Final Report
Project NT048

NATIONAL INSTITUTE FOR FOREST PRODUCTS INNOVATION

Implementation of single-step genomic selection in eucalypts

2021

Launceston Centre
Funded by the Australian Government, Tasmanian Government & Industry Partners.
nifpi.org.au

NATIONAL INSTITUTE FOR
FOREST PRODUCTS INNOVATION
LAUNCESTON

Implementation of single-step genomic selection in eucalypts plantations

Prepared for

National Institute for Forest Products Innovation

Launceston

by

Dr Richard Kerr, Dr Josquin Tibbits, Dr Tony McRae,

Dr Ed Breen, Prof Hans Daetwyler

Publication: Implementation of single-step genomic selection in eucalypts Plantations

Project No: NIF111-1819 [NT048]

IMPORTANT NOTICE

ISBN: 978-1-922718-04-4

Researcher/s:

Dr Richard Kerr, Dr Tony McRae
Tree Breeding Australia Limited
PO BOX 1811, Mt Gambier SA

Dr Josquin Tibbits, Dr Ed Breen,
Prof Hans Daetwyler
Agriculture Victoria Research
AgriBio, 5 Ring Road,
Bundoora, Victoria

This work is supported by funding provided to Forest and Wood Products Australia (FWPA) to administer the **National Institute for Forest Products Innovation** program by the Australian Government Department of Agriculture, Water and Environment and the Tasmanian Government.

Australian Government
Department of Agriculture, Water and the Environment

Tasmanian Government

Forest and Wood Products Australia
Level 11, 10-16 Queen St, Melbourne, Victoria, 3000
T +61 3 9927 3200 F +61 3 9927 3288
E info@nifpi.org.au
W www.nifpi.org.au

Executive Summary

This project aimed to continue development and application of single-step genomic selection in the Australian hardwood breeding programs supported by Tree Breeding Australia (TBA) and its members for the benefit of the national hardwood forest industries. The project has enabled further adoption of genomics technology into tree breeding through the generation of new foundational data sets for the *Eucalyptus nitens* breeding program and the addition of new genotype and sequence data sets for the *Eucalyptus globulus* breeding program. These data sets now capture most of the variation segregating in these breeding programs and will underpin future steps in delivering routine genomic prediction including in the important step of designing low-cost assays for routine genotyping. The research has developed and tested methods for imputation and pedigree recovery and has applied these to non-eucalypt breeding pedigrees.

Importantly, the outputs of this project have already impacted operational breeding with the data created in this project joined with previous data sets and incorporated into TREEPLAN runs in 2021. This implementation realised the goal for joint use of data sets created through different technology platforms.

This project has achieved its main objectives of building foundational data sets and developing and testing methodologies further setting up the Australian forest industries for a non-disruptive implementation of genomic breeding into the national forest tree breeding programs. Implementation of single-step prediction in 2021 TREEPLAN evaluations indicates increasing impact on breeding decision making. This is being driven by the growing size of the genomics data sets which is improving EBV accuracy and increasingly enabling the inclusion of new unmeasured germplasm for early selection. Increased EBV accuracy and early selection impact both the rate of genetic gain and the efficiency of breeding operations and these efficiencies flow through to industry with better germplasm available for planting which increases plantation estate productivity, resilience, and profitability.

Table of Contents

Executive Summary i
Introduction 1
Methodology 2
Results 4
E. nitens sample collections 4
E. nitens discovery collection 4
E. nitens training collection 4
Sequence *E. nitens* discovery collection 8
Development of SNP Discovery Pipeline 9
Generating SNP sets for use in the project 9
Generating EGLOB HD SNP Set 1 9
Generating ENIT HD SNP Set 1 12
Generating EGLOB LD SNP Set 1 12
Development of imputation pipelines 13
Pipeline 1 13
Pipeline 2 14
Pipeline 3 15
Imputation testing 15
Imputation Tests 1 17
Imputation Tests 2 18
Imputation Tests 3 19
Investigation of the intersection of low- and high-density SNP sets 20
Intersection 1 20
Intersection 2 20
Intersection 3 21
Operational demonstration of genomic selection in *E. globulus* 22
Construction of a consolidated GRM 22
Single-step analysis in *E. globulus* with a consolidated GRM 26
Run 2021 TREEPLAN *E. globulus* data with 2020 version of GRM 32
Single-step analysis in *E. nitens* 34
Development of pedigree forensics pipelines 37
SEQUOIA 39
GRM-NRM comparison tool 39
Discussion 41
Conclusions 44
Recommendations 45
References 46
Appendix 1 47
Appendix 2 48

Introduction

The primary objective of genomic integration is to increase the profitability of forest growing activities in Australia by increasing the rate and deployment of genetic gain through fast-tracking the selection of new parents to reduce the generation interval in breeding programs for *E. globulus* and *E. nitens.* The operationalisation of single-step genomic prediction methods into existing tree breeding systems is key to achieving this. To operationalise single-step genomic prediction we put forward the following objectives:

1. Build foundational datasets upon which genomic selection can sustainably operate (reference genome built, medium-density sequencing of core pedigree, defining industry standard SNP sets).
2. Build core methodologies and workflows needed to implement genomic selection in an operational setting (algorithms, imputation methodologies).
3. Implement genomic selection at an operational scale in collaboration with industry partners (e.g. breeders, growers, deployers). Operationalising entails high-throughput genotyping, defining standard operating procedures and ensuring end-users have been well educated in the concepts and protocols regarding genomic prediction.

This NIFPI project has enabled TBA and its research partner, Agriculture Victoria Research (AVR) to begin and complete activities associated with these objectives. It is stressed that a complete operationalisation of single-step genomic prediction in both eucalypt species will span multiple research partitions.

Research under this NIFPI project has had a particular focus on objectives 1 and 2 with less attention directed to objective 3. Our goal has been to establish and expand base resource data sets and analysis pipelines that will be used in multiple downstream steps to underpin implementation of genomics into the TBA tree improvement programs.

We have focused on the analysis of the genomic resources, existing prior to, and generated during, this project, for the *Eucalyptus globulus* breeding program. While the genomic resources for this program are the most extensive for any of the TBA breeding programs, they are still short of the 15,000 genotypes required for a full base training set and it will remain a key objective to continue to expand these base resources in future projects. As the project progressed it was agreed between project partners that the analysis focus should shift to the data sets available for *E. globulus* as the most extensive data sets are available in this program and focus on these data would enable more extensive testing and simplify transfer between species in future projects. This change in focus was agreed from the initial milestone statements, which had a stronger focus on the analysis of *E. nitens* data sets.

Methodology

(Note: we are using the Methodology section to briefly outline the sequence of activities in the project. The Results section will then describe in more detail methodology and results specific to each of the activities listed below).

The project was implemented using an adaptive approach with continuous review of project needs and priorities. This approach led to changes in project activities and details of specific project implementations based on assessment of the resources available for implementation of specific steps. Some adjustments were direct results of changes arising from COVID-19 impacts which included delays in data generation in the middle of 2020 due to the Victorian lockdowns. Overall, COVID-19 impacts were restricted to changing within project priorities and did not impact the overall project delivery or success.

The first two activities aimed to create founder and discovery collection for *E. nitens.* These activities were based on the ongoing positive impact similar collections made in *E. globulus* in earlier projects have had and the aim was to develop comparable core data collections. The data sets were based on a detailed analysis of the pedigree and phenotypic records which drove germplasm collection activities. The germplasm collections were then assayed by whole genome sequence (WGS) methods to ensure a comprehensive survey and capture of the gene diversity in the Australian *E. nitens* breeding population. Prior to this project no accessible WGS data existed for *E. nitens.* These foundational data sets will have a long useful life and will be used in multiple activities in this and future projects.

1. ***E. nitens* sample Collections** - Identify and sample foliage from the combined national *E. nitens* breeding population (comprising populations belonging to Forico and Sustainable Timber Tasmania).
    a. Foliage for the Discovery (Founder) Collection
    b. Foliage for the Training Collection
2. **Sequence *E. nitens* discovery collection -** Generation of ~8-10x raw sequencing coverage for 384 individuals.

Our next activity was to design a SNP discovery pipeline that could be applied across all species (not just eucalypts).

3. **Development of SNP Discovery Pipeline**

Following the development of the SNP Discovery Pipeline we were then able to generate high-density SNP sets in both *E. globulus* and *E. nitens*. Additionally, a commercial SNP chip was used to generate a low-density SNP set in *E. globulus*. The low-density set provided us with the necessary data to test imputation pipelines and **G** matrix consolidation.

4. **Generating SNP sets for use in the project**
    a. Generating the *E. globulus* high density (HD) set: EGLOB HD SNP Set 1
    b. Generating the *E. nitens* high density (HD) set: ENIT HD SNP Set 1
    c. Generating the *E. globulus* low-density (LD) set: EGLOB LD SNP Set 1

When processing data at the level of SNP loci, it will be necessary to infer data that is missing or is error prone. Imputation refers to the process of inferring the missing data or replacing error prone data with substituted values of higher accuracy. A major component of this research partition has been to test various imputation methods, make recommendations and to build imputation pipeline prototypes, which can be adopted for routine use by industry. Three imputation pipelines developed and tested.

5. **Development and testing of imputation pipelines**
    a. Pipeline 1 – “filling-in” within a specific assay
    b. Pipeline 2 – imputing from low coverage WGS data to high-density SNP sets
    c. Pipeline 3 – imputing from low-density SNP chips to high-density SNP sets

In our initial planning we indicated that we would demonstrate imputation from a low-density, commercial grade SNP chip to a high-density SNP set (e.g., imputing calls made from the EGLOB LD SNP Set 1 to calls on the EGLOB HD SNP Set 1). And furthermore, demonstrate the construction of a **G** matrix incorporating assay data from both low- and high-density platforms. As the project progressed, it became apparent that with the genomic resources currently available this demonstration would not be possible. The imputation testing indicated much larger reference panels would be needed to ensure successful imputation from low- to high-density SNP sets. Obtaining reference panels of appropriate size was outside the scope of this project, in any species, yet alone in *E. nitens* (which was originally going to be the target species for this demonstration). The project team decided on an alternative strategy, which was to construct a **G** matrix by incorporating data from the EGLOB LD SNP Set 1 and all other SNP assay data generated using either low- or high pass whole genome sequencing. Rather than impute up to a high-density set, we obtained the intersection of SNP between those featured on EGLOB LD SNP Set 1 and those SNP discovered *de novo* from our WGS work. This combined SNP set was used in TREEPLAN runs in 2021.

6. **Investigation of the intersection of low- and high-density SNP sets**
    a. Intersection level 1 – match Euc72K probes to RaGOO assembly (the latest up-to-date assembly available to AVR)
    b. Intersection level 2 – SNP passing level 1 that are included in EGLOB HD set 1
    c. Intersection level 3 – Intersect *de novo* WGS SNP with Euc72K SNP using *E. grandis* genome assembly Version 2.0 (both WGS SNP and Euc72K reference this assembly)

The next activity was an operational demonstration of genomic selection in *E. globulus*. The intersection of *de novo* WGS SNP with Euc72K SNP resulted in 17,103 SNP that were used in the construction of a GRM for use in a TREEPLAN analysis. All available assayed individuals including approximately 2,000 recently assayed juveniles (new progeny that have not been measured), were included in the GRM.

7. **Operational demonstration of genomic selection in *E. globulus***
    a. Construction of a consolidated GRM
    b. Single-step analyses in *E. globulus* with a consolidated GRM

Though we did not demonstrate imputation and consolidation of **G** matrices in *E. nitens* we were able to demonstrate the impact of more genomic data in this species. This led to the next activity.

8. **Single-step analyses in *E. nitens***

The final activity completed in this NIFPI project was to build and test a pedigree forensics pipeline. Part of this development was to investigate the appropriateness of a public domain software package called SEQUOIA for undertaking pedigree forensics using genotype call data on individual SNP. Another aspect of the development was to write a software tool that would undertake pedigree forensics by comparing relationship coefficients obtained in the **G** and **A** matrices. In-silico simulation of scenarios and the testing of software was a feature of this investigation.

9. **Development of a pedigree forensics pipeline**
    a. Build and test pipeline using in-silico simulation
    b. Prototype pipeline with real data

Results

E. nitens sample collections

E. nitens discovery collection

Individuals in the discovery collection were identified by computing the "contribution matrix", which is a lower triangular matrix containing the fraction of genes that individuals (in this case the founders) passed to descendants (in this case the named, 2nd generation individuals). In most breeding programs a few founders contribute most of the genes to the later generations. This is certainly true in *E. nitens*. One hundred and forty-two founders (native mothers) have contributed 97% of the genes to second-generation, named genotypes. Our strategy was to sample as many as possible of the direct descendants of these 142 founders, as the native mothers themselves have since been destroyed. Unfortunately, most of the 1st generation breeding trials have also been destroyed and the best hope of recovering foliage is from the older breeding facilities where clones of 1st generation selections remain archived.

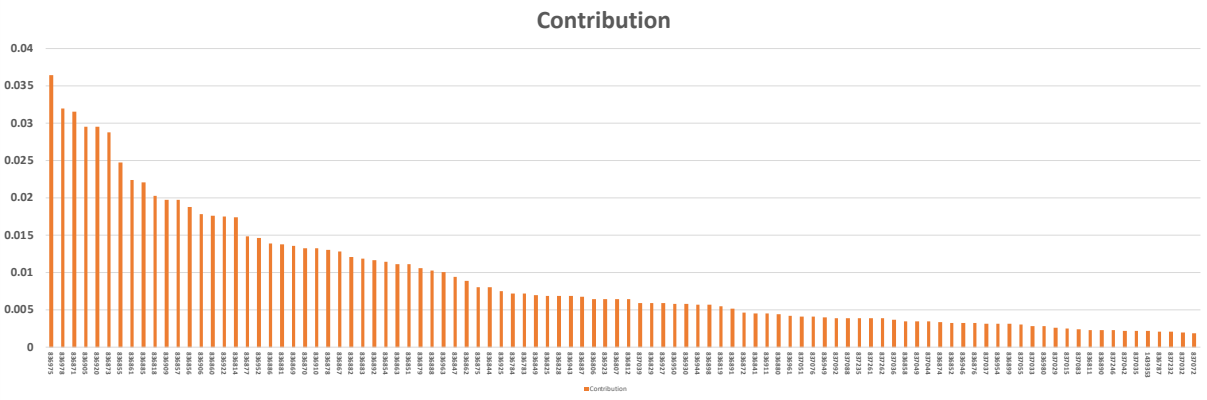


Figure 1 Important founders, sorted by their fractional contributions to named, 2nd generation genotypes in the national E. nitens pedigree.

Lists of 173 and 27 descendants were distributed to Forico and Sustainable Timber Tasmania (STT), respectively. Forico were able to sample 160 from their list and SST were able to sample 23 from theirs. Both organisations were asked to provide foliage from genotypes they wished to include in the discovery collection. Forico provided foliage for a further 118 genotypes and SST provided foliage for a further 50 genotypes.

There were 351 genotypes sampled in total by partner organisations for the *E. nitens* discovery collection. The samples were sent to AVR facilities in Bundoora for DNA extraction, library creation and sequencing. DNA libraries were generated for 339 samples. It is typical for the library creation step to fail in a small minority of cases. It was decided to include 45 recent *E. globulus* parents in the Collection to make up the numbers for a complete plate (384 wells).

E. nitens training collection

A training collection will need to comprise a substantial number of individuals (two to three thousand) that have been measured phenotypically, and DNA assayed. There is already a sizeable number of measured individuals assayed via the Gondwana assay (1381). However, the Gondwana assay is a very low-density assay (less than 3,000 SNP) and the existing number of measured and assayed individuals is perhaps adequate for training this SNP set. We have set in place a plan for obtaining a training data set for the eventual industry standard, high-density SNP set, once it has been finalised. The plan entails

- retrospective genotyping (i.e. sampling older, first generation progeny trials)
- resample progeny already assayed with the Gondwana SNP set, for assaying with the future high-density SNP set
- assume existing training individuals (albeit for a very small SNP set) will also be relevant for training a much larger SNP set

The latter individuals will only be useful for training a high-density SNP set if imputation from low- to high-density SNP sets is successful. Research completed to-date in *E. globulus* suggests that a sizeable reference population will be required for successful low- to high-density imputation.

Table 1 shows the distribution of the majority of the 1381 *E. nitens* individuals assayed using the Gondwana SNP set in terms of trials they were measured in and which selection criteria the measurements map to. Selection criteria with no measurements mapped to them have been highlighted in grey. It can be seen there are currently no assayed individuals measured in dry site types (all ages) or high site type (late age), nor individuals measured for some of the wood quality traits such as BD_disc (all ages) and BD_core (early age) and Cellulose_core (latter ages). In addition, there is nothing measured for: Stem straightness; Branch quality; KPY; Pilodyn; and Acoustic Velocity. Our retrospective genotyping strategy was to target trials from which we could obtain samples from individuals with measurements on “missing” traits.

Table 1 Distribution of measured progeny assayed with Gondwana chip by trial and measured phenotypic trait.

	06011 Hudlers	01031 Cold Hardy Trial Middlesex Spur 9	97045 Open Pollinated Trial Blythe Rd	98041 2nd Gen Selection West Ridgley	95051 2nd Gen Selection Jacksons East	81043 Hampshire Extension	97071 Clonal Trial Kelatier	97041 OP Trial Loudwater Rd	86015 West Ridgely NSW	95052 2nd Gen Selection Guide Rd	81041 Hampshire Seed Orchard	Number observ-ations
trial	398	269	206	62	50	33	30	29	26	23	20	1146
ASFrost_Cold_01		269										269
ASFrost_Cold_02_04												0
BD_core_04_07												0
BD_core_08_15	396	269		44	35					15		759
BD_core_16_29			198									198
BD_disc_04_07									1			1
BD_disc_08_15												0
BD_disc_16_29												0
Cellulose_Core_04_07												0
Cellulose_Core_08_15	397	268		44	35					15		759
Cellulose_Core_16_29			204									204
DBH_Dry_02_03												0
DBH_Dry_04_07												0
DBH_Dry_08_15												0
DBH_Cold_02_03			206									206
DBH_Cold_04_07		269	206									475
DBH_Cold_08_15		269	204									473
DBH_Cold_16_29		269	203									472
DBH_High_02_03				62					26			88
DBH_High_04_07					50		30		26	22		128
DBH_High_08_15				62	50		30			23		165
DBH_High_16_29												0
DBH_Normal_02_03								29				29
DBH_Normal_04_07	398					33					20	451
DBH_Normal_08_15	398							29				427
DBH_Normal_16_29												0
Ht_Dry_01												0
Ht_Dry_02_03												0
Ht_Cold_01												0
Ht_Cold_02_03		179										179
Ht_Cold_04_07		269										269
Ht_High_01												0
Ht_High_02_03												0
Ht_Normal_01						33						33
	398	269	206	62	50	33	30	29	26	23	20	1146

The majority of 1st and 2nd generation trials in *E. nitens* have been felled or destroyed, which limits our capacity to retrospectively genotype. However, the STT managed Meunna 1st generation trial is still active, and it makes sense to sample from this trial as there are potentially high value progeny in this trial. Table 2 shows the number of high value progeny against selection criteria for this trial. A high value progeny is a progeny with multiple observations. One high value progeny is selected per family. There are over 400 potential families from which to sample one progeny. Sampling these progenies will help fill in the missing gaps in the training collection. These individuals have been measured for the missing traits such as stem straightness, acoustic velocity, and latter age growth in a high site type.

Table 2 The numbers of high value progeny measured for various selection criteria in Meunna base population trial.

	RP25208 Meunna base pop trial
ASFrost_Cold_01	
BD_core_08_15	102
BD_disc_04_07	
Cellulose_Core_08_15	101
Cellulose_Core_16_29	
DBH_Dry_02_03	
DBH_Dry_04_07	
DBH_Cold_02_03	
DBH_Cold_04_07	
DBH_Cold_08_15	
DBH_Cold_16_29	
DBH_High_02_03	
DBH_High_04_07	416
DBH_High_08_15	416
DBH_High_16_29	413
DBH_Normal_02_03	
DBH_Normal_04_07	
DBH_Normal_08_15	
DBH_Normal_16_29	
Ht_Dry_01	
Ht_Dry_02_03	
Ht_Cold_01	
Ht_Cold_02_03	
Ht_Cold_04_07	
Ht_High_01	415
Ht_Normal_02_03	
KPY_disc_04_07	
KPY_core_08_15	
KPY_core_16_29	
Pilodyn_04_07	274
Acvel_16_29	339
BRQS_04_07	
BRQS_08_15	
STEMST_04_07	
STEMST_08_15	
ZSTEMST_04_07	416
ZSTEMST_08_15	416
	416

Our collaborating industry partner Forico has an ongoing marker assisted selection (MAS) project with Gondwana Genomics. As part of that project, they have assayed up to 11,000 juveniles (new progeny that have not been assessed). Table 3 shows the numbers of juveniles that have been recently measured or are scheduled to be measured. The strategy is to re-sample approximately 2,000 of these individuals. The 2,000 individuals will eventually be assayed for the high-density SNP set once it has been finalised. Having a sizeable number of individuals assayed for both panels (the Gondwana panel and the high-density SNP set designed by AVR) will provide the necessary data to test and implement imputation from the low-density Gondwana panel to the high-density SNP set.

Table 3 Juveniles in Forico's E. nitens tree improvement program, which have been assayed and are awaiting assessment.

Trial	Count	To be assessed
17012 Basils Spur 5 *E. nitens* Progeny Trial	3682	By 2022
13012 Parrawe 11 Progeny Trial	1816	Before June 30 2021
17011 Kingsclere Spur 1 Progeny Trial	1434	Has been assessed (yet to be entered)
19011 *E. nitens* Progeny Trial, Rogetta rd, Hampshire	961	By 2023

In summary, the training collection will be provided by the following 3 sampling efforts

1. The LINK sample group achieved by resampling 2000 progeny in the recent "MAS" trials (17012, 13012, 17011). These individuals will be assayed with the eventual high-density SNP set and have already been assayed with the Gondwana SNP set. They provide us with the data to enable imputation from a low-density SNP set to a high-density SNP set.
2. The EXISTING group which encompasses approximately 1417 progeny that have been assayed for the Gondwana SNP set and have been measured and are currently mapped into TREEPLAN systems and potentially the remaining ~6,000 progeny that are yet to be measured. The hope is that genotypes for the high-density SNP set can be obtained via imputation.
3. The RETROSPECTIVE group encompassing 420 high value progeny from the Meunna base population that will provide the "missing" phenotypic data.

It is likely that gaps in the phenotypic data will remain. It is unlikely there will ever be assayed individuals with growth data observed on a dry site type. From our discussions with project personnel associated with *E. nitens* breeding (Kelsey Joyce and Dean Williams) dry site types no longer have any relevance. That is, it is unlikely material will ever be deployed to a dry site type.

Sequence *E. nitens* discovery collection

The whole genome sequencing (WGS) was completed using a combination of MiSEQ (PE 300+300) and NovoSEQ (PE 150+150) sequencing. The 339 *E. nitens* samples in the discovery collection yielded an average raw read coverage of 12.09x, with a range between 0.03x and 279x. While the coverage range is wider than generally seen most of the samples had coverage between 6x and 12x with only a small number (12) with coverage below 4x. Having a range of coverage is normal from the generation of sequence data. For the *E. globulus* parent selections the average coverage was 12.63x (range 2.8x to 103x) with 5 samples having coverage less than 4x. Coverage results for each sample is provided in Appendix 1.

Raw sequencing data in the form of files of zip archived fastq format DNA sequence reads are currently stored on the AVR BASC computer system. These have been generated and flagged for a storage period of four years. All raw fastq format sequencing data is available for transfer to TBA at their request via the AVR SFTP server. TBA are currently reviewing their genomic data storage needs and will soon come to a decision on how best to archive such data.

Development of SNP Discovery Pipeline

SNP discovery requires pipelines that can analyse *de novo* whole genome sequence. The main pipeline implemented and developed in this work is based on the GATK workflows. GATK stands for Genome Analysis Tool-Kit and is a collection of command-line tools for analysing high-throughput sequencing data with a primary focus on variant discovery. The tools can be used individually or chained together into complete workflows. The final pipelines implemented in this work followed as closely as possible the developed GATK end-to-end workflows, called GATK Best Practices and were developed using the available *E. globulus* genomic resources. We were able to successfully generate a high-density SNP set in *E. globulus*, which we refer to as HD SNP Set 1, using this pipeline. An outline of the pipeline for creation of HD SNP Set 1 is shown in Figure 2.

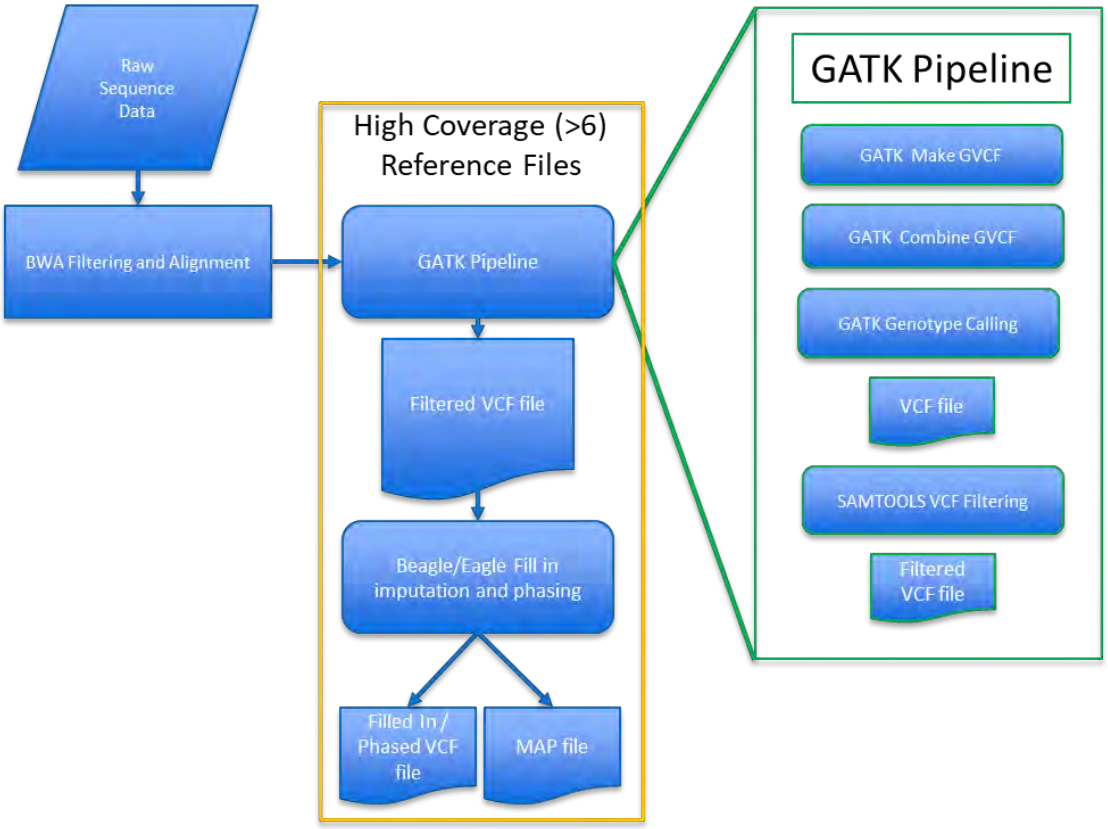


Figure 2 Pipeline for SNP discovery to create HD SNP Set1. Variant discovery followed GATK's "best practices" workflow for germline short variant discovery, by running CombineGVCFs to consolidate gVCFs and GenotypeGVCFs to perform joint genotyping. Due to the current lack of a E. globulus dbSNP, GATK4 BQSR was not used.

Generating SNP sets for use in the project

Generating EGLOB HD SNP Set 1

The *de novo* SNP discovery used a set of 963 whole genome sequence data sets from 963 individual *E. globulus* trees where the average genome sequence coverage was greater than 6. The raw variant call format (VCF) data file was filtered using SAMtools and BCFtools (Danecek *et al*. 2021) to a set of bi-allelic markers polymorphic at a minor allele frequency of 1% and for sites with less than 10% missing data. Fill in imputation and phasing of this data set was then done using Beagle (Beagle 5.1; Browning *et al.* 2018; Browning and Browning 2007) for imputation and Eagle (Eagle 2.4.1; Loh *et al.* 2016) for phasing. The sequence data input was generated in previous projects and was aligned

against the interim RaGOO (Alonge *et al.* 2019) chromosome scaffolded reference genome being assembled by AVR with the complete chloroplast and mitochondrial genomes added (EX46-HAC-Mod1-P.R2.RaGOO_EX46CP-EX46MT). The resulting data set is referred to as HD SNP Set1. A summary of SNP discovery is provided in Table 4. The inter SNP distance provides an indication of the expected SNP density in any region with a summary bar plot of inter SNP distances shown in Figure 3. Inter SNP distances for *E. globulus* are very short with the majority of SNP being found within 100-2000 bp of each other. This high SNP density and low expected linkage disequilibrium (LD) has significant implications for how well we would expect imputation to work. Native *E.globulus* has a high effective population size and our sampled individuals hopefully represent adequately the diversity of the species. Thus, our expectation is for linkage disequilibrium (non-random association between alleles at different loci) to be low.

Table 4 Summary of SNP discovery for HD SNP Set1 based on GATK best practice variant discovery pipeline and whole genome sequence data from 963 discovery collection (E. globulus) samples with a mean sequence coverage of greater than six.

Chromosome	Biallelic SNP (MAF > 1%, MISS < 10%)	Length (bp)	SNP/kbp
Chr1	538,669	42,805,116	13
Chr2	676,961	54,643,196	12
Chr3	748,284	68,413,973	11
Chr4	480,014	41,230,448	12
Chr5	736,052	63,447,923	12
Chr6	628,416	54,369,819	12
Chr7	600,141	55,219,908	11
Chr8	907,963	73,639,819	12
Chr9	480,284	39,381,439	12
Chr10	518,270	42,238,892	12
Chr11	553,622	47,166,220	12
TOTAL	**6,868,676**	**582,556,479**	**12**

This SNP set is the most comprehensive variant data set discovered to date for *E. globulus.* While it has been extensively filtered by imposing MAF and missingness thresholds, it is possible that variant positions with complex inheritance patterns (e.g. SNP variants in regions with underlying INDEL polymorphism) and some artefacts remain in the data set. Some of these will become apparent as they will prove difficult to impute or fail in Mendelian segregation tests. Despite this shortcoming, the SNP set represents a significant improvement on the approximately 800K SNP being used in previous studies/projects and is now the best available data set for this species that we know of. HD SNP Set 1 was completed using fill in imputation pipeline using Beagle and Eagle. This fill in imputation should be applied routinely to high coverage WGS sequencing data generated on parents and key breeding individuals to continue to expand HD SNP Set1. This SNP set defines an industry standard SNP set. A principal component analysis (PCA) of the SNP genotypes is shown in Figure 4 and shows that the HD SNP Set1 captures the expected diversity of the breeding program. Provenance structure is particularly evident in the first-generation individuals (beige coloured points). There is a strong indication of four clusters, perhaps representing major evolutionary events such as the migration of individuals into the Otways, King Island and Western Tasmania, followed by migration to Southern and Eastern Tasmanian, then finally to the Furneax group.

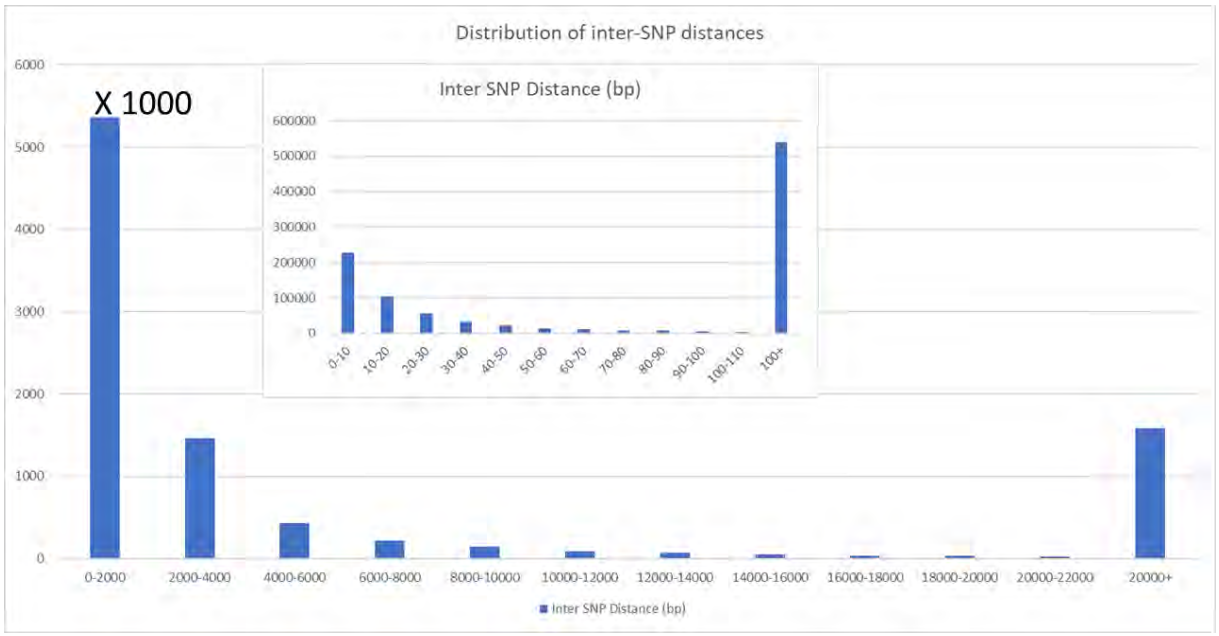


Figure 3 Bar plot showing the distribution of inter SNP distances for variants discovered using GATK pipeline from 963 Eucalyptus globulus trees with greater than 6x average coverage. Inset bar plot shows breakdown of the 0-2000 bin for inter SNP distances between 0-100 + bp showing that the vast majority of inter SNP distances are between 100 and 2000 bp. Very few genome regions exist with inter SNP distances greater than 10,000 bp.

Figure 4 PCA based on LD SNP Set1 showing coloration by generation.

Generating ENIT HD SNP Set 1

The GATK best practice variant discovery pipeline was applied to the 339 *E. nitens* samples sequenced. A total of 32 million biallelic SNP were discovered after applying filtering with MAF > 1% and a missing call rate < 10%. This high number of discovered SNP is partly due to the excellent average read depth that was obtained in the sequencing step.

Table 5 Number of discovered SNP per chromosome using the 339 E. nitens samples

Chromosome	Biallelic SNP (MAF > 1%, MISS < 10%)
Chr1	2,134,081
Chr2	3,025,165
Chr3	4,649,007
Chr4	2,161,302
Chr5	4,183,287
Chr6	2,636,023
Chr7	3,416,312
Chr8	4,186,791
Chr9	1,981,997
Chr10	2,035,209
Chr11	2,407,509
TOTAL	**32,816,683**

Generating EGLOB LD SNP Set 1

A low-density (LD) SNP set was obtained by assaying a sample of the *E. globulus* breeding population with the Euc72K Chip. In this sense we are not discovering SNP *de novo* but assaying individuals with an existing, commercial SNP set, which resulted from SNP discovery work carried out by other research groups.

The samples comprised parents not yet assayed and 3rd generation progeny mainly from Green Triangle trial sites. The number of samples in this collection was restricted to 1056, which amounts to 11 plates of 96.

The collection was augmented with 480 foliage samples (5 plates of 96) that had been previously collected in previous projects and assayed using high to medium coverage WGS.

The consignment of samples was sent to Thermo Fisher for processing with the Euc72K Chip (68,055 Eucalyptus and 4,147 Corymbia Features). A file was received back from Thermo Fisher indicating 1508 samples had passed Dish Quality Control (DQC), sample quality control call rate (QC CR) and Plate quality control. A total of 50,865 markers were recommended as “Best and Recommended”.

The raw data file was then filtered using BCFtools (Danecek, 2021) to a set of bi-allelic markers polymorphic at a minor allele frequency of 1% and with less than 10% missing data. As a result of this step the number of useable SNP was reduced to approximately 33,000.

Development of imputation pipelines

Imputation is the process of replacing missing or error prone data with substituted values that are of higher accuracy. In genomics studies it generally refers to methods that infer unobserved SNP locus genotypes using a reference population which has no missing data, and on which accurate calls have been obtained. Imputation methodologies are broad in scope and cover:

- to fill in the randomly missing SNP genotypes following the application of an assay
- to infer non-assayed SNP genotypes for samples assayed for a subset of sites
- to improve, through inference, genotypes at sites with some, but inadequate, information such as is obtained by low and middle coverage sequencing data

The quality of imputed datasets is largely dependent on the software used, as well as the specifics of the reference populations chosen and the underlying patterns of variation in the population under study.

In the tree breeding programs of TBA there are several points of application for imputation and this work seeks to specifically investigate and implement specific solutions to the major application points. The three major application points of imputation are:

- a simple “filling-in” within a specific assay where one attempts to fill in the missing calls (**Pipeline 1**)
- for imputation from Skim Whole Genome Sequencing (SWGS) data to a common high-density target set of SNP (**Pipeline 2**)
- imputation from a low-density chip assay (e.g. the Euc72K chip) to a common high density target set of SNP (**Pipeline 3**)

These three applications present different underlying problems and different pipelines will be required in application.

While the goal is to develop imputation pipelines that can be applied across all breeding programs, species specific application of imputation in each breeding program will require the development of the appropriate background datasets and tests to validate efficacy for application. The current NIFPI project has allowed us to specifically focus on application in *E. globulus* using the genomic resources available for that breeding program, which at this time are the most extensive for any of the TBA breeding programs. From our discussions with Agriculture Victoria Research (AVR), it was decided that, for the foreseeable future, imputation pipelines are best handled using AVR computing resources. Transfer of these pipelines to TBA computing infrastructure at this stage would be inefficient and counterproductive. All outputs and scripts are available for TBA to use as project outputs.

Pipeline 1

Pipeline 1 is essentially for “filling-in” and phasing. With most assays not every marker will be successfully called across and within individual samples. Fill-in imputation attempts to recover the missing calls. Phasing occurs after fill-in imputation and determines which alleles were co-inherited on the same chromosome. The pipeline consists of scripts that apply the public domain software

- Beagle (Beagle 5.1; Browning *et al.* 2018; Browning and Browning 2007), for the fill-in imputation step, and
- Eagle (Eagle 2.4.1; Loh *et al.* 2016), for the phasing step.

Pipeline 1 was used in the final stage of the process to derive EGLOB HD SNP Set 1 (see Figure 2) and in deriving EGLOB LD SNP Set 1 post filtering with BCFTools.

Pipeline 1 should be applied routinely to high coverage WGS data generated on parents and key breeding individuals to continue to expand EGLOB HD SNP Set1; and routinely to any low-density SNP chips, which invariably have small amounts of missing data.

Pipeline 2

Pipeline 2 is for the imputation of skim whole genome sequencing SWGS; (in this study defined as <6x coverage, but generally referring to <1x coverage data sets). Due to the low read depth associated with SWGS, there are few loci with enough reads to accurately call genotypes and the set of loci called generally differs from individual to individual. SWGS genotypes can be improved by imputing missing genotypes and improving genotype accuracy of loci with insufficient reads. A reference panel of phased high coverage (high accuracy) haplotypes are used to assist in the calling of sites that have not been, or inadequately, read.

The pipeline consists of scripts that implement the software package GLIMPSE (Glimpse V1.1.0; Rubinacci *et al.* 2021), which was designed specifically for the imputation of low-coverage sequencing data sets.

Pipeline 2 was used to derive EGLOB HD SNP Set 2 (genotype call data on > 6 million SNP for 4515 samples). EGLOB HD SNP Set 1 was used as the reference panel. A diagrammatic summary of the process to create EGLOB HD SNP Set 2 is shown in Figure 5.

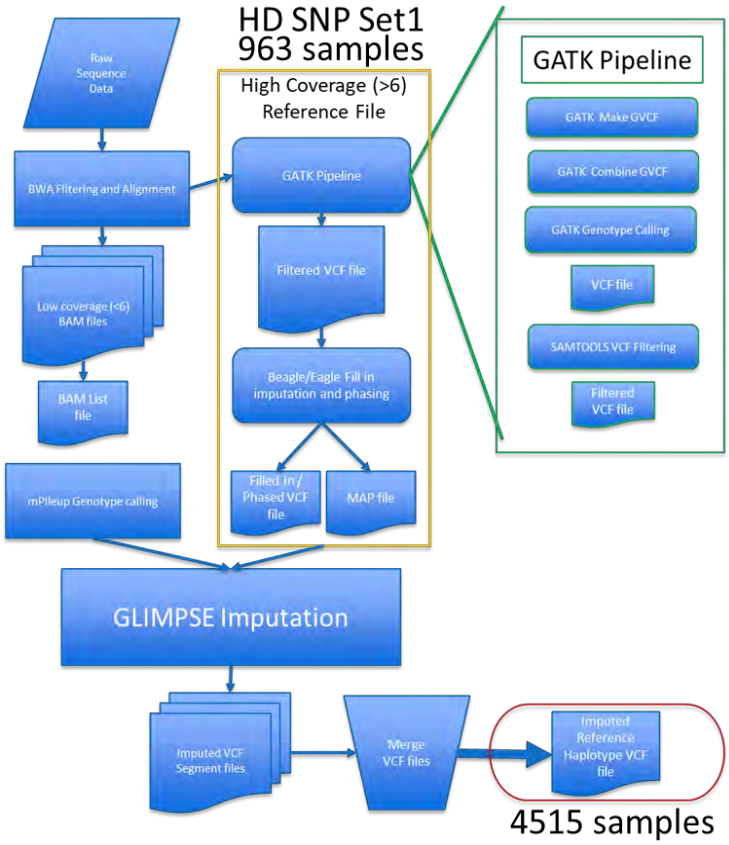


Figure 5 Pipeline for creation of HD SNP Set2.

The process of application should be to first expand EGLOB HD SNP Set 1 to incorporate data from new, high coverage samples, such as parents. An aspect of TBA's genomic selection strategy is to routinely assay new parents using medium to high coverage WGS and then to apply Pipeline 2 to impute any new low coverage WGS data. The GATK data bases for the high-coverage samples can be appended with new data for new samples and are therefore an extensible resource created as an output of this project.

Pipeline 3

Pipeline 3 is a more traditional application imputing low density chip assay genotype data to a high-density target. The public domain software "Minimac" is the primary engine in this pipeline (Das et al, 2016; Fuchsberger et al, 2015; Howie et al. 2012). The Minimac approach is to employ a reference panel of more densely typed individuals, such *E. globulus* parents assayed using high coverage WGS. The reference panel is then used to find haplotype segments that are shared among the target individuals which have been processed with the low density, commercial assay (for example, the Euc72K chip). Minimac operates on the premise of restricting the search for matching haplotypes to a small set of likely haplotype configurations, as determined by the reference panel. Minimac requires the target samples to be pre-phased. Minimac has undergone several revisions and the current version which we are using is Minimac3.

Imputation testing

Pipeline 3 will underpin our effort to construct genomic relationship matrices using DNA assay data from different marker panels. It was important for us to determine how well we could impute up to a high-density marker set, given the genomic resources currently available to us, in terms of reference and target sets. Therefore, we implemented several investigations of the Minimac approach used in Pipeline 3. The approach entailed applying a sub-sampling method to simulate target samples, and target SNP sets, of various sizes, and then impute back testing various reference data set sizes. We sub-sampled both chip and WGS derived data.

An overview of the imputation testing is given in Figure 6.

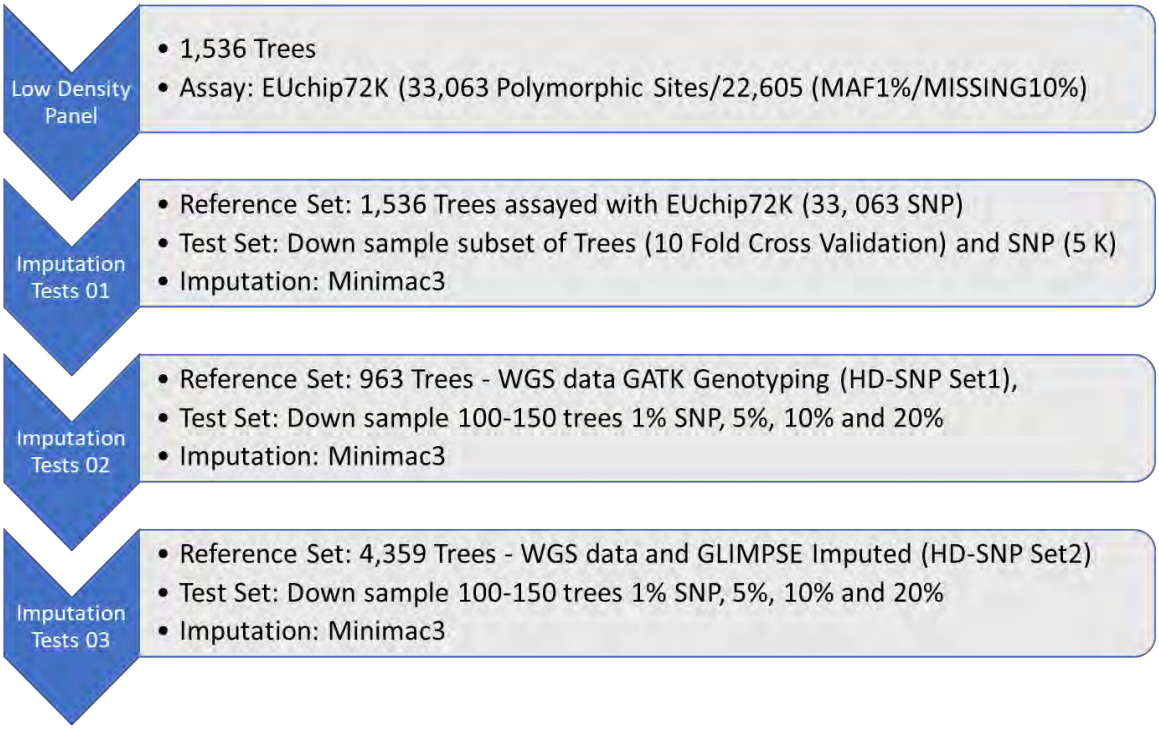


Figure 6 Overview of imputation tests run to investigate imputation options. Tests 01 used the EUchip72K data set to develop the pipeline steps, whereas steps 02 and 03 investigate the potential imputation accuracies that are likely given various SNP subsets.

For all imputation tests the accuracy of imputation was calculated as the Pearson correlation and concordance of imputed genotypes from each respective iteration and raw genotypes from the relevant sub-sampled data set. Concordance was calculated as the proportion of imputed genotypes matching full data set genotypes.

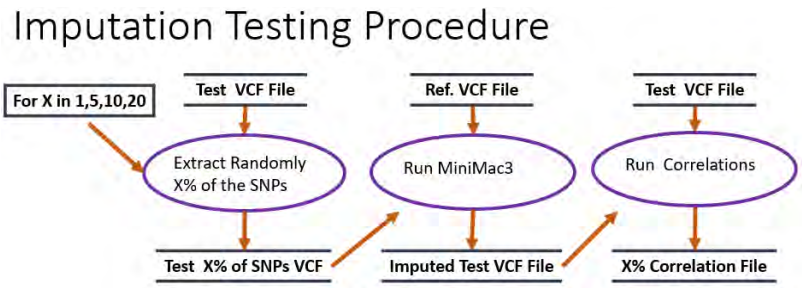
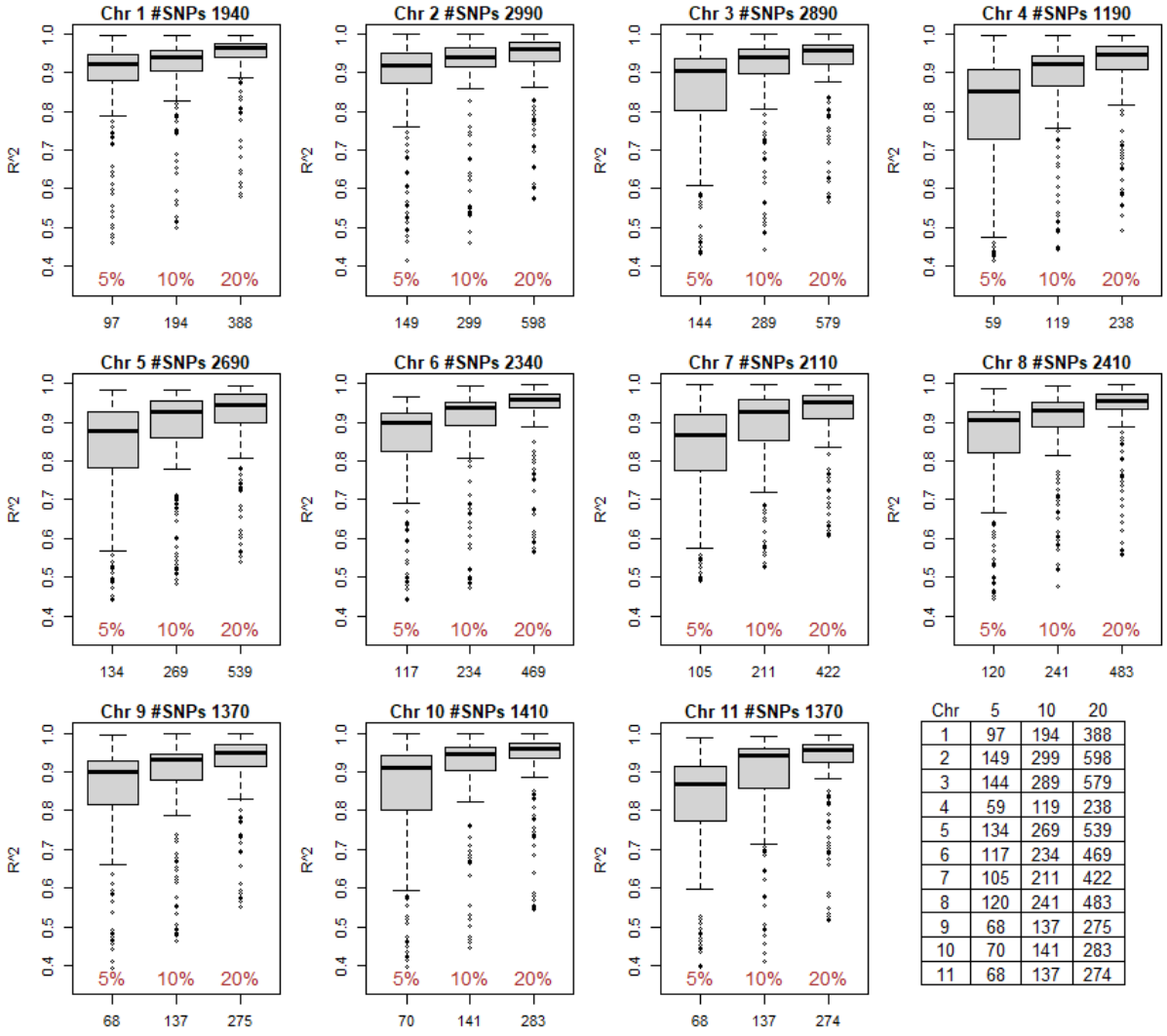


Figure 7 Imputation and testing.

The imputation and correlation procedure is schematically given in Figure 7, whereby a Test VCF File containing M samples and N SNPs, was reduced to 5 Test X% of SNP VCF files, which contained all samples, but with SNP panels reduced to just 1%, 5%, 10% and 20% of the original number of SNPs respectively. These panels were then imputed back up to their original number of SNPs, using Pipeline 3, to produce an Imputed Test CVF File. Then a sample wise Pearson's correlation and concordance between the Imputed Test CVF File and Test VCF File was recorded.

Imputation Tests 1

The first set of imputation tests (Imputation Tests 01) use EGLOB LD SNP Set 1 as the data set and were primarily aimed at testing Pipeline 3 using Minimac3, rather than testing the efficiency of imputation per se. The data set (EGLOB LD SNP Set 1) is not fully representative of the breeding program diversity, with the bulk of samples derived from a small number of generation 3 families. Results of the imputation, by chromosome, are shown in Figure 8, with imputation accuracies above 0.9 per sample seen with even the lowest sub-sample of 5% (only a hundred or so markers). The limited diversity in the data probably accounts for the high accuracy seen in these results and is unlikely to be reflective of what will be possible in the HD sets. Larger population samples are used in generating HD sets and faster LD breakdown is likely to be observed which will impact imputation accuracy. In addition, the reference panels for use in Pipeline 3 with HD sets have lower, overall genotype accuracies, due to their creation from WGS rather than from chip assays.

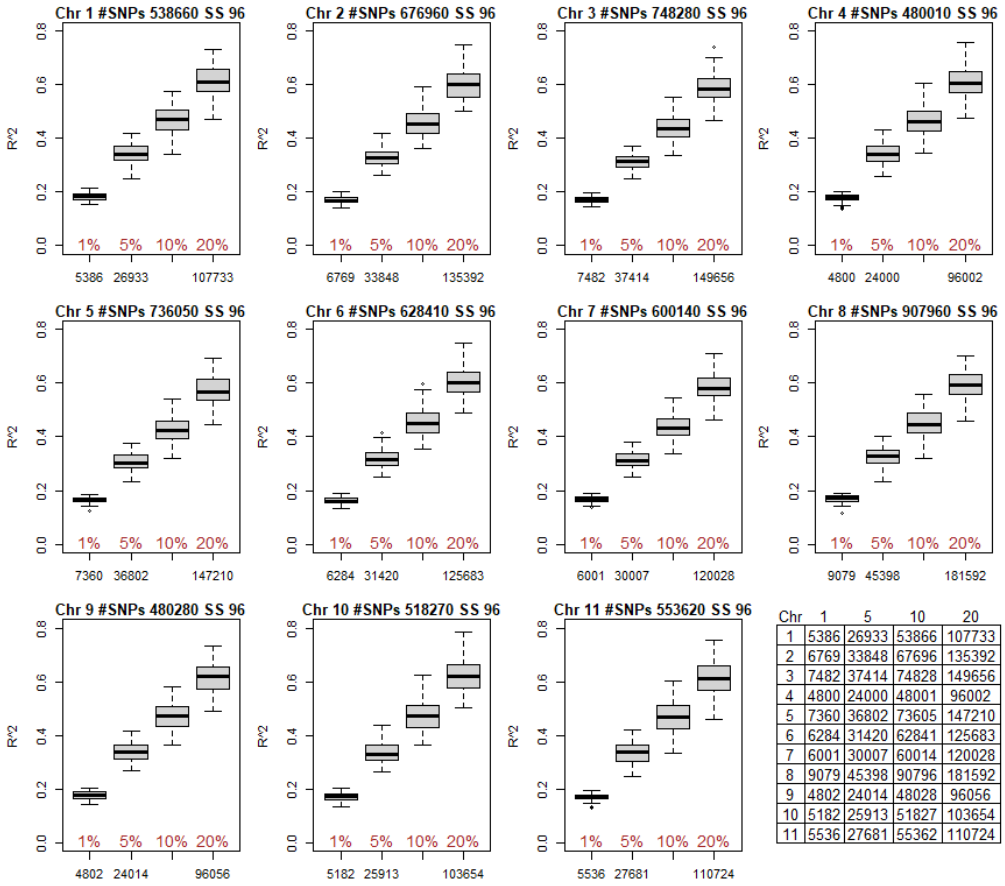


Chr	5	10	20
1	97	194	388
2	149	299	598
3	144	289	579
4	59	119	238
5	134	269	539
6	117	234	469
7	105	211	422
8	120	241	483
9	68	137	275
10	70	141	283
11	68	137	274

Figure 8 Imputation Tests 01 - results for each chromosome showing the squared correlation (R2) between imputed and reference genotypes based on sub-sampling different SNP panels (5%, 10% and 20%) for the 11 chromosomes.

Imputation Tests 2

The second set of imputation tests (Imputation Tests 02) use EGLOB HD SNP Set1 as the data set and were primarily aimed at investigating Pipeline 3 imputation across a high-density SNP set based on WGS rather than chip data. The data set (EGLOB HD SNP Set 1) is more representative of the breeding program diversity with samples more evenly derived from across generations. There is structure in the data reflecting the population history of the breeding program and the trees are far less related overall than was seen in EGLOB LD SNP Set 1. Results of the imputation, by chromosome, are shown in Figure 9, with imputation accuracies increasing from a very low level for small SNP sub-selections used in training to a maximum of ~0.6 per sample with the largest sub-sample of 20%. This lower accuracy is more likely reflective of the current state of imputation accuracy achievable in the HD sets with the rapid LD breakdown meaning that very large marker numbers are required to impute genome wide to high accuracy. The lower overall genotype accuracies and the smaller number of individuals in the reference set are also likely factors coming into play in this data set. It is likely there are sub-sets of markers that are readily imputable consistently within this set and a research task will be to investigate to find and characterise these sets.



Chr	1	5	10	20
1	5386	26933	53866	107733
2	6769	33848	67696	135392
3	7482	37414	74828	149656
4	4800	24000	48001	96002
5	7360	36802	73605	147210
6	6284	31420	62841	125683
7	6001	30007	60014	120028
8	9079	45398	90796	181592
9	4802	24014	48028	96056
10	5182	25913	51827	103654
11	5536	27681	55362	110724

Figure 9 Imputation Tests 02 - results for each chromosome showing the squared correlation (R2) between imputed and reference genotypes based on sub-sampling different SNP panels (1, 5%, 10% and 20%) for the 11 chromosomes.

Imputation Tests 3

The third set of imputation tests (Imputation Tests 03) use EGLOB HD SNP Set 2 as the data set and were primarily aimed at investigating the addition of samples from low coverage WGS to create a larger training set, and its impact on the final achieved imputation accuracies. The data set (EGLOB HD SNP Set 2) is very representative of the breeding program diversity with samples evenly derived from across generations and including most parents that have been used in the program linking the germplasm pools within and across generations. As for EGLOB HD SNP Set 1, there is structure in the data reflecting the population history of the breeding program and the trees are less related overall than was seen in EGLOB LD SNP Set1. Results of the imputation for Chromosome 1 are shown in Figure 10, with imputation accuracies increasing by around 20% with the larger training set. This accuracy improvement reflects the importance of training set size and indicates that increasing the remaining set to between 10 and 20 thousand trees will make reasonably high imputation accuracy achievable in the HD sets. This larger training set requirement is reflective of what is observed in human data sets and is a point of departure from what is observed in domestic animal and plant species where relatively small training sets are required to drive high imputation accuracy to the whole genome level. This result indicates that continued genotyping using a WGS approach would be a valid strategy in the eucalypts as it will serve to develop a training set of a more appropriate scale to drive imputation within the breeding program as well as improve the overall genotype accuracy within samples as each allele will be sampled increasingly frequently across the program. Low pass WGS in the blue gum program is currently very cost competitive compared to available chip platforms and returns a significantly higher information content and data value compared to genotyping with low density assays. This makes the data re-useable and accumulative to drive future advances in outcomes from imputation. While there is a trade-off between within individual marker (SNP) genotype accuracy and SNP number it seems that this trade-off decreases as the number of samples assayed with low coverage WGS increases. This is because information shared between related trees can be used to improve the underlying site by site genotype accuracies as shown with the development of EGLOB HD SNP Set2 using the Pipeline 2 (GLIMPSE) imputation procedure. As it is an active area of research it is also probable that methods for low pass WGS will continue to improve over the coming years.

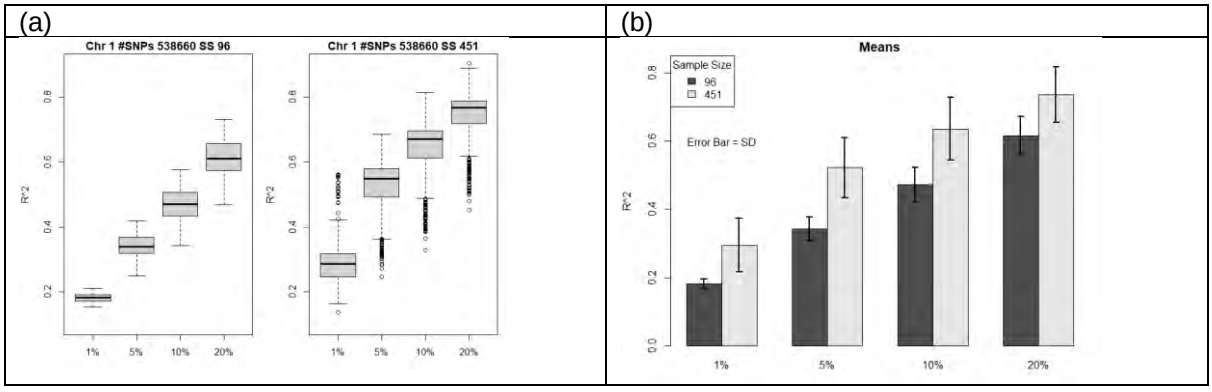


Figure 10 Imputation Tests 03 - sample size effect: (a) results for Chromosome 1, showing the squared correlation (R2) between imputed and reference genotypes based on sub-sampling different SNP panels (1, 5%, 10% and 20%). Left plot (SS 96), had 96 test samples and 864 reference samples, the right plot (SS 451) 451 test and 4059 reference samples (approx. 4.7X more samples). (b) Show means and standard deviations for data given in (a).

Investigation of the intersection of low- and high-density SNP sets

Intersection 1 (matching SNP in common to Euc72K chip and AgVic sequence assembly)

The ability to use data sets from different platforms relies upon the ability to match markers, in terms of their genomic referencing, that are featured on each of the platforms. This is best achieved through use of a common reference genome sequence. The *E. grandis* reference sequence was the basis of the referencing of SNP on the Euc72K chip. Thus, it made sense to base the referencing of SNP discovered *de novo* in our sequencing work on this same reference sequence.

The first step was to match the EuChip72K SNP probes to the EX46-HAC-Mod1-P.R2.RaGOO_EX46CP-EX46MT genome assembly (or RaGOO assembly for short) using the DNA sequence alignment tool NUCLEAR (GYDLE, https://gydle.com/innovations). This genome assembly used the *E. grandis* reference genome sequence for chaining scaffolds into super-scaffolds. The alignment reports back all matches across the entire genome for each flanking sequence unique to each SNP. The list of alignments was then filtered for single alignments and for alignments with 100 coverage of the flanking sequence (allowing mismatches). These alignment results are summarised in Table 6. Of the 68,055 probes on the chip only 41,321 could be unambiguously mapped to the RaGOO assembly. More flanking sequences did map but these were filtered out due to incomplete mappings or due to multiple mapping locations across the genome. Of the 68,055 probes 33,019 returned polymorphic signal across a panel of approximately 1500 trees sampled from the breeding program. When combined there are only 20,053 SNP that map to the RaGOO assembly and have a polymorphic signal.

Table 6 Summary of Euc72K Chip flanking sequence alignments against the Eucalyptus globulus reference genome assembly EX46-HAC-Mod1-P.R2.RaGOO_EX46CP-EX46MT.

SNP Group	Number
On Euc72K chip	68,055
Polymorphic in TBA breeding population	33,019
Mapped to reference genome	41,321
Polymorphic and mapped	20,053

Intersection 2 (matching SNP in common to Euc72K chip and EGLOB HD SNP set 1)

The second level of intersection is finding those SNP on the Euc72K Chip that are successfully matched to the RaGOO assembly, and which are also identified in the set of SNP discovered overall, and then of these, are included in EGLOB HD SNP Set 1.

There were 6.87 million bi-allelic SNP (MAF >1%, Maximum Missing <10%) discovered across the *E. globulus* genome using the set of 963 accessions with an average coverage of greater than 6 (see Table 4). The encodings used on the Euc72K Chip SNP vary in most cases with the encodings on these *de novo* discovered SNP. A minority of 29.2% of SNP have the exact encoding, while 69.3% of the SNP have an opposing, or opposite and opposing strand encoding of the REF and ALT bases (i.e. the REF base in one set is the ALT base in the other AND/OR it is called on the opposing strand). Opposite and opposing strand encoding are common issues when comparing SNP genotyping platforms and generally reflects the different reference genomes used. For joint usage these opposing and opposite strand encodings need to be unified. For a very small number of SNP (1.5%) the ALT or REF bases do not agree between the two sets regardless of strand or REF/ALT encoding pattern and these represent cases where there is likely an error in one set or the other. These positions should be excluded from analysis.

The EGLOB HD SNP Set 1 is a subset of these overall variants and has a total of 6.8 million SNP that are bi-allelic and with a minor allele frequency of greater than 1%. As the EGLOB HD SNP Set 1 is a filtered set many true SNPs will have been left out of this set, either because they are found in genomic regions where there is also small INDEL variation or because the frequency is below 1% in our discovery population (and therefore in the breeding program). Small INDEL variation is common across the *E. globulus* genome between accessions and many SNP are of very low MAF. Surprisingly the final intersection of mapped Euc72K SNP and *de novo* discovered SNP in the EGLOB HD SNP Set1 reduces to 2,732 SNP positions after all filtering and matching. This is a very small subset and presents a significant challenge to cross platform data integration via this route, without imputation fully established as a routine operation.

Intersection 3 (matching SNP in common to Euc72K chip and SNP discovered using a pre-RaGOO assembly)

To combine assay results from multiple genotyping assays we instead called WGS SNP against the *Eucalyptus grandis* genome assembly Version 2.0 (https://phytozome-next.jgi.doe.gov/info/Egrandis_v2_0). This strategy aimed to limit the loss of SNP due to strandedness and REF/ALT allele state as both the WGS SNP and the Euc72K would be referencing the same genome assembly. The downside would be the increased genetic distance between the reference genome sequence and the WGS data which can increase the rate of missing data due to dropout from divergent sequences which do not map. Alignments of samples with a coverage >4 were used for a *de novo* SNP discovery and then the GLIMPSE imputation pipeline was used to genotype remaining samples with coverage <4, essentially following the same procedures as described above. Table 7 summarises the intersection process. The Euc72K manifest included 67,683 SNP loci of which 33,660 remained after processing the 1508 individuals directly assayed with the chip (see section on EGLOB LD SNP Set1). The processing of individuals identified SNP with low MAF and/or high missingness. Of these 33,660 SNP, 18,582 were identified in the *de novo* SNP discovery process. That is, they were shown to be biallelic, and matched the exact REF/ALT state and strandedness as reported in the Euc72K manifest. In the final intersection set there were 17,103 SNP loci that could be used in a joint chip and WGS analysis where exact matching between the SNP in the WGS genotypes and the reported Euc72K genotypes was maintained.

Table 7 Numbers of SNP after each filtering step.

Chromo-some	72K SNP Manifest	72K – Filtered (MAF 1%, Miss30 %)	Orzenil vs EGRA *de novo* manifest allele match	Final Data Allele Match
1	6764	3222	1652	1379
2	8943	4348	2467	2191
3	7948	3765	2357	2137
4	3911	2096	968	940
5	7183	3088	2067	1860
6	6927	3595	1990	1769
7	6572	3155	1695	1630
8	6572	3367	1940	1822
9	4243	2395	1122	1120
10	4457	2444	1180	1157
11	4163	2185	1144	1098
TOTAL	**67683**	**33660**	**18582**	**17103**

Operational demonstration of genomic selection in E. globulus

Construction of a consolidated GRM

AVR supplied a dosage file consisting of genotype calls made on 7,296 individuals for 17,103 SNP. Dosages are real numbers ranging between 0 and 2 (3 is used for a missing call). The closer the number is to an integer (0, 1 or 2), the more accurate is the call. A value of 0.5 for example is indicative the call is equally likely to be a 0 or a 1. The 7,296 individuals included approximately 2,000 juvenile progeny that have been more recently assayed using low pass WGS, in addition to all individuals assayed to date using either high and low pass WGS and the Euc72K chip.

The genomic relationship coefficients in an initial build were compared to relationship coefficients derived using only pedigree. This comparison provides an initial check on the integrity of the GRM.

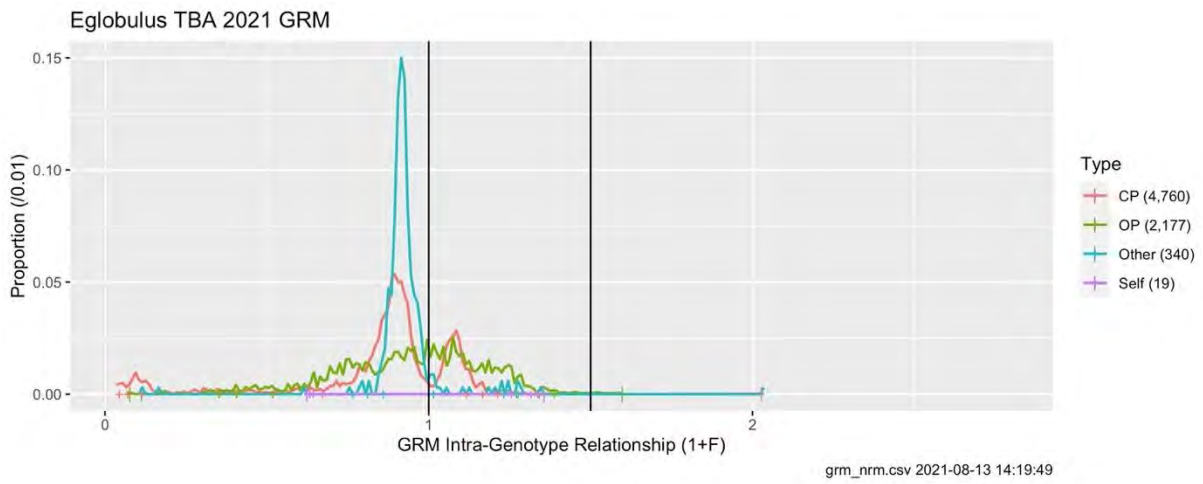


Figure 11 Frequency distributions of GRM intra-genotype coefficients, by type (CP=individual is result of cross pollination, OP=individual is result of open pollination, Self=individual is a selfed tree).

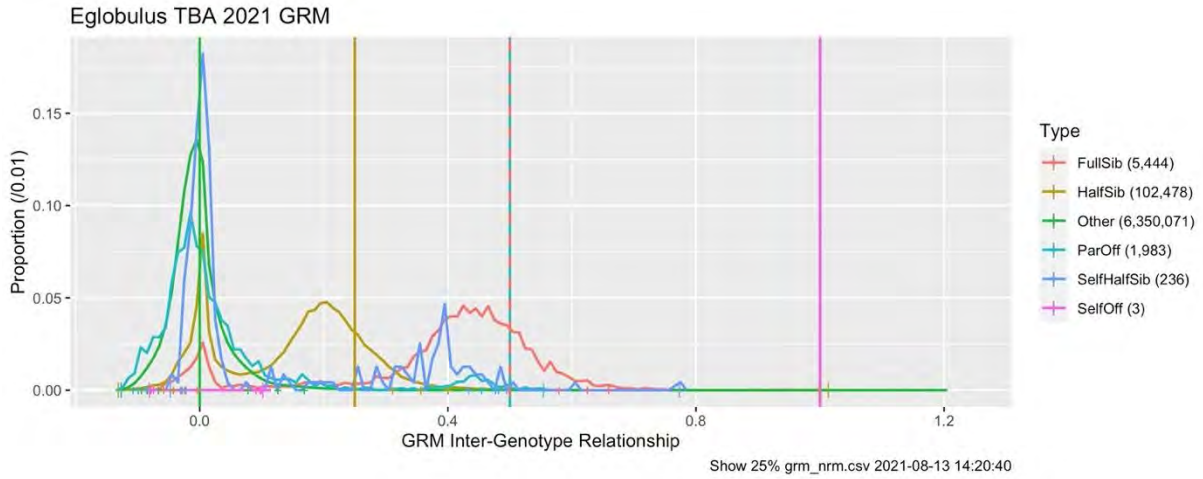


Figure 12 Frequency distributions of GRM inter-genotype coefficients, by relationship determined from the pedigree.

Figure 11 plots the frequency distribution of the diagonal elements of the GRM (intra-genotype coefficients) for different classes of individual according to mating type (result of cross-, open-, or self-pollination). Figure 12 plots the frequency distribution of the off-diagonal elements of the GRM (inter-genotype coefficients) for different classes of individual according to relational type (full-sibs, half-sibs,

etc). It appears the GRM coefficients are generally underestimated. The mode of the GRM coefficients between recognised half-sibs is below the theoretical, expected value of 0.25, while that for recognised full-sibs is below the theoretical, expected value of 0.5. Diagonal GRM elements for non-inbred trees are generally lower than the expected value of 1.

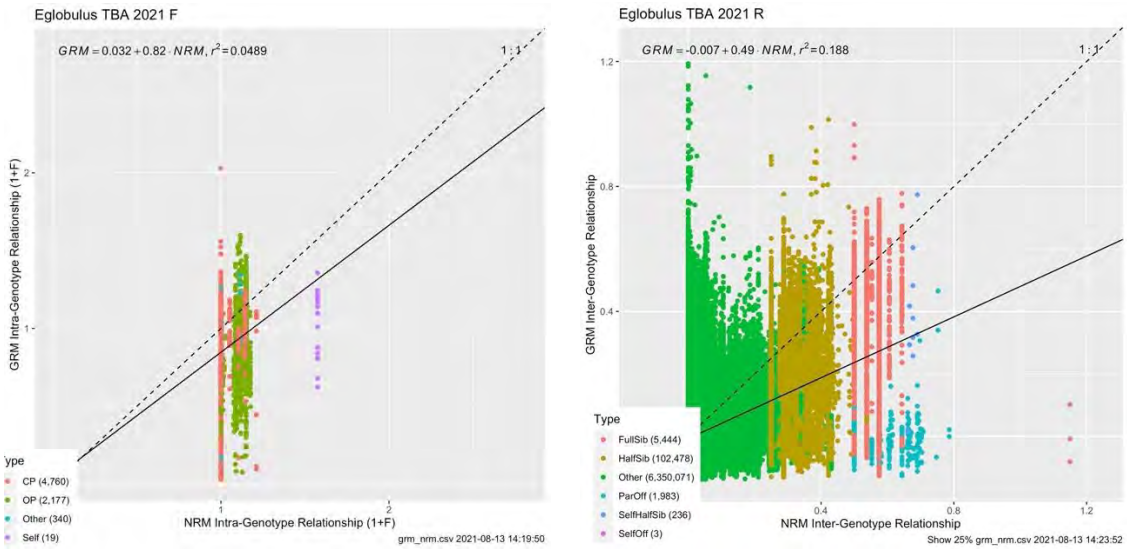


Figure 13 X-Y plots of intra- and inter-genotype GRM coefficients against NRM coefficients.

Figure 13 shows X-Y plots of GRM coefficients against NRM coefficients at the intra-and inter-genotype levels. The slopes of the plotted trend lines have values less than 1 indicating the GRM coefficients are generally lower than the corresponding NRM coefficients.

A heat map of the coefficients of the 1,235 assayed individuals that belong to either generation 0 or generation 1 and are not the result of inter-subrace crossing is displayed in Figure 14. A clustering algorithm was applied to the coefficients prior to generating the heatmap. It would appear there are between 5 and 6 main clusters, supporting the results from the principal components analysis (see Figure 4). A small cluster aligns with individuals representing the Wilson's Promontory provenance, which was not evident in the PCA plot.

Figure 14 Heat map of the GRM coefficients for 1235 trees representative of a base population.

Several hypotheses were put forward for causing the observed underestimation of GRM coefficients:

1. SNP allele frequencies, which are used to centre and scale the genomic relationships, are not fully reflective of the "true" base population allele frequencies because currently they are computed using genotype call data on all assayed trees
2. Population structure is not accounted for in the computation of the GRM
3. Ascertainment bias exists in the sense that the allele frequencies for chosen SNP are not fully reflective of the true spectra of allele frequencies in the genome.

We tested hypothesis 1 by computing allele frequencies using only 1,235 generation 1 individuals (which are offspring of randomly selected native mothers). We tested hypothesis 2 by incorporating a vector of cluster membership into the van Raden methodology for computing GRM coefficients. The revised methodology uses the allele frequencies specific to each cluster. Hypothesis 3 was tested by deliberately fixing the value of the scaling factor. The value for the scaling factor is the average heterozygosity in the population ($2\sum p_i(1-p_i)$), which was estimated as 0.28. Ascertainment bias may have prevented the inclusion of SNP with very low MAF, which if included, may lead to lower average heterozygosities. We fixed the value for the scaling factor at values less than that observed.

The strategies for testing hypotheses 1 and 2 did not significantly help in making GRM coefficients align better to the NRM coefficients. The strategy for testing hypothesis 3 did lead to a better alignment.

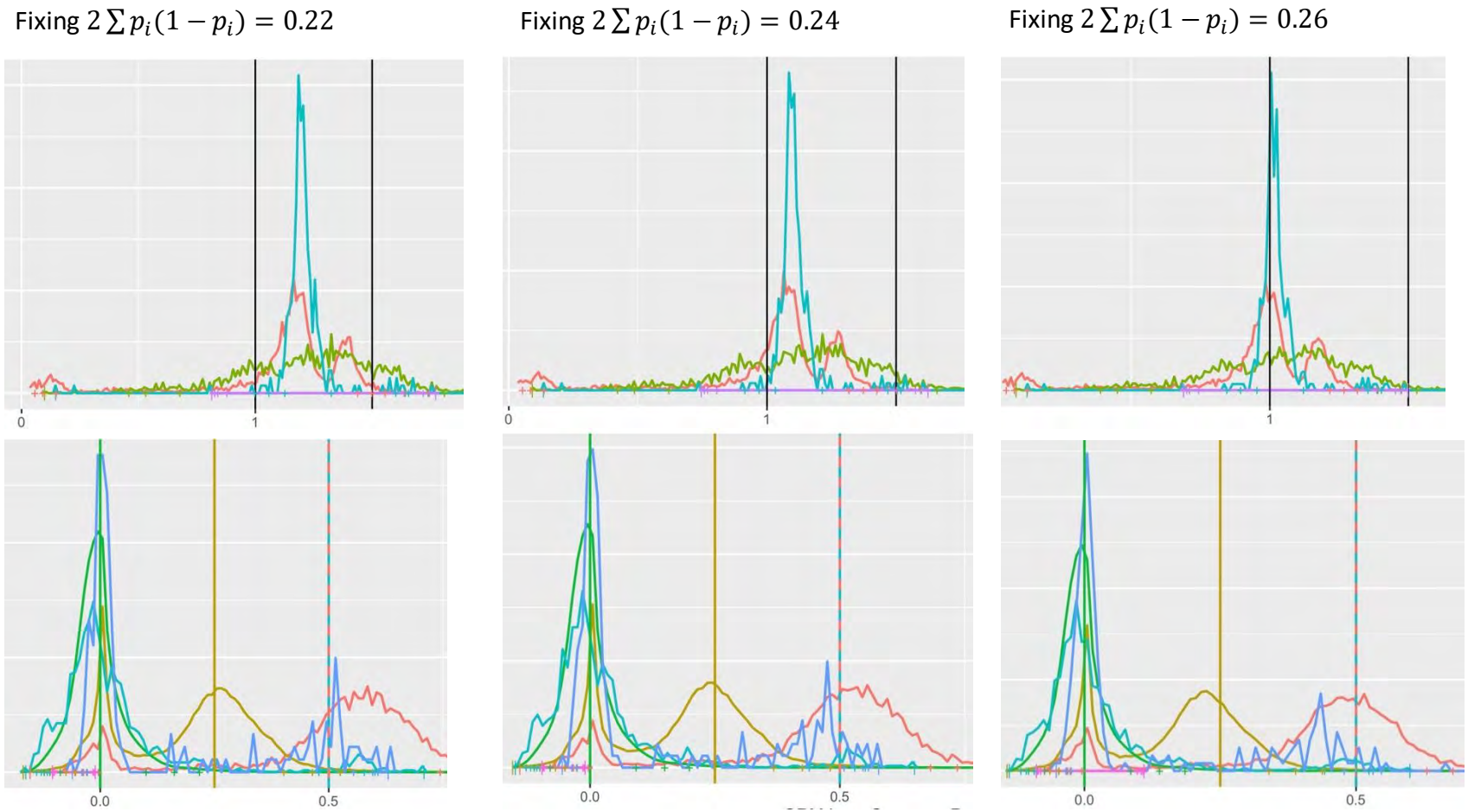
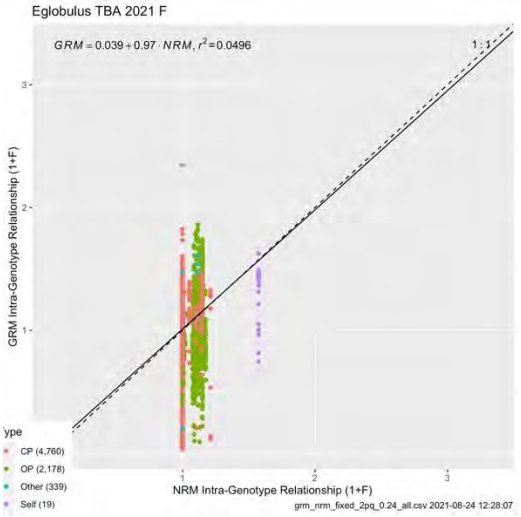
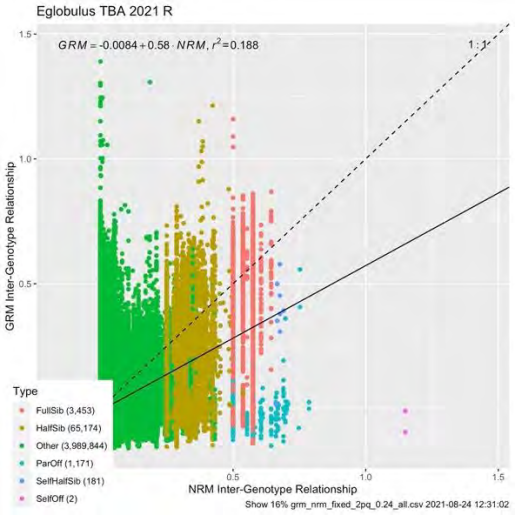


Figure 15 Frequency distributions of intra- (top row) and inter-genotype (bottom row) genomic relationship coefficients, when deliberately fixing average heterozygosity ($2\sum p_i(1-p_i)$) at values less than that observed. The species in this case E. globulus.





Single-step analysis in *E. globulus* with a consolidated GRM

Single step analyses in *E. globulus* that have been run between years 2018 and 2020 have used a **G** matrix constructed with approximately 800,000 SNP. These SNP sets were defined in previous projects such as PNC408-1516 'Single-Step TREEPLAN Incorporating genomic data in TREEPLAN evaluations to increase genetic gain', using genome assemblies which have since been superseded. At the time of publishing the final report for PNC408-1516, we had assayed approximately 4,300 individuals, of which approximately 2,900 were included in the final **G** matrix. Many had been excluded because of poor co-call rates. Low co-call rates were an impediment to reliable **G** matrix construction, and this was the impetus to develop imputation methods to improve genotype reliability from low coverage WGS data sets. Prior to this project we were not using imputation to fill-in missing calls and relying on the KGD method (Dodds et al. 2015) for **G** matrix construction.

For the 2021 single-step analysis in *E. globulus* we constructed a **G** matrix based on the intersection of SNP between the EGLOB LD and HD SNP Sets referenced to the *Eucalyptus grandis* genome assembly. A SNP dosage file was generated that contained allele dosages for 7,296 samples, across 17,103 SNP. This increased number of samples (7,296) reflects the inclusion of WGS samples that had been previously excluded due to poor genotype call rates, the additional samples for which WGS has been completed in this project, the 1,508 (1,056 new) samples assayed with the Euc72K chip and the 2,000 juvenile progeny recently assayed using low-coverage WGS. At the time of the 2021 analysis the system contained 722,685 observations for 18 selection criteria (SC), measured on stems at 341,824 positions across 174 trials. There were 347,573 genotypes and 7,168 families in the pedigree. The selection criteria are correlated to varying degrees to 7 breeding objective traits (BOT). Multiple $NPV Indices have been defined by the economic weighting of BOT. For the 2021 analysis a new **G** matrix was constructed using standard "van Raden" methodology. It was aligned to the **A** matrix using the methodology suggested by Legarra et al. (2014) and then weighted with the **A** matrix using a lambda (λ) of 0.2.

Data, pedigree, the new **G**-matrix and parameters were extracted from DATAPLAN for the TREEPLAN system 'EGlob_May2021' (SystemID=1000). The prediction error variances (PEV) of the genetic effects in the TREEPLAN single-step model were computed using a trial version of the Linear Mixed Models Toolbox (LMT) software supplied by Dr Vinzent Boerner. This software has more advanced algorithms for PEV computation than software currently used by TBA (SSSADI). Accuracies

($r_{u\hat{u}}$) of EBV for selection criteria, breeding objective traits and NPV $Indices were computed as a function of the PEV and either the diagonals of the **H**-matrix, or the **A** matrix, for the values 1 + F in the following equation:

$$r_{u\hat{u}} = \sqrt{1 - \frac{PEV}{(1+F)\sigma_a^2}}.$$

X-Y plots showing the selection criteria trait (SCT) accuracies with and without the **G** matrix are shown in Figure 16 for assayed trees and in Figure 17 for non-assayed trees. Plots showing the breeding objective trait (BOT) accuracies, for both assayed and non-assayed trees are shown in Figure 18. Points have been coloured to denote trees in different generation by parent status categories (e.g. “Gen-0.parent” and “Gen-0.non-parent” denote parents and non-parents in generation 0, respectively). Table 8 shows the distribution of individuals among the various categories. Gen-2 non-parents have been assayed the most, and there are very few Gen-0 individuals assayed. Table 9 contains the mean EBV accuracies and percentage change in the mean (%), when using the **H** or **A** matrices for the BOT: VOL_GTR, Density and Kraft Pulp Yield. Results for the other regional based volume BOT (e.g VOL_TAS, VOL_WA etc) are like VOL_GTR and not shown.

Non-parents have lower accuracy than parents, and individuals in earlier generations have lower accuracy than individuals in later generations. The X-Y plots and tables show that the individuals that benefit most from a DNA assay are the earlier generation non-parents. The improvement in accuracy is largest for the few Gen-0 non-parents, particularly for a SCT such as Predicted Pulp Yield (PPY) which has a lower incidence of measurement. Accuracies of BOT reflect the accuracies of the SCT that are more highly correlated to them (accuracy of EBV for Kraft Pulp Yield is mostly a reflection of the accuracy for PPY). Hence, we see a 35% improvement in accuracy for Gen-0 non-parents for Kraft Pulp Yield (see Table 9).

Table 8 Distribution of assayed individuals that comprise the 2021 GRM, by generation and parent-status (in E. globulus).

	Parent	Non-parent	Total
Gen-0	7	11	18
Gen-1	349	871	1220
Gen-2	97	3067	3164
Gen-3	0	746	746
			5148

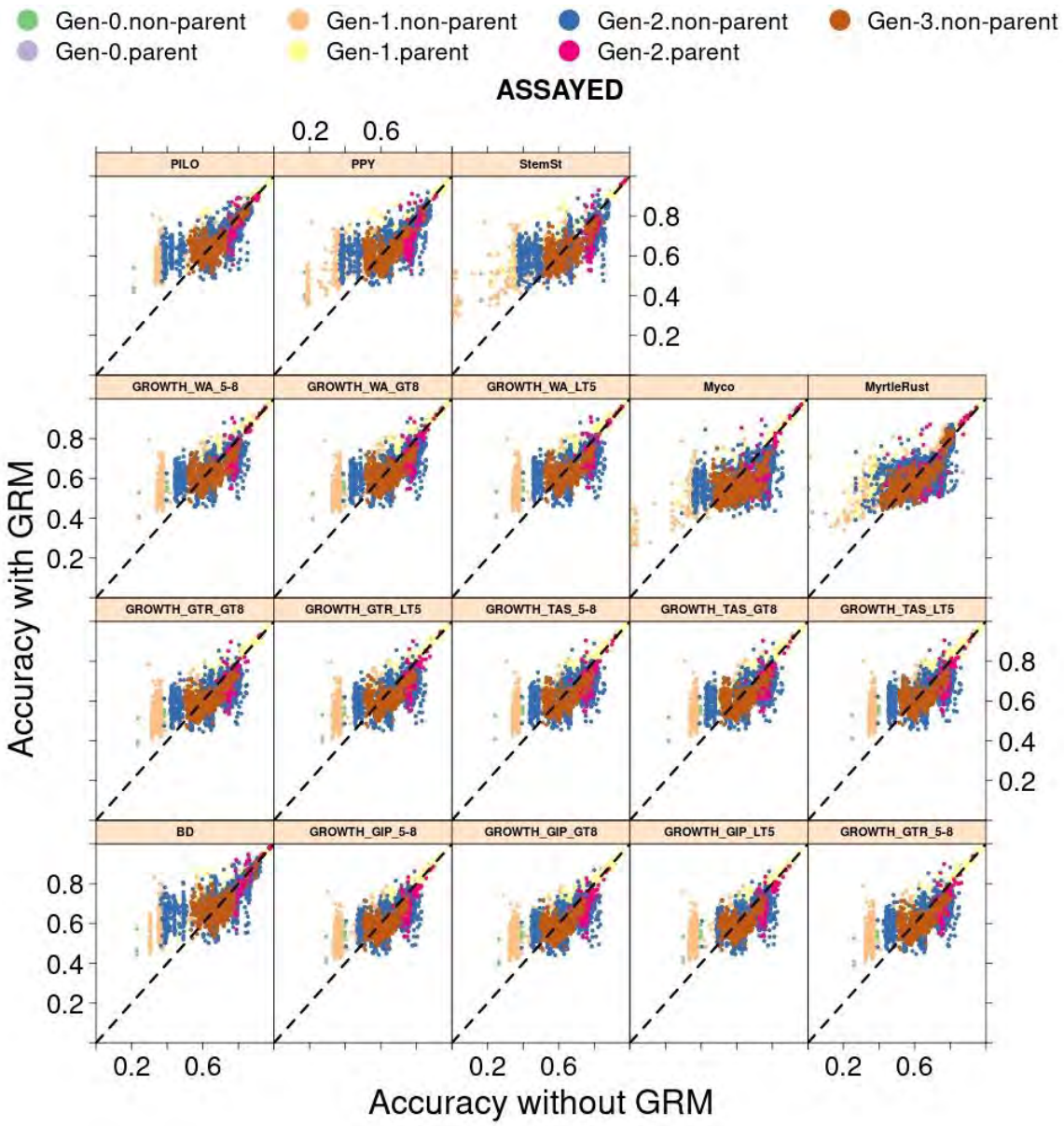


Figure 16 Accuracies of EBV for assayed trees, for selection criteria traits (SCT) computed with and without the 2021 genomic relationship matrix (GRM) in E. globulus.

In Figure 16 we generally are seeing large bands of increased accuracy for Gen-1 non-parents (bisque colour), and for those Gen-2 and Gen-3 non-parents (dark blue and brown colours) that initially have low accuracy. There is less tendency to observe a marked improvement in accuracy for parents as these generally have high *prior* accuracy. However most Gen-1 parents (yellow colour) do have increased accuracy when the GRM is used, possibly due to recovery, or discrimination, of previously unknown paternal relationships.

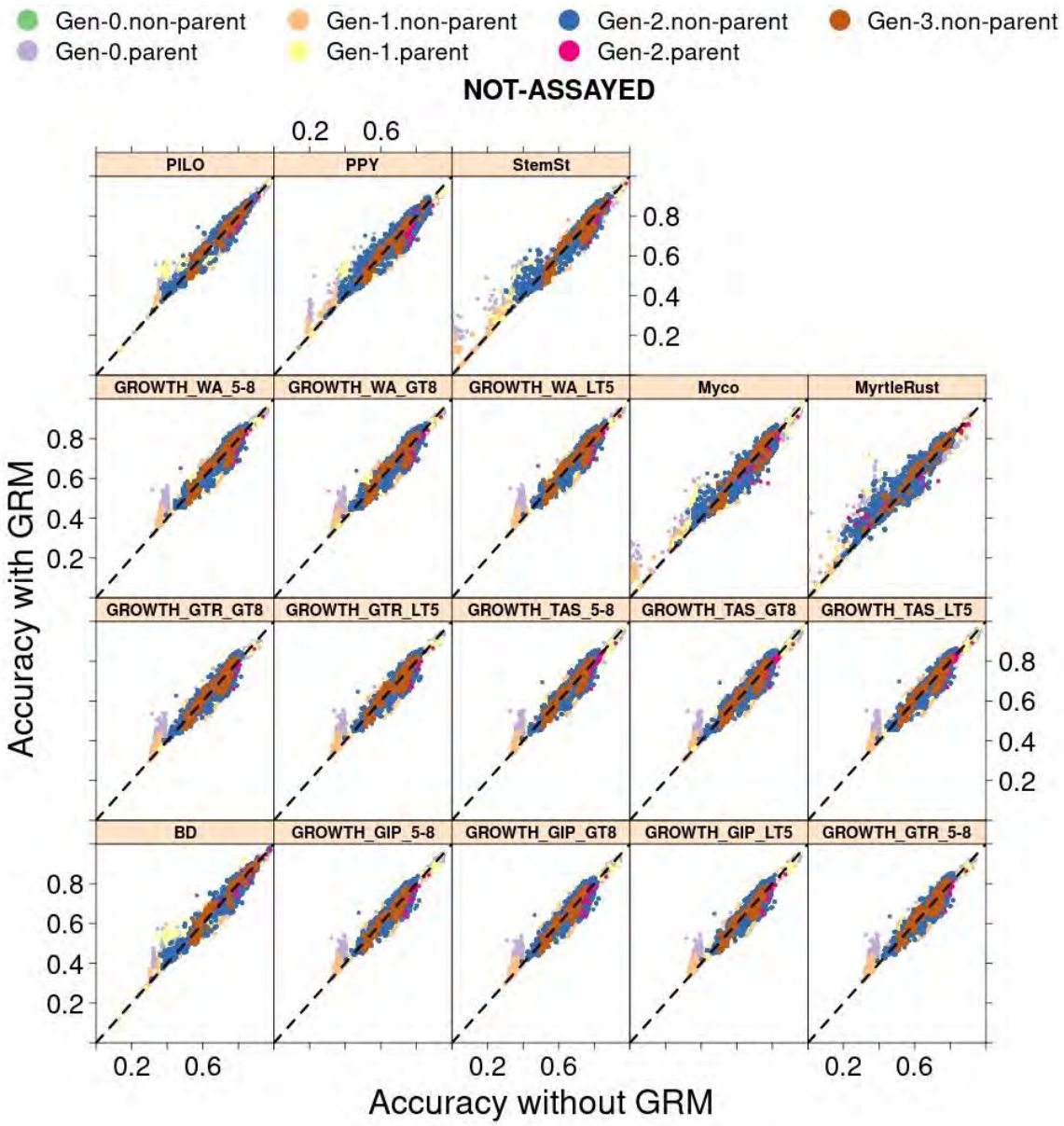


Figure 17 Accuracies of EBV for non-assayed trees, for selection criteria traits (SCT) computed with and without the 2021 genomic relationship matrix (GRM) in E. globulus.

In Figure 17 (accuracies associated with non-assayed trees) we observe distinct increases in accuracy across all SCT for Gen-0 parents (mauve colour) and Gen-1 non-parents (bisque colour), and distinct increases for Gen-1 parents for some SCT such as basic density (BD), PILO and PPY.

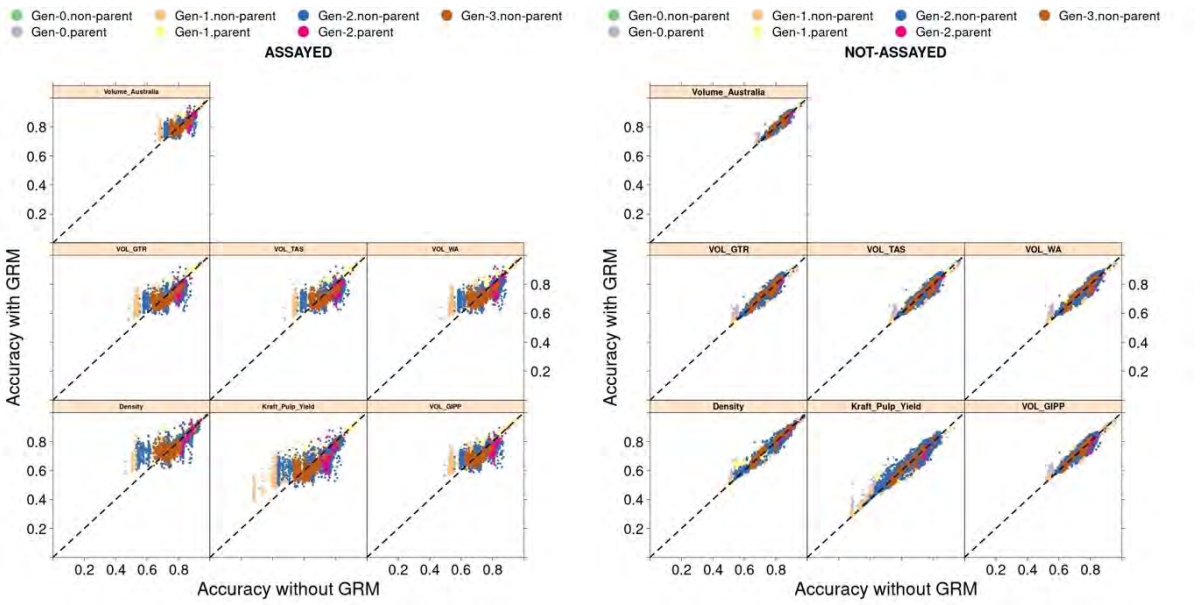


Figure 18 Accuracies of EBV for breeding objective traits (BOT) computed with and without the 2021 genomic relationship matrix (GRM) in E. globulus. The left plot shows accuracies for assayed trees and the right plot shows accuracies for non-assayed trees.

Figure 18 shows that the changes in accuracy for BOT mirror those observed for SCT (Figure 17). The tables below (Table 9) show that on average, there is a small average decrease in accuracy for later generational categories, such as Gen-2 parents and Gen-2 non-parents and Gen-3 non-parents, when the **H** matrix is used. This decrease may reflect the impact of the realised relationships being reflected rather than the average expected relationships. It may also reflect accumulation of pedigree errors. It is in these categories of trees that we desire an increased accuracy from assaying because it is from these categories that new selections are made. In *E. nitens* (results to follow) we do not observe this happening and in fact the largest increases in accuracy are observed for these categories of trees. In *E. nitens* there is a much larger population of assayed trees and having an insufficient cohort of trees assayed may be a reason we observe the opposite trend in *E. globulus*. What is clear is that the largest changes (increases) in accuracy are observed in the trees with the lowest initial accuracy whilst trees with higher initial accuracies seem to have the same or decreased accuracy.

Table 9 Mean EBV accuracies for VOL_GTR, Density and Kraft Pulp Yield and percentage change in the mean (%), when using the H (based on the 2021 GRM) and A matrices in the mixed model equations (in E. globulus).

	Assayed						Non-Assayed					
VOL_GTR												
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.85	0.82	3.51	0.64	0.54	18.00	0.71	0.70	0.82	0.63	0.62	0.15
Gen-1	0.83	0.81	2.48	0.70	0.65	7.98	0.70	0.70	-0.36	0.68	0.68	-0.29
Gen-2	0.77	0.81	-4.84	0.74	0.78	-4.82	0.79	0.80	-1.30	0.78	0.77	1.01
Gen-3				0.68	0.70	-2.26				0.76	0.76	0.61
Density												
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.89	0.85	4.80	0.67	0.53	25.57	0.74	0.74	0.91	0.67	0.67	0.02
Gen-1	0.90	0.88	2.45	0.75	0.69	9.85	0.68	0.69	-0.33	0.69	0.70	-0.44
Gen-2	0.88	0.89	-1.01	0.82	0.83	-0.62	0.84	0.85	-0.61	0.81	0.80	0.83
Gen-3				0.74	0.75	-0.54				0.82	0.81	1.12
Kraft Pulp Yield												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.73	0.66	10.71	0.55	0.41	34.67	0.54	0.52	1.98	0.45	0.45	0.49
Gen-1	0.74	0.70	5.57	0.61	0.53	15.49	0.54	0.54	0.13	0.51	0.51	-0.04
Gen-2	0.69	0.75	-7.59	0.70	0.73	-3.47	0.67	0.69	-1.78	0.68	0.68	1.32
Gen-3				0.61	0.62	-2.21				0.66	0.66	-0.06

Figure 19 indicates that use of a GRM can result in substantial increases in the accuracy of a NPV $Index EBV. The infrastructure in place in DATAPLAN to convert selection criteria trait EBV to BOT EBV and then finally to EBV for NPV $Indices provides a practical means for operational breeders to incorporate genomics into breeding decisions. Mean increases in accuracy for the specific generational class by parent status categories are provided in Table 10.

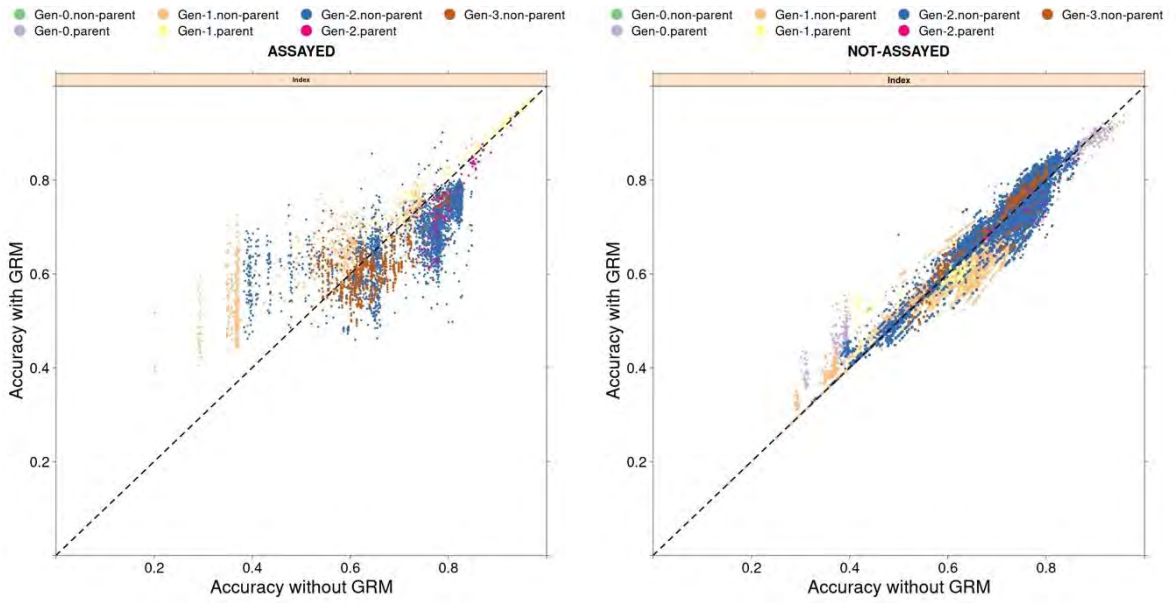


Figure 19 Accuracies of EBV for a NPV $Index (titled 'Index') computed with and without the 2021 genomic relationship matrix (GRM) in E. globulus. The left plot shows accuracies for assayed trees and the right plot shows accuracies for non-assayed trees.

Table 10 Mean EBV accuracies for NPV Index called 'Index', and percentage change in the mean (%), when using the H (based on the 2021 GRM) and A matrices in the mixed model equations (in E. globulus).

	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.81	0.74	8.86	0.53	0.36	48.82	0.60	0.59	1.61	0.50	0.50	0.13
Gen-1	0.81	0.78	3.44	0.63	0.55	15.32	0.57	0.57	-0.53	0.56	0.56	-0.63
Gen-2	0.76	0.80	-5.58	0.69	0.73	-6.02	0.74	0.76	-1.80	0.72	0.71	1.32
Gen-3				0.61	0.63	-4.03				0.72	0.71	1.40

Run 2021 TREEPLAN *E. globulus* data with 2020 version of GRM.

To investigate the implications of using a GRM constructed with more individuals but using a smaller SNP set derived from both chip and WGS assay methods, we extracted the 2020 version of the GRM from DATAPLAN to use with the same TREEPLAN system 'EGlob_May2021'. The 2020 GRM contained coefficients among 2,882 individuals and was based on genotype calls for 856,011 SNP. There was a high proportion of missing genotype calls (37%) and no imputation was used. PEV and accuracies were again computed using the LMT software.

Figure 20 shows the X-Y plots of accuracies computed with and without the 2020 GRM, for assayed and non-assayed trees. They generally show that the increases in accuracy for those categories of trees, which were shown previously to have higher accuracy with the 2021 GRM, are substantially less relative to the 2021 scenario. For example, there is a 4.1% increase in accuracy for Gen-1 non-parents for the BOT Kraft Pulp Yield when using the **H** matrix based on the 2020 GRM (see Table 11). The corresponding increase when the **H** matrix is based on the 2021 GRM is 15.5% (see Table 9). Similar patterns occur for the other BOT and for the $NPV Indices.

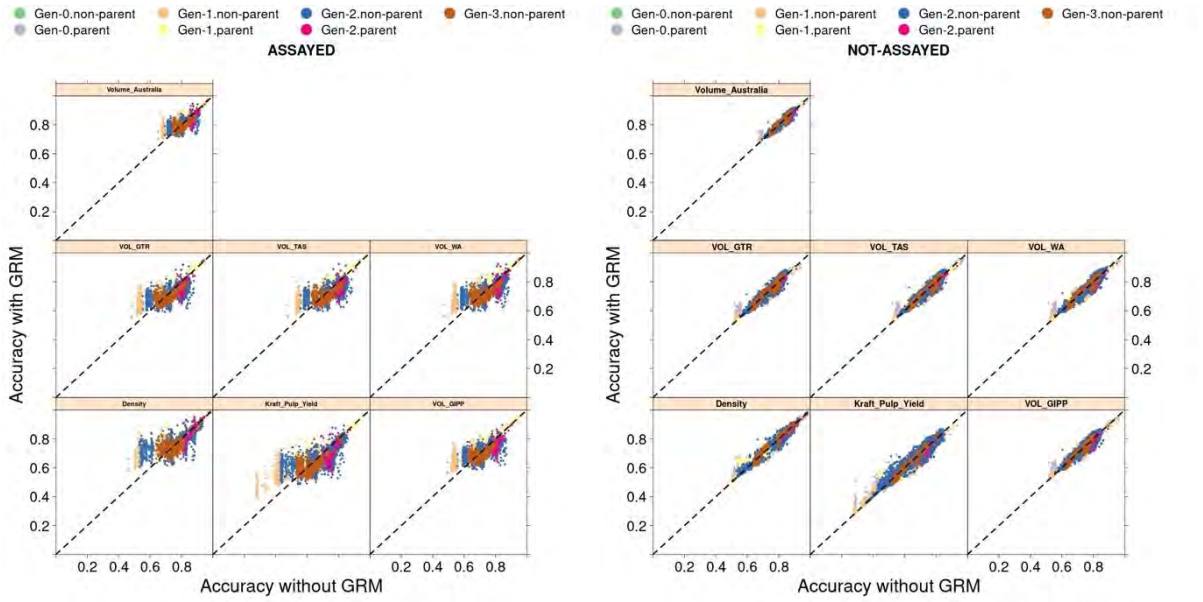


Figure 20 Accuracies of EBV for breeding objective traits (BOT) computed with and without the 2020 genomic relationship matrix (GRM) in E. globulus. The left plot shows accuracies for assayed trees and the right plot shows accuracies for non-assayed trees.

Table 11 Mean EBV accuracies for VOL_GTR, Density and Kraft Pulp Yield and percentage change in the mean (%), when using the H (based on the 2020 GRM) and A matrices in the mixed model equations (in E. globulus).

VOL_GTR												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.83	0.82	1.56	0.58	0.55	5.73	0.71	0.70	0.15	0.62	0.63	-0.11
Gen-1	0.81	0.81	0.26	0.66	0.65	2.38	0.70	0.70	-0.09	0.68	0.68	-0.14
Gen-2	0.79	0.82	-3.07	0.78	0.80	-2.61	0.80	0.80	-0.03	0.77	0.77	0.01
Gen-3				0.76	0.78	-1.57				0.75	0.75	-0.11
Density												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.86	0.84	2.26	0.59	0.54	8.27	0.74	0.74	0.24	0.67	0.67	-0.01
Gen-1	0.88	0.88	0.35	0.71	0.69	2.90	0.69	0.69	0.03	0.70	0.70	-0.11
Gen-2	0.88	0.90	-1.55	0.84	0.85	-0.75	0.85	0.85	0.18	0.80	0.80	0.10
Gen-3				0.83	0.84	-0.46				0.81	0.81	0.16
Kraft Pulp Yield												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.70	0.67	4.16	0.46	0.42	9.54	0.53	0.53	0.16	0.45	0.45	-0.52
Gen-1	0.71	0.70	0.76	0.56	0.53	4.10	0.54	0.54	-0.21	0.51	0.51	-0.33
Gen-2	0.72	0.75	-4.72	0.73	0.75	-2.52	0.67	0.67	0.19	0.67	0.67	-0.12
Gen-3				0.70	0.71	-1.51				0.65	0.65	-0.46

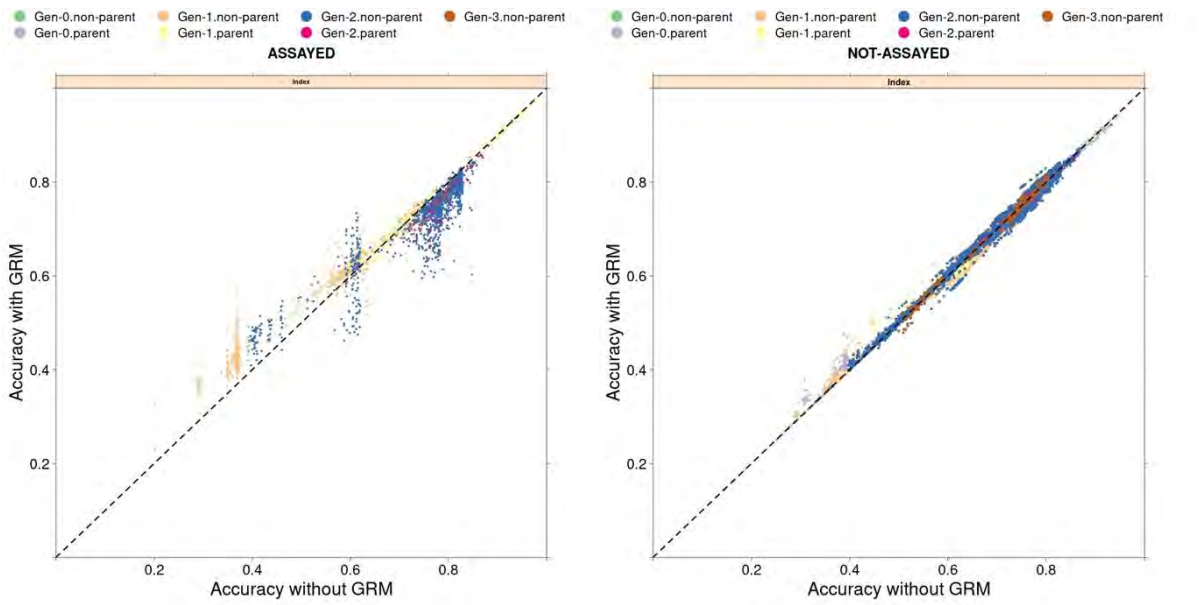


Figure 21 Accuracies of EBV for 'Index' computed with and without the 2020 genomic relationship matrix (GRM). The left plot shows accuracies for assayed trees and the right plot shows accuracies for non-assayed trees.

Table 12 Mean EBV accuracies for NPV Index called 'Index', and percentage change in the mean (%), when using the H (based on the 2020 GRM) and A matrices in the mixed model equations (in E. globulus).

	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.78	0.74	4.40	0.43	0.37	15.36	0.59	0.59	0.37	0.49	0.50	-0.11
Gen-1	0.79	0.78	0.54	0.58	0.55	4.90	0.57	0.57	-0.02	0.56	0.56	-0.22
Gen-2	0.78	0.81	-3.64	0.73	0.76	-3.27	0.75	0.75	0.19	0.71	0.71	0.11
Gen-3				0.72	0.73	-1.98				0.70	0.70	0.04

In Table 12 there is a 4.9% increase in accuracy for Gen-1 non-parents for the NPV Index called 'Index' when using the **H** matrix based on the 2020 GRM. The corresponding increase when the **H** matrix is based on the 2021 GRM is 15.5% (see Table 10).

There are many contributing factors to the better performance of the 2021 GRM. They include the following

- Improved reference genome assembly. Prior to 2021 we based SNP discovery on an assembly consisting of many thousands of contigs, but in 2021 we had a chromosome level assembly
- Improved SNP discovery pipeline based on GATK
- Imputation is now used to fill in a large proportion of missing genotype calls for those individuals assayed using whole genome sequencing
- Approximately 1000 individuals with highly accurate genotype calls made using the Euc72K chip were added to the data set

Though the 2021 GRM was based on a much smaller SNP set, the value of another 1000 assayed individuals, and having all missing genotype calls filled in using imputation appears to outweigh the negatives of a smaller SNP set. Efforts to continue increasing the number of assayed individuals should remain a high priority as the current set is still only about half to a third the scale required to achieve high cross-generation imputation accuracy.

Single-step analysis in E. nitens

In June 2021 a **G** matrix constructed for 12,386 individuals and from an unknown number of SNP was received from Gondwana Genomics. This **G** matrix was imported into DATAPLAN and flagged for use with the current national *E. nitens* TREEPLAN analysis system. This system contains 697,868 observations for 54 selection criteria (SC), measured on stems at 199,438 positions. There are 210,409 genotypes and 3,879 families in the pedigree. The selection criteria are correlated to varying degrees to 30 breeding objective traits (BOT). Multiple $NPV Indices have been defined by the economic weighting of BOT. Table 13 shows the distribution of assayed individuals among the various categories. Gen-3 non-parents have been assayed the most. As for *E. globulus* very few Gen-0 individuals have been assayed (only 1 in this case). The distinguishing feature of *E. nitens* is that most of the assayed trees (Gen-2 and Gen-3 non-parents) do not have observations, and the size of the training set (individuals with observations and have been assayed) is much smaller than in *E. globulus*.

Data, pedigree, the 2021 **G**-matrix and parameters were extracted from DATAPLAN for the TREEPLAN system '202005_GRM_2021' (SystemID=1003). The prediction error variances (PEV) of the genetic effects in the TREEPLAN single-step model were computed using a trial version of the Linear Mixed Models Toolbox (LMT) software. Accuracies were then derived from the PEV and the relevant diagonal elements of either the **H** or **A** matrices.

Table 13 Distribution of assayed individuals that comprise the 2021 GRM, by generation and parent-status (in E. nitens).

	Parent	Non-parent	Total
Gen-0	1	0	1
Gen-1	211	390	601
Gen-2	352	3807	4,159
Gen-3	0	7625	7,625
Total	**564**	**11,822**	**12,386**

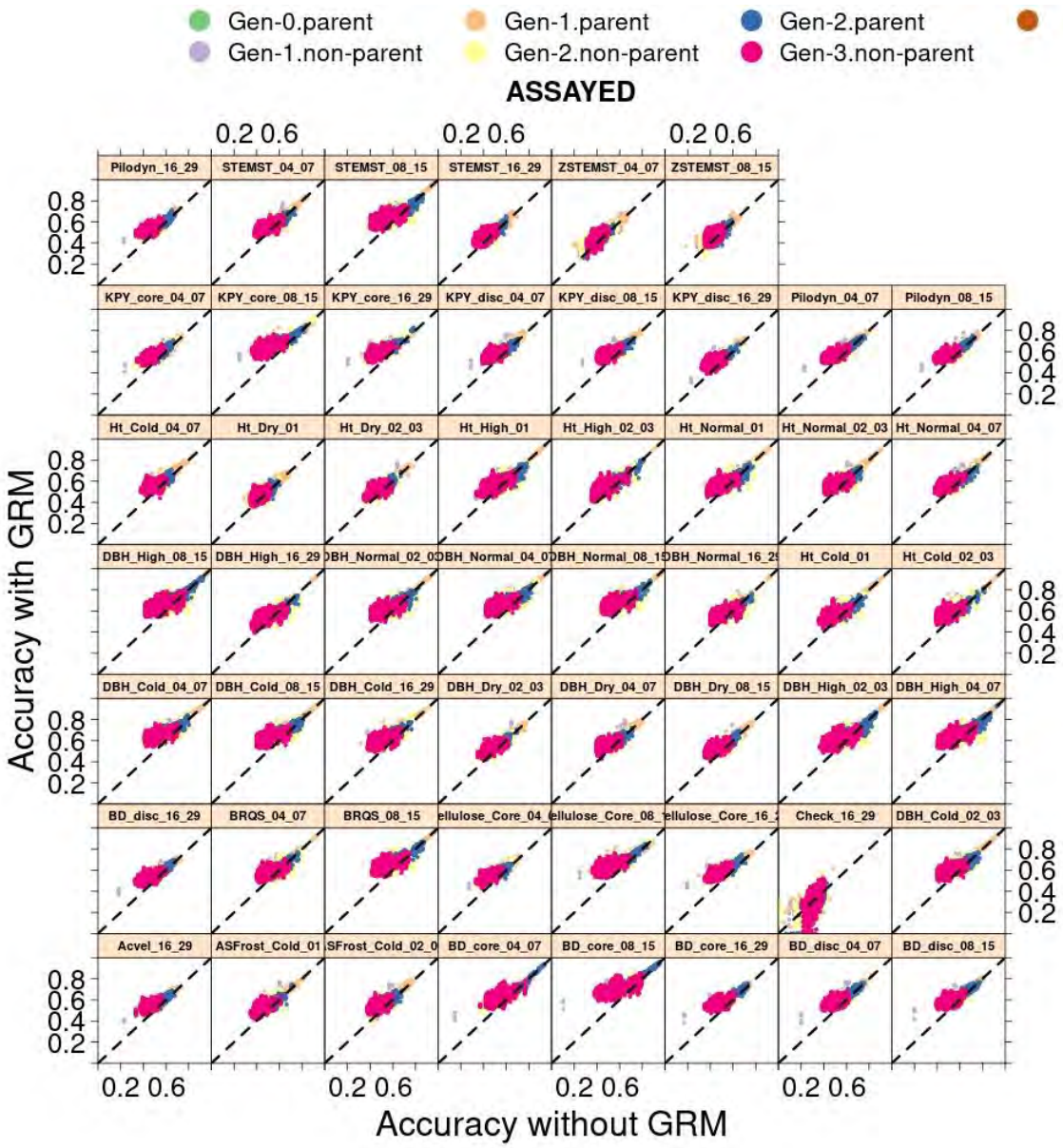


Figure 22 Accuracies of EBV for assayed trees, for selection criteria traits (SCT) computed with and without the 2021 genomic relationship matrix (GRM) in E. nitens.

In Figure 22 we see that there is a widespread improvement in accuracy for most of the assayed trees, which are dominated by the Gen-3 non-parent category (red colour). Examining the actual shift in the mean accuracies for selected selection criteria traits (SCT) we see the percent improvements are substantial (see Table 14). They range between 11 and 38 % improvement for non-parents and between 2 and 14 % for parents.

Table 14 Mean EBV accuracies for selected SCT, and percentage change in the mean (%), when using the H and A matrices in the mixed model equations (in E. nitens).

Pilodyn_08_15												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.66	0.62	5.86				0.60	0.60	0.20	0.32	0.32	0.31
Gen-1	0.69	0.66	3.58	0.61	0.54	13.75	0.54	0.54	-0.04	0.51	0.51	-0.15
Gen-2	0.66	0.63	3.72	0.60	0.54	11.42	0.52	0.51	0.55	0.51	0.51	0.41
Gen-3				0.57	0.47	20.42				0.42	0.42	0.07
KPY_core_08_15												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.78	0.75	5.08				0.54	0.54	1.45	0.31	0.31	2.11
Gen-1	0.74	0.65	13.78	0.66	0.51	30.96	0.50	0.49	2.11	0.47	0.47	1.12
Gen-2	0.71	0.62	13.94	0.70	0.58	19.68	0.55	0.53	2.78	0.53	0.52	2.57
Gen-3				0.64	0.46	37.93				0.42	0.42	0.18
DBH_Cold_08_15												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.85	0.83	2.72				0.64	0.64	0.27	0.38	0.38	0.67
Gen-1	0.81	0.76	5.75	0.67	0.56	19.82	0.56	0.56	-0.18	0.54	0.55	-0.09
Gen-2	0.72	0.67	7.51	0.69	0.61	12.52	0.62	0.62	-0.25	0.61	0.61	-0.34
Gen-3				0.63	0.49	28.96				0.44	0.44	-0.60
Cellulose_Core_08_15												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.83	0.81	2.54				0.56	0.55	1.45	0.32	0.31	3.39
Gen-1	0.74	0.65	13.43	0.67	0.53	26.97	0.52	0.51	1.75	0.49	0.48	1.08
Gen-2	0.71	0.62	13.80	0.69	0.58	18.90	0.55	0.53	2.90	0.54	0.53	2.52
Gen-3				0.64	0.47	36.06				0.42	0.42	0.73
BD_disc_08_15												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.66	0.64	3.26				0.59	0.59	0.19	0.32	0.32	0.35
Gen-1	0.73	0.70	3.12	0.63	0.55	14.97	0.55	0.55	-0.52	0.50	0.50	-0.22
Gen-2	0.69	0.67	3.34	0.63	0.57	11.01	0.55	0.55	-0.12	0.53	0.53	0.28
Gen-3				0.60	0.49	21.20				0.43	0.43	-0.60
DBH_Normal_08_15												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.73	0.56	28.61				0.73	0.73	0.03	0.43	0.43	-0.21
Gen-1	0.86	0.83	3.81	0.71	0.61	17.14	0.63	0.63	-0.50	0.62	0.62	-0.26
Gen-2	0.77	0.73	5.43	0.73	0.65	11.33	0.69	0.69	-0.78	0.65	0.66	-0.66
Gen-3				0.67	0.53	26.77				0.53	0.53	-0.42
BD_core_08_15												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.91	0.89	1.90				0.61	0.61	0.61	0.34	0.33	1.38
Gen-1	0.89	0.87	2.59	0.72	0.56	28.02	0.56	0.56	0.47	0.52	0.52	0.40
Gen-2	0.89	0.86	2.92	0.76	0.67	14.46	0.60	0.60	0.34	0.58	0.58	0.36
Gen-3				0.73	0.58	25.69				0.44	0.44	0.14

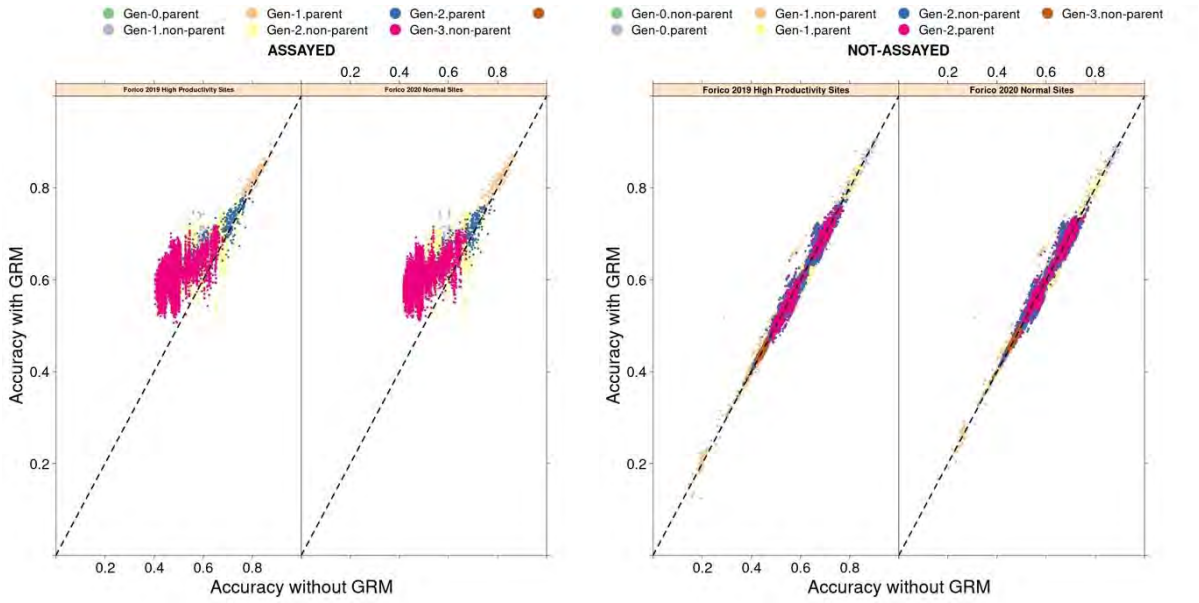


Figure 23 Accuracies of EBV for NPV indices 'Forico 2019 High Productivity Sites' and 'Forico 2020 Normal Sites' computed with and without the 2021 genomic relationship matrix (GRM) in E. nitens. The left plot shows accuracies for assayed trees and the right plot shows accuracies for non-assayed trees.

Table 15 Mean EBV accuracies for NPV Indices called 'Forico 2019 High Productivity Sites' and 'Forico 2020 Normal Sites' and percentage change in the mean (%), when using the H (based on the 2021 GRM) and A matrices in the mixed model equations (in E. nitens).

Forico 2019 High Productivity Sites												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.68	0.59	15.36				0.64	0.64	0.26	0.38	0.37	0.59
Gen-1	0.77	0.73	4.67	0.66	0.57	14.61	0.57	0.57	-0.19	0.54	0.54	-0.08
Gen-2	0.70	0.66	6.50	0.66	0.59	11.82	0.60	0.60	-0.10	0.58	0.58	0.13
Gen-3				0.61	0.49	25.58				0.44	0.44	-0.21
Forico 2020 Normal Sites												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.68	0.57	18.88				0.67	0.67	0.37	0.39	0.39	0.80
Gen-1	0.77	0.74	4.08	0.66	0.58	13.90	0.59	0.59	-0.10	0.56	0.56	0.12
Gen-2	0.69	0.65	5.48	0.65	0.59	11.15	0.60	0.60	0.02	0.59	0.59	0.24
Gen-3				0.61	0.49	24.52				0.47	0.47	0.14

Table 15 shows that there has been substantial improvement in the accuracy of predicted EBV for indices in current use by Forico, when using a **H** matrix. Gen-3 non-parents have an increase of 25%.

Development of pedigree forensics pipelines

Pedigree forensics is the detection of errors in field-based pedigrees based on genomic information. For example, an assumed parent-offspring relationship is detected as a mismatch given the results of DNA assays on both individuals. Pedigree forensics entails the recovery of the "true" parent if the true parent has also been assayed. The forensics pipeline also must include scope for undertaking quality control of the DNA assay itself, because the quality of the forensic checking is only as good as the

quality of the assay. Figure 24 shows a schematic of how a pedigree forensics pipeline will work. The schematic distinguishes between facilities, processes, actions and software. At the entry point is "raw" data, which are data supplied from a genomics service provider that is yet to be translated into actual SNP genotype calls. It may be raw sequence read data such as BAM files (or possibly FASTQ files), or it may be the CEL data that many chip manufacturers such as Thermo Fisher provide. In many cases it will not be the TBA's task to work with such data, but perhaps a research partner. Regardless of who does it, a major process is the task of converting raw data into genotype calls, and to filter samples and SNP based on various criteria such as minor allele frequency (MAF) and missingness.

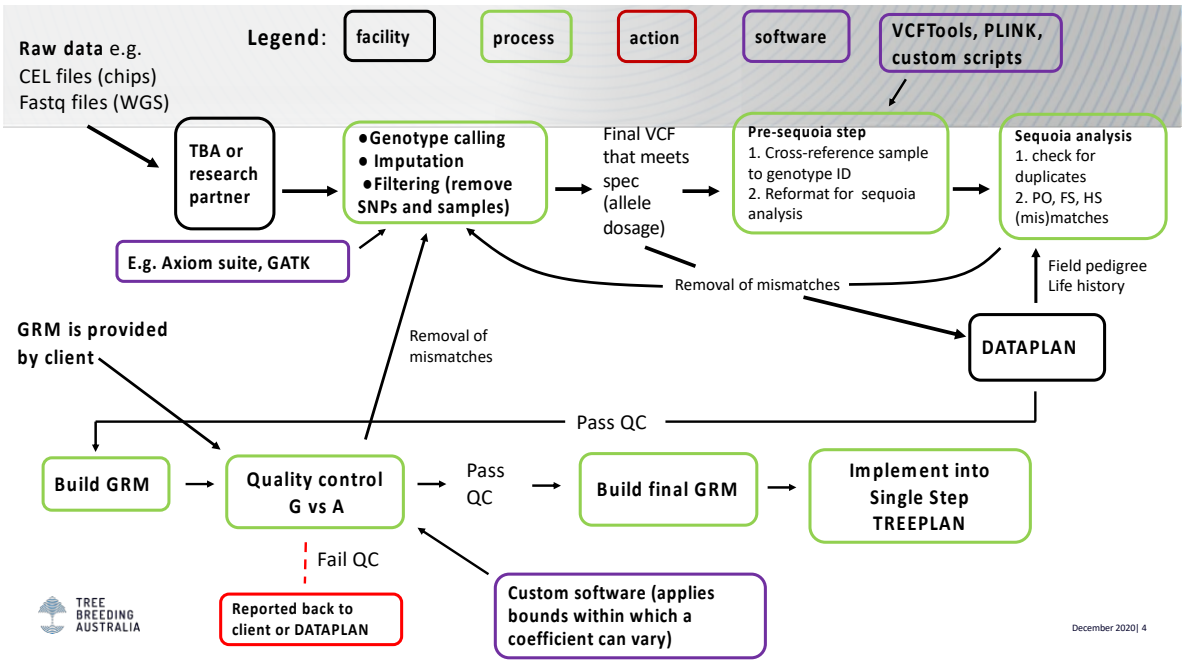


Figure 24 Flow chart describing the placement of pedigree error detection and recovery steps within the overall TREEPLAN single-step pipeline

The result of this process is the collection of genotype calls for the submitted samples, often stored in variant call format (VCF). It is probable that TBA must make provision for the storage of VCF files, or some equivalent format, in DATAPLAN. Variance call format has a very broad scope and can cover many types of assay data. An option discussed at this stage is for the TBA to accept VCF files that meet a certain specification (e.g. genotype calls are stored in the "allele dosage" format).

From discussions amongst the project personnel, it has become obvious that pedigree error detection and pedigree recovery will need to be undertaken outside the DATAPLAN framework. As is the case for imputation, transfer of these activities to within TBA computing infrastructure would be best done when the processes have been more fully worked through and finalised.

Pedigree forensics will occur at two stages within the overall TREEPLAN pipeline:

- Forensics applied at the level of SNP data using a software tool known as Sequoia, which occurs outside the DATAPLAN framework
- Forensics applied at the level of the **G** matrix, within the framework of DATAPLAN

SEQUOIA

SEQUOIA (Huisman, 2017) is a recently developed software package designed to turn information on hundreds or thousands of SNP into a multi-generational pedigree, using full-likelihood based methodology. It can be used to recover an unknown pedigree or purely as tool to flag “mismatches”, i.e. instances where a field-based pedigree does not agree with the pedigree inferred from the SNP data. Its core function however is to assign individuals as parents, when those individuals have been assayed. It can cluster half-siblings that share an unsampled parent and can assign grand-parents to half-sib ships.

At the core of the SEQUOIA software is the **Sequoia** function for running parentage assignment and full pedigree reconstruction.

- If no iterations are specified (MaxSibIter=0), the function only performs parentage assignment
- If one or more iterations are specified (MaxSibIter>0) it will attempt to find pairs of likely full- and half-siblings
- It then clusters the pairs into sibships, assigning a ‘dummy parent’ to each sibship
- It tries to replace dummy parents with genotyped individuals where possible

The **PedCompare** function in the SEQUOIA package is useful for comparing a field-based and genetically inferred pedigree. It identifies mismatches for those individuals which have genotyped parents assigned to them based on SNP data, that do not match the parents supplied from the field-based pedigree. Table 16 shows a snippet of output from the PedCompare function. In this example there are some full sibs showing a mismatch in terms of its assigned male parent. SEQUOIA works by creating dummy female and male parents and attempts to assign actual assayed individuals to these dummies. In this case there have been no matches. The “true” male parent must still not be assayed. SEQUOIA uses the codes “GD” and “GG” to denote the individuals is assayed and the parent is a potential dummy (GD) or is assayed (GG).

Table 16 Example of SEQUOIA output from the PedCompare function.

Id	FieldPED FP	FieldPED MP	Dummy FP	Dummy MP	Assigned FP	Assigned MP	FP cat	MP cat	Status of FP	Status of MP
14471	11847	9560	F0037	M0060	11847	nomatch	GG	GD	Match	Mismatch
14472	11847	9560	F0037	M0060	11847	nomatch	GG	GD	Match	Mismatch
14476	11847	9560	F0037	M0060	11847	nomatch	GG	GD	Match	Mismatch
14479	11847	9560	F0037	M0060	11847	nomatch	GG	GD	Match	Mismatch
14480	11847	9560	F0037	M0060	11847	nomatch	GG	GD	Match	Mismatch

It is possible to run SEQUOIA as a stand-alone FORTRAN program outside the R framework. This may be the desirable strategy to take if implementing SEQUOIA within the TREEPLAN pipeline as TBA are already accustomed to running FORTRAN executables in the pipeline. Also, when the data set becomes large (> 10,000 individuals) we may struggle to read the genetic data into R. Compiling the stand-alone FORTRAN with all the debugging options enabled will help us to understand where and why the program occasionally fails. Using either the R or standalone version within DATAPLAN would require SNP level data to be also accessible from within DATAPLAN.

GRM-NRM comparison tool

Comparing the constructed GRM with the NRM (limited to the assayed individuals, so it has the same dimensions as the GRM) is an alternative method for detecting mismatches between a field-based pedigree and a pedigree inferred from the SNP data. A custom PERL script has been written that performs this comparison, once FORTRAN programs have been used to construct both the NRM and

GRM. This tool could prove useful in situations where TBA has received only a constructed GRM from a 3rd party and does not have access to the SNP level data.

The tool should be run as a second stage QA process, once the first stage QA process using SEQUOIA has been completed, or in lieu of the first stage QA process, if TBA received a constructed GRM, rather than SNP level data.

A limited GRM-NRM comparison is performed, in the sense that only the following relationships are examined

- The female parent- and male parent-offspring pairings in a CP family
- All possible pairings among the full-sibs in the CP family
- The female parent-offspring pairing in an OP family
- All possible pairings among the half-sibs in the OP family

Hill and Weir have published useful articles on the variance expected in genomic relationships (Hill and Weir, 2011; Hill and Weir 2012). These papers develop theory to predict the variance in genomic relationship coefficients as a function of genetic map length, the number of chromosomes and the relational type (first, second, third degree relative etc). This theory is used to predict the expectations of variance in half- and full-sib relationships. In theory there is no variance in the genomic relationship between parent and offspring and should not deviate from 0.5. However, due to genotyping errors and the finite sampling of the genome, variance is observed. Simulation may be one way to derive what would be typical given an assumed genotyping error rate and sampling protocol.

An initial quality control step will include checking for unintentional duplication of samples, and for dubious SNPs (based on call rates, MAF, missingness etc) that have made it through from the genotype calling and imputation stage. An initial pedigree error detection step can also be undertaken that does not rely on iterating within the SEQUOIA function, which can be time consuming. We have now fully tested the SEQUOIA software for implementing these steps and recommend its operational application. There is now a prototype SEQUOIA “mini-pipeline” within the overall pedigree forensics pipeline. An important part of the SEQUOIA mini-pipeline is the extraction of field-based pedigree and life history data from DATAPLAN and the matching of sample IDS contained in the VCF file to DATAPLAN genotype IDs (genotype ID here meaning an “individual ID” and applies to both the ortet and its clonal replicates).

At this stage the plan is to have SEQUOIA check for mismatches between the field-based pedigree stored in DATAPLAN and the pedigree inferred from the SNP data (e.g. assigned parents in CP and/or OP families that are unlikely based on SNP data). We have devised algorithms that can determine whether the mismatch is the result of a family-based error (i.e. the parent has been misassigned leading to all assayed sibs in the family also surfacing with red flags), or a genotype-level error (i.e. the genotype has been mis-assigned and its assayed sibs are surfacing clean, indicating the assigned parents are correct). The offending individuals are then removed from any subsequent downstream processing (the construction of the GRM). With time and experience we can then begin to get more sophisticated and begin to explore how automatic changing of the DATAPLAN pedigree can be implemented.

Discussion

One of the highlight outcomes of the current work has been the success of the single-step analysis in *E. nitens*. Firstly, from the aspect of demonstrating the direct portability of the single-step methodology across species, and secondly, from the aspect of demonstrating the large improvements in accuracy possible with the technology. The target species for the development of the single-step methodology had been *E. globulus*. We anticipated very early in our planning that each species would have its own unique features and the platform for incorporating genomics had to be flexible. We identified single-step BLUP technology as that platform. Once we had it trialled in the pilot species it was seamlessly carried over to its use in *E. nitens*. We also recognised that we could not afford to get too distracted with specific SNP genotyping platforms, anticipating that in the future SNP genotyping technologies will change. TBA recognised that low and medium coverage whole genome sequencing, and SNP chips are both valid platforms for obtaining genomic data and they all have their advantages and disadvantages. We elected not to get trapped into building databases for storing DNA information at the locus level but focused only on the storage of genomic relationship coefficients. This strategy has served us well, in that TBA has been able to work efficiently and collaboratively with multiple providers including one 3rd party genomic services provider that does not wish to share individual locus information but can deliver accurate genomic relationships.

The single-step BLUP methodology, aided by a good working relationship between TBA, its member (Forico) and a 3rd party genomics services provider (Gondwana Genomics), has resulted in a 25% increase in accuracy for the main $NPV index that Forico uses to identify new selections. This is an outstanding result and provides an indication of what lies ahead for *E. globulus*. There are now over 12,000 individuals assayed in *E. nitens,* which is more than double what is available in *E. globulus* (~5,000). Though the results are not as strong in *E. globulus*, it is likely that doubling the number of assayed trees, and by refinements of the technology, we will soon see increases in accuracy similar to those obtained in *E. nitens*.

The story with *E. nitens* is constantly evolving. The Gondwana SNP panel is working extremely well and although there is no strong pressing need for Forico to consider an alternative SNP genotyping platform, TBA must be cognisant that other members may wish to do so. Because of the activities completed in this project TBA is well placed to proceed with a single-step analysis in *E. nitens* based on genomic data obtained from multiple platforms. TBA has access to a database of 32 million SNP that are segregating in the Australian *E. nitens* breeding program. From this database it can design a high-density assay that smaller assay sets (including the present Gondwana one) can impute up to. We have the materials in place to research and test imputation pipelines specific to this species. From experience gained in imputation development in *E. globulus*, TBA has learnt that large training sets are required to drive high imputation accuracy to the whole genome level. Thus, we anticipate a large training set will be required in *E. nitens*. Over 4,000 trees have been identified for this purpose.

While *E. nitens* remains in a wait and see position regarding **G** matrix construction using multiple SNP genotyping platforms, TBA has shifted its focus directly onto *E. globulus*. This NIFPI project has demonstrated that in principle the strategy will work. For the strategy to work there must be significant development of genomic resources, provided ideally through whole genome sequencing. Because these resources have been developed in *E. globulus,* TBA and AVR were able to quickly adapt to unforeseen issues and re-prioritise strategy. As a result of this project the Australian eucalypt breeding industry has access to the following: state of the art sequence alignment and variance discovery pipelines; fully tested imputation pipelines specific to various applications points; and SNP filtering tools. This advanced genomics tool kit allowed AVR to quickly provide TBA with an intersection between the SNP on a commercially available SNP chip (Euk72K) built using non-Australian genetic resources and SNP discovered *de novo* in the core pedigree of the Australian breeding population.

TBA was then able to quickly adapt to a position of demonstrating the consolidation of genomic data provided from two platforms into a single **G** matrix. The result of this consolidation was a substantial lift in the accuracy of EBV, relative to when applying an unconsolidated **G** matrix.

This is an excellent position to be in, while the Australian industry waits for more medium coverage whole genome sequencing (MWGS) of elite parents to come on-line. Additional MWGS is necessary to boost the size of the imputation reference training set to between 10 and 20 thousand trees. Once this training set size has been obtained, then consolidation in terms of imputing from low density chip sets to high-density SNP is likely to be realised.

It should also be stated that the correct modelling of provenance structure will play an important role in the national TREEPLAN analysis of *E. globulus*. This issue is less evident in *E. nitens.* Strong provenance structure may be one of several contributing reasons why we are seeing less gain in accuracy from using genomics in *E. globulus*. The actual quantum of genomic data is one other reason. Another is that in *E. nitens* breeding there is a predominance of open-pollinated crossing and the gain from using genomics is largely in part due to the recovery of paternal, and hence full-sib, relationships. A complex simulation study that attempts to model the underlying factors that cause provenance structure will be required to fully test the planned analysis technique of using meta-founders in the construction of the **H** matrix. Such a study was outside the scope of this NIFPI project. TBA will endeavour to fulfill its duty in providing a solution to the provenance problem in single-step analysis.

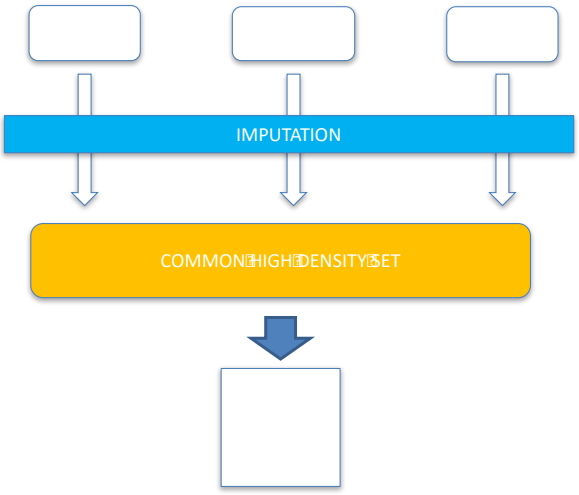


Figure 25 Different assays may use different marker sets and the lists of individuals may differ from assay to assay. Imputation is used to impute all individuals up to a common high-density SNP set.

In general terms, our strategy for allowing different providers, using different assays, to provide genomic data to the national genetic evaluation analysis of a species such as *E. globulus* has not changed from when we formulated it over 4 years ago (see Figure 25). Implementation of genomic breeding approaches in tree breeding operations requires the development of genotyping assays that are accurate, cheap, and high throughput. TBA is supportive of a culture that encourages some competition amongst providers to provide the most efficient and cheapest assay. The development and refinement of genotyping methods is crucial to realising genetic gain from genomics. Hence TBA fully expects assays to change with time. We believe this NIFPI project has provided TBA with the needed resources to future proof existing assay efforts and adapt to new ones. We have already shown that different assays can be combined, but, yet to fully realise high accuracy imputation from

the different, low-density assays to a common high-density set. TBA strongly recommends that continued genotyping using a WGS approach will serve to develop a training set of more appropriate scale to successfully drive this imputation. Low pass WGS in the blue gum program is currently very cost competitive compared to available chip platforms and returns a significantly higher information content and data value compared to genotyping with low density assays. In summary TBA is adhering to its stated position of using a **G** matrix computed using genotype calls made for a high-density SNP set and using imputation to fill-in those calls when the routine assay is based on a low-density set.

Another significant output of this NIFPI project has been the development of pedigree forensics as a routine practise in TBA operations. Two pedigree forensics pipelines are now in operation, one based on SNP level data and one based on comparing coefficients between the **G** and **A** matrices. The approaches are complementary to each other, and the latter approach provides a means to undertake forensics if SNP level data are not available. The pipelines have been applied in a separate NIFPI project in the Radiata Pine breeding pedigree and will be applied to the eucalypt programs once the requisite data is collated. The results obtained in *P. radiata* had a large and immediate impact, in that pedigree error detection has assumed a much greater role in that species improvement program. This has been flagged as an important follow up research focus.

Conclusions

- Single-step analysis is performing extremely well in *E. nitens* with up to 25% improvement in accuracy of predicting EBV for juvenile progeny. This should translate to a 25% increase in genetic gain.
- The indications are that similar gains can be made in *E. globulus* with the increased data from this project showing a substantial lift compared to earlier single-step runs with fewer trees. The quantum of assayed trees still needs to be substantially increased and problems in the analysis due to provenance structure still need to be resolved.
- TBA has not deviated from the position that **G** matrices based on high-density SNP sets is key for allowing different assays from different providers to be used in a common genetic evaluation analysis.
- The use of low-density, low-cost SNP genotyping assays, and the imputation of the assay results to larger SNP sets, are key for making genomic selection operational in tree improvement programs.
- A suite of imputation pipelines has been developed for operational use in Australian eucalypt breeding programs.
- Larger imputation reference panels are required for successful application of the imputation pipeline that imputes from low-density SNP chips to a high-density set.
- Initial versions of the pedigree forensics and recovery pipelines have been developed for use in eucalypt and non-eucalypt species as a result of this NIFPI project.

Recommendations

- A general presentiment within the group is that the development of an across species high-density (HD) SNP set, that will eventually be translated into a cost-effective chip or assay, is a sound idea. The target species would include, and not be limited to, the four main commercial species grown in Australia. The high-density (HD) SNP sets discovered in this project (EGLOB HD SNP sets 1 and 2, ENITEN HD SNP set 1) represent a good starting point for the definition of the final HD SNP set, but need further refinement. This refinement would be based on 3 areas of research:

    1. More complex filtering based on
        a. Checking Mendelian inheritance using trio and family data
        b. Sample and SNP based genotype probabilities
        c. Hardy-Weinberg expectations
    2. The idea of selecting “tag SNP”, which are SNP that specifically tag the diversity in the species. This is achieved by studying LD patterns and haplotype block structure. Tag SNP will impute more consistently and reliably that random SNP.
    3. The idea of targeting SNP that are “causal” via GWAS studies. The success in *E. nitens* is partly the result of the hottest 1000 projects that sought to identify associated SNP. We should try and replicate this result in *E. globulus*.

- Continued use of whole genome sequencing to develop the data sets needed to investigate all the above.
- The pattern of accuracy improvement (and decline) seen in *E. globulus* is intriguing and represents a challenge as to what may be happening. The project team would recommend a dedicated work package for untangling the potentially many confounding factors at play.
    - Addressing the Meta-Founder issue. The propagation of ancestral relationships through the pedigree may increase general relatedness in later generational progeny (which are the cohorts in which we are seeing a decrease in accuracy).
    - Examining if a high incidence of pedigree errors could be contributing to the problem. Perhaps later generations are not as related as we assume they are?
    - The scaling/correction of marker genotype calls via appropriate estimation of the population allele frequencies.
- A work package developing better computer simulation models that will aid in the preceding points and will help in elucidating better breeding strategy.
- We would recommend a consolidated project that would bring all major commercial species up to an even level in terms of genomic resources and training and reference populations. We feel Australia is in a unique position to lead the rest of the world in terms of operationalising genomic selection in forest tree breeding.

References

Alonge et al. (2019) RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biology 20: 224.

Browning, B.L., Zhou, Y. and Browning, S.R. (2018). A one-penny imputed genome from next generation reference panels. Am J Hum Genet 103: 338-348.

Browning, S.R. and Browning B.L. (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-1097.

Danecek, P. et al. (2021) Twelve years of SAMtools and BCFtools. GigaScience 10

Das et al. (2016) Next-generation genotype imputation service and methods. Nature Genetics 48: 1284-1287.

Dodds, K.G. et al. (2015) Construction of relatedness matrices using genotyping-by-sequencing data. BMC Genomics 16:1047

Fuchsberger, C. Abecasis, G. and Hinds, D.A. (2015) minimac2: faster genotype imputation. Bioinformatics 31:782-784.

Hill, W. and Weir, B. (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. 93:47-64.

Hill, W. and Weir, B. (2012) Variation in actual relationship among descendants of inbred individuals. Genet. Res. 94:267-274.

Howie, B, Fuchsberger, C., Stephens, M., Marchini, J. and Abecasi, G.R. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics. 44:955

Huisman, J. (2017) Pedigree reconstruction from SNP data: parentage assignment, sibship clustering. *Mol. Ecol. Resources*

Loh et al. (2016) Reference-based phasing using the Haplotype Reference Consortium panel. Nature Genetics 48: 1443-1450.

Rubinacci et al. (2021) Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nature Genetics 53:120-126.

Appendix 1

Table 17 Estimated mean raw sequence coverage per sample based on read numbers. Calculation is done assuming PE 150+150 sequencing and a genome size of 530 Mbp.

GenotypeID	AverageCoverage	GenotypeID	AverageCoverage	GenotypeID	AverageCoverage	GenotypeID	AverageCoverage	GenotypeID	AverageCoverage
Nid2088026	7.183	Nid2238180	10.74	Nid2087916	27.80	Nid2087575	10.71	Gid834803	6.53
Nid1201041	9.461	Nid2087547	7.79	Nid2087702	0.07	Nid1176439	5.46	Gid8999146	5.94
Nid2163193	11.455	Nid8907831	8.10	Nid2087669	10.02	Nid8907599	12.66	Gid128140	4.60
Nid1417028	14.492	Nid2087561	11.67	Nid8687197	13.84	Nid2087576	10.61	Gid1682325	4.60
Nid1198032	5.147	Nid1411492	6.59	Nid2087917	10.87	Nid2087590	7.80	Gid124391	4.59
Nid8907239	5.557	Nid8909972	10.30	Nid2087703	7.84	Nid8684215	9.48	Gid1682260	4.41
Nid8494032	6.205	Nid8907799	8.52	Nid2226016	16.56	Nid2087578	13.96	Gid422820	4.07
Nid2171890	15.096	Nid2204333	13.69	Nid2087704	8.58	Nid8494395	15.32	Gid124523	3.64
Nid2166124	1.156	Nid2167283	14.61	Nid1203596	8.09	Nid2227638	15.53	Gid2399077	3.42
Nid2088041	16.834	Nid2169230	8.37	Nid1176087	5.31	Nid2225092	8.18	Gid507378	3.20
Nid2087670	19.608	Nid2087564	7.79	Nid1178399	23.37	Nid2087612	5.14	Gid2399016	2.93
Nid2087657	6.299	Nid8683370	11.43	Nid2088056	5.87	Nid1201900	11.26	Gid127728	2.80
Nid8687185	17.571	Nid8493115	9.29	Nid2087919	11.09	Nid8494416	10.28		
Nid2088042	10.578	Nid2220456	14.20	Nid2087705	7.73	Nid2234436	9.39		
Nid2088029	15.795	Nid2179613	6.88	Nid1410825	4.66	Nid8687569	13.48		
Nid1195792	11.093	Nid2167519	6.42	Nid2087720	11.44	Nid2146875	7.55		
Nid2087672	2.707	Nid1411303	6.37	Nid2087707	7.53	Nid8907652	12.67		
Nid1176977	7.975	Nid8494383	9.01	Nid2087721	6.37	Nid2234452	12.40		
Nid8911830	11.614	Nid2087580	8.96	Nid2087708	5.93	Nid2208378	12.70		
Nid8910469	5.500	Nid8686517	4.41	Nid2170429	7.69	Nid2087843	12.49		
Nid8494249	6.593	Nid8683587	7.80	Nid1204673	5.81	Nid2087596	6.94		
Nid2171225	6.739	Nid2087581	9.28	Nid1174653	8.68	Nid1204814	15.45		
Nid1205057	4.616	Nid2087603	7.75	Nid8684327	11.10	Nid1178117	11.34		
Nid2088045	15.276	Nid1176233	10.69	Nid2087724	8.32	Nid1198206	8.05		
Nid1418621	13.560	Nid8907627	9.05	Nid1179051	12.18	Nid2087844	11.53		
Nid8907946	6.367	Nid2087831	13.48	Nid8911389	7.82	Nid8907868	12.21		
Nid8685125	8.373	Nid8685709	5.00	Nid8686000	16.47	Nid2087618	5.09		
Nid2226235	12.039	Nid2087834	8.13	Nid2087740	5.42	Nid2087846	3.21		
Nid2171916	26.591	Nid2087607	5.14	Nid2087727	17.96	Nid1175589	6.07		
Nid2088047	16.172	Nid1195514	6.46	Nid1179280	11.97	Nid2087860	9.79		
Nid2087676	7.451	Nid1198210	4.47	Nid1195420	8.99	Nid2087847	10.49		
Nid1416463	6.841	Nid8684893	5.08	Nid8684559	6.10	Nid8686122	7.36		
Nid8687665	8.777	Nid2087835	8.19	Nid2226728	13.72	Nid2226387	17.57		
Nid2166799	9.082	Nid2087621	15.58	Nid2225682	8.56	Nid2224075	8.22		
Nid2088048	16.476	Nid2087608	8.61	Nid2087514	8.65	Nid2087848	13.15		
Nid2087677	15.699	Nid8907431	8.90	Nid8687029	7.04	Nid2087849	10.91		
Nid2219012	11.420	Nid2087836	5.60	Nid2087515	7.01	Nid2221793	8.96		
Nid2199568	7.503	Nid2087622	1.33	Nid2087970	23.54	Nid1416637	5.02		
Nid2087691	8.468	Nid2087589	8.21	Nid2087743	23.76	Nid1413513	8.40		
Nid1175702	25.227	Nid2087850	11.84	Nid2087516	8.01	Nid1411600	18.85		
Nid2087693	10.433	Nid8909573	6.79	Nid1178244	9.21	Nid8683429	13.84		
Nid2087714	23.406	Nid2225779	15.92	Nid8685634	4.35	Nid2165960	7.10		
Nid2087694	12.241	Nid8907648	8.39	Nid2087530	9.36	Nid2088021	11.85		
Nid1198531	6.737	Nid8686093	9.41	Nid2087517	9.35	Nid1203549	6.94		
Nid8494518	8.992	Nid2167559	10.26	Nid2087959	10.68	Nid2180140	10.79		
Nid2087716	6.159	Nid2087852	9.50	Nid2087531	13.18	Nid2087865	7.97		
Nid2087696	6.906	Nid2087839	9.60	Nid2087518	9.69	Nid2087638	14.60		
Nid8687030	11.384	Nid2087625	9.54	Nid1179285	7.85	Nid1197430	6.65		
Nid2087944	8.426	Nid8687153	10.59	Nid8909254	8.10	Nid2088024	6.08		
Nid2087717	9.607	Nid2166975	6.33	Nid2161831	7.88	Nid2170147	6.43		
Nid2087697	8.008	Nid2088010	153.63	Nid2087746	11.29	Nid2088025	8.72		
Nid1196222	7.820	Nid8686997	6.38	Nid2087532	8.64	Gid829523	103.50		
Nid8686433	11.217	Nid2087854	16.01	Nid2087519	7.08	Gid821569	36.86		
Nid2087731	12.064	Nid2237663	9.06	Nid1204029	8.09	Gid2652694	31.95		
Nid2087718	8.741	Nid2087641	5.47	Nid2225260	11.73	Gid2398606	31.23		
Nid2087732	8.357	Nid1203092	6.72	Nid2087760	11.10	Gid422857	26.04		
Nid2087719	7.850	Nid2165952	13.17	Nid1178709	6.52	Gid269992	23.23		
Nid1175937	7.659	Nid2162980	7.81	Nid8910343	279.83	Gid1986176	20.14		
Nid8911417	6.733	Nid2087642	9.35	Nid2087761	21.63	Gid2652574	19.23		
Nid2167227	14.696	Nid1420896	10.63	Nid1179288	26.12	Gid833933	19.03		
Nid2087734	9.119	Nid1177331	10.15	Nid1178723	10.60	Gid2652365	18.68		
Nid2087520	7.727	Nid8910452	10.00	Nid1179536	8.79	Gid422447	17.77		
Nid2087962	4.328	Nid2087857	3.95	Nid2087977	10.16	Gid422520	17.68		
Nid2087521	11.190	Nid2087643	9.36	Nid2188448	5.19	Gid423234	17.53		
Nid2235157	7.851	Nid2087644	8.27	Nid1197770	12.72	Gid828128	17.20		
Nid1178711	10.880	Nid1176488	9.90	Nid1178039	8.46	Gid821576	16.74		
Nid8683983	9.767	Nid2201486	1.66	Nid8907789	11.54	Gid2398599	16.27		
Nid2087737	0.036	Nid2175221	17.02	Nid8687527	12.49	Gid423129	15.86		
Nid2087524	9.103	Nid2088016	6.86	Nid2087539	8.80	Gid1987479	15.36		
Nid2187623	12.773	Nid1177989	6.63	Nid8683131	0.15	Gid422916	15.10		
Nid2087525	8.899	Nid2088030	12.71	Nid2166448	8.10	Gid829522	14.80		
Nid1175956	5.151	Nid2088017	9.56	Nid8908250	9.70	Gid423141	14.66		
Nid2087980	8.544	Nid1176523	8.27	Nid8683119	13.30	Gid422442	14.52		
Nid1416066	6.086	Nid8908804	11.97	Nid2219316	23.63	Gid833904	14.04		
Nid1196933	7.384	Nid2171900	24.63	Nid8685445	3.06	Gid833743	13.32		
Nid2087754	11.302	Nid2088019	7.59	Nid8683574	17.47	Gid423047	13.27		
Nid2087540	205.037	Nid2087648	10.04	Nid2087769	7.43	Gid8999285	13.12		
Nid2087527	8.800	Nid1200910	10.12	Nid2087555	8.50	Gid833814	12.94		
Nid8911892	7.506	Nid8494024	173.14	Nid1204065	12.21	Gid834442	12.46		
Nid2170709	6.638	Nid2154325	7.57	Nid2168563	8.92	Gid834270	11.99		
Nid2168823	9.079	Nid2087890	8.09	Nid2087998	32.05	Gid833493	11.43		
Nid2087541	8.173	Nid1704465	8.86	Nid2087570	9.13	Gid2399031	10.92		
Nid2087528	8.849	Nid2087664	11.33	Nid2087557	8.34	Gid422816	10.20		
Nid1201796	6.529	Nid1413067	8.12	Nid1416999	7.71	Gid833572	9.94		
Nid1196260	6.852	Nid8493215	9.34	Nid2236726	11.97	Gid833494	9.80		
Nid8683548	11.624	Nid2238071	11.32	Nid2087999	6.93	Gid1986057	9.49		
Nid2087757	6.205	Nid2088036	17.38	Nid1176662	10.69	Gid422631	9.47		
Nid2087543	8.610	Nid1202090	8.70	Nid2087559	9.83	Gid2398686	9.30		
Nid2087544	10.660	Nid8908576	10.11	Nid1201406	6.24	Gid8999525	9.08		
Nid2235881	10.509	Nid2219000	16.00	Nid2087807	8.46	Gid2653385	8.54		
Nid2223408	7.502	Nid2154328	0.04	Nid2087573	8.38	Gid124284	8.41		
Nid2187643	7.920	Nid2087913	7.14	Nid1200361	9.68	Gid9000590	8.40		
Nid2087986	3.307	Nid2087700	9.09	Nid1196065	5.49	Gid8999654	8.27		
Nid2087545	8.688	Nid2088052	11.83	Nid2187231	9.00	Gid2650950	7.87		
Nid8910355	5.895	Nid2087701	6.51	Nid8493566	9.77	Gid9000560	7.61		
Nid2087546	12.549	Nid2087681	11.28	Nid2166697	6.34	Gid123690	7.35		

Appendix 2

Table 18 Mean EBV accuracies for VOL_WA, VOL_TAS, VOL_GIPPS, VOL_AUST and percentage change in the mean (%), when using the H (based on the 2021 GRM) and A matrices in the mixed model equations (in E. globulus).

VOL_WA												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.87	0.84	3.33	0.65	0.54	19.30	0.73	0.72	0.77	0.65	0.65	0.08
Gen-1	0.85	0.83	2.19	0.71	0.66	8.06	0.70	0.70	-0.45	0.69	0.69	-0.35
Gen-2	0.79	0.83	-4.51	0.76	0.79	-4.80	0.82	0.83	-1.56	0.80	0.79	0.88
Gen-3				0.69	0.71	-2.26				0.79	0.79	1.15
VOL_TAS												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.87	0.84	3.35	0.65	0.54	18.58	0.73	0.73	0.68	0.63	0.63	0.05
Gen-1	0.85	0.83	2.00	0.72	0.67	7.24	0.72	0.72	-0.53	0.70	0.70	-0.45
Gen-2	0.77	0.81	-4.94	0.75	0.79	-4.88	0.82	0.83	-1.46	0.80	0.79	0.84
Gen-3				0.69	0.71	-2.41				0.77	0.77	0.76
VOL_GIPPS												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.85	0.82	3.09	0.64	0.55	16.96	0.73	0.72	0.68	0.64	0.64	0.06
Gen-1	0.82	0.81	1.82	0.70	0.65	7.33	0.71	0.71	-0.67	0.69	0.69	-0.39
Gen-2	0.76	0.81	-6.01	0.73	0.77	-4.51	0.81	0.82	-1.71	0.78	0.77	0.78
Gen-3				0.68	0.69	-2.77				0.75	0.75	0.27
VOL_AUST												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.92	0.90	2.07	0.75	0.68	9.89	0.82	0.82	0.41	0.75	0.75	0.03
Gen-1	0.90	0.88	1.42	0.80	0.77	4.48	0.80	0.80	-0.37	0.79	0.79	-0.26
Gen-2	0.85	0.88	-2.99	0.83	0.85	-2.80	0.87	0.88	-0.99	0.86	0.85	0.55
Gen-3				0.78	0.79	-1.29				0.85	0.84	0.61

Table 19 Mean EBV accuracies for VOL_WA, VOL_TAS, VOL_GIPPS, VOL_AUST and percentage change in the mean (%), when using the H (based on the 2020 GRM) and A matrices in the mixed model equations (in E. globulus).

VOL_AUST												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.91	0.90	0.97	0.65	0.63	3.25	0.82	0.82	0.09	0.75	0.75	-0.02
Gen-1	0.89	0.88	0.18	0.77	0.76	1.29	0.80	0.80	-0.04	0.79	0.79	-0.07
Gen-2	0.86	0.88	-1.97	0.76	0.76	-0.85	0.88	0.88	-0.08	0.85	0.85	0.05
Gen-3				0.11	0.11	-0.39				0.84	0.84	0.02
VOL_WA												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.86	0.84	1.58	0.53	0.50	6.18	0.73	0.72	0.16	0.65	0.65	-0.05
Gen-1	0.83	0.83	0.28	0.67	0.65	2.36	0.70	0.70	-0.05	0.69	0.69	-0.12
Gen-2	0.81	0.83	-2.93	0.70	0.71	-1.41	0.83	0.83	-0.14	0.79	0.79	0.05
Gen-3				0.10	0.10	-0.71				0.78	0.78	0.05
VOL_TAS												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.86	0.84	1.58	0.53	0.50	6.24	0.73	0.73	0.14	0.63	0.63	-0.08
Gen-1	0.83	0.83	0.24	0.68	0.67	2.15	0.72	0.72	-0.09	0.70	0.70	-0.15
Gen-2	0.79	0.81	-3.06	0.70	0.71	-1.43	0.83	0.83	-0.16	0.79	0.79	0.04
Gen-3				0.10	0.10	-0.67				0.76	0.76	-0.05
VOL_GIPPS												
	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.84	0.82	1.48	0.53	0.50	5.82	0.72	0.72	0.16	0.64	0.64	-0.03
Gen-1	0.81	0.81	0.33	0.66	0.65	2.27	0.71	0.71	-0.06	0.69	0.69	-0.11
Gen-2	0.78	0.81	-3.20	0.68	0.69	-1.32	0.82	0.82	-0.18	0.77	0.77	0.06
Gen-3				0.10	0.10	-0.68				0.74	0.74	-0.15