frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 26 March 2024
DOI 10.3389/fgene.2024.1361418

Check for updates

OPEN ACCESS

EDITED BY
Leonardo Alfredo Ornella,
Cubiqfoods SL, Spain

REVIEWED BY
Harry Lamb,
University of Queensland, Australia
Zenaida V. Magbanua,
Mississippi State University, United States

*CORRESPONDENCE
Natalia Cristina Aguirre,
✉ aguirre.natalia@inta.gob.ar

†These authors have contributed equally to this work

RECEIVED 26 December 2023
ACCEPTED 19 February 2024
PUBLISHED 26 March 2024

CITATION
Aguirre NC, Villalba PV, García MN, Filippi CV, Rivas JG, Martínez MC, Acuña CV, López AJ, López JA, Pathauer P, Palazzini D, Harrand L, Oberschelp J, Marcó MA, Cisneros EF, Carreras R, Martins Alves AM, Rodrigues JC, Hopp HE, Grattapaglia D, Cappa EP, Paniego NB and Marcucci Poltri SN (2024), Comparison of ddRADseq and EUChip60K SNP genotyping systems for population genetics and genomic selection in *Eucalyptus dunnii* (Maiden). *Front. Genet.* 15:1361418. doi: 10.3389/fgene.2024.1361418

Comparison of ddRADseq and EUChip60K SNP genotyping systems for population genetics and genomic selection in *Eucalyptus dunnii* (Maiden)

Natalia Cristina Aguirre[1]*, Pamela Victoria Villalba[1†], Martín Nahuel García[1†], Carla Valeria Filippi[1,2], Juan Gabriel Rivas[1], María Carolina Martínez[1], Cintia Vanesa Acuña[1], Augusto J. López[3], Juan Adolfo López[3], Pablo Pathauer[4], Dino Palazzini[4], Leonel Harrand[5], Javier Oberschelp[5], Martín Alberto Marcó[5], Esteban Felipe Cisneros[6], Rocío Carreras[6], Ana Maria Martins Alves[7], José Carlos Rodrigues[7], H. Esteban Hopp[1], Dario Grattapaglia[8], Eduardo Pablo Cappa[4,9], Norma Beatriz Paniego[1] and Susana Noemí Marcucci Poltri[1]

[1]Instituto de Agrobiotecnología y Biología Molecular, UEDD INTA-CONICET, Hurlingham, Argentina, [2]Laboratorio de Bioquímica, Departamento de Biología Vegetal, Facultad de Agronomía, Universidad de la República, Montevideo, Uruguay, [3]Estación Experimental Agropecuaria de Bella Vista, Instituto Nacional de Tecnología Agropecuaria, Bella Vista, Argentina, [4]Instituto de Recursos Biológicos, Instituto Nacional de Tecnología Agropecuaria, Hurlingham, Argentina, [5]Estación Experimental Agropecuaria de Concordia, Instituto Nacional de Tecnología Agropecuaria, Concordia, Argentina, [6]Facultad de Ciencias Forestales, Universidad Nacional de Santiago del Estero (UNSE), Santiago del Estero, Argentina, [7]Centro de Estudos Florestais e Laboratório Associado TERRA, Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda, Lisboa, Portugal, [8]Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), Recursos Genéticos e Biotecnologia, Brasilia, Brazil, [9]Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, Argentina

Eucalyptus dunnii is one of the most important *Eucalyptus* species for short-fiber pulp production in regions where other species of the genus are affected by poor soil and climatic conditions. In this context, *E. dunnii* holds promise as a resource to address and adapt to the challenges of climate change. Despite its rapid growth and favorable wood properties for solid wood products, the advancement of its improvement remains in its early stages. In this work, we evaluated the performance of two single nucleotide polymorphism, (SNP), genotyping methods for population genetics analysis and Genomic Selection in *E. dunnii*. Double digest restriction-site associated DNA sequencing (ddRADseq) was compared with the EUChip60K array in 308 individuals from a provenance-progeny trial. The compared SNP set included 8,011 and 19,008 informative SNPs distributed along the 11 chromosomes, respectively. Although the two datasets differed in the percentage of missing data, genome coverage, minor allele frequency and estimated genetic diversity parameters, they revealed a similar genetic structure, showing two subpopulations with little differentiation between them, and low linkage disequilibrium. GS analyses were performed for eleven traits using Genomic Best Linear Unbiased Prediction (GBLUP) and a conventional pedigree-based model (ABLUP). Regardless of the SNP dataset, the predictive ability (PA) of GBLUP was better than that of ABLUP for six traits (Cellulose content, Total and Ethanolic extractives, Total and Klason lignin content and Syringyl and Guaiacyl lignin monomer ratio). When contrasting the SNP datasets used to estimate PAs, the GBLUP-EUChip60K model gave higher and significant PA values for six traits, meanwhile, the values estimated using ddRADseq gave higher values for three other traits. The PAs correlated positively with narrow sense heritabilities, with the highest correlations shown by the ABLUP and GBLUP-EUChip60K. The two genotyping methods, ddRADseq and EUChip60K, are generally comparable for population genetics and genomic prediction, demonstrating the utility of the former when subjected to rigorous SNP filtering. The results of this study provide a basis for future whole-genome studies using ddRADseq in non-model forest species for which SNP arrays have not yet been developed.

KEYWORDS

double digest restriction-site associated DNA sequencing, genotyping by sequencing, SNP array, genomic prediction, ABLUP, GBLUP

1 Introduction

The genus *Eucalyptus* comprises more than 700 species of native trees from Australia, New Guinea, Timor, Indonesia and the Philippines (Ladiges et al., 2003; Thornhill et al., 2019), with some of them exhibiting excellent growth and adaptability in different environments. The most widely planted species include *E. grandis*, *E. saligna*, *E. pellita*, *E. urophylla*, *E. globulus*, *E. dunnii*, *E. nitens*, *E. tereticornis* and *E. camaldulensis* (Grattapaglia and Kirst, 2008). Despite its limited natural distribution in Australia, *E. dunnii* (Maiden, 1905) has shown relevant genetic variability for growth, adaptation to different soils and tolerance to abiotic stresses such as frost, drought and high summer humidity, when planted as an exotic (Darrow, 1994; Jovanovic et al., 2000; Clarke, 2009; Thomas et al., 2009; Shi et al., 2016). As a result of these advantages, *E. dunnii* plantations have been established in subtropical areas such as southern China, South Africa, South America and Australia, where it is considered to be an alternative to other species, particularly in the context of climate change (Marcó and White, 2002; Marcó, 2005; Maseko et al., 2007; Hardner et al., 2016; López et al., 2016; Shi et al., 2016; Gallo et al., 2018; Resquin et al., 2019). However, genetic improvement of *E. dunnii* is still at an early stage.

Molecular breeding programs facilitate the selection of the best individuals in the early stages of development, before the phenotypic traits of interest have been expressed. It therefore has significant potential to accelerate the rate of genetic gain in a shorter time, which is extremely important in forest species with long generation times (Neale and Kremer, 2011; Varshney et al., 2017). Numerous studies have been conducted in forest trees demonstrating that Genomic Selection (GS) equals or outperforms phenotypic selection for traits related to growth and wood properties. This enhances the rate of genetic improvement over time by intensifying selection, significantly reducing generation intervals, and improving the precision of breeding values (Grattapaglia, 2022).

In GS, the underlying assumption is that all markers might be tied to genomic loci influencing the trait under study and therefore can be used to select the best individuals for breeding (Meuwissen et al., 2001) through the application of novel statistical methods based on whole genome regression (Cappa et al., 2019). Unlike marker-assisted selection (MAS), which applies stringent hypothesis testing to declare the association of markers with variation in the target trait, GS relies on capturing all loci that cause phenotypic variation among individuals with dense marker coverage (Hayes and Goddard, 2010; Isik, 2014).

By building prediction models based on the analysis of a population with both phenotypes and genotypes data (Training population, TP), the sum of marker effects can be used to predict the genomic estimated breeding values (GEBVs) of individuals that were only genotyped (Selection candidates, SC). It is a consensus now that GS depends largely on the existence of genetic relatedness between the TP and SC and some degree of Linkage Disequilibrium (LD) between markers and causal loci (Desta and Ortiz, 2014; Tan et al., 2017).

A commonly used GS approach is the or Genomic Best Linear Unbiased Prediction (GBLUP, Strandén and Garrick, 2009; VanRaden, 2008). This method predicts breeding values using a genomic-based relationship matrix between individuals (***G***-matrix). By using the ***G***-matrix instead of the conventional matrix of expected pedigree relationships (***A***-matrix), GBLUP predicts the GEBVs more accurately, thus increasing the genetic gain and selection accuracy for the next-generation of the breeding cycle (Nejati-Javaremi et al., 1997; Villanueva et al., 2005).

In recent decades, different types of molecular markers have been developed for genetic analysis, such as Single Nucleotide Polymorphism (SNP), Simple Sequence Repeats (SSR) or Microsatellites, Insertions/Deletions (InDels), Structural Variants (SV), etc (Fuentes-Pardo and Ruzzante, 2017). SNPs have become the markers of choice due to their abundance, stability, codominance, low cost per datapoint and ease in assay design, automation and data interpretation (Bajgain et al., 2016; Torkamaneh et al., 2018).

The ideal genotyping technique for GS would be Whole Genome Sequencing (WGS) providing the full sampling of genetic variants among individuals (Fuentes-Pardo and Ruzzante, 2017). Nevertheless, this is still too expensive to accommodate experiments with large sample sizes, especially for forest tree breeding programs that operate on low budgets. SNP arrays and reduced representation sequencing (RRS) have been used in practice to obtain genome-wide SNP data in a cost-effective way for molecular breeding studies. Both methodologies have their advantages and disadvantages (Pavan et al., 2020).

For commercially relevant species of *Eucalyptus*, multispecies SNP genotyping platforms were developed including the Illumina Infinium based EUChip60K (Silva-Junior et al., 2015) and the second generation Euc72K array based on the Axiom technology[1]. These two systems allow genotyping over 60,000 genome-wide SNPs, have approximately 28,000 SNPs in common and satisfy essential requirements of high precision, throughput, data reproducibility and low genotyping cost (Silva-Junior et al., 2015). The SNPs included in these arrays were discovered from low-depth WGS data of 240 individuals of 12 different species and detected using a reference genome of *E. grandis* v1.0 (Myburg et al., 2014). *E. dunnii* was represented by only 12 unrelated individuals in which 17,014 SNPs were converted with Minor Allele Frequency (MAF) > 0.01 (Silva-Junior et al., 2015). Although these arrays have been very successful worldwide for they accommodate several different eucalypt species, SNP discovery from a small number of individuals is subject to ascertainment bias (Albrechtsen et al., 2010; Li and Kimmel, 2013).

Medium-density SNP arrays with several tens of thousands of SNP, have been developed for all mainstream forest tree species, such as for *Populus trichocarpa* (Geraldes et al., 2013), *Picea* ssp. (Pavy et al., 2013; Bernhardsson et al., 2021), *Pinus* ssp. (Plomion et al., 2015; Perry et al., 2020; Caballero et al., 2021; Graham et al., 2022; Jackson et al., 2022; Kastally et al., 2022), *Pseudotsuga menziesii* (Howe et al., 2020), *Araucaria angustifolia* (Silva et al., 2020), *Eucalyptus* ssp. (EUChip60K and Euc72K). In the case of *Eucalyptus*, the two SNP arrays have been used in a large number of studies (e.g., Telfer et al., 2015; Müller et al., 2019; Marco de Lima et al., 2019; Mhoswa et al., 2020; Ballesta et al., 2018; Mostert-O'Neill et al., 2020; Jurcic et al., 2021; Paludeto etal., 2021; Valenzuela et al., 2021; Tan and Ingvarsson, 2022; Duarte et al., 2023). Given its multispecies nature, this platform generally provides between 10,000 and 30,000 informative SNPs in all planted species. Due to its fixed content it is unclear whether population genomics analyses ultimately exclude relevant variants in unsampled genomic regions.

As a practical alternative to SNP array development, especially for orphan species, SNP genotyping based on RRS strategies have been used. SNP genotyping following Restriction Enzyme based RRS (REbRRS) techniques are approaches that combine genome reduction and sampling of both coding and non-coding regions without the need for prior genomic information (Davey et al., 2011; Andrews et al., 2016). These techniques rely on Next-Generation Sequences (NGS) of a reduced genome portion of several individuals analysed simultaneously (multiplexed), do not require a reference genome or prior knowledge of polymorphisms, and combine marker discovery and genotyping in a single protocol. Therefore, they provide a rapid, high-throughput, and cost-effective strategy for performing genome-wide analyses. Furthermore, they can be applied to non-model species and unique germplasm sets to obtain exclusive polymorphism information (Davey et al., 2011; Andrews et al., 2016), to sample alternative genomic regions providing complementary data to SNP array data, especially in plants (Deschamps et al., 2012).

REbRRS embraces a group of similar protocols that include Genotyping by Sequencing (GBS, Elshire et al., 2011), Restriction site Associated DNA sequencing (RADseq, Baird et al., 2008), and double digest RADseq (ddRADseq, Peterson et al., 2012), being widely adopted in the conservation and breeding area (Andrews et al., 2016; Fuentes-Pardo and Ruzzante, 2017; Campbell and Dupuis, 2018; Wright et al., 2019). In comparison to fixed content SNP arrays, RRS techniques require bioinformatics analysis of sequence data to detect variants and declare genotypes (Nielsen et al., 2011). Missing data, loci sampling and genotype reproducibility issues across experiments are common features of these methods due to polymorphisms in enzyme cleavage sites and variation in the sequence coverage across individuals and loci (Andrews et al., 2016), potentially causing bias in population genetic statistics. To mitigate these limitations, stringent filtering for high call rates and imputation are highly recommended (Money et al., 2015; Andrews et al., 2016; Bajgain et al., 2016; Torkamaneh et al., 2018). Both methods are powerful means to study the genome and provide sufficient resolution to perform different molecular genetic approaches, despite the fact that chip-based SNP genotyping requires less computational knowledge and data processing resources than the REbRRS method, resulting in less missing data and higher reproducibility (Bajgain et al., 2016).

REbRRS methods have been applied to forest species mainly for linkage mapping, QTL detection, marker development, phylogenetics, phylogeography, parentage analysis, association mapping, genomic selection, population genetics, genome scanning, among others (Parchman et al., 2018). In particular, ddRADseq combined with reference genomic SNP calling yielded a higher number of reliable markers compared to other REbRRS methods such as GBS and RADseq in beech and oak Ulaszewski et al. (2021). In *Eucalyptus*, a few studies have used REbRRS methods, including GBS (Grattapaglia et al., 2011; Durán et al., 2018; Klápště et al., 2021), DArT-seq (Sansaloni et al., 2011; Marco de Lima et al., 2019), and ddRADseq (Aguirre et al., 2019). Specifically, for *E. dunnii* ddRADseq has been optimized (Aguirre et al., 2019), and there is a scale-up protocol (Aguirre et al., 2023).

In this work, we were interested in evaluating the comparative performance of two high-throughput genotyping systems, ddRADseq and the SNP platform EUChip60K, for population genetics analyses and GS. To our knowledge, this is the first report where a restriction enzyme-based RRS method, ddRADseq, is compared to SNP array data for population genetic parameters and GS analyses in a tree species, and specifically in *Eucalyptus*.

2 Materials and methods

2.1 *Eucalyptus dunnii* breeding population

The *E. dunnii* breeding population (1,520 trees) under study was established in 1991 (31° 45′S, 58° 15′W, 40 m. a.s.l., Entre Ríos province, Argentina) with a complete block design (Marcó and White, 2002). This population was composed of 72 open-pollinated (OP) families, of which 60 were from four native origins in New

[1] https://www.thermofisher.com/order/catalog/product/551134; accessed on 13 December 2023

South Wales (NSW) state in Australia. The remaining 12 families came from a local provenance or seed sources of known Australian origin (Moleton, NSW), where they were selected by their superiority in stem straightness and volume (Supplementary Table S1).

2.2 Phenotypic characterization of *E. dunnii* breeding population

E. dunnii trees were measured for growth traits by Diameter at Breast Height at six and 20 years old (DBH6, and DBH20) and Stem Straightness at 6 years old (SS6; López et al., 2016; Marcó and White, 2002). At 20 years old, the intensity of growth stresses was evaluated by measuring the Log End Split Index (LESI20, López et al., 2016), and Wood basic Density (WD20, Kg/m^3) was estimated by water immersion. Estimates of six wood chemical properties at 20 years were obtained using Near Infrared spectroscopy (NIR) at the Instituto Superior de Agronomia (ISA, Portugal). These included: Cellulose content (CEL20), Total and Ethanolic extractives (TE20 and EE20), Total and Klason Lignin content (TL20 and KL20) and Syringyl and Guaiacyl lignin monomer ratio (S/G20), as described by Rodrigues et al. (1998). Details of the traits measured are summarized in Supplementary Table S2. Phenotypic traits data were adjusted to normal distributions and standardized if required, except for the SS6 categorical variable, which was transformed to a continuous variable using a Normal Score (*stats* R package, R Core Team, 2023). The experimental design effect was removed using breedR (Muñoz and Sánchez, 2014), with an individual tree mixed linear model using restricted maximum likelihood inference (REML, Patterson and Thompson, 1971).

2.3 Genotypic data of *E. dunnii* breeding population

DNA was extracted from lyophilized leaves of 308 *E. dunnii* individuals using a CTAB method (Hoisington et al., 1994) with modifications for the species (Marcucci Poltri et al., 2003), quantified with Qubit 2.0 fluorometer (Thermo Fisher Scientific), and quality verified by both Nanodrop (Thermo Fisher Scientific) and 1% agarose gel electrophoresis (as described in Aguirre et al., 2019; 2023).

2.3.1 SNP ddRADseq dataset

A ddRADseq genotyping protocol optimized for *E. dunnii* (Protocol 2 from Aguirre et al., 2019; Aguirre et al., 2023) was applied to the breeding population, by constructing 13 libraries/pool of 24 samples each (including four extra samples, as required by the protocol), at the Unidad de Genómica, IABiMo-INTA, Argentina. The libraries were sequenced using a NextSeq 500 instrument (Illumina, Inc., San Diego, CA, USA). Sequencing was carried out with a 150-cycle high-output kit NextSeq and set up for 75 bp paired-end (PE) reads (Illumina Inc.).

To search for loci and SNP markers in the ddRADseq data, Stacks v1.48 software (Catchen et al., 2013) was used, as described by Aguirre et al. (2019) for "with reference analysis". In summary, sequences were filtered by quality with *process_radtags* (removing barcodes, adapters, reads without enzyme cutting site, and with Phred quality value mean below 10, also truncating them to 66 bp). The loci and SNPs were identified using the *ref_map.pl* pipeline, where reads were previously mapped against the *E. grandis* reference genome v2.0 (Myburg et al., 2014) using Bowtie2 (Langmead and Salzberg, 2012) with default parameters. In detail, a minimum of three reads was used to define an allele (tag or stack) within an individual (-m 3), two bases of difference between alleles were allowed to build a locus within an individual (-M 2), and two different bases between loci were allowed to build the loci catalogue (-n 2) between individuals. Stacks or alleles with great depth of sequences were removed since it is very likely that they came from repetitive regions of the genome (-t). As a diploid species, only loci made up of two stacks (-X "ustacks: -max_locus_stacks 2″) were considered. Additionally, the *rxstacks* program was applied as described by Aguirre et al. (2019); however, in this case, the loci logarithm of the likelihoods was filtered up to −20 (minus 20). Finally, the *populations* component was executed using a filter of defined loci with a minimum of six reads (-m 6), as the call of heterozygous loci is more robust as the ddRADseq read depth increases (from 3 to 6; Rochette and Catchen, 2017).

2.3.2 SNP EUChip60K dataset

DNA samples from 308 individuals were lyophilized in 96-well plates and sent to the NEOGEN (USA, © Neogen Corporation) for genotyping with the EUChip60K (Silva-Junior et al., 2015). For allelic designation, the genotyping module of GenomeStudio 2.0 program was used (Illumina, San Diego, CA, USA). A cluster file optimized for the Maidenaria section of subgenus *Symphyomyrtus* (i.e.,: *E. globulus*, *E. nitens* and *E. dunnii*) and a technical filter for quality parameters were used as suggested by Silva-Junior et al. (2015).

Because the SNP coordinates were provided based on the *E. grandis* reference genome v1, oligonucleotide sequences of chip probes were mapped against the *E. grandis* v2.0 reference genome using Bowtie2 (Langmead and Salzberg, 2012; with default parameters). Next, to compare with the ddRADseq dataset, the SNP coordinates of the EUChip60K dataset were converted to the *E. grandis* v2.0 genome using our own script in bash/R language.

2.4 Genomic datasets quality filter and imputation

The proportion of total and per genotype missing data, observed heterozygosity per individual, and genetic relationships between them were calculated using PLINK v1.9 (Chang et al., 2015). An individual was eliminated from both datasets using VCFtools software (Danecek et al., 2011) if it showed at least one of the following occurrences in at least one SNP dataset: a high proportion of missing data (more than 60%), high heterozygosity values (outside the range of population distribution, higher than three times the standard deviation), or unexpected genetic relationships with other individuals (greater than expected for an OP population, such as father or mother/child relationship and/or very close to clones, --king-cutoff 0.354). SNPs were then filtered out using a

MAF of 0.01 by PLINK v1.9. As a final quality control of the filtered data, pairwise genetic distances were calculated with each genotyping dataset using the snpgdsIBS option of the SNPRelate R package (Zheng et al., 2012). Both datasets were correlated with a Mantel test (Mantel, 1967; vegan R package; Dixon, 2003).

Imputation of missing data, in both genotypic datasets, was performed using the LinkImpute program (Money et al., 2015). After imputation, ddRADseq and EUChip60K datasets were merged with the BCFtools tool (Li et al., 2009) to generate a third joint SNP dataset (hereinafter called ddRADseq + EUChip60K or combined data).

2.5 Population genetics analyses

2.5.1 SNP distribution and MAF in the three datasets

To compare the performance of the applied genotyping methodologies and the combined data, SNP distributions and their allele frequencies along the *E. grandis* genome were evaluated by CMplot[2] and synbreed R packages (Wimmer et al., 2012).

2.5.2 Linkage disequilibrium estimation

The LD between each SNP pair was estimated using TASSEL software (Bradbury et al., 2007) considering SNPs with MAF ≥0.01 and no correction for population structure or relatedness as Estopa et al. (2023). Patterns of LD decay for each dataset were plotted in a 10 Kbp window using R software according to the method of Hill and Weir (1988) and based on the physical distance of the *E. grandis* v2 genome (Myburg et al., 2014).

2.5.3 Population genetic structure and diversity parameter estimation

To estimate genetic structure and genetic diversity parameters, each dataset was LD pruned (r^2 greater than 0.2) by implementing the --indep-pairwise function of the PLINK v1.9 program, using 2 Mb windows with an overlap between them of 200 kb. This filter was used to eliminate redundant information, obtain a more accurate estimation, and reduce the computational demand of statistical analyses.

Population genetic structure was estimated by Discriminant Analysis of Principal Components (DAPC; Jombart et al., 2010) using adegenet 2.0.0 R package (Jombart, 2008). Because DAPC requires defining the number of groups in advance, SNP data was transformed using PCA and a k-means clustering algorithm. Successive k-means were run using find. clusters function of adegenet, and optimal grouping was chosen through the lowest Bayesian Information Criterion (BIC; Schwarz, 1978) value. For these population structure analyses, random sub-sampling of 800 SNPs was applied to each of the three genomic datasets filtered by LD. Subsequently, F_{ST} was calculated between the genetic groups defined by DAPC for each of the three datasets using the populations module (option: -fstats) of the Stacks program (Catchen et al., 2013). The significance of each F_{ST} value was calculated through bootstrap resampling implemented in the said population module (--fst_correction *p*_value -k --bootstrap_fst). This calculation was used with the default parameters, which were a resampling number of 10,000 times and a *p*-value less than 0.05, to report the F_{ST} value.

The following population genetic diversity statistics were calculated for each dataset and genetic structure group: allele frequencies p and q, expected (He) and observed (Ho) Heterozygosity and Polymorphic Information Content (PIC) with the popgen function in the snpReady R package (Granato et al., 2018).

2.6 Genomic selection

2.6.1 Genomic selection models

For the GS proof-of-concept, a single-trait model was used with the corresponding ***A***- or ***G***-matrix for the conventional ABLUP and GBLUP models and 11 phenotypic traits evaluated. For the ABLUP model, the additive relationship matrix ***(A)*** was calculated using the getA function in pedigreemm R software (Vazquez et al., 2010). For the GBLUP model the genetic relationship matrix ***(G)*** was calculated using the function A. mat in the rrBLUP program (Endelman, 2011). ABLUP and GBLUP models were applied by kin.blup function (rrBLUP R package), thus solves mixed models of the form:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zg} + \boldsymbol{\varepsilon}$$

Where $\boldsymbol{\beta}$ is a vector of fixed effects, $\boldsymbol{g}$ is a vector of random genotypic values with covariance $\boldsymbol{G} = \boldsymbol{Var}\ (\boldsymbol{g})$, and the residuals follow $\boldsymbol{Var}\ (\boldsymbol{\varepsilon}) = \boldsymbol{R_i}\sigma^2_e$, with $\boldsymbol{R_i} = \boldsymbol{1}$ by default. For all models and phenotypic traits, the number of trees with genotypic data was the same (280 trees). However, the number of individuals with both phenotypic and genotypic data varied by trait (Supplementary Table S2). For ABLUP only the genotyped individuals in the trial were predicted to make the results comparable to those obtained with GBLUP.

2.6.2 Validation of the model

A leave-one-out (LOO) cross-validation strategy was performed for all traits, where in each case the entire population except one individual was used as the TP and the phenotype of the excluded individual was predicted. Pearson's correlations (stats R package, cor function, R Core Team, 2023) between the phenotypic records corrected for environmental effects and the predicted values were used to obtain the predictive ability (PA) of each model. The significance of Pearson's correlation was determined using a two-tailed *t*-test with an alpha level of 0.05.

2.6.3 Heritability

For the estimation of variance components and heritability, ABLUP results from the kin.blup function of the rrBLUP package were used (Endelman, 2011).

3 Results

3.1 Genotyping data of *E. dunnii* breeding population

3.1.1 SNP ddRADseq dataset

Sequencing of all 13 library pools on the NextSeq instrument yielded 383.5 million PE (57.6 Gb) passing filter reads (mean quality

[2] https://github.com/YinLiLin/R-CMplot

greater than 30 Phred index). The average number of PE reads per pool of 24 samples was 24,224,266.5.

After all sequence quality filters, an average of 1,009,344.5 PE reads per sample was finally obtained. However, there was a large variation in the number of PE reads between samples (112,342 to 3,116,508). This variation is expected and mainly due to the variation in the number of reads within sample pools (Aguirre et al., 2019).

A mean of ~80% of reads per individual mapped to the *E. grandis* v2.0 reference genome and 530,885 SNPs at 195,010 loci (75bp each locus) were found in the SNP calling analysis. After applying the first loci and SNPs quality filters with Stacks software, a raw ddRADseq dataset was obtained, with 42,058 SNPs in 16,123 polymorphic loci. Each locus was defined by a depth of at least 6 reads, a likelihood greater than −20, a MAF greater than 0.01 and the presences in at least 50% of the 308 individuals.

3.1.2 SNP EUChip60K dataset

Following the standard genotyping and quality control procedures (Illumina, 2010), all 308 individuals were successfully genotyped.

3.2 Genomic datasets quality filter and imputation

The overall proportions of missing data for the EUChip60K (64,639 SNPs) and ddRADseq (42,058 SNPs) datasets were 0.11 and 0.34, respectively. The proportion per individual ranged from 0.13 to 0.87 with a mean of 0.34 ± 0.15 for ddRADseq and between 0.10 and 0.14 with a mean of 0.11 ± 0.006 for the EUChip60K data. Twenty-eight individuals were eliminated from both datasets: 18 of them showed a high proportion of missing data (60%) in ddRADseq dataset; 10 of them presented heterozygosity values greater than 0.4 (mean of 308 individuals: 0.29 ± 0.04) and a closer than expected genetic relationship with another individual (two of them in both datasets and eight only in the EUChip60K dataset).

SNPs with more than 20% of missing data and MAF below 0.01 were filtered out. The final datasets were composed of 280 individuals with 8,170 ddRADseq SNPs and a total of 13% missing data and the same 280 individuals 19,045 EUChip60K with SNPs and a total of 3% missing data. The Mantel test correlation between the genetic distance matrices obtained with the two datasets was found to be r = 0.69 (Significance 0.001). LinkImpute, the software/algorithm used to impute, estimates accuracy before processing by sub-sampling the existing data, removing and imputing (Money et al., 2015). Imputation accuracies of 0.89 and 0.84 were achieved for the ddRADseq (8,170 SNPs) and EUChip60K (19,045 SNPs) dataset.

Finally, both datasets were joined, creating the combined dataset. Since the imputation could modify SNPs allele frequencies, the three datasets were again filtered by MAF, giving a total of 8,011, 19,008 and 27,019 SNPs for ddRADseq, EUChip60K, and the combined dataset, respectively.

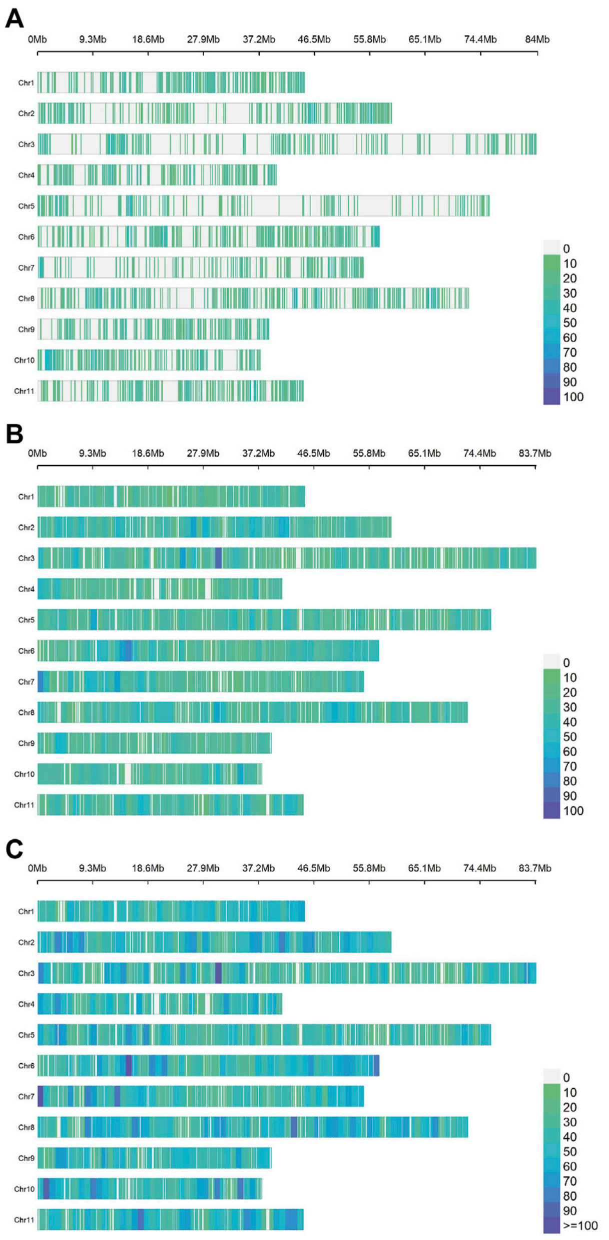

FIGURE 1
Density and distribution of the SNPs per chromosome for each dataset of the Eucalyptus dunnii population. The chromosomes are plotted from 1 (top) to 11 (bottom) on the ordinate axis and their sizes are given in Mb on the abscissa axis. The colour scale varies from zero SNPs (green) to more than 100 SNPs per 1 Mb (deep blue). From top to bottom: **(A)** ddRADseq (7,831 SNPs); **(B)**. EUChip60K (18,678 SNPs); **(C)**. ddRADseq + EUChip60K (26,509 SNPs).

3.3 Population genetics analyses

3.3.1 SNPs distribution and MAF in the three datasets

The average number of SNPs per chromosome was 712, 1,698 and 2,410 for ddRADseq, EUChip60K and the combined dataset respectively, distributed along the 11 chromosomes (Figure 1).

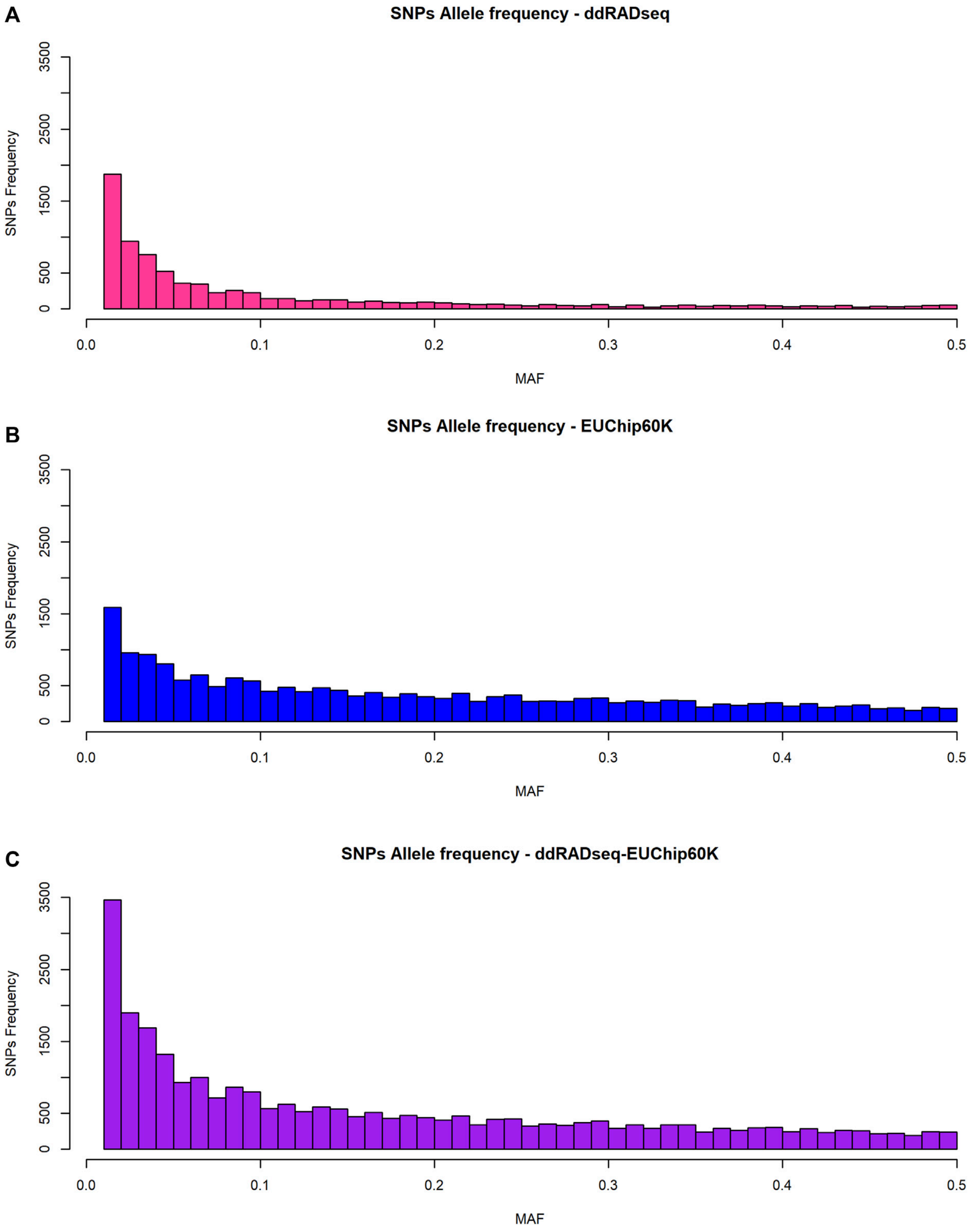

FIGURE 2
Allele frequency distribution of SNPs. **(A)** ddRADseq; **(B)** EUChip60K; **(C)** ddRADseq + EUChip60K.

Examining the mean distances between contiguous SNPs (Supplementary Table S3) and, given the lower number of markers, the ddRADseq dataset shows the largest distances (75,831 ± 20,108 bp). The EUChip60K data showed on average less than half that distance (33,198.04 ± 2,217.10 bp). The combined datasets showed the lowest inter-SNPs distance (average: 23,009.75 bp ± 2,217.10 bp). In Figure 1, the ddRADseq dataset reflected a more clustered pattern of SNP distribution than the EUChip60K dataset, again consistent with the higher mean distances between SNPs and its higher standard deviation.

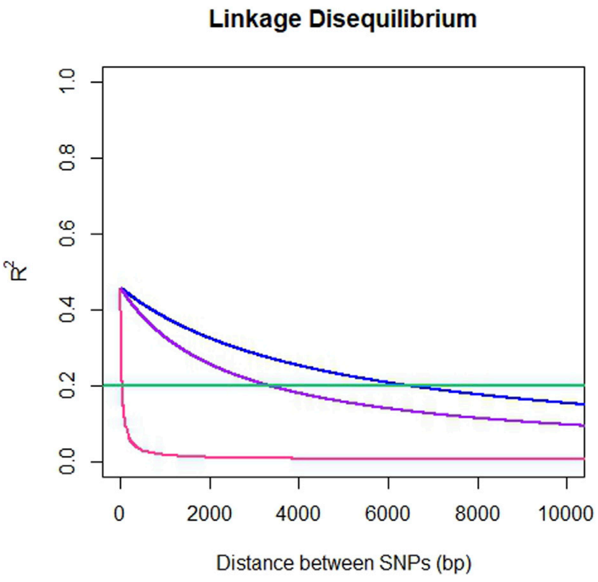

FIGURE 3
Pattern of linkage disequilibrium decay in E. dunnii. Pairwise SNP distances up to 10 Kbp *versus* r^2 for ddRADseq (pink), EUChip60K (blue) and ddRADseq + EUChip60K (purple). Abline at 0.2 r^2 in green.

In relation to the sum of the extreme SNP distance for all chromosomes (Total), it is observed that, with the combined dataset (coverage of 611.09 Mb), this value closely approximates the sizes of the *E. grandis* genome (640 Mb). It is important to note that these distances are relative, since they were estimated based on the *E. grandis* reference genome (Myburg et al., 2014) (Figure 1) which has a larger size compared to the *E. dunnii* genome (~530 Mb, Grattapaglia & Bradshaw, 1994). These results suggest that both genotyping methods cover the whole genome, as shown in the SNP density and distribution scheme per chromosome (Figure 1).

The allele frequency spectrum of SNP data generated by the two genotyping systems showed a striking difference (Figure 2). ddRADseq showed a strong bias towards low-frequency alleles, with 68% of SNPs (5,469 of 8,011 SNPs) having a MAF <0.1, consistent with the fact that the vast majority of variants will be rare. The EUChip60K data showed the expected MAF distribution based on the preselection made during chip development aimed to enrich SNPs with higher frequency, reflecting the coverage bias in fixed-content SNP chip data (37% of the SNPs with MAF <0.1, 7,070 of 19,008 SNPs). The combined datasets provided a MAF distribution that should be of interest for the GS approach (46% of the SNPs with MAF <0.1, 12539 of 27,019 SNPs).

3.3.2 Linkage disequilibrium estimation

The average genome-wide LD for SNP pairs (r^2) for ddRADseq, EUChip60K and ddRADseq + EUChip60K were 0.025, 0.032 and 0.033, respectively. LD was observed to fall below the 0.2 r^2 threshold at a distance of 37 bp for the ddRADseq dataset, 6,387 bp (6.4 Kbp) for EUChip60K and 3,298 bp (3.3 Kbp) for the combined dataset (Figure 3).

3.3.3 Population genetic structure and diversity parameters estimation

To ensure more reliable estimates of population genetic structure parameters, SNP pairs displaying an r^2 value greater than 0.2 were pruned. The LD-pruned datasets consisted of 4,848 SNPs for ddRADseq, 13,385 SNPs for EUChip60K, and 17,611 SNPs for the combined set.

Population genetic structure analysis with the DAPC method identified two genetic groups based on the lowest BIC value (Figure 4; Supplementary Figures 1A–C). Only two individuals differed in group assignment when datasets were compared. These two individuals were assigned to group one by EUChip60K and combined datasets, but belonged to group two when the ddRADseq dataset was applied. Group one was composed of 43 (ddRADseq) or 45 (EUChip60k and combined data-set) of the 52 individuals coming from local provenance seeds (Australian origin: Moleton, NSW; Supplementary Table S1). The group two consisted of the remaining 239 (ddRADseq) or 237 (EUChip60k and combined data-set) individuals, depending on the dataset considered (Supplementary Figures 1D–F). However, the F_{ST} estimates between these two genetic groups were low, irrespective of the SNP dataset used (F_{ST} = 0.0148, *p*-value <0.05 for ddRADseq; F_{ST} = 0.0155, *p*-value <0.05 for EUChip60K, and F_{ST} = 0.0148, *p*-value <0.05 for the combined dataset). Population genetic statistics were estimated with each SNP dataset (Table 1). Higher diversity was estimated with the SNP chip data, consistent with the higher MAF observed for the SNPs sampled. No significant difference was seen in diversity measures between the two groups found in the structure analysis.

TABLE 1 Population genetic diversity parameters estimated with the three genotyping datasets. Total pop.: parameters calculated for whole population; Group 1 and Group 2: parameters calculated for each genetic group defined by DAPC analysis; p: average major allele frequency; q: average minor allele frequency; He: expected heterozygosity; Ho: observed heterozygosity; PIC: polymorphic information content.

	ddRADseq			EUChip60K			ddRADseq+EUChip60K		
	Total pop.	Group 1	Group 2	Total pop.	Group 1	Group 2	Total pop.	Group 1	Group 2
p	0.89	0.89	0.89	0.80	0.80	0.80	0.82	0.82	0.82
q	0.11	0.11	0.11	0.20	0.20	0.20	0.18	0.18	0.18
He	0.17	0.14	0.17	0.28	0.25	0.28	0.25	0.22	0.25
Ho	0.15	0.13	0.15	0.29	0.27	0.29	0.26	0.24	0.26
PIC	0.14	0.12	0.14	0.23	0.20	0.23	0.21	0.18	0.21

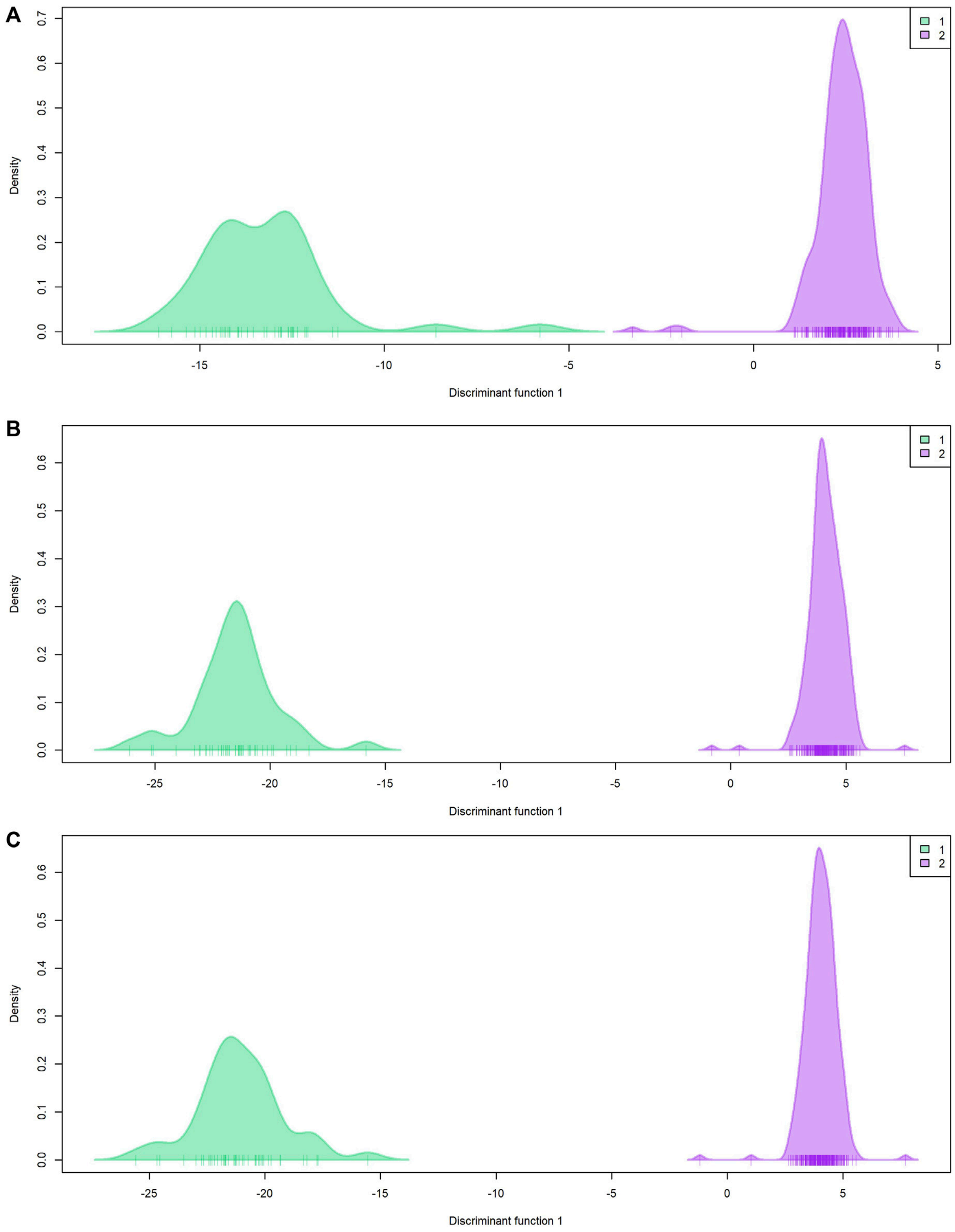

FIGURE 4
Population genetic structure by DAPC. Plot of density by discriminant function 1. **(A)** ddRADseq; **(B)** EUChip60K; **(C)** ddRADseq + EUChip60K. Group 1 (green) composed by 43 **(A)** or 45 **(B, C)** individuals; group 2 (violet) composed by 239 **(A)** or 237 **(B, C)** individuals.

3.4 Genomic selection

The predictive ability (PA) of GBLUP (all genomic datasets) was superior to ABLUP for six of the traits evaluated (Figure 5, Supplementary Table S4), specifically wood quality traits estimated by NIR (EE20, TE20, KL20, TL20, S/G20, CEL20). However, for the growth trait DBH6, despite GBLUP demonstrating superiority, the PA value was nearly zero and not

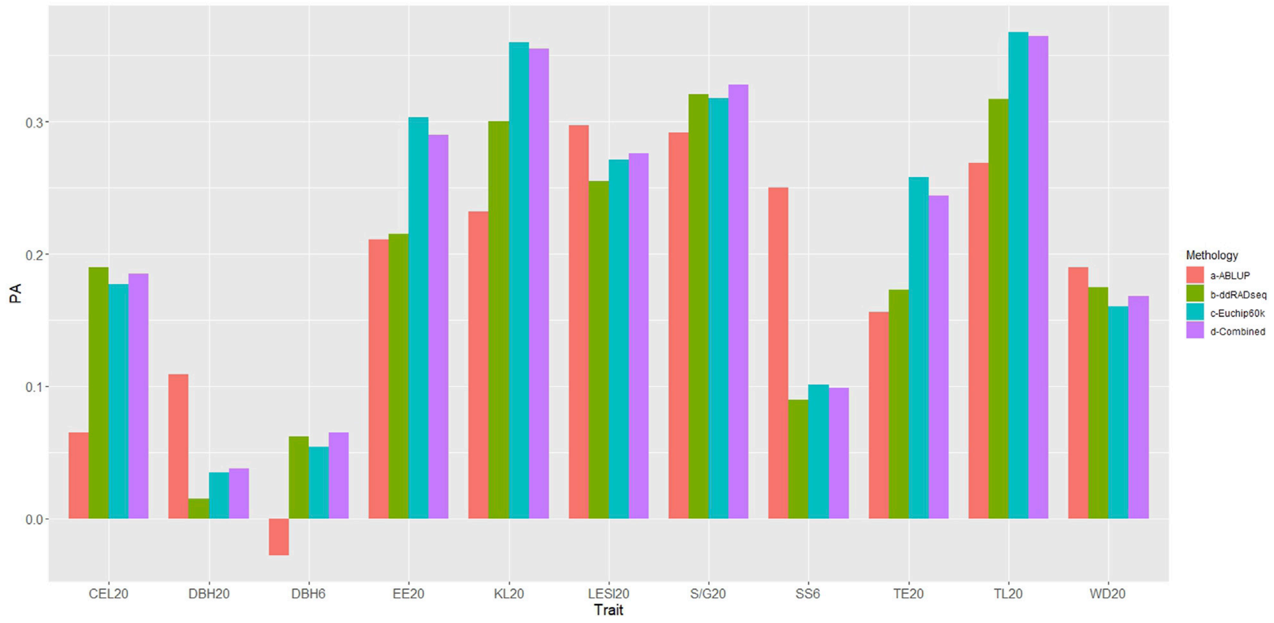

FIGURE 5
Predictive abilities obtained with ABLUP and GBLUP (for each genomic dataset). PA: Predictive ability. Traits under study: DBH: diameter at breast height, SS: stem shape, LESI: log end split index, EE: ethanolic extractives, TE: total extractives, KL: Klason lignin, TL: total lignin, SG: Syringyl/Guaiacyl, CEL: cellulose, WD: basic density. Methodology: a- ABLUP (red), b-ddRADseq (green), c- EUChip60K (blue), d-ddRADseq + EUChip60K (violet).

statistically significant (p > 0.05). In contrast, ABLUP presented higher PA values for SS6, DBH20, LESI20 and WD20.

When comparing the GBLUP PA obtained with the ddRADseq and EUChip60K SNP datasets, the latter yielded higher and significant PA values for six traits (SS6, LESI20, EE20, TE20, KL20 and TL20), while ddRADseq dataset yielded higher PA values for three traits (S/G20, CEL20 and WD20). For both DBH traits PA values were close to zero and not significant (p > 0.05) with the two datasets. For six of the 11 traits PAs obtained with the EUChip60K data were higher than those obtained with ddRADseq data and for some traits such EE20, KL20, TE20 and TL20 the differences were substantial.

Contrasting the results of the combined dataset with each independent dataset, it is observed that for three traits (DBH20, LESI20 and S/G20), the former showed slightly higher PA values, although overall the PA values obtained with the combined datasets mostly close to the highest PA value obtained with one of the independent datasets.

The mean squared errors (MSEs) were similar between ABLUP and GBLUP, with the three datasets, EUChip60K and the combined dataset showing slightly lower average MSEs (ABLUP: 0.904, ddRADseq: 0.898, EUChip60K: 0.885 and ddRADseq + EUChip60K: 0.885).

Heritabilities estimated with the ABLUP model varied between 0.251 for CEL20 and 0.834 for LESI20 (Supplementary Table S4). Four traits showed high (>0.5) h^2 values (LESI20, EE20, KL20 and TL20) while the remaining traits had moderate values ($0.15 < h^2 < 0.50$). Pearson's correlations between h^2 and PA were 0.819, 0.732, 0.805 and 0.796 for pedigree, ddRADseq, EUChip60K and ddRADseq + EUChip60K data respectively.

4 Discussion

The present study aimed to evaluate SNP data obtained with two alternative genotyping methods, ddRADseq and fixed-content SNP array, for estimating population genetic parameters and modeling GS for wood quality and growth traits in a breeding population of *E. dunnii,* a forest tree. This *Eucalyptus* species is important in the context of climate change, due to its growth advantages on some environmental conditions. Having access to a high-density, low-cost, flexible, and accurate genotyping platform is essential for the successful application of GS. The number of informative markers is expected to be directly proportional to the predictive power of a GS model, by more accurately capturing relatedness between training set and selection candidates and increasing the likelihood that loci controlling the target quantitative trait will be in LD with at least one marker (Meuwissen and Goddard, 2010).

4.1 ddRADseq application

Due to the absence of a reference genome for *E. dunnii*, we initially evaluated SNP discovery and genotyping with both a *de novo* and a reference-based analysis to compare the results (Stacks program; Catchen et al., 2013). Both analyses can be applied to identify SNPs with high accuracy after applying stringent bioinformatics and quality filters (Aguirre et al., 2019). After implementing the first quality filters 42,058 SNPs were found in 16,123 polymorphic loci with a reference-based method. In contrast to the reference-based analysis, which only considers reads that map on the *E. grandis* genome v2.0 (80% of reads), the *de novo* analysis uses all reads for marker identification. As expected, the *de novo* analysis recovered more SNPs and loci (55,338 SNPs in 22,629 polymorphic loci). These results are encouraging to explore these *E. dunnii* SNPs that could not be detected using the *E. grandis* reference. On the other hand, *de novo* discovery requires stricter criteria and parameters when defining loci, due to the higher number of false positives obtained (Rochette and Catchen, 2017). Thus, only SNPs detected by the reference-based analysis were considered for the goal of comparing ddRADseq with EUChip60K datasets in downstream applications. Additionally, based on previous work with imputation strategies for ddRADseq data (Merino, 2018), a further filter was applied to this SNPs dataset, leaving only those with call rate >80% (Aguirre et al., 2019). These *de novo* analysis results suggest that ddRADseq has the potential to be widely applicable to forest tree species that do not have a reference genome. Nevertheless, it should be pointed out that for species without a reference genome, the sequencing data should aim for longer paired-end reads than the 75 bp long ones obtained for *E. dunnii* in this study. In addition, the number of optimal reads per sample required depends on the optimal number of loci and depth of coverage driven by the project's goals and the genetic nature of the species under study (Andrews et al., 2016). Reliable *de novo* locus discovery and genotyping in diploids requires high coverage (10–20x or >20x; Andrews et al., 2016; Rochette and Catchen, 2017). According to the previous setup of the ddRADseq protocol for *E. dunnii* (Aguirre et al., 2019), a minimum of 700,000 reads per sample is required to achieve 10x depth coverage, although it is better to guarantee a minimum average of 1 million reads per sample to ensure good results for all samples. In our hands, as well as filtering for quality of ddRADseq loci and removing individuals with a high proportion of missing data, extreme heterozygosity and unexpected relatedness (in both datasets), this minimum average number of reads per sample was employed to acquire sturdy data for the study.

In the *Eucalyptus* genus, despite its high cost, WGS was applied in some cases (Kainer et al., 2019; Yong et al., 2021). RRS has been applied for association studies by using target re-sequencing of specific genes (Ghosh Dasgupta et al., 2021; Candotti et al., 2023). Besides, sequence capture, where *E. grandis* reference genome (Bartholomé et al., 2014; Myburg et al., 2014) was used to design probes and capture genomic regions to be sequenced, was evaluated (de Moraes et al., 2018). It is worth mentioning that, as the latter methodologies, ddRADseq has the potential and allows the discovery and detection of all kinds of DNA variation (e.g., copy number variants or CNV, microsatellites or SSRs, SNP, InDels, plastid DNA; Aguirre et al., 2019; Meger et al., 2019; Aballay et al., 2021) and the inclusion of all types of variants could improve the predictive ability in GS (Lyra et al., 2019; Della Coletta et al., 2021).

4.2 EUChip60K microarray application

As expected, the number of polymorphic SNPs obtained with the EUChip60K in the 308 individuals (19,011 SNPs, with Call rate >80% and MAF> 0.01) was slightly higher than the number originally reported based on genotyping only 12 individuals (17,014 SNPs; Call rate 98.8%, MAF >0.01) (Silva-Junior et al., 2015). As pointed out by Resende et al. (2017), less than 50% of the 64 thousand SNPs available in the EUChip60K are typically polymorphic in line with the multi-species nature of the EUChip60K, in which not all SNPs were designed to be informative in each single species, but rather that the chip would provide approximately 15,000 to 30,000 useful SNP in each one of almost 20 eucalypt species (Silva-Junior et al., 2015). The highest proportions of informative SNPs in the EUChip60K are generally found in species that were more represented in the sequencing data used for SNP discovery. This was the case of the work of Cappa et al. (2019), who obtained 33,398 SNPs (MAF >0.01) for 999 trees of hybrids between *E. grandis* × *E. urophylla* and *E. grandis* × *E. camaldulensis*. Conversely, for less represented species like *E. dunnii* in this work, Jurcic et al. (2021), reported 11,284 with a more rigorous MAF>0.05, and Suontama et al. (2019) reported 12,236 SNPs in 691 individuals of *E. nitens*.

4.3 Comparison of genotyping methodologies

4.3.1 Missing data and imputation

ddRADseq was chosen to evaluate against the standard EUChip60K data because it is the REbRRS method that typically yields a higher number of reliable markers, as observed for beech and oak (Ulaszewski et al., 2021), when comparing RADseq, GBS and ddRADseq. When comparing the ddRADseq and EUChip60K datasets, the latter had a much lower proportion of total and per sample missing data. This was expected due to the EUChip60K design and DNA hybridisation-based methodology (de Moraes et al., 2018). However, in order to overcome missing data, imputation methods were applied as reported for *Picea glauca* (Gamal El-Dien et al., 2015). The LD-kNNi imputation algorithm was applied to both genomic datasets for the *E. dunnii* population. High accuracies of genotype assignment to missing data were obtained (0.8949 for ddRADseq dataset with 8,170 SNPs; and 0.8443 for EUChip60K dataset with 19,045 SNPs). Likewise, an algorithm also based on the use of nearest-neighbour genotype information, kNN-Fam, together with the Expectation Maximisation algorithm, showed the highest accuracies in imputing GBS data for GS in *P. glauca* (Gamal El-Dien et al., 2015) when compared to the mean imputation and singular value decomposition methods. On the other hand, Rutkoski et al. (2013) compared four imputation methods for application in GS and found that the random forest regression method produced superior accuracy, followed by the kNNi method, and the lowest accuracy was the mean imputation method. Furthermore, they concluded that including markers with a large proportion of missing data almost always led to higher GS accuracies after imputing, even when the order of the markers is not known (Rutkoski et al., 2013). Another study in *E. cladocalyx* also applied the LD-kNNi algorithm, implemented in 5.2 (Trait Analysis Association, Evolution and Linkage; Bradbury et al., 2007), which allowed them to impute data and apply GS in this non-model species (Ballesta et al., 2020).

4.3.2 Minor allele frequencies

Eucalyptus species is primarily outcrossing and has wide pollen/seed dispersal. As a result, high proportions of polymorphic loci with rare alleles are observed in natural populations (Byrne, 2008). The breeding population genotyped in the present work is likely to retain a high proportion of such rare variants as it is only one generation removed from natural stands. SNPs with MAF >0.01 were retained for the downstream analyses since 4,011 of the 8,011 SNPs in the ddRADseq dataset had a frequency <0.05. SNPs with a MAF >0.01 were also used in GS in *Eucalyptus* (Cappa et al., 2019; Suontama et al., 2019). The inclusion of rare variants has the potential to contribute to the accuracy of GS prediction models, although the overall contribution of rare SNPs to the variance of quantitative traits in breeding populations has been questioned (Liu et al., 2015), and some reports suggest that low MAF SNPs, do not influence genomic predictions (Zhu et al., 2017; Zhang et al., 2019; Trujano-Chavez et al., 2021).

The multispecies strategy used for the development of EUChip60K somewhat mitigated the ascertainment bias towards more common SNPs (Silva-Junior et al., 2015). However, when comparing the MAF distributions obtained in the datasets of the two genotyping methods, EUChip60K SNPs showed a higher average allele frequency and a lower proportion of rarer SNPs than ddRADseq in this *E. dunnii* population. This corroborates the expectation that the EUChip60K targets more common polymorphisms in the population than ddRADseq. This same trend was observed in other studies comparing RRS and SNP arrays for the same sample set. Negro et al. (2019) working with maize observed that array data showed a uniform MAF distribution, while GBS data presented an excess of rare alleles with an "L" shaped MAF distribution. The authors justify these differences because maize microarrays (50K and 600K) were developed based on sequencing 27 and 30 lines respectively, while SNPs from GBS data were detected in 247 lines, allowing for a greater discovery of rare alleles. Otherwise, in *Eucalyptus*, as expected, the proportion of rare variants was significantly higher with the sequence capture method than with the EUChip60K (de Moraes et al., 2018).

4.3.3 SNP density

With respect to the distribution of SNPs along the genome, it was observed that SNPs obtained from ddRADseq analysis showed a more clustered and less homogeneous pattern than those obtained with the chip data. This is consistent with the expectations based on the design of the EUchip60K. To develop it, 240 tree genomes from 12 species were sequenced at a depth of 3.5× each, resulting in a total of 46,997,586 raw SNP variants. The SNPs were filtered using multivariable metrics, retaining only variant SNPs of high quality that displayed polymorphism in the largest number of species. This resulted in an array containing 60,904 SNPs, with a homogeneous genome-wide coverage of 96% (1 SNP per 12–20 kb) as reported by Silva-Junior et al., in 2015. Similar pattern was seen when EUChip60K data was compared to sequence capture data (de Moraes et al., 2018) and also in maize, where SNPs from GBS showed higher SNP density in telomeric regions, while the 50K microarray data showed a more homogeneous distribution, and the 600K microarray showed a higher density of markers in pericentromeric regions (Negro et al., 2019). These results demonstrate the benefit of pre-selecting polymorphic loci when developing a SNP array, despite the inherent limitation of variable levels of ascertainment bias. Similarly, in *E. dunnii* no SNPs were found common to both genotyping methods, which is consistent with a study in *E. globulus* (Durán et al., 2018), where it was observed that of the 2,597 SNPs obtained with the GBS method, only 24 SNPs were common to the 13,669 polymorphic SNPs presented by EUChip60K.

4.3.4 Linkage disequilibrium

Pairwise estimates of LD (r^2) between all SNPs (MAF ≥0.01) and all chromosomes were independently estimated for the three datasets. A rapid LD decay in *Eucalyptus* genus was reported in several studies, presenting values of LD that dropped below 0.2 between 5.7 Kbp to 637.7 Kbp (Silva-Junior and Grattapaglia, 2015; Muller et al., 2019; Estopa et al., 2023). *E. dunnii* showed a rapid LD decay between this range (EUChip60K dataset).

The genome-wide LD decay to an r^2 below 0.2 was significantly faster for ddRADseq (37 bp) compared to EUChip60K (6.4 Kbp). This difference can be explained by the small distances between SNPs within the same locus in the ddRADseq dataset, which were generated for 75 bp of the Illumina read length (average of 2.6 SNPs every 75bp genomic region or locus).

Similar results were obtained when comparing sequence capture with EUChip60K. The same trend was observed by de Moraes et al. (2018) for two MAF thresholds (0.05 and 0.1), falling below the 0.2 r^2 value at lower distances (50-100Kbp) for the sequence-capture SNP dataset than with the SNP array (250 Kbp) for a MAF of 0.05. They explain that these LD decay differences between datasets could be due to the intensive pre-selection step for the SNPs included in the SNP array (1 SNP every 12–20 Kbp, Silva-Junior et al., 2015) likely resulting in ascertainment bias.

4.3.5 Genetic diversity

The maintenance of genetic diversity is key to the viability of a population, particularly in a breeding program where long-term sustainable genetic gain with GS should be in balance with genetic diversity (Grattapaglia, 2022). According to the MAF distribution, for *E. dunnii* a lower average heterozygosity (He) was observed with ddRADseq (0.17) than with the microarray (0.28). This difference is consistent with the observed allele frequency spectrum of the two datasets where the SNPs genotyped with the EUChip60k have a higher average allele frequency which will result in higher heterozygosity. Negro et al. (2019) observed the same trend in maize, where He for GBS (Elshire et al., 2011) was 0.27 and for the 50K and 600K microarrays was 0.35 and 0.34, respectively. The observed heterozygosity for *E. dunnii* (Ho: 0.29) with the EUChip60K dataset was similar to that found in *E. cladocalyx* with the same SNP-chip (Ho: 0.22; Ballesta et al., 2020). As expected it was lower than the estimate for an *E. dunnii* seed orchard using nine multiallelic SSR markers (Ho: 0.66; Zelener et al., 2005).

4.3.6 Genetic structure

The population genetic structure detected with the two datasets was similar despite the differences in allele frequency distributions, and only two individuals differed in the genetic groups assignment. Such correspondence between the population structure obtained with the GBS and SNP-chip datasets was also observed by Negro et al. (2019) in maize and by Elbasyoni et al. (2018) in winter wheat. Such population genetic structure detected by DAPC analysis in *E. dunnii* population refers to two groups with little but significant genetic differentiation between them. This is likely due to the genetic composition of the smallest group (trees selected for growth and stem straightness in a local commercial plantation with a narrow genetic base) and the composition of the largest genetic group, which has many families from different geographical sources of seeds, resulting in dissimilar allele frequencies (Alqudah et al., 2019). However, the population genetic structure detected was very low (F_{ST} = 0.0148) although significant. As suggested by Yu et al. (2006), kinship relationships are able to capture the underlying genetic structure except in cases where there is an obvious regional difference (Cappa et al., 2013). In the case of the *E. dunnii* population under study, it is derived from seeds from a very narrow geographic region in Australia, corresponding to the distribution of the species, suggesting high gene flow and therefore little genetic differentiation.

4.3.7 Genomic selection

This study compared the performance of two genomic datasets and their combination, in building a genomic selection model for a breeding population of *E. dunnii.* They were contrasted with the traditional approach using pedigree information (ABLUP). The evaluation was based on their predictive ability for 11 growth and wood quality traits. When considering the present results, it must be taken into account that the size of the population studied is rather small, a factor that affects the accuracy of the prediction (Grattapaglia, 2022).

Several studies have also applied GS in *Eucalyptus* using EUChip60K data (Müller and Neves, 2017; Durán et al., 2018; Cappa et al., 2019; Suontama et al., 2019; Jurcic et al., 2021; Duarte et al., 2023), but none of them used the ddRADseq genotyping method. This present work is the first to compare ddRADseq and SNP array data for the application of GS in forest trees and the first to apply GBLUP using ddRADseq in *E. dunnii.*

Genomic approaches are expected to perform better than pedigree-based approaches because they use more accurate kinship information (Cappa et al., 2019). Our results showed that the GBLUP (with any of the three data sets) outperformed the ABLUP for six out of 11 traits. However, the ABLUP approach performed better than GBLUP for four traits, two of which were growth traits. This may be due to an overestimation of additive variation by the ABLUP approach, which cannot disentangle the non-additive variation (Muñoz and Sanchez, 2014; Gamal El-Dien et al., 2016; Cappa et al., 2019).

Previous studies have used the EUChip60K platform and applied GBLUP and five other Bayesian GS methods to predict traits in different *Eucalyptus* populations. For instance, Müller and Neves (2017) showed predictive abilities for DBH of 0.16 and 0.44 for populations of *E. benthamii* (n = 505) and *E. pellita* (n = 732), respectively. In contrast, our study found lower and no significant PA values for DBH at 6 and 20 years in the Ubajay population of *E. dunnii* (PA with chip: 0.054 and 0.035 DBH6 and DBH20, respectively, with GBLUP). These results suggest that the *E. dunnii* population has lower additive genetic variation for this trait. Another study by Durán et al. (2018) applied GBLUP and three other Bayesian GS models to predict stem volume and wood density traits in a clonal population of *E. globulus* using the EUChip60K microarray. This population had a similar size (310 trees) to the *E. dunnii* population (280 trees). The study found a higher PA value, using GBLUP, for wood density (0.63) in *E. globulus* compared to *E. dunnii* (PA WD20: 0.160 with chip data). The difference in accuracy between the two populations might be due to *E. globulus* having closer kinship relationships and involving a smaller number of families (40 full-sib families and 13 half-sib families, produced by crossing 23 parents). This corroborates the well documented fact in a number of studies that effective population size and relationship are the main drivers of genomic prediction (Grattapaglia, 2022; Isik, 2022).

In *E. benthamii,* Estopa et al. (2023) compared different genomic prediction models with ABLUP in a population of 780 individuals from 77 families genotyped with the EUChip60K and phenotyped for five traits, including wood density, extractives content, and lignin content. They found that the PAs for ABLUP were lower than for GBLUP for all five traits, which is consistent with the results of the present work. For lignin content, the PA values were 0.23 for ABLUP and 0.34 for GBLUP, which are similar to the results of the present work (0.269 for ABLUP and 0.368 for GBLUP). For extractive content, the PA values were 0.16 for ABLUP and 0.18 for GBLUP. In comparison to *E. dunnii*, ABLUP was similar (0.156) and GBLUP was lower (0.258). For wood density, the PA values were 0.27 for ABLUP and 0.43 for GBLUP, which were higher than in the present work (ABLUP: 0.190 and GBLUP: 0.160).

Genomic selection in *E. dunnii* using EUChip60K data has only been applied in a few studies. In a preliminary study by Naidoo et al. (2018), GBLUP was applied to an *E. dunnii* population in South Africa. The study analyzed 9,102 SNP markers in 840 offspring from 89 half-sib families and applied GBLUP to predict five phenotypic traits. The results showed PA values of 0.38 for diameter at breast height and 0.51 for wood density. However, much lower values were found in the present study, which could be due to different environments, different origins, and a small number of genotyped individuals. Jones et al. (2019) investigated whether combining data from different trails could improve the accuracy of the GS model in *E. dunnii.* The study found that accuracy for diameter at breast height increased by 86% and tree height by 290% (from 0.18 to 0.72). Jurcic et al. (2021) applied GS in *E. dunnii* using multiple-trait multiple-site single-step GBLUP (ssGBLUP models) for DBH6 and SS6. The h^2 of DBH6 obtained by ABLUP was 0.262 (s.d.: 0.039), which is similar to the present work results (0.242), and the PA was 0.324, but near zero in the present work. For SS6, the h^2 was 0.19 and the PA was 0.350, while in the present work, the h^2 was 0.413 and PA 0.25 for the same trait. However, Jurcic et al. (2021) applied a model using both genomic and pedigree information, based on a different number of individuals in the population (1,520 trees), and two additional trials, such that the PA and h^2 values are not directly comparable.

In general, it can be concluded that models including genomic data are promising for the application in breeding programs, in particular in *E. dunnii*, as they show higher PA for most of the traits, compared to ABLUP. These models can be used to generate a ranking of *E. dunnii* individuals based on the priorities of the breeding program, such as selecting individuals with high wood quality and higher growth. The choice of genotyping platform is a key element that can affect the performance of GS (Elbasyoni et al., 2018). Sequencing-based genotyping methods in principle provide a large number of molecular markers, but often have a high proportion of missing data requiring rigorous filtering that often result in an operationally lower number of SNPs when compared to chip-based data. SNP arrays, on the other hand, provide large number of markers with very little missing data, but due to their fixed content may suffer from ascertainment bias in allele frequencies and do not allow the discovery of population-specific variants (Albrechtsen et al., 2010; Li and Kimmel, 2013; Bajgain et al., 2016). Elbasyoni et al. (2018) compared the performance of these two genotyping platforms for GS in 299 lines of winter hard wheat (*Triticum aestivum L.*), one of the few studies that compared the performance of these genotyping methods for GS in crops. They observed that GBS, imputing 10% (10,775 SNPs) and 50% (39,674 SNPs) of missing data, showed similar or even higher genomic prediction accuracy than the microarray data (19,515 SNPs) for all agronomic traits, depending on the percentage of missing data imputed from the starting GBS. In contrast, for *E. dunnii*, although ddRADseq showed slightly higher PA for some traits, the EUChip60K data provided higher PA values. This suggests that the performance of different genotyping platforms can vary depending on the species and population being studied. In the present work, it was observed that EUChip60K provided higher PA for most of the traits compared to ddRADseq. A similar trend was observed in the study by de Moraes et al. (2018), where the performance of sequence capture and EUChip60K was compared for GS. The study found that the microarray method showed higher PA for most traits.

There are various factors that affect the accuracy of prediction models, and one of them is the genotyping density (Grattapaglia, 2022). The EUChip60K dataset was found to be the most effective in predicting most of the traits evaluated in this study. This could be related to its higher marker density, which can better explain phenotypic variation compared to ddRADseq for most of the traits evaluated. Nevertheless, for two traits, the ddRADseq + EUChip60K dataset showed a higher PA, indicating that the number of markers is not the only factor that influences the PA. Kinship also plays an important role and studies on forest trees demonstrate that moderate genotyping densities of around 10,000 to 15,000 data points are sufficient for reasonable predictive power (Grattapaglia, 2022). A similar observation was made by de Moraes et al. (2018), where the combined datasets of sequence capture and EUChip60K did not improve the accuracy of the model.

Growth traits, such as diameter at breast height, are likely to be related to fitness and are controlled by a large number of genes (Falconer and Mackay, 1996) with a large interaction with the environment therefore expressing a low heritability (Nunes et al., 2016). Diameter at breast height DBH showed low to moderate h^2, consistent with reports in other eucalypts with values between 0.11 and 0.41 (Gallo et al., 2018; Cappa et al., 2019; Marco de Lima et al., 2019; Jurcic et al., 2021). Chemical traits, on the other hand, are often related to specific biosynthesis pathway, likely controlled by fewer loci, less influenced by the environment with higher heritability (Gion et al., 2011). The heritability estimated from the pedigrees for growth and wood quality traits were moderate to high, with LESI20 and TL20 showing higher values than most wood quality traits estimated from NIR analysis or growth, in agreement with other evaluations in *Eucalyptus* (Resende et al., 2017; Tan et al., 2017; Cappa et al., 2019; Marco de Lima et al., 2019), and *E. dunnii* (Jurcic et al., 2021).

Heritability for wood density, S/G ratio and extractive content in the *E. dunnii* population were generally estimated in the same range as in previous studies of other *Eucalyptus* species (Stackpole et al., 2011; Makouanzi et al., 2017; Resende et al., 2017; Tan et al., 2017; Varghese et al., 2017; Gallo et al., 2018; Marco de Lima et al., 2019; Cappa et al., 2019; Suontama et al., 2019; Paludeto etal., 2021). For *E. dunnii*, Gallo et al. (2018) observed high broad-sense individual heritability estimates (0.64) for Klason lignin. NIR estimates of KL yielded moderate to high narrow heritability values in half-sib progeny of *E. camaldulensis* (0.21), *E. globulus* (0.27) and *E. urophylla* (0.76) (Stackpole et al., 2011; Hein et al., 2012; Varghese et al., 2017). Similar results were found for *E. dunnii* in this paper, where narrow h^2 showed high values (KL: 0.669 and TL: 0.726), which are also in the range of the literature (de Moraes et al., 2018; Cappa et al., 2019; Marco de Lima et al., 2019). These results suggest that this trait has a high level of genetic control with a better possibility of obtaining significant genetic gains. This is also evidenced by the highest PAs obtained by genomic selection in this study. A positive correlation between heritability and predictive ability was observed in *E. dunnii* regardless of the genotyping data used. Furthermore, this correlation has already been demonstrated by Grattapaglia and Resende (2010) by means of simulations. This trend has also been observed for *Eucalyptus* (de Moraes et al., 2018), pine (Resende et al., 2012), animals (Hayes et al., 2014) and crops (Poland et al., 2012; Crossa et al., 2013).

In summary, when comparing the ddRADseq and EUChip60K methodologies, we observed differences in the percentage of missing data, genome-wide marker coverage, minor allele frequency ratios and in the estimation of genetic diversity parameters. Furthermore, no major differences were observed in the estimation of genetic structure and linkage disequilibrium.

Regarding their performance in GS, the inclusion of any of the three genomic data sets in the prediction models increases the predictive ability of the estimates compared to traditional methods. This trend was observed for most of the traits evaluated that showed significant values (six out of ten), all of them being wood quality traits. This indicates an advantage of using genomic data for selection. When comparing the ddRADseq and EUChip60K datasets, the EUChip60K yielded higher predictive abilities in most cases, although ddRADseq provided slightly higher predictions for some traits.

Both genotyping methods, ddRADseq and EUChip60K, are generally comparable for diversity analysis and genomic prediction, demonstrating the usefulness of the former provided that it undergoes rigorous SNP filtering. The results of this study provide a foundation for future whole-genome studies using ddRADseq in non-model forest species for which SNP arrays have not been developed.

Data availability statement

The data presented in the study are deposited in the European Variation Archive (EVA) in the EMBL-EBI repository, accession number PRJEB73817 (https://www.ebi.ac.uk/eva/?eva-study=PRJEB73817).

Author contributions

NA: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Visualization, Writing–original draft, Writing–review and editing, Software. PV: Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing–review and editing. MG: Formal Analysis, Investigation, Methodology, Software, Visualization, Writing–review and editing. CF: Investigation, Methodology, Software, Writing–review and editing. JR: Formal Analysis, Investigation, Resources, Writing–review and editing. MM: Conceptualization, Funding acquisition, Project administration, Visualization, Writing–review and editing. CA: Investigation, Resources, Visualization, Writing–review and editing. AL: Investigation, Resources, Writing–review and editing. JL: Data curation, Investigation, Resources, Writing–review and editing. PP: Investigation, Resources, Writing–review and editing. DP: Investigation, Resources, Writing–review and editing. LH: Investigation, Resources, Writing–review and editing. JO: Investigation, Resources, Writing–review and editing. MM: Conceptualization, Funding acquisition, Resources, Writing–review and editing. EC: Funding acquisition, Project administration, Writing–review and editing. RC: Project administration, Writing–review and editing. AM: Methodology, Resources, Writing–review and editing. JR: Methodology, Resources, Writing–review and editing. HH: Conceptualization, Funding acquisition, Writing–review and editing. DG: Conceptualization, Funding acquisition, Visualization, Writing–review and editing. EC: Conceptualization, Software, Visualization, Writing–review and editing. NP: Conceptualization, Funding acquisition, Supervision, Visualization, Writing–review and editing. SM: Conceptualization, Funding acquisition, Project administration, Supervision, Visualization, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by the BIOTECH II platform (grant number 373-780), which funded EUChip60K genotyping and NIR phenotyping; Programa de Sustentabilidad y Competitividad Forestal (PROMEF), Unidad para el Cambio Rural (UCAR), which funded staff salaries; Programa Nacional Forestal de INTA (PNFOR-1104064), which funded trial maintenance and phenotyping of growth traits; Proyecto de Investigación Científica y Tecnológica del Ministerio de Ciencia, Tecnología e Innovación (PICT 2014-0795) and Proyecto INTA 2019-PE-E6-I114-001, which funded the genotyping by ddRADseq; Programa Nacional de Biotecnología de INTA (PNBIO 1131042), which funded DNA extraction and set up ddRADseq protocol. NA was supported by doctoral fellowships from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). This work used computational resources from BIOCAD, IABiMo-UEDD INTA-CONICET, Programa de Sustentabilidad y Competitividad Forestal/BID 2853/Oc-Ar, Consorcio Argentino de Tecnología Genómica, MinCyT PPL 2011 004.

Acknowledgments

We thank the Genomics Unit (Unidad de genomica, UGB) of the Institute of Agrobiotechnology and Molecular Biology (IABiMo, INTA) for technical assistance in library preparation and NGS data generation.

Conflict of interest

The 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/fgene.2024.1361418/full#supplementary-material

References

Aballay, M. M., Aguirre, N. C., Filippi, C. V., Valentini, G. H., and Sánchez, G. (2021). Fine-tuning the performance of ddRAD-seq in the peach genome. *Sci. Rep.* 11, 6298. doi:10.1038/s41598-021-85815-0

Aguirre, N., Filippi, C., Zaina, G., Rivas, J., Acuña, C., Villalba, P., et al. (2019). Optimizing ddRADseq in non-model species: a case study in *Eucalyptus dunnii* maiden. *Agronomy* 9 (9), 484. doi:10.3390/agronomy9090484

Aguirre, N. C., Filippi, C. V., Vera, P. A., Puebla, A. F., Zaina, G., Lia, V. V., et al. (2023). "Double digest restriction-site associated DNA sequencing (ddRADseq) technology," in *Plant genotyping. Methods in molecular biology*. Editor Y. Shavrukov (New York, NY: Springer-Nature publisher. Humana). doi:10.1007/978-1-0716-3024-2_4

Albrechtsen, A., Nielsen, F. C., and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* 27 (11), 2534–2547. doi:10.1093/molbev/msq148

Alqudah, A. M., Sallam, A., Stephen Baenziger, P., and Börner, A. (2019). GWAS: fast-forwarding gene identification and characterization in temperate Cereals: lessons from Barley - a review. *J. Adv. Res.* 22, 119–135. doi:10.1016/j.jare.2019.10.013

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., and Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* 17 (2), 81–92. doi:10.1038/nrg.2015.28

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3 (10), e3376–e3377. doi:10.1371/journal.pone.0003376

Bajgain, P., Rouse, M. N., and Anderson, J. A. (2016). Comparing genotyping-by-sequencing and single nucleotide polymorphism chip genotyping for quantitative trait loci mapping in wheat. *Crop Sci.* 56 (1), 232–248. doi:10.2135/cropsci2015.06.0389

Ballesta, P., Bush, D., Silva, F. F., and Mora, F. (2020). Genomic predictions using low-density SNP markers, pedigree and gwas information: a case study with the non-model species *Eucalyptus cladocalyx*. *Plants* 9, 99. doi:10.3390/plants9010099

Ballesta, P., Serra, N., Guerra, F. P., Hasbún, R., and Mora, F. (2018). Genomic prediction of growth and stem quality traits in *Eucalyptus globulus* Labill. at its southernmost distribution limit in Chile. *Forests* 9 (12), 779. doi:10.3390/f9120779

Bartholomé, J., Mandrou, E., Mabiala, A., Jenkins, J., Nabihoudine, I., Klopp, C., et al. (2014). High-resolution genetic maps of Eucalyptus improve *Eucalyptus grandis* genome assembly. *New Phytol.* 206 (4), 1283–1296. doi:10.1111/nph.13150

Bayer, M., Morris, J. A., Booth, C., Booth, A., Uzrek, N., and Russell, J. (2019). "Exome cap-ture for variant discovery and analysis in barley," in *Barley, Methods in molecular biology*. Editor HarwoodW (NewYork: Humana Press), pp283–310. doi:10.1007/978-1-4939-8944-7_18

Bernhardsson, C., Zan, Y. J., Chen, Z. Q., Ingvarsson, P. K., and Wu, H. X. (2021). Development of a highly efficient 50K single nucleotide polymorphism genotyping array for the large and complex genome of Norway spruce (*Picea abies* L. Karst) by whole genome resequencing and its transferability to other spruce species. *Mol. Ecol. Resour.* 21, 880–896. doi:10.1111/1755-0998.13292

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23 (19), 2633–2635. doi:10.1093/bioinformatics/btm308

Burridg, A. J., Winfield, M. O., Wilkinson, P. A., Przewieslik-Allen, A. M., Edwards, K. J., and Barker, G. L. A. (2022). The use and limitations of exome capture to detect novel variation in the hexaploid wheat genome. *Front. Plant Sci.* 13, 841855. doi:10.3389/fpls.2022.841855

Byrne, M. (2008). "Phylogeny, diversity and evolution of eucalypts," in *Plant genome: biodiversity and evolution, volume 1, part E*. Editors A. K. Sharma and A. Sharma (Enfield: Science Publishers), 303–346.

Caballero, M., Lauer, E., Bennett, J., Zaman, S., McEvoy, S., Acosta, J., et al. (2021). Toward genomic selection in *Pinus taeda*: integrating resources to support array design in a complex conifer genome. *Appl. Plant Sci.* 9 (6), e11439. doi:10.1002/aps3.11439

Campbell, B., Dupuis, S., Dupuis, J. R., and Sperling, F. A. H. (2018). Would an RRS by any other name sound as RAD? *Methods Ecol. Evol.* 9 (9), 1920–1927. doi:10.1111/2041-210X.13038

Candotti, J., Christie, N., Ployet, R., Mostert-O'Neill, M. M., Reynolds, S. M., Neves, L. G., et al. (2023). Haplotype mining panel for genetic dissection and breeding in Eucalyptus. *Plant J.* 113 (1), 174–185. doi:10.1111/tpj.16026

Cappa, E. P., de Lima, B. M., da Silva-Junior, O. B., Garcia, C. C., Mansfield, S. D., and Grattapaglia, D. (2019). Improving genomic prediction of growth and wood traits in Eucalyptus using phenotypes from non-genotyped trees by single-step GBLUP. *Plant Sci.* 284, 9–15. doi:10.1016/j.plantsci.2019.03.017

Cappa, E. P., El-Kassaby, Y. A., Garcia, M. N., Acuña, C., Borralho, N. M., Grattapaglia, D., et al. (2013). Impacts of population structure and analytical models in genome-wide association studies of complex traits in forest trees: a case study in *Eucalyptus globulus*. *PLoS ONE* 8, e81267. doi:10.1371/journal.pone.0081267

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22 (11), 3124–3140. doi:10.1111/mec.12354

Cezard, T., Cunningham, F., Hunt, S. E., Koylass, B., Kumar, N., Saunders, G., et al. (2021). The European Variation Archive: a FAIR resource of genomic variation for all species. *Nucleic Acids Res.* gkab960. doi:10.1093/nar/gkab960

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4 (1), 7. doi:10.1186/s13742-015-0047-8

Clarke, C. (2009). "The profitable pulp mill," in Australian Forest Genetics Conference (Australia).

Covarrubias-Pazaran, G. (2016). Genome assisted prediction of quantitative traits using the R package sommer. *PLoS ONE* 11, 01567444. doi:10.1371/journal.pone.0156744

Crossa, J., Beyene, Y., Kassa, S., Pérez, P., Hickey, J. M., Chen, C., et al. (2013). Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3* 3, 1903–1926. doi:10.1534/g3.113.008227

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

Darrow, W. K. (1994). "The effect of drought on eucalypt species growing on shallow soils in South Africa," in *I. Effect on mortality and growth* (South Africa: Institute for Commercial Forestry Research). (No. 7/94). Bulletin series. Available at: https://books.google.com.ar/books?id=−4nEzQEACAAJ.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12 (7), 499–510. doi:10.1038/nrg3012

Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B., and Hirsch, C. N. (2021). How the pan-genome is changing crop genomics and improvement. *Genome Biol.* 22, 3. doi:10.1186/s13059-020-02224-8

de Moraes, B. F. X., dos Santos, R. F., de Lima, B. M., Aguiar, A. M., Missiaggia, A. A., da Costa Dias, D., et al. (2018). Genomic selection prediction models comparing sequence capture and SNP array genotyping methods. *Mol. Breed.* 38, 115. doi:10.1007/s11032-018-0865-3

Deschamps, S., Llaca, V., and May, G. D. (2012). Genotyping-by-sequencing in plants. *Biology1* 1, 460–483. doi:10.3390/biology1030460

Desta, Z. A., and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends Plant S. C.* 19 (9), 592–601. doi:10.1016/j.tplants.2014.05.006

Dixon, P. (2003). VEGAN, A package of R functions for community ecology. *J. Veg. Sci.* 14 (6), 927–930. doi:10.1111/j.1654-1103.2003.tb02228.x

Duarte, D., Dutour, J., Jurcic, E. J., Villalba, P. V., Centurión, C., and Cappa, E. P. (2023). Genomic selection comes to life: unraveling its potential in an advanced four-generation *Eucalyptus grandis* population. *Agrocienc Urug* 27 (NE2), e1250. doi:10.31285/AGRO.27.1250

Durán, R., Zapata-Valenzuela, J., Balocchi, C., and Valenzuela, S. (2018). Efficiency of EUChip60K pipeline in fingerprinting clonal population of *Eucalyptus globulus*. *Trees* 32, 663–669. doi:10.1007/s00468-017-1637-0

Elbasyoni, I. S., Lorenz, A. J., Guttieri, M., Frels, K., Baenziger, P. S., Poland, J., et al. (2018). A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. *Plant Sci.* 270, 123–130. doi:10.1016/j.plantsci.2018.02.019

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 (5), 193799–e19410. doi:10.1371/journal.pone.0019379

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255. doi:10.3835/plantgenome2011.08.0024

Estopa, R. A., Paludeto, J. G. Z., Müller, B. S. F., de Oliveira, R. A., Azevedo, C. F., de Resende, M. D. V., et al. (2023). Genomic prediction of growth and wood quality traits in *Eucalyptus benthamii* using different genomic models and variable SNP genotyping density. *New For.* 54, 343–362. doi:10.1007/s11056-022-09924-y

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to quantitative genetics*. United States: Longman Group Limited.

Fuentes-Pardo, A. P., and Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: advantages, limitations and practical recommendations. *Mol. Ecol.* 26, 5369–5406. doi:10.1111/mec.14264

Gallo, R., Barcellos Pantuza, I., dos Santos, G. A., Vilela de Resende, M. D., Xavier, A., Ferreira Simiqueli, G., et al. (2018). Growth and wood quality traits in the genetic selection of potential *Eucalyptus dunnii* Maiden clones for pulp production. *Industrial Crops Prod.* 123, 434–441. doi:10.1016/j.indcrop.2018.07.016

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Chen, C., Porth, I., and El-Kassaby, Y. A. (2015). Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. *BMC Genomics* 16, 370. doi:10.1186/s12864-015-1597-y

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., and El-Kassaby, Y. A. (2016). Implementation of the realized genomic relationship matrix to open-pollinated white spruce family testing for disentangling additive from nonadditive genetic effects. *G3 Gene Genomes| Genet.* 6 (3), 743–753. doi:10.1534/g3.115.025957

Geraldes, A., Difazio, S. P., Slavov, G. T., Ranjan, P., Muchero, W., Hannemann, J., et al. (2013). A 34K SNP genotyping array for *Populus trichocarpa*: design, application to the study of natural populations and transferability to other Populus species. *Mol. Ecol. Resour.* 13, 306–323. doi:10.1111/1755-0998.12056

Ghosh Dasgupta, M., Abdul Bari, M. P., Shanmugavel, S., Dharanishanthi, V., Muthupandi, M., Kumar, N., et al. (2021). Targeted re-sequencing and genome-wide association analysis for wood property traits in breeding population of *Eucalyptus tereticornis* × *E. grandis*. *Genomics* 113 (6), 4276–4292. doi:10.1016/j.ygeno.2021.11.013

Gion, J.-M., Carouché, A., Deweer, S., Bedon, F., Pichavant, F., Charpentier, J.-P., et al. (2011). Comprehensive genetic dissection of wood properties in a widely-grown tropical tree: Eucalyptus. *BMC Genomics* 12, 301. doi:10.1186/1471-2164-12-301

Graham, N., Telfer, E., Frickey, T., Slavov, G., Ismael, A., Klápštõe, J., et al. (2022). Development and validation of a 36K SNP array for radiata pine (*Pinus radiata* D.don). *Forests* 13, 176. doi:10.3390/f13020176

Granato, I. S. C., Galli, G., de Oliveira Couto, E. G., e Souza, M. B., Mendonça, L. F., and Fritsche-Neto, R. (2018). snpReady: a tool to assist breeders in genomic analysis. *Mol. Breed.* 38 (8), 102. doi:10.1007/s11032-018-0844-8

Grattapaglia, D. (2022). Twelve years into genomic selection in forest trees: climbing the slope of enlightenment of marker assisted tree breeding. *Forests* 13 (10), 1554. doi:10.3390/f13101554

Grattapaglia, D., and Bradshaw, H. D., Jr. (1994). Nuclear DNA content of commercially important Eucalyptus species and hybrids. *Can. J. For. Res.* 24 (5), 1074–1078. doi:10.1139/x94-142

Grattapaglia, D., de Alencar, S., Pappas, G., Ziegler, D., Dumermuth, E., Antz, S., et al. (2011). Genome-wide genotyping and SNP discovery by ultra-deep Restriction-Associated DNA (RAD) tag sequencing of pooled samples of *E. grandis* and *E. globulus*. *BMC Proc.* 5 (Suppl. 7), P45. doi:10.1186/1753-6561-5-S8-P45

Grattapaglia, D., and Kirst, M. (2008). Eucalyptus applied genomics: from gene sequences to breeding tools. *New Phytol.* 179 (4), 911–929. doi:10.1111/j.1469-8137.2008.02503.x

Grattapaglia, D., and Resende, M. D. V. (2010). Genomic selection in forest tree breeding. *Tree Genet. Genomes* 7, 241–255. doi:10.1007/s11295-010-0328-4

Grattapaglia, D., Silva-Junior, O. B., Kirst, M., de Lima, B. M., Faria, D. A., and Pappas, G. J. (2011). High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. *BMC Plant Biol.* 11, 65. doi:10.1186/1471-2229-11-65

Hardner, C. M., Healey, A. L., Downes, G., Herberling, M., and Gore, P. L. (2016). Improving prediction accuracy and selection of open-pollinated seed-lots in *Eucalyptus dunnii* maiden using a multivariate mixed model approach. *Ann. For. Sci.* 73 (4), 1035–1046. doi:10.1007/s13595-016-0587-9

Hayes, B., and Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. *Genome* 53 (11), 876–883. doi:10.1139/G10-076

Hayes, B. J., MacLeod, I. M., Daetwyler, H. D., Bowman, P. J., Chamberlian, A. J., Vander Jagt, C. J., et al. (2014). "Genomic prediction from whole genome sequence in livestock: the 1000 bull genomes project," in *Proceedings of the 10th world congress on genetics applied to livestock production* (Vancouver), 1–6.

Hein, P. R. G., Bouvet, J. M., Mandrou, E., Vigneron, P., Clair, B., and Chaix, G. (2012). Age trends of microfibril angle inheritance and their genetic and environmental correlations with growth, density and chemical properties in *Eucalyptus urophylla* S.T. Blake wood. *Ann. For. Sci.* 69, 681–691. doi:10.1007/s13595-012-0186-3

Hill, W. G., and Weir, B. S. (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Popul. Biol.* 33 (1), 54–78. doi:10.1016/0040-5809(88)90004-4

Hoisington, D., González de León, D., and Khairallah, M. (1994). *Laboratory protocols: CIMMYT applied molecular genetics laboratory protocols (2da edició)*. México: CIMMYT.

Howe, G. T., Jayawickrama, K., Kolpak, S. E., Kling, J., Trappe, M., Hipkins, V., et al. (2020). An Axiom SNP genotyping array for Douglas-fir. *BMC Genomics* 21, 9. doi:10.1186/s12864-019-6383-9

Illumina (2010). *Infinium genotyping data analysis – a guide for analyzing infinium genotyping data using the genomestudio genotyping module*. United States: Illumina Inc. Available at: https://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf.

Isik, F. (2014). Genomic selection in forest tree breeding: the concept and an outlook to the future. *New For.* 45, 379–401. doi:10.1007/s11056-014-9422-z

Isik, F. (2022). "Genomic prediction of complex traits in perennial plants: a case for forest trees," in *Genomic prediction of complex traits. Methods in molecular biology*. Editors N. Ahmadi and J. Bartholomé (New York, NY: Humana), 2467. doi:10.1007/978-1-0716-2205-6_18

Jackson, C., Christie, N., Reynolds, S. M., Marais, G. C., Tii-kuzu, Y., Caballero, M., et al. (2022). A genome-wide SNP genotyping resource for tropical pine tree species. *Mol. Ecol. Resour.* 22 (2), 695–710. doi:10.1111/1755-0998.13484

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24 (11), 1403–1405. doi:10.1093/bioinformatics/btn129

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 11 (1), 94. doi:10.1186/1471-2156-11-94

Jones, N., Naidoo, S., Kanzlera, A., and Myburg, A. (2019). "Genomic prediction by combining data across *Eucalyptus dunnii* populations," in *IUFRO tree biotechnology 2019 meeting. Forests, technology and society. 23 to 29 of june* (Raleigh, USA: North Carolina State University).

Jovanovic, T., Arnold, R., and Booth, T. (2000). Determining the climatic suitability of *Eucalyptus dunnii* for plantations in Australia, China and central and South America. *New For.* 19 (3), 215–226. doi:10.1023/a:1006662718206

Jurcic, E. J., Villalba, P. V., Pathauer, P. S., Palazzini, D. A., Gpj, O., Harrand, L., et al. (2021). Single-step genomic prediction of *Eucalyptus dunnii* using different identity-by-descent and identity-by-state relationship matrices. *Hered. (Edinb).* 127 (2), 176–189. doi:10.1038/s41437-021-00450-9

Kainer, D., Padovan, A., Degenhardt, J., Krause, S., Mondal, P., Foley, W. J., et al. (2019). High marker density GWAS provides novel insights into the genomic architecture of terpene oil yield in Eucalyptus. *New Phytol.* 223 (3), 1489–1504. doi:10.1111/nph.15887

Kastally, C., Niskanen, A. K., Perry, A., Kujala, S. T., Avia, K., Cervantes, S., et al. (2022). Taming the massive genome of Scots pine with PiSy50k, a new genotyping array for conifer research. *Plant J.* 109, 1337–1350. doi:10.1111/tpj.15628

Klápště, J., Ashby, R. L., Telfer, E. J., Graham, N. J., Dungey, H. S., Brauning, R., et al. (2021). The use of "genotyping-by-sequencing" to recover shared genealogy in genetically diverse Eucalyptus populations. *Forests* 12, 904. doi:10.3390/f12070904

Ladiges, P. Y., Udovicic, F., and Nelson, G. (2003). Australian biogeographical connections and the phylogeny of large genera in the plant family Myrtaceae. *J. Biogeogr.* 30 (7), 989–998. doi:10.1046/j.1365-2699.2003.00881.x

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

Li, B., and Kimmel, M. (2013). Factors influencing ascertainment bias of microsatellite allele sizes: impact on estimates of mutation rates. *Genetics* 195 (2), 563–572. doi:10.1534/genetics.113.154161

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinforma. Appl. Note* 25 (16), 2078–2079. doi:10.1093/bioinformatics/btp352

Liu, H., Meuwissen, T., Sørensen, A. C., and Berg, P. (2015). Upweighting rare favourable alleles increases long-term genetic gain in genomic selection programs. *Genet. Sel. Evol.* 47, 19. doi:10.1186/s12711-015-0101-0

López, J., Borralho, N., López, A. J., Marcó, M., and Harrand, L. (2016). Variación genética del índice de rajado de rollizos en *Eucalyptus dunnii* Maiden. *Ciencia Investigación For.* 22 (2), 23–34. doi:10.52904/0718-4646.2016.454

Lu, M., Krutovsky, K. V., Nelson, C. D., Koralewski, T. E., Byram, T. D., and Loopstra, C. A. (2016). Exome genotyping, linkage disequilibrium and population structure in loblolly pine (*Pinus taeda* L.). *BMC Genomics* 17, 730. doi:10.1186/s12864-016-3081-8

Lyra, D. H., Galli, G., Alves, F. C., Granato, Í. S. C., Vidotti, M. S., Bandeira E Sousa, M., et al. (2019). Modeling copy number variation in the genomic prediction of maize hybrids. *Theor. Appl. Genet.* 132 (1), 273–288. doi:10.1007/s00122-018-3215-2

Maiden, J. H. (1905). On a new species of Eucalyptus from northern New South Wales. *Proc. Linn. Soc. N. S. W.* 30, 336–338. doi:10.5962/bhl.part.12906

Makouanzi, G., Chaix, G., and Nourissier, S. (2017). Genetic variability of growth and wood chemical properties in a clonal population of *Eucalyptus urophylla×Eucalyptus grandis* in the Congo. *South. For. J. For. Sci.* 1–8. doi:10.2989/20702620.2017

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27 (2), 209–220.

Marcó, M., and White, T. L. (2002). Genetic parameter estimates and genetic gains for *Eucalyptus* grandis and *E. Dunnii* in Argentina. *For. Genet.* 9 (3), 205–215.

Marcó, M. A. (2005). Eucalyptus de Rápido crecimiento para usos sólidos. *INTA* 5 (8), 178–179.

Marco de Lima, B., Cappa, E. P., Silva-Junior, O. B., Garcia, C., Mansfield, S. D., and Grattapaglia, D. (2019). Quantitative genetic parameters for growth and wood properties in *Eucalyptus* "urograndis" hybrid using near-infrared phenotyping and genome-wide SNP-based relationships. *PLoS ONE* 24, e0218747. doi:10.1371/journal.pone.0218747

Marcucci Poltri, S. N., Zelener, N., Rodriguez Traverso, J., Gelid, P., and Hopp, H. E. (2003). Selection of a seed orchard of *Eucalyptus dunnii* based on genetic diversity criteria calculated using molecular markers. doi:10.1093/treephys/23.9.625

Maseko, B., Burgess, T. I., Coutinho, T. A., and Wingfield, M. J. (2007). Two new Phytophthora species from South African Eucalyptus plantations. *Mycol. Res.* 111 (Pt 11), 1321–1338. doi:10.1016/j.mycres.2007.08.011

Meger, J., Ulaszewski, B., Vendramin, G. G., and Burczyk, J. (2019). Using reduced representation libraries sequencing methods to identify cpDNA polymorphisms in European beech (*Fagus sylvatica* L). *Tree Genet. Genomes* 15, 7. doi:10.1007/s11295-018-1313-6

Merino, G. (2018). *Imputación de Genotipos Faltantes en Datos de Secuencación Masiva*. Córdoba, Argentina: Universidad Nacional de Córdoba.

Meuwissen, T., and Goddard, M. (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185 (2), 623–631. doi:10.1534/genetics.110.116590

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157 (4), 1819–1829. doi:10.1093/genetics/157.4.1819

Mhoswa, L., O'Neill, M. M., Mphahlele, M. M., Oates, C. N., Payn, K. G., Slippers, B., et al. (2020). Genome-wide association study for resistance to the insect pest *Leptocybe invasa* in *Eucalyptus grandis* reveals genomic regions and positional candidate defense genes. *Plant Cell. Physiology* 61 (7), 1285–1296. doi:10.1093/pcp/pcaa057

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G. Y., and Myles, S. (2015). LinkImpute: fast and accurate genotype imputation for nonmodel organisms. *G3 Genes., Genomes, Genet.* 5 (11), 2383–2390. doi:10.1534/g3.115.021667

Mostert-O'Neill, M. M., Reynolds, S. M., Acosta, J. J., Lee, D. J., Borevitz, J. O., and Myburg, A. A. (2020). Genomic evidence of introgression and adaptation in a model subtropical tree species, *Eucalyptus grandis*. *Mol. Ecol.* doi:10.1111/mec.15615

Müller, B. S. F., de Almeida Filho, J. E., Lima, B. M., Garcia, C. C., Missiaggia, A., Aguiar, A. M., et al. (2019). Independent and Joint-GWAS for growth traits in Eucalyptus by assembling genome-wide data for 3373 individuals across four breeding populations. *New Phytol.* 221 (2), 818–833. Epub 2018 Sep 25. PMID: 30252143. doi:10.1111/nph.15449

Müller, B. S. F., Neves, L. G., de Almeida Filho, J. E., Resende, M. F. R., Muñoz, P. R., Dos Santos, P. E. T., et al. (2017). Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of *Eucalyptus*. *BMC Genomics* 18, 524. doi:10.1186/s12864-017-3920-2

Muñoz, F., and Sánchez, L. (2014). *breedR: statistical methods for forest genetic resources analysts*. Available at: https://github.com/famuvie/breedR.

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., et al. (2014). The genome of *Eucalyptus grandis*. *Nature* 510 (7505), 356–362. doi:10.1038/nature13308

Naidoo, R., Jones, N., Kanzler, A., and Myburg, A. (2018). "Genomic selection modelling of growth and wood properties in *Eucalyptus dunnii*," in *Cirad - FRA, IUFRO - AUT, MUSE - FRA. 2018. Eucalyptus 2018: managing Eucalyptus plantation under global changes* (Montpellier, France: Abstracts book Montpellier CIRAD), 225. doi:10.19182/agritrop/00023

Neale, D., and Kremer, A. (2011). Forest tree genomics: growing resources and applications. *Nat. Rev. Genet.* 12, 111–122. doi:10.1038/nrg2931

Negro, S. S., Millet, E. J., Madur, D., Bauland, C., Combes, V., Welcker, C., et al. (2019). Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. *BMC Plant Biol.* 19 (1), 318. doi:10.1186/s12870-019-1926-4

Nejati-Javaremi, A., Smith, C., and Gibson, J. P. (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Animal Sci.* 75 (7), 1738–1745. doi:10.2527/1997.7571738x

Nielsen, R., Paul, J., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451. doi:10.1038/nrg2986

Nunes, A. C. P., Santos, G. A., Resende, M. D. V., Silva, L. D., Higa, A., and Assis, T. F. (2016). Estabelecimento de zonas de melhoramento para clones de eucalipto no Rio Grande do Sul. *Sci. For.* 44, 563–574. doi:10.18671/scifor.v44n111.03

Paludeto, J. G. Z., Grattapaglia, D., Estopa, R. A., and Tambarussi, E. V. (2021). Genomic relationship–based genetic parameters and prospects of genomic selection for growth and wood quality traits in *Eucalyptus benthamii*. *Tree Genet. Genomes* 17, 38. doi:10.1007/s11295-021-01516-9

Parchman, T. L., Jahner, J. P., Uckele, K. A., Galland, L. M., and Eckert, A. J. (2018). RADseq approaches and applications for forest tree genetics. *Tree Genet. Genomes* 14 (3), 39. doi:10.1007/s11295-018-1251-3

Patterson, H. D., and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58 (3), 545–554. doi:10.1093/biomet/58.3.545

Pavan, S., Delvento, C., Ricciardi, L., Lotti, C., Ciani, E., and D'Agostino, N. (2020). Recommendations for choosing the genotyping method and best practices for quality control in crop genome-wide association studies. *Front. Genet.* 11, 447. doi:10.3389/fgene.2020.00447

Pavy, N., Gagnon, F., Rigault, P., Blais, S., Deschênes, A., Boyle, B., et al. (2013). Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. *Mol. Ecol. Resour.* 13 (2), 324–336. doi:10.1111/1755-0998.12062

Perry, A., Wachowiak, W., Downing, A., Talbot, R., and Cavers, S. (2020). *Development of a single nucleotide polymorphism array for population genomic studies in four European pine species.* doi:10.1111/1755-0998.13223

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE* 7 (5), e37135. doi:10.1371/journal.pone.0037135

Plomion, C., Bartholomé, J., Lesur, I., Boury, C., Rodríguez-Quilón, I., Lagraulet, H., et al. (2015). *High-density SNP assay development for genetic analysis in maritime pine (*Pinus pinaster*).* doi:10.1111/1755-0998.12464

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

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

Raghavan, V., Kraft, L., Mesny, F., and Rigerte, L. (2022). A simple guide to *de novo* transcriptome assembly and annotation. *Brief. Bioin-form* 23, bbab563. bbab563. doi:10.1093/bib/bbab563

R Core Team (2023). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Resende, M. F. R., Munoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., et al. (2012). Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics* 190, 1503–1510. doi:10.1534/genetics.111.137026

Resende, R. T., Resende, M. D. V., Silva, F. F., Azevedo, C. F., Takahashi, E. K., Silva-Junior, O. B., et al. (2017). Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus. *New Phytol.* 213 (3), 1287–1300. doi:10.1111/nph.14266

Resquin, F., Navarro-Cerrillo, R. M., Carrasco-Letelier, L., and Cecilia Rachid Casnati (2019). Influence of contrasting stocking densities on the dynamics of above-ground biomass and wood density of *Eucalyptus benthamii, Eucalyptus dunnii*, and *Eucalyptus grandis* for bioenergy in Uruguay. *For. Ecol. Manag.* 438, 63–74. doi:10.1016/j.foreco.2019.02.007

Rochette, N. C., and Catchen, J. M. (2017). Deriving genotypes from RAD-seq short-read data using Stacks. *Nat. Protoc.* 12 (12), 2640–2659. doi:10.1038/nprot.2017.123

Rodrigues, J., Faix, O., and Pereira, H. (1998). Determination of lignin content of *Eucalyptus globulus* wood using FTIR spectroscopy. *Holzforschung* 52 (1), 46–50. doi:10.1515/hfsg.1998.52.1.46

Rutkoski, J. E., Poland, J., Jannink, J. L., and Sorrells, M. E. (2013). Imputation of unordered markers and the impact on genomic selection accuracy. *G3 Genes., Genomes, Genet.* 3 (3), 427–439. doi:10.1534/g3.112.005363

Sansaloni, C., Petroli, C., Jaccoud, D., Niediek, T., Gudermann, F., and Lütkemeyer, D. (2011). Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. *BMC Proc.* 5 (Suppl. 7), P54. doi:10.1186/1753-6561-5-S8-P54

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statistics* 6 (2), 461–464. doi:10.1214/aos/1176344136

Shi, T., Arnold, R. J., Kang, W., Duan, F., Qian, Y., Xie, H., et al. (2016). Genetic variation and gains for two generations of *Eucalyptus dunnii* in China. *Aust. For.* 79 (1), 15–24. doi:10.1080/00049158.2015.1086720

Silva, P. I. T., Silva-Junior, O. B., Resende, L. V., Sousa, V. A., Aguiar, A. V., and Grattapaglia, D. (2020). A 3K Axiom SNP array from a transcriptome-wide SNP resource sheds new light on the genetic diversity and structure of the iconic subtropical conifer tree *Araucaria angustifolia* (Bert.) Kuntze. *PLoS ONE* 15 (8), e0230404. doi:10.1371/journal.pone.0230404

Silva-Junior, O. B., Faria, D. A., and Grattapaglia, D. (2015). A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 Eucalyptus tree genomes across 12 species. *New Phytol.* 206 (4), 1527–1540. doi:10.1111/nph.13322

Silva-Junior, O. B., and Grattapaglia, D. (2015). Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of Eucalyptus grandis. *New Phytol.* 208 (3), 830–845. doi:10.1111/nph.13505

Stackpole, D. J., Vaillancourt, R. E., Alves, A., Rodrigues, J., and Potts, B. M. (2011). Genetic variation in the chemical components of *Eucalyptus globulus* wood. *Genes. Genomes Genet.* 1, 151–159. doi:10.1534/g3.111.000372

Strandén, I., and Garrick, D. J. (2009). Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J. Dairy Sci.* 92, 2971–2975. doi:10.3168/jds.2008-1929

Suontama, M., Klápště, J., Telfer, E., Graham, N., Stovold, T., Low, C., et al. (2019). Efficiency of genomic prediction across two *Eucalyptus nitens* seed orchards with different selection histories. *Heredity* 122 (3), 370–379. doi:10.1038/s41437-018-0119-5

Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B., and Ingvarsson, P. K. (2017). Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC Plant Biol.* 17, 110. doi:10.1186/s12870-017-1059-6

Tan, B., and Ingvarsson, P. K. (2022). Integrating genome-wide association mapping of additive and dominance genetic effects to improve genomic prediction accuracy in Eucalyptus. *Plant Genome* 15 (2), e20208. doi:10.1002/tpg2.20208

Telfer, E. J., Stovold, G. T., Li, Y., Silva-Junior, O. B., Grattapaglia, D. G., and Dungey, H. S. (2015). Parentage reconstruction in *Eucalyptus nitens* using SNPs and microsatellite markers: a comparative analysis of marker data power and robustness. *PLoS ONE* 10, e0130601. doi:10.1371/journal.pone.0130601

Thomas, D., Henson, M., Joe, B., Boyton, S., and Dickson, R. (2009). Review of growth and wood quality of plantation-grown *Eucalyptus dunnii* maiden. *Aust. For.* 72 (1), 3–11. doi:10.1080/00049158.2009.10676283

Thornhill, A. H., Crisp, M. D., Külheim, C., Lam, K. E., Nelson, L. A., Yeates, D. K., et al. (2019). A dated molecular perspective of eucalypt taxonomy, evolution and diversification. *Botany* 32 (1), 29–48. doi:10.1071/SB18015

Torkamaneh, D., Boyle, B., and Belzile, F. (2018). Efficient genome-wide genotyping strategies and data integration in crop plants. *Theor. Appl. Genet.* 131 (3), 499–511. doi:10.1007/s00122-018-3056-z

Trujano-Chavez, M. Z., Valerio-Hernández, J. E., López-Ordaz, R., and Ruíz-Flores, A. (2021). Frecuencia de alelo menor en predicción genómica para características de crecimiento en bovinos Suizo Europeo. *Rev. Bio Ciencias* 8, e1052. doi:10.15741/revbio.08.e1052

Ulaszewski, B., Meger, J., and Burczyk, J. (2021). Comparative analysis of SNP discovery and genotyping in *Fagus sylvatica* L. And *Quercus robur* L. Using RADseq, GBS, and ddRAD. *Methods. For.* 12, 222. doi:10.3390/f12020222

Valenzuela, C. E., Ballesta, P., Ahmar, S., Fiaz, S., Heidari, P., Maldonado, C., et al. (2021). Haplotype- and SNP-based GWAS for growth and wood quality traits in *Eucalyptus cladocalyx* trees under arid conditions. *Plants* 10, 148. doi:10.3390/plants10010148

Van Raden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91 (11), 4414–4423. doi:10.3168/jds.2007-0980

Varghese, M., Harwood, C. E., Bush, D. J., Baltunis, B., Kamalakannan, R., Suraj, P. G., et al. (2017). Growth and wood properties of natural provenances, local seed sources and clones of *Eucalyptus camaldulensis* in southern India: implications for breeding and deployment. *New*48, 67–82. doi:10.1007/s11056-016-9556-2

Varshney, R. K., Roorkiwal, M., and Sorrells, M. E. (2017). "Genomic selection for crop improvement: new molecular breeding strategies for crop improvement," in *Genomic selection for crop improvement: New molecular breeding strategies for crop improvement*, 1–258.

Vazquez, A. I., Bates, D. M., Rosa, G. J., Gianola, D., and Weigel, K. A. (2010). Technical note: an R package for fitting generalized linear mixed models in animal breeding. *J. Anim. Sci.* 88 (2), 497–504. Epub 2009 Oct 9. PMID: 19820058. doi:10.2527/jas.2009-1952

Villanueva, B., Pong-Wong, R., Fernández, J., and Toro, M. A. (2005). Benefits from marker-assisted selection under an additive polygenic genetic model. *J. Anim. Sci.* 83 (8), 1747–1752. doi:10.2527/2005.8381747x

Wimmer, V., Albrecht, T., Auinger, H. J., and Schön, C. C. (2012). synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics* 28 (15), 2086–2087. doi:10.1093/bioinformatics/bts335

Wright, B., Farquharson, K. A., McLennan, E. A., Belov, K., Hogg, C. J., and Grueber, C. E. (2019). From reference genomes to population genomics: comparing three reference-aligned reduced-representation sequencing pipelines in two wildlife species. *BMC Genomics* 20 (1), 453. doi:10.1186/s12864-019-5806-y

Yong, W. T. L., Ades, P. K., Runa, F. A., Bossinger, G., Sandhu, K. S., Potts, B. M., et al. (2021). Genome-wide association study of myrtle rust (*Austropuccinia psidii*) resistance in *Eucalyptus obliqua* (subgenus Eucalyptus). *Tree Genet. Genomes* 17, 31. doi:10.1007/s11295-021-01511-0

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38 (2), 203–208. doi:10.1038/ng1702

Zelener, N., Poltri, S. N. M., Bartoloni, N., López, C. R., and Hopp, H. E. (2005). Selection strategy for a seedling seed orchard design based on trait selection index and genomic analysis by molecular markers: a case study for Eucalyptus dunnii. *Tree Physiol.* 25 (11), 1457–1467. doi:10.1093/TREEPHYS/25.11.1457

Zhang, H., Yin, L., Wang, M., Yuan, X., and Liu, X. (2019). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Front. Genet.* 10, 189. doi:10.3389/fgene.2019.00189

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28 (24), 3326–3328. doi:10.1093/bioinformatics/bts606

Zhu, B., Zhang, J., Niu, H., Guan, L., Guo, P., Xu, L., et al. (2017). Effects of marker density and minor allele frequency on genomic prediction for growth traits in Chinese Simmental beef cattle. *J. Integr. Agric.* 16 (4), 911–920. ISSN 2095-3119. doi:10.1016/S2095-3119(16)61474-0