Genome wide selection in *Citrus* breeding

I.B. Gois[1], A. Borém[1], M. Cristofani-Yaly[2], M.D.V. de Resende[3,4], C.F. Azevedo[4], M. Bastianel[2], V.M. Novelli[2] and M.A. Machado[2]

[1]Departamento de Fitotecnia, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[2]Instituto Agronômico de Campinas, Centro APTA Citros Sylvio Moreira, Cordeirópolis, SP, Brasil
[3]Embrapa Florestas, Colombo, PR, Brasil
[4]Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, MG, Brasil

Corresponding author: I.B. Gois
E-mail: itamarafloresta@gmail.com

Genet. Mol. Res. 15 (4): gmr15048863
Received June 8, 2016
Accepted July 20, 2016
Published October 17, 2016
DOI http://dx.doi.org/10.4238/gmr15048863

ABSTRACT. Genome wide selection (GWS) is essential for the genetic improvement of perennial species such as *Citrus* because of its ability to increase gain per unit time and to enable the efficient selection of characteristics with low heritability. This study assessed GWS efficiency in a population of *Citrus* and compared it with selection based on phenotypic data. A total of 180 individual trees from a cross between Pera sweet orange (*Citrus sinensis* Osbeck) and Murcott tangor (*Citrus sinensis* Osbeck x *Citrus reticulata* Blanco) were evaluated for 10 characteristics related to fruit quality. The hybrids were genotyped using 5287 DArT_seq™ (diversity arrays technology) molecular markers and their effects on phenotypes were predicted using the random regression - best linear unbiased predictor (rr-BLUP) method. The predictive ability, prediction bias, and accuracy of GWS were estimated to verify

its effectiveness for phenotype prediction. The proportion of genetic variance explained by the markers was also computed. The heritability of the traits, as determined by markers, was 16-28%. The predictive ability of these markers ranged from 0.53 to 0.64, and the regression coefficients between predicted and observed phenotypes were close to unity. Over 35% of the genetic variance was accounted for by the markers. Accuracy estimates with GWS were lower than those obtained by phenotypic analysis; however, GWS was superior in terms of genetic gain per unit time. Thus, GWS may be useful for *Citrus* breeding as it can predict phenotypes early and accurately, and reduce the length of the selection cycle. This study demonstrates the feasibility of genomic selection in *Citrus*.

Key words: Molecular markers; DarT_seq; Linkage disequilibrium; Early selection; Selective accuracy

INTRODUCTION

The *Citrus* genus is a major food and economic crop worldwide, and many strategies have been adopted to increase its productivity and quality. Breeding is the preferred strategy to address these issues. When referring to *Citrus* breeding, it is important to note that the citrus plant itself consists of two parts, the canopy and the rootstock, which are often from two different species and require targeted studies to improve their individual quality and interaction. Phenotypic selection is generally applied, but this is costly and time consuming, especially when the desired traits are expressed at later stages (Machado et al., 2011).

Marker-assisted selection (MAS) was proposed by Lande and Thompson (1990) to obtain faster gains and increase selection efficiency with respect to selection based on phenotypic data only. MAS utilizes phenotypic data and molecular markers in genetic linkage with certain loci controlling desired quantitative traits (quantitative trait loci, QTL). QTL markers are selected following appropriate statistical modeling and are subjected to testing for type II error, i.e., the probability of accepting a false hypothesis (Resende et al., 2014).

In *Citrus* breeding, the MAS strategy has been adopted to select several traits of economic importance such as resistance to biotic and abiotic stress (Ito et al., 2014) and physicochemical characteristics of fruit. However, the identification of QTL associated with such characteristics only explains a small fraction of the observed genetic variation in each character (Siviero et al., 2002; Siviero et al. 2006; Gussen et al., 2011; Asins et al., 2012); therefore, selection based on these markers is not feasible. The rationale for not using MAS is that traits under selection are governed by many loci with small effects that do not reach statistical significance when used in traditional QTL analysis, and this in turn leads to the identification of only a small number of QTL with major effects (Kemper and Goddard 2012; Resende et al., 2014).

Genome-wide selection (GWS) was first proposed by Meuwissen et al. (2001) and is a form of molecular-marker assisted selection that simultaneously predicts the genetic effects of hundreds or thousands of DNA markers on phenotypes (Resende et al., 2008). These markers, which are in linkage disequilibrium (LD) with the QTL, can have both large and small effects and can account for nearly all of the genetic variation within a quantitative trait (Resende et al., 2008). Additionally, whole genome prediction tends to be more accurate because it

better accommodates the variation created by Mendelian segregation during gamete formation (Resende et al., 2008; Zapata-Velenzuela et al., 2013).

The increasing availability and decreasing costs of molecular markers spanning the whole genome has led to the implementation of large-scale genomic selection in breeding programs for different species (Misztal et al., 2009; Zhong et al., 2009; Daetwyler et al., 2010; Grattapaglia and Resende, 2011; Resende et al., 2012; Resende Júnior et al., 2012). GWS has promise for use in breeding programs because of the efficient selection of traits with low heritability, permitting the best use of available genetic resources by selecting appropriate genetic crosses, and increasing genetic gain per unit time as it enables early identification of the best individuals for selection (Legarra et al., 2008; Daetwyler et al., 2013; Resende et al., 2014).

In *Citrus* species, the juvenile period is long, ranging from 1 to 20 years; although flowering and fruiting may occur within 3-7 years depending on the species. Therefore, there is a delay in the hybridization process and in the selection of desired characteristics in breeding programs. In addition, large areas are required for the development of hybrids, which not only increases the cost of plant maintenance in the field but also limits the number of families and individuals per family that can be evaluated (Talon and Gmitter Junior, 2008; Gmitter Junior et al., 2012). GWS can minimize these undesirable consequences as selection can be performed in juvenile plants, which reduces the interval between generations, increases the intensity of selection and, therefore, the gain per unit time and cost (Resende et al., 2008; Wong and Bernardo, 2008; Heffner et al., 2009; Grattapaglia and Resende, 2011; Kemper and Goddard, 2012).

GWS has rarely been applied to fruit trees, being restricted to apples (Kumar et al., 2012), pears (Iwata et al., 2013), and grapes (Viana et al., 2016). To our knowledge, the present study represents the first attempt to investigate the feasibility of genomic selection in *Citrus*. This research was carried out to apply GWS to a biparental hybrid population of *Citrus* using DArT_seq™ markers and phenotypes of 10 traits, to assess its selection efficiency, and compare it to selection based on phenotypic data.

MATERIAL AND METHODS

Characterization of the population and phenotypic traits

The genetic effects of the markers were predicted in a hybrid population of *Citrus*, which is used as canopy. The population composed by 180 individuals was generated by crossing Pera sweet orange (*Citrus sinensis* Osbeck) and Murcott tangor (*Citrus sinensis* Osbeck x *Citrus reticulata* Blanco).

The experiment was established in 2004 at Cordeirópolis, SP, in a completely randomized design with three clonal replicates. Phenotypic evaluation was performed in 2012 for the following characteristics: fruit mass (g), longitudinal fruit diameter (cm), transverse fruit diameter (cm), ratio of longitudinal fruit diameter (cm)/transverse fruit diameter (cm), number of segments per fruit, number of seeds per fruit, juice yield (mL), soluble solids (ºBrix), ratio of soluble solids/acidity, and number of fruits per box.

A sample from each replicate, consisting of five fruits each, was used to obtain phenotypic data. Fruit mass was measured using a digital balance with a capacity of 15 kg and accuracy of 5 g; the longitudinal and transverse fruit diameters were measured using a graduated ruler; the number of segments per fruit and seed number were directly counted; the number of fruits per box was obtained by dividing capacity of the box (40.8 kg) by the average

weight of the fruit; juice yield was measured after crush extraction in an OIC model OTTO 1800 and the juice/fruit ratio was calculated. The soluble solids content (ºBrix) was determined by direct reading on a refractometer (B & S model RFM 330) and acidity was determined by titrating 25 mL juice against normal sodium hydroxide solution with phenolphthalein as an indicator. The soluble solids/acidity ratio, which indicates the degree of fruit ripening, was also calculated. All analyses were performed in the Quality and Postharvest Laboratory of the Center APTA *Citrus* Sylvio Moreira/IAC, Cordeirópolis, SP.

Statistical model for the analysis of phenotypic data

Phenotypic data were analyzed to estimate genetic variance and heritability using the REML/BLUP method, also called the mixed model methodology (Henderson, 1973), using the Selegen REML/BLUP software (Resende, 2002). The following model was fit: y = Xu + Zg + e; where, y is the vector of observations; u is the scalar related to overall mean (fixed effect); g is the vector of genotypic effects (assumed to be random); and e is the residual vector (random). The capital letters (X, Z) represent incidence matrices for these effects. The structures of means and variances associated with the model are: $y \mid u, V \sim N(Xu, V)$; $g \mid \sigma_g^2 \sim N(0, I\sigma_g^2)$; $e \mid \sigma_e^2 \sim N(0, I\sigma_e^2)$; and $Cov(g, e) = 0$; where, V is the phenotypic covariance matrix; I is an identity matrix; σ_g^2 and σ_e^2 are the genotypic and residual variances, respectively. The mixed model equations for the best linear unbiased predictor (BLUP) method to obtain individual breeding values was:

$$\begin{bmatrix} \hat{u} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I\lambda \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \end{bmatrix} \quad \text{(Equation 1)}$$

where:

$$\lambda = \sigma_e^2 / \sigma_g^2 = \frac{1-h^2}{h^2} \text{ is a penalization parameter; and} \quad \text{(Equation 2)}$$

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \text{ is the individual broad-sense heritability.} \quad \text{(Equation 3)}$$

Variance components were estimated by restricted maximum likelihood (REML) (Patterson and Thompson, 1971) iterating the mixed model equations. The prediction accuracy of genotypic values for clone selection was calculated from the inverse of the left-hand side of the mixed model equations (Resende, 2002).

Genotyping and quality control of the markers

Molecular assessment of 180 individuals in the hybrid *Citrus* population was performed using the diversity arrays technology (DArT) method (Jaccoud et al., 2001), which reduces genome complexity by using a combination of restriction enzymes and a next-generation sequencing technique called DArT_seq™. Total DNA was extracted, quantified on

a Nanodrop-8000 spectrophotometer, normalized to 100 ng/μL, distributed onto plates, and submitted to genotyping on the Diversity Arrays Technology platform (DArT P/L, Australia).

The method used to obtain the DArT_seqTM markers involved reducing genome complexity using the restriction enzymes *Pst*I and *Taq*I, and subsequent sequencing. *Pst*I and adapters specific for 96 different bar codes were linked to the restriction fragments. The resulting products were amplified and their quality was checked. Samples were sequenced on an Illumina Hiseq2000 machine. The *Pst*I adapters included a sequencing primer such that the sequences were always read from the *Pst*I restriction sites. The resulting sequences (FASTQ files) were filtered for quality, with a confidence cut-off of 90%. Sequences were aligned with the Clementine tangerine reference genome available at https://phytozome.jgi.doe.gov. This procedure yielded 27,960 DArT_seq™ markers, which were coded "0", if absent, or "1", if present. All markers were analyzed in F1 progeny hybrid seeds for expected Mendelian segregation patterns.

Quality control of the markers was performed by eliminating low polymorphic loci [minor allele frequency (MAF) ≤ 5%] and/or loci with low call rate in the genotyped individuals (call rate ≤ 95%). MAF was calculated using the equation: MAF = min[p, (1 - p)], where *p* is the allele frequency in a locus; and call rate was calculated using the expression: Call Rate = 1 - M / P, where *M* refers to individuals with missing genotypes and *P* refers to individuals with present genotypes (not missing).

Predicting genetic effects of markers using random regression (RR)-BLUP

The RR-BLUP method (Meuwissen et al., 2001) was used to predict the genetic effects of the markers. This method fits the random effects of markers as a covariate with effects on phenotypes (Resende et al., 2012).

The estimators associated with this random regression procedure promote shrinkage on the marker effects as a function of the penalty parameter λ, which makes it possible to estimate a higher number of parameters than the number of data points. The penalty parameter is given by: $\lambda = \frac{\sigma_e^2}{\sigma_{gi}^2} = \frac{\sigma_e^2}{\sigma_g^2 / n_Q}$, where σ_{gi}^2 is the genetic variance associated with the marker locus *i*; σ_g^2 is the trait genetic variance; σ_e^2 is the residual variance; and n_Q is associated with the number of marker loci and, for DArT dominant markers, is represented by $n_Q = \Sigma_{pi}(1 - pi)$ (Resende et al., 2012).

The following general linear mixed model was designed to predict the effects of the markers: $y = Xu + Wm_a + Sm_d + e$; where *y* is the vector of phenotypic observations; *u* is the vector of fixed effects (overall means); m_a is the vector of additive genetic marker effects; m_d is the vector of dominant genetic marker effects; *e* is the vector of random errors; *X*, *W,* and *S* are the incidence matrices for *u*, m_a, and m_d, respectively.

Parameterizations of additive and dominant marker effects with dominant markers such DArT were performed according to the method described by Viana and Resende (2014). In an allogamous population in Hardy-Weinberg equilibrium (HWE), appropriate parameterization in the *W* matrix, associated with additive effects, is '0' for the absence of the marker (mm type) and 1.78 for the presence of the marker (MM or Mm types). In the *S* matrix along with the effects of dominance, parameterization is as follows: '0' for the absence of the marker (mm) and 0.89 for the presence of marker (MM or Mm). Assuming HWE, the additive genetic variation of the character in the population is given by: $\sigma_\alpha^2 = \sum_{i=1}^{n} p_i\, q_i\, \alpha_i^2$; where p_i is the frequency G_{MM+Mm}, e α_i is the additive effect at locus *i*; and dominant genetic variation is

given by: $\sigma_d^2 = \sum_{i=1}^{n}(p_i^2\ q_i^2\ \alpha_i^2)$; where d_i is the dominant effect for locus *i*.

The means and variance structure is defined as: $m_a \sim N(0, WW'\sigma_{ma}^2)$; md ~ N(0, SS' σ_{md}^2); e ~ N(0, I σ_e^2), in that σ_{ma}^2 is the marker additive genetic variance given by $\sigma_{ma}^2 = \frac{\sigma_a^2}{\sum_{i-1}^{n}(p_i\ q_i)}$; σ_{md}^2 is the marker variance due to dominance deviations and is defined as $\sigma_{md}^2 = \frac{\sigma_d^2}{\sum_{i-1}^{n}(p_i\ q_i)^2}$; σ_e^2 are residual variances.

The mixed model equations to predict m_a and m_d using the RR-BLUP method are:

$$\begin{bmatrix} X'X & X'W & X'S \\ W'X & W'W + I\frac{\sigma_e^2}{\sigma_{ma}^2} & W'S \\ S'X & S'W & S'S + I\frac{\sigma_e^2}{\sigma_{md}^2} \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{m}_a \\ \hat{m}_d \end{bmatrix} = \begin{bmatrix} X'Y \\ W'Y \\ S'Y \end{bmatrix} \quad \text{(Equation 4)}$$

The genomic values (GV) of individuals were estimated using the equation: $GV = Wm_a + Sm_d$. All analyses were performed on the R software, version 3.1.2 (R Development Core Team, 2012) using the RR-BLUP package (Endelman, 2011).

Predictive ability, prediction bias, and accuracy of GWS

The predictive ability (r_{yg}) was obtained by correlating the corrected phenotypic values with predicted genomic values and was determined by cross-validation through a Jackknife procedure. Prediction bias (*b*) was calculated as the regression coefficient of phenotypic values on the predicted genomic values, wherein '*b*' values close to 1 indicate that predictions are not biased and are effective in predicting the actual magnitude of difference between the individuals being assessed (Resende et al., 2012).

The experimental accuracy of GWS was obtained by the estimator: $r_{gg} = r_{yg}/h$, in which r_{gg} corresponds to the accuracy of GWS; r_{yg} the predictive ability; and *h* the square root of the heritability of the character being studied (Legarra et al., 2008).

The expected accuracy was calculated using the formula proposed by Resende et al. (2008): $r_{ggE} = \sqrt{\frac{r_{mq}^2(Nhm_m^2)}{1+[(N-1)h_m^2]}}$, where in $r_{mq}^2 = \frac{n_m}{n_m + M_e}$ is the proportion of variance explained by the markers; *N* is the number of genotyped and phenotyped individuals; n_m is the number of markers; h_m^2 is the individual heritability of a locus given by $h_m^2 = \frac{r_{mq}^2 h^2}{M_e}$, and h^2 is the heritability of the character. For calculating M_e (effective number of chromosome segments) the approach proposed by Goddard et al. (2011) was used: $Me = 2NeL$, where *Ne* is the effective population number (2, in the present paper) and *L* is the length of the *Citrus* genome in Morgans.

The estimated number of QTL controlling each character was calculated by the expression $n_{QTL} = \frac{Nr_{mq}^2 h^2 (r_{mq}^2 - r_{gg}^2)}{r_{gg}^2}$. These expressions were obtained using accuracy formulae proposed by Resende et al. (2008), wherein the term 'r_{mq}^2' was estimated from the number of markers that maximized the accuracy of GWS for each character.

Efficiency of GWS

The superiority of GWS over selection based on phenotypic data only in a perennial species, is mainly related to the gain per unit time (Resende et al., 2012). To confirm this, the gain in selective efficiency of GWS, compared to selection based only on phenotypes,

(Equation 4)

was calculated using the expression: $G_{Ef} = \frac{r_{gg} t_f}{r_{yy} t_{GWS}} - 1$, where r_{gg} is the accuracy of GWS; r_{yy} is the accuracy of selection based on phenotype means; t_f is the average time required for the completion of the selection cycle based on phenotype means; and t_{GWS} is the average time required to complete the selection cycle based on the GWS (Resende et al., 2012).

RESULTS AND DISCUSSION

After quality control, 5287 markers were found to be useful and their effects were estimated. The estimated genome size of the sweet orange is, according to Jarrell et al. (1992), 1700 cM. Therefore, the density of markers used in the analyses was ~3 markers per cM, which corresponds to a distance of 0.32 cM between markers. This value was used to calculate the proportion of variance that could be explained by the markers (r^2_{mq}), which, in turn, is related to the expected magnitude of LD (r^2) (Resende et al., 2012). Figure 1 shows the expected proportion of variance accounted for by the markers (r^2_{mq}) depending on the number of markers used. As shown, r^2_{mq} increases with the number of markers as this value depends on the recombination frequency, which is in turn, a function of the distance between the markers (Grattapaglia and Resende, 2011). Assuming that LD between marker pairs is approximately equal to LD between markers and the QTL (Zhong et al., 2009), the number of markers used in this study yield high LD.

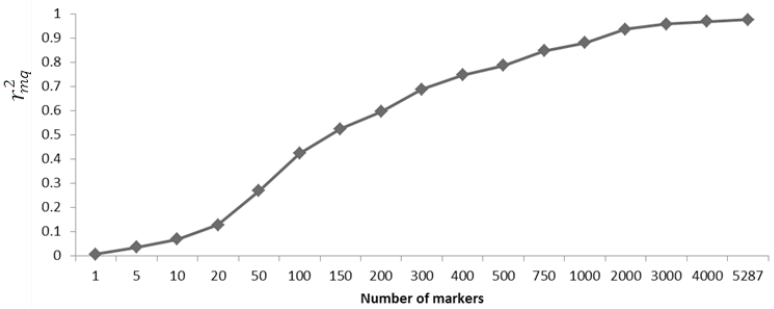

Figure 1. Expected proportion of genetic variance explained by markers (r^2_{mq}) with increasing number of markers.

Estimates of genomic heritability (r^2_G), predictive ability (r_{yg}), prediction bias (b), accuracy of GWS (r_{gg}), QTL number estimates (n_{QTL}), and proportion of genetic variance explained by the markers (r^2_{mq} when selected based on the magnitudes of their effects are presented in Table 1.

The fraction of trait heritability attributable to the markers (ω) varied from 20% (ratio of soluble solids/acidity) to 53% (ratio of longitudinal/transverse fruit diameter). The number of markers necessary to provide these fractions ranged from 100 for soluble solid content to 2000 for the number of seeds (Table 1). As the number of markers increased, the accuracy values described by these markers tended to remain constant once a maximum value was reached.

The observed missing heritability, defined as the fraction of heritability not accounted for by markers, is probably due to incomplete LD among the alleles causing genetic variation and the markers used for heritability estimation. LD may be lost when the alleles causing genetic variation have lower MAF values than the marker (Kemper and Goddard, 2012).

Table 1. Estimates of genomic heritability (h^2_G, predictive ability (r_{yg}), prediction bias (b), accuracy of GWS (r_{gg}), quantitative trait loci (QTL) number estimates (n_{QTL}), and proportion of genetic variance explained by the markers (r^2_{mq} when selected based on the magnitudes of their effects in a breeding *Citrus* population.

Trait	h^2	h^2_G	w	n_m	B	r_{yg}	r_{yy}	r_{gge}	r_{ggo}	r^2_{mq}	n_{QTL}
DL	0.77	0.24	0.31	1000	0.95	0.64	0.88	0.79	0.73	0.54	99
DT	0.76	0.25	0.33	750	0.90	0.63	0.87	0.77	0.72	0.53	97
DL/ DT	0.49	0.26	0.53	1000	0.90	0.63	0.70	0.72	0.91	0.84	11
FM	0.81	0.28	0.35	300	0.88	0.64	0.90	0.72	0.72	0.53	68
F/B	0.68	0.22	0.32	400	0.98	0.64	0.82	0.72	0.77	0.61	47
NG	0.62	0.27	0.44	200	0.81	0.59	0.79	0.74	0.73	0.54	34
NS	0.80	0.22	0.28	2000	0.95	0.62	0.89	0.81	0.69	0.49	131
JY	0.67	0.21	0.31	1000	0.94	0.56	0.82	0.77	0.69	0.49	110
SS	0.76	0.27	0.36	100	0.74	0.53	0.87	0.63	0.61	0.38	49
Rt	0.81	0.16	0.20	750	1.02	0.53	0.90	0.78	0.59	0.36	219

h_2: Heritability based on phenotypes; h^2_G: Heritability based on markers; $w = h^2_G / h_2$: Proportion of heritability based on phenotypes captured by markers; n_m: number of markers that maximized the selective accuracy of GWS; b: bias of prediction; r_{yg}: Predictive ability; r_{yy}: accuracy of phenotypic selection based on the REML/BLUP method; r_{gge}: expected accuracy of GWS; r_{ggo}: observed accuracy of GWS; r^2_{mq}: genetic variance explained for the markers; n_{QTL}: estimated number of QTL controlling the trait in the population; DL: longitudinal fruit diameter (cm); DT: transverse fruit diameter (cm); DL/DT: ratio of longitudinal fruit diameter/transverse fruit diameter; FM: fruit mass (g); F/B: number of fruits per box; NG: number of segments per fruit; NS: number of seeds per fruit; JY: juice yield (mL); SS: soluble solids content; Rt: ratio of soluble solids/acidity.

The predictive ability of GWS, which reflects the ability of molecular information to consistently predict a phenotype (Cavalcanti et al., 2012), ranged from 0.53 for the ratio of soluble solids/acidity and soluble solids traits, to 0.64, for the longitudinal diameter of fruit, fruit mass, and number of fruits per box (Table 1). These magnitudes are similar to those reported for forest trees species (Resende et al., 2012, Resende Junior et al., 2012) and dairy cattle breeding (Hayes et al., 2009).

The regression coefficient (b), which is a measure of prediction bias, showed values close to unity, implying there was no prediction bias. However, the characteristic soluble solids had a 'b' value of 0.74, indicating that genetic variance was overestimated (Resende et al., 2012).

Estimates of GWS accuracy ranged from 0.59 to 0.91, implying moderate-to-high accuracy. According to the classification proposed by Resende and Duarte (2007), accuracy estimates <50% are defined as low, while those between 50 and 70% are defined as moderate, and those between 70 and 90% are defined as high. Conversely, estimates of accuracy based on phenotypic selection ranged from 0.70 to 0.90, implying high accuracy. Comparison of these values shows that accuracy estimates based on phenotypic selection were superior to those obtained by GWS for all characteristics, except for the ratio of the longitudinal/transverse fruit diameter trait.

The experimental accuracy values were close to the expected accuracy values (Table 1). The formula used to obtain expected accuracy values is important in the design of GWS studies as it can be used to calculate the number of individuals and markers required to obtain the desired accuracy. Using the formula for expected accuracy, it was observed that for a characteristic with the lowest heritability (ratio of longitudinal/transverse diameter; 0.49), at least 250 individuals and 4000 markers would be necessary to obtain an accuracy value greater than 0.70. Conversely, for traits with high heritability, such as fruit mass and ratio of soluble solids/acidity, the use of 180 individuals and 1000 markers yielded a much higher accuracy of 0.81. According to Grattapaglia and Resende (2011), accuracy calculated using the deterministic approach is directly proportional to the product of heritability and the ratio between the number of individuals used and the number of QTL governing the trait. In a simulation study, these authors observed that decreased accuracy with decreasing heritability

could be compensated for by using a larger number of individuals in the training population.

The proportion of variance explained by the markers ranged from 36% (ratio of soluble solids/acidity) to 84% (ratio of longitudinal fruit diameter/transverse fruit diameter). The genetic variance not explained by these markers is, therefore, probably due to incomplete LD between these markers and the alleles causing genetic variation in this character. GWS could explain a greater proportion of the observed variance in traits compared with methods based on the identification of individual QTL. This was observed by Viana et al. (2016) who compared MAS and genomic selection in a grape breeding population consisting of 143 individuals obtained from a cross between (*Vitis rupestris* x *Vitis arizonica/girdiana*) x *Vitis vinifera.* The efficiency of genomic selection was 1.7 to 11.6-fold higher than that obtained by traditional QTL analysis.

In *Citrus* culture, the use of MAS also permitted the identification of a few QTL with major effects on traits of interest. Siviero et al. (2002) found only one QTL associated with the quantitative inheritance of characteristics (number of fruits per plant and number of seeds per fruit) in a F1 progeny derived from a cross between *Citrus sunki* and *Poncirus trifoliata* 'Rubidoux'. A study on root-rot resistance caused by *Phytophthora* in progeny derived from a cross between *C. sunki* and *P. trifoliata* (Siviero et al., 2006) provided two QTL maps for the species *P. trifoliate,* which could explain 16-24% of the genetic variation. In addition, only one QTL was mapped for the species *C. sunki*, which could only account for 14% of the variation. Asins et al. (2012) performed a QTL analysis in a segregating population of *C. grandis* and *C. clementine* to evaluate resistance to *Citrus tristeza* virus and identified a QTL that contributed to about 24% of the total genetic variance.

Considering the five markers with large effects for each trait, only 21.5-31.36% of the variance was explained. Those values only account for part of the genetic variation and further confirm the quantitative nature of these characteristics, which are governed by many genes of small effect and, probably, by some genes with moderately large effects (Kemper and Goddard, 2012).

The estimated number of QTL controlling a trait varied from 11 (ratio of longitudinal fruit diameter/transverse fruit diameter) to 219 (ratio of soluble solids/acidity) (Table 1). There have been no reports in the literature on the number of loci that control these traits, as the studies conducted so far with MAS permitted the detection of only one or more loci with major effects. These results confirm the quantitative nature of the traits studied and reiterate that the accuracy of GWS is inversely proportional to the number of QTL identified (Wong and Bernardo, 2008). In a study conducted by Resende et al. (2012) in two populations of *Eucalyptus*, 200 markers with large effects were sufficient to account for more than 80% of the heritability of the traits analyzed. An increase in the number of markers did not result in a linear increase in the accuracy of GWS (Figure 2).

The accuracy of GWS depends on the proportion of the variance that can be explained by the markers, and the accuracy of predictions of marker effects that are in LD with the QTL. Thus, the main factors that modify the accuracy of GWS are the heritability of the trait, the number of loci that control this trait and the distribution of its effects, the number of individuals in the training population, and the extent of LD between the markers and the QTL, which, in turn, depends on the effective population size and the number of markers used for the analyses (Grattapaglia and Resende, 2011).

The importance of the number of individuals used to estimate the effect of the loci is shown by the accuracy formula. Therefore, the higher the number of individuals used, the more accurate the genomic value estimates for those individuals (Hayes et al., 2009). Thus, the use of only 180 individuals to predict the effects of the markers probably contributed to

the moderate accuracy. Of note, the number of individuals used in this study was acceptable for the selection of superior clones in full-sib families since this number represents 99.45% of the genetic variability of the family. This fraction is given by $N_{ef} / 2$, where $N_{ef} = 2n / (n + 1)$ is the effective population size of a full-sib family and n is the number of individuals per family.

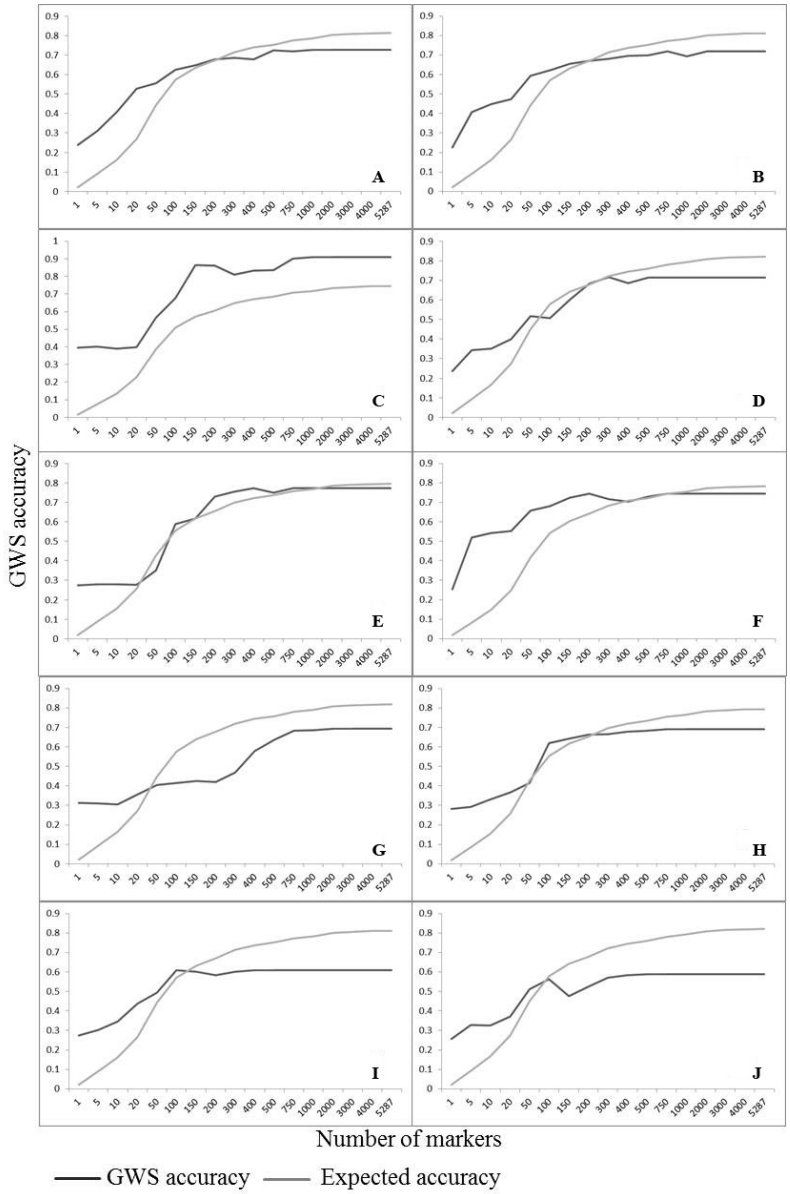

Figure 2. Accuracy of genome wide selection (GWS) and expected accuracy with an increase in the number of markers for traits evaluated in a population of *Citrus*. **A.** Longitudinal fruit diameter; **B.** transverse fruit diameter; **C.** ratio of longitudinal fruit diameter/transverse fruit diameter; **D.** fruit mass; **E.** number of fruits per box; **F.** number of segments per fruit; **G.** number of seeds; **H.** juice yield; **I.** soluble solids content; **J.** ratio of soluble solids/acidity.

According to Resende et al. (2012), when more individuals per family and more families are analyzed, the power of detection increases asymptotically and more QTL are revealed. Zhao et al. (2013) analyzed accuracy values based on the predicted effects of genetic markers in the entire maize population consisting of 11 families or those within each family. Those authors reported that the estimates of accuracy were substantially reduced when individual GV were predicted within a family compared with those predicted in the entire population. However, when the genetic effects were estimated and the GV was predicted within each family (for large families), the accuracy was satisfactory, as seen by the results presented here. Cavalcanti et al. (2012) confirmed that within families, GWS is an interesting alternative that can be used to increase the breeding efficiency of cashew. In that study, the authors obtained a high accuracy value (0.86) for a characteristic with high heritability with only 74 individuals and a limited number of markers (238).

One of the assumptions of GWS is that the phases of LD between markers and QTL are identical in the populations used for training and selection (Hayes et al., 2009). As genotyping and phenotyping were performed using individuals of one family, it is likely that they did not cover the extent and pattern of the LD present in all *Citrus* populations used in the breeding program; therefore, the use these results in other families or populations is not recommended.

The small effective size of the *Citrus* population used in this study (*Ne* = 2) and the large number of markers used have contributed favorably to the accuracy values obtained. The impact of the effective size and number of markers used to predict GWS accuracy has been demonstrated by Grattapaglia and Resende (2011) using simulated data. Resende et al. (2012) applied GWS to two *Eucalyptus* breeding populations and found that higher accuracy estimates were observed in the population that had the lowest *Ne*. This is because an increase in the effective size reduces the accuracy of genomic predictions due to an increase in the number of chromosome segments and because LD between the marker and the causative mutation only remains consistent if the two occur very close to each other on the chromosome (Kemper and Goddard, 2012).

GWS requires a high density of markers in order to maximize the number of QTL that are in LD with at least one marker (Heffner et al., 2009; Resende et al., 2014). Resende et al. (2008) demonstrated the importance of using a high density of markers to account for markers of both large and small effects, and thereby to increase the probability of explaining all of the genetic variation for the trait of interest. The results of this study are consistent with those claims since the proportion of variance explained by the markers was higher than that in studies that used a smaller number of markers. Table 2 compares the efficiency of GWS with selection based solely on phenotypic data.

When reducing the duration of the selection cycle by 25%, genetic gain with GWS was higher than that obtained by phenotypic selection for all traits, except for soluble solids and the ratio of soluble solids/acidity (Table 2). When time was reduced by 50%, GWS was superior (ranging from 31 to 160%) for all traits. When time was reduced by 75%, GWS improved (ranging from 162 to 420%). The higher gains in efficiency obtained for the ratio of longitudinal fruit diameter/transverse fruit diameter with GWS further confirm the importance of GWS in the selection of characteristics that exhibit low heritability.

Heffner et al. (2009) stated that even when only moderate accuracy is achieved using GWS, it is possible to obtain a genetic gain that is superior to that obtained by phenotypic selection, as GWS reduces the duration of the selection cycle. This was clearly observed in the present study, in which the duration of the selection cycle was reduced.

Table 2. Gains in efficiency of GWS for selection based solely on phenotypic data in a breeding population of *Citrus*.

Trait	Reduction in the selection cycle		
	25%	50%	75%
DL	0.11	0.66	2.32
DT	0.10	0.66	2.31
DL/ DT	0.73	1.60	4.20
FM	0.07	0.60	2.20
F/B	0.25	0.88	2.76
NG	0.23	0.85	2.70
NS	0.03	0.55	2.10
JY	0.12	0.68	2.37
SS	-0.07	0.40	1.80
Rt	-0.13	0.31	1.62

DL: longitudinal fruit diameter (cm); DT: transverse fruit diameter (cm); DL/DT: ratio of longitudinal fruit diameter/ transverse fruit diameter; FM: fruit mass (g); F/B: number of fruits per box; NG: Number of segments per fruit; NS: Number of seeds per fruit; JY: juice yield (mL); SS: soluble solids content; Rt: ratio of soluble solids/acidity.

Wong and Bernardo (2008) reported that GWS implementation shortened the selection cycle of oil palm from 19 to 6 years, which authors claimed could also be extrapolated to other perennial species. That study used a population of 50 individuals, which is the typical size for an oil palm breeding population. GWS yielded higher gains per unit time and reduced the costs even with this small number of individuals. Importantly, GWS has been shown to be a promising tool for genetic improvement when applied to traits of economic interest that are difficult to measure and/or have a high cost of evaluation (tolerance to cold, drought, disease resistance) (Grattapaglia and Resende, 2011). These kinds of traits should be studied using genomic selection in *Citrus* breeding.

The results presented here suggest that GWS can be useful for *Citrus* breeding as it can predict the phenotypes accurately and timeline, and reduce the length of the selection cycle. Thus, GWS is a promising tool for improving *Citrus* culture. This may be the first study to demonstrate the feasibility of genomic selection in *Citrus*.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for the research fellowship of the first author. Research supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (Processes #2007/08435-5 and #2011/18605-0) and Instituto Nacional de Ciência e Tecnologia (INCT) de Genômica para Melhoramento de Citros (Process #573848/2008-4).

REFERENCES

Asins MJ, Fernandez-Ribacoba J, Bernet GP, Gadea J, et al. (2012). The position of the major QTL for Citrus tristeza virus resistance is conserved among *Citrus grandis, C. aurantium* and *Poncirus trifoliata. Mol. Breed.* 29: 575-587. http://dx.doi.org/10.1007/s11032-011-9574-x

Cavalcanti JJV, Resende MDV, Santos FHC and Pinheiro CR (2012). Predição simultânea dos efeitos de marcadores moleculares e seleção genômica ampla em cajueiro. *Rev. Bras. Frutic.* 34: 840-846. http://dx.doi.org/10.1590/S0100-29452012000300025

Daetwyler HD, Pong-Wong R, Villanueva B and Woolliams JA (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185: 1021-1031. http://dx.doi.org/10.1534/genetics.110.116855

Daetwyler HD, Calus MPL, Pong-Wong R, de Los Campos G, et al. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193: 347-365. http://dx.doi.org/10.1534/genetics.112.147983

Endelman JB (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4: 250-255. http://dx.doi.org/10.3835/plantgenome2011.08.0024

Gmitter Junior FG, Chen C, Machado MA, Souza AA, et al. (2012). *Citrus* genomics. *Tree Genet. Genomes* 8: 611-626. http://dx.doi.org/10.1007/s11295-012-0499-2

Goddard ME, Hayes BJ and Meuwissen THE (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed. Genet.* 128: 409-421. http://dx.doi.org/10.1111/j.1439-0388.2011.00964.x

Grattapaglia D and Resende MDV (2011). Genomic selection in forest tree breeding. *Tree Genet. Genomes* 7: 241-255. http://dx.doi.org/10.1007/s11295-010-0328-4

Gussen O, Uzun A, Seday U and Kafa G (2011). QTL analysis and regression model for estimating fruit setting in young *Citrus* trees based on molecular markers. *Sci. Hortic. (Amsterdam)* 130: 418-424. http://dx.doi.org/10.1016/j.scienta.2011.07.010

Hayes BJ, Bowman PJ, Chamberlain AJ and Goddard ME (2009). Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92: 433-443. http://dx.doi.org/10.3168/jds.2008-1646

Heffner EL, Sorrells ME and Jannink JL (2009). Genomic selection for crop improvement. *Crop Sci.* 49: 1-12. http://dx.doi.org/10.2135/cropsci2008.08.0512

Henderson CR (1973). Maximum likelihood estimation of variance components. Unpublished manuscripts, Animal Science Dept., Cornell University.

Ito TM, Polido PB, Rampim MC, Kaschuk G, et al. (2014). Genome-wide identification and phylogenetic analysis of the AP2/ERF gene superfamily in sweet orange (*Citrus sinensis*). *Genet. Mol. Res.* 13: 7839-7851. http://dx.doi.org/10.4238/2014.September.26.22

Iwata H, Hayashi T, Terakami S, Takada N, et al. (2013). Potential assessment of genome-wide association study and genomic selection in Japanese pear *Pyrus pyrifolia. Breed. Sci.* 63: 125-140. http://dx.doi.org/10.1270/jsbbs.63.125

Jaccoud D, Peng K, Feinstein D and Kilian A (2001). Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Res.* 29: E25. http://dx.doi.org/10.1093/nar/29.4.e25

Jarrell DC, Roose ML, Traugh SN and Kupper RS (1992). A genetic map of citrus based on the segregation of isozymes and RFLPs in an intergeneric cross. *Theor. Appl. Genet.* 84: 49-56. http://dx.doi.org/10.1007/BF00223980

Kemper KE and Goddard ME (2012). Understanding and predicting complex traits: knowledge from cattle. *Hum. Mol. Genet.* 21 (R1): R45-R51. http://dx.doi.org/10.1093/hmg/dds332

Kumar S, Bink MCAM, Volz RK, Bus VGM, et al. (2012). Towards genomic selection in apple (*Malus* x *domestica* Borkh.) breeding programmes: Prospects, challenges and strategies. *Tree Genet. Genomes* 8: 1-14. http://dx.doi.org/10.1007/s11295-011-0425-z

Lande R and Thompson R (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124: 743-756.

Legarra A, Robert-Granié C, Manfredi E and Elsen JM (2008). Performance of genomic selection in mice. *Genetics* 180: 611-618. http://dx.doi.org/10.1534/genetics.108.088575

Machado MA, Cristofani-Yaly M and Bastianel M (2011). Breeding, genetic and genomic of citrus for disease resistance. *Rev. Bras. Frutic.* 33: 158-172. http://dx.doi.org/10.1590/S0100-29452011000500019

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

Misztal I, Legarra A and Aguilar I (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.* 92: 4648-4655. http://dx.doi.org/10.3168/jds.2009-2064

Patterson HD and Thompson R (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58: 545-554. http://dx.doi.org/10.1093/biomet/58.3.545

R Development Core Team (2012). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Resende MDV (2002). Genética Biométrica e estatística no melhoramento de plantas perenes. Brasília: Embrapa Informação tecnológica.

Resende MDV and Duarte JB (2007). Precisão e controle de qualidade em experimentos de avaliação de cultivares. *Pesqui. Agropecu. Trop.* 37: 182-194.

Resende MDV, Lopes OS, Silva RL and Pires IE (2008). Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. *Pesq. Flor. Bra.* 56: 63-77.

Resende MDV, Resende MFR, Jr., Sansaloni CP, Petroli CD, et al. (2012). Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytol.* 194: 116-128. http://dx.doi.org/10.1111/j.1469-8137.2011.04038.x

Resende MDV, Silva FF, Resende MFR, Junior. and Azevedo CF (2014). Genome-wide selection. In: Biotechnology and Plant Breeding (Borem A and Fritsche-Neto R, eds.). Elsevier.

Resende Júnior MFR, Muñoz P, Resende MDV, Garrick DJ, et al. (2012). Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics* 190: 1503-1510. http://dx.doi.org/10.1534/genetics.111.137026

Siviero A, Cristofani M, Boava LP and Machado MA (2002). Mapeamento de QTLs associados à produção de frutos e sementes em híbridos de *Citrus sunki vs Poncirus trifoliata. Rev. Bras. Frutic.* 24: 741-743. http://dx.doi.org/10.1590/S0100-29452002000300045

Siviero A, Cristofani M, Furtado EL, Garcia AAF, et al. (2006). Identification of QTLs associated with citrus resistance to Phytophthora gummosis. *J. Appl. Genet.* 47: 23-28. http://dx.doi.org/10.1007/BF03194595

Talon M and Gmitter Junior FG (2008). *Citrus* genomics. *Intern. J. Plant Genomics*: 1-17.

Viana AP and Resende MDV (2014). Seleção Genômica Ampla (GWS). In: Genética Quantitativa no Melhoramento de Fruteiras (Viana AP, Resende MDV, eds.). Interciência, Rio de Janeiro.

Viana AP, Resende MDV, Riaz S and Walker MA (2016). Genome selection in fruit breeding: application to table grapes. *Sci. Agric.* 73: 142-149. http://dx.doi.org/10.1590/0103-9016-2014-0323

Wong CK and Bernardo R (2008). Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theor. Appl. Genet.* 116: 815-824. http://dx.doi.org/10.1007/s00122-008-0715-5

Zapata-Velenzuela J, Whetten RW, Neale D, Mckeand S, et al. (2013). Genomic estimated breeding values using genomic relationship matrices in a cloned population of Loblolly Pine. *Genes Genom. Genet* 3: 909-916.

Zhao Y, Gowda M, Liu W, Wurschum T, et al. (2013). Choice of shrinkage parameter and prediction of genomic breeding values in elite maize breeding populations. *Plant Breed.* 132: 99-106. http://dx.doi.org/10.1111/pbr.12008

Zhong S, Dekkers JCM, Fernando RL and Jannink JL (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. *Genetics* 182: 355-364. http://dx.doi.org/10.1534/genetics.108.098277