Genomic prediction with allele dosage information in highly polyploid species

Lorena G. Batista¹ · Victor H. Mello¹ · Anete P. Souza² · Gabriel R. A. Margarido¹

Received: 24 June 2021 / Accepted: 6 November 2021 / Published online: 20 November 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Key message Including allele, dosage can improve genomic selection in highly polyploid species under higher frequency of different heterozygous genotypic classes and high dominance degree levels.

Abstract Several studies have shown how to leverage allele dosage information to improve the accuracy of genomic selection models in autotetraploid. In this study, we expanded the methodology used for genomic selection in autotetraploid to higher (and mixed) ploidy levels. We adapted the models to build covariance matrices of both additive and digenic dominance effects that are subsequently used in genomic selection models. We applied these models using estimates of ploidy and allele dosage to sugarcane and sweet potato datasets and validated our results by also applying the models in simulated data. For the simulated datasets, including allele dosage information led up to 140% higher mean predictive abilities in comparison to using diploidized markers. Including dominance effects were highly advantageous when using diploidized markers, leading to mean predictive abilities which were up to 115% higher in comparison to only including additive effects. When the frequency of heterozygous genotypes in the population was low, such as in the sugarcane and sweet potato datasets, there was little advantage in including allele dosage information in the models. Overall, we show that including allele dosage can improve genomic selection in highly polyploid species under higher frequency of different heterozygous genotypic classes and high dominance degree levels.

Introduction

Polyploids are organisms with more than two sets of chromosomes. The number of sets of chromosomes in an organism is named its ploidy level. Polyploids are classified into two major categories of auto and allopolyploids. Allopolyploids result from the combination of distinct parental genomes and are characterized by preferential pairing of chromosomes, with bivalent chromosome formation in meiosis and disomic inheritance at each locus. In contrast, autopolyploids have more than two homologs per homology group, often leading to the formation of multivalent chromosomes and polysonic inheritance (Soltis and Soltis 2000).

Many economically important species are autopolyploids. Among these, a high ploidy level (> 4) is observed in a number of species such as sweet potato, sugarcane, and some ornamental flowers and forage crops. Sweetpotato, an autohexaploid, is the fourteenth most important food crop in the world regarding production volume (FAOSTAT, 2020), and sugarcane, with ploidy levels ranging up to 16 (Garcia et al. 2013), accounts for 80% of the worldwide sugar production (CIRAD) and has potential to become the main crop for bioenergy production. The main bottleneck in breeding programs for these species is the long process for selection of cultivars. A traditional sugarcane breeding program is usually divided into several phases of selection, each consisting of large experiments that are usually conducted for more than one crop cycle (Cheavegatti-Gianotto et al. 2011; Zhou 2013), taking up to 12 years from the initial crosses until commercial cultivar release (Park et al. 2007). Sweetpotato breeding programs follow a similar breeding scheme, with selection of cultivars taking up to 10 years (Katayama et al. 2017). In this context, there is a pressing need for the deployment of strategies to reduce experimental costs and time for selection of cultivars.
Genomic selection is a viable way of achieving improvements in breeding programs in terms of time and costs (Hefner et al. 2009a). Genomic selection consists of using a representative population that is both genotyped and phenotyped (i.e., the training population) to predict the effect of genetic markers widely spread throughout the genome. The predicted effects are then used to predict the breeding or genotypic value of genotyped individuals (Meuwissen et al. 2001). This allows selection to be carried based on predicted breeding values, reducing the need for further costly phenotypic evaluations and shortening the time needed for selection of the best genotypes. Genomic selection has been successfully implemented in several crop breeding programs (Bernardo and Yu 2007; Hefner et al. 2009a; Crossa et al. 2010; Resende et al. 2012; Duhnen et al. 2017) and can potentially increase genetic gain in sugarcane and sweet potato breeding programs (Voss-Fels et al. 2021; Hayes et al. 2021). Although genomic selection can greatly improve breeding programs, its implementation demands a relatively large set of genetic markers to be consistently obtained at feasible costs, a process which is hindered in complex genomes such as those of highly autopolyploid species.

Due to the complexity of their genomes, genetic studies in autopolyploid species were historically mostly carried using either dominant or diploidized markers (Dufresne et al. 2014), that is, polymorphisms that are either detected in a presence/absence fashion or polymorphisms where all heterozygous genotypes are collapsed into a single class. When using only dominant or diploidized markers, information on the different categories of heterozygous genotypes is effectively lost. However, several new tools are now available that allow estimating the allele dosage (i.e., the quantitative genotypes) of markers (Serang et al. 2012; Blischak et al. 2018; Gerard et al. 2018; Clark et al. 2019), and information of all possible genotypic classes can now potentially be used in genomic studies of polyploids.

In autotetraploid, several studies have shown how to leverage allele dosage information to improve the accuracy of genomic selection models (Slater et al. 2016; de Bem Oliveira et al. 2018; Hawkins and Yu 2018; Endelman et al. 2018; Amadou et al. 2020). However, to our knowledge, no studies so far have expanded these methodologies to specifically address organisms with higher ploidy levels. In this study, our objective was to determine whether the inclusion of allele dosage information improves genomic prediction in highly autopolyploid species. In order to achieve that, we generalized genomic prediction models used in autotetraploid and assess the accuracy of genome-wide prediction when incorporating allele dosage information in sugarcane and sweet potato datasets, two highly autopolyploid species. In order to validate our results, we also assess the accuracy of prediction in four simulated datasets.

Material and methods

Genetic material and field experiments

The sugarcane dataset consisted of a segregating F1 progeny of 179 individuals derived from the crossing of two commercial cultivars, IACSP95-3018 (female) and IACSP93-3046 (male). IACSP95-3018 is a promising clone used as a parent in the breeding program at IAC (Instituto Agronômico de Campinas), and IACSP93-3046 has a high level of sucrose, good tillering, and an erect stool habit, being recommended for mechanical harvest. The first field experiment was set in Sales de Oliveira, SP, Brazil, in 2007. A randomized complete block design with four replicates was used and evaluations were carried in the harvest years of 2008 (plant cane) and 2009 (ratoon cane). The full-sib progeny was then clonally propagated for the second field experiment that was set in Ribeirão Preto, SP, Brazil, in 2011. A randomized complete block design with three replicates was used and evaluations were carried in 2012 (plant cane), 2013, and 2014 (ratoon cane), each year corresponding to a different harvest. Both parents were included in each block of the two experiments. All replicates were used to collect phenotypes for stalk diameter (cm), stalk height (cm) and stalk weight (kg) in both experiments. Also, two blocks in each experiment were used to collect phenotypes for soluble solids content (Brix), sucrose content, and fiber percentage.

The sweet potato dataset consisted of phenotypic records on 282 accessions of Ipomoea batatas made available by Jackson et al. (2018), which are part of a broader group of 731 accessions randomly selected from the USDA germplasm bank in Griffin, Georgia, United States. For this study, we chose the accessions based on the availability of genotypic data. These materials have origin in more than 30 countries in eight geographic regions (Africa, Australia, Caribbean, Central America, East Asia, North America, Pacific Islands, and South America). The accessions were planted in field trials and phenotyped in the years 2012, 2013, and 2014. Each sweet potato entry was planted in two replications of a single row, five-plant plots arranged in a randomized complete block experimental design. In this study, we used phenotypic data from the stele colorimetry analysis. The stele colorimetry data included values of the green–red coordinate (a), the yellow-blue coordinate (b), color saturation (C), lightness (L), and hue angle (h).

Distribution and correlation between phenotypic measurements for sugarcane and sweet potato can be found in Fig S1.1 and Fig S1.2 of Supplementary Material 1.

Genotyping

For the sugarcane population, parents, and F1 progeny were genotyped using the genotyping-by-sequencing protocol of Elshire et al. (2011). Reduced representation libraries were...
prepared using the PstI restriction enzyme. PstI is a rare-cutting enzyme, because its restriction site has a length of 6 bp, allowing lower genomic coverage and a higher genotyping depth (Poland and Rife 2012). Four lanes containing 96-plex libraries were sequenced using the Illumina GAIIx and, subsequently, another four lanes with the same 96-plex libraries were sequenced using the Illumina NextSeq500 platform.

The genotyping-by-sequencing protocol used for the sweet potato accessions is described by Wadl et al. (2018), where a modified genotyping-by-sequencing protocol optimized for highly heterozygous and polyploid genomes was used (GBSPoly). They used a combination of CviAII and TseI restriction enzymes for preparing the libraries (restriction sites with 4 and 5 bp, respectively). Libraries were multiplexed with 96 pooled samples. In this study, we used the raw read data the authors in Wadl et al. (2018) made available in the NCBI database with accession code SRP152827.

For the sugarcane dataset, we called variants using a modified version of the TASSEL-GBS pipeline (Pereira et al. 2018). This version provides exact read counts of the alleles for each single nucleotide polymorphism (SNP). We used default values in all plugins of the pipeline, except for the MergeDuplicateSNPs plugin, in which we used the argument callHets and set the misMat argument value to 0.3. These values were chosen to allow a greater number of heterozygous SNPs to be kept in subsequent steps. The sequenced reads were then aligned to the methyl-filtrated assembly of the sugarcane genome (Grativol et al. 2014), using the software Bowtie2 v2.3.5 (Langmead and Salzberg 2012).

The sweet potato raw reads were first aligned to the two ancestral reference genomes I. trifida and I. triloba (Shiotani 1988; Oracion et al. 1990; Freyre et al. 1991) using Bowtie2 v2.3.5 (Langmead and Salzberg 2012). We then used the HaplotypeCaller tool in the GATK software (version 4.1.4) to call SNPs, indels, and copy number variants. Only SNPs were kept in subsequent steps.

For both species, we used the read count information of each SNP to estimate their ploidy level and call sample genotypes using the software SuperMASSA and VCF2SM (Serang et al. 2012; Pereira et al. 2018). For sugarcane, ploidy levels ranging from two to 20 were evaluated and only SNPs with ploidy estimates between six and 14 were kept (Garcia et al. 2013). We also filtered for a minimum mean read depth per individual of 50 reads, maximum mean read depth per individual of 500 reads, minimum posterior probability of genotype configuration of 0.8, minimum posterior probability of each genotype assignment of 0.5, and minimum call rate of 50%. For sweet potato, ploidy levels ranging from four to eight were evaluated and only SNPs with a ploidy estimate of six were chosen. We used a minimum mean read depth per individual of 45 reads, maximum mean read depth per individual of 200 reads and the remaining arguments were the same as for sugarcane.

We used the R package updog (Gerard et al. 2018) to reestimate the genotypes of the SNPs that met the filtering criteria in both species. The updog package has the advantage of accounting for allelic bias, overdispersion, and sequencing errors when estimating SNP genotypes, given a predetermined ploidy level. For sweet potato, SNP sets resulting from the alignment with each of the reference genomes were merged, and redundant SNPs (i.e., SNPs that had genotype calls identical to another SNP for all individuals) were removed.

Finally, we performed a chi-squared segregation test on the population genotype class frequencies. For the sugarcane F1 progeny, based on the estimates of SNP genotypes in the parents, we tested the goodness-of-fit of marker genotypes to a hypergeometric distribution of gametes (Mollinari and Serang 2015). For the sweet-potato diversity panel, we tested the goodness-of-fit of marker genotypes to the distribution expected under Hardy–Weinberg equilibrium. Using the Bonferroni correction with a global alpha value of 0.05, we selected only SNPs for which the null hypothesis of adherence to the expected distributions was not rejected.

**Phenotypic mixed model analysis**

Adjusted phenotypic means (i.e., BLUEs—best linear unbiased estimates) for each individual were obtained using a two-stage analysis (Damesa et al. 2017). All analyses for the sugarcane and sweet potato datasets were performed using ASReml-R v4.1.0.160 (Butler et al. 2009). Stage one consisted of a within-site (sugarcane) or within-year (sweet potato) analysis, where the genotype effect was considered fixed. The remaining effects were considered as random (sweet potato: replicates—sugarcane: harvest effects, blocks-within-harvest effects, and genotype×harvest interaction effects). The covariance matrix (Ω) for the vector of genotype effects (u g) in site j was obtained from the inverse of the coefficient matrix of the mixed model equations, returned as $\hat{C}_{fixed}$ in the asreml object (Endelman et al. 2018). Stage two was a joint analysis considering the two sites (sugarcane) or the three years (sweet potato), using the following linear models for the sugarcane and sweet potato datasets, respectively:

$$\hat{u}_{ij} = \mu + g_i + s_j + (gs)_{ij} + e_{ij},$$

$$\hat{u}_{ij} = \mu + g_i + t_j + (gt)_{ij} + e_{ij},$$

where $\hat{u}_{ij}$ is the genotype effect estimate obtained in the stage one analyses, the parameter $\mu$ is the intercept, $g_i$ is a fixed effect of genotypes, $s_j$ is a random effect of sites, $(gs)_{ij}$ is a random effect for the genotype×site, $t_j$ is a random effect of years, $(gt)_{ij}$ is a random effect for the genotype×year interaction, and the variance of the residual $e_{ij}$ is $(ao^2)^{-1}$, where $ao^2$ is the $i$th diagonal element of $\hat{C}_{fixed}$ from the stage one analysis (Damas e et al. 2017). The BLUEs of the
genotypes obtained after this stage were subsequently used to fit the genomic selection models.

**Phenotypic variance partitioning**

For phenotypic variance partitioning and to obtain estimates of heritability for the traits in the sugarcane dataset, the phenotypic values were first standardized and the following random linear model was used for each trait:

\[ y_{ijkl} = \mu + g_i + s_j + h_k + b_{ijk} + (gs)_{ij} + (gh)_k + (gsh)_{ijk} + e_{ijkl}, \]

where \( \mu \) is the intercept, \( g_i \) is the effect of genotypes, \( s_j \) is the effect of sites, \( h_k \) is the effect of harvest, \( b_{ijk} \) is the effect of replicates within sites and harvests, \( (gs)_{ij} \) is the effect of the genotype × site interaction, \( (gh)_k \) is the effect of the genotype × harvest interaction, \( (gsh)_{ijk} \) is the effect of the genotype × site × harvest interaction, and \( e_{ijkl} \) is the residual effect.

The estimates of heritability were obtained using:

\[ h^2 = \frac{\sigma^2_{g} + \sigma^2_{g} + \sigma^2_{g} + \sigma^2_{e}}{\sigma^2_{g} + \sigma^2_{s} + \sigma^2_{h} + \sigma^2_{e}}, \]

where \( \sigma^2_{g} \), \( \sigma^2_{s} \), \( \sigma^2_{h} \), and \( \sigma^2_{e} \) are the genotypic variance, the variance of the genotype × site interaction, the variance of the genotype × harvest interaction, and the residual variance, respectively. The values \( s, h, c_{sh}, r \) correspond to the number of sites, the number of harvests, the number of combinations of sites and harvests, and the total (combined) number of replicates of both experiments, respectively.

For sweet potato, the following random linear model was used for each trait:

\[ y_{ijkl} = \mu + g_i + tj + b_{ij} + (gt)_{ij} + e_{ijkl}, \]

where \( \mu \) is the intercept, \( g_i \) is the effect of genotypes, \( t_j \) is the effect of years, \( b_{ij} \) is the effect of replicates within years, \( (gt)_{ij} \) is the effect of the genotype × year interaction, and \( e_{ijkl} \) is the residual effect.

The estimates of heritability were obtained using:

\[ h^2 = \frac{\sigma^2_{g} + \sigma^2_{t} + \sigma^2_{e}}{\sigma^2_{g} + \sigma^2_{t} + \sigma^2_{e}}, \]

where \( \sigma^2_{g} \), \( \sigma^2_{t} \), and \( \sigma^2_{e} \) are the genotypic variance, the variance of the genotype × year interaction, and the residual variance, respectively. The values \( r \) and \( h \) correspond to the number of years total (combined) number of replicates in all years, respectively.

**Genomic selection models**

We incorporated allele dosage information in our genomic selection models by expanding and adapting the GBLUP methodology for autotetraploid species proposed by Endelman et al. (2018). In sugarcane, besides the higher ploidy, the model also has to account for different ploidy levels among SNPs. In order to achieve that, we expanded the theory by adapting the estimation of the genomic covariance matrix of both the additive values (\( G \)) and the digenic dominance values (\( D \)).

Genomic predictions were obtained using the following linear model:

\[ \hat{g}_i = \mu + a_i + e_i, \]

where \( \hat{g}_i \) is the BLUE of the \( i \)th individual obtained with the two-stage phenotypic analysis, \( \mu \) is the intercept, \( a_i \) is the random effect of genotypes, and \( e_i \) is the random residual effect.

We used two covariance structures in the genomic selection model: i) \( IV_r + GV_a \), and ii) \( IV_r + GV_a + DV_d \), where \( I \) is the identity matrix, \( V_r \) is the residual variance, \( V_a \) is the additive genetic variance, and \( V_g \) is the dominance genetic variance. All analyses were performed using ASReml-R v4.1.0.160 (Butler et al. 2009).

**Genomic covariance matrix of additive values (\( G \))**

Consider a matrix \( X \) with \( n \) rows and \( m \) columns, the rows corresponding to the individuals in the population and the columns corresponding to SNPs, where each element \( x_{ij} \) corresponds to the dosage of the alternative allele for the \( j \)-th SNP in the \( i \)-th individual. If \( p_j \) is the frequency of the alternative allele at the \( j \)-th locus, we can obtain an \( n \times m \) matrix \( P \) where the values in the \( j \)-th column all correspond to \( p_j \). For hexaploid sweet potato, subtracting 6 from \( p_j \) results in the matrix \( W \) of centered genotypes. The \( G \) matrix is then obtained by the formula:

\[ G = \frac{WW^T}{\sum_j 6p_j(1 - p_j)} \]

For sugarcane, because the SNPs have different ploidy levels, the same value of allele dosage for one SNP does not represent the same genotype for other SNPs with different ploidy. For example, for a hexaploid SNP, an allele dosage of six represents a homozygous genotype, while for an octoploid SNP the same value represents one of the possible heterozygotes.

To account for the different ploidy levels between SNPs, we used the following formula:

\[ Z = 2X\Phi^{-1}, \]

where \( \Phi \) is an \( m \times m \) diagonal matrix of ploidy values, such that each diagonal element \( \varphi_j \) corresponds to the ploidy of the \( j \)-th SNP. The resulting matrix \( Z \), with the same dimensions of \( X \), has all its elements varying from 0 to 2, where 0 represents loci that are homozygous for the reference allele and 2 represents loci that are homozygous for the alternative allele, the values in between corresponding to heterozygous loci.

The subsequent steps to obtain \( G \) are the same as for diploids (VanRaden 2008a, b). Subtracting 2\( P \) from \( Z \) results in...
the matrix \( W \) of centered genotypes. The \( G \) matrix is then obtained by the formula:

\[
G = \frac{WW^T}{\sum_j 2p_j(1 - p_j)}
\]

Genomic covariance matrix of digenic dominance values (\( D \))

We first introduce the expansion of the digenic dominance values in the autotetraploid model to a hexaploid scenario. Higher ploidy levels can be parametrized in a similar fashion. Considering a hexaploid locus with two alleles \( B \) and \( b \), the digenic effect for each allele pair can be obtained as demonstrated by Endelman et al. (2018), with the following set of equations:

\[
\begin{align*}
\beta_{BB} &= q^2 \beta \\
\beta_{Bb} &= -pq \beta \\
\beta_{bb} &= p^2 \beta,
\end{align*}
\]

where \( p \) is the allele frequency of \( B \), \( q \) is the allele frequency of \( b \), with \( q = 1 - p \), and \( \beta \) is the digenic dominance effect. Also, we have that:

\[
\beta = \beta_{BB} - 2\beta_{Bb} + \beta_{bb}.
\]

For a hexaploid locus, seven genotypic classes are possible in a population (i.e., allele dosages ranging from 0 to 6). For each genotypic class, different combinations of digenic effects are present. For example, for the genotypic class \( BBbBbb \), there are 6 possible combinations of two \( B \) alleles, 8 possible combinations of a \( B \) allele with a \( b \) allele, and 1 possible combination of two \( b \) alleles. By replacing each digenic effect with their corresponding values in (Eq. 1), we obtain the total digenic dominance coefficient for each genotype class. Table 1 shows the combinations of digenic effects and the total digenic dominance coefficient for each possible genotype class of a hexaploid locus.

The formula to obtain the total digenic dominance for a given biallelic hexaploid locus can then be generalized as:

\[
\delta = \left( 15p^2 - 5xp + \frac{1}{2}x(x - 1) \right) \beta,
\]

where \( \delta \) is the total digenic dominance and \( x \) is the dosage of the \( B \) allele.

We used the same process described for hexaploid loci to obtain equations for other levels of ploidy. Table 2 shows the generalized formulas to obtain the total digenic dominance for even ploidies from six through 14.

The formulas in Table 2 can then be generalized as:

\[
Q\beta = \left( (P \odot P)C - P(\Phi - 1) \right) \odot X + \frac{1}{2}X \odot (X - 1) \beta
\]

Formulas for the total digenic dominance for different levels of ploidy

Table 2

| Ploidy | Total digenic dominance |
|-------|-------------------------|
| 6     | \((15p^2 - 5xp + \frac{1}{2}x(x - 1))\beta\) |
| 8     | \((28p^2 - 7xp + \frac{1}{2}x(x - 1))\beta\) |
| 10    | \((45p^2 - 9xp + \frac{1}{2}x(x - 1))\beta\) |
| 12    | \((66p^2 - 11xp + \frac{1}{2}x(x - 1))\beta\) |
| 14    | \((91p^2 - 13xp + \frac{1}{2}x(x - 1))\beta\) |

where \( \odot \) represents the Hadamard product, \( C \) is an \( m \times m \) diagonal matrix where each diagonal element \( c_j \) corresponds to \( \left( \frac{q_j}{2} \right) \), and \( P \), \( \Phi \) and \( X \) are as previously defined. For sugarcane, to account for different ploidy levels the matrix \( Z \) replaces \( X \) in the formula and the matrix \( \Phi \) and \( C \) have \( q_j = 2 \) for every \( j \).

Finally, the genomic covariance matrix of digenic dominance values (\( D \)) was obtained with:

\[
D = \sum_{QQ'}c_{QQ'}(1 - p_j)^2.
\]

Model and marker set comparisons

We compared two models for the genotype effects, one using only the additive \( G \) matrix (G model) and one using both the \( G \) and \( D \) matrices (G + D model). We also investigated the effect of using two different sets of genotypic information: (i) a fully informative model considering SNP markers with ploidy and allele dosage estimates, and (ii) diploidized SNP markers. The diploidized SNP set was obtained by setting the values of all heterozygous loci in \( X \) to 1 and homozygous genotypes for the reference and alternative alleles to 0 and 2, respectively. By doing so, all heterozygous genotypes were effectively merged in a single class, regardless
of their dosage. As a consequence, large deviations from Hardy–Weinberg Equilibrium (HWE) are likely to be present on diploidized data. In order to account for that, for diploidized markers, the \( G \) and \( D \) matrices were obtained using the Natural and Orthogonal Interactions approach used for diploids (Vitezica et al. 2017). We chose this methodology as it generalizes orthogonality even to populations that are not in HWE.

The models were compared in terms of predictive ability. For that, 1,000 cross-validation runs were carried out, such that in each run 10% of the population was sampled and used as the validation set, while the remaining 90% were used as the training set (Stone 1974). We measured predictive ability as the correlation between predicted genotypic values and BLUEs of the individuals in the validation set.

**Simulated datasets**

**Population structure and founder genotypes**

Stochastic simulations of two populations with different genotype class distributions were used to validate the accuracy of prediction of genomic selection models using allele dosage estimates for additive and dominance effects. One population was simulated with a nearly uniform distribution of all possible genotypic classes (Population 1). The second population was simulated with a higher frequency of simplex and homozygous genotypes which, in consequence, resulted in a higher prevalence of rare alleles (Population 2).

Genome simulation parameters were chosen to match the sweet potato genome. An autohexaploid genome consisting of 90 chromosomes (15 homology groups) was simulated and these chromosomes were assigned a genetic length of 1.43 Morgans and a physical length of \( 2 \times 10^7 \) base pairs (Wu et al. 2018). Sequences for each chromosome were generated using the Markovian Coalescent Simulator (Chen et al. 2009) and AlphaSimR v1.0.3 (Gaynor et al. 2021). Recombination rate was inferred from genome size (i.e. 1.43 Morgans / \( 2 \times 10^7 \) base pairs = \( 7.15 \times 10^{-9} \) per base pair), and the mutation rate was set to \( 2 \times 10^{-9} \) and \( 2 \times 10^{-7} \) per base pair and effective population size were set to 50 for Populations 1 and 2, respectively. The probability of quadrivalent formation was set to 0.15 (Mollinari et al. 2020).

Simulated genome sequences were used to produce 50 founder genotypes. This was accomplished by randomly sampling gametes from the simulated genome to assign as sequences for the founders. Sites that were segregating in the founders’ sequences were randomly selected to serve either as causal loci or markers. For Population 1 we simulated a total of 1,000 segregating sites per homology group, of which 250 were selected as causal loci and 750 were selected as markers (3,750 causal loci and 11,250 markers in total). For Population 2 we simulated a total of 6,330 segregating sites per homology group, of which 1,580 were selected as causal loci and 4,750 were selected as markers (23,700 causal loci and 71,250 markers in total). Out of the 23,700 causal loci, a total of 3,750 sites with a high frequency of simplex and homozygous genotypes in the population were selected. The causal loci were selected by setting the additive allele effect of non-selected causal loci as zero, which would in turn also zero the dominance effects as detailed in the next section. Out of the 71,250 markers, a total of 11,250 sites with a high frequency of simplex and homozygous genotypes in the population were selected. The allele frequencies and genotype distribution of QTL and markers in both populations are shown in Fig S1.1 and Fig S1.2 of Supplementary Material 1.

**Phenotype simulation**

AlphaSimR defines an individual’s raw genotype dosage \( x \) as the number of copies of the “1” allele at a locus, which is then scaled in accordance with the ploidy level. The scaled dosages make inputs in the package invariant to ploidy level. The scaled additive genotype dosages \( x_A \) are given by the formula:

\[
x_A = \left( x - \frac{ploidy}{2} \right) \left( \frac{2}{ploidy} \right)
\]

And the scaled dominance genotype dosages \( x_D \) are given by the formula:

\[
x_D = x(ploidy - x) \left( \frac{2}{ploidy} \right)^2
\]

For autopolyploid organisms, this scaled dominance genotype dosage is consistent with a digenic dominance model (i.e. no higher-order dominance effects are simulated) (Gaynor et al. 2021).

The true additive value of the simulated trait is then determined by the summing of its causal loci additive allele effects multiplied by the scaled additive genotype dosages. Additive allele effects were sampled from a standard normal distribution.

In the same way, the true dominance value of the simulated trait is determined by the summing of its causal loci dominance allele effects multiplied by the scaled dominance genotype dosages. The dominance effect \( d \) at a locus is the dominance degree \( \delta \) at that locus times the absolute value of its additive allele effect \( a \):

\[
d = \delta |a|
\]

In this study, the dominance degrees were sampled from a normal distribution with variance 0.2 (Werner et al. 2020) and a mean of either 0.3 (low dominance) or 1 (high
dominance). The additive and dominance effects were then scaled to achieve a desired genotypic variance of 1.

A phenotype was then simulated by summing the additive and dominance values and subsequently adding random error in order to achieve a heritability of 0.5.

**Population simulation**

For each population structure (Populations 1 and 2) and dominance level (low dominance and high dominance), we simulated F1 populations with 300 individuals formed by randomly crossing the founder genotypes. Each of the four simulation scenarios (two populations x two dominance degree levels) was replicated 20 times. For each replicate, we deployed genomic selection models using a k-fold cross-validation scheme with k = 10. We measured predictive ability as the correlation between true and estimated genotypic values in the validation set.

Additionally, for Population 1 we simulated datasets with the same population size as the sugarcane dataset (170). For this scenario, we used a high dominance degree level and heritability close to the average heritability we found for sugarcane traits (0.3). This scenario was also replicated 20 times.

**Results**

We were able to obtain a large number of SNPs with estimates of ploidy and allele dosage in both sugarcane and sweet potato. In both species, most of the genotypes were either homozygous or had only one copy of the alternative allele. The genomic selection models showed low prediction ability in the sugarcane dataset and moderate to high predictive ability in the sweet potato dataset. Overall, the prediction ability values in both datasets showed little sensitivity to including ploidy and allele dosage information or dominance effects in the model. These results were replicated in simulated datasets where the marker genotype distribution was similar to the real datasets (Population 2). In other simulated populations, which had a higher frequency of heterozygous markers (Population 1), the highest values of predictive ability were achieved when including ploidy and allele dosage information in the models (full ploidy models). In Population 1, including digenic dominance effects in full ploidy models was only advantageous when the dominance level was high. When using diplodized markers, including dominance effects increased predictive ability regardless of the dominance level.

**Genotyping**

In sugarcane, a total of 6,589 SNPs were kept after filtering for mean read depth, posterior probability of genotypes and ploidy estimates, call rate, and segregation distortion in the progeny. A total of 11 individuals did not have any sequenced reads and were considered not genotyped, thus being used in phenotypic analyses but not for genomic selection. A summary of ploidy and allele dosage estimates of the SNPs is shown in Fig. 1. The majority of the SNPs had ploidy estimates of ten (31.18%) and eight (28.93%), followed by 17.88% of SNPs with ploidy estimates of 12, 15.59% with an estimated ploidy of six, and 6.43% with ploidy 14. Within each ploidy level, most of the genotypes were either homozygous for the reference allele or had only one copy of the reference allele, with allele dosages of zero and one accounting for more than 50% of the total number of genotype calls for ploidy levels from six to 12. For ploidy 14, dosage estimates were more evenly distributed among different levels, but there was still an excess of dosages equal to zero and one.

In sweet potato, we identified a total of 77,837 SNPs that were kept after filtering for mean read depth, posterior probability of genotypes and ploidy estimates, call rate, and segregation distortion in the progeny.
segregation according to Hardy–Weinberg Equilibrium. A summary of allele dosage estimates of the SNPs is shown in Fig. 2. Most of the genotypes were either homozygous for the reference allele (53%) or had only one copy of the reference allele (13%), with allele dosages of zero and one (for both the reference and alternative alleles) accounting for more than 76% of the total number of genotype calls.

Genomic selection

Sugarcane

Overall, the predictive abilities of genomic selection in sugarcane were low, regardless of the model or marker set utilized. Figure 3 shows the distribution of the predictive ability values in the sugarcane dataset over different cross-validation runs of the G and G + D models, when using all the makers with full ploidy and allele dosage information and using diploidized makers.

For Brix and Diameter, the G and G + D models using ploidy and allele dosage estimates showed higher mean predictive abilities than when using diploidized markers. For Brix, mean predictive abilities were invariant to the inclusion of dominance effects in the models, both when using dosage information (0.24 ± 0.01) or diploidized markers (0.18 ± 0.02). Whereas for Diameter, the mean predictive abilities slightly decreased when dominance effects were included, regardless of the marker set used. When using dosage information, mean predictive abilities were 0.19 ± 0.01 (G model) and 0.16 ± 0.01 (G + D model) and when using diploidized markers, mean predictive abilities were 0.17 ± 0.01 (G model) and 0.14 ± 0.02 (G + D model).

For fiber percentage and stalk height, the G + D model using diploidized markers resulted in the highest mean predictive abilities (0.22 ± 0.02 and 0.26 ± 0.02, respectively). When dosage information was used, the mean predictive abilities were equal for the G and G + D models (0.14 ± 0.01 for fiber and 0.24 ± 0.02 for weight). For the G model using diploidized markers mean predictive abilities were 0.17 ± 0.01 and 0.21 ± 0.02 for fiber and weight, respectively. Dominance effects accounted for a significant share of the genotypic variance for both traits.

In order to better understand the low values of predictive ability we observed in the sugarcane dataset, we performed a phenotypic variance partitioning analysis and obtained estimates of heritability for the evaluated traits. In general, the genotypic variance had a relatively small or intermediate magnitude for all the traits, with correspondingly low or intermediate heritability values. Figure 4 shows the partitioning of the phenotypic variance into its main components. The residual variance had a large magnitude for all of the traits, corresponding to 36, 35, 49, 58, 48, and 34% of the phenotypic variation observed for Brix, sucrose content, fiber percentage, stalk diameter, stalk weight and stalk height, respectively. The effect of genotypes had an intermediate magnitude for stalk diameter and a small magnitude for...
For traits Brix, sucrose content, fiber percentage, stalk diameter, stalk weight, and stalk height, the heritability coefficients were 0.31, 0.35, 0.37, 0.55, 0.41, and 0.36, respectively.

We also performed the partition of genetic variance into additive and digenic dominance components using the G and G + D models. Figure 5 shows the variance partitioning for each trait, model, and marker set. Our results show that the digenic dominance effects didn’t explain any portion of the genetic variance regardless of the model and marker set used.

**Sweet potato**

The predictive abilities of the genomic selection models in sweetpotato were moderate to high and the distribution
of predictive ability values was nearly equivalent between models and marker sets. Figure 6 shows the distribution of the predictive ability values in the sweet potato dataset over different cross-validation runs of the G and G + D models when using all the makers with full ploidy and allele dosage information and using diploidized makers.

The values of mean predictive ability for the green–red coordinate (a), the yellow-blue coordinate (b), and color saturation (C) were similarly high and barely differed between marker sets and models. The G and G + D models using either diploidized markers or full dosage information had nearly equal mean predictive ability for all three traits: 0.67 ± 0.01, 0.70 ± 0.01, and 0.72 ± 0.01 for a, b, and C, respectively. For lightness (L) and hue angle (h), the mean predictive ability values were lower than for the other three colorimetry traits. Predictive abilities did not differ when including dosage information or dominance effects in the model. Both models using diploidized markers and markers with dosage information had nearly equal mean predictive abilities of approximately 0.59 ± 0.02 and 0.58 ± 0.01 for L and h, respectively.

The heritability coefficients estimated through the phenotypic analysis were very high for all traits. For colorimetry traits, a, b, C, L, and h the estimated heritabilities were 0.98, 0.95, 0.96, 0.85, and 0.98, respectively. For the genotypic analysis, the partition of genetic variance into additive and digenic dominance components using the G and G + D model is shown in Fig. 7. For sweet potato, the digenic dominance effects did not explain any portion of the genetic variance for traits a, C and L. For trait b, when using dosage information, a very small portion (1%) of the variance was attributed to digenic dominance effects. For trait h, when using allele dosage information and when using diploidized markers, the percentage of variance explained by dominance effects was 5%.

Simulations

In the simulated datasets, the highest predictive abilities were achieved when including full ploidy and dosage information. Figure 8 shows the distribution of the predictive ability values in the simulated datasets over different cross-validation runs of the G and G + D models when using all the makers with full ploidy and allele dosage estimates (Full ploidy) and diploidized markers.

When using dosage information, including digenic dominance effects was only advantageous under a high dominance degree and when the genotype frequencies in the population were more evenly distributed (Population 1). In this scenario, when using full ploidy markers, the G and G + D models had mean predictive abilities of 0.32 ± 0.02 and 0.48 ± 0.02, respectively. The mean predictive ability of the G + D model when using diploidized markers (0.43 ± 0.02) was lower than that of the G + D model using
dosage information. The G model using diploidized markers had the lowest mean predictive ability value (0.22 ± 0.03).

For Population 1 with a lower dominance degree, when using full ploidy markers the mean predictive ability of the G + D model (0.48 ± 0.02) was nearly equal but slightly lower than that of the G model (0.49 ± 0.02). When using diploidized markers there was a clear advantage of including dominance in the models, with mean predictive abilities of 0.20 ± 0.02 and 0.43 ± 0.02 for the G and the G + D models, respectively.

When the frequency of non-simplex heterozygous genotypes in the population was low (Population 2) the values of mean predictive ability for the different models and markers were similar in all simulated scenarios. For the low dominance degree level, the mean predictive abilities were approximately 0.42 ± 0.02 for both the G and G + D models using dosage information and 0.42 ± 0.02 and 0.40 ± 0.02 when using diploidized markers. For the high dominance degree level, the mean predictive abilities were approximately 0.45 ± 0.02 for both the G and G + D models using full dosage information and approximately 0.45 ± 0.03 with the less informative diploidized markers.

Results for genotypic variance partitioning in the simulated datasets are shown in Fig. 9. The results show that the G + D model when using allele dosage information is able to capture at least part of the variation due to digenic dominance effects, regardless of the population and dominance degree level. For Population 1, the proportion of variance explained by digenic dominance effects was 23 and 12% for high and low levels of dominance, respectively. For Population 2, the proportion of variance explained by digenic dominance effects was 10...
and 8% for high and low levels of dominance, respectively. When using diploidized markers, the G + D model was not able to capture any variance associated with dominance effects in Population 2. The proportions of variance due to additive effects when using diploidized markers were 24% (Pop. 1 / High dominance), 43% (Pop. 1 / Low dominance), 33% (Pop. 2 / High dominance), and 50% (Pop. 2 / Low dominance). The proportions of variance due to additive effects when using allele dosage information were 32% (Pop. 1 / High dominance), 27% (Pop. 1 / Low dominance), 26% (Pop. 2 / High dominance), and 42% (Pop. 2 / Low dominance).
Results for the additional simulated scenario with 170 individuals are shown in Fig S1.5 and Table S1.1 of Supplementary Material 1. Predictive abilities in this dataset were lower than the predictive abilities obtained in larger simulated datasets Fig S1.5. When using allele dosage information, predictive abilities were invariant to the inclusion of dominance in the models. The mean predictive abilities for the G and G + D models when including dosage information were of 0.29 ± 0.02 and 0.28 ± 0.02, respectively. When using diploidized markers, there was a clear improvement when dominance effects were included. The mean predictive abilities for the G and G + D models when including dosage information were 0.16 ± 0.02 and 0.27 ± 0.02, respectively. Variance partitioning results show that the variance due to dominance effects was only captured by the G + D model when diploidized markers were used (Table S1.1).

**Discussion**

We present our discussion in two sections. First, we discuss the results we obtained implementing genomic prediction in the sugarcane and sweet potato datasets. Second, we discuss the results we obtained with the simulated datasets and compare those with what we obtained with the real data. In both sections, we also show how models could potentially be improved to address the limitations in our study.

**Genomic prediction in sugarcane and sweet potato**

The values of prediction ability for sugarcane were low, while for sweet potato we were able to obtain moderate to high values of predictive ability. Regardless of the prediction ability magnitude, for both species, there was no significant improvement in predictions when including allele dosage information or dominance effects in the model. For sugarcane, we believe the low heritability and the size of the population were the main reasons why prediction models had low performance. In both species, the high number of homozygous and simplex markers are likely playing a role in the low sensitivity of the models to including dosage information and digenic dominance effects.

We were able to obtain high-quality genotypic data in sugarcane. We identified 6,550 SNPs with high mean read depths, high posterior probability of genotypes, and ploidy estimates. Our SNP set exceeds in marker count many genetic studies in sugarcane (Bundock et al. 2009; Gouy et al. 2013; Costa et al. 2016; Yang et al. 2017; Gutierrez et al. 2018). However, the phenotypic variance partitioning analysis showed that, for all traits, most of the variation observed in the field experiments did not stem from differences between the individuals in the F1 progeny, as the variance components associated with the effect of genotypes and genotype × environment interactions had low magnitude in comparison to other experimental sources of variation.
These low values of genotypic variability resulted in low to intermediate values of heritability, which in turn are usually associated with lower predictive ability (Combs and Bernardo 2013; Lian et al. 2014). For all of the traits we evaluated, several studies have reported higher heritability coefficients when analyzing data from sugarcane cultivar trials (Milligan et al. 1990; Gravois and Milligan 1992; Tenas et al. 2016). This indicates that implementing genomic selection in sugarcane is likely to be more advantageous than our results may suggest.

The small training population size (170) in the sugarcane dataset might also be playing a key role in explaining the low values of predictive ability we observed. This idea is supported by comparing predictive abilities of the models when including or not including digenic dominance effects. For most of the traits, there was a small reduction in the predictive ability when digenic dominance effects were included. Including digenic dominance effects result in estimating an additional parameter (the variance of the digenic dominance effects), thus requiring more observations for accurate estimates to be obtained (Button et al. 2013). With a small population size, the estimates of dominance effects were likely not accurate, and the predictive ability of the model decreased. Indeed, the G model showed lower values of AIC (Akaike Information Criterion) than the G + D model (Table S1.2 of Supplementary Material 1), indicating a better fit when digenic dominance effects were not included in the model. In this case, the inclusion of dominance likely led to a model fit in the training set that did not translate into a better fit in the validation set (i.e., overfitting), resulting in lower prediction accuracy.

In both datasets, a large proportion of the SNP calls corresponded to either homozygous or single-dosage genotypes. In breeding populations, this can either occur when the polymorphisms genotyped represent relatively recent mutations in the genomes or when selective pressure has led to the near fixation of genotyped loci (Robertson 1960). In highly polyploid species such as sugarcane and sweet potato, even with very intense selective pressure, the fixation of favorable alleles is extremely difficult as deleterious alleles may have a high number of copies (Gallais 2003). Hence, the presence of recent mutations is the likely explanation for the genotype frequencies we observed. In the literature, similar results were observed for a population of chrysanthemum, an autohexaploid species (van Geest et al. 2017). These results challenge the standard belief that highly autopolyploid organisms are generally highly heterozygous and thus raise the need for further genomic studies in these species.

This low frequency of higher-dosage genotypes is potentially masking the advantages of including allele dosage information in genomic selection models. As mostly only one class of heterozygous genotype is present, the marker sets with dosage information are not much more informative than their diploidized counterparts. We verified the effect of small population size and low number of heterozygous classes in our prediction models by using simulated datasets, and we showed that it indeed affects the sensitivity of prediction models to the two different marker sets. In the following section, we discuss these results more thoroughly.

**Genomic prediction in simulated datasets**

The simulation results demonstrated that genomic prediction including allele dosage information and digenic dominance effect leads to higher predictive abilities only when there is a substantial presence of different heterozygous genotypic classes in the population (Population 1). As mentioned in the previous section, this is likely the main reason why predictions did not improve when including allele dosage information for the real datasets we used in this study. When the simulated populations had a higher frequency of homozygous and simplex genotypes (Population 2), and therefore a similar genotype distribution to the sugarcane and sweet potato datasets, we observed the performance of genomic prediction models to also be invariant to the inclusion of allele dosage and dominance effects.

With this, the results demonstrate that the simulated populations are a good proxy for better understanding the results we obtained in the real datasets. In addition to that, results from the variance partitioning analysis in the simulated data showed that G + D models when including allele dosage information were able to capture at least part of the variance associated with digenic dominance effects, even in Population 2. This indicates that the absence of variation due to dominance for most traits in the sweet potato datasets might indeed be due to mostly additive gene action. That is not the case for sugarcane. As we speculated, results from the variance partitioning in the smaller simulated dataset showed that the G + D model using dosage information was not able to estimate digenic dominance effects with a population size of 170. As for the G + D model using diploidized markers, the variance partitioning analysis in Population 2 showed that dominance effects are not captured by the model when there is a much higher frequency of homozygous and simplex genotypes in the population.

Our results also demonstrate that the use of diploidized markers is a good alternative when allele dosage estimates are not available. This is true even with a sizeable presence of different heterozygous genotypic classes (Population 1). In these simulated scenarios, we observed that the performance of the G + D model using diploidized markers nearly matched the performance of the G + D model when considering allele dosage information, regardless of the dominance level. This is important because when using genotyping-by-sequencing techniques in autopolyploids, accurate genotype calls with allele dosage demand a high sequencing depth.
(Uitdewilligen et al. 2015; Bastien et al. 2018). When only low-depth sequencing data is available, making diploidized genotype calls can be an efficient way of using the data without having to obtain allele dosage estimates (Matias et al. 2019). Our results show that, in these situations, if dominance effects are included in the prediction model, the performance loss for using diploidized markers is not drastic.

Including dominance in the model is also important when using the allele dosage information. However, in this case, including digenic dominance effects is only advantageous when the dominance degree is high. When allele dosage information is included and the dominance degree is low, the G model performs just as well as the G + D model. In contrast, under high dominance degree levels, the performance loss when using the G model rather than the G + D model is significant. To date, little is known about the magnitude of the dominance gene action that is present in the traits of highly autopolyploid species. More research is still needed for breeders to have an estimate of the dominance level of traits in autopolyploid breeding populations. In the current context, our results show that the G + D model should be preferred, as it is the best performing model regardless of the dominance level and frequency of non-simplex heterozygote classes in the population.

Generally, autopolyploid crop varieties are clonally propagated and the genotypes in vegetatively propagated crops are typically heterozygous (Grüneberg et al. 2009). The genetic value of heterozygous genotypes is a function of additive and non-additive gene action (Falconer and Mackay 1996a, b). Non-additive gene action comprises both dominance and epistatic effects. For clonally propagated species, both additive and non-additive gene action are transmitted across generations in the selection process (Bernardo 2010). Therefore, genomic selection models for cultivar selection in these species should aim to include both dominance and epistatic effects for traits that are likely affected by non-additive variation. The importance of including dominance effects in genomic models for clonally propagated crops has also been demonstrated for selection of parents in recurrent selection breeding programs (Werner et al. 2020). In this study, we investigated only two of many possible ways of including dominance effects in prediction models for highly polyploid species.

It is important to notice that we validated our models using simulated phenotypes consisting of only additive and digenic dominance effects. We were able to demonstrate that our fully informative dosage-aware analysis performs better than other simpler genomic prediction models when it comes to these two simulated effects. However, higher-order dominance effects (i.e., interactions between more than two alleles) could also be present in autopolyploid species; hence further improvements in predictions could be achieved by expanding genomic prediction models to include these effects. Moreover, it is still unclear how much of the genotypic variation in highly autopolyploid species is explained by digenic dominance effects. In autotetraploid, Endelman et al. (2018) and Amadeu et al. (2020) have observed digenic dominance effects to explain only a small portion of the genotypic variance. In their case, there was little advantage to including digenic dominance effects in genomic predictions.

**Conclusion**

We showed that estimates of ploidy and allele dosage can improve genomic prediction in highly polyploid species. This is mostly true when there is a substantial number of heterozygous genotypes in the population. When the frequency of heterozygous genotypes in the population is low, such as in the sugarcane and sweet potato datasets, there is little advantage in including allele dosage information in the models. Our simulation results also show that using diploidized markers in the absence of allele dosage estimates can nearly match the performance of fully informative marker sets. However, this is true only when including dominance effects in the genomic prediction models. With the full dosage information available, digenic dominance effects can significantly improve genomic prediction, provided that the trait being predicted has a high mean dominance degree and that the population has a high frequency of heterozygous genotypes.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00122-021-03994-w.

**Author Contribution Statement** LGB, APS, and GRM conceived the study. APS provided the genotyping by sequencing raw read data for the sugarcane population. LGB and VHM performed the SNP calling in the sugarcane and sweet potato datasets. LGB expanded and implemented the genomic selection models and designed the plant breeding program simulations. All authors read and approved the manuscript.

**Funding** This study was supported in part by the Brazilian National Council for Scientific and Technological Development (CNPq) and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES) Finance Code 001.

**Declarations**

**Conflict of interest** The authors certify that they have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript.

**Availability of data and code** The sugarcane and sweet potato datasets as well as the code for obtaining genomic covariance matrices of additive and digenic dominance effects can be found on the Github repository https://github.com/Lorenagbb/GS_HighlyPolyploid. The code for generating all four simulated datasets can also be found on the repository.
References

Amadeu RR, Ferrão LFV, de Oliveira IB, Benevenuto J, Endelman JB et al (2020) Impact of dominance effects on autotetraploid genomic prediction. Crop Sci 60:656–665

Bastien M, Boudhrioua C, Fortin G, Belzile F (2018) Exploring the potential and limitations of genotyping-by-sequencing for SNP discovery and genotyping in tetraploid potato. Genome 61:449–456

Bernardo R (2010) Breeding for Quantitative Traits in Plants. Stemma Press, Woodsbury, MN

Bernardo R, Yu J (2007) Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Sci 47:1082–1090

Blischak PD, Kubatko LS, Wolfe AD (2018) SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. Bioinformatics 34:407–415

Bundock PC, Eliott FG, Ablett G, Benson AD, Casu RE et al (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. Plant Biotechnol J 7:347–354

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) ASReml-R reference manual. 160

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J et al (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365

Cheavegatti-Gianotto A, de Abreu HMC, Arruda P, Bespalhok Filho B, Bundock PC, Eliott FG, Ablett G, Benson AD, Casu RE et al (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. Plant Biotechnol J 7:347–354

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) ASReml-R reference manual. 160

Center de coopération internationale en recherche agronomique pour le développement (CIRAD). https://www.cirad.fr/en/our-research/tropical-supply-chains/sugar cane/context-and-issues

Clark LV, Lipka AE, Sacks EJ (2019) polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids. G3: Genes. Genomes, Genomics 9:663–673

Combs E, Bernardo R (2013) Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. Plant Genome 6

Costa EA, Anoni CO, Mancini MC, Santos FRC, Marconi TG et al (2016) QTL mapping including codominant SNP markers with ploidy level information in a sugarcane progeny. Euphytica 211:1–16

Crossa J, de los Campos, P. Pérez, D. Gianola, J. Burgueño, et al (2010) Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics 186:713

Damesa TM, Möhring J, Worku M, Piepho H-P (2017) One Step at a Time: Stage-Wise Analysis of a Series of Experiments. Agron J 109:845–857

de Bem Oliveira I, Resende MF, Ferrao F, Amadeu R, Endelman J et al (2018) Genomic prediction of autotetraploids: influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. bioRxiv

Dufrène F, Stift M, Vergilino R, Mable BK (2014) Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. Mol Ecol 23:40–69

Duhnen A, Gras A, Teyssèdre S, Romestant M, Claustres B et al (2017) Genomic Selection for Yield and Seed Protein Content in Soybean: A Study of Breeding Program Data and Assessment of Prediction Accuracy. Crop Sci 57:1325

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLOS ONE 6:1e19379

Endelman JB, Carley CAS, Betheke PC, Coombs JJ, Clough ME et al (2018) Genetic Variance Partitioning and Genome-Wide Prediction with Allele Dosage Information in Autotetraploid Potato. Genetics 209:77

Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics. Longman, Harlow, UK

Food and Agriculture Organization of the United Nations (1997) FAOSTAT statistical database. FAO, Rome

Freyre R, Iwanaga M, Orjeda G (1991) Use of Ipomoea trifida (HBK.) G. Don germ plasm for sweet-potato improvement. 2. Fertility of synthetic hexaploids and triploids with 2 a gametes of I. trifida, and their interspecific crossability with sweet potato. Genome 34:209–214

Gallais A (2003) Quantitative genetics and breeding methods in autopolyploid plants. Editions Quae

Garcia AAC, Mollinari M, Marcon TG, Serang OR, Silva RR et al (2013) SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Sci Rep 3:3399

Gaynor RC, Gorjanc G, Hickey JM (2021) AlphaSimR: an R package for breeding program simulations. G3 Genes Genomics Genetics 11

Gerard D, Ferrão LFV, Garcia AAF, Stephens M (2018) Genotyping Polyploids from Messy Sequencing Data. Genetics 210:789

Gouy M, Roussele Y, Bastianelli D, Lecomte P, Bonnal L et al (2013) Experimental assessment of the accuracy of genomic selection in sugarcane. Theor Appl Genet 126:2575–2586

Gratvol C, Regulski M, Bertalan M, McCombie WR, Da Silva FR et al (2014) Sugarcane genome sequencing by methylation filtration provides tools for genomic research in the genus Saccharum. Plant J 79:162–172

Gravois KA, Milligan SB (1992) Genetic Relationships between Fiber and Sugarcane Yield Components 32:62–67

Grüneberg W, Mwanga R, Andrade M, Espinoza J (2009) Selection methods. Part 5: breeding clonally propagated crops. Plant Breed Farm Participation, pp 275–322

Gutierrez AF, Hoy JW, Kimbenga CA, Baisakh N (2018) Identification of Genomic Regions Controlling Leaf Scalld Resistance in Sugarcane Using a Bi-parental Mapping Population and Selective Genotyping by Sequencing. Front Plant Sci 9:877

Hawkins C, Yu L-X (2018) Recent progress in alfalfa (Medicago sativa L.) genomics and genomic selection. Crop J

Hayes BJ, Wei X, Joyce P, Atkin F, Deomano E et al (2021) Accuracy of genomic prediction of complex traits in sugarcane. Theor Appl Genet 134:1455–1462

Heffner EL, Sorrells ME, Jannink J-L (2009a) Genomic Selection for Crop Improvement All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval syst. Crop Sci 49:1–12

Jackson DM, Harrison HF, Jarret RL, Wadl PA (2018) Color analysis of any information storage and retrieval syst. Crop Sci 49:1–12

Katayama K, Kobayashi A, Sakai T, Kuranouchi T, Kai Y (2017) Recent progress in sweetpotato breeding and cultivars for diverse applications in Japan. Breed Sci 67:3–14

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359

Liang L, Jacobson A, Zhong S, Bernardo R (2014) Genomewide Prediction Accuracy within 969 Maize Biparental Populations. Crop Sci 54:1514

Matias FI, Xavier Meireles KG, Nagamatsu ST, Lima Barrios SC, Borges do Valle C et al (2019) Expected genotype quality and...
diploidized marker data from genotyping-by-sequencing of Urochloa spp. Tetraploids. Plant Genome 12:190002

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

Milligan SB, Gravois KA, Bischoff KP, Martin FA (1990) Crop Effects on Broad-Sense Heritabilities and Genetic Variances of Sugarcane Yield Components. Crop Sci 30:344

Mollinari M, Olukolu BA, da Pereira GS, Khan A, Gemenet D et al (2020) Unraveling the Hexaploid Sweetpotato Inheritance Using Ultra-Dense Multilocus Mapping. G3: GenesGenomesGenetics 10:281

Mollinari M, Serang O (2015) Quantitative SNP genotyping of polyploids with MassARRAY and other platforms. In: Batley J (ed) Plant Genotyping. Methods in Molecular Biology (Methods and Protocols) 1245

Oracion M, Niwa K, Shiotani I (1990) Cytological analysis of tetraploid hybrids between sweet potato and diploid Ipomoea trifida (HBK) Don. Theor Appl Genet 80:617–624

Park S, Jackson P, Berding N, Inmam-Bamber G (2007) Conventional Breeding Practices within the Australian Sugarcane Breeding Program 29:10

Pereira GS, Garcia AAF, Margarido GRA (2018) A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. BMC Bioinformatics 19:398–398

Poland JA, Rife TW (2012) Genotyping-by-Sequencing for Plant Breeding and Genetics. The Plant Genome Journal 5:92

Resende MDV, Resende MFR, Sansaloni CP, Petrolci CD, Missiaglia AA 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

Robertson A (1960) A theory of limits in artificial selection. Proc R Soc London Ser B Biol Sci 153:234–249

Serang O, Mollinari M, Garcia AAF (2012) Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. PLOS ONE 7:e30906

Shiotani I (1988) Genomic structure and the gene flow in sweet potato and related species, pp. 61–73

Slater AT, Cogan NOI, Forster JW, Hayes BJ, Daetwyler HD (2016) Improving genetic gain with genomic selection in autotetraploid potato. Plant Genome 9

Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J Roy Stat Soc: Ser B (methodol) 36:111–133

Soltis PS, Soltis DE (2000) The role of genetic and genomic attributes in the success of polyploids. Proc Nat Acad Sci 97(13):7051–7057

Tena E, Mekbib F, Ayana A (2016) Heritability and Correlation among Sugarcane (Saccharum spp.) Yield and Some Agronomic and Sugar Quality Traits in Ethiopia. Am J Plant Sci 07:1453–1477

Uitdewilligen JG, Wolters AMA, Bjorn B, Born TJ, Visser RG et al (2015) Correction: a next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PloS one 10:e0141940

van Geest G, Voorrips RE, Esselink D, Post A, Visser RG et al (2017) Conclusive evidence for hexasonic inheritance in chrysanthemum based on analysis of a 183 k SNP array. BMC Genomics 18:1–12

Van Raden PM (2008a) Efficient Methods to Compute Genomic Predictions. J Dairy Sci 91:4414–4423

Vitezica ZG, Legarra A, Toro MA, Varona L (2017) Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. Genetics 206:1297–1307

Voss-Fels KP, Wei X, Ross EM, Frisch M, Aitken KS et al (2021) Strategies and considerations for implementing genomic selection to improve traits with additive and non-additive genetic architectures in sugarcane breeding. Theor Appl Genet 134:1493–1511

Werner CR, Gaynor RC, Sargent DJ, Lillo A, Gorjanc G et al. (2020) Genomic selection strategies for clonally propagated crops. bioRxiv 2020.06.15.152017

Wu S, Lau KH, Cao Q, Hamilton JP, Sun H et al (2018) Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for genetic improvement. Nat Commun 9:1–12

Yang X, Sood S, Glynn N, Islam MS, Comstock J et al. (2017) Constructing high-density genetic maps for polyploid sugarcane (Saccharum spp.) and identifying quantitative trait loci controlling brown rust resistance. Mol Breed 37

Zhou M (2013) Conventional Sugarcane Breeding in South Africa: Progress and Future Prospects. Am J Plant Sci 04:189–197

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.