Determination of the optimal number of markers and individuals in a training population necessary for maximum prediction accuracy in F_2 populations by using genomic selection models

L.A. Peixoto, L.L. Bhering and C.D. Cruz

Departamento de Biologia, Universidade Federal de Viçosa, Viçosa, MG, Brasil

Corresponding author: L.A. Peixoto
E-mail: leoazevedop@gmail.com

Genet. Mol. Res. 15 (4): gmr15048874
Received June 10, 2016
Accepted October 13, 2016
Published November 21, 2016
DOI http://dx.doi.org/10.4238/gmr15048874

ABSTRACT. Genomic selection is a useful technique to assist breeders in selecting the best genotypes accurately. Phenotypic selection in the F_2 generation presents with low accuracy as each genotype is represented by one individual; thus, genomic selection can increase selection accuracy at this stage of the breeding program. This study aimed to establish the optimal number of individuals required to compose the training population and to establish the amount of markers necessary to obtain the maximum accuracy by genomic selection methods in F_2 populations. F_2 populations with 1000 individuals were simulated, and six traits were simulated with different heritability values (5, 20, 40, 60, 80 and 99%). Ridge regression best linear unbiased prediction was used in all analyses. Genomic selection models were set by varying the

number of individuals in the training population (2 to 1000 individuals) and markers (2 to 3060 markers). Phenotypic accuracy, genotypic accuracy, genetic variance, residual variance, and heritability were evaluated. Greater the number of individuals in the training population, higher was the accuracy; the values of genotypic and residual variances and heritability were close to the optimum value. Higher the heritability of the trait, higher is the number of markers necessary to obtain maximum accuracy, ranging from 200 for the trait with 5% heritability to 900 for the trait with 99% heritability. Therefore, genomic selection models for prediction in F_2 populations must consist of 200 to 900 markers of major effect on the trait and more than 600 individuals in the training population.

Key words: Genomic prediction; Heritability; Prediction ability; Breeding; Quantitative genetics

INTRODUCTION

Plant selection has been carried out by mankind since early history; however, it intensified at the beginning of the last century, when breeding programs of major crops were established (Allard, 1999). Over the past 100 years, methods and technologies associated with plant selection have shown tremendous progress (Borém and Miranda, 2013). Selection methods have evolved from mass selection (Borém and Miranda, 2013), which involves simple phenotypic selection of individuals; or combined selection (Borém and Miranda, 2013), which takes into account the information between and within families; to reciprocal recurrent selection (Ordas et al., 2012), which includes a gradual increase in the frequency of favorable alleles by repeated selection cycles, without lowering the genetic variability of the population.

With the advent of molecular markers in the 80s, it was possible to improve selection accuracy by means of marker assisted selection (MAS) (He et al., 2014). Although MAS has led to a significant improvement in plant breeding, this technique is only effective for qualitative traits or for traits governed by a few genes, such as in the case of sudden death syndrome in soybean (Lightfoot, 2015), wheat rust (Yaniv et al., 2015), tolerance to salinity (Ashraf et al., 2012), and rice bacterial blight (Pandey et al., 2013). Within 30 years from its advent, techniques involving molecular markers evolved from employing isoenzymes (Dirlewanger et al., 1998), RAPD (Lynch and Milligan, 1994), RFLP (Langer and Maixner, 2004), AFLP (Frascaroli et al., 2013), microsatellites (Soldati et al., 2013) to single nucleotide polymorphism (SNP) (Belaj et al., 2012).

Among the variations found in the genome, SNP variations are the most widely distributed and abundant in the genome. With the development of SNP genotyping platforms and with the improvement of statistical methods, Meuwissen et al. (2001) presented a new approach based on multiple regression using markers as covariates, also known as genomic selection. The objective of genomic selection is to identify possible markers in linkage disequilibrium with the gene regions of interest. Since this pioneering study, several authors have used this technique to predict the genetic value in several plant species, such as corn (Beyene et al., 2015), soybeans (Zhang et al., 2016), wheat (Bassi et al., 2016), forest species (Cros et al., 2015), sugarcane (Gouy et al., 2013), and rice (Spindel et al., 2015).

Although there are several studies on genomic selection, only a few show the effect of different factors affecting prediction accuracy in genomic selection, such as the number of markers and individuals in the training population. Isidro et al. (2015) evaluated five criteria to determine the optimal number of individuals to compose the training population for five traits in wheat. The authors found that greater the number of individuals in the training population, greater was the value of prediction accuracy. However, other effects must also be taken into account, such as plant architecture and population structure. de Los Campos et al. (2013) carried out marker selection based on their importance to the trait as indicated by the results of the GWAS analysis, through meta-analysis. They found that marker selection was effective in humans, since accuracy was 7.5% higher for genomic selection models using 5k SNPs when compared with models using 400k SNPs.

The F_2 generation is one of the most important stages in a plant breeding program because greater genetic variability and heterosis are found at this stage (Tang et al., 1993). Moreover, in the F_2 generation, it is possible to estimate the allelic frequency for each gene by the Mendelian segregation, to evaluate possible deviations from the Hardy-Weinberg equilibrium (Falconer and Mackay, 1996), to estimate genotypic and environmental variance, and consequently, to estimate heritability (Tang et al., 1996). The use of genomic selection in F_2 populations is still restricted to a few studies (Ren et al., 2015) and little is known about how the factors affecting the prediction accuracy can affect the estimate of genetic parameters in F_2 populations. Therefore, the objectives of this study were to establish the optimal number of individuals and markers required to compose the training population in genomic selection models in order to capture maximum genetic variance and consequently achieve greater accuracy in F_2 populations.

MATERIAL AND METHODS

Data simulation

Simulation of F_2 populations was performed using the simulation module of the GENES software (Cruz, 2013). This allowed for information on the genome, parents genotypes, populations of controlled crossings, and quantitative trait data to be generated.

Genome simulation

The simulated genome comprised 15 linkage groups, similar to a diploid species 2n = 2x = 30. Each linkage group was simulated with 200 cM, comprised 200 markers, spaced equidistantly (1 cM), totaling 3060 markers. These markers were assumed to be codominant and biallelic. Furthermore, 4 markers per linkage group were considered responsible for the control of phenotypic expression of quantitative traits, which were randomly inserted into the genome.

Parent simulation

Contrasting homozygous parents were simulated, i.e., parent 1 was coded as carrying allele A_1 (code 2), and parent 2 was coded as carrying the alternative allele A_2 (code 0) for all existing markers. Thus, the cross between parent 1 and 2 generated the F_1 population with all markers being heterozygous and in an approximation stage ($A_1B_1//A_2B_2$).

Simulation of the mapping population

F_2 populations were generated by the selfing of individuals from the F_1 population. Each individual of the F_1 population produced 5000 gametes, and when 2 of these gametes met at random, the first individual of the F_2 population was generated. This process was repeated until the formation of all individuals in each population.

Each gamete was formed based on the following criteria: the allele of the first marker was randomly chosen (A_1 or A_2) to start gamete formation (initialization allele); the allele of the second marker was chosen taking into account the distance to the first gene, i.e, the crossing-over frequencies were counted, and the choice of which allele (B_1 or B_2) to constitute the gamete was based on the probabilities of each gamete $P(A_1B_1)$ and $P(A_2B_2)$, which are parental gametes, and $P(A_1B_2)$ and $P(A_2B_1)$, which are recombinant gametes. This process was carried out for every gene. Null interference, i.e., the crossing-over that occurred between genes A and B, was considered to not interfere with the following crossing-over between genes B and C. This ensured that all gametes formed were different owing to the random choice of the allele in the first gene, and to the probability conditioned to each allele for the next genes. Since all genes were simulated equidistantly at 1 cm, the recombination frequency was 1% for all genes, i.e., the probability of each gamete was: $P(A_1B_1) = P(A_2B_2) = 0.49$, and $P(A_1B_2) = P(A_2B_1) = 0.005$.

The F_2 population simulation was encoded with 0, 1 and 2, in which 0 corresponded to homozygote individuals (A_2A_2), 1 corresponded to heterozygote individuals (A_1A_2), and 2 corresponded to homozygote individuals (A_1A_1) for a given locus.

Simulation of quantitative traits

For the simulation of quantitative traits, a value corresponding to the probability generated by a binomial distribution, of parameter p = q = 0.5, and n = 59 (generating a probability family of 60 elements) was first assigned as the importance of each locus. This value, which denominates the proportion of the genetic variance, explained by each QTL (PGV/QTL), reflects the importance of the locus to the genotypic mean, and consequently to the proportion of genetic variance of the trait explained by each QTL.

Each trait was simulated as being controlled by 60 QTLs distributed equidistantly in the genome (4 QTLs per linkage group). The effect of each QTL was defined as: $A_1A_1 = \mu + a$; $A_1A_2 = \mu + d$; $A_2A_2 = \mu - a$, in which a is the additive effect of each gene, and d is the dominant effect of each gene. Since the value of d was defined as null, the mean degree of dominance (d/a) was equal to zero for all loci, i.e., all loci only presented an additive effect.

The genotypic value (GV) of each individual was defined by the equation:

$$GV = \sum_{i=1}^{n} (PGV / QTL_i \ x \ effect \ of \ QTL_i) \qquad \text{(Equation 1)}$$

The environmental effect was not correlated with the genotypic value, and was estimated following an $N(0,\sigma^2)$distribution. The value of σ^2 is calculated from the heritability of the trait and by the value of genotypic variance (σ_g^2). It was simulated that traits with heritability of 5, 20, 40, 60, 80 and 99%. σ_g^2 was calculated as being the variance of the genotypic value of individuals of the F_2 population. Thus, the phenotypic value was calculated as:

$$PV = u + VG + EA \quad \text{(Equation 2)}$$

In which u = 100 is the mean, and PV is the phenotypic value.

Data analysis

After population generation, the mapping process was carried out, starting with the analysis of individual loci segregation. Chi-square (χ^2) tests were used to check if the markers were segregated per what is expected for an F_2 population. It was verified if Linkage Groups were restored, with chromosome size, distance between markers and order of markers, and it was concluded that it was an F_2 population with the desired simulation properties.

The ridge regression best linear unbiased prediction (RR-BLUP) method of genomic selection used in the analysis, which aims at estimating the effect for each of the covariates (SNP markers) included in the model. RR-BLUP assumes that all SNPs control phenotypic expression of QTL, and assumes homogeneous variance.

Genomic selection models ranging from 2 to 1000 individuals in the training population (TP) and in all the available markers, i.e., 3060 markers, were to verify the optimum number of individuals required to comprise the TP. The validation population was always composed of 200 individuals randomly chosen in the population. Trend graphs were plotted for genetic variance, residual variance, heritability, and genotypic and phenotypic accuracy.

To evaluate the number of markers necessary to capture the entire genetic variance, and consequently achieve greater accuracy, the number of markers in the genomic selection models varied from 2 to 3060. The TP comprised 800 individuals, and the validation population consisted of 200 individuals. Marker selection was carried out and the marker with the lowest effect was deleted from the original matrix and not used in subsequent analysis, i.e., the following model would have one marker less. This marker selection occurred until the model comprised only 2 markers. Trend graphs were plotted for genetic variance, residual variance, heritability and phenotypic accuracy.

Phenotypic and genotypic accuracies were estimated by the Pearson's correlation between the phenotypic value and the genomic estimate breeding value (GEBV), and between the true genotypic value and the GEBV, respectively.

Genetic variance (σ_g^2) was estimated according to Falconer and Mackay (1996):

$$\sigma_g^2 = 2\sum_{i=1}^{n} p * q * \alpha_i^2 \quad \text{(Equation 3)}$$

in which α_i^2 is the allele substitution effect for each locus. Heritability (h^2) was estimated as:

$$h^2 = \frac{\sigma_g^2}{\left(\sigma_g^2 + \sigma^2\right)} \quad \text{(Equation 4)}$$

Software and hardware information

Simulations were carried out using the GENES software (Cruz, 2013), while the analyses of segregation and genomic selection tests were performed using the statistical R software (R Core Team, 2015). The RR-BLUP package was used to run the RR-BLUP model.

Two high-performance computers (Intel Xeon, processor E5-26 12° generation 3.30 GHz, 64 and 96 Gb RAM, 1024 Gb hard drive) were used to perform the genomic selection analysis.

RESULTS

In order to evaluate how the size of the TP and the number of markers influenced the genomic estimate breeding value prediction using RR-BLUP, the size of the TP ranged from 2 to 1000 individuals, and from 2 to 3060 SNP markers.

Marker segregation test

The segregation test was carried out in order to verify whether the simulation process established a population with genetic characteristics of an F_2 population, as proposed by Falconer and Mackay (1996). It was observed that the allelic and genotypic frequencies were close to the expected value for an F_2 population, for all the markers that control the trait (QTL - Table 1) and the other simulated markers (**Table S1**).

The evaluation of the Hardy-Weinberg equilibrium was carried out using the chi-square test (χ^2), which indicated the expected segregation for an F_2 population (Table 1; and **Table S1** shows the P value).

The proportion of the genetic variance of the trait explained by each QTL followed a binomial distribution, as expected in the simulation process (Table 1). The values of the additive effects ranged between the QTL, and they were high in QTL located on the median chromosomes, and low in the QTL located in the first (chromosome 1 over 5) and in the last chromosomes (chromosome 16 over 20) (Table 1).

Evaluation of the training population size

The genotypic variances presented similar behavior regardless of the simulated heritability (Figure 1). Similar results were observed for the residual variance. It was observed that the higher the number of individuals used in the reference population, the greater was the ability of the model to estimate variance components in a manner similar to the simulated parameters. It was also observed that heritability influences the number of individuals to compose the TP used to estimate accurate variance components (genotypic and residual).

The genotypic and phenotypic accuracies showed higher estimates in the TP with a high number of individuals (Figure 2). It was also observed that the higher the heritability of the trait, the higher were the estimates of phenotypic and genotypic accuracies (Figure 2). However, the TP with more than 600 individuals provided a small gain in phenotypic and genotypic accuracies. Genotypic accuracy had lower values than the phenotypic accuracy for all the heritabilities evaluated. However, the higher the value of the simulated heritability, the closer were the estimates of phenotypic and genotypic accuracies (Figure 2).

The estimated heritability showed similar behavior to the genotypic and residual variances, i.e., it presented variable values when few individuals were used in the TP (Figure 2). With the increase in the number of individuals in the TP, more stable values of the estimated heritability were observed, and these values were closer to the values of simulated heritability, except for the traits with 80 and 99% heritability, whose estimated heritability values were lower than the simulated heritability values (Figure 2).

Table 1. Segregation test, minor allele frequency (MAF), P value associated with the χ^2 test of the evaluation of the Hardy-Weinberg equilibrium (HWE P value), and effect of markers associated with the quantitative trait loci (QTL).

QTL	AA	Aa	aa	Total	maf	HWE P value	PGV/QTL	AE (u+a)
M42	248	503	249	1000	0.4995	0.849491	0.00000000	0.000000
M83	220	525	255	1000	0.4825	0.104833	0.00000000	0.000000
M123	249	488	263	1000	0.4930	0.451512	0.00000000	0.000000
M165	254	502	244	1000	0.4950	0.896830	0.00000000	0.000000
M246	253	478	269	1000	0.4920	0.166462	0.00000000	0.000000
M287	275	480	245	1000	0.4850	0.215878	0.00000000	0.000000
M327	290	487	223	1000	0.4665	0.494414	0.00000000	0.000000
M369	255	506	239	1000	0.4920	0.698262	0.00000000	0.000001
M450	234	491	275	1000	0.4795	0.605211	0.00000000	0.000004
M491	248	499	253	1000	0.4975	0.950199	0.00000002	0.000022
M531	245	499	256	1000	0.4945	0.952612	0.00000011	0.000109
M573	231	507	262	1000	0.4845	0.635811	0.00000049	0.000486
M654	255	495	250	1000	0.4975	0.752424	0.00000194	0.001942
M695	238	503	259	1000	0.4895	0.838532	0.00000702	0.007021
M735	257	502	241	1000	0.4920	0.892912	0.00002310	0.023069
M777	255	502	243	1000	0.4940	0.895725	0.00006920	0.069208
M858	234	529	237	1000	0.4985	0.066591	0.00019000	0.190321
M899	247	520	233	1000	0.4930	0.203601	0.00048100	0.481401
M939	263	489	248	1000	0.4925	0.490986	0.00112300	1.123269
M981	253	509	238	1000	0.4925	0.564308	0.00242400	2.423897
M1062	251	490	259	1000	0.4960	0.528386	0.00484800	4.847794
M1103	264	494	242	1000	0.4890	0.715601	0.00900300	9.003046
M1143	232	508	260	1000	0.4860	0.595299	0.01555100	15.55072
M1185	265	517	218	1000	0.4765	0.251149	0.02501600	25.01637
M1266	246	491	263	1000	0.4915	0.575321	0.03752500	37.52455
M1307	269	493	238	1000	0.4845	0.679807	0.05253400	52.53438
M1347	273	487	240	1000	0.4835	0.430338	0.06869900	68.69880
M1389	254	498	248	1000	0.4970	0.900241	0.08396500	83.96520
M1470	248	497	255	1000	0.4965	0.850723	0.09596000	95.96023
M1511	254	481	265	1000	0.4945	0.230923	0.10257800	102.5782
M1551	249	501	250	1000	0.4995	0.949546	0.10257800	102.5782
M1593	275	470	255	1000	0.4900	0.059366	0.09596000	95.96023
M1674	234	497	269	1000	0.4825	0.879831	0.08396500	83.96520
M1715	260	499	241	1000	0.4905	0.958649	0.06869900	68.69880
M1755	239	509	252	1000	0.4935	0.565527	0.05253400	52.53438
M1797	252	505	243	1000	0.4955	0.749867	0.03752500	37.52455
M1878	241	501	258	1000	0.4915	0.942279	0.02501600	25.01637
M1919	239	495	266	1000	0.4865	0.769225	0.01555100	15.55072
M1959	253	519	228	1000	0.4875	0.221634	0.00900300	9.003046
M2001	252	508	240	1000	0.4940	0.609637	0.00484800	4.847794
M2082	238	505	257	1000	0.4905	0.743092	0.00242400	2.423897
M2123	255	480	265	1000	0.4950	0.206994	0.00112300	1.123269
M2163	254	504	242	1000	0.4940	0.796736	0.00048100	0.481401
M2205	229	499	272	1000	0.4785	0.996183	0.00019000	0.190321
M2286	254	495	251	1000	0.4985	0.752043	0.00006920	0.069208
M2327	280	501	219	1000	0.4695	0.855905	0.00002310	0.023069
M2367	268	492	240	1000	0.4860	0.630126	0.00000702	0.007021
M2409	256	504	240	1000	0.4920	0.793981	0.00000194	0.001942
M2490	255	508	237	1000	0.4910	0.605591	0.00000049	0.000486
M2531	244	502	254	1000	0.4950	0.896830	0.00000011	0.000109
M2571	244	492	264	1000	0.4900	0.621650	0.00000002	0.000022
M2613	256	502	242	1000	0.4930	0.894419	0.00000000	0.000004
M2694	242	496	262	1000	0.4900	0.809997	0.00000000	0.000001
M2735	257	504	239	1000	0.4910	0.792309	0.00000000	0.000000
M2775	260	501	239	1000	0.4895	0.938444	0.00000000	0.000000
M2817	261	491	248	1000	0.4935	0.572781	0.00000000	0.000000
M2898	241	539	220	1000	0.4895	0.013079	0.00000000	0.000000
M2939	255	499	246	1000	0.4955	0.951607	0.00000000	0.000000
M2979	232	505	263	1000	0.4845	0.728628	0.00000000	0.000000
M3021	236	510	254	1000	0.4910	0.520283	0.00000000	0.000000

M = marker. The number following the letter M stands for the number of the marker in the original table; PGV/QTL = proportion of genetic variance of each trait explained by each QTL; AE = additive effect.

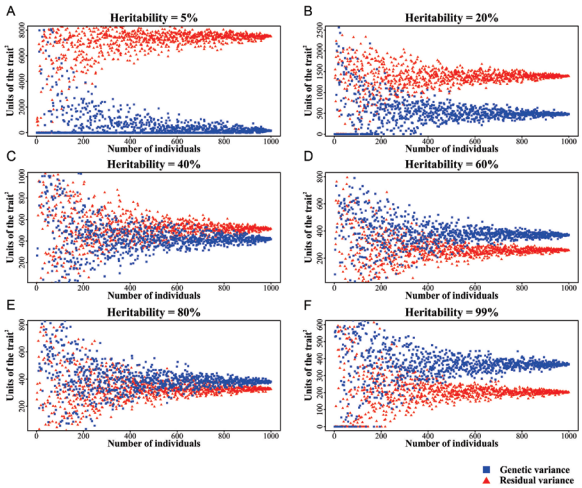

Figure 1. Trend of genotypic variance (blue) and residual variance (red) in function of the number of individuals in the training population, for traits with different heritability: **A.** 5%; **B.** 20%; **C.** 40%; **D.** 60%; **E.** 80%; **F.** 99%.

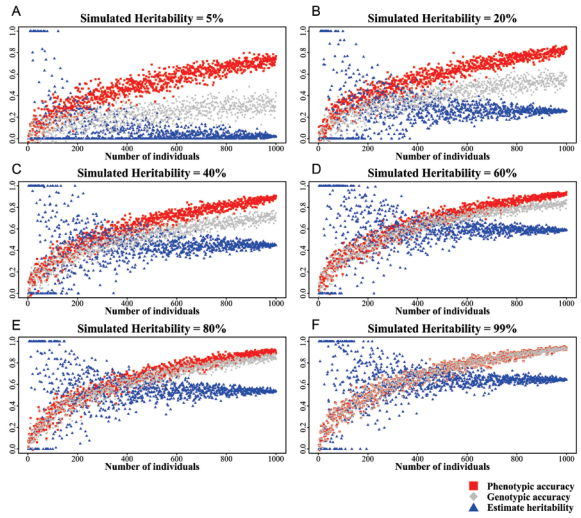

Figure 2. Trend of phenotypic accuracy, genotypic accuracy, and heritability in function of the number of individuals in the training population for traits with different heritability: **A.** 5%; **B.** 20%; **C.** 40%; **D.** 60%; **E.** 80%; **F.** 99%.

Evaluation of the number of markers necessary to obtain genomic prediction in an F_2 population

The genotypic variance was quadratic, i.e., it increased up to a certain number of markers, and then it gradually decreased with the increase in the number of markers (Figure 3). The optimal number of markers ranged according to the heritability of the trait (Table 2). It was also found that the higher the trait heritability, the greater the number of markers necessary to obtain the best genotypic variance estimate.

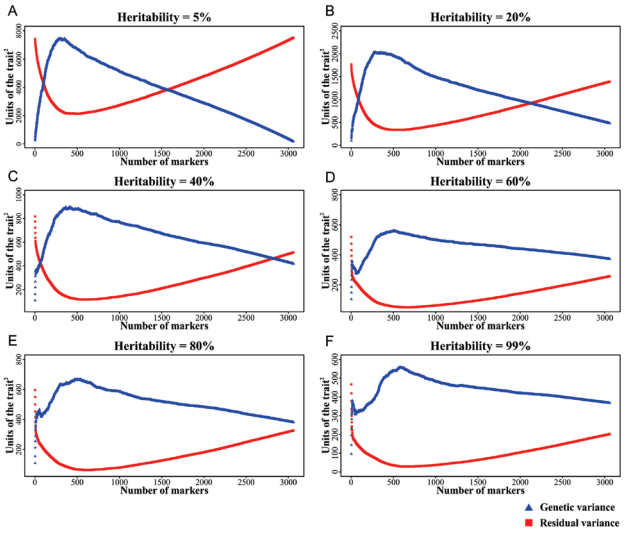

Figure 3. Trend of genotypic variance and residual variance in function of the number of markers used for the training of the genomic selection model for traits with different heritabilities: **A.** 5%; **B.** 20%; **C.** 40%; **D.** 60%; **E.** 80%; **F.** 99%.

Table 2. Number of markers (NM) necessary to obtain the optimal value (OV) of genotypic variance (σ_g^2), residual variance (σ^2), heritability (h^2), and accuracy in an F_2 population.

		σ_g^2	α^2	h^2	Accuracy
OV	C1	7466	2126	0.77	0.91
	C2	2035	337	0.85	0.94
	C3	897	115	0.88	0.96
	C4	560	50	0.91	0.98
	C5	669	62	0.91	0.98
	C6	557	28	0.95	0.99
NM*	C1	240-397	372-554	307-487	545-739
	C2	258-415	479-634	408-579	589-741
	C3	332-525	518-709	489-647	693-864
	C4	410-573	602-757	522-711	704-870
	C5	428-588	532-685	500-666	639-827
	C6	486-674	585-746	557-710	729-884

*The interval corresponds to the 5% best values for each parameter evaluated. C1 to C6 correspond to each trait simulated by varying the heritability value (5, 20, 40, 60, 80 and 99%).

Residual variance presented a quadratic trend with positive concavity, i.e., it decreased up to a certain number of markers, and then increased with an increase in the number of markers (Figure 3). The optimal number of markers for residual variance increased with the increase in the heritability of the trait (Table 2).

The value of the estimated heritability increased exponentially up to an optimal number of markers, and then decreased linearly (Figure 4). The number of markers for the maximum heritability point increased with an increase in the simulated heritability for each trait (Table 2). The decrease in estimated heritability with an increase in the number of markers was lower for traits with higher simulated heritability.

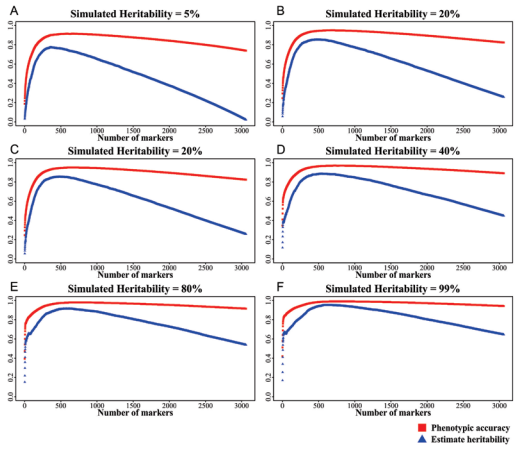

Figure 4. Trend of phenotypic accuracy and heritability in the function of the number of markers used for genomic selection model training for traits with different heritabilities: **A.** 5%; **B.** 20%; **C.** 40%; **D.** 60%; **E.** 80%; **F.** 99%.

Prediction accuracy of the training population presented an exponential increase up to a maximum point, and then a slight linear decrease (Figure 4). This decrease was lower for traits of high simulated heritability. The optimal number of markers to obtain greater accuracy increased with an increase in the heritability value of the trait (Table 2).

DISCUSSION

Marker segregation test

Using the allelic frequency, the gene frequency, and the Hardy-Weinberg equilibrium, it was verified that the simulated population indeed represented a population with all the characteristics of an F_2 population, i.e., (A)p = (a)q = 0.5, (AA)p^2 = (aa)q^2 = 0.25, and (Aa)2pq = 0.5.

The great importance of recovering all the information from an F_2 population using the simulation process is that the genetic variance, environmental variance, and heritability are easy to estimate in this type of population. According to Falconer and Mackay (1996), genetic variance in F_2 population is estimated as follows:

$$\sigma_g^2 = 2pq\alpha^2 + (2pqd)^2 \qquad \text{(Equation 5)}$$

in which d value was simulated as 0 for all loci, and thus the genetic variance is equal to the additive variance. This is easily calculated, since α^2 is the variance of the markers calculated using the RR-BLUP method. Thus, heritability can be estimated from the equation proposed by Falconer and Mackay (1996) for an F_2 population:

$$h^2 = \frac{\sigma_g^2}{(\sigma_g^2 + \sigma^2)} \qquad \text{(Equation 6)}$$

and σ^2 is the residual variance of the markers estimated by the RR-BLUP method.

Consequently, all the genetic and environmental parameters were accurately calculated, and hence, these parameters were the criteria for choosing the best genomic selection model, i.e., the model composed of the ideal number of individuals in the training population, and the number of markers required to accurately train the model.

Training population size versus estimated genetic value

Usually, the increase in the number of individuals in the TP increases the prediction accuracy of the genetic value (Desta and Ortiz, 2014). However, despite the increase in accuracy, when more than 600 individuals was used in the TP, this increment was very low, making it almost null for traits with a heritability of 80-99% in the present study.

Besides the number of individuals in the TP, population structure may influence the prediction by the genomic selection methods. Studies on oat (Asoro et al., 2011), corn (Ogutu et al., 2012), and beet (Würschum et al., 2013) showed that the use of a structured population together with large enough TP considerably increases prediction accuracy. Therefore, all the results of this study are valid for an F_2 population. Other studies are required to evaluate other types of populations, such as backcrossing, RILs, half-sib families, and full-sib families, since each type of population has a different structure, influencing the prediction by genomic selection methods.

Isidro et al. (2015), when evaluating several methods to optimize the choice of individuals to compose the TP, found that the population structure and the trait architecture are the factors that most influence the TP performance. Thus, it is difficult to verify a standard size of the TP for the several possible heritabilities and different population types. Using the current study with an F_2 population, it can be concluded that 600 individuals are enough, regardless of the trait architecture. However, for traits with low heritability, accuracy values are less stable, i.e., depending on the individuals of the TP, accuracy is higher or lower, and when heritability increases, the accuracy value is constant, regardless of the number of individuals of the TP. This was verified in this study, since the analyses were repeated 50 times for each TP size. Thus, prior knowledge of the trait under study may help researchers to design the experiment in order to obtain accurate results through genomic selection, and consequently reduce the cost of the breeding program.

Markers density versus estimated breeding value

It was found that a number of markers ranging from 200 to 900 is enough to capture all the genotypic variance of an F_2 population, and consequently achieve maximum accuracy. This value varies depending on the trait heritability, since the higher the heritability, the greater is the number of markers necessary to obtain maximum accuracy. This fact can be explained by the effect of each QTL and their influence on the genomic selection methods. All traits were simulated with 60 QTL; however, the higher the heritability of the trait, the higher is the effect of each QTL. One of the characteristics of genomic selection methods is the capture of minor effect markers, mainly because the RR-BLUP method uses the same variance for all markers. This means that RR-BLUP cannot capture the full effect of major effect QTL, which thereby requires more markers to explain the genotypic variance of the trait. An alternative to improve the variance capture of QTL for major effects is the use of Bayesian methods, which assume specific variance for each marker, such as Bayes A and Bayes B (Gianola et al., 2009).

Erbe et al. (2013) evaluated Brown Swiss cattle by genotyping the animals using 777k chips, and observed that the genetic variance estimated via genomic selection models increased to 20k, becoming constant after this number of markers. They concluded that even with a population of infinite individuals for training and a large number of markers, it would not be possible to increase the accuracy for this population. Poland et al. (2012) found that 1827 SNPs were enough to capture all the genetic variance in wheat populations. In our study, the marker of lower effect was deleted in each iteration, and was observed that it was not necessary to use several markers to explain the genotypic variance of the trait, since by the simulation process, only 60 markers explained all the variation of the trait. Thus, prior knowledge of the trait may be important for the development of low density chips specific for a given trait or species. The development of this type of chip is important to reduce genotyping costs. In animal breeding, low density chips for cattle (Heaton et al., 2002; Boichard et al., 2012) and pigs (Wellmann et al., 2013) have been developed. Moreover, the genotyping cost of a low density chip is much lower than that of a high density chip (Habier et al., 2009).

In addition, the high accuracy for models using a small number of markers (200 to 900 SNPs), which was verified in this study, can be explained by the fact that individuals of the training population are highly correlated, since all of them are descended from the same parent, i.e., the individuals of the F_2 population share alleles identical by descent (Poland et al., 2012). However, despite the reduced number of markers for each studied trait, these markers are different for each trait, making the construction of a multi-trait low-density chip very difficult (Habier et al., 2009). Therefore, in further studies, it is necessary to seek strategies that enable simultaneous marker selection for several traits, and thus build a multi-trait low-density chip.

The number of markers used in this study were low. However, several studies have shown that a small number of markers can be used for high-accuracy genomic selection for many traits (Bhering et al., 2015; Spindel et al., 2015). In the future, it is necessary to test the optimal number of markers and individuals, using a large number of markers.

CONCLUSION

The ideal number of individuals to compose the training population is strongly correlated with the heritability of the trait. However, a training population comprising more than 600 individuals ensures maximum accuracy, regardless of the heritability for an F_2 population.

A genomic selection model that uses 300-800 markers is enough to capture all the genetic variance, and to decrease the residual variance, in order to obtain the maximum prediction accuracy of an F_2 population.

ACKNOWLEDGMENTS

We are thankful to CAPES (Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), FAPEMIG (Fundação de Amparo à Pesquisa de Minas Gerais), and Universidade Federal de Viçosa for financial support. We also thank the Biometric Lab (Universidade Federal de Viçosa, Brazil) where all analyses were performed by remote access.

REFERENCES

Allard RW (1999). Principles of plant breeding. John Wiley & Sons, New York.

Ashraf M, Akram NA, Mehboob-Ur-Rahman and Foolad MR (2012). Marker-assisted selection in plant breeding for salinity tolerance. *Methods Mol. Biol.* 913: 305-333.

Asoro FG, Newell MA, Beavis WD, Scott MP, et al. (2011). Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. *Plant Genome* 4: 132-144. http://dx.doi.org/10.3835/plantgenome2011.02.0007

Bassi FM, Bentley AR, Charmet G, Ortiz R, et al. (2016). Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). *Plant Sci.* 242: 23-36. http://dx.doi.org/10.1016/j.plantsci.2015.08.021

Belaj A, del Carmen Dominguez-García M, Atienza SG, Urdíroz NM, et al. (2012). Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. *Tree Genet. Genomes* 8: 365-378. http://dx.doi.org/10.1007/s11295-011-0447-6

Beyene Y, Semagn K, Mugo S, Tarekegne A, 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. http://dx.doi.org/10.2135/cropsci2014.07.0460

Bhering LL, Junqueira VS, Peixoto LA, Cruz CD, et al. (2015). Comparison of methods used to identify superior individuals in genomic selection in plant breeding. *Genet. Mol. Res.* 14: 10888-10896. http://dx.doi.org/10.4238/2015.September.9.26

Boichard D, Chung H, Dassonneville R, David X, et al.; Bovine LD Consortium (2012). Design of a bovine low-density SNP array optimized for imputation. *PLoS One* 7: e34130. http://dx.doi.org/10.1371/journal.pone.0034130

Borém A and Miranda GV (2013). Melhoramento de Plantas. UFV, Viçosa.

Cros D, Denis M, Sánchez L, Cochard B, et al. (2015). Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). *Theor. Appl. Genet.* 128: 397-410. http://dx.doi.org/10.1007/s00122-014-2439-z

Cruz CD (2013). GENES: a software package for analysis in experimental statistics and quantitative genetics. *Acta Sci. Agron.* 35: 271-276. http://dx.doi.org/10.4025/actasciagron.v35i3.21251

de Los Campos G, Vazquez AI, Fernando R, Klimentidis YC, et al. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* 9: e1003608. http://dx.doi.org/10.1371/journal.pgen.1003608

Desta ZA and Ortiz R (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci.* 19: 592-601. http://dx.doi.org/10.1016/j.tplants.2014.05.006

Dirlewanger E, Pronier V, Parvery C, Rothan C, et al. (1998). Genetic linkage map of peach [*Prunus persica* (L.) Batsch] using morphological and molecular markers. *Theor. Appl. Genet.* 97: 888-895. http://dx.doi.org/10.1007/s001220050969

Erbe M, Gredler B, Seefried FR, Bapst B, et al. (2013). A function accounting for training set size and marker density to model the average accuracy of genomic prediction. *PLoS One* 8: e81046. http://dx.doi.org/10.1371/journal.pone.0081046

Falconer D and Mackay T (1996). Introduction to Quantitative Genetics. Longman Scientific & Technical, Harlow, UK.

Frascaroli E, Schrag TA and Melchinger AE (2013). Genetic diversity analysis of elite European maize (*Zea mays* L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs. *Theor. Appl. Genet.* 126: 133-141. http://dx.doi.org/10.1007/s00122-012-1968-6

Gianola D, de los Campos G, Hill WG, Manfredi E, et al. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* 183: 347-363. http://dx.doi.org/10.1534/genetics.109.103952

Gouy M, Rousselle Y, Bastianelli D, Lecomte P, et al. (2013). Experimental assessment of the accuracy of genomic selection in sugarcane. *Theor. Appl. Genet.* 126: 2575-2586. http://dx.doi.org/10.1007/s00122-013-2156-z

Habier D, Fernando RL and Dekkers JC (2009). Genomic selection using low-density marker panels. *Genetics* 182: 343-353. http://dx.doi.org/10.1534/genetics.108.100289

He J, Zhao X, Laroche A, Lu Z-X, et al. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci.* 5: 484. http://dx.doi.org/10.3389/fpls.2014.00484

Heaton MP, Harhay GP, Bennett GL, Stone RT, et al. (2002). Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. *Mamm. Genome* 13: 272-281. http://dx.doi.org/10.1007/s00335-001-2146-3

Isidro J, Jannink J-L, Akdemir D, Poland J, et al. (2015). Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* 128: 145-158. http://dx.doi.org/10.1007/s00122-014-2418-4

Langer M and Maixner M (2004). Molecular characterisation of grapevine yellows associated phytoplasmas of the stolbur-group based on RFLP-analysis of non-ribosomal DNA. *VITIS-Journal of Grapevine Research* 43: 191-199.

Lightfoot DA (2015). Two Decades of Molecular Marker-Assisted Breeding for Resistance to Soybean Sudden Death Syndrome. *Crop Sci.* 55: 1460-1484. http://dx.doi.org/10.2135/cropsci2014.10.0721

Lynch M and Milligan BG (1994). Analysis of population genetic structure with RAPD markers. *Mol. Ecol.* 3: 91-99. http://dx.doi.org/10.1111/j.1365-294X.1994.tb00109.x

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

Ogutu JO, Schulz-Streeck T and Piepho H-P (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proc.* 6 (Suppl 2): S10. http://dx.doi.org/10.1186/1753-6561-6-S2-S10

Ordas B, Butron A, Alvarez A, Revilla P, et al. (2012). Comparison of two methods of reciprocal recurrent selection in maize (*Zea mays* L.). *Theor. Appl. Genet.* 124: 1183-1191. http://dx.doi.org/10.1007/s00122-011-1778-2

Pandey MK, Rani NS, Sundaram RM, Laha GS, et al. (2013). Improvement of two traditional Basmati rice varieties for bacterial blight resistance and plant stature through morphological and marker-assisted selection. *Mol. Breed.* 31: 239-246. http://dx.doi.org/10.1007/s11032-012-9779-7

Poland J, Endelman J, Dawson J, Rutkoski J, et al. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5: 103-113. http://dx.doi.org/10.3835/plantgenome2012.06.0006

R Core Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Ren R, Ray R, Li P, Xu J, et al. (2015). Construction of a high-density DArTseq SNP-based genetic map and identification of genomic regions with segregation distortion in a genetic population derived from a cross between feral and cultivated-type watermelon. *Mol. Genet. Genomics* 290: 1457-1470. http://dx.doi.org/10.1007/s00438-015-0997-7

Soldati MC, Fornes L, Van Zonneveld M, Thomas E, et al. (2013). An assessment of the genetic diversity of *Cedrela balansae* C. DC. (Meliaceae) in Northwestern Argentina by means of combined use of SSR and AFLP molecular markers. *Biochem. Syst. Ecol.* 47: 45-55. http://dx.doi.org/10.1016/j.bse.2012.10.011

Spindel J, Begum H, Akdemir D, Virk P, 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. http://dx.doi.org/10.1371/journal.pgen.1004982

Tang B, Jenkins JN, McCarty J and Watson C (1993). F2 hybrids of host plant germplasm and cotton cultivars: II. Heterosis and combining ability for fiber properties. *Crop Sci.* 33: 706-710. http://dx.doi.org/10.2135/cropsci1993.0011183X003300040013x

Tang B, Jenkins J, Watson C, McCarty J, et al. (1996). Evaluation of genetic variances, heritabilities, and correlations for yield and fiber traits among cotton F2 hybrid populations. *Euphytica* 91: 315-322. http://dx.doi.org/10.1007/BF00033093

Wellmann R, Preuß S, Tholen E, Heinkel J, et al. (2013). Genomic selection using low density marker panels with application to a sire line in pigs. *Genet. Sel. Evol.* 45: 28. http://dx.doi.org/10.1186/1297-9686-45-28

Würschum T, Reif JC, Kraft T, Janssen G, et al. (2013). Genomic selection in sugar beet breeding populations. *BMC Genet.* 14: 85. http://dx.doi.org/10.1186/1471-2156-14-85

Yaniv E, Raats D, Ronin Y, Korol AB, et al. (2015). Evaluation of marker-assisted selection for the stripe rust resistance gene Yr15, introgressed from wild emmer wheat. *Mol. Breed.* 35: 1-12. http://dx.doi.org/10.1007/s11032-015-0238-0

Zhang J, Song Q, Cregan PB and Jiang G-L (2016). Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). *Theor. Appl. Genet.* 129: 117-130. http://dx.doi.org/10.1007/s00122-015-2614-x

Supplementary material

Table S1. Segregation test, minor allele frequency (MAF), P value associated with the χ^2 test of the evaluation of the Hardy-Weinberg equilibrium (hwe. P value) for all SNPs.