frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 December 2023
DOI 10.3389/fpls.2023.1303417

Check for updates

OPEN ACCESS

EDITED BY
Leif Skot,
Aberystwyth University, United Kingdom

REVIEWED BY
Jun Yan,
China Agricultural University, China
Margaret L. Worthington,
University of Arkansas, United States

*CORRESPONDENCE
Anete Pereira de Souza
anete@unicamp.br

†These authors have contributed equally to this work and share first authorship

RECEIVED 28 September 2023
ACCEPTED 15 November 2023
PUBLISHED 12 December 2023

CITATION
Martins FB, Aono AH, Moraes ACL, Ferreira RCU, Vilela MM, Pessoa-Filho M, Rodrigues-Motta M, Simeão RM and Souza AP (2023) Genome-wide family prediction unveils molecular mechanisms underlying the regulation of agronomic traits in *Urochloa ruziziensis*. *Front. Plant Sci.* 14:1303417. doi: 10.3389/fpls.2023.1303417

COPYRIGHT
© 2023 Martins, Aono, Moraes, Ferreira, Vilela, Pessoa-Filho, Rodrigues-Motta, Simeão and Souza. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Genome-wide family prediction unveils molecular mechanisms underlying the regulation of agronomic traits in *Urochloa ruziziensis*

Felipe Bitencourt Martins[1†], Alexandre Hild Aono[1†], Aline da Costa Lima Moraes[2], Rebecca Caroline Ulbricht Ferreira[1], Mariane de Mendonça Vilela[3], Marco Pessoa-Filho[4], Mariana Rodrigues-Motta[5], Rosangela Maria Simeão[3] and Anete Pereira de Souza[1,2*]

[1]Center for Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Campinas, São Paulo, Brazil, [2]Department of Plant Biology, Biology Institute, University of Campinas (UNICAMP), Campinas, São Paulo, Brazil, [3]Embrapa Gado de Corte, Brazilian Agricultural Research Corporation, Campo Grande, Mato Grosso, Brazil, [4]Embrapa Cerrados, Brazilian Agricultural Research Corporation, Brasília, Brazil, [5]Department of Statistics, University of Campinas (UNICAMP), Campinas, São Paulo, Brazil

Tropical forage grasses, particularly those belonging to the *Urochloa* genus, play a crucial role in cattle production and serve as the main food source for animals in tropical and subtropical regions. The majority of these species are apomictic and tetraploid, highlighting the significance of *U. ruziziensis*, a sexual diploid species that can be tetraploidized for use in interspecific crosses with apomictic species. As a means to support breeding programs, our study investigates the feasibility of genome-wide family prediction in *U. ruziziensis* families to predict agronomic traits. Fifty half-sibling families were assessed for green matter yield, dry matter yield, regrowth capacity, leaf dry matter, and stem dry matter across different clippings established in contrasting seasons with varying available water capacity. Genotyping was performed using a genotyping-by-sequencing approach based on DNA samples from family pools. In addition to conventional genomic prediction methods, machine learning and feature selection algorithms were employed to reduce the necessary number of markers for prediction and enhance predictive accuracy across phenotypes. To explore the regulation of agronomic traits, our study evaluated the significance of selected markers for prediction using a tree-based approach, potentially linking these regions to quantitative trait loci (QTLs). In a multiomic approach, genes from the species transcriptome were mapped and correlated to those markers. A gene coexpression network was modeled with gene expression estimates from a diverse set of *U. ruziziensis* genotypes, enabling a comprehensive investigation of molecular mechanisms associated with these regions. The heritabilities of the evaluated traits ranged from 0.44 to 0.92. A total of 28,106 filtered SNPs were used to predict phenotypic measurements, achieving a mean predictive ability of 0.762. By employing feature selection

techniques, we could reduce the dimensionality of SNP datasets, revealing potential genotype-phenotype associations. The functional annotation of genes near these markers revealed associations with auxin transport and biosynthesis of lignin, flavonol, and folic acid. Further exploration with the gene coexpression network uncovered associations with DNA metabolism, stress response, and circadian rhythm. These genes and regions represent important targets for expanding our understanding of the metabolic regulation of agronomic traits and offer valuable insights applicable to species breeding. Our work represents an innovative contribution to molecular breeding techniques for tropical forages, presenting a viable marker-assisted breeding approach and identifying target regions for future molecular studies on these agronomic traits.

KEYWORDS

feature selection, forage grasses, gene coexpression networks, genomic prediction, machine learning, major importance markers, RNA-Seq

1 Introduction

Pastures composed of tropical forage grasses, particularly those belonging to the *Urochloa* genus, serve as the main food source for livestock animals in tropical and subtropical regions. These pastures play a significant role in the economic sectors associated with beef and dairy production, as well as seed markets (Jank et al., 2014; Ferreira et al., 2021). The genetic improvement of *Urochloa* species is recent, starting approximately 40 years ago, and presents challenges due to varying ploidy levels, high heterozygosity, and a prevalent mode of reproduction through apomixis (Ferreira et al., 2021; Simeão et al., 2021). Among the main goals of breeding programs are the development of cultivars that exhibit tolerance to biotic stresses, adaptability to future climate changes, and increased productivity with enhanced nutritional value to optimize animal performance (Pereira et al., 2018b; Simeão et al., 2021).

These goals can be expedited through the incorporation of genomic selection (GS) into breeding cycles. GS employs statistical models to perform genomic predictions (GPs) of plant performance based on genetic markers, mainly single nucleotide polymorphisms (SNPs) (Daetwyler et al., 2013). Although the estimation of GP models has already demonstrated feasibility in other important polyploid crops (de Bem Oliveira et al., 2020; Pincot et al., 2020; Ferrão et al., 2021; Haile et al., 2021; Juliana et al., 2022; Petrasch et al., 2022), this methodology has only recently started to be tested in *Urochloa* spp. (Matias et al., 2019a; Aono et al., 2022). Therefore, efforts must be directed toward the establishment of high-quality marker panels and large-scale phenotyping (Simeão et al., 2021). Fortunately, two *Urochloa* spp. genomes, specifically *U. ruziziensis* (2n=2x=18), have recently become available (Pessoa-Filho et al., 2019; Worthington et al., 2021), facilitating the identification of many SNPs with the potential to enhance the accuracy of GP analyses in *Urochloa* spp.

Traditionally, GP models employ a dense dataset of molecular markers to compute genomic estimated breeding values at the individual level (Meuwissen et al., 2001). However, in the case of *U. ruziziensis* and other forage species, such as alfalfa and ryegrass, it is a common practice to employ the family (full or half-siblings) as the basic unit for phenotyping and selection (Simeão et al., 2012; Simeão et al., 2016a; Simeão et al., 2016b; Biazzi et al., 2017; Cericola et al., 2018; Jia et al., 2018; Murad Leite Andrade et al., 2022). This practice makes the development of genome-wide family prediction (GWFP) approaches highly advantageous. By considering family groups as the measurement unit, there is a reduction in genotyping efforts, as well as the costs associated with developing GP models (Zou et al., 2016; Rios et al., 2021; Murad Leite Andrade et al., 2022). Furthermore, the implementation of GWFP can improve the predictive ability of selection, increasing the rate of genetic gains for complex traits, as demonstrated in studies on loblolly pine and alfalfa (Rios et al., 2021; Murad Leite Andrade et al., 2022).

To identify family-pool markers, sequencing approaches can be employed to generate a large number of SNP markers (Elshire et al., 2011; Poland et al., 2012). Genotyping-by-sequencing (GBS) is a cost-effective and high-throughput genotyping method that can be used to identify SNPs even in the absence of a reference genome. However, it is important to ensure a reasonable sequencing depth to minimize the occurrence of missing data points (Thakral et al., 2022). GBS has been employed in several studies on family-pool genotyping (Futschik and Schloötterer, 2010; Bélanger et al., 2016; Cericola et al., 2018; Schneider et al., 2022) due to its advantages and straightforward applicability in obtaining allele counts from sequencing reads (Byrne et al., 2013). Consequently, in the context of family-pool GP, the use of allele counts derived from GBS allows for direct inference without the need for estimating allelic dosages (Guo et al., 2018).

In addition to the application of GP models in GS approaches, family-pool markers can also be employed in genome-wide association studies (GWAS). Unlike selection-based applications, GWAS aims to identify loci that are associated with a greater extent of genetic variation, thereby enhancing the understanding of the

genetic architecture underlying complex traits (Ashraf et al., 2014; Zhang et al., 2014; Fè et al., 2015). In this sense, adopting a family-based approach provides a more comprehensive perspective on the genetic variations related to the configuration of traits across different families. Once these genomic associations have been assessed, additional omics approaches can be employed to further elucidate the biological mechanisms triggered by adjacent genes and their association with the configuration of complex traits (Scossa et al., 2021).

Traditionally, data generated from various levels of biological information, such as genomics, transcriptomics, and proteomics, have been analyzed separately. However, more recently, the integration of data followed by appropriate statistical analysis has emerged as a promising approach to unravel the biological implications of different traits in humans (Yang et al., 2014), microorganisms (Borin et al., 2018; Rosolen et al., 2022), animals (Parker Gaddis et al., 2016; Mateescu et al., 2017), and plants (Francisco et al., 2021; Cardoso-Silva et al., 2022). Despite the economic importance of *U. ruziziensis* and the availability of molecular data resources, no study incorporating multiomics has been conducted on *U. ruziziensis* or any species of the *Urochloa* genus.

Although assessing different aspects, GP and GWAS possess complementary advantages, providing robust information for the identification of potential candidate genes related to agronomically important traits. Methodologies originally used for GP have been applied in GWAS to detect loci associated with the trait of interest (Goddard et al., 2016; Wang et al., 2020; Wolc and Dekkers, 2022). Conversely, association studies have demonstrated their usefulness in enhancing GP (Zhang et al., 2014; Bian and Holland, 2017; Jeong et al., 2020). To further enhance the outcomes of association and prediction studies, researchers have explored the integration of machine learning (ML) algorithms. Despite the controversial incorporation of ML in GP, with some studies highlighting its advantages (Ma et al., 2018; Waldmann et al., 2020; Aono et al., 2022) and others refuting them (Crossa et al., 2019; Montesinos-López et al., 2019; Zingaretti et al., 2020), numerous investigations consistently demonstrate that ML-based strategies incorporating feature selection (FS) techniques effectively reduce marker density. These methods not only maintain or enhance prediction accuracy but also enable the identification of polymorphisms associated with phenotypes (Li et al., 2018; Aono et al., 2020; Pimenta et al., 2021; Aono et al., 2022).

In this study, we assessed the feasibility of family-based genotyping in autotetraploid *U. ruziziensis* (2n = 4x = 36) and investigated the GWFP capability to predict biomass production and growth traits in both wet and dry seasons. We employed traditional statistical methods as well as ML algorithms to analyze the data. To enhance prediction accuracy, we employed FS strategies to identify subsets of SNP markers with increased predictive power. Furthermore, we used an ML tree-based approach to estimate the importance of these variations in prediction. The most significant markers were then used as a guide to map RNA-Seq assembled genes, which were considered putatively associated with the investigated traits. To gain a deeper understanding of the molecular mechanisms underlying the regulation of these traits in the different seasons investigated, we expanded the set of identified genes by constructing a gene coexpression network (GCN). Our study not only brings innovation to GWFP, but also proposes a means of integrating genomic and transcriptomic data. Moreover, our findings contribute to the expansion of knowledge on the biological processes influencing the investigated agronomic traits. The outcomes of this work offer valuable resources for future studies and breeding programs targeting the *Urochloa* genus.

2 Materials and methods

2.1 *Urochloa ruziziensis* phenotyping

The progenies used in this study were generated as part of the *Urochloa* breeding program of the Brazilian Agricultural Research Corporation (Embrapa) Beef Cattle (EBC), located in Campo Grande, Mato Grosso do Sul State, Brazil (20°27'S, 54°37'W, 530 m), as described by Simeão et al. (2012), Simeão et al. (2016a), Simeão et al. (2016b). In 2010, seven sexual autotetraploid-induced accessions (R30, R38, R41, R44, R46, R47 and R50) were replicated 20 times to create an open pollination randomized field organized into 26 lines and 12 columns spaced by 2 meters. In 2012, out of the 140 plants, 59 were selected to form breeding progenies and compose the experiment of the study. This selection was based on their viable seed production and flowering synchrony. A total of 1,180 individuals (20 seeds from each of the 59 plants selected) were planted in a randomized block design, with one plant per plot spaced 1.5 m apart (Simeão et al., 2016a, b). From the 59 half-sibling progenies, 50 were chosen based on the criterion of selecting the progenies with more plants that succeeded in the field.

The phenotypic evaluations were performed considering nine clippings at 15 cm height: (1) March 2012; (2) January 2013; (3) April 2013; (4) May 2013; (5) September 2013; (6) October 2013; (7) November 2013; (8) December 2013; and (9) January 2014. According to the climatological water balance assessed through the available water capacity (AWC) metric (Supplementary Figure S1) (Simeão et al., 2016a, b), six clippings were performed in the wet season (1-3,7-9) and three in the dry season (4-6). In addition to the nine clippings, we had a total sum (T) evaluation for each phenotype in the period.

The agronomic traits evaluated in all clippings were green matter yield (GM) and dry matter yield (DM), both measured in grams per family, and regrowth (RG), with scores varying from 0 to 6 as described by Figueiredo et al. (2012). In addition, in clippings 2 and 5, approximately 200 g of leaves and stems from each plant were used to estimate leaf dry matter yield (LDM) and stem dry matter yield (SDM). Considering that clipping 1 was discarded from the analysis, we evaluated 33 combinations of agronomic traits and clippings (clippings 2-9 for GM, DM and RG, and clippings 2 and 5 also for SDM and LDM) (Figure 1), which we considered different phenotypes.

For each combination of agronomic traits and clippings, we employed the following linear mixed-effects model:

$$y = Xr + Zg + e$$

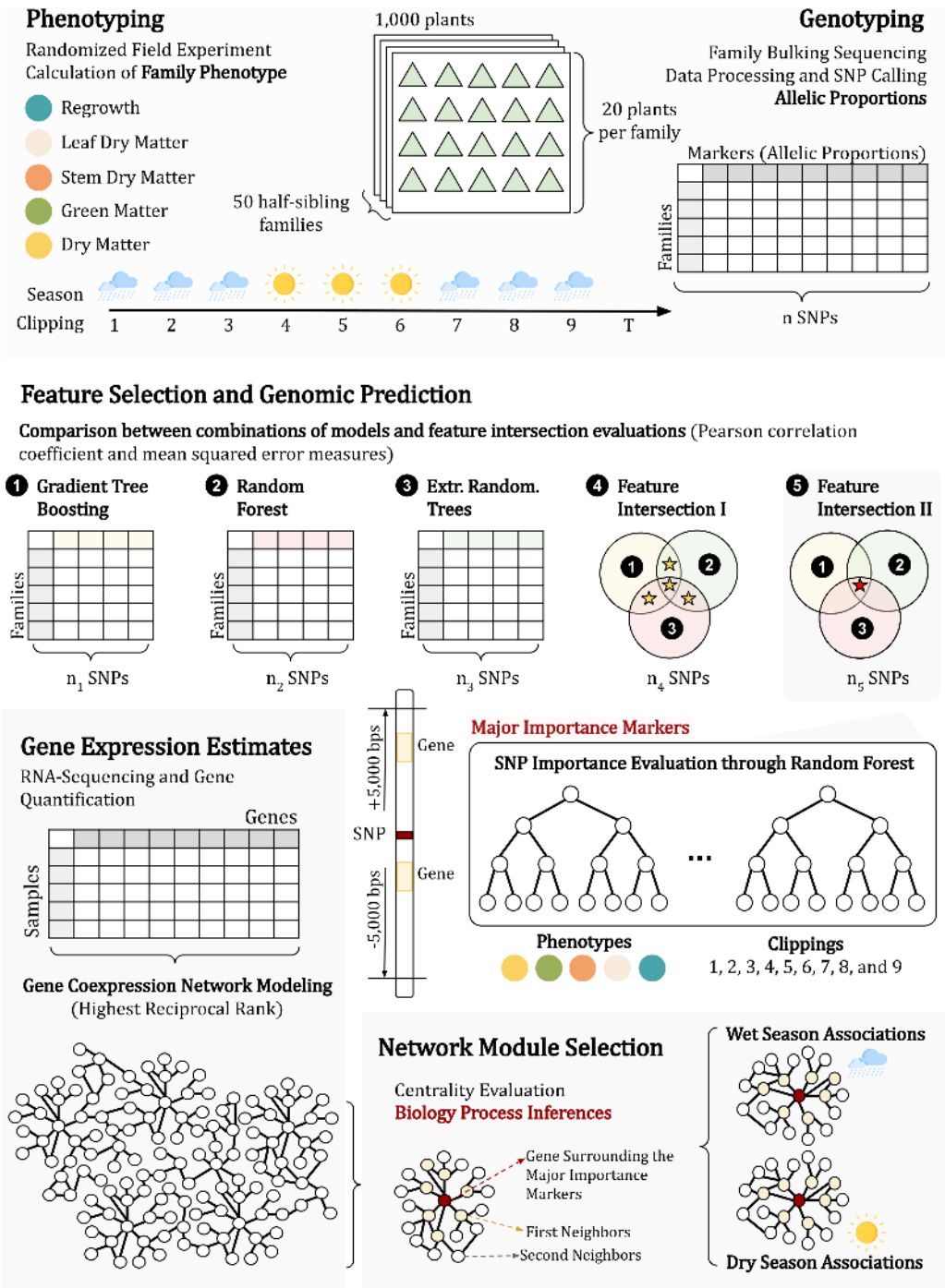

FIGURE 1
The approach established in this research can be divided into three main parts: (i) phenotyping and genotyping the population (1); (ii) identifying phenotypically associated markers through genomic prediction (2 and 3); and (iii) investigating the genes physically linked to the markers in a coexpression network (4, 5 and 6).

where y represents the phenotypic measurements, X is the design matrix of the fixed repetition effects r, Z is the design matrix of the random genotypic effects g, and e is the random residual vector. All the statistical analyses were performed using the software Selegen - REML/BLUP (Resende, 2002; Colombari-Filho et al., 2013). Narrow-sense trait heritability estimates were corrected using the Wright's coefficient of relationship, as described by Simeão et al., (2016a), Simeão et al., (2016b).

To obtain family measurements, we calculated the average of each trait per family and scaled the results between 0 and 1 with the Min-Max technique. To perform a data descriptive analysis of family traits, we used boxplots to assess the distribution and

outliers, computed Pearson's correlation among all phenotype clippings and performed a principal component analysis (PCA) to assess population structure. The descriptive analysis was performed in R (R Core Team, 2021), and all PCA and plots were performed with the package pcaMethods (Stacklies et al., 2007) and the package ggplot2 (Wickham and Chang, 2016), respectively.

2.2 Genotyping

Genomic DNA of all individuals was extracted using the DNeasy Plant kit (QIAGEN) and pooled according to each family, totaling 50 samples. GBS libraries were constructed following the method proposed by Poland et al. (2012) using a combination of a rare cutting enzyme (EcoT22I) and a frequent cutting enzyme (MspI). Subsequently, libraries were sequenced as 150-bp single-end reads using the High Output v2 Kit (Illumina, San Diego, CA, USA) in a NextSeq 500 platform (Illumina, San Diego, CA, USA). The raw sequence data have been submitted to the NCBI Sequence Read Archive (SRA) under accession number PRJNA973612.

We performed quality evaluation of GBS raw sequence reads using FastQC version 0.11.5 (Andrews, 2010) and SNP calling using the TASSEL-GBS pipeline (Glaubitz et al., 2014) modified for polyploids (Pereira et al., 2018a). The reads were aligned to the *U. ruziziensis* genome assembly (Pessoa-Filho et al., 2019; GenBank Assembly GCA_015476505.1) using the BowTie 2.3.1 aligner (Langmead and Salzberg, 2012), and only uniquely mapped reads were employed. SNP markers were filtered using VCFtools v0.1.17 (Danecek et al., 2011) with the following criteria: a minimum sequencing depth of 20 reads, no more than 25% missing data per site, biallelic SNPs only, and removal of redundant (same genotypes in all samples) markers from the sets (Figure 1).

The allele frequency for each marker was estimated as the ratio between the number of reads for the alternative allele and the total number of reads. Missing data were replaced by the site mean of allele frequency. Furthermore, a PCA was performed on the complete genotype data to assess population structure.

2.3 Genomic prediction and feature selection

To create subsets of markers for each phenotype, three FS techniques were applied to the SNP data using the Python 3 library scikit-learn v1.0.2 (Pedregosa et al., 2011): gradient tree boosting (FS-1) (Chen & Guestrin, 2016), extremely randomized trees (FS-2) (Geurts et al., 2006), and random forest (FS-3) (Breiman, 2001). For the FS-1 technique, we employed the mean squared error (MSE) as the loss function, set the learning rate to 0.1, and considered 100 boosting stages. The criterion for assessing split quality was based on the MSE with improvement score by Friedman. We established that a minimum of 2 samples was required to split an internal node, while a minimum number of 1 sample was required for a leaf node. Furthermore, we constrained the maximum number of nodes within the trees to 3. For FS-2 and FS-3, the forest consisted of 100 trees, employing the MSE as the quality measurement function. The minimum number of samples required to split an internal node and the minimum number of samples required to form a leaf node were consistent with those of FS-1. In FS-3, the trees had no node limit, and bootstrapping was employed. Then, to perform modeling, we created feature intersection (FI) datasets by evaluating the intersection of the FS methods, considering markers that were selected by at least two FS techniques (FI-1) and markers that were selected by all three FS techniques (FI-2), similar to the approach proposed by Aono et al. (2020) and Aono et al. (2022) (Figure 1).

As GP strategies, we estimated different models considering six regression approaches across the 33 combinations of traits and clippings, as well as both the reduced (FI-1 and FI-2) and complete versions of the dataset. As conventional GP models, we employed the semiparametric reproducing kernel Hilbert space (RKHS) with a Gaussian kernel (GK) as the covariance function using the R package BGGE v0.6.5 (Granato et al., 2018) and Bayesian ridge regression (BRR) with the R package BGLR v1.0.9 (Perez and de los Campos, 2014). Both models were estimated using 20,000 iterations with a thinning of 5 and a burn-in of 2,000. Additionally, we evaluated four ML algorithms using Python 3 with the scikit-learn library v1.0.2 (Pedregosa et al., 2011): (i) support vector machine (SVM) (Cristianini and Shawe-Taylor, 2000); (ii) random forest (RF) (Breiman, 2001); (iii) adaptive boosting (AB) (Freund and Schapire, 1997); and (iv) multilayer perceptron (MLP) neural network (Popescu et al., 2009). For SVM regression, a radial basis function was used as the kernel, with the gamma coefficient defined as $1/(p \times \sigma_Z^2)$, where p represents the number of loci and σ_Z^2 the variance of the genotype matrix Z. The RF regression was performed with the same parameters as those described for FS-3. For AB, we employed a decision tree regressor as the base estimator, used a linear loss function to assign weights, and limited the maximum boosting interaction to 50 estimators. Finally, the MLP neural network was constructed with a single hidden layer comprising 100 neurons activated by the rectified linear unit (ReLU) function. We employed a quasi-Newton method to optimize the weights and applied a regularization term of 0.001 strength in the L2 regularization term.

The evaluation of the previously described models for GP was performed using a k-fold (k=5) cross validation strategy, repeated 100 times. Two metrics were measured: predictive ability (PA), quantified as the Pearson correlation coefficient, and MSE (Figure 1).

To compare the models, the phenotype clippings and the datasets, we used ANOVAs with multiple comparisons by Tukey's tests implemented in the agricolae R package (De Mendiburu and De Mendiburu, 2020) (Figure 1). For PA, we considered the best scenario to be that in which Tukey's test had "a" or "a" combined with other letters, such as "ab" or "abc", which represents the highest values. On the other hand, for MSE, a scenario is better when its MSE value is lower. Therefore, we considered the best scenarios those with the higher letter or combined with other letters (i.e., "f", "ef" or "def").

2.4 Major importance markers

After identifying the best dataset of markers for each phenotype clipping, we used the random forest algorithm (Breiman, 2001) to estimate the impurity importance of each SNP marker. This estimation was performed considering the Gini importance, which quantifies the normalized total reduction in the criterion (MSE) achieved by each feature (the sum of the feature importance across all markers is equal to 1). To obtain a more refined subset comprising only the markers most likely associated with agronomic traits, we established the major importance set by selecting the top 3 Gini importance markers in each phenotype clipping (Figure 1). In cases where the sum of these three values did not reach 0.5, we continued selecting additional values until the condition was satisfied. Furthermore, a PCA was performed using the major importance markers dataset.

2.5 Transcriptome assembly, quantification and annotation

Previous RNA-Seq data of 11 genotypes of *U. ruziziensis* were used to assess gene expression (Hanley et al., 2021; NCBI BioProject PRJNA513453). Raw data were quality-trimmed using Trimmomatic v0.39 (Bolger et al., 2014). The Illumina adapters, the first 12 bases of the read, and the leading and trailing bases with quality less than 3 were trimmed; the sliding window of 4 bases was set to cut the read when quality/base was less than 20 and only reads with more than 75 bases were kept. Then, the filtered reads were *de novo* assembled by Trinity v2.5.1 (Grabherr et al., 2011) considering a minimum contig length of 300 bases, and assembly integrity was evaluated using the Trinity.pl package utility (Figure 1).

SALMON 1.1.0 (Patro et al., 2017) was used to quantify transcript expression, which was subsequently summarized at the gene level using the tximport R package (Soneson et al., 2015). We retained only genes with more than one transcript per million (TPM) in at least three of the 11 samples, disregarding genes with low-level expression. The longest isoform for each gene was selected, and BUSCO v5.2.2 (Manni et al., 2021) was used to evaluate the annotation completeness against the Viridiplantae database. Finally, we aligned the filtered assembly to the UniProt database (Bateman et al., 2020) using Blastx and Blastn 2.10.0 (Altschul et al., 1990) with an e-value cutoff of 1e-10. Gene Ontology (GO) terms were retrieved using Trinotate software (Bryant et al., 2017), which performed functional annotation (Figure 1).

2.6 Genes linked with markers and GO enrichment

To identify genes physically linked to major importance markers (section 2.4), we conducted alignments between the genes derived from the transcriptome assembly (section 2.5) against the *U. ruziziensis* genome (Pessoa-Filho et al., 2019). Therefore, genes that aligned in a window of 5,000 bp up- and downstream of the marker position were considered physically linked. The alignment was performed using Blastn 2.10.0 (Altschul et al., 1990) with a minimum query coverage of 75% and an E-value cutoff of 1e-6. To visualize the gene position within the genome, we constructed a physical map using MapChart v2.32 (Voorrips, 2002), including information regarding the phenotype and the seasonal associations, as well as Gini importance (Figure 1). In addition, a circular map was constructed using the R package circlize v0.4.14 (Gu et al., 2014) to show the associated genes that were duplicated.

Finally, to obtain a functional profile of the genes linked to the markers, biological process GO term enrichment analysis was performed. This step was achieved with the R package topGO (Alexa and Rahnenfuhrer, 2022), and GO terms with p values< 0.01 in Fisher's exact test were considered significantly enriched.

2.7 Coexpression network

We modeled a GCN using the transcript quantifications normalized in transcripts per million (TPM) and the highest reciprocal rank (HRR) (Mutwil et al., 2009) approach, considering a limit of 30 edges. From the GCN, we selected the genes associated with the agronomic traits and included highly correlated genes that were not considered in the network ranking (Pearson correlation coefficient ≥ 0.9 and a maximum p value of 0.01 with Bonferroni correction). From this defined gene set, we selected the first and second gene neighbors in the GCN. To evidence the gene associations with the two seasons, we highlighted genes related to phenotype clippings 2,3,7,8 and 9, considering them as components of a wet-season associated network. Similarly, genes associated with clippings 4, 5 and 6 were selected to form the dry-season associated network.

Network visualization and evaluation were performed using Cytoscape software v3.9.1 (Shannon et al., 2003). For each gene, we calculated the degree centrality measure with the methods of Barabási and Oltvai, 2004, and considered the genes with outlier values as hubs. Finally, biological process GO term enrichment analyses were performed for the selected genes, including first and second neighbors, to produce a general and seasonal functional profile of the metabolic pathways associated with the agronomic traits with the same method described in 2.6 (Figure 1).

3 Results

3.1 Phenotypic and genotypic data analyses

In our study, we evaluated five important traits for forage grasses (GM, DM, RG, LDM, and STM) across various clippings selected based on wet and dry seasons. Individual measurements were averaged at the subfamily level, and we excluded data from the first clipping. The descriptive analysis of subfamily based phenotypic data did not reveal any discernible patterns concerning the dispersion and skewness of the traits (Supplementary Figure S2). We did not identify any outliers in 17

out of the 33 traits. Despite the absence of any apparent similarity in phenotypic dispersion between the phenotypes evaluated, the correlation analysis yielded significant values for all the comparisons conducted (Supplementary Figure S3). We observed an average R Pearson correlation coefficient of 0.72 (Supplementary Figure S3), with the strongest correlations (~1) observed between the same clippings of GM and DM. Additionally, early clippings (2 and 3) tended to be less correlated with all other measures. This pattern was particularly more pronounced for GM, DM, and SDM. In contrast, SDM in clipping 2 exhibited the lowest correlation with all other phenotypes (Supplementary Figure S3). The progeny mean narrow-sense heritabilities for all phenotype clippings showed a mean value of 0.79, ranging from 0.44 (SDM in clipping 2) to 0.92 (LDM in clipping 5) (Supplementary Table S1).

The GBS experiment generated ~720 million reads, which were processed into 1.3 million tags using the Tassel pipeline. We identified a total of 77,413 SNP markers in this step. After applying quality filters, estimating allele frequencies, and imputing missing genotypes, we retained 28,106 of these markers. This final dataset of markers is referred to as the "complete data" (CD).

By using the phenotypic and genotypic data, we performed PCAs, plotting the dispersion of subfamilies using the scores of the first two principal components (PCs) (Supplementary Figures S4, S5). Although arising from different sources of variation (the proportion of variance explained by the first two PCs was 85.2% and 57.2% for the phenotypic and genotypic data, respectively), similar patterns could be observed. To corroborate such a similarity, we colored the samples from the genotypic PCA scatter plot using PC1 of the phenotypic data. Even without a pronounced presence of 3 groups, as in the phenotypic PCA, the coloring in the genotypic PCA evidenced a clear association between both PCA results (Supplementary Figure S5).

3.2 Genome-wide family prediction

The predictive performance of the GP models at the family level using the CD was assessed through the consideration of two conventional approaches (RKHS and BRR) across 33 phenotypes. Employing a 100-times 5-fold CV strategy, the RKHS model exhibited slightly superior results compared to BRR, with a mean PA of ~0.762 and mean MSE of ~0.025, contrasted to a mean PA of ~0.745 and a mean MSE of 0.026 in BRR. We observed a maximum PA of ~0.875 in the DM-8 trait and a minimum PA of ~0.490 in SDM-2. Aiming to achieve higher performance levels, we evaluated four ML algorithms (SVM, RF, AB and MLP). Among these models, SVM exhibited the best overall performance, with a mean PA of ~0.759 and a mean MSE of ~0.026; however, it did not surpass the performance of the RKHS approach. By considering Tukey's test results for MSE, it became evident that the RKHS model significantly outperformed SVM, emerging as the superior approach in 30 traits compared to 13 of SVM (Supplementary Tables S2-S4). Our results indicate that when using CD for prediction, the ML algorithms were unable to outperform the performance of conventional models.

To increase our predictive accuracy and assess potential associations between traits and markers, we selected specific subsets of SNPs for each of the 33 traits based on the intersections established between FS sets. Each FS approach yielded a distinct quantity of markers: FS-1 selected sets with quantities ranging from 129 to 175 markers (mean of ~150, 0.53% of the CD); FS-2 from 484 to 1154 (mean of ~848, 3% of the CD); and FS-3 from 563 to 853 (mean of ~699, 2.5% of the CD). By considering the intersection approaches established, we obtained FI-1 with SNP quantities ranging from 76 to 122 markers (mean of ~102, 0.36% of the CD) and FI-2 with quantities varying from 5 to 23 markers (mean ~11, 0.04% of the CD) (Supplementary Table S5). In addition to obtaining more restricted sets, these markers selected by FI have more evidence of trait associations, as they were selected by multiple algorithms. In this sense, model performances using the CD were contrasted with the use of models created from the datasets selected by FI-1 and FI-2.

The employment of the FI datasets increased the performance of all models for all traits. This improvement was particularly pronounced in the AB and RF models, which presented the highest levels of accuracy, overcoming RKHS in both FI sets. Among the six models evaluated, the FI-1 approach presented an improved overall performance when compared to FI-2, being considered by Tukey's test the best approach in 168 (FI-2 = 100) and 136 (FI-2 = 89) scenarios for PA and MSE, respectively (Supplementary Table S6). However, individual results for the best models in each scenario were similar, as indicated by the best model in FI-1 (AB with a mean PA of ~0.894 and a mean MSE of ~0.013) and FI-2 (RF with a mean PA of ~0.893 and a mean MSE of ~0.013) (Supplementary Tables S2-S4). Furthermore, when analyzing the clippings of a phenotype, we observed that the best performances for clippings in the combinations AB-FI-1 and RF-FI-2 varied in GM and DM, but for RG (clipping 3), SDM (clipping 5) and LDM (clipping 5), the results were equivalent (Supplementary Tables S2-3 and 7).

In this sense, we observed that for the prediction task, both combinations AB-FI-1 and RF-FI-2 can be employed with comparable performance levels. However, for investigating trait –marker associations and catalogs of putative associated genes, FI-2 represents a more restrictive approach. With sets (mean of ~11 markers) approximately ten times smaller than the sets of FI-1 (mean of ~102 markers), FI-2 markers provide a group of markers with a probable reduced number of false positive associations. Therefore, we considered the combination RF-FI-2 as the most promising approach to be employed in our datasets. In addition to the significant decrease in marker density through FI-2, the RF algorithm demonstrated high efficiency for prediction with a PA increase of 6.9% and an MSE reduction of 22.6% when compared to the RKHS using the FI-2 dataset or 17% when compared to the RKHS using the CD dataset (Table 1).

3.3 Major importance markers

Given that the FS strategies employed in our study relied on ML algorithms estimated through a combination of decision trees, and

TABLE 1 Comparison of RKHS and RF model predictive ability and mean squared error for all phenotype clippings using the FI-2 datasets.

Phenotype	Clipping	Predictive ability				Mean squared error			
		RKHS	RF	Diff.	Diff. (%)	RKHS	RF	Diff.	Diff. (%)
Green Matter	2	0.857	0.850	-0.007	-0.8%	0.018	0.018	0	0.0%
	3	0.870	0.869	-0.001	-0.1%	0.016	0.017	0.001	6.3%
	4	0.894	0.912	0.018	2.0%	0.015	0.013	-0.002	-13.3%
	5	0.866	0.917	0.051	5.9%	0.016	0.012	-0.004	-25.0%
	6	0.863	0.903	0.040	4.6%	0.018	0.012	-0.006	-33.3%
	7	0.844	0.923	0.079	9.4%	0.019	0.010	-0.009	-47.4%
	8	0.881	0.913	0.032	3.6%	0.020	0.015	-0.005	-25.0%
	9	0.897	0.944	0.047	5.2%	0.012	0.008	-0.004	-33.3%
	T	0.867	0.937	0.070	8.1%	0.018	0.009	-0.009	-50.0%
	Mean	0.871	0.908	0.037	4.2%	0.017	0.013	-0.004	-24.6%
Regrowth	2	0.868	0.896	0.028	3.2%	0.015	0.014	-0.001	-6.7%
	3	0.940	0.956	0.016	1.7%	0.011	0.008	-0.003	-27.3%
	4	0.859	0.911	0.052	6.1%	0.017	0.012	-0.005	-29.4%
	5	0.826	0.869	0.043	5.2%	0.016	0.013	-0.003	-18.8%
	6	0.833	0.884	0.051	6.1%	0.024	0.016	-0.008	-33.3%
	7	0.862	0.896	0.034	3.9%	0.015	0.011	-0.004	-26.7%
	8	0.860	0.872	0.012	1.4%	0.014	0.013	-0.001	-7.1%
	9	0.813	0.843	0.030	3.7%	0.017	0.014	-0.003	-17.6%
	T	0.886	0.932	0.046	5.2%	0.013	0.008	-0.005	-38.5%
	Mean	0.861	0.895	0.035	4.1%	0.016	0.012	-0.004	-22.8%
Dry Matter	2	0.537	0.638	0.101	18.8%	0.044	0.035	-0.009	-20.5%
	3	0.802	0.840	0.038	4.7%	0.014	0.013	-0.001	-7.1%
	4	0.935	0.941	0.006	0.6%	0.010	0.009	-0.001	-10.0%
	5	0.882	0.895	0.013	1.5%	0.014	0.015	0.001	7.1%
	6	0.884	0.917	0.033	3.7%	0.016	0.011	-0.005	-31.3%
	7	0.849	0.913	0.064	7.5%	0.017	0.011	-0.006	-35.3%
	8	0.911	0.915	0.004	0.4%	0.016	0.016	0	0.0%
	9	0.787	0.925	0.138	17.5%	0.022	0.011	-0.011	-50.0%
	T	0.912	0.943	0.031	3.4%	0.013	0.009	-0.004	-30.8%
	Mean	0.833	0.881	0.048	6.5%	0.018	0.014	-0.004	-19.8%
Leaf Dry Matter	2	0.835	0.875	0.040	4.8%	0.019	0.014	-0.005	-26.3%
	5	0.899	0.924	0.025	2.8%	0.012	0.011	-0.001	-8.3%
	T	0.900	0.952	0.052	5.8%	0.011	0.007	-0.004	-36.4%
	Mean	0.878	0.917	0.039	4.5%	0.014	0.011	-0.003	-23.7%
Stem Dry Matter	2	0.664	0.818	0.154	23.2%	0.033	0.021	-0.012	-36.4%
	5	0.904	0.921	0.017	1.9%	0.011	0.012	0.001	9.1%
	T	0.694	0.835	0.141	20.3%	0.025	0.015	-0.010	-40.0%
	Mean	0.754	0.858	0.104	15.1%	0.023	0.016	-0.007	-22.4%
	Overall Mean	0.839	0.892	0.052	6.9%	0.018	0.013	-0.004	-22.6%

that the top-performing models for FI-1 and FI-2 were AB and RF, respectively, we employed an additional approach to assess marker –trait associations using decision tree structures. We ranked the markers based on RF scores obtained from the FI-2 selected markers. We selected the top three Gini importance markers for each trait, and if the sum of importance for these top three markers did not reach at least 0.5 (out of a total of 1.0), we continued selecting markers from the ranking until we reached half of the total importance score. This process allowed us to compile a list of markers with the highest feature importance, thus preventing underrepresentation of importance across traits. From the 283 FI-selected markers across the 33 traits, we identified a subset of 69 markers with significant predictive relevance. Notably, only for SDM clipping 5, we had to select four markers instead of three (Supplementary Table S8).

Furthermore, we performed a PCA to evaluate the subfamily dispersion considering this set of 69 major importance markers. The first two PCs explained 67.5% of the data variance, an intermediate value between the complete set of SNPs (57.2%) and the phenotypic data (85.2%) (Supplementary Figure S6). Although the values of the first PCs seem to be inverted in such a PCA when compared to the others performed, we observed a similar dispersion pattern (Supplementary Figures S4, S5). As we expected, the scatter plot displayed a group formation visually closer to the phenotypic PCA. Since the markers were selected through associations with the traits, there was a strong relation between the major importance data PC1 and the samples colored using the phenotypic PC1 values (Supplementary Figure S7).

To assess the physical distribution of the FS-selected markers, we constructed a physical map for *U. ruziziensis* using the values obtained from the species' genome. In addition to the set of 69 major importance markers, we incorporated all the FI-2 markers into the constructed map (Figure 2). Regarding the distribution of these markers, we observed associations across all chromosomes without a clear pattern, except for the presence of extensive regions with little or no markers, primarily located in the central regions of chromosomes 1, 2, 5, 7 and 8. We speculate that these regions correspond to the centromeric regions (Figure 2). Chromosome 1 presented the highest number of associations when considering both FI and major importance marker sets, with relatively consistent representativeness. However, it was also the chromosome with the highest number of identified SNPs (Table 2). On the other hand, chromosomes 5 and 6 presented the lowest presence of associations, while chromosome 4 experienced a significant change in representativeness, with a 7% reduction from FI to major importance (Table 2). Furthermore, especially in chromosomes 1, 4 and 7, we observed regions characterized by a high density of minor importance markers near major importance markers, which may suggest the presence of QTL regions associated with agronomic traits.

The major importance set was composed of various markers associated with more than one trait. As evidenced in the physical map, the marker associated with more trait clippings is on chromosome 7, position 42,826,434. This marker was associated with four of the five phenotypes evaluated and was selected for nine clippings, three of which had a Gini importance higher than 0.4 and in six Gini importance between 0.2 and 0.4 (Figure 2). Other markers were associated with various trait clippings, such as a marker on chromosome 1 position 69,834,400, which was associated with six trait clippings, and three other markers with four associations in chromosomes 1 and 3 (Figure 2).

When evaluating the markers for each of the five traits without separating them by clippings, we analyzed the intersections of sets to quantify markers associated with multiple traits (Supplementary Figure S7). Despite the variation in marker quantities between the FI-2 and major importance sets, the logical relationships among the trait sets remained consistent: GM, DM, and LDM shared the highest number of markers, while RG and SDM had a higher proportion of exclusive markers. Interestingly, SDM and RG exhibited generally lower correlations with the other traits as well.

3.4 Marker genes associated with phenotypes

To obtain a set of genes expressed by the species and subsequently assess their coexpression, we employed a previously published transcriptome of 11 *U. ruziziensis* genotypes. The sequencing of the libraries produced a total of ~1.7 billion reads, with 95.5% (Supplementary Table S9) being retained and used for *de novo* assembly. The resulting transcriptome encompassed 575,524 transcripts, of which 223,593 were categorized as unigenes, featuring a transcript N50 length of 1,227 bp. Following filtration based on expression levels, the dataset was reduced to 288,487 transcripts, representing 49,445 unigenes. The evaluation of assembly completeness was performed by comparing the 49,445 unigenes against the Viridiplantae database. From the 425 total BUSCO groups searched, we found 297 complete sequences (69.8%), 48.2% as a single copy and 21.6% as duplicated copies, in addition to 74 (17.4%) and 54 (12.8%) fragmented and missing sequences, respectively.

In the process of functional annotation, we aligned the transcripts to the UniProt database and obtained 197,045 associated GO terms. Among these, 6,156 were unique GO terms. This collection of genes and GO terms was then employed to perform a biological process GO term enrichment analysis of the genes linked to the major importance markers.

After aligning transcripts with the reference genome of *U. ruziziensis* and considering a window of 5,000 bp up/downstream of the marker positions, we mapped a total of 217 genes (264 considering genes with multiple copies) in close physical proximity to 58 markers (Figure 2 and Supplementary Table S8). We did not detect genes linked to all markers, such as on chromosome 1, where no genes were found within a marker region associated with six traits clippings, or on chromosome 5, where out of the four major importance markers, two lacked associated genes (Figure 2).

As previously stated, we identified genes with multiple copies that are linked to more than one major importance marker region. There were 22 genes meeting this criterion, and they are highlighted in red in Figure 2. To facilitate a more comprehensive investigation of these genes, we represented their distribution in a circular map that illustrates their genomic positions. Additionally, we combined

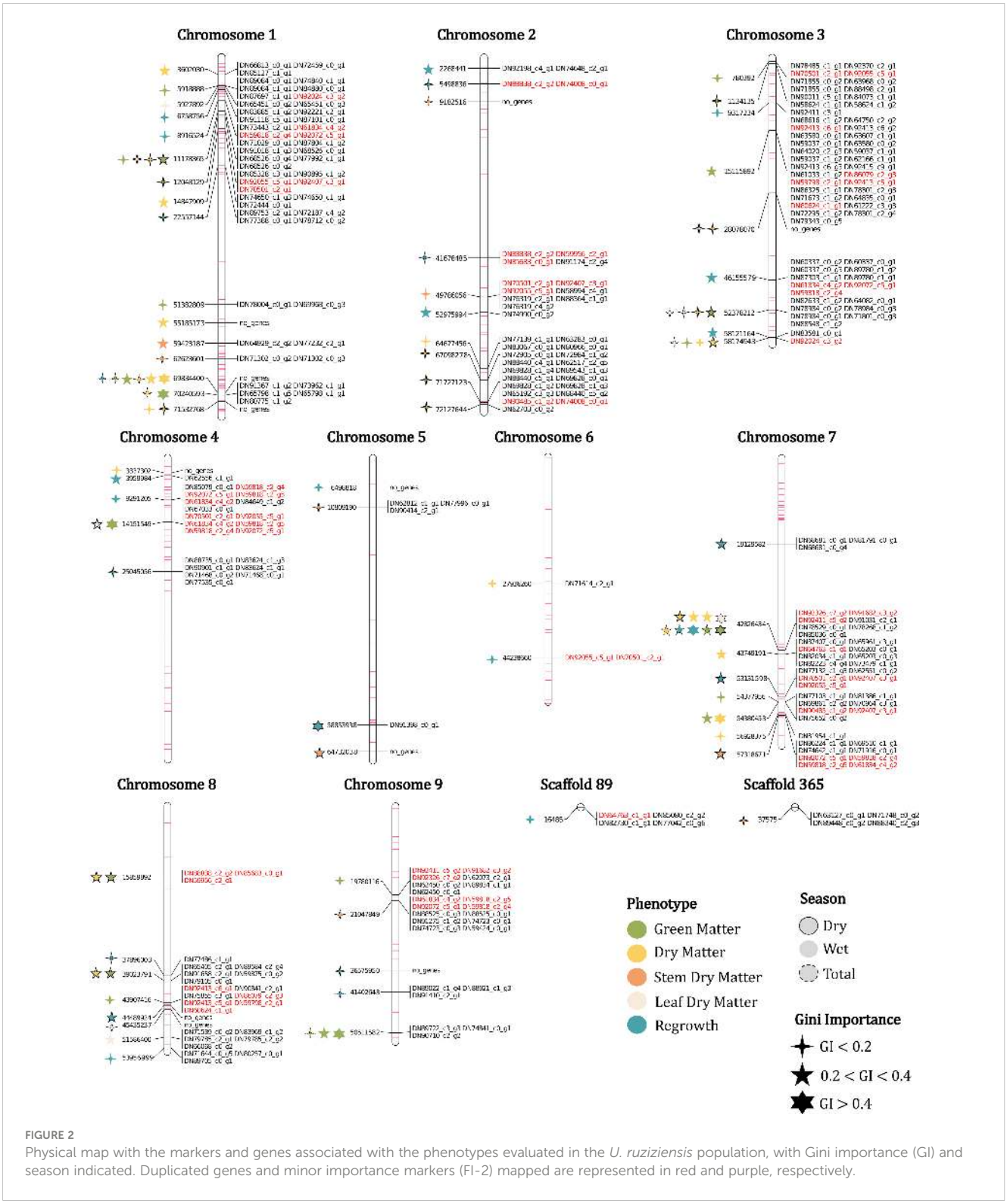

FIGURE 2
Physical map with the markers and genes associated with the phenotypes evaluated in the *U. ruziziensis* population, with Gini importance (GI) and season indicated. Duplicated genes and minor importance markers (FI-2) mapped are represented in red and purple, respectively.

information about copy number variation, trait/season associations, and Gini importance (Figure 3). Among these genes, we identified five genes with six copies. Notably, three of these genes are found together, and collectively, they are associated with seven different trait categories. Furthermore, we identified genes with 4, 3 and 2 copies, all linked to all the evaluated traits, albeit with varying levels of importance, demonstrating no clear pattern.

Regarding the functional annotation of the genes associated with the phenotypes, we identified proteins/enzymes and GO terms for 100 of the 217 genes (Supplementary Table S8). In the region associated with more traits, on chromosome 7 (position 42,826,434), seven annotated genes were mapped, some of which were cinnamoyl-CoA reductase 1, and DEAD-box ATP-dependent RNA helicase 25. Furthermore, on chromosome 1 (position 11,178,365), which is

TABLE 2 Number and percentage of SNP markers identified/selected in each chromosome considering the complete data (CD), feature intersection (FI-2) and top Gini importance datasets.

Chromosome	Complete Data	Feature Intersection - 2	Top Gini Importance
1	4722 (16.8%)	69 (23.4%)	16 (23.2%)
2	3565 (12.7%)	36 (12.2%)	10 (14.5%)
3	3249 (11.6%)	28 (9.5%)	9 (13%)
4	3552 (12.6%)	42 (14.2%)	5 (7.2%)
5	1384 (4.9%)	15 (5.1%)	4 (5.8%)
6	2508 (8.9%)	23 (7.8%)	2 (2.9%)
7	3796 (13.5%)	37 (12.5%)	8 (11.6%)
8	2127 (7.6%)	18 (6.1%)	8 (11.6%)
9	2426 (8.6%)	22 (7.5%)	5 (7.2%)
Scaffolds	777 (2.8%)	5 (1,7%)	2 (2.9%)
Total	28106 (100%)	295 (100%)	69 (100%)

associated with four traits, there are genes annotated to the multidomain protein RHM2/MUM4 which is involved in UDP-D-glucose to UDP-L-rhamnose conversion (Supplementary Table S8). Considering the genes with multiple copies, only three had functional annotation, which translates into AIM25-altered inheritance rate of mitochondria protein 25, cinnamoyl-CoA reductase 1 and E3 ubiquitin-protein ligase SINAT5.

Beyond specific protein annotation, to obtain a general functional profile of the proteins identified, we performed an enrichment analysis of the biological process GO terms and obtained a profile with 18 significant terms (p value< 0.01). The enrichment analysis identified terms associated with various phenotype clippings, such as "lignin biosynthetic process", "auxin efflux" and "flavonol biosynthetic process" (Supplementary Table S10).

3.5 Coexpression network

To provide deeper insights into the functional patterns of genes associated with the agronomic traits evaluated, we modeled a GCN using the gene quantifications from the *U. ruziziensis* accessions. From a total of 49,445 genes, we identified significant interactions between 14,141 genes, represented as nodes in the network structure, connected by 17,812 edges (Supplementary Figure S8). Within this GCN, we found 54 genes from the 217 genes associated with the major importance markers. As we restricted the GCN created to the top 30 gene associations, we expanded the collection of 54 selected genes to more than 109 by considering correlations with a minimum Pearson coefficient of 0.9 and a Bonferroni corrected p value of 0.01. This group of 153 genes was considered directly associated with the traits evaluated.

The potential of a GCN to elucidate metabolic pathways lies in its ability to identify genes that, despite not being selected by the prediction methodology, exhibit coexpression with them. To this end, we extended the set of 153 genes previously selected to the GCN first (308 genes) and second gene neighbors (2233 genes), creating a comprehensive agronomic trait network comprising a total of 2704 genes (nodes) and 3453 edges (Figure 4).

The functional profile of the general agronomic traits network was determined through an enrichment analysis of biological process GO terms, which revealed 11 significant terms (p< 0.01) for the gene set excluding the second neighbors and 16 terms when considering all genes in the network (Supplementary Table S11). When examining the restricted set, which excluded the second neighbors, we found enriched terms related to hormones, such as auxin efflux and abscisic acid transport, as well as biosynthetic processes involving molecules such as flavonoids. In the broader set that included second neighbors, we identified terms associated with DNA metabolism, including mismatch repair, DNA replication, and DNA duplex unwinding. Additionally, other enriched terms were linked to responses to stress, such as response to chitin and regulation of circadian rhythm.

To further explore the differences in functional gene patterns associated with the different seasons, we separated the general agronomic trait network into two seasonal parts. The genes associated with the traits in clippings 2,3,7,8 and 9 were selected for the wet season-associated network, and the genes associated with the traits in clippings 4, 5 and 6 were selected for the dry season-associated network. The wet and dry-season networks encompassed 33 and 22 genes associated with the major importance markers, 58 and 54 highly correlated genes, 102 and 231 first neighbors, 1322 and 1359 second neighbors, and a total of 1515 and 1666 genes (nodes) with 1801 and 2205 edges, respectively (Figures 4A, B). Comparing the seasonal functional profiles, we found shared terms such as flavonol biosynthetic process, auxin efflux and mitotic recombination-dependent replication fork processing. Additionally, we discovered season-specific terms such as abscisic acid transport, isoleucine biosynthetic process, response to nematode and chaperone-mediated protein folding for the wet season. In contrast, the dry season featured enriched terms such as pyridoxal phosphate biosynthetic process, response to water deprivation and response to chitin, all of which are related to stress response (Supplementary Tables S12, S13).

Another remarkable aspect of using GCNs to investigate the regulation of metabolic pathways lies in their ability to define hub genes, which possess a high number of connections in the network, as determined by the degree metric. The hub genes play an important role in regulating the functionality of numerous other genes, thereby potentially influencing the expression of the phenotypes that we are studying. In our modeled agronomic traits network, we found 235 hub genes (degree > 4), of which 14 had a degree > 40. Considering the seasonal networks, there were 107 and 158 hubs in the wet and dry season networks, respectively. Among the highest degree hub genes (>40), we found some present in both season networks, such as the genes that encode the 60S ribosomal protein L9 and the 14-3-3 protein zeta (Supplementary Table S14). While specific to the wet season, we found hub genes of the proteins ELF4-LIKE 4, SUV2 and lipid-transfer DIR1 (Supplementary Table S15) and to the dry season, 60S ribosomal L9, fatty acid-binding and 3-hydroxyacyl dehydratase FabZ (Supplementary Table S16).

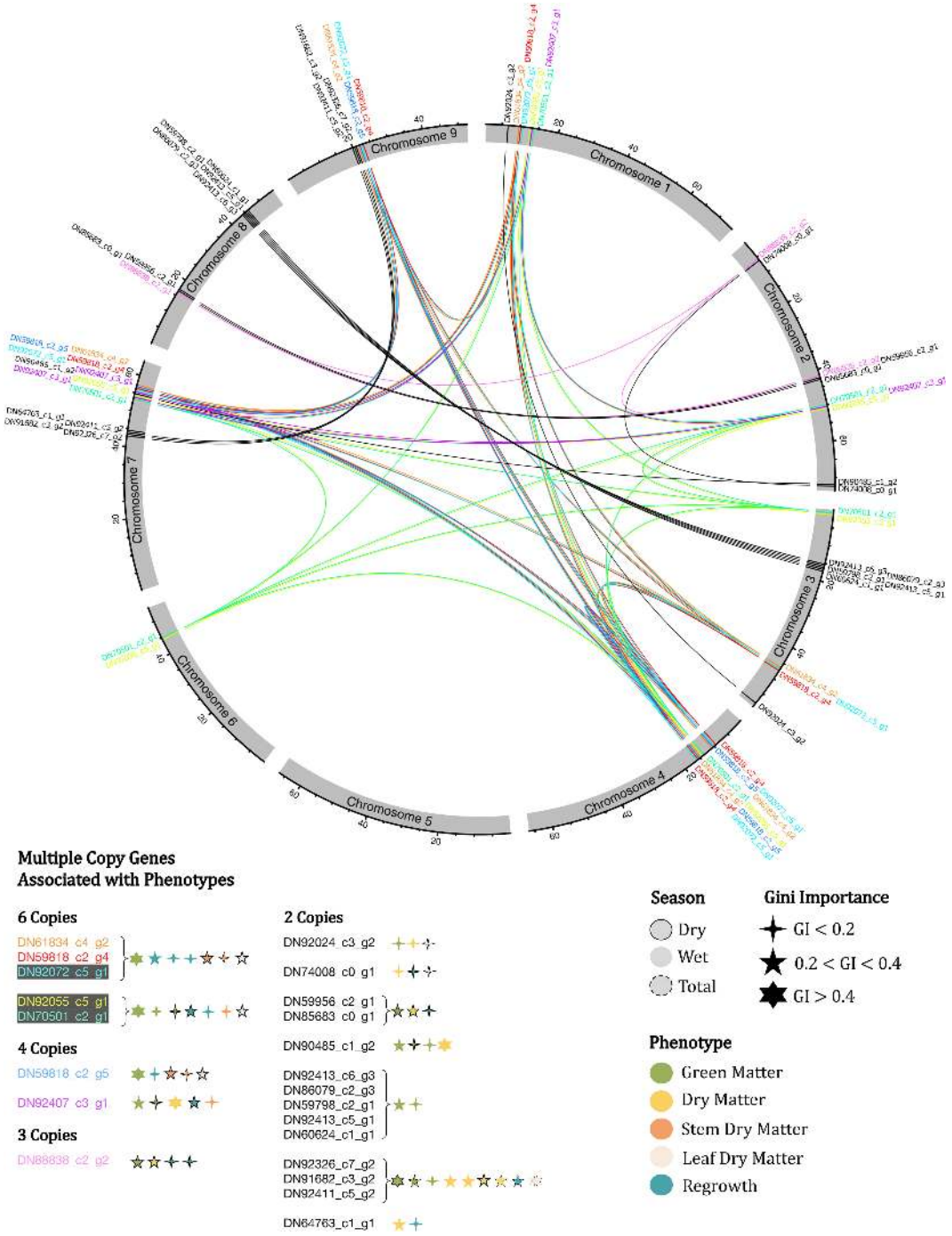

FIGURE 3
Circular map of the *U. ruziziensis* genome, indicating multiple copy genes identified as associated with the phenotypes evaluated, indicating Gini importance and season. The same genes are indicated with the same color, except genes with two copies.

4 Discussion

4.1 Genome-wide family prediction

Recent advances in omics approaches and computational methods for polyploid species have enabled the emergence of studies in important *Urochloa* breeding areas. These include genome assembly (Pessoa-Filho et al., 2019; Worthington et al., 2021), contaminant identification (Martins et al., 2021), transcriptomics (Vigna et al., 2016a; Salgado et al., 2017; Hanley et al., 2021; Jones et al., 2021; Worthington et al., 2021), linkage and QTL mapping (Ferreira et al., 2016; Thaikua et al., 2016; Vigna et al., 2016b; Worthington et al., 2016; Worthington et al., 2019; Worthington et al., 2021), GWAS (Matias et al., 2019b), and GS/GP (Matias et al., 2019a; Aono et al., 2022). Even with the recent progress, there are no studies employing integrative methodologies

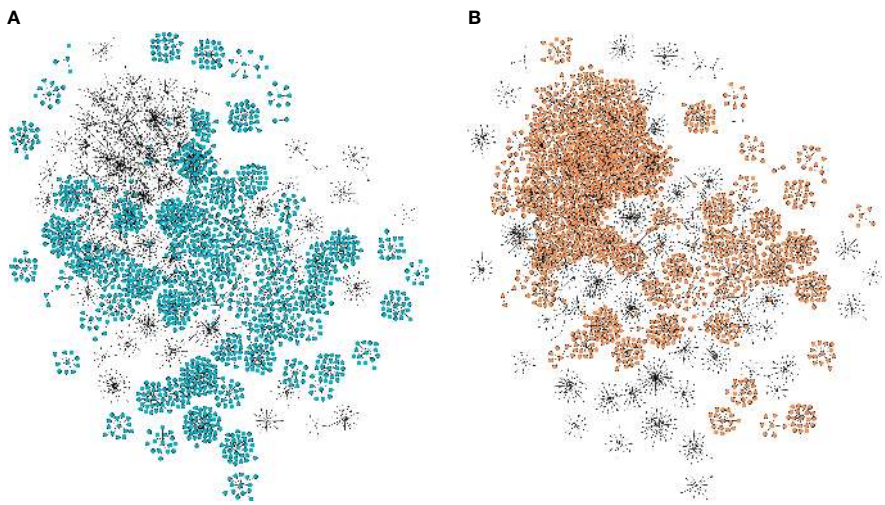

FIGURE 4
Selected and correlated gene coexpression network with first and second neighbors. Each node represents a gene, and each connection represents their correlation. **(A)** Genes associated with wet season trait clippings. **(B)** Genes associated with dry season trait clippings.

in the genus. Therefore, by leveraging the limited genetic resources available and employing biocomputational techniques, we have pioneered the first multiomic approach in the *Urochloa* genus, thereby expanding the molecular knowledge available for breeding.

In our initial approach to evaluating GWFP, we employed the complete marker dataset (CD) with conventional parametric and semiparametric GP models (BRR and RKHS). Our findings consistently yielded either higher or, at the very least, equivalent PA when compared to the achievements reported in other GWFP studies. While we achieved a high mean PA of ~0.8 for the DM clippings, in the case of alfalfa, the authors observed values below 0.7 for the same trait in both 10-fold and leave-one-group-out cross-validation scenarios (Murad Leite Andrade et al., 2022). Additionally, for ryegrass, an even lower PA value of 0.34 was observed in a leave-one-family-out cross-validation scenario (Guo et al., 2018). If we consider the GWFP results for other phenotypes, such as rust resistance and heading date in alfalfa (Murad Leite Andrade et al., 2022), as well as lignin content, stiffness and diameter in loblolly pine (Rios et al., 2021), they consistently exhibited smaller PA values compared to our results, which presented a mean PA of ~0.762 across the 33 traits. We attribute this PA in our predictions primarily to the relatively small population size and limited genetic diversity among our samples. This combination has been previously reported to enhance predictive accuracy, as demonstrated in wheat (Edwards et al., 2019). Moreover, increasing the population with genetically distant samples tends to increase the complexity of the prediction task, subsequently reducing its accuracy, as highlighted in previous studies (Lorenz and Smith, 2015; Berro et al., 2019).

The performance we achieved seems even more promising when compared to GP conducted at the individual level in tropical forages. Studies with *U.ruziziensis* interspecific hybrids (Matias et al., 2019a), *U. decumbens* (Aono et al., 2022), *M. maximum* (de C. Lara et al., 2019; Aono et al., 2022), and *P. virgatum* (Lipka et al., 2014) have yielded PAs ranging from values close to zero to maximum values of 0.7 when evaluating several agronomic and morphological traits and employing different cross-validation schemes. In addition to the well-known advantages of genomic selection, such as its potential to shorten the breeding process (Simeão-Resende et al., 2014) and reduce phenotyping costs (Crossa et al., 2017), the use of GWFP offers other advantages, particularly in forage breeding programs, which typically rely on family- or plot-level phenotyping for conventional breeding (Rios et al., 2021). Furthermore, when GWFP is conducted using ML models combined with FS/FI strategies, it has the potential to significantly lower genotyping costs.

Moreover, for the application of GWFP in breeding programs or for future research, we recommend prioritizing experimental designs featuring a greater number of families rather than increasing the number of individuals within each family. Studies with tetraploid full-sibling families have concluded that using six individuals is sufficient to effectively capture family variation, both in terms of genotyping and phenotyping (de Bem Oliveira et al., 2020; Rios et al., 2021). Furthermore, it is worth noting that to a certain extent, enlarging the training population holds the potential to enhance the performance of GWFP, as indicated by previous research (Fè et al., 2015).

The application of ML algorithms in GP has been extensively explored across various species and phenotypes (Grinberg et al., 2016; Lello et al., 2018; Liang et al., 2020; Chung et al., 2021; Islam et al., 2021; Sandhu et al., 2021). Although there is no concrete empirical evidence supporting the superiority of ML over linear methods (Zingaretti et al., 2020; Varshney, 2021), ML techniques have consistently demonstrated superior or at least equivalent performance compared to conventional models in diverse scenarios (Bellot et al., 2018; Abdollahi-Arpanahi et al., 2020; Liang et al., 2021; Wang et al., 2022). ML methods have the potential to outperform conventional GP models, especially when handling intricate phenotypes influenced by significant dominance and epistatic effects (Wang et al., 2018; Tong and Nikoloski, 2021).

Moreover, there are no studies investigating the applicability of ML methods in GWFP. Thus, we evaluated four classical ML algorithms (SVM, RF, AB and MLP). Surprisingly, none of these algorithms was able to outperform the RKHS model. Although SVM demonstrated competitive performances for PA, it was not as good for MSE. RF and AB exhibited intermediate performances, whereas MLP markedly underperformed (Supplementary Tables S2-4). The poor performance of MLP may be attributed to the limited sample size of our dataset and the lack of hyperparameter tuning. Neural network methods are well known for their need for substantial datasets and meticulous hyperparameter tuning to achieve high prediction accuracy (Bellot et al., 2018; Montesinos-López et al., 2021).

4.2 Major importance markers

Our study goes beyond the applicability analysis of GWFP in *U. ruziziensis*. We also aimed to investigate the metabolic regulation of agronomic traits. As an initial step to achieve this objective, we aimed to establish potential marker-phenotype associations. To this end, the strong performance improvement observed using the FS/FI approach indicates that the selected sets of markers are likely to be near QTLs, and can therefore be used to define genomic regions involved in phenotypic variation (Steinfath et al., 2010; Heer et al., 2018; Zhou et al., 2019; Aono et al., 2020; Pimenta et al., 2021; Aono et al., 2022; Pimenta et al., 2022). Additionally, by utilizing allele proportions for genotyping the family, this approach can be extended to other crop species that employ the family as the unit for conducting GS.

In contrast to other approaches aimed at identifying genotype-phenotype associations, FS techniques do not rely on specific biparental populations (RILs, NILs, F2, etc.), which are necessary for QTL mapping (Mohan et al., 1997; Dhingani et al., 2015). Moreover, FS techniques have the ability to uncover nonlinear and complex associations, addressing a limitation of linear models used in GWAS (Korte and Farlow, 2013).

ML models based on decision trees offer good prediction interpretability since it is possible to assess feature importance. In the context of GP, these models can rank markers based on their association strength with the modeled phenotype (Azodi et al., 2019; Bayer et al., 2021; Medina et al., 2021). Therefore, given that the best model for each FI dataset type was equivalent, we computed the RF Gini importance for the more restricted FI-2 datasets and selected the most significant features to create an even smaller and more reliable set of putatively agronomic trait-associated markers. By using half-sibling families' bulks as a representation of the genetic variability available for breeding, genotyping similar agronomic traits in various clippings and selecting only the most influential markers in the predictions, we were able to minimize the limitations of the method due to small sample size and obtain a reliable set of markers.

In this major importance set, we identified markers associated with multiple phenotypes. Notably, the number of shared markers was more pronounced for GM, DM and LDM, which is in accordance with the observed correlations among phenotype clippings, where SDM and RG exhibited lower correlations with the other traits (Supplementary Figures S3, S6). The high overall correlation, with a mean of 0.72, among traits and the overlap of the identified markers were as expected. This is because the assessed biomass characteristics are highly similar and likely influenced by the same metabolic processes. The DM phenotype was determined by drying the GM material, while SDM/LDM phenotyping involved separating the DM into stems and leaves. Furthermore, biomass production is dependent on the plant's growth capacity (RG). Consequently, the narrow-sense heritabilities of these traits within the families were also very similar. As discussed in other studies, modeling performance is strongly influenced by heritability (Wang et al., 2018; Xu et al., 2018; Murad Leite Andrade et al., 2022). Therefore, our prediction performances did not vary significantly and were correlated with the heritabilities (Supplementary Tables S1-S3).

In the absence of genome annotation, we employed RNA-Seq data in a multiomic approach to map genes physically associated with the major importance markers. Considering the similarity of the agronomic traits employed and the potential involvement of the same biological processes in their regulation, we then conducted a functional analysis that considered the annotation of all genes collectively. This allowed for an overview of the most influential processes governing biomass production and growth.

The enrichment of GO terms related to the annotated genes provides evidence of the methodological capacity to identify QTL regions influencing the evaluated agronomic traits (Supplementary Table S10). Associated with various phenotype clippings, terms related to the lignin biosynthetic process stood out. Previous research has established its significant impact on plant development (Yoon et al., 2015; Bahri et al., 2020). Mutants of lignin biosynthesis genes have shown phenotypes of dwarfism/reduced plant growth (Schilmiller et al., 2009; Li et al., 2009; Song and Wang, 2011), altered morphology (Elkind et al., 1990; Jones et al., 2001; Franke et al., 2002), and tissue browning (Bout and Vermerris, 2003; Xu et al., 2011; Saballos et al., 2012). Furthermore, terms associated with auxin efflux were identified, which are known for their importance in growth regulation. Auxin hormone effects depend on concentration and are primarily produced in meristematic and specific regions (Blakeslee et al., 2005). The transport and distribution of auxin within plant tissues constitutes an essential aspect of its function in plant organogenesis and morphogenesis (Woodward and Bartel, 2005). This transport is facilitated by influx and efflux carrier proteins, providing essential directional and positional cues for various developmental processes, including vascular differentiation, apical dominance, organ development, and tropic growth (Benková et al., 2003; Blancaflor and Masson, 2003; Friml et al., 2003; Blilou et al., 2005; Grieneisen et al., 2007).

Furthermore, the flavonol biosynthetic process, which is another enriched term identified in our results, is known to regulate plant growth and development by controlling auxin transport. Its effects are primarily observed in root elongation, quantity and gravitropic response (Jacobs and Rubery, 1988; Brown et al., 2001; Santelia et al., 2008; Grunewald et al., 2012). Flavonols can influence auxin transportation by different

mechanisms, such as modulating the transcription of genes encoding auxin transport proteins (Peer et al., 2004), acting as kinase inhibitors that regulate the phosphorylation status of auxin transport proteins (Agullo et al., 1997; Peer and Murphy, 2007), or altering the cellular redox state (Fernández-Marcos et al., 2013). In addition, flavonols have antioxidant functions, acting in response to stress such as UV radiation, wounding, drought, metal toxicity, and nutrient deprivation. These conditions lead to the accumulation of reactive oxygen species (ROS), which can damage cellular components and consequently impact plant development (Winkel-Shirley, 2001; Baskar et al., 2018; Agati et al., 2020). The list of terms associated with plant growth, development and stress response continues with the folic acid biosynthetic process (Stakhova et al., 2000; Gorelova et al., 2017), galactolipid metabolic process (Jouhet et al., 2007; Kobayashi et al., 2007; Botté et al., 2011), and cellular response to cold.

The enriched terms provided an overview of the biological function of the identified genes. However, for the genes with multiple copies, limited information was generated, as only three out of the 22 genes had functional annotation. Nevertheless, these three genes appear to have a significant impact on the evaluated agronomic traits. One of these genes, DN91682_c3_g2 (cinnamoyl-CoA reductase 1), which was identified in two copies, is involved in the lignin biosynthetic process (Lauvergeat et al., 2001), circadian rhythm, and response to cold (Carpenter et al., 1994). The second gene, DN64763_c1_g1 (E3 ubiquitin-protein ligase SINAT5), also found in two copies, is known to play key roles in multiple plant developmental stages and several abiotic stress responses (Shu and Yang, 2017). Furthermore, although it has been reported in yeast, the gene DN92072_c5_g1 (AIM25-altered inheritance rate of mitochondria protein 25), which was found in six copies linked to major importance markers, acts in the cellular response to heat and oxidative stress (Aguilar-Lopez et al., 2016) (Figure 3) (Supplementary Table S8).

As a result of diverse mechanisms, such as whole-genome duplication, tandem duplication, and transposon-mediated duplication, plant genomes have an abundance of duplicated genes (Panchy et al., 2016). These duplicate copies can persist for several reasons: insufficient time for the accumulation of deleterious mutations or selection pressure to preserve redundant functions (Panchy et al., 2016). This pressure can arise from four mechanisms: gene dosage increase (Ohno, 1970), subfunctionalization (Force et al., 1999), gene balance (Freeling and Thomas, 2006), and paralog interference (Baker et al., 2013). Beyond identifying multiple copies of genes associated with agronomic traits, further investigation into the mechanisms influencing their retention and how these copies interact and impact the trait may provide valuable insights for improving breeding methods to achieve higher genetic gains.

4.3 Gene coexpression network

We conducted additional multiomic investigations to gain a deeper understanding of the metabolic pathways and regulatory mechanisms that govern the evaluated agronomic traits. We modeled a GCN and isolated the previously identified genes and their coexpressed neighbors (Figure 4). This integration has been successfully employed in different species and has produced noteworthy results (Calabrese et al., 2017; Schaefer et al., 2018; Yan et al., 2020; Francisco et al., 2021). The ability of such networks to simulate complex biological systems and uncover novel biological associations has transformed molecular biology research (D'haeseleer et al., 2000; Liu et al., 2020), enabling the exploration of regulatory relationships, inference of metabolic pathways, and transfer of annotations (Rao and Dixon, 2019). Following the "guilt-by association" principle, GCNs typically reveal interactions among genes with correlated biological functions (Oliver, 2000; Wolfe et al., 2005; Childs et al., 2011). Furthermore, this strategy can contribute to initiatives aimed at exploring targets for molecular perturbations, such as CRISPR. These inferences hold the potential to reveal genes capable of enhancing the loss or gain of functions, thereby influencing phenotypes relevant to breeding programs.

In this context, we could expand our set of identified genes through coexpression analysis, providing broader insight into the metabolic pathways influencing the observed phenotypes. Moreover, the annotated genes within these modules can serve as a basis to infer the biological functions of the unannotated genes. Our network modeling has extended our understanding of genes associated with the previously discussed enriched terms. It has also enabled the identification of new genes involved in biological processes related to DNA integrity, stability and metabolism. These genes act in mismatch repair, telomere capping, and duplex unwinding, all of which are known to impact the normal growth and development of plants to varying degrees (Tuteja, 2003; Kim and Kim, 2018; Karthika et al., 2020). Additionally, our network also expanded the genes involved in regulating the abscisic acid (ABA) transport. Modulating hormone levels within tissues and cells is critical for maintaining a balance between defense mechanisms and growth processes, especially in suboptimal environments. This regulation also plays an important role in controlling stomatal closure (Seo and Koshiba, 2011; Chen et al., 2020). Furthermore, the network has elucidated genes involved in regulating the circadian rhythm. Such a process not only allows plants to adapt to daily environmental changes but also enables them to anticipate and prepare for these challenges in advance (Millar, 2016; Kim et al., 2017; Creux and Harmer, 2019). Notably, the gene ELF4-LIKE 4, a key player in the circadian rhythm (Doyle et al., 2002), stands out as one of the hub genes with the highest degree value in the network (Supplementary Table S12). Finally, we also identified genes related to response to chitin, an important component of the plant immune system activated in the presence of pathogens such as fungi, arthropods, and nematode egg shells (Kombrink et al., 2011; Sánchez-Vallet et al., 2015).

Furthermore, by separating the general agronomic trait network into two seasonal parts, we were able to investigate the potential impact of metabolic processes on plant development and production during both wet and dry periods. In our findings, we identified enriched terms related to auxin efflux and flavonol biosynthetic processes in both networks. These results have already been discussed in the context of auxin transport regulation, indicating the importance of the hormone regardless

of the season. During the wet season, in addition to the previously mentioned abscisic acid transport, we observed terms associated with plant development, such as the isoleucine biosynthetic process (Yu et al., 2013) and the response to nematodes, which are pathogens capable of modifying Plant Physiol., development, metabolism, and immunity (Eves-van den Akker, 2021). In contrast, within the dry network, we found enriched terms related to responses to water deprivation and protein transport. This provides evidence of the metabolic mechanisms required to deal with abiotic stress.

Network degree analysis provided a means to identify hub genes, which are the most highly connected genes in the network. These hubs typically encompass genes with broad regulatory functions or associations with essential roles in biological processes (Carlson et al., 2006; Reverter and Chan, 2008; Amrine et al., 2015). In our analysis, in addition to the previously mentioned ELF4-LIKE 4 protein, we identified several ribosomal protein genes as hubs, such as 40S S6 and S15a-2, 60S L9 and L14-2, 54S L12, and Ubiquitin-40S S27a-1. The heterogeneity of ribosome composition is well-known and forms the foundation of the specialized ribosome theory, which states that different groups of ribosomes are tailored to translate specific sets of mRNAs (Gilbert, 2011; Xue and Barna, 2012; Genuth and Barna, 2018). Although the major discussion in this field is concentrated in elucidating how changes in ribosome composition might facilitate the translation of specific groups of mRNAs (Norris et al., 2021), our results indicate another intriguing aspect of this theory. Although the precise connection between the observed hub ribosomal proteins and the translation of the genes linked to the hubs has yet to be established, we hypothesize that their coexpression may result from a coregulatory mechanism that ensures the availability of specific tailored ribosomes in sufficient quantities for translating the mRNAs of these linked genes. Regarding the relationship between ribosomal proteins and the characteristics evaluated in this research, it has been reported that *A. thaliana* mutants in these proteins are often smaller and have simplified/aberrant vasculature and polarity defects (Van Lijsebettens et al., 1994; Ito et al., 2000; Fujikura et al., 2009; Horiguchi et al., 2011), which can directly impact attributes related to regrowth and biomass production.

In addition, among the highest degree hubs in the network, we found genes associated with lipid metabolism. These specific genes encode important proteins, including the lipid-transfer protein DIR1, fatty acid-binding protein, 3-hydroxyacyl-ACP dehydratase, and 3-Ketoacyl-CoA Synthase 4. These proteins play roles in fatty acid biosynthesis (Supplementary Table S14). Fatty acids, which are common components of complex lipids, are reported to have important roles in plant biology, including cell structure and response to different stresses such as temperature changes (Routaboul et al., 2000; Iba, 2002; Hou et al., 2006), salinity, drought (Mikami and Murata, 2003; Gigon et al., 2004; Zhang et al., 2005), exposure to heavy metals (Verdoni et al., 2001; Chaffai et al., 2007; Maksymiec, 2007), and pathogens (Kachroo et al., 2003; Nandi et al., 2005). Fatty acids, as integral components of cellular membranes, suberin, and cutting waxes (Beisson et al., 2007), contribute to stress resistance by modulating membrane fluidity, releasing α-linolenic acid (Grechkin, 1998), serving as precursors for bioactive molecules (Hou et al., 2016), and acting as modulators of plant defense gene expression (Kachroo et al., 2003).

Another gene with broad activity identified as a hub is 2-oxoglutarate/Fe(II)-dependent dioxygenase (2-ODD) (Supplementary Table S14). This highly versatile enzyme facilitates numerous oxidative reactions, playing a crucial role in biosynthetic pathways for normal organismal function and the production of high value specialized metabolites (Farrow and Facchini, 2014). Its roles extend across various pathways, including DNA repair, histone demethylation, posttranslational modifications, auxin and salicylic acid catabolism, and biosynthesis of gibberellin, ethylene, flavonoid and glucosinolate. 2-ODD is reported to have a significant impact on plant growth and development (Farrow and Facchini, 2014).

Another important aspect of the methodology employed lies in its ability to identify regions associated with known genes linked to specific traits. Equally important is its capacity to elucidate unannotated genes that should be investigated. In our results, more than half of the identified genes linked to the major importance markers lacked functional annotation. Remarkably, some of these unannotated genes seemed to be highly important, as they were observed to have multiple copies and associations with various traits (Figure 3). When we expanded our analysis to the GCN, even more unannotated genes emerged, including important hub genes evidenced by their high degree values (Supplementary Table S14). These genes/regions are important targets to expand the knowledge on the metabolic regulation of agronomic traits and represent valuable information that can be applied in species breeding.

Our work is innovative in different aspects and represents a significant advance in the field of molecular breeding techniques applicable to tropical forages. This study marks the first exploration of the applicability of GWFP in a *Urochloa* species, being the first time that FS and ML algorithms have been employed in GWFP. These techniques not only enhance prediction metrics but also drastically reduce the number of makers required for accurate prediction. Furthermore, employing a multiomic approach, we integrated the selected markers with transcriptome data to construct a coexpression network capable of providing insights into the regulation of plant growth and biomass production in the species. The results demonstrate the great potential of molecular breeding in reducing breeding costs, expediting the release of new cultivars, and facilitating metabolic investigations, even in orphan species with high genomic complexity, such as tropical forages.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

FM: Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing. AA: Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing. AM: Writing –

original draft, Writing – review & editing. RF: Writing – original draft, Writing – review & editing. MV: Resources, Writing – review & editing. MP: Resources, Writing – review & editing. MM: Supervision, Writing – review & editing. RS: Conceptualization, Resources, Supervision, Writing – review & editing. AS: Conceptualization, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by grants from the Fundação de Amparo à Pesquisa de do Estado de São Paulo (FAPESP), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES - Computational Biology Programme and Financial Code 001), Embrapa and UNIPASTO. FM received a PhD fellowship from CAPES (88882.329502/2019-01). AA received a PhD fellowship from FAPESP (2019/03232-6); RF received a PD fellowship from FAPESP (2018/19219-6); and AS received a research fellowship from CNPq.

Acknowledgments

We would like to acknowledge the Fundação de Amparo à Pesquisa de do Estado de São Paulo (FAPESP), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). We also acknowledge the Brazilian Agricultural Research Corporation (Embrapa Gado de Corte) for providing the populations used in this study. This manuscript was previously posted to bioRxiv (https://doi.org/10.1101/2023.09.25.559305).

Conflict of interest

Authors MV, MP, and RS were employed by the company Brazilian Agricultural Research Corporation (Embrapa).

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1303417/full#supplementary-material

References

Abdollahi-Arpanahi, R., Gianola, D., and Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics selection evolution: GSE* 52 (1), 12. doi: 10.1186/s12711-020-00531-z

Agati, G., Brunetti, C., Fini, A., Gori, A., Guidi, L., Landi, M., et al. (2020). Are flavonoids effective antioxidants in plants? Twenty years of our investigation. *Antioxidants* 9, 1098. doi: 10.3390/antiox9111098

Aguilar-Lopez, J. L., Laboy, R., Jaimes-Miranda, F., Garay, E., DeLuna, A., and Funes, S. (2016). Slm35 links mitochondrial stress response and longevity through TOR signaling pathway. *Aging* 8 (12), 3255–3271. doi: 10.18632/aging.101093

Agullo, G., Gamet-Payrastre, L., Manenti, S., Viala, C., Rémésy, C., Chap, H., et al. (1997). Relationship between flavonoid structure and inhibition of phosphatidylinositol 3-kinase: a comparison with tyrosine kinase and protein kinase C inhibition. *Biochem. Pharmacol.* 53, 1649–1657. doi: 10.1016/S0006-2952(97)82453-7

Alexa, A., and Rahnenfuhrer, J. (2022). *topGO: Enrichment Analysis for Gene Ontology. R package version 2.48.0.*

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215 (3), 403–410). doi: 10.1016/s0022-2836(05)80360-2

Amrine, K. C. H., Blanco-Ulate, B., and Cantu, D. (2015). Discovery of core biotic stress responsive genes in Arabidopsis by weighted gene co-expression network analysis. *PloS one* 10 (3), e0118731. doi: 10.1371/journal.pone.0118731

Andrews, S. (2010) *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Aono, A. H., Costa, E. A., Rody, H. V. S., Nagai, J. S., Pimenta, R. J. G., Mancini, M. C., et al. (2020). Machine learning approaches reveal genomic regions associated with sugarcane brown rust resistance. *Sci. Rep.* 10, 20057. doi: 10.1038/s41598-020-77063-5

Aono, A. H., Ferreira, R., Moraes, A., Lara, L., Pimenta, R., Costa, E. A., et al. (2022). A joint learning approach for genomic prediction in polyploid grasses. *Sci. Rep.* 12 (1), 12499. doi: 10.1038/s41598-022-16417-7

Ashraf, B. H., Jensen, J., Asp, T., and Janss, L. L. (2014). Association studies using family pools of outcrossing crops based on allele-frequency estimates from DNA sequencing. TAG. Theoretical and applied genetics. *Theoretische und angewandte Genetik* 127 (6), 1331–1341. doi: 10.1007/s00122-014-2300-4

Azodi, C. B., Pardo, J., VanBuren, R., de los Campos, G., and Shiu, S.-H. (2019). Transcriptome-based prediction of complex traits in maize. *Plant Cell* 32 (1), 139–151. doi: 10.1105/tpc.19.00332

Bahri, B. A., Daverdin, G., Xu, X., Cheng, J. F., Barry, K. W., Brummer, E. C., et al. (2020). Natural variation in lignin and pectin biosynthesis-related genes in switchgrass (Panicum virgatum L.) and association of SNP variants with dry matter traits. *Bioenerg. Res.* 13, 79–99. doi: 10.1007/s12155-020-10090-2

Baker, C. R., Hanson-Smith, V., and Johnson, A. D. (2013). Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science* 342 (6154), 104–108. doi: 10.1126/science.1240810

Barabási, A. L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272

Baskar, V., Venkatesh, R., and Ramalingam, S. (2018). "Flavonoids (Antioxidants systems) in higher plants and their response to stresses," in *Antioxidants and Antioxidant Enzymes in Higher Plants*. Eds. D. Gupta, J. Palma and F. Corpas (Cham: Springer). doi: 10.1007/978-3-319-75088-0_12

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., et al. (2020). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49 (D1), D480–D489. doi: 10.1093/nar/gkaa1100

Bayer, P. E., Petereit, J., Danilevicz, M. F., Anderson, R., Batley, J., and Edwards, D. (2021). The application of pangenomics and machine learning in genomic selection in plants. *Plant Genome* 14 (3), p.e20112. doi: 10.1002/tpg2.20112

Beisson, F., Li, Y., Bonaventure, G., Pollard, M., and Ohlrogge, J. B. (2007). The acyltransferase GPAT5 is required for the synthesis of suberin in seed coat and root of Arabidopsis. *Plant Cell* 19 (1), 351–368. doi: 10.1105/tpc.106.048033

Bélanger, S., Esteves, P., Clermont, I., Jean, M., and Belzile, F. (2016). Genotyping-by-Sequencing on Pooled Samples and its Use in Measuring Segregation Bias during the Course of Androgenesis in Barley. *Plant Genome* 9 (1). doi: 10.3835/plantgenome2014.10.0073

Bellot, P., de los Campos, G., and Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits? *Genetics* 210 (3), 809–819. doi: 10.1534/genetics.118.301298

Benková, E., Michniewicz, M., Sauer, M., Teichmann, T., Seifertová, D., Jürgens, G., et al. (2003). Local, efflux-dependent Auxin gradients as a common module for plant organ formation. *Cell* 115 (5), 591–602. doi: 10.1016/s0092-8674(03)00924-3

Berro, I., Lado, B., Nalin, R. S., Quincke, M., and Gutiérrez, L. (2019). Training population optimization for genomic selection. *Plant Genome* 12, 190028. doi: 10.3835/plantgenome2019.04.0028

Bian, Y., and Holland, J. B. (2017). Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. *Heredity* 118 (6), 585–593. doi: 10.1038/hdy.2017.4

Biazzi, E., Nazzicari, N., Pecetti, L., Brummer, E. C., Palmonari, A., Tava, A., et al. (2017). Genome-wide association mapping and genomic selection for alfalfa (*Medicago sativa*) forage quality traits. *PloS One* 12, e0169234. doi: 10.1371/journal.pone.0169234

Blakeslee, J. J., Peer, W. A., and Murphy, A. S. (2005). Auxin transport. *Curr. Opin. Plant Biol.* 8 (5), 494–500. doi: 10.1016/j.pbi.2005.07.014

Blancaflor, E. B., and Masson, P. H. (2003). Plant gravitropism. Unraveling the ups and downs of a complex process. *Plant Physiol.* 133 (4), 1677–1690. doi: 10.1104/pp.103.032169

Blilou, I., Xu, J., Wildwater, M., Willemsen, V., Paponov, I., Friml, J., et al. (2005). The PIN Auxin efflux facilitator network controls growth and patterning in Arabidopsis roots. *Nature* 433, 7021 39–44. doi: 10.1038/nature03184

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* 30 (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Borin, G. P., Carazzolle, M. F., Dos Santos, R., Riaño-Pachón, D. M., and Oliveira, J. (2018). Gene co-expression network reveals potential new genes related to sugarcane bagasse degradation in *Trichoderma reesei* RUT-30. *Front. bioengineering Biotechnol.* 6. doi: 10.3389/fbioe.2018.00151

Botté, C. Y., Yamaryo-Botté, Y., Janouškovec, J., Rupasinghe, T., Keeling, P. J., Crellin, P., et al. (2011). Identification of plant-like galactolipids in Chromera velia, a photosynthetic relative of malaria parasites. *J. Biol. Chem.* 286 (34), 29893–29903. doi: 10.1074/jbc.M111.254979

Bout, S., and Vermerris, W. (2003). A candidate-gene approach to clone the sorghum Brown midrib gene encoding caffeic acid O-methyltransferase. *Mol. Genet. Genom.* 269, 205–214. doi: 10.1007/s00438-003-0824-4

Breiman, L. (2001) *Bagging predictors. Mach. Learn.* doi: 10.1007/BF00058655.

Brown, D. E., Rashotte, A. M., Murphy, A. S., Normanly, J., Tague, B. W., Peer, W. A., et al. (2001). Flavonoids act as negative regulators of auxin transport in *vivo* in Arabidopsis. *Plant Physiol.* 126, 524–535. doi: 10.1104/pp.126.2.524

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., et al. (2017). A tissue-mapped axolotl *de novo* transcriptome enables identification of limb regeneration factors. *Cell Rep.* 18 (3), 762–776. doi: 10.1016/j.celrep.2016.12.063

Byrne, S., Czaban, A., Studer, B., Panitz, F., Bendixen, C., and Asp, T. (2013). Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing. *PloS One* 8 (3), e57438. doi: 10.1371/journal.pone.0057438

Calabrese, G. M., Mesner, L. D., Stains, J. P., Tommasini, S. M., Horowitz, M. C., Rosen, C. J., et al. (2017). Integrating GWAS and co-expression network data identifies bone mineral density genes SPTBN1 and MARK3 and an osteoblast functional module. *Cell Syst.* 4, 46–59. doi: 10.1016/j.cels.2016.10.014

Cardoso-Silva, C. B., Aono, A. H., Mancini, M. C., Sforça, D. A., da Silva, C. C., Pinto, L. R., et al. (2022). Taxonomically restricted genes are associated with responses to biotic and abiotic stresses in sugarcane (*Saccharum* spp.). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.923069

Carlson, M. R., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., and Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics* 7 (1). doi: 10.1186/1471-2164-7-40

Carpenter, C. D., Kreps, J. A., and Simon, A. E. (1994). Genes encoding glycine-rich arabidopsis thaliana proteins with RNA-binding motifs are influenced by cold treatment and an endogenous circadian rhythm. *Plant Physiol.* 104 (3), 1015–1025. doi: 10.1104/pp.104.3.1015

Cericola, F., Lenk, I., Fè, D., Byrne, S., Jensen, C. S., Pedersen, M. G., et al. (2018). Optimized use of low-depth genotyping-by-sequencing for genomic prediction among multi-parental family pools and single plants in perennial ryegrass (Lolium perenne L.). *Front. Plant Sci.* 9, 369. doi: 10.3389/fpls.2018.00369

Chaffai, R., Elhammadi, M. A., Seybou, T. N., Tekitek, A., Marzouk, B., and El Ferjani, E. (2007). Altered fatty acid profile of polar lipids in maize seedlings in response to excess copper. *J. Agron. Crop Sci.* 193 (3), 207–217. doi: 10.1111/j.1439-037x.2007.00252.x

Chen, T., and Guestrin, C. (2016). "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. (New York: ACM), 785–794.

Chen, K., Li, G., Bressan, R. A., Song, C., Zhu, J., and Zhao, Y. (2020). Abscisic acid dynamics, signaling, and functions in plants. *J. Integr. Plant Biol.* 62 (1), 25–54. doi: 10.1111/jipb.12899

Childs, K. L., Davidson, R. M., and Buell, C. R. (2011). Gene coexpression network analysis as a source of functional annotation for rice genes. *PloS One* 6, e22196. doi: 10.1371/journal.pone.0022196

Chung, C. W., Hsiao, T. H., Huang, C. J., Chen, Y. J., Chen, H. H., Lin, C. H., et al. (2021). Machine learning approaches for the genomic prediction of rheumatoid arthritis and systemic lupus erythematosus. *BioData Min.* 14, 52. doi: 10.1186/s13040-021-00284-5

Colombari-Filho, J. M., Resende, M. D. V., Morais, O. P., Castro, A. P., Guimarães, E. P., Pereira, J. A., et al. (2013). Upland rice breeding in Brazil: a simultaneous genotypic evaluation of stability, adaptability and grain yield. *Euphytica* 192, 117–129. doi: 10.1007/s10681-013-0922-2

Creux, N., and Harmer, S. (2019). Circadian rhythms in plants. *Cold Spring Harb. Perspect. Biol.* 11 (9), a034611. doi: 10.1101/cshperspect.a034611

Cristianini, N., and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods* (Cambridge University Press).

Crossa, J., Martini, J. W. R., Gianola, D., Pérez-Rodríguez, P., Jarquin, D., Juliana, P., et al. (2019). Deep kernel and deep learning for genome-based prediction of single traits in multienvironment breeding trials. *Front. Genet.* 10. doi: 10.3389/fgene.2019.01168

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

Daetwyler, H. D., Calus, M. P., Pong-Wong, R., de Los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193 (2), 347–365. doi: 10.1534/genetics.112.147983

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27 (15), 2156–2158. doi: 10.1093/bioinformatics/btr330

de Bem Oliveira, I., Amadeu, R. R., Ferrão, L. F. V., and Muñoz, P. R. (2020). Optimizing whole-genomic prediction for autotetraploid blueberry breeding. *Heredity (Edinb).* 125 (6), 437–448. doi: 10.1038/s41437-020-00357-x

de C. Lara, L. A., Santos, M. F., Jank, L., Chiari, L., Vilela, M., de, M., et al. (2019). Genomic selection with allele dosage in panicum maximum Jacq. *G3 Genes|Genomes|Genetics* 9 (8), 2463–2475. doi: 10.1534/g3.118.200986

De Mendiburu, F., and De Mendiburu, M. F. (2020). *Package A'gricolae'. R package version*. 1–2.

D'haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707–726. doi: 10.1093/bioinformatics/16.8.707

Dhingani, R. M., Umrania, V. V., Tomar, R. S., Parakhia, M. V., and Golakiya, B. (2015). Introduction to QTL mapping in plants. *Ann. Plant Sci.* 4 (04), 1072–1079.

Doyle, M. R., Davis, S. J., Bastow, R. M., McWatters, H. G., Kozma-Bognar, L., Nagy, F., et al. (2002). The ELF4 gene controls circadian rhythms and flowering time in Arabidopsis thaliana. *Nature* 419, 74–77. doi: 10.1038/nature00954

Edwards, S. M., Buntjer, J. B., Jackson, R., Bentley, A. R., Lage, J., Byrne, E., et al. (2019). The effects of training population design on genomic prediction accuracy in wheat. *Theor. Appl. Genet.* 132, 1943–1952. doi: 10.1007/s00122-019-03327-y

Elkind, Y., Edwards, R., Mavandad, M., Hedrick, S. A., Ribak, O., Dixon, R. A., et al. (1990). Abnormal plant development and down-regulation of phenylpropanoid biosynthesis in transgenic tobacco containing a heterologous phenylalanine ammonia-lyase gene. *Proc. Natl. Acad. Sci.* 87 (22), 9057–9061. doi: 10.1073/pnas.87.22.9057

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6, e19379. doi: 10.1371/journal.pone.0019379

Eves-van den Akker, S. (2021). Plant–nematode interactions. *Curr. Opin. Plant Biol.* 62, 102035. doi: 10.1016/j.pbi.2021.102035

Farrow, S. C., and Facchini, P. J. (2014). Functional diversity of 2-oxoglutarate/Fe(II)-dependent dioxygenases in plant metabolism. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00524

Fè, D., Cericola, F., Byrne, S., Lenk, I., Ashraf, B. H., Pedersen, M. G., et al. (2015). Genomic dissection and prediction of heading date in perennial ryegrass. *BMC Genomics* 16, 921. doi: 10.1186/s12864-015-2163-3

Fernández-Marcos, M., Sanz, L., Lewis, D. R., Muday, G. K., and Lorenzo, O. (2013). "Control of auxin transport by reactive oxygen and nitrogen species," in *Polar Auxin Transport, Signaling and Communication in Plants*, vol. 17 . Eds. R. Chen and F. Baluska (Berlin: Springer-Verlag), 103–117.

Ferrão, L., Amadeu, R. R., Benevenuto, J., de Bem Oliveira, I., and Munoz, P. R. (2021). Genomic selection in an outcrossing autotetraploid fruit crop: lessons from blueberry breeding. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.676326

Ferreira, R. C. U., Cançado, L. J., Do Valle, C. B., Chiari, L., and de Souza, A. P. (2016). Microsatellite loci for Urochloa decumbens (Stapf) R.D. Webster and cross-amplification in other Urochloa species. *BMC. Res. Notes* 9, 152. doi: 10.1186/s13104-016-1967-9

Ferreira, R. C. U., da Costa Lima Moraes, A., Chiari, L., Simeão, R. M., Vigna, B. B. Z., and de Souza, A. P. (2021). An overview of the genetics and genomics of the *Urochloa* species most commonly used in pastures. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.770461

Figueiredo, U. J., Nunes, J. A. R., and do Valle, C. B. (2012). Estimation of genetic parameters and selection of Brachiaria humidicola progenies using a selection index. *Crop Breed. Appl. Biotechnol.* 12 (4), 237–244. doi: 10.1590/s1984-70332012000400002

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151 (4), 1531–1545. doi: 10.1093/genetics/151.4.1531

Francisco, F. R., Aono, A. H., da Silva, C. C., Gonçalves, P. S., Scaloppi Junior, E. J., Le Guen, V., et al. (2021). Unravelling rubber tree growth by integrating GWAS and biological network-based approaches. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.768589

Franke, R., Hemm, M. R., Denault, J. W., Ruegger, M. O., Humphreys, J. M., and Chapple, C. (2002). Changes in secondary metabolism and deposition of an unusual lignin in the ref8 mutant of Arabidopsis. *Plant J.* 30 (1), 47–59. doi: 10.1046/j.1365-313x.2002.01267.x

Freeling, M., and Thomas, B. C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 16 (7), 805–814. doi: 10.1101/gr.3681406

Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55, 119–139. doi: 10.1006/jcss.1997.1504

Friml, J., Vieten, A., Sauer, M., Weijers, D., Schwarz, H., Hamann, T., et al. (2003). Efflux-dependent auxin gradients establish the apical-basal axis of Arabidopsis. *Nature* 426, 147–153. doi: 10.1038/nature02085

Fujikura, U., Horiguchi, G., Ponce, M. R., Micol, J. L., and Tsukaya, H. (2009). Coordination of cell proliferation and cell expansion mediated by ribosome-related processes in the leaves of Arabidopsis thaliana. *Plant J.* 59 (3), 499–508. doi: 10.1111/j.1365-313X.2009.03886.x

Futschik, A., and Schloütterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* 186 (1), 207–218. doi: 10.1534/genetics.110.114397

Genuth, N. R., and Barna, M. (2018). The discovery of ribosome heterogeneity and its implications for gene regulation and organismal life. *Mol. Cell* 71 (3), 364–374. doi: 10.1016/j.molcel.2018.07.018

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn. J.* 63, 3–42. doi: 10.1007/s10994-006-6226-1

Gigon, A., Matos, A.-R., Laffray, D., Zuily-Fodil, Y., and Pham-Tthi, A.-T. (2004). Effect of drought stress on lipid metabolism in the leaves of Arabidopsis thaliana (Ecotype Columbia). *Ann. Bot.* 94 (3), 345–351. doi: 10.1093/aob/mch150

Gilbert, W. V. (2011). Functional specialization of ribosomes? *Trends Biochem. Sci.* 36 (3), 127–132. doi: 10.1016/j.tibs.2010.12.002

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PloS One* 9, e90346. doi: 10.1371/journal.pone.0090346

Goddard, M. E., Kemper, K. E., MacLeod, I. M., Chamberlain, A. J., and Hayes, B. J. (2016). Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proc. Biol. Sci.* 283 (1835), 20160569. doi: 10.1098/rspb.2016.0569

Gorelova, V., Ambach, L., Rébeillé, F., Stove, C., and van der Straeten, D. (2017). Folates in plants: Research advances and progress in crop biofortification. *Front. Chem.* 5, 21. doi: 10.3389/fchem.2017.00021

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29 (7), 644–652. doi: 10.1038/nbt.1883

Granato, I., Cuevas, J., Luna-Vázquez, F., Crossa, J., Montesinos-López, O., Burgueño, J., et al. (2018). BGGE: A new package for genomic-enabled prediction incorporating genotype × Environment interaction models. *G3 Genes|Genomes|Genetics* 8 (9), 3039–3047. doi: 10.1534/g3.118.200435

Grechkin, A. (1998). Recent developments in biochemistry of the plant lipoxygenase pathway. *Progress Lipid Res* 37 (5), 317–352. doi: 10.1016/s0163-7827(98)00014-9

Grieneisen, V. A., Xu, J., Maree, A. F., Hogeweg, P., and Scheres, B. (2007). Auxin transport is sufficient to generate a maximum and gradient guiding root growth. *Nature* 449, 1008–1013. doi: 10.1038/nature06215

Grinberg, N. F., Lovatt, A., Hegarty, M., Lovatt, A., Skøt, K. P., Kelly, R., et al. (2016). Implementation of genomic prediction in Lolium perenne (L.) breeding populations. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00133

Grunewald, W., De Smet, I., Lewis, D. R., Löfke, C., Jansen, L., Goeminne, G., et al. (2012). Transcription factor WRKY23 assists auxin distribution patterns during Arabidopsis root development through local control on flavonol biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* 109, 1554–1559. doi: 10.1073/pnas.1121134109

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. *Bioinformatics* 30 (19), 2811–2812. doi: 10.1093/bioinformatics/btu393

Guo, X., Cericola, F., Fè, D., Pedersen, M. G., Lenk, I., Jensen, C. S., et al (2018). Genomic prediction in Tetraploid ryegrass using allele frequencies based on genotyping by sequencing. *Front. Plant Sci.* 9, 1165. doi: 10.3389/fpls.2018.01165

Haile, T. A., Walkowiak, S., N'Diaye, A., Clarke, J. M., Hucl, P. J., Cuthbert, R. D., et al. (2021). Genomic prediction of agronomic traits in wheat using different models and cross-validation designs. *Theor. Appl. Genet.* 134 (1), 381–398. doi: 10.1007/s00122-020-03703-z

Hanley, S. J., Pellny, T. K., de Vega, J. J., Castiblanco, V., Arango, J., Eastmond, P. J., et al. (2021). Allele mining in diverse accessions of tropical grasses to improve forage quality and reduce environmental impact. *Ann. Bot.* 128 (5), 627–637. doi: 10.1093/aob/mcab101

Heer, K., Behringer, D., Piermattei, A., Bässler, C., Brandl, R., Fady, B., et al. (2018). Linking dendroecology and association genetics in natural populations: Stress responses archived in tree rings associate with SNP genotypes in silver fir (Abies albaMill.). *Mol. Ecol.* 27 (6), 1428–1438. doi: 10.1111/mec.14538

Horiguchi, G., Mollá-Morales, A., Pérez-Pérez, J. M., Kojima, K., Robles, P., Ponce, M. R., et al. (2011). Differential contributions of ribosomal protein genes to Arabidopsis thaliana leaf development. *Plant J.* 65 (5), 724–736. doi: 10.1111/j.1365-313X.2010.04457.x

Hou, G., Ablett, G. R., Pauls, K. P., and Rajcan, I. (2006). Environmental effects on fatty acid levels in soybean seed oil. *J. Am. Oil Chemists' Soc.* 83 (9), 759–763. doi: 10.1007/s11746-006-5011-4

Hou, Q., Ufer, G., and Bartels, D. (2016). Lipid signalling in plant responses to abiotic stress. *Plant Cell Environ.* 39 (5), 1029–1048. doi: 10.1111/pce.12666

Iba, K. (2002). Acclimative response to temperature stress in higher plants: Approaches of Gene Engineering for Temperature Tolerance. *Annu. Rev. Plant Biol.* 53 (1), 225–245. doi: 10.1146/annurev.arplant.53.100201.160729

Islam, M. S., McCord, P. H., Olatoye, M. O., Qin, L., Sood, S., Lipka, A. E., et al. (2021). Experimental evaluation of genomic selection prediction for rust resistance in sugarcane. *Plant Genome* 14 (3). doi: 10.1002/tpg2.20148

Ito, T., Kim, G. T., and Shinozaki, K. (2000). Disruption of an Arabidopsis cytoplasmic ribosomal protein S13-homologous gene by transposon-mediated mutagenesis causes aberrant growth and development. *Plant J.* 22 (3), 257–264. doi: 10.1046/j.1365-313x.2000.00728.x

Jacobs, M., and Rubery, P. H. (1988). Naturally-occurring auxin transport regulators. *Science* 241, 346–349. doi: 10.1126/science.241.4863.346

Jank, L., Barrios, S. C., do Valle, C. B., Simeão, R. M., and Alves, G. F. (2014). The value of improved pastures to Brazilian beef production. *Crop Pasture Sci.* 65, 1132–1137. doi: 10.1071/CP13319

Jeong, S., Kim, J. Y., and Kim, N. (2020). GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. *Sci. Rep.* 10 (1), 19653. doi: 10.1038/s41598-020-76759-y

Jia, C., Zhao, F., Wang, X., Han, J., Zhao, H., Liu, G., et al. (2018). Genomic prediction for 25 agronomic and quality traits in alfalfa (Medicago sativa). *Front. Plant Sci.* 9, 1220. doi: 10.3389/fpls.2018.01220

Jones, C., De Vega, J., Worthington, M., Thomas, A., Gasior, D., Harper, J., et al. (2021). A comparison of differential gene expression in response to the onset of water stress between three hybrid Brachiaria genotypes. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.637956

Jones, L., Ennos, A. R., and Turner, S. R. (2001). Cloning and characterization of irregular xylem4 (irx4): a severely lignin-deficient mutant of Arabidopsis. *Plant J.* 26 (2), 205–216. doi: 10.1046/j.1365-313x.2001.01021.x

Jouhet, J., Maréchal, E., and Block, M. A. (2007). Glycerolipid transfer for the building of membranes in plant cells. *Prog. Lipid Res.* 46 (1), 37–55. doi: 10.1016/j.plipres.2006.06.002

Juliana, P., He, X., Marza, F., Islam, R., Anwar, B., Poland, J., et al. (2022). Genomic selection for wheat blast in a diversity panel, breeding panel and full-sibs panel. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.745379

Kachroo, A., Lapchyk, L., Fukushige, H., Hildebrand, D., Klessig, D., and Kachroo, P. (2003). Plastidial fatty acid signaling modulates salicylic acid– and jasmonic acid–mediated defense pathways in the Arabidopsis ssi2 mutant. *Plant Cell* 15 (12), 2952–2965. doi: 10.1105/tpc.017301

Karthika, V., Babitha, K. C., Kiranmai, K., Shankar, A. G., Vemanna, R. S., and Udayakumar, M. (2020). Involvement of DNA mismatch repair systems to create genetic diversity in plants for speed breeding programs. *Plant Physiol. Rep.* 25, 185–199. doi: 10.1007/s40502-020-00521-9

Kim, M. K., and Kim, W. T. (2018). Telomere structure, function, and maintenance in plants. *J. Plant Biol.* 61 (3), 131–136. doi: 10.1007/s12374-018-0082-y

Kim, J., Kim, H.-S., Choi, S.-H., Jang, J.-Y., Jeong, M.-J., and Lee, S. (2017). The importance of the circadian clock in regulating plant metabolism. *Int. J. Mol. Sci.* 18 (12), 2680. doi: 10.3390/ijms18122680

Kobayashi, K., Kondo, M., Fukuda, H., Nishimura, M., and Ohta, H. (2007). Galactolipid synthesis in chloroplast inner envelope is essential for proper thylakoid biogenesis, photosynthesis, and embryogenesis. *Proc. Natl. Acad. Sci.* 104 (43), 17216–17221. doi: 10.1073/pnas.0704680104

Kombrink, A., Sánchez-Vallet, A., and Thomma, B. P. H. J. (2011). The role of chitin detection in plant–pathogen interactions. *Microbes Infection* 13 (14–15), 1168–1176. doi: 10.1016/j.micinf.2011.07.010

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9, 29. doi: 10.1186/1746-4811-9-29

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9 (4), 357–359. doi: 10.1038/nmeth.1923

Lauvergeat, V., Lacomme, C., Lacombe, E., Lasserre, E., Roby, D., and Grima-Pettenati, J. (2001). Two cinnamoyl-CoA reductase (CCR) genes from Arabidopsis thaliana are differentially expressed during development and in response to infection with pathogenic bacteria. *Phytochemistry* 57 (7), 1187–1195. doi: 10.1016/s0031-9422(01)00053-x

Lello, L., Avery, S. G., Tellier, L., Vazquez, A. I., de los Campos, G., and Hsu, S. D.H. (2018). Accurate genomic prediction of human height. *Genetics* 210 (2), 477–497. doi: 10.1534/genetics.118.301267.

Li, X., Yang, Y., Yao, J., Chen, G., Li, X., Zhang, Q., et al. (2009). FLEXIBLE CULM 1 encoding a cinnamyl-alcohol dehydrogenase controls culm mechanical strength in rice. *Plant Mol. Biol.* 69, 685–697. doi: 10.1007/s11103-008-9448-8

Li, B., Zhang, N., Wang, Y.-G., George, A. W., Reverter, A., and Li, Y. (2018). Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet.* 9. doi: 10.3389/fgene.2018.00237

Liang, M., Chang, T., An, B., Duan, X., Du, L., Wang, X., et al. (2021). A stacking ensemble learning framework for genomic prediction. *Front. Genet.* 12. doi: 10.3389/fgene.2021.600040

Liang, M., Miao, J., Wang, X., Chang, T., An, B., Duan, X., et al. (2020). Application of ensemble learning to genomic selection in chinese simmental beef cattle. *J. Anim. Breed. Genet.* 138 (3), 291–299. doi: 10.1111/jbg.12514

Lipka, A. E., Lu, F., Cherney, J. H., Buckler, E. S., Casler, M. D., and Costich, D. E. (2014). Accelerating the switchgrass (Panicum virgatum L.) breeding cycle using genomic selection approaches. *PloS One* 9 (11), e112227. doi: 10.1371/journal.pone.0112227

Liu, C., Ma, Y., Zhao, J., Nussinov, R., Zhang, Y. C., Cheng, F., et al. (2020). Computational network biology: data, models, and applications. *Phys. Rep.* 846, 1–66. doi: 10.1016/j.physrep.2019.12.004

Lorenz, A. J., and Smith, K. P. (2015). Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop Sci.* 55 (6), 2657–2667. doi: 10.2135/cropsci2014.12.0827

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., et al. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* 248, 1307–1318. doi: 10.1007/s00425-018-2976-9

Maksymiec, W. (2007). Signaling responses in plants to heavy metal stress. *Acta Physiologiae Plantarum* 29 (3), 177–187. doi: 10.1007/s11738-007-0036-3

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* 38 (10), 4647–4654. doi: 10.1093/molbev/msab199

Martins, F. B., Moraes, A. C. L., Aono, A. H., Ferreira, R. C. U., Chiari, L., Simeão, R. M., et al. (2021). A semi-automated SNP-based approach for contaminant identification in Biparental polyploid populations of tropical forage grasses. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.737919

Mateescu, R. G., Garrick, D. J., and Reecy, J. M. (2017). Network analysis reveals putative genes affecting meat quality in Angus cattle. *Front. Genet.* 8. doi: 10.3389/fgene.2017.00171

Matias, F. I., Alves, F. C., Meireles, K. G. X., Barrios, S. C. L., do Valle, C. B., Endelman, J. B., et al. (2019a). On the accuracy of genomic prediction models considering multi-trait and allele dosage in Urochloa spp. *interspecific tetraploid hybrids. Mol. Breed.* 39 (7), 1–16. doi: 10.1007/s11032-019-1002-7

Matias, F. I., Vidotti, M. S., Meireles, K. G. X., Barrios, S. C. L., do Valle, C. B., Carley, C. A. S., et al. (2019b). Association mapping considering allele dosage: an example of forage traits in an interspecific segmental allotetraploid Urochloa spp. panel. *Crop Sci.* 59, 2062–2076. doi: 10.2135/cropsci2019.03.0185

Medina, C. A., Kaur, H., Ray, I., and Yu, L.-X. (2021). Strategies to increase prediction accuracy in genomic selection of complex traits in alfalfa (Medicago sativa L.). *Cells* 10 (12), 3372. doi: 10.3390/cells10123372

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Mikami, K., and Murata, N. (2003). Membrane fluidity and the perception of environmental signals in cyanobacteria and plants. *Prog. Lipid Res.* 42 (6), 527–543. doi: 10.1016/s0163-7827(03)00036-5

Millar, A. J. (2016). The intracellular dynamics of circadian clocks reach for the light of ecology and evolution. *Annu. Rev. Plant Biol.* 67, 595–618. doi: 10.1146/annurev-arplant-043014-115619

Mohan, M., Nair, S., Bhagwat, A., Krishna, T. G., Yano, M., Bhatia, C. R., et al. (1997). Genome mapping, molecular markers and marker-assisted selection in crop plants. *Mol. Breed.* 3 (2), 87–103. doi: 10.1023/A:1009651919792

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Barrón-López, J. A., Martini, J. W., Fajardo-Flores, S. B., et al. (2021). A review of deep learning applications for genomic selection. *BMC Genomics* 22, 19. doi: 10.1186/s12864-020-07319-x

Montesinos-López, O. A., Montesinos-López, A., Tuberosa, R., Maccaferri, M., Sciara, G., Ammar, K., et al. (2019). Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01311

Murad Leite Andrade, M. H., Acharya, J. P., Benevenuto, J., de Bem Oliveira, I., Lopez, Y., Munoz, P., et al. (2022). Genomic prediction for canopy height and dry matter yield in alfalfa using family bulks. *Plant Genome* 15 (3), e20235. doi: 10.1002/tpg2.20235

Mutwil, M., Usadel, B., Schutte, M., Loraine, A., Ebenhoh, O., Persson, S., et al. (2009). Assembly of an interactive correlation network for the arabidopsis genome using a novel heuristic clustering algorithm. *Plant Physiol.* 152 (1), 29–43. doi: 10.1104/pp.109.145318

Nandi, A., Moeder, W., Kachroo, P., Klessig, D. F., and Shah, J. (2005). Arabidopsis ssi2-Conferred Susceptibility to Botrytis cinerea Is Dependent on EDS5 and PAD4. *Mol. Plant-Microbe Interactions®* 18 (4), 363–370. doi: 10.1094/mpmi-18-0363

Norris, K., Hopes, T., and Aspden, J. L. (2021). Ribosome heterogeneity and specialization in development. *WIREs RNA* 12 (4). doi: 10.1002/wrna.1644

Ohno, S. (1970). *Evolution by Gene Duplication* (New York: Springer-Verlag).

Oliver, S. (2000). Guilt-by-association goes global. *Nature* 403, 601–602. doi: 10.1038/35001165

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. (2016). Evolution of gene duplication in plants. *Plant Physiol.* 171 (4), 2294–2316. doi: 10.1104/pp.16.00523

Parker Gaddis, K. L., Null, D. J., and Cole, J. B. (2016). Explorations in genome-wide association studies and network analyses with dairy cattle fertility traits. *J. dairy Sci.* 99 (8), 6420–6435. doi: 10.3168/jds.2015-10444

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14 (4), 417–419. doi: 10.1038/nmeth.4197

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Peer, W. A., Bandyopadhyay, A., Blakeslee, J. J., Makam, S. N., Chen, R. J., Masson, P. H., et al. (2004). Variation in expression and protein localization of the PIN family of auxin efflux facilitator proteins in flavonoid mutants with altered Auxin transport in Arabidopsis thaliana. *Plant Cell* 16, 1898–1911. doi: 10.1105/tpc.021501

Peer, W. A., and Murphy, A. S. (2007). Flavonoids and Auxin transport: modulators or regulators? *Trends Plant Sci.* 12, 556–563. doi: 10.1016/j.tplants.2007.10.003

Pereira, J. F., Azevedo, A. L. S., Pessoa-Filho, M., Romanel, E. A. D. C., Pereira, A. V., Vigna, B. B. Z., et al. (2018b). Research priorities for next-generation breeding of tropical forages in Brazil. *Crop Breed. Appl. Biotechnol.* 18, 314–319. doi: 10.1590/1984-70332018v18n3n46

Pereira, G. S., Garcia, A. A. F., and Margarido, G. R. A. (2018a). A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. *BMC Bioinform.* 19, 398. doi: 10.1186/s12859-018-2433-6

Perez, P., and de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198 (2), 483–495. doi: 10.1534/genetics.114.164442

Pessoa-Filho, M., Sobrinho, F. S., Fragoso, R. R., Silva Junior, O. B., and Ferreira, M. E. (2019). "A phased diploid genome assembly for the forage grass urochloa ruziziensis based on single-molecule real-time sequencing," in *Plant and Animal Genome Conference*. (Livingston, NJ: Scherago), 27. Available at: https://www.embrapa.br/en/busca-de-publicacoes/-/publicacao/1107378/a-phased-diploid-genome-assembly-for-the-forage-grass-urochloa-ruziziensis-based-on-single-molecule-real-time-sequencing.

Petrasch, S., Mesquida-Pesci, S. D., Pincot, D. D. A., Feldmann, M. J., López, C. M., Famula, R., et al. (2022). Genomic prediction of strawberry resistance to postharvest fruit decay caused by the fungal pathogen Botrytis cinerea. *G3 (Bethesda)* 12 (1), jkab378. doi: 10.1093/g3journal/jkab378

Pimenta, R. J. G., Aono, A. H., Burbano, R. C. V., da Silva, M. F., dos Anjos, I. A., de Andrade Landell, M. G., et al. (2022). Multiomic investigation of sugarcane mosaic virus resistance in sugarcane. *Cold Spring Harbor Lab*. doi: 10.1101/2022.08.18.504288

Pimenta, R. J. G., Aono, A. H., Burbano, R. C. V., Coutinho, A. E., da Silva, C. C., Dos Anjos, I. A., et al. (2021). Genome-wide approaches for the identification of markers and genes associated with sugarcane yellow leaf virus resistance. *Sci. Rep.* 11, 15730. doi: 10.1038/s41598-021-95116-1

Pincot, D. D. A., Hardigan, M. A., Cole, G. S., Famula, R. A., Henry, P. M., Gordon, T. R., et al. (2020). Accuracy of genomic selection and long–term genetic gain for resistance to *Verticillium* wilt in strawberry. *Plant Genome* 13, e20054. doi: 10.1002/tpg2.20054

Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PloS One* 7 (2), e32253. doi: 10.1371/journal.pone.0032253

Popescu, M. C., Balas, V., Perescu-Popescu, L., and Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Trans. Circuits Syst.* 8, 579–588.

Rao, X., and Dixon, R. A. (2019). Co-expression networks for plant biology: why and how. *Acta Biochim. Biophys. Sin.* 51 (10), 981–988. doi: 10.1093/abbs/gmz080

R Core Team (2021). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing). Available at: https://www.R-project.org/.

Resende, M. D. V. (2002). *Software Selegen – REML/BLUP* (Colombo-Brazil: Embrapa Florestas).

Reverter, A., and Chan, E. K. F. (2008). Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics* 24 (21), 2491–2497. doi: 10.1093/bioinformatics/btn482

Rios, E. F., Andrade, M. H. M. L., Resende, M. F. R. C.OMMAJ.R.X.X.X, Kirst, M., de Resende, M. D. V., de Almeida Filho, J. E., et al. (2021). Genomic prediction in family

bulks using different traits and cross-validations in pine. *G3 Genes|Genomes|Genetics* 11 (9). doi: 10.1093/g3journal/jkab249

Rosolen, R. R., Aono, A. H., Almeida, D. A., Ferreira Filho, J. A., Horta, M., and De Souza, A. P. (2022). Network analysis reveals different cellulose degradation strategies across *Trichoderma harzianum* strains associated with XYR1 and CRE1. *Front. Genet.* 13. doi: 10.3389/fgene.2022.807243

Routaboul, J.-M., Fischer, S. F., and Browse, J. (2000). Trienoic fatty acids are required to maintain chloroplast function at low temperatures. *Plant Physiol.* 124 (4), 1697–1705. doi: 10.1104/pp.124.4.1697

Saballos, A., Sattler, S. E., Sanchez, E., Foster, T. P., Xin, Z., Kang, C., et al. (2012). Brown midrib2 (Bmr2) encodes the major 4-coumarate:coenzyme A ligase involved in lignin biosynthesis in sorghum (Sorghum bicolor (L.) Moench). *Plant J.* 70 (5), 818–830. doi: 10.1111/j.1365-313x.2012.04933.x

Salgado, L. R., Lima, R., Santos, B. F. D., Shirakawa, K. T., Vilela, M. D. A., Almeida, N. F., et al. (2017). *De novo* RNA sequencing and analysis of the transcriptome of signalgrass (Urochloa decumbens) roots exposed to aluminum. *Plant Growth Regul.* 83, 157–170. doi: 10.1007/s10725-017-0291-2

Sánchez-Vallet, A., Mesters, J. R., and Thomma, B. P. H. J. (2015). The battle for chitin recognition in plant-microbe interactions. *FEMS Microbiol. Rev.* 39 (2), 171–183. doi: 10.1093/femsre/fuu003

Sandhu, K., Aoun, M., Morris, C., and Carter, A. (2021). Genomic selection for end-use quality and processing traits in soft white winter wheat breeding program with machine and deep learning models. *Biology* 10 (7), 689. doi: 10.3390/biology10070689

Santelia, D., Henrichs, S., Vincenzetti, V., Sauer, M., Bigler, L., Klein, M., et al. (2008). Flavonoids redirect PIN-mediated polar auxin fluxes during root gravitropic responses. *J. Biol. Chem.* 283, 31218–31226. doi: 10.1074/jbc.M710122200

Schaefer, R. J., Michno, J. M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., et al. (2018). Integrating coexpression networks with GWAS to prioritize causal genes in maize. *Plant Cell* 30, 2922–2942. doi: 10.1105/tpc.18.00299

Schilmiller, A. L., Stout, J., Weng, J.-K., Humphreys, J., Ruegger, M. O., and Chapple, C. (2009). Mutations in the Cinnamate 4-hydroxylase gene impact metabolism, growth and development in Arabidopsis. *Plant J.* 60 (5), 771–782. doi: 10.1111/j.1365-313x.2009.03996.x

Schneider, M., Shrestha, A., Ballvora, A., and Léon, J. (2022). High-throughput estimation of allele frequencies using combined pooled-population sequencing and haplotype-based data processing. *Plant Methods* 18 (1), 34. doi: 10.1186/s13007-022-00852-8

Scossa, F., Alseekh, S., and Fernie, A. R. (2021). Integrating multi-omics data for crop improvement. *J. Plant Physiol.* 257, 153352. doi: 10.1016/j.jplph.2020.153352

Seo, M., and Koshiba, T. (2011). Transport of ABA from the site of biosynthesis to the site of action. *J. Plant Res.* 124, 501–507. doi: 10.1007/s10265-011-0411-4

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13 (11), 2498–2504. doi: 10.1101/gr.1239303

Shu, K., and Yang, W. (2017). E3 ubiquitin ligases: ubiquitous actors in plant development and abiotic stress responses. *Plant Cell Physiol.* 58 (9), 1461–1476. doi: 10.1093/pcp/pcx071

Simeão, R. M., Resende, M. D. V., Alves, R. S., Pessoa-Filho, M., Azevedo, A. L. S., Jones, C. S., et al. (2021). Genomic selection in tropical forage grasses: current status and future applications. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.665195

Simeão, R. M., do Valle, C. B., Alves, G. F., Moreira, D. A. L., da Silva, D. R., Araújo D de, F., et al. (2012). Melhoramento de Brachiaria ruziziensis tetraploide sexual na Embrapa: métodos e avanços. *Embrapa Campo Grande. Documentos* 194, 1–32.

Simeão, R., Silva, A., Valle, C., Resende, M. D., and Medeiros, S. (2016a). Genetic evaluation and selection index in tetraploid Brachiaria ruziziensis. *Plant Breed* 135 (2), 246–253. doi: 10.1111/pbr.12353

Simeão, R. M., Valle, C. B., and Resende, M. D. V. (2016b). Unravelling the inheritance,QSTand reproductive phenology attributes of the tetraploid tropical grass Brachiaria ruziziensis(Germain et Evrard). *Plant Breed.* 136 (1), 101–110. doi: 10.1111/pbr.12429

Simeão-Resende, R. M., Casler, M. D., and Resende, M. D. V. (2014). Genomic selection in forage breeding: accuracy and methods. *Crop Sci.* 54, 143–156. doi: 10.2135/cropsci2013.05.0353

Soneson, C., Love, M. I., and Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* 4. doi: 10.12688/f1000research.7563.1

Song, J., and Wang, Z. (2011). RNAi-mediated suppression of the phenylalanine ammonia-lyase gene in Salvia miltiorrhiza causes abnormal phenotypes and a reduction in rosmarinic acid biosynthesis. *J. Plant Res.* 124, 183–192. doi: 10.1007/s10265-010-0350-5

Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcaMethods a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* 23, 1164–1167. doi: 10.1093/bioinformatics/btm069

Stakhova, L., Stakhov, L., and Ladygin, V. (2000). Effects of exogenous folic acid on the yield and amino acid content of the seed of Pisum sativum L. and Hordeum vulgare L. *Appl. Biochem. Microbiol.* 36, 85–89. doi: 10.1007/BF02738142

Steinfath, M., Gärtner, T., Lisec, J., Meyer, R. C., Altmann, T., Willmitzer, L., et al. (2010). Prediction of hybrid biomass in Arabidopsis thaliana by selected parental SNP and metabolic markers. *Theor. Appl. Genet.* 120, 239–247. doi: 10.1007/s00122-009-1191-2

Thaikua, S., Ebina, M., Yamanaka, N., Shimoda, K., Suenaga, K., and Kawamoto, Y. (2016). Tightly clustered markers linked to an apospory-related gene region and quantitative trait loci mapping for agronomic traits in Brachiaria hybrids. *Grassl. Sci.* 62, 69–80. doi: 10.1111/grs.12115

Thakral, V., Yadav, H., Padalkar, G., Kumawat, S., Raturi, G., Kumar, V., et al. (2022). Recent advances and applicability of GBS, GWAS, and GS in polyploid crops. *Genotyping by Sequencing Crop Improvement*, 328–354. doi: 10.1002/9781119745686.ch15

Tong, H., and Nikoloski, Z. (2021). Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *J. Plant Physiol.* 257, 153354. doi: 10.1016/j.jplph.2020.153354

Tuteja, N. (2003). Plant DNA helicases: the long unwinding road. *J. Exp. Bot.* 54 (391), 2201–2214. doi: 10.1093/jxb/erg246

Van Lijsebettens, M., Vanderhaeghen, R., De Block, M., Bauw, G., Villarroel, R., and Van Montagu, M. (1994). An S18 ribosomal protein gene copy at the Arabidopsis PFL locus affects plant development by its specific expression in meristems. *EMBO J.* 13 (14), 3378–3388. doi: 10.1002/j.1460-2075.1994.tb06640.x

Varshney, R. K. (2021). Plant Genome special issue: Advances in genomic selection and application of machine learning in genomic prediction for crop improvement. *Plant Genome* 14 (3). doi: 10.1002/tpg2.20178

Verdoni, N., Mench, M., Cassagne, C., and Bessoule, J.-J. (2001). Fatty acid composition of tomato leaves as biomarkers of metal-contaminated soils. *Environ. Toxicol. Chem.* 20 (2), 382–388. doi: 10.1002/etc.5620200220

Vigna, B. B. Z., de Oliveira, F. A., de Toledo-Silva, G., da Silva, C. C., do Valle, C. B., and de Souza, A. P. (2016a). Leaf transcriptome of two highly divergent genotypes of Urochloa humidicola (Poaceae), a tropical polyploid forage grass adapted to acidic soils and temporary flooding areas. *BMC Genomics* 17, 910. doi: 10.1186/s12864-016-3270-5

Vigna, B. B. Z., Santos, J. C. S., Jungmann, L., do Valle, C. B., Mollinari, M., Pastina, M. M., et al. (2016b). Evidence of allopolyploidy in Urochloa humidicola based on cytological analysis and genetic linkage mapping. *PloS One* 11, e0153764. doi: 10.1371/journal.pone.0153764

Voorrips, R. E. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. *J. Hered.* 93 (1), 77–78. doi: 10.1093/jhered/93.1.77

Waldmann, P., Pfeiffer, C., and Mészáros, G. (2020). Sparse convolutional neural networks for genome-wide prediction. *Front. Genet.* 11. doi: 10.3389/fgene.2020.00025

Wang, Z., Chapman, D., Morota, G., and Cheng, H. (2020). A multiple-trait bayesian variable selection regression method for integrating phenotypic causal networks in genome-wide association studies. *G3 (Bethesda Md.)* 10 (12), 4439–4448. doi: 10.1534/g3.120.401618

Wang, X., Shi, S., Wang, G., Luo, W., Wei, X., Qiu, A., et al. (2022). Using machine learning to improve the accuracy of genomic prediction of reproduction traits in pigs. *J. Anim. Sci. Biotechnol.* 13, 60. doi: 10.1186/s40104-022-00708-0

Wang, Y., Sun, G., Zeng, Q., Chen, Z., Hu, X., Li, H., et al. (2018). Predicting growth traits with genomic selection methods in Zhikong scallop (Chlamys farreri). *Mar. Biotechnol.* 20, 769–779. doi: 10.1007/s10126-018-9847-z

Wickham, H., and Chang, W. (2016). *Package 'ggplot2'* (Vienna: R Foundation for Statistical Computing). doi: 10.1007/978-3-319-24277-4

Winkel-Shirley, B. (2001). It takes a garden. How work on diverse plant species has contributed to an understanding of flavonoid metabolism. *Plant Physiol.* 127, 1399–1404. doi: 10.1104/pp.010675

Wolc, A., and Dekkers, J. (2022). Application of Bayesian genomic prediction methods to genome-wide association analyses. *Genetics selection evolution: GSE* 54 (1), 31. doi: 10.1186/s12711-022-00724-8

Wolfe, C. J., Kohane, I. S., and Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinform.* 6, 227–227. doi: 10.1186/1471-2105-6-227

Woodward, A. W., and Bartel, B. (2005). Auxin: regulation, action, and interaction. *Ann. Bot.* 95 (5), 707–735. doi: 10.1093/aob/mci083

Worthington, M., Ebina, M., Yamanaka, N., Heffelfinger, C., Quintero, C., Zapata, Y. P., et al. (2019). Translocation of a parthenogenesis gene candidate to an alternate carrier chromosome in apomictic Brachiaria humidicola. *BMC Genomics* 20 (1). doi: 10.1186/s12864-018-5392-4

Worthington, M., Heffelfinger, C., Bernal, D., Quintero, C., Zapata, Y. P., Perez, J. G., et al. (2016). A parthenogenesis gene candidate and evidence for segmental allopolyploidy in apomictic Brachiaria decumbens. *Genetics* 203 (3), 1117–1132. doi: 10.1534/genetics.116.190314

Worthington, M., Perez, J. G., Mussurova, S., Silva-Cordoba, A., Castiblanco, V., Jones, C., et al. (2021). A new genome allows the identification of genes associated with natural variation in aluminium tolerance in *Brachiaria* grasses. *J. Exp. Bot.* 72, 302–319. doi: 10.1093/jxb/eraa469

Xu, B., Escamilla-Treviño, L. L., Sathitsuksanoh, N., Shen, Z., Shen, H., Percival Zhang, Y.-H., et al. (2011). Silencing of 4-coumarate:coenzyme A ligase in switchgrass leads to reduced lignin content and improved fermentable sugar yields for biofuel production. *New Phytol.* 192 (3), 611–625. doi: 10.1111/j.1469-8137.2011.03830.x

Xu, Y., Wang, X., Ding, X., Zheng, X., Yang, Z., Xu, C., et al. (2018). Genomic selection of agronomic traits in hybrid rice using an NCII population. *Rice* 11, 32. doi: 10.1186/s12284-018-0223-4

Xue, S., and Barna, M. (2012). Specialized ribosomes: A new frontier in gene regulation and organismal biology. Nature Reviews. *Mol. Cell Biol.* 13 (6), 355–369. doi: 10.1038/nrm3359

Yan, Z., Huang, H., Freebern, E., Santos, D. J., Dai, D., Si, J., et al. (2020). Integrating RNA-Seq with GWAS reveals novel insights into the molecular mechanism underpinning ketosis in cattle. *BMC Genomics* 21, 489. doi: 10.1186/s12864-020-06909-z

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., and Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* 5, 3231. doi: 10.1038/ncomms4231

Yoon, J., Choi, H., and An, G. (2015). Roles of lignin biosynthesis and regulatory genes in plant development. *J. Integr. Plant Biol.* 57 (11), 902–912. doi: 10.1111/jipb.12422

Yu, H., Zhang, F., Wang, G., Liu, Y., and Liu, D. (2013). Partial deficiency of isoleucine impairs root development and alters transcript levels of the genes involved in branched-chain amino acid and glucosinolate metabolism in Arabidopsis. *J. Exp. Bot.* 64 (2), 599–612. doi: 10.1093/jxb/ers352

Zhang, M., Barg, R., Yin, M., Gueta-Dahan, Y., Leikin-Frenkel, A., Salts, Y., et al. (2005). Modulated fatty acid desaturation via overexpression of two distinct ω-3 desaturases differentially alters tolerance to various abiotic stresses in transgenic tobacco cells and plants. *Plant J.* 44 (3), 361–371. doi: 10.1111/j.1365-313x.2005.02536.x

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., et al. (2014). Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PloS One* 9 (3), e93017. doi: 10.1371/journal.pone.0093017

Zhou, W., Bellis, E. S., Stubblefield, J., Causey, J., Qualls, J., Walker, K., et al. (2019). Minor QTLs mining through the combination of GWAS and machine learning feature selection. *BioRxiv*, 712190. doi: 10.1101/712190

Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., et al. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00025

Zou, C., Wang, P., and Xu, Y. (2016). Bulked sample analysis in genetics, genomics and crop improvement. *Plant Biotechnol. J.* 14 (10), 1941–1955. doi: 10.1111/pbi.12559