fernando RL, Dekkers JCM (2009). Genomic selection using low-density marker panels. *Genetics* **182**: 343–353.

Habier D, Fernando RL, Garrick DJ (2013). Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics* **194**: 597–607.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* **12**: 186.

Hayashi T, Iwata H (2010). EM algorithm for Bayesian estimates of genomic breeding values. *BMC Genetics* **11**: 3.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009). Invited review: genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* **92**: 433–443.

Heffner EL, Lorenz AJ, Jannink JL, Sorrells ME (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop Sci* **50**: 1681–1690.

Heffner EL, Jannink JL, Iwata H, Souza E, Sorrells ME (2011). Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci* **51**: 2597–2606.

Heslot N, Rutkoski J, Poland J, Jannink JL, Sorrells ME (2013). Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. *PLoS One* **8**: e74612.

Higashide T, Yasuba KI, Suzuki K, Nakano A, Ohmori H (2012). Yield of Japanese tomato cultivars has been hampered by a breeding focus on flavor. *HortScience* **47**: 1408–1411.

Iwata H, Hayashi T, Terakami S, Takada N, Saito T, Yamamoto T (2013). Genomic prediction of trait segregation in a progeny population: a case study of Japanese pear (*Pyrus pyrifolia*). *BMC Genet* **14**: 81.

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

Massman JM, Jung HJG, Bernardo R (2013). Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. *Crop Sci* **53**: 58–66.

Meuwissen T, Goddard M (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* **185**: 623–631.

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

Onogi A, Iwata H (2016). VIGoR: variational Bayesian inference for genome-wide regression. *J Open Res Softw* **4**: e11.

Onogi A, Ideta O, Inoshita Y, Ebana K, Yoshioka T, Yamasaki M *et al.* (2015). Exploring the areas of applicability of whole-genome prediction methods for Asian rice (*Oryza sativa* L.). *Theor Appl Genet* **128**: 41–53.

Park T, Casella G (2008). The Bayesian LASSO. *J Am Stat Assoc* **103**: 681–686.

Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE (2013). Genomic predictability of interconnected biparental maize populations. *Genetics* **194**: 493–503.

Rutkoski J, Singh RP, Huerta-Espino J, Bhavani S, Poland J, Jannink JL *et al.* (2015a). Efficient use of historical data for genomic selection: a case study of stem rust resistance in wheat. *Plant Genome* **8**. doi:10.3835/plantgenome2014.09.0046

Rutkoski J, Singh RP, Huerta-Espino J, Bhavani S, Poland J, Jannink JL *et al.* (2015b). Genetic gain from phenotypic and genomic selection for quantitative resistance to stem rust of wheat. *Plant Genome* **8**. doi:10.3835/plantgenome2014.10.0074

Shirasawa K, Asamizu E, Fukuoka H, Ohyama A, Sato S, Nakamura Y *et al.* (2010). An interspecific linkage map of SSR and intronic polymorphism markers in tomato. *Theor Appl Genet* **121**: 731–739.

Shirasawa K, Fukuoka H, Matsunaga H, Kobayashi Y, Kobayashi I, Hirakawa H *et al.* (2013). Genome-wide association studies using single nucleotide polymorphism markers developed by re-sequencing of the genomes of cultivated tomato. *DNA Res* **20**: 593–603.

Sim SC, Durstewitz G, Plieske J, Wieseke R, Ganal MW, Van Deynze A *et al.* (2012). Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. *PLoS One* **7**: e40563.

Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E *et al.* (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.

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**: 635–641.

VanRaden PM (2008). Efficient methods to compute genomic predictions. *J Dairy Sci* **91**: 4414–4423.

Yamamoto E, Matsunaga A, Onogi A, Kajiya-Kanegae H, Minamikawa M, Suzuki A *et al.* (2016). A simulation-based breeding design that uses whole-genome prediction in tomato. *Sci Rep* **6**: 19454.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)