Final Report
Project NS038

NATIONAL INSTITUTE FOR
FOREST PRODUCTS INNOVATION

Development of state-of-the art genomic resources for pine breeding to enable single-step genomic selection

2022

Mount Gambier Centre
Funded by the Australian Government, South Australian Government & Industry Partners.
nifpi.org.au

NATIONAL INSTITUTE FOR
FOREST PRODUCTS INNOVATION
MT GAMBIER

Development of state-of-the art genomic resources for pine breeding to enable single-step genomic selection

Prepared for

National Institute for Forest Products Innovation

Mount Gambier

by

Dr Richard Kerr, Dr Josquin Tibbits, Dr Ed Breen,
Prof Hans Daetwyler, Dr Tony McRae

Publication: Development of state-of-the art genomic resources for pine breeding to enable single-step genomic selection

Project No: NIF101-1819 [NS038]

IMPORTANT NOTICE

Whilst all care has been taken to ensure the accuracy of the information contained in this publication, the National Institute for Forest Products Innovation and all persons associated with it (NIFPI) as well as any other contributors make no representations or give any warranty regarding the use, suitability, validity, accuracy, completeness, currency or reliability of the information, including any opinion or advice, contained in this publication. To the maximum extent permitted by law, FWPA disclaims all warranties of any kind, whether express or implied, including but not limited to any warranty that the information is up-to-date, complete, true, legally compliant, accurate, non-misleading or suitable.

To the maximum extent permitted by law, FWPA excludes all liability in contract, tort (including negligence), or otherwise for any injury, loss or damage whatsoever (whether direct, indirect, special or consequential) arising out of or in connection with use or reliance on this publication (and any information, opinions or advice therein) and whether caused by any errors, defects, omissions or misrepresentations in this publication. Individual requirements may vary from those discussed in this publication and you are advised to check with State authorities to ensure building compliance as well as make your own professional assessment of the relevant applicable laws and Standards.

ISBN:978-1-922718-01-3

Researcher/s:
Dr Richard Kerr, Dr Tony McRae
Tree Breeding Australia Limited
PO BOX 1811 Mount Gambier, SA

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, Fisheries and Forestry and the Government of South Australia in partnership with the University of South Australia.

Australian Government
Department of Agriculture, Fisheries and Forestry

Government of South Australia
Department of Primary Industries and Regions

Executive Summary

Breeding has traditionally been one of the main activities used to maximise fibre production. The adoption of genomics into Australian Pine breeding programs will dramatically improve the gains made from breeding. This project formed the important first steps on the path to adoption, namely in initiating a *de novo* genome assembly, building a foundational genomics data set that characterises the Australian breeding germplasm diversity, in rolling out the first DNA assay to be used routinely in conifers in Australia, and to initiate pipelines that will correct historical and current mistakes made in the definition of breeding program pedigrees. These innovations will allow the local plantation growing sector to be internationally competitive.

Building genome assemblies was identified as a key priority in the Forest and Wood Products Australia commissioned research review for 2019-2024 for the forest industries. Assemblies are important for aligning sequence data generated for the purposes of SNP discovery. The project has initiated a *de novo* assembly of the *P. radiata* genome with American based Dovetail Genomics, in preference to accessing an assembly undertaken by New Zealand; a strategy that will result in much greater intellectual freedom for Australian researchers.

For building the foundational genomics data set the project identified over 3000 founder trees contributing genes to the breeding program. The top 200 of these contributed over 85% of genes and in this project, we successfully located and sampled megagametophytes from 118 of these founder trees, representing over 65% of the genes in the breeding program. These high contribution founders were assayed using whole genome sequencing of gDNA extracted from either 8 or 4 haploid megagametophytes (depending on contribution) generating a comprehensive data set which represents more than half the diversity in the breeding program. The generated data set will be useful in numerous applications including in the development of high- and low-throughput genotyping assays that adequately represent the Australian breeding program diversity.

This project tested a low-density array developed and made accessible by a New Zealand collaboration between government, RPBC and SCION. This application showed that most features (SNP targets) are polymorphic in the Australian breeding population. A follow up consignment of 768 samples assayed (960 samples in total have been assayed) created a small trial data set which was used to trial a single-step genomic selection run in *P. radiata*. While EBV accuracies did not noticeably increase, which is not unexpected given the modest amount of genomic data supplied to the analysis, the trial run did demonstrate that the implementation of single-step analysis methodology into TREEPLAN in previous projects is directly translated to the radiata pine program.

This project also implemented the pedigree error detection and recovery pipeline developed in the parallel NIFPI project NIF111-1819 which was focused on eucalypts. This pipeline discovered a substantial number of mismatches between the recorded field-based pedigree and the pedigree inferred from the SNP data. This finding points to potential systematic problems in maintaining the identity of individuals, and their links to parental identifiers, across the 60 odd years of tree breeding and the many points of transfer of genetic material (grafts to arboreta, pollen applied to cones, seed extraction, seed transferred to nurseries, plants transferred to trial sites etc). While pedigree errors are to be expected in a long running breeding program this finding has led TBA to develop a strategy for fixing historical errors and for improving pedigree recording which should substantially improve EBV accuracy. This strategy will enable identification and isolation of pedigree errors and to their eventual correction. This process will likely lead to a deeper understanding of the causes of pedigree errors and offer a new avenue to drive continual improvement in breeding operations.

Overall, the project has given the Australian radiata pine industry some immediate and practical deliverables. These include the inclusion of any available genomic data into routine genetic evaluation. This inclusion occurs without any disruption to current practises. This lifts the onus off the operational breeders as they do not need to worry about how they are to incorporate genomic information. Another immediate benefit is the appropriation of the genomic information to correct mistakes made in recording breeding program pedigrees and to audit the identities of elite breeding material. This will result in immediate lifts in the rate of genetic gain of between 15 and 20%.

COVID IMPACTS:

This project was significantly disrupted by COVID restrictions imposed at various times in Victoria and nationally through 2020 and 2021. These restrictions caused significant delays to scheduled laboratory works which delayed delivery of the main data sets. Despite these delays the main objectives of this project have been met with generation and delivery of the data sets achieved before the project end date.

Table of Contents

Introduction

Tree Breeding Australia's key objective in the genomics space is to operationalise single-step genomic prediction (Legarra *et al*., 2009, Christensen and Lund, 2010). The single-step procedure is not disruptive and allows traditional sources of information (pedigree and phenotypic measurements on all trees) to be used in conjunction with DNA level data that may be only gathered on small subsets of trees. In terms of its operationalisation for the *P. radiata* breeding program, a first critical step is the complete characterisation of the Australian breeding germplasm. This will underpin all future genomics works and will ensure current and future high-throughput genotyping assays can adequately capture this characterisation. The following objectives are within the context of the wider objective.

1. Build foundational datasets upon which genomic selection can be implemented. These datasets include a genome wide SNP discovery dataset and a training data set obtained by genotyping the core pedigree of the TBA breeding population with a high-density assay.
2. Build on existing core methodologies and workflows needed to implement genomic selection in an operational setting (revise solving algorithms, imputation methodologies and methodologies for undertaking pedigree forensics).
3. Implement genomic selection at an operational scale in collaboration with industry partners (e.g. breeders, growers, deployment managers). The validation of the New Zealand SNP chip, or a recommendation to develop our own low-density array, for use in routine assays is the main outcome in this objective. The validation will involve the demonstration of the imputation step where genotypes for the high-density assay are imputed from the results of the low-density assay.

This NIFPI project has enabled TBA and its research partner, Agriculture Victoria Research (AVR) to focus on activities primarily associated with objective 1. It is stressed that a complete operationalisation of single-step genomic prediction in radiata pine and other conifer species will span multiple research partitions. It was not the intention of this project to genotype the core pedigree of the TBA breeding population, but only to gather the genomic resources necessary to complete this activity in a future project. Part of this "gathering" step was to undertake a thorough analysis of the breeding program genetics and to identify founder and high value progeny which are to be the basis of major collections of foliage samples from thousands of individuals. These collections will be sourced from across the national estate of trials and arboreta and from seed stores with collections stored in Mount Gambier. These collections will be maintained as ready for DNA extraction and processing at a future date.

This project (NIF101-1819) has been running alongside its sister project NIF111-1819 'Implementation of single-step genomic selection in eucalypts'. The sister project had more of a deliberate focus on development of imputation technologies and pedigree forensics. The forensics pipeline, once developed, was then applied to the available SNP assay data in radiata pine with a substantial number of pedigree errors in the national database discovered. The errors were more prevalent in one of the trials sampled. This finding prompted a thorough testing of the forensics pipeline and a more rigorous checking of the software via in-silico simulation. What was initially thought to be a minor issue became a higher priority and a significant portion of this final report details the findings.

There are two points that need to be discussed in relation to a shift in the project strategic plan. Firstly, a project outcome was to test the SNP chip developed by SCION by assaying a small cohort of trees. The positive outcome of this test led to a second round of assays on a larger cohort of individuals. Secondly, the project application had outlined a strategy for gaining early access to the New Zealand funded reference genome assembly for radiata pine. The strategy entailed us coordinating the collection of Hi-C data via an American genomics company (Dovetail Genomics). This data was then to aid the chromosome level assembly of a reference genome for radiata pine. Access to this level of assembly would have aided us in our SNP discovery work. However, despite extensive effort, ongoing intellectual property issues between the New Zealand partners made this strategy impossible to proceed with within the project timeline. The resources allocated to this activity were redirected into the assaying of approximately 780 individuals with the NZ SNP chip. Regarding the loss of a reference genome to aid us in our SNP discovery, we sought another option. Namely to align our generated sequence data to the *Pinus taeda* V2.0 reference genome.

Methodology

The project was structured by aligning activities and deliverables to objectives 1 and 3 defined in the Introduction.

Objective 1 – build foundational data to underpin genomic selection implementation

The first activity was updated during the project and was changed from providing a reference genome Hi-C dataset to completing

1. **Step 1 in development of a radiata Pine genome assembly**
    a. Engage Dovetail to deliver agreed data and analysis

The second activity was to collect all material necessary to build the foundational genomics dataset.

2. **Radiata pine breeding program sample collections**
    a. Identify and sample seeds from key founder individuals (Founder Collection)
    b. Identify and sample foliage from the core pedigree of the breeding population (Parent and Key Progeny Collections)

The Parent and Key Progeny Collections were made across the estate trial network and involved considerable in-kind support from the industry partners. DNA has not been extracted from the collections at this point. With historic material (parents of first- and second-generation progeny) becoming increasingly harder to recover and with many first- and second-generation progeny trials reaching maturity, it was considered prudent to make these collections a key activity in this project. Hopefully they will not need not be repeated in future years as they are tedious and costly compared to nursery-based collections that will be the basis of operational genomics. Once the founder collection was completed the plan was to extract mega-gametophyte tissue from the seeds of each founder and extract DNA from the tissue for whole genome sequencing. Megagametophyte tissue is a maternal nutritive tissue and is haploid and has been selected for sequencing as the haploid signal can be effectively used in data analysis and SNP (variant) discovery. The activities were the

3. **Generation of ~0.8x raw whole genome sequencing coverage for megagametophytes sampled from the Founder Collection trees**
4. **Analysis of this genomic data set for variant discovery (deferred due to COVID delays)**

The third objective was:

Objective 3 – Implement genomic selection at an operational scale

Toward objective 3 we wanted to test the newly available SNP chip developed in New Zealand, as a potential vehicle for providing a low/medium-cost, low-density array, for use by the Australian industry.

5. **Testing of an industry standard low-cost, low-density SNP array**

The development entailed the assaying of an initial consignment of 192 samples with the NZ-based chip. A positive outcome of this initial testing led us to consider a second consignment of 768 samples. With close to 1000 individuals assayed, including assays of many founder parents, additional work packages were added to make use of these data sets and to substitute for the deferment of activity 4 above. This included applying the recently developed pedigree forensics pipeline to the SNP assay results and use of these data sets in a single-step TREEPLAN run to demonstrate the applicability of this pipeline developed in earlier projects.

For the pedigree pipeline a thorough testing was required as, in its development several issues arose. These included making informed decisions regarding the number and type of SNP to be used in the pipeline; testing how efficient the software was in detecting pedigree errors of different types, and how well the pipeline could recover true parentage given missing data, and the false positive/negative rates. An in-depth simulation study was undertaken to address these questions.

6. **In silico analysis of the pedigree forensics pipeline**
    a. An introduction to SEQUOIA
    b. An introduction to **G**-**A** matrix comparison
    c. Testing by simulation

After this testing phase we put the SNP assay results to work by firstly applying them to the pedigree forensics pipeline. Secondly, by trialling a single-step TREEPLAN analysis that incorporates a genomic relationship matrix (GRM or **G** matrix) based on the genotype calls made with the SNP chip.

7. **Putting the SNP chip assay results to work**
    a. Running SEQUOIA
    b. Building a **G** matrix
    c. Checking the **G** against the **A** matrix
    d. Running a single-step analysis

Results

Step 1 in development of a radiata Pine genome assembly

Engage Dovetail to deliver agreed data and analysis

Dovetail Genomics was engaged to provide services to undertake a first phase assembly of the radiata pine genome under project; ID: DEP2874 Monterey Pine, *Pinus radiata*, Proximity Ligation + Scaffolding arising from the quote Q-03770. These services include construction one Omni-C library per 3 gigabases of the organism's genome (*P. radiata* genome size is ~ 21 Gbp). Dovetail will scaffold the draft assembly (minimum N50 of 100kb is required) through the HiRise software pipeline using the proximity ligation data and assess library quality through sequencing ~2M PE75bp reads and mapping these data back to the draft assembly. The total run time for this project will be approximately 52 weeks from the receipt of the sample, which was shipped from TBA in June 2021. These services were provided at a significant discount of 47.6% off the listed retail price.

HiRise Assembly deliverables include:

- The HiRise assembly in FASTA format
- A report summarizing key assembly statistics, features of the proximity ligation library, and a linkage density plot of the proximity ligation library data
- A table detailing the breaks made to the input scaffolds
- A table describing the position of the input assembly scaffolds within the final HiRise scaffolds
- BAM file(s) containing alignments of the proximity ligation library read pairs mapped to the draft assembly

All HiRise deliverables will be shared with TBA upon delivery to AVR.

Radiata pine breeding program sample collections

Founder collection

Our goal is to generate a compendium of breeding diversity based on complete genome characterisation of the founder trees of the national TBA breeding population. To identify the key founders in the TBA radiata breeding population, we computed the "contribution" matrix. This matrix contains the fraction of genes that each founder has transmitted to a descendent. In this case the descendants we targeted were the named, 2nd generation individuals, because they represent the current cohort of breeding parents. Manipulation of the entries allows us to rank founders on their total contributions to this cohort and to determine the percentage of the genetics that we can account for. The ranking of founder contributions is shown in Figure 1. Well known genotypes such as 'NZ850-055' and 'A12038' top the rankings and these two genotypes alone account for approximately 10% of the genetics in the breeding program. In total around 3000 founding trees were identified.

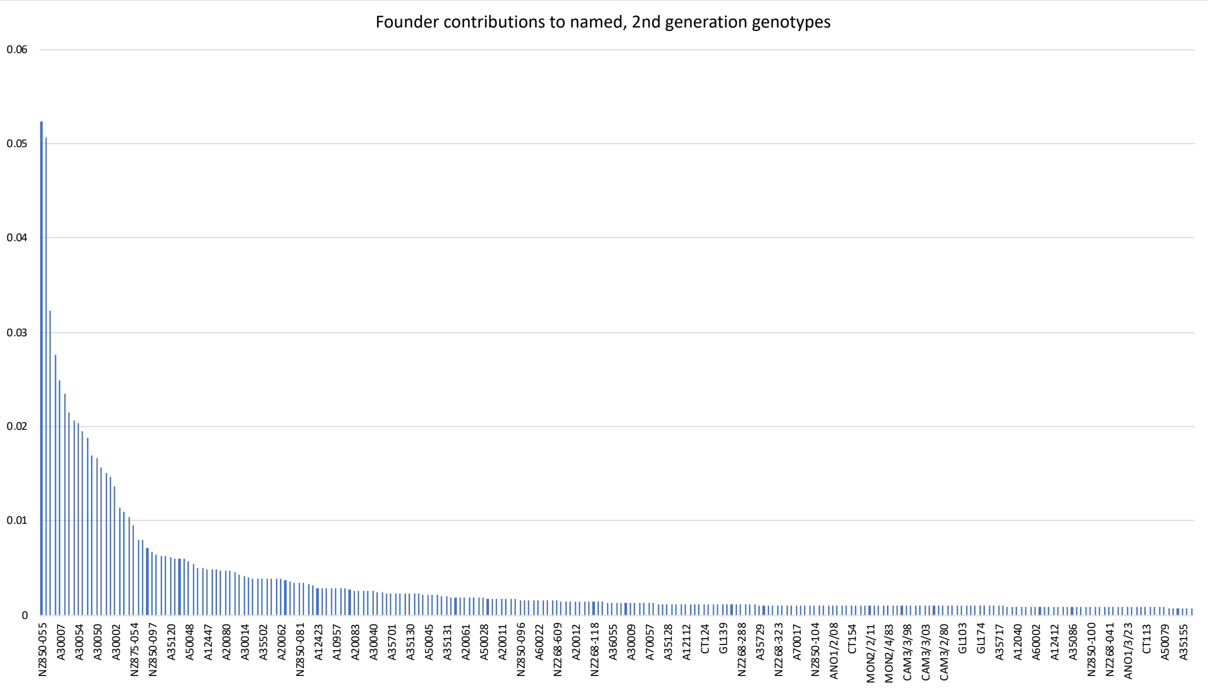

Figure 1 Important founders, sorted by their fractional contributions to named, 2nd generation genotypes in the national P. radiata breeding population

The top 30 ranked founders account for approximately 50% of the genetics in the program, while the top 100 ranked founders account for approximately 73% of the genetics. After exhaustive searches of seed stores, the TBA, its members and collaborators were able to retrieve seed samples from 25 of the top 30 ranked founders, and 65 of the top 100 ranked founders, which combined account for 60% of the genetics in the program. The founder collection includes another 53 trees from outside the top 100 ranked founders. The total of 125 trees in the collection account for ~65% of the genetics. There remains a possibility of recovering seeds from two top ranked, NZ originating founders (within the top 34 ranked founders). A substantial number of other founder trees were also identified as available in arboreta and trials, however, none were cone bearing and these have been earmarked for collection in a few years once new cones have time to develop and ripen.

Identify and sample foliage from the core pedigree of breeding population

Identification and sampling foliage from the core pedigree is a sensible first step to undertake in the implementation of genomic prediction in a breeding program. The core pedigree provides cross program and cross generational connectivity amongst genotyped trees and assists with propagation of information throughout the program and with correct pedigree recovery. It will also underpin a planned future research partition where a high-density (200-800 thousand SNP) assay will be applied to the core pedigree of the TBA breeding population, namely all parents of controlled-pollinated (CP) crosses and one progeny from each cross. This assay will help us to understand the underlying structures (e.g. major haplotype blocks, founder variation) specific to the TBA radiata breeding population. Such knowledge will also future proof deployment of single-step genomic selection in radiata pine. Our own high-density assay will also allow us to deploy any future, low-cost, low-density assay developed by overseas genomics service providers. The core pedigree will be represented by the founder collection and two further sample collections: the Key Progeny and the Parent Collections.

Key Progeny Collection

The Key Progeny collection aims to sample widely across the diversity of the program targeting 'high value progeny' (about 2,000). We define 'high value progeny' as progeny with observations measured across all traits and site types, and the progeny needed to be "high-value" in the sense they have been measured for at least 4 traits. As a first step we sampled all grafted genotypes in the NGRC breeding arboretum, as this was an efficient means to sample many high value progeny. These genotypes have been sourced from all site types, have been measured for multiple traits, and are current candidate parents for breeding. Table 1 summarises the Key Progeny collection and shows that approximately 500 genotypes have now been sampled from the NGRC. The collection was then augmented by deliberately sampling more recent progeny trials across site types: TAS, WA, MVAL/NSW, CGIPP, CVIC. It was also important that TBA sampled progeny derived from crosses between native land race material with TBA breeding population parents. One hundred and fifty samples from this type of material was added to the collection.

Table 1 Trials sampled for the Progeny Collection to date

Trial	Site-Type	Count
NGRC	All site types	497
Caroline (BRGT1301)	GTR	226
Connorville (BR0801)	TAS - low elevation	198
Moogara (BRGT1304)	TAS - high elevation	204
Bundaleer (BRGT1403)	MVAL/NSW	150
Jarrahwood (BRGT1404)	WA	166
Mt Mercer (BRGT1302)	CGIPP	206
Heywood's (BRGT1303)	CVIC	206
Native land race hybrids	All site types	150
	TOTAL	**2000**

Parent Collection

The Parent Collection aims to include foliage samples taken from all parents used in the breeding program. As a first pass collection, we targeted both parents of any individual in the Key Progeny Collection. Many individuals in the Progeny Collection which are now located in the NGRC were crossed more than two decades ago and it is becoming increasingly difficult to locate the parents for such individuals. To undertake this collection TBA and its members regularly met via virtual conferencing to discuss the sourcing of hard-to-find parents.

Table 2 summarises the results of these efforts. Some parents were available in the older facilities in South Australia such as the Walshes breeding arboretum and the Glenburnie seed orchard. Hancock Victorian Plantations (HVP) were able to locate many of the historic parents in their facilities. Forest Products Commission (FPC) have some parents in their facilities. Many parents of progeny in the newer trials such as Heywood's, Mt Mercer, Bundaleer and Jarrahwood have been cloned into the NGRC. This point demonstrates that there is cross-over between the various collections. There are individuals in both the Parent and Key Progeny collections and there are individuals in both the Founder and Parent collections. Because of the high value of this collection TBA and its members will continue to source parents that have not yet been sampled.

Table 2 The Parent Collection: - a summary

Facility	Number sampled
Walshes breeding arboretum, SA	39
Glenburnie seed orchard, SA	4
Various HVP facilities in Victoria	73
Various FPC facilities in WA	12
Key Progeny in NGRC that are also parents	93
Total	209
Parents confirmed as unavailable/lost	46
Parents yet to be sourced	**250**

Generation of ~0.8x raw sequencing coverage for 4 or 8 megagametophytes per founder sampled in the Founder Collection

Of the 125 seed-lots collected and shipped to the AVR laboratory (AgriBio, La Trobe University), 119 were used for megagametophyte tissue isolation. Megagametophyte tissue isolation is based on seed germination and dissection of the megagametophyte from the developing embryo. Of these 119 seed-lots, 118 yielded at least one megagametophyte and 106 yielded the target of either 4 or 8 megagametophytes (based on parental contribution). With 8 megagametophytes there is a 0.992 probability of sampling both alleles at least once and with 4 the probability is 0.875. A total of 904 megagametophytes were isolated. A subset of 552 megagametophyte tissue samples, representing 118 founder genotypes, were used in library construction. Overall, this collection sampled ~63% of the founding genes in the Australian Radiata Pine breeding program with the top 44 founding genotypes representing 57.7% of the founding genes.

The generation of whole genome sequencing for all batches of sequencing libraries was completed using the Illumina NovoSeq workflow system. DNA was extracted using a modified CTAB method and shotgun libraries constructed using the KAPA™ HyperPrep (Roche) method. Libraries were sequenced on multiple runs of the Illumina NovoSeq and Illumina MiSeq sequencing instruments and fastq files were generated using standard Illumina base calling workflows.

Analysis of this genomic data set including (but not limited to) filtering, alignment and SNP variant discovery

Overall sequence data was generated for the 552 megagametophyte samples to an average nominal coverage depth of 0.84. Results by founder sample are summarised in Appendix 1. The outputs of this work are the raw sequence files which have been made available to TBA. The actual files are stored on the AVR BASC for a period of 4 years. These raw data are available to TBA upon request and can be shared using the AVR SFTP server TAWNY.

Testing an industry standard low-cost, low-density SNP assay

The strategy for delivering this outcome was to first validate a Thermo Fisher based chip assay developed in NZ through a research program jointly funded by the Radiata Pine Breeding Company (RPBC) and the Ministry of Business, Innovation and Employment (MBIE) of New Zealand. Scion Strategic Science Investment Funding also supported the research. The plan, if the outcome of the testing was negative, would be to proceed to immediate recommendation for developing our own Australian derived low-cost assay. If the Scion chip proved to have utility, then the development of a new lower cost assay can be completed without the pressure to have an available working assay. Significant advantages and savings are likely to be available if a new chip assay is eventually developed. A new chip will exploit the technological advances that have occurred since the Scion chip was designed; and make use of the genomic resources developed in this project, which are based specifically on Australian germplasm. There may be scope for the new chip to be multi-purpose in the sense the chip allows for multi-species hybridisations. This has proven to be an effective approach used by AVR to drive down genotyping costs.

An initial consignment of 192 foliage samples, collected from founder genotypes, first-generation parents and native landrace material was sent to Australian Genome Research Foundation (AGRF) laboratories for DNA extraction. The extracted DNA was shipped to the Thermo Fisher laboratory in California, USA for genotyping using the Scion chip. TBA received back the called SNP genotypes in variant call format (VCF). Scion have control over the raw data received directly back from Thermo Fisher and are responsible for quality control decisions made (which SNP/samples to reject). In Figure 2, the left plot shows the distribution of frequencies of the alleles denoted as the reference allele. The distribution is in expectation with theory, in that it has a U-shape with more low frequency alleles as compared to high-frequency alleles. It is unclear if there is significant ascertainment bias arising from the design process.

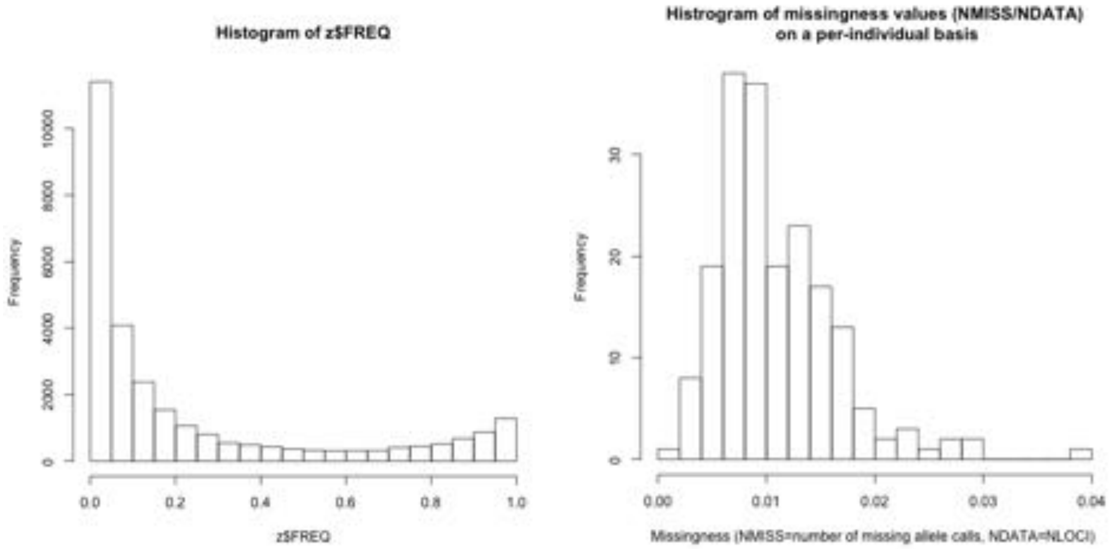

Figure 2 Histogram of the allele frequencies (left plot) and histogram of the missingness values on a per-individual basis (right plot)

Figure 2, right plot shows the distribution of missingness values on a per-individual basis (fraction of loci not called per individual). It generally shows a high call-rate, which is a feature of a well-designed chip. We concluded the chip yielded satisfactory results and decided to send a second consignment of samples for processing.

The second consignment consisted of 490 foliage samples separated from the NGRC derived Progeny Collection samples and 224 foliage samples separated from the Caroline derived Progeny Collection samples (see Table 1), plus a further 48 foliage samples supplied from HVP which were collected from founder genotypes they had archived. Six known duplicate samples were included to make up a total of 768 samples. Examination of the VCF received back from Scion revealed the distribution of allele frequencies and the distribution of missingness was like the first consignment of 192 samples (see Figure 3).

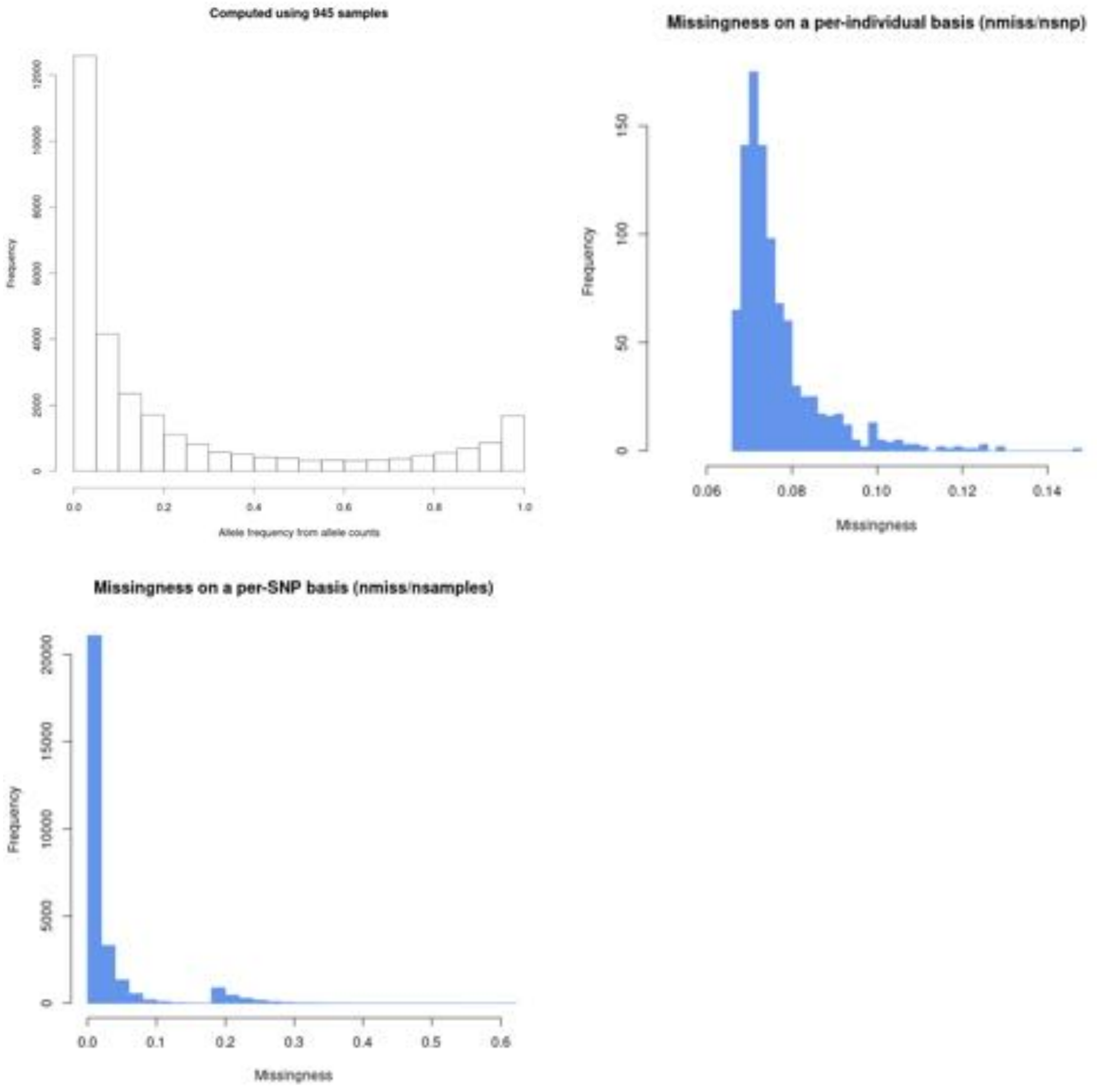

Figure 3 Histogram of the allele frequencies (top left plot), histogram of the missingness values on a per-individual basis (right plot) and histogram of the missingness values on a per-SNP basis (right bottom plot), when considering all individuals in both consignments

The deliberate placement of duplicates in the second consignment afforded the opportunity to check on the concordance rate between the original and the duplicate. The concordance rate is the number of identical genotype calls made on the original and the duplicate divided by the number of called genotypes. Concordance rates were generally very high with one sample showing a high number of mismatches, most likely indicative of mis-called genotypes due to a poor assay for that sample.

Table 3 A check of the concordance and discordance between duplicates and the original

Duplicate	Original	Number matches	Concordance rate	Number of non-matches	Discordance rate
a551114-4396658-072021-911_P11.CEL (TBA-761)	a551114-4396658-072021-911_H16.CEL (TBA-582)	27586	0.965	180	0.006
a551114-4396658-072021-911_F15.CEL (TBA- 764)	a551114-4400667-081521-006_I15.CEL (TBA-369)	28058	0.981	87	0.003
a551114-4396658-072021-911_N11.CEL (TBA-760)	a551114-4397097-072421-802_B10.CEL (TBA-339)	27727	0.969	173	0.006
a551114-4416684-041822-458_P12.CEL (TBA-762)	a551114-4397097-072421-802_D04.CEL (TBA-559)	26066	0.911	842	0.029
a551114-4396658-072021-911_D13.CEL (TBA-763)	a551114-4397097-072421-802_L08.CEL (TBA-488)	27771	0.971	146	0.005
a551114-4397097-072421-802_P12.CEL (TBA-520)	a551114-4396658-072021-915_J20.CEL (TBA-520)	21295	0.745	6955	0.243

It was found that the two VCF files received for each consignment were not compatible for merging using standard tools (such as VCF-merge). The names of the contigs changed between 2019 and 2020, and the lists of SNP in each file were not identical. There were approximately 28,500 SNP assayed in both consignments with approximately 2,000 SNP unique to each consignment, and with 26,600 SNP in common. The union of SNP in both consignments amounted 30,560 SNP. A custom script was written to merge the two files. Due to duplication of samples in the first consignment and samples with missing genotype identifiers, the total number of useable samples was reduced from 958 to 945.

In-silico analysis of pedigree forensics pipeline

Given that we added a new plan for undertaking pedigree forensics on the available SNP data, a sub-project dedicated to thoroughly testing the software used in pedigree forensics was undertaken. A preliminary run-through of the SNP data with the software indicated a substantial number of errors in the pedigree and the project team wanted to be sure the software was reliable and robust before sharing these results more widely. There are two main software packages that were tested: SEQUOIA, which implements likelihood-based methodology at the SNP level; and the **G/A** matrix comparison tool, which operates at the level of relationship coefficients.

Introduction to 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 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 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 grandparents 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, the function only performs parentage assignment
- If one or more iterations are specified 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 wherever possible

SEQUOIA's author advises using a subset of between 300 and 700 SNP with

- decent call rates (> 0.9)
- in low linkage disequilibrium with each other ($r^2 < 0.2$)
- high minor allele frequencies (MAF > 0.3)

To investigate these parameter settings we subset the SNP file using PLINK. PLINK software has a function that will output a subset of SNP from the list of available SNP given a set of criteria and can be quickly used to produce many independent SNP data draws. Running PLINK requires data to be in a PLINK readable format, which is achievable via tools that convert from VCF to PLINK formats. The switches used in the filtering step are typically

- --maf 0.3 --indep 50 5 2

where

- The maf switch specifies only SNP with minor allele frequencies (MAF) greater than 0.3 are selected.
- The three parameters in the --indep switch are: window size in variant count (50), a variant count to shift the window at the end of each step (5), and a variance inflation factor (VIF) threshold (2). At each step, all variants in the current window with VIF exceeding the threshold are removed.
- A VIF of 1 would imply that the SNP is completely independent of all other SNPs. The PLINK manual advises values between 1.5 and 2 should probably be used; if this threshold is too low and/or the window size is too large, too many SNPs may be removed

The best way to get the SNP genotype data supplied to SEQUOIA is by converting it to a "raw" format which can be achieved using a PLINK switch (--encodeA)

When running SEQUOIA, it is important to provide files containing:

- A field-based pedigree
    - id, dam, sire (use NA if unknown)
- and "life history" data, covering
    - The individual's identity, Sex (1=female, 2=male, 3=unknown, 4=hermaphrodite) and Birth-Year (Planting-Year for trees)

The field-based pedigree is easily obtained from DATAPLAN but required some "massaging" because SEQUOIA does not handle probabilistic parentage. The life history data are not mandatory, but in our initial testing poor results were obtained unless life history data was supplied. Such data is very easy to obtain from DATAPLAN by querying the location of the ortet and the year of planting of the trial.

The core SEQUOIA function requires some key parameters to be set. These are

- **MaxSibIter** This parameter specifies the maximum number of iterations of sibship clustering and can have values:
    - -1 Only check for duplicates

  - 0 check for duplicates + parentage assignment
  - *x* > 0 as above plus at most *x* iterations of full pedigree reconstruction
- When **MaxSibIter** is <= 0, the program is quite fast and can only run in several minutes. When **MaxSibIter** is > 0, and the number of SNP is in the thousands and not hundreds, convergence can take several days.
- **Err** The genotyping error rate assumed, equal across all SNP.
- **Complex** The complexity of the mating system considered. The default is "full", which considers the full range of possible relationships including relatives mating each other, but assumes the organism is dioecious (i.e. an individual cannot change sex). Setting Complex="herm" allows individuals to change sex. Setting Complex="herm2" is similar to "herm" but completely ignores the dam vs sire role ("herm" does make this distinction). With "herm2" no conclusions can be drawn from whether individuals are assigned as maternal or paternal half-siblings. TBA has found that better results are obtained with "herm2", even though it does occasionally want to make the female parent the male parent and vice versa. These sex role assignment errors are generally easily corrected.

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 broader genetic evaluation 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.

It is simple to run SEQUOIA with **MaxSibIter** set to 0, to test for duplicates and parentage mis-assignment and this would be fast and not that disruptive to a 'typical' TREEPLAN run. A TREEPLAN run strategy could be to remove from the GRM those individuals that have mis-assignment with the field pedigree. These individuals are flagged in DATAPLAN for follow-up work with more computing intensive SEQUOIA runs (setting **MaxSibIter** > 0) and other investigative work. The aim will be to semi-automate the recovery and updating of the field parentage records.

A typical sequence of steps when running SEQUOIA within the R framework is as follows

1. Import the data
   Geno <- GenoConvert(InFile = "input_data.raw", InFormat="raw")
2. Import the life history data
   lifehist <- read.table("life_hist.txt",header=T)
3. Import the field-pedigree
   fieldped <- read.table("field_ped.txt",header=T)
4. Run a simple parentage assignment
   ParOUT <- SEQUOIA(GenoM = Geno, LifeHistData = lifehist, MaxSibIter = 0, Err=0.01, Complex='herm2')
5. Compare the pedigree inferred from the SNP data to the field pedigree
   chk <- PedCompare(Ped1 = fieldped, Ped2 = ParOUT$Pedigree)

6. Output the mismatches

   ```
   write.csv(chk$Mismatch, 'Mismatch.csv')
   ```

7. If doing sibship clustering in follow up SEQUOIA runs

   ```
   SeqOUT <- SEQUOIA(GenoM = Geno, LifeHistData = lifehist, MaxSibIter = 20, Err=0.01, Complex='herm2')
   ```

The PedCompare function in the R package is useful for comparing a field-based and genetically inferred pedigree. It identifies mismatches for those individuals which have genotyped parents, assigned based on SNP data, that do not match the parents supplied from the field-based pedigree.

Introduction to the G-A matrix comparison tool

Comparing the constructed **G** matrix with the **A** matrix (limited to the assayed individuals, so it has the same dimensions as the **G** matrix) is an alternative method for detecting mismatches between a field-based pedigree and a pedigree inferred from the SNP data. A custom PERL script was written that performs this comparison using outputs from FORTRAN programs that construct both the **A** and **G** matrices. This tool is likely to prove useful in situations where TBA has received only a constructed **G** matrix from a 3rd party and does not have access to the SNP level data and may in fact be a better tool to implement in a TREEPLAN run for exclusion of samples with pedigree errors over the SEQUOIA approach detailed above. The tool could be run as a second stage quality assurance (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 **G** matrix, rather than SNP level data.

A limited **G**-**A** 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 (2011, 2012) have published useful articles on the variance expected in genomic relationships. 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 these should not deviate from 0.5. However, due to genotyping errors and the finite sampling of the genome, variance in parent offspring relationships is observed. Simulation may be one way to derive what would be typical given an assumed genotyping error rate and sampling protocol.

Testing Pedigree Recovery via simulation

Simulation was used to get a better feel for the features of both approaches to pedigree forensics, to test limitations, and in the case of SEQUOIA to understand better the implications of setting different values to the main parameters. In the case of simulated data, the known number of chromosomes and map length, given the recombination rate assumed in the coalescent simulation, was used to derive the expectations.

The simulation was designed to mimic a typical “3-generation” generic forest tree breeding population:

- Generation 0
    - A set of 200 native mothers and 800 unknown fathers comprised the base
    - The coalescent simulator ‘msprime’ was used to generate SNP level data on 1000 founders assumed to derive from a single population
    - 21,907 SNP across 10 chromosomes were assumed sampled, of which 1906 were QTL
- Generation 1
    - 200 OP families generated: each mother generates 60 OP progeny, which were tested at year 5
    - 12 progeny in 40 families were targeted for DNA assaying. Deliberately not assaying every family.

  - Mothers of assayed progeny were assayed.
  - 200 new parents were selected on the basis of an index value
- Generation 2
  - 200 parents were crossed at year 8 using a partial diallel design to generate 600 families
  - Each family generated 20 CP progeny
  - Progeny were assessed at year 13
  - 5 progeny within each of 60 randomly selected families were targeted for DNA assaying
  - Parents of assayed progeny were deliberately not assayed (in order to provide cases of mismatch where the true parent was not assayed)
  - 200 new parents were selected to breed generation 3
- Generation 3
  - 200 parents were crossed using a partial diallel design to generate 600 families
  - Each family generated 20 CP progeny
  - 5 progeny within each of 60 randomly selected families were targeted for DNA assaying
  - Parents of assayed progeny were also assayed

A range of pedigree error types were introduced to represent the range of pedigree error types that may be encountered in a real system. Table 4 summarises the pedigree errors that were introduced into the pedigree by deliberately changing either one or both true parents to other individuals. Errors set at the family level will apply to all sibs in the family with the misassigned parent or parents. These errors will occur in real life when the wrongly identified pollen is applied to the female parent or vice-versa, or when the identity of a seed-lot is wrongly assigned and both parents are wrong. In the simulation we expect to see cohorts of full- and half-sibs still maintaining their correct relationship to each other but their pedigree-based relationships to other relatives via their parents to not align with what the SNP genotypes are inferring. Under this simulated scenario, no errors of type 8 and 10 were obtained.

Table 4 A set of 12 error types introduced into the pedigree that are applied at the family level

Family type	Error code	True Mum assayed	True Dad assayed	Wrong Mum	Wrong Dad	Wrong Mum assayed	Wrong Dad assayed
CP	1	0	0	1	1	0	0
CP	2	0	0	1	0	1	NA
CP	3	0	0	1	1	1	1
CP	4	1	1	1	1	0	0
CP	5	1	1	1	0	1	NA
CP	6	1	1	1	1	1	1
CP	7	1	0	1	0	1	NA
OP	8	0	0	1	0	0	NA
OP	9[#]	0/1	0	1 (true mum and wrong mum are swapped)	0	0/1	NA
OP	10	0	0	1 (wrong mum still has her own family)	0	1	NA
OP	11	1	0	1 (true mum and wrong mum are swapped)	0	1	NA
OP	12	1	0	1 (wrong mum still has her own family)	0	1	NA

[#] When swapping 4 mothers between different OP families 2 mothers were assayed, 2 were not

Errors were also set at the genotype level (these will have the code 13). In the simulation this was achieved by assigning the SNP genotype data of individual X to individual Y. Hence what you think is individual Y is individual X, and a pedigree-based relationship coefficient between Y and any of its relatives, including its assumed full- and half-sibs, will not agree with what the SNP genotype data is inferring.

A couple of each type of family-based errors and 26 genotype-based errors were implanted into the simulated data set obtained at generation 3. There was a total of 166 errors implanted (142 individuals will have errors due to family-based errors and 25 due to genotype-based errors and 1 with both type of error). The data was then processed through SEQUOIA. Several SNP filtering options were used (by changing the VIF and window size) to get different sized SNP sets:

```
> plink -bfile tbasim --maf 0.3 --indep 100 5 1.5   → resulted in 680 SNP
> plink -bfile tbasim --maf 0.3 --indep 50 5 2.0    → resulted in 1105 SNP
> plink -bfile tbasim --maf 0.1 --indep 50 5 2.0    → resulted in 2217 SNP
```

A SEQUIOA run was also tested using all available 21,907 SNP.

Table 5 shows the results of the SEQUOIA runs on the simulated data, when MaxSibIter is either set to 0 or 20, and for various SNP set sizes. The runs for 1105 are not shown as they were almost identical to when there were 680 SNP. The best results were obtained for a SNP set size of 680, which confirms the recommendation that a SNP set size of under 700 is sufficient, if all SNP meet a high MAF (e.g. 0.3) and are in linkage equilibrium with each other. Error detection got progressively worse when expanding the SNP set size up to maximum size of 21,907 SNP.

SEQUOIA appears to have a high success rate at detecting a wrong mother when the seed is open-pollinated, regardless of whether the true mother is assayed (error types 11,12) or not (error types 9 and 10). A parentage assignment run (MaxSibIter=0) is sufficient for detecting these types of error.

Table 5 Results of SEQUOIA in terms of detecting known errors in the simulated pedigree, for different SNP sets

Error type	# Individuals expected to have errors	# errors detected (680 SNP, MaxSibIter= 0)	# errors detected (680 SNP, MaxSibIter= 20)	# errors detected (2217 SNP, MaxSibIter= 0)	# errors detected (2217 SNP, MaxSibIter= 20)	# errors detected (21907 SNP, MaxSibIter= 0)	# errors detected (21907 SNP, MaxSibIter= 20)
1	10	0	10	0	5	0	1
2	2	0	0	0	0	0	0
3	10	0	10	0	10	0	7
4	6	6	6	6	6	0	0
5	10	10	10	10	10	1	5
6	10	10	10	10	10	0	0
7	6	5	6	5	5	4	4
9	26	24	24	24	24	7	7
11	36	36	36	36	36	15	15
12	24	24	24	24	24	5	5
13	25	10	14	10	14	4	5

In their raw format the output from SEQUOIA is not conducive for helping a breeder obtain some clues as to the possible causes for the pedigree error. A custom script was written that parses the SEQUOIA output and reads a complete pedigree file for the population, as well as a locations file (the location of the ortet for all individuals) and summarises the information into a tabular format. Some examples of parsed SEQUOIA output are shown in Table 6.

Table 6 Examples of errors detected by sequoia and parsed by custom script that collates information of sibs (sibs are shown with inferred parent or parents in parentheses and their location in brackets)

Genotype id	Error code	Progeny type	Trial	Mum id	Dad id	Inferred mum	Inferred dad	Number sibs	Number assayed	Number mismatched	Sibs with similar error
4066	11	OP	01_000	51	0	52	0	60	12	12	4068 (52) [01_000], 4094 (52) [01_000], 4108 (52) [01_000], 4073 (52) [01_000], 4117 (52) [01_000], 4114 (52) [01_000], 4077 (52) [01_000], 4105 (52) [01_000], 4103 (52) [01_000], 4112 (52) [01_000], 4089 (52) [01_000]
7424	12	OP	01_000	102	0	108	0	120	24	12	7426 (108) [01_000], 7476 (108) [01_000], 7468 (108) [01_000], 7473 (108) [01_000], 7454 (108) [01_000], 7474 (108) [01_000], 7478 (108) [01_000], 7440 (108) [01_000], 7465 (108) [01_000], 7461 (108) [01_000], 7429 (108) [01_000]
27065	4	CP	Not planted	85	5885	20650	16401	20	5	5	27080 (20650 x 16401) [NA], 27075 (20650 x 16401) [NA], 27078 (20650 x 16401) [NA], 27067 (20650 x 16401) [NA]
26084	5	CP	Not planted	15409	24381	15212	MATCH	20	5	5	26090 (24381 x 15212) [NA], 26100 (24381 x 15212) [NA], 26092 (24381 x 15212) [NA], 26091 (24381 x 15212) [NA]
31307	6	CP	Not planted	10924	164	4273	21680	20	5	5	31313 (4273 x 21680) [NA], 31319 (4273 x 21680) [NA], 31310 (4273 x 21680) [NA], 31317 (4273 x 21680) [NA]
16628	7	CP	01-008	29	6731	2922	MATCH	20	5	5	16632 (2922) [01_008], 16635 (2922) [01_008], 16639 (2922) [01_008], 16630 (2922) [01_008]
17343	1	CP	01_008	111	9292	MATCH	M0010	20	5	5	17350 (M0010) [01_008], 17353 (M0010) [01_008], 17346 (M0010) [01_008], 17352 (M0010) [01_008]
14688	3	CP	01_008	4706	10917	F0035	M0032	20	5	5	14695 (F0035 x M0032) [01_008], 14693 (F0035 x M0032) [01_008], 14700 (F0035 x M0032) [01_008], 14697 (F0035 x M0032) [01_008]
14135	13	CP	01_008	5038	6720	8807	M0010	10	5	1	

For tree with genotype_id 4066 Table 6 shows that the correct mother (52) has been inferred from the SNP data. In this OP family there are 60 sibs, of which 12 have been assayed and all 12 have also been mismatched. This would lead a breeder to conclude that the assumed mother (51) has been wrongly assigned. For tree with genotype_id 7424 it is a similar story, but there are 120 sibs, 24 of which have been assayed, and of those assayed, 12 have been identified as having a different mother (108). A slightly different conclusion could be reached: perhaps there was one crossing event and the mother (85) was correctly assigned, and another crossing event when it was incorrectly assigned as the mother.

A parentage assignment run is also sufficient for finding the true parents, when both true parents are assayed and when both assumed parents are either not assayed (error type 4) or assayed (error type 6). When only 1 parent is falsified in a CP family (the mother) and the false mother is assayed, and the father is assayed (error type 5) or not assayed (error type 7), a parentage assignment run is also again sufficient.

When both true parents are not assayed and either the falsified parents are not assayed (error type 1) or assayed (error type 3), a parentage assignment run is not sufficient for finding the errors. Sibship clustering is required to form sibships and SEQUOIA can then determine that the parentage of these sibships has been wrongly assigned. In the case of error type 1 (both true and falsified parents are not assayed) SEQUOIA does not get the story completely right. In the example of genotype 17343 It suggests that the assigned mother could be right, as it probably has no information on the mother, but it is saying the father is wrong, presumably because the father is the assigned parent of assayed progeny in other families and there are inconsistencies when comparing the sibs of those families with the sibs of this family. It proposes a dummy male parent (M0032).

In the case of error type 3 (true parents are not assayed and falsified parents are assayed) SEQUOIA is unequivocal in proposing two dummy parents, which cannot be matched to any genotyped individuals.

SEQUOIA was able to detect 14 of the 25 imposed genotype-based errors (type 13). In most cases this occurred because the parents of the actual genotype had been assayed. In very few instances was SEQUOIA able to detect a genotype-based error even if the parents of the actual genotype were not assayed. Genotype 14135 was an example. SEQUOIA was able to determine that the correct mother was 8807 even though it has not been assayed but can determine that the assigned father is wrong. There were 5 sibs in this CP family that were assayed but this is the only sib that has a mismatch suggesting the family pedigree is correct, but this one genotype has been mis-labelled at some point.

The **G**-**A** matrix comparison method was also applied to the simulated data set. In general, the **G**-**A** matrix comparison performed well, backing up the findings of SEQUOIA and will provide a useful complement to SEQUOIA, or an alternative method of pedigree error detection if SNP level data are not available. Notably the **G**-**A** matrix comparison does not detect type 1 errors because parents are not assayed, and the sibs in the family remain true full-sibs, even though their parents are misassigned. **G**-**A** matrix comparison did detect type 2 errors (mother is wrong and is assayed, true mother is not assayed), where SEQUOIA did not. **G**-**A** matrix comparison does detect type 3 errors (both mother and father are wrong and both false parents are assayed.

As expected, **G**-**A** matrix comparison as it stands does not detect errors of the type where false parents are not assayed, but the true parents are (type 4). The program could scan for individuals that have a **G** matrix coefficient in the range 0.47 to 0.53 with all sibs in the focal CP family and propose these as the true parents. The **G**-**A** matrix comparison was successful in detecting all other error types, except for most instances of error type 9 when the false female parent was not assayed. Table

7 summarises the result of pedigree error detection using the **G**-**A** matrix comparison and compares these with results obtained from SEQUOIA.

Table 7 Comparing pedigree error detection using sequoia with a method based on comparing the G with the A matrix

Error type	# individuals expected to have errors	# errors detected SEQUOIA (680 SNP, MaxSibIter=20)	# errors detected using GRM-NRM comparison
1	10	10	0
2	2	0	2
3	10	10	10
4	6	6	0
5	10	10	10
6	10	10	10
7	6	6	6
9	26	24	4
11	36	36	36
12	24	24	24
13	25	14	22

Putting the SNP chip assay results to work

The availability of a substantial number of individuals with DNA assay data led us to consider undertaking initial quality control and pedigree forensics, building a draft **G** matrix, checking the **G** against the **A** matrix and running a TREEPLAN single-step analysis.

Quality control and pedigree forensics

The VCFTOOL utility was used for preliminary filtering of the SNP. A total of 8147 SNP with low MAF (< 0.01) were removed, and a total of 829 SNP with high missingness (> 0.5 as faintly seen in the bottom plot in Figure 3) were also removed, leaving 21,584 SNP.

The public domain software PLINK was used to convert the genotype call data in VCF to raw format for entry into the SEQUOIA R package. A SEQUOIA R function was used to convert the raw data to a SNP genotype matrix.

Based on the work completed with simulated data, from which we determined pedigree forensics are best undertaken using a limited number of independent SNP, a subset of SNP from the final 21,584 for the 945 individuals was selected. This was achieved in two steps. A pruning step was first performed using the public domain software PLINK. The PLINK help documentation suggests the following switches:

- --allow-extra-chrom

This flag is needed because a large number of contig names are used in lieu of a finite set of chromosome labels

- --indep 200 5 1.5

This flag is used to produce a pruned subset of markers that are in approximate linkage disequilibrium with each other. The three parameters are: window size in variant count (50), a variant count to shift the window at the end of each step (5), and a variance inflation factor (VIF) threshold (1.5). At each step, all variants in the current window with VIF exceeding the threshold are removed. A VIF of 1 would imply that the SNP is completely independent of all other SNPs. The PLINK manual advises values between 1.5 and 2 should probably be used; if this threshold is too low and/or the window size is too large, too many SNPs may be removed.

- --maf 0.3

This flag is used to produce a pruned subset of markers that have a minor allele frequency greater than 0.3, which is recommended for undertaking pedigree forensics. The PLINK pruning step resulted in a subset of 2296 markers.

The data for 2296 SNP on 945 individuals were converted to RAW format using the PLINK --encodeA flag and imported into R and translated into a DataFrame using the SEQUOIA GenoConvert function.

A field-based pedigree and life-history data were also imported from DATAPLAN and converted to R DataFrames FieldPed and LifeHist, respectively. SEQUOIA provides a SnpStats function to estimate the genotyping error rate per SNP, conditional on the provided field-based pedigree and an assumed error structure (probabilities of observing a genotype conditional on actual genotype and per-locus error rate E). The SEQUOIA manual recommends dropping SNPs with an error rate higher than 0.1. This reduced the number of SNP to 1517 SNP, which though more than double the recommended number of between 300 and 700 SNP, is still manageable in terms of computing run time.

Two SEQUOIA runs were then completed.

Run 1 on 945 samples using a subset of 1517 SNP, with no iteration, was used to identify and remove seven duplicates. The intentional duplicates had already been removed so these were accidental duplicates that we were not aware of. These samples were identified and removed from the main VCF file, which stores the complete set of SNP genotypes.

Run 2 on 938 samples using a subset 1517 SNP, with iteration allowed, in combination with the PedCompare function, was used to flag mismatches between the field-based pedigree and the pedigree inferred with SNP data. The complete set of results is a large Table, containing 329 progeny with some type of misassignment (either 1 or both parents are mismatched) and is presented in Appendix 2. It is a hard table to digest and not easy for breeders to obtain some sense if a serious systematic error has occurred at some time point or location. Was there a particular time or epoch of the breeding program in which a noticeable number of errors occurred? Are there trials with noticeably more errors than what is considered typical? Did errors occur during trial establishment, seedling establishment in, or transfer from, the nursery, or either at crossing or grafting time? Table 8 shows, by trial, the number of families with at least one mismatched parent. In trial BRGT1301 there are 59 families with a mismatch, and most of the assayed sibs within those families are mismatched (64 out of 86), probably indicating a systematic error of some kind with the establishment of this trial. Field notes from this trial also indicate that sample tracking errors were likely for this trial showing the importance of collecting and archiving trial establishment records.

Table 8 Number of families with at least one mismatched parent by trial, and the number of assayed sibs within those families and the number of mismatched sibs within the families

Trial id	Number of families with one or more mismatches	Number assayed sibs within the families	Number mismatched sibs within the families
BRGT1301	59	86	64
BR9601	28	80	78
RES1295	12	19	18
BR9606	11	29	19
BR9705	11	20	17
BR9617	10	15	14
Q14/1.38	10	10	10
BR9611	7	14	11
BR9703	7	23	23
BR0903	5	10	5
BR9615	4	9	7
RAD238	4	4	4
BR0901	3	8	7
BR0904	2	2	2
BR9613	2	6	2
BR9614	2	8	7
BR9713	2	2	2
BRGT1201	2	5	5
GT0001	2	4	2
GT0002	2	5	4
VRC070	2	2	2
BR0803	1	3	3
BR9604	1	5	3
BR9701	1	1	1
BR9707	1	4	3
BRGT1303	1	2	1
BRGT1403	1	2	2
GT9506	1	1	1
RAD137	1	1	1
VRC071	1	2	1
VRC095	1	1	1
Total	**197**	**383**	**320**

In some instances, SEQUOIA has been able to locate the correct parent (see Table 9). In the case of trial BRGT1301, about a half of those families with a mismatch could be assigned either the correct female parent or correct male parent or both, because they had been assayed. It is not unexpected that many true parents could be recovered in the trial BRGT1301 (Caroline) because most of the planted progeny were generated from crossing undertaken in the NGRC and virtually all genotypes in the NGRC have been assayed.

Table 9 Number of cases when a "correct" parent could be assigned to a family with mismatched progeny

Trial ID	Number of cases when the correct mum could be inferred	Number of cases when the correct dad could be inferred
BRGT1301	30	33
BR9601	19	21
Q14/1.38	10	0
RES1295	10	4
BR9617	9	4
BR9606	8	9
BR9705	7	9
BR9703	6	5
BR9611	4	5
BR9613	2	2
BR9614	2	2
RAD238	2	4
BR0803	1	0
BR0901	1	2
BR0904	1	2
BR9604	1	0
BR9615	1	4
BR9701	1	1
BR9707	1	0
BR9713	1	1
BRGT1201	1	0
GT0002	1	1
VRC095	1	0

A summary of the pipeline completed so far is presented in Figure 4. A feature of the pipeline is the central, filtered VCF file that will most likely be stored in DATAPLAN. Several cycles of filtering, of both SNP and samples, based on various criteria, will be necessary before the **G** matrix construction step. In this case we did not fix pedigree errors identified using SEQUOIA and went ahead and built the **G** matrix to determine if the **A** matrix **G** matrix comparison analysis supported/confirmed the mismatches determined by SEQUOIA.

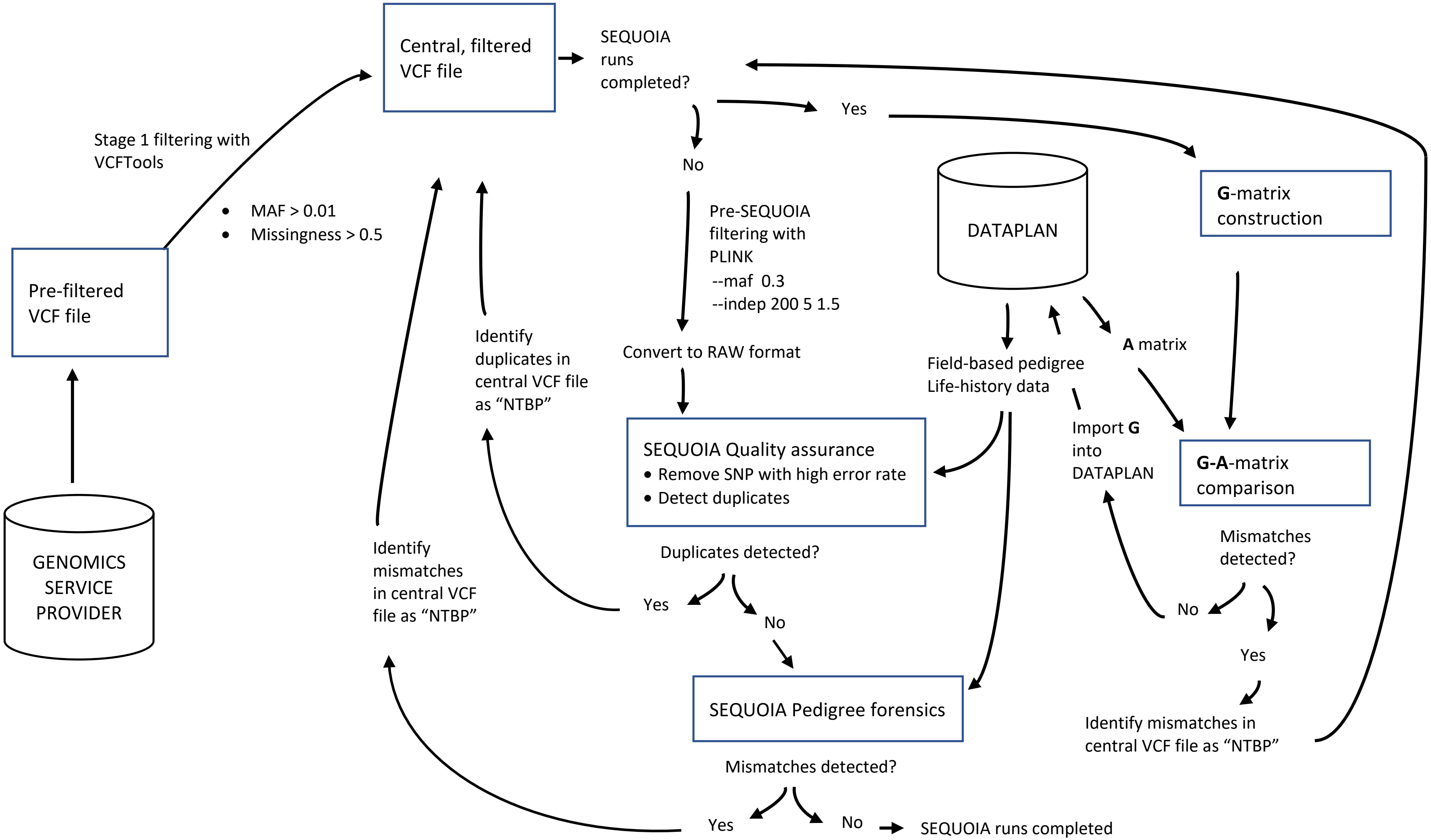

Figure 4 Pipeline for quality control, pedigree forensics and G-matrix building (“NTBP” denotes NOT TO BE PROCESSED)

Constructing a G matrix

The KGD method (Dodds et al., 2015) is applicable when imputation is not used to fill in the missing genotype calls, and slightly different SNP sets are used among different cohorts of individuals (as is the case here). Elements of the **G** matrix are calculated using only those SNPs which are scored in both corresponding individuals. Genotype calls for the filtered set of SNP (21,584 SNP), for the 938 individuals, were used to compute the **G** matrix.

Figure 5 shows the distributions of the genomic relationship coefficients for three relational types: parent-offspring, full-sibs and half-sibs. In each case there is a distinct mixture of two normal distributions: one distribution is centred near the expected values of 0.5, 0.5 and 0.25, respectively, and the other smaller distribution centred around zero, which indicates the individuals are not related. The fact that the main distribution is not centred exactly around the theoretical expectation indicates our factors for centralising the coefficients are not quite correct. With more data these normalising factors will become better estimated.

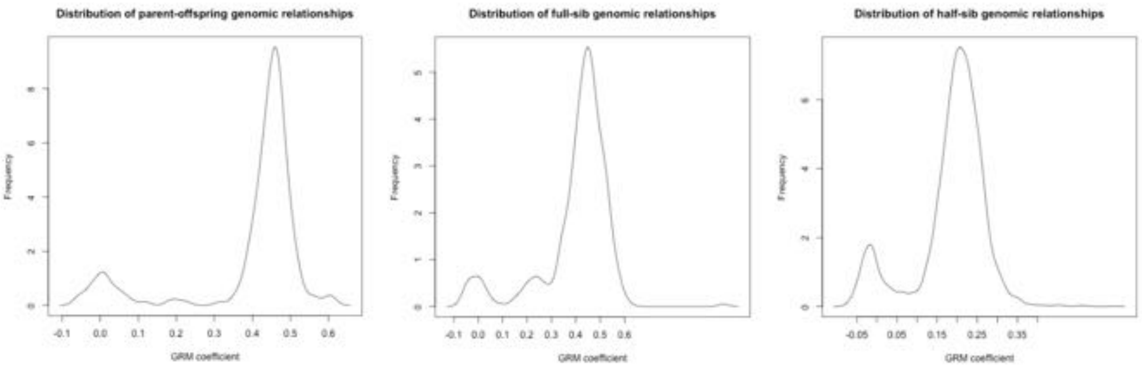

Figure 5 Distribution of genomic relationships for three relational types: parent-offspring; full-sibs; and half-sibs

Checking the G matrix against the A matrix

There were 882 instances where the **G** matrix coefficient did not agree with the **A** matrix coefficient when inspecting coefficients only for the relational types considered (see Table 10). Most of these instances (780) indicate the assumed half-sib relationship is incorrect.

Table 10 Number of instances when a G matrix coefficient doesn't agree with the A matrix coefficient for the four relational types considered

Relational type	**Number of G-A matrix mismatches**
Parent-offspring in a CP family	29
Full sibs in a CP family	66
Mother-offspring in a HS family	7
Half-sibs in a HS family	780
Total	**882**

Most of the **G**-**A** matrix discrepancies support the findings of the SEQUOIA analysis. There were 74 instances of **G**-**A** matrix discrepancies that have no obvious connection to the SEQUOIA analysis results. From a simulation study we know that SEQUOIA had trouble detecting errors where one parent is wrong and is assayed and the other parent is correct and when an assayed individual X is in truth individual Y, which is not assayed, particularly when the parents of Y have not been assayed either. It is probable that the **G**-**A** matrix comparison tool is detecting these types of errors.

Of the 329 progeny that had some type of mis-assignment as indicated from the SEQUOIA analysis, 224 were also detected as having some type of **G**-**A** matrix discrepancy. That is, the **G**-**A** matrix discrepancy analysis supported or confirmed most cases from the SEQUOIA analysis. We also know from the simulation study that a **G**-**A** matrix discrepancy analysis cannot detect errors where the true parents are not assayed and false parents are not assayed and when false parents are not assayed, but the true parents are. When inspecting progeny that were flagged as having mismatched parents by the SEQUOIA analysis but were not flagged in the **G**-**A** discrepancy analysis, it would appear this is the case: their parents as stated by the field-based pedigree, were not assayed.

Full pedigree forensics for the correction of the field-based pedigree is likely to remain a stand-alone operation with an ‘in line’ pedigree error identification step to be applied in TREEPLAN runs to simply remove individuals from the analysis with incorrect pedigree assignment. Correction of the field pedigree should be implemented with caution and should wherever possible also source other information that could support changes, such as trial records and further testing of relatives. Correcting the field-based pedigree is also likely to be best implemented as an iterative process with corrections made over many rounds of checking until no errors are detected. Results from pedigree correction should also be used to identify operational practices that are high risk and to assist with practice improvement in operational breeding to reduce the likelihood of pedigree errors accumulating. As genotyping using arrays becomes more accessible the opportunity for fully correcting and checking pedigree will increase.

Single-step analysis

A **G** matrix constructed for 938 individuals based on a set of 21,584 SNP, was imported into DATAPLAN and flagged for use with the current national *P. radiata* TREEPLAN analysis system. This system contains 1,987,755 observations for 34 selection criteria (SC), measured on stems at 576,542 positions. There are 573,434 genotypes and 7,817 families in the pedigree. The selection criteria are correlated to varying degrees to 10 breeding objective traits (BOT). Multiple $NPV Indices have been defined by the economic weighting of BOT.

The prediction error variances (PEV) of the genetic effects in 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. 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 BOT accuracies with and without the **G** matrix are shown in Figure 6, for assayed and non-assayed trees. Points have been coloured to denote trees in different generation by parent status classes (e.g. Gen-0.parent and Gen-0.non-parent denote parents and non-parents in generation 0, respectively). As expected, due to the small number of assayed trees there have been no dramatic shifts in accuracy yet. This is expected as the proportion of the **H** matrix corresponding to assayed trees is very small (938/573,434 = 0.16 percent). There are small, discernible improvements in accuracy for BRANCH and SWEEP for Generation 1 non-parents that have been assayed. The improvements in accuracy for BOT then transfer to a small improvement in accuracy for the $NPV Index value (see Figure 7).

Table 11 to Table 14 contain the mean EBV accuracies and percentage change in the mean (%), when using the **H** or **A** matrices, for each of the BOT: MAI, SWEEP, STIFFNESS and BRANCH, respectively, and confirm the negligible changes observed in the X-Y plots. The greatest change in accuracy is for assayed, generation-0, non-parents when the BOT is BRANCH. The improvement is 5%.

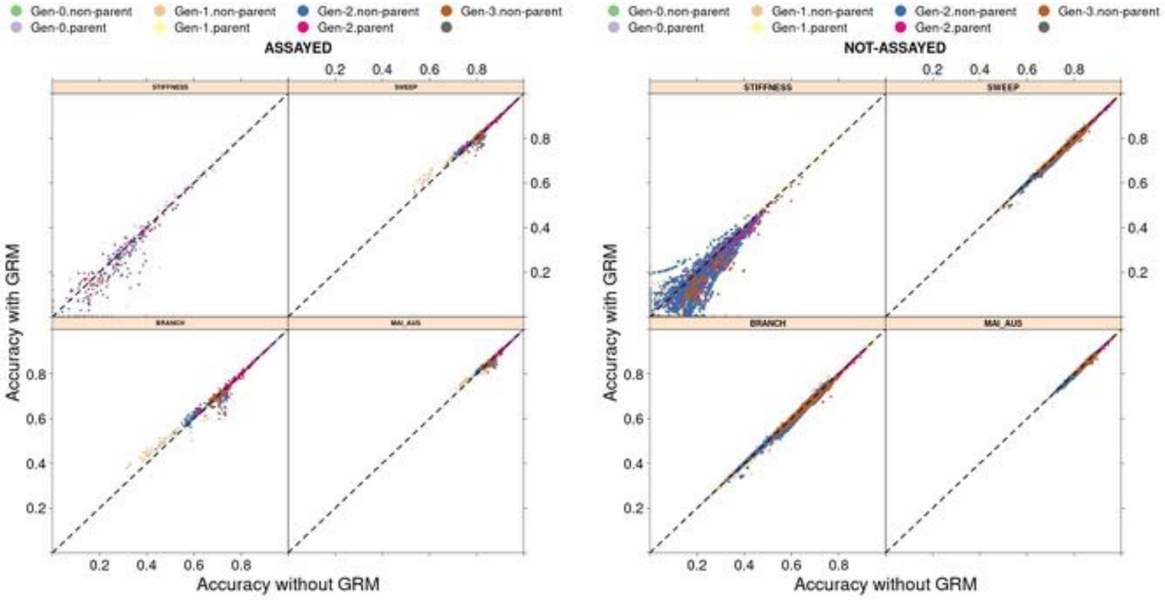

Figure 6 Accuracies of EBV for breeding objective traits (BOT) computed with and without the genomic relationship matrix (GRM). The left plot shows accuracies for assayed trees and the right shows accuracies for non-assayed trees.

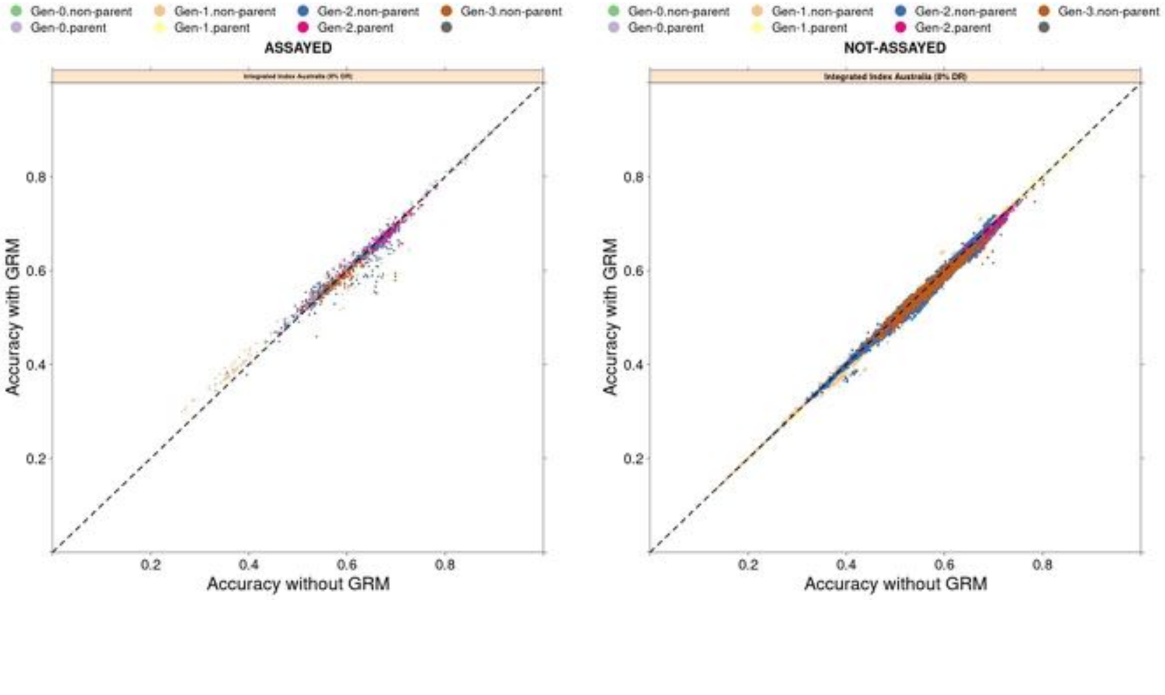

Figure 7 Accuracies of EBV for a NPV $Index, computed with and without the genomic relationship matrix (GRM). The left plot shows accuracies for assayed trees and the right shows accuracies for non-assayed trees

Table 11 Mean EBV accuracies for MAI, and percentage change in the mean (%), when using the H and A matrices in the mixed model equations

	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.96	0.95	0.08	0.78	0.78	0.90	0.85	0.85	0.01	0.74	0.74	0.01
Gen-1	0.89	0.89	-0.12	0.81	0.81	0.64	0.89	0.89	-0.04	0.80	0.80	-0.01
Gen-2	0.90	0.90	-0.05	0.86	0.86	-0.09	0.90	0.90	-0.27	0.84	0.84	-0.23
Gen-3				0.85	0.85	-0.26				0.84	0.84	-0.28

Table 12 Mean EBV accuracies for SWEEP, and percentage change in the mean (%), when using the H and A matrices in the mixed model equations

	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.96	0.96	0.07	0.65	0.63	2.42	0.83	0.83	0.01	0.57	0.57	0.01
Gen-1	0.85	0.85	-0.13	0.72	0.71	1.67	0.87	0.87	-0.08	0.71	0.71	-0.02
Gen-2	0.86	0.86	0.01	0.80	0.80	0.02	0.88	0.88	-0.37	0.76	0.76	-0.42
Gen-3				0.80	0.80	-0.33				0.76	0.77	-0.55

Table 13 Mean EBV accuracies for STIFFNESS, and percentage change in the mean (%), when using the H and A matrices in the mixed model equations

	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.31	0.31	0.90				0.00	0.00	0.54			
Gen-1	0.11	0.11	2.77	0.02	0.03	-27.67	0.10	0.10	0.33	0.00	0.00	-5.82
Gen-2	0.16	0.17	-6.77	0.06	0.08	-18.80	0.18	0.20	-10.39	0.01	0.02	-24.27
Gen-3				0.02	0.03	-49.81				0.00	0.01	-41.64

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

	Assayed						Non-Assayed					
	Parent			Non-parent			Parent			Non-parent		
	H	**A**	%	**H**	**A**	%	**H**	**A**	%	**H**	**A**	%
Gen-0	0.89	0.89	0.19	0.48	0.46	4.99	0.61	0.61	0.02	0.37	0.37	0.02
Gen-1	0.75	0.75	-0.25	0.56	0.54	2.88	0.75	0.75	-0.13	0.55	0.55	-0.04
Gen-2	0.77	0.77	-0.10	0.69	0.69	-0.19	0.79	0.80	-0.61	0.65	0.65	-0.66
Gen-3				0.69	0.70	-0.77				0.66	0.67	-0.81

Discussion

This project was motivated by TBA's goal of having single-step genomic selection becoming routine in the Australian radiata pine breeding program. The technology has already been successfully implemented into the Australian blue gum breeding program in previous projects with significant increases in EBV accuracy, and is now being realised in both the *E. nitens* and *E. globulus* breeding programs with **G** matrices derived from thousands of assayed trees. Matching the success in these species was always going to be a more challenging task in a conifer species like *P. radiata* due to the mammoth size of its genome. The primary activity of this project was to initiate the development of foundational genomic resources, in the knowledge that follow-up research partitions would require these and be necessary to fully realise routine implementation of the breeding strategy referred to as genomic selection via single-step BLUP.

Our plan is multipronged. Firstly, TBA knew in advance that a reference genome is critical for the implementation of genomics methods into breeding and immediately for aligning and mapping *de novo* whole genome sequence generated in the project. Our initial strategy was to contribute to this effort by forging international collaboration to build a single super high-quality assembly. Our initial plan was to commission a Hi-C analysis of the genome of the individual used in the radiata pine assembly project being undertaken in New Zealand and to contribute this to the NZ effort. This was to be contributed to both improve the assembly and to secure TBA early access to their draft genome assembly. Despite expressions of goodwill by all parties and interest, this strategy did not eventuate with institutional barriers leading to an amicable postponement of collaborative intentions and the project team making the decision to initiate an Australian based *de novo* genome assembly effort. Costs have dramatically fallen in the last two years, and because AVR secured a substantial discount from Dovetail Genetics, undertaking this *de novo* assembly will cost a small fraction of what it costed five years ago. This work in now in progress and will extend beyond the life of this project.

The second line of attack was to generate a foundational dataset based on whole genome sequencing of the genomes of the most important founders in the radiata breeding program. TBA successfully sourced seeds from a majority of the most important founders such that we have sampled and sequenced approximately 65% of the genetics in the program. The haploid mega-gametophyte tissue cultivated from these seeds was our preferred source of DNA, as the haploid signal can be effectively used in data analysis and SNP discovery. The TBA now has a whole genome sequence database from which it can launch future phases of SNP discovery.

A third line of attack was to undertake major foliage collections that will wait in storage until a future research partition can be initiated that will fund assaying of the collections with the high-density SNP set. These collections are costly to undertake, and the pressure is to undertake them as soon as possible because 1st and 2nd generation progeny trials will soon vanish along with the parents used in generating these progenies. It important to capture the genomic relationships among 1st and 2nd generation individuals, to better train the single-step genomic selection model and to provide more accurate estimates of the genetic trend. Research in livestock genetics has shown the negative effects of only using later generational material, which in most cases is the only option available in livestock breeding. Meyer et al. (2018) demonstrated the huge discrepancy between the true genetic trend and trend computed from EBVs in single-step BLUP model that omitted data from earlier generations.

A fourth line of attack was to actively begin single-step genomic selection in radiata pine by trialling a currently available low-density SNP chip developed in New Zealand. The initial testing of the SNP chip was successful in the sense that the SNP loci contained on the chip were segregating in the Australian breeding population. A second consignment of samples was promptly ordered boosting the number of assayed individuals to 945, a few of which were lost to subsequent analysis due to unintended duplication of samples and samples not passing QC. This number of assayed individuals provided an early opportunity to road-test the pedigree forensics pipeline developed jointly by this project and a sister project targeting eucalypts (NIF111-1819). This then led to the discovery of putative systemic errors in the field-based pedigrees stored in TBA's centralised database DATAPLAN and has prompted TBA to initiate further research effort into finding and fixing pedigree errors. This is a major research finding and demonstrates the added benefit of using genomic data in operational breeding. The availability of assayed individuals also enabled the first single-step, national radiata pine TREEPLAN run. As expected, the modest size of the **G** matrix, relative to the size of the complete pedigree, did not lead to substantive increases in accuracy of EBV prediction, but the successful completion of the run does demonstrate the portability of the methodology and this run can serve as a benchmark against which future progress can be compared.

Conclusions

1. Our strategy for incorporating genomic data into forest tree genetic evaluation, namely the adoption of single-step BLUP methodology into TREEPLAN, has been an outstanding success. Since 2017, when an initial pilot single-step run was completed, using *E. globulus* as the target species, TBA has overseen the introduction of single-step genomic step into five species, three of which are conifer species.

2. The New Zealand derived SNP chip is currently our best choice for a low-density, medium cost, dual purpose SNP chip. It has been shown equally useful as an assay for undertaking pedigree forensics, and as an assay for providing data to build a **H** matrix for use in single-step analysis.

3. TBA has amassed on behalf of the Australian industry, a compendium of breeding diversity based on complete genome characterisation of the founder trees of the national breeding population.

4. A comprehensive database of SNP, discovered by examining allele variation in the Australian breeding population, will lead to a more fully functional single-step genomic selection program in the Australian radiata pine industry. Having an Australian designed high-density SNP set should allow TBA to deploy future, low-cost, low-density assay developed by local or overseas genomics service providers

5. Foliage collections comprising high value progeny and their parents have now been completed. These collections will define the training population needed to drive future genomic EBV predictions.

6. Accumulated historical pedigree errors in the radiata pine program appear to be significant enough to require addressing. Finding and correcting these errors will enable reconciliation of genomic and phenotypic data sources and should lead to improvements in EBV accuracy and better selections.

7. The Hi-C analysis of the genome of the genotype targeted in the NZ based genome assembly effort did not proceed as planned. All concerned parties amicably agreed the strategy was not workable within the project time frame, but we are still looking for ways to collaborate in the future.

8. A first pass *de novo* assembly of the radiata genome based on an Australian genotype “96R5114” – a selfed progeny of an elite parent originating in Victoria, has instead been initiated. The task of generating a finished chromosome scale assembly is an enormous undertaking but is becoming increasingly possible due to decreasing costs, breakthroughs in long read sequencing technologies and analysis algorithm development. Having an Australian assembly will ensure freedom to operate in the genomics space and place the Australian industry in a strong position to define its own future.

Recommendations

1. The immediate application of low-density SNP assays to take advantage of the major foliage collections. The results of the assays can be immediately used in TREEPLAN providing a short-term boost to the value of the single-step analysis.
2. TBA must proceed with the definition of a high-density SNP assay. Its application over the major foliage collections will define the training population to drive the development of ultra-low cost, high throughput genotyping assays and enable testing of imputation methods, required for future proofing the technology.
3. Experience in other plant and animal systems and application of genetic theory show that to achieve sufficient accuracy of prediction of genetic values in the radiata pine breeding program it will require a training population (set of trees with both phenotypes and genotypes) of around 20,000. The TBA must continue to push for increased sampling of past, present and future genetic trials, to increase the training population.
4. Development of a purpose designed ultra-low cost, high throughput genotyping assay will be a key driver of genomics adoption and we recommend the development of a multi-species solution fit for purpose for Australian tree breeding should be a priority for development to accelerate adoption of these technologies into routine breeding and deployment activities across all target species.

References

Hill, W.G and Weir, B (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genet. Res*.

Hill, W.G and Weir, B (2012) Variation in actual relationship among descendants of inbred individuals. *Genet. Res.*

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

Legarra *et al*. (2009) A relationship matrix including full pedigree and genomic information. *J Dairy Sci*.

Acknowledgements

Researcher’s Disclaimer (if required)

Appendix 1 - A summary of the sequencing data

Table 15 Summary of sequencing data generated per sample founder genotype

Gid	Genotype Name	Reads	Target Megas	Actual Megas	Average Coverage per Mega	Contribution	Sum Contribution
36074	NZ850-055	193,583,975	8	4	0.66	5.2%	5.2%
36015	A12038	621,354,803	8	5	1.69	5.1%	10.3%
36069	A70052	341,545,644	8	8	0.58	3.2%	13.5%
36016	A12349	509,393,051	8	8	0.87	2.8%	16.3%
10218	A30007	14,899,544	8	3	0.07	2.5%	18.8%
36023	A30026	3,050,931,736	8	5	8.32	2.3%	21.1%
36018	A20055	292,187,442	8	9	0.44	2.1%	23.3%
11097	NZ850-007	13,541,423	8	3	0.06	2.1%	25.3%
36028	A30054	320,638,920	8	8	0.55	2.0%	27.4%
36019	A20064	331,909,853	8	8	0.57	2.0%	29.3%
36070	A70053	224,297,981	8	4	0.76	1.9%	31.2%
11106	NZ850-091	482,494,220	4	8	0.82	1.7%	32.9%
36027	A30050	763,070,378	8	4	2.60	1.7%	34.5%
10179	A20058	466,761,928	8	8	0.80	1.6%	36.1%
36078	NZ850-121	1,570,521,769	8	5	4.28	1.5%	37.6%
10223	A30012	310,820,014	8	8	0.53	1.5%	39.1%
36021	A30002	968,142,233	8	8	1.65	1.4%	40.4%
36047	A52039	160,932,590	8	2	1.10	1.1%	41.6%
11104	NZ850-089	19,111,277	8	3	0.09	1.1%	42.7%
10237	A30028	738,297,876	8	8	1.26	1.0%	43.7%
10142	A12374	210,919,170	8	2	1.44	0.8%	44.5%
10258	A30055	1,027,888,066	8	8	1.75	0.8%	45.3%
10148	A12419	268,640,079	8	4	0.92	0.7%	46.0%
10300	A35102	193,454,454	8	8	0.33	0.6%	46.7%
36029	A35078	11,384,316	8	6	0.03	0.6%	47.3%
10232	A30022	371,968,322	8	8	0.63	0.6%	47.9%
10305	A35120	19,078,189	8	5	0.05	0.6%	48.5%
36024	A30037	363,466,504	8	8	0.62	0.6%	49.1%
36075	NZ850-082	37,009,263	8	2	0.25	0.6%	49.7%
36044	A50048	384,411,408	8	7	0.75	0.6%	50.3%
10245	A30036	12,755,946	8	5	0.03	0.5%	50.8%
10298	A35080	33,601,542	8	6	0.08	0.5%	51.4%
10348	A36008	3,973,542	8	1	0.05	0.5%	51.8%
10151	A12447	817,538,510	8	8	1.39	0.5%	52.3%
10160	A20002	188,057,284	8	7	0.37	0.5%	52.8%
36049	A52051	1,319,354,136	8	8	2.25	0.5%	53.3%
10188	A20080	807,611,228	8	8	1.38	0.5%	53.8%
10079	A10935	486,435,578	8	9	0.74	0.5%	54.2%
36045	A50178	1,150,996,224	8	6	2.62	0.5%	54.7%
11098	NZ850-037	408,165,253	8	8	0.70	0.4%	55.1%
36022	A30014	927,622,998	8	6	2.11	0.4%	55.5%
10086	A10956	525,004,869	8	5	1.43	0.4%	55.9%

Gid	Genotype Name	Reads	Target Megas	Actual Megas	Average Coverage per Mega	Contribution	Sum Contribution
36020	A20085	352,361,487	8	5	0.96	0.4%	56.3%
36031	A35502	291,988,510	8	8	0.50	0.4%	56.7%
36026	A30047	43,087,352	8	7	0.08	0.4%	57.1%
11034	A60027	608,311,158	8	8	1.04	0.4%	57.5%
10184	A20070	303,453,270	8	8	0.52	0.4%	57.9%
10182	A20062	358,114,125	8	4	1.22	0.4%	58.2%
10250	A30043	476,761,930	8	8	0.81	0.4%	58.6%
10410	A50015	789,093,370	4	8	1.35	0.3%	59.0%
11100	NZ850-081	33,927,707	4	4	0.12	0.3%	59.3%
276754	A35506	357,783,250	4	4	1.22	0.3%	59.6%
10226	A30016	292,321,023	4	4	1.00	0.3%	60.0%
10311	A35132	7,991,676	4	2	0.05	0.3%	60.3%
10426	A50080	130,335,292	4	2	0.89	0.3%	60.6%
10087	A10957	261,594,257	4	4	0.89	0.3%	60.9%
11103	NZ850-087	285,740,009	4	2	1.95	0.3%	61.1%
36043	A50047	341,462,858	4	4	1.16	0.3%	61.4%
36025	A30040	21,704,192	4	4	0.07	0.3%	61.6%
10194	A20088	63,136,494	4	4	0.22	0.2%	61.9%
10328	A35701	6,414,339	4	2	0.04	0.2%	62.1%
10307	A35124	5,056,539	4	1	0.07	0.2%	62.3%
10309	A35130	144,906,335	4	4	0.49	0.2%	62.6%
36042	A50045	152,934,930	4	2	1.04	0.2%	62.8%
10310	A35131	189,793,787	4	5	0.52	0.2%	63.0%
10125	A12236	155,713,334	4	3	0.71	0.2%	63.2%
10181	A20061	475,483,865	4	6	1.08	0.2%	63.4%
10183	A20069	180,197,499	4	4	0.61	0.2%	63.5%
10213	A30001	339,153,372	4	4	1.16	0.2%	63.7%
10222	A30011	161,205,640	4	4	0.55	0.2%	63.9%
11108	NZ850-096	552,064,327	4	4	1.88	0.2%	64.1%
10329	A35702	28,917,329	4	2	0.20	0.2%	64.2%
316204	NZ268-609	167,166,337	4	3	0.76	0.2%	64.4%
36068	A70029	325,415,096	4	4	1.11	0.2%	64.5%
10178	A20056	408,169,334	4	4	1.39	0.1%	64.7%
36040	A50006	53,843,326	4	4	0.18	0.1%	64.8%
10389	A36055	46,450,791	4	4	0.16	0.1%	64.9%
276753	A35162	51,519,945	4	1	0.70	0.1%	65.1%
10401	A50001	245,395,203	4	4	0.84	0.1%	65.2%
10220	A30009	257,832,734	4	4	0.88	0.1%	65.3%
11319	A35149	112,505,391	4	4	0.38	0.1%	65.4%
10145	A12403	256,380,123	4	3	1.17	0.1%	65.6%
11016	A60004	6,030,083	4	1	0.08	0.1%	65.7%
10306	A35123	15,603,850	4	2	0.11	0.1%	65.8%
10111	A12112	316,209,209	4	4	1.08	0.1%	65.9%
11315	A35137	10,245,273	4	3	0.05	0.1%	66.0%
10436	A50269	126,474,667	4	3	0.57	0.1%	66.1%

Gid	Genotype Name	Reads	Target Megas	Actual Megas	Average Coverage per Mega	Contribution	Sum Contribution
10423	A50077	29,825,030	4	4	0.10	0.1%	66.2%
36037	A35737	154,688,503	4	4	0.53	0.1%	66.3%
10318	A35154	6,936,006	4	2	0.05	0.1%	66.4%
36030	A35165	13,795,386	4	3	0.06	0.1%	66.5%
10248	A30041	227,100,743	4	4	0.77	0.1%	66.6%
10299	A35086	69,659,954	4	4	0.24	0.1%	66.7%
10425	A50079	125,080,048	4	4	0.43	0.1%	66.7%
36050	A52052	873,250,007	4	4	2.98	0.1%	66.8%
10266	A30067	352,955,847	4	4	1.20	0.1%	66.9%
10216	A30005	4,414,069	4	2	0.03	0.1%	66.9%
10238	A30029	87,726,388	4	4	0.30	0.1%	67.0%
10240	A30031	304,963,996	4	4	1.04	0.1%	67.1%
319096	NZ268-426	207,033,069	4	4	0.71	0.1%	67.1%
10191	A20084	84,565,527	4	4	0.29	0.0%	67.2%
10254	A30048	776,185,496	4	4	2.65	0.0%	67.2%
11314	A35134	203,193,516	4	4	0.69	0.0%	67.3%
10424	A50078	239,144,301	4	4	0.82	0.0%	67.3%
10235	A30025	257,560,866	4	4	0.88	0.0%	67.3%
10252	A30045	329,155,099	4	4	1.12	0.0%	67.4%
10350	A36010	68,177,423	4	4	0.23	0.0%	67.4%
81476	NZ850-077	281,136,455	4	4	0.96	0.0%	67.4%
10221	A30010	180,244,195	4	4	0.61	0.0%	67.5%
10225	A30015	158,091,861	4	4	0.54	0.0%	67.5%
10227	A30017	378,485,316	4	4	1.29	0.0%	67.5%
10431	A50176	133,705,670	4	4	0.46	0.0%	67.6%
10320	A35163	75,541,866	4	4	0.26	0.0%	67.6%
10324	A35507	74,970,591	4	4	0.26	0.0%	67.6%
10325	A35508	54,225,655	4	4	0.18	0.0%	67.6%
10407	A50010	190,340,562	4	4	0.65	0.0%	67.6%
10295	A35016	50,903,212	4	4	0.17	0.0%	67.7%
10353	A36013	70,983,793	4	4	0.24	0.0%	67.7%

Appendix 2 - Detailed results of SEQUOIA run in *P. radiata*

Genotype id	Progeny type	Trial	Mum id	Dad id	Inferred mum	Inferred dad	Number sibs	Number assayed	Number mis matched	Details of other sibs with mismatches
4665537	CP	RES1295	10328	42576	F0019	OK	94	1	1	
8703289	CP	BRGT1301	10410	99912	F0061	10410	181	1	1	
8704122	CP	BRGT1301	10421	99281	OK	M0064	160	1	1	
42689	OP	RAD137	11099	0	36015	0	1435	1	1	
8704245	CP	BRGT1301	11103	99281	OK	36069	161	1	1	
545253	CP	BR9617	36015	41707	F0007	175437	34	1	1	
207135	CP	BR9705	36015	41779	F0010	M0024	596	2	2	8701931 (F0010 x M0024) [BRGT1301]
104281	CP	BR9601	36015	41996	F0021	M0024	539	6	6	344797 (F0007 x M0016) [BR9611] 679255 (F0007 x M0016) [BR9701] 8703094 (F0021 x M0024) [BRGT1301] 206221 (F0007 x M0016) [BR9705] 187834 (F0007 x M0016) [BR9703]
99085	CP	BR9601	36015	42180	F0007	OK	219	1	1	
101170	CP	BR9601	36015	42199	F0007	M0011	283	2	2	174394 (F0007 x M0011) [BR9606]
913750	CP	BR9614	36015	42658	F0007	M0008	345	5	5	915317 (F0007 x M0008) [BR9614] 913751 (F0007 x M0008) [BR9614] 914866 (F0007 x M0008) [BR9614] 914865 (F0007 x M0008) [BR9614]
913582	CP	BR9614	36015	42721	F0007	M0018	515	3	2	915100 (F0007 x M0018) [BR9614]
544782	CP	BR9617	36015	277780	F0007	OK	130	1	1	
543979	CP	BR9617	36015	277786	F0007	OK	144	1	1	
543457	CP	BR9617	36015	277796	F0007	OK	125	1	1	
8703059	CP	BRGT1301	36016	42571	F0053	36016	78	1	1	
2413156	CP	GT0002	36018	36069	36069	11104	140	2	2	2446674 (F0019 x M0024) [GT0001]
104654	CP	BR9601	36021	42001	F0045	M0051	332	2	2	918163 (F0045 x M0051) [BR9613]
210392	CP	BR9707	36021	42016	F0045	OK	400	4	3	909189 (F0045) [BR9713] 8704591 (F0045) [BRGT1301]
8704130	CP	BRGT1301	36021	42016	F0045	41776	400	4	1	
8701919	CP	BRGT1301	36044	36015	OK	M0052	806	3	1	
8704580	CP	BRGT1301	36044	41996	OK	M0052	440	3	1	
206947	CP	BR9705	36044	41996	F0021	M0052	440	3	2	8704469 (F0021 x M0052) [BRGT1301]
8702149	CP	BRGT1301	36044	42571	F0053	M0052	62	1	1	
8702142	CP	BRGT1301	36044	99449	OK	M0052	128	1	1	
8702139	CP	BRGT1301	36044	99912	OK	M0052	219	1	1	
4664526	CP	RES1295	36069	41709	OK	M0006	321	1	1	
4125118	CP	BR0903	36069	41996	OK	M0064	528	3	1	
207548	CP	BR9705	36069	41996	F0021	M0064	528	3	2	779702 (F0021 x M0064) [BR9702]
8701911	CP	BRGT1301	36069	42588	F0009	M0064	370	1	1	
101682	CP	BR9601	36069	42661	F0014	M0064	290	3	3	184426 (F0014 x M0064) [BR9615] 345145 (F0014 x M0064) [BR9611]
8703443	CP	BRGT1301	36069	42662	OK	M0064	238	1	1	
4664459	CP	RES1295	36074	42001	F0037	OK	20	1	1	
4662768	CP	RES1295	36074	42577	F0039	OK	18	1	1	
8704590	CP	BRGT1301	36077	10328	10425	M0064	110	1	1	
8702141	CP	BRGT1301	41707	36016	OK	M0069	57	4	2	8739090 (M0069) [BRGT1303]
709193	CP	GT9506	41707	36047	OK	M0049	373	1	1	
8702958	CP	BRGT1301	41707	41996	OK	M0043	587	3	2	8737698 (M0043) [BRGT1303]
103894	CP	BR9601	41707	42721	OK	M0018	138	2	2	545183 (M0047) [BR9617]
8704466	CP	BRGT1301	41709	42575	OK	M0006	45	1	1	
103177	CP	BR9601	41709	42576	F0019	M0006	392	4	4	544671 (F0019 x M0006) [BR9617] 8703062 (F0019 x M0006) [BRGT1301] 103178 (F0019 x M0006) [BR9601]
171764	CP	BR9606	41709	42577	F0039	M0006	91	1	1	
544330	CP	BR9617	41709	42661	F0014	M0006	185	1	1	
8702269	CP	BRGT1301	41709	277802	OK	M0006	101	1	1	
8702152	CP	BRGT1301	41709	345843	F0042	OK	143	1	1	
343882	CP	BR9611	41710	277814	277781	OK	145	2	1	

Genotype id	Progeny type	Trial	Mum id	Dad id	Inferred mum	Inferred dad	Number sibs	Number assayed	Number mis matched	Details of other sibs with mismatches
345219	CP	BR9611	41710	277814	F0007	M0022	145	2	1	
100321	CP	BR9601	41728	42270	F0040	M0021	77	2	2	103884 (F0040 x M0021) [BR9601]
8141350	CP	BRGT1403	41731	36047	36047	M0021	15	2	2	8143851 (36047 x M0021) [BRGT1403]
100310	CP	BR9601	41731	36069	36069	M0021	42	2	2	101778 (36069 x M0021) [BR9601]
187855	CP	BR9703	41731	41996	F0021	M0021	341	8	8	206920 (F0021 x M0021) [BR9705] 778329 (F0021 x M0021) [BR9702] 912134 (F0021 x M0021) [BR9713] 188066 (F0021 x M0021) [BR9703] 680072 (F0021 x M0021) [BR9701] 8141604 (F0021 x M0021) [BRGT1403] 909836 (F0021 x M0021) [BR9713]
188581	CP	BR9703	41731	42577	F0039	M0021	244	2	2	204850 (F0039 x M0021) [BR9705]
179083	CP	BR9615	41776	36047	F0020	M0049	235	1	1	
176339	CP	BR9606	41776	36049	36049	M0022	123	1	1	
170242	CP	BR9606	41776	41996	F0021	M0022	270	3	3	347130 (42012 x M0022) [BR9611] 912676 (F0021 x M0022) [BR9614]
8703302	CP	BRGT1301	41776	42146	42146	M0022	55	1	1	
100500	CP	BR9601	41776	42576	F0019	M0022	182	5	5	915178 (F0019 x M0022) [BR9614] 170752 (F0019 x M0022) [BR9606] 178962 (F0019 x M0022) [BR9615] 103831 (F0019 x M0022) [BR9601]
6204267	CP	BRGT1201	41776	100567	F0020	OK	36	4	4	6206236 (F0020) [BRGT1201] 6207143 (F0020) [BRGT1201] 6207272 (F0020) [BRGT1201]
8701818	CP	BRGT1301	41776	206221	F0037	M0022	55	1	1	
8703074	CP	BRGT1301	41779	42146	F0010	OK	541	2	1	
207419	CP	BR9705	41779	42590	F0010	42689	315	1	1	
8701817	CP	BRGT1301	41839	185891	F0070	M0060	121	1	1	
346397	CP	BR9611	41848	41996	F0021	M0041	365	2	2	4112027 (F0021 x M0041) [BR0901]
335245	CP	RAD238	41977	42083	F0033	M0040	58	1	1	
334789	CP	RAD238	41977	42566	42566	M0040	58	1	1	
101865	CP	BR9601	41996	42576	F0018	M0016	318	8	8	180843 (F0019 x M0016) [BR9615] 2096791 (F0019 x M0016) [BR0803] 1395153 (F0019 x M0043) [BR0602] 4107698 (F0018 x M0043) [BR0901] 4127982 (F0019 x M0016) [BR0904] 1395152 (F0018 x M0043) [BR0602] 344731 (F0019 x M0016) [BR9611]
676941	CP	BR9701	41996	42586	F0016	M0016	233	1	1	
101752	CP	BR9601	41996	42661	F0014	M0043	147	4	3	103582 (F0014 x M0043) [BR9601] 543314 (F0014 x M0043) [BR9617]
8704247	CP	BRGT1301	42000	42575	OK	M0060	63	1	1	
168765	CP	BR9606	42001	36015	F0010	M0024	565	4	4	343362 (F0037 x M0024) [BR9611] 4664965 (F0037 x M0024) [RES1295] 8702160 (F0037 x M0024) [BRGT1301]
8703327	CP	BRGT1301	42001	42576	F0037	M0060	23	1	1	
8704593	CP	BRGT1301	42001	42577	F0039	M0051	62	1	1	
2096506	CP	BR0803	42016	41776	F0020	OK	486	3	3	8702285 (F0020) [BRGT1301] 4110090 (F0020) [BR0901]
1144280	CP	BR9604	42016	41895	F0035	OK	371	5	3	1144717 (F0035) [BR9604] 8704468 (F0035) [BRGT1301]
919460	CP	BR9613	42016	41895	F0007	M0018	371	5	1	
4665131	CP	RES1295	42120	42084	41707	OK	118	1	1	
4108402	CP	BR0901	42120	345477	F0022	OK	73	5	5	4125335 (F0022) [BR0903] 4123880 (F0022) [BR0903] 4129174 (F0022)

Genotype id	Progeny type	Trial	Mum id	Dad id	Inferred mum	Inferred dad	Number sibs	Number assayed	Number mis matched	Details of other sibs with mismatches
										[BR0904] 4111449 (F0022) [BR0901]
8701379	CP	BRGT1301	42120	345843	F0042	OK	111	2	1	
8701814	CP	BRGT1301	42120	345843	OK	175437	111	2	1	
8701930	CP	BRGT1301	42126	277849	F0043	OK	127	1	1	
344654	CP	BR9611	42139	36049	36049	M0026	206	2	2	543736 (36049 x M0026) [BR9617]
169212	CP	BR9606	42139	41996	F0021	M0026	131	2	2	176392 (F0021 x M0026) [BR9606]
207213	CP	BR9705	42139	42016	F0035	M0026	90	1	1	
918938	CP	BR9613	42139	42421	F0013	M0026	170	1	1	
8703303	CP	BRGT1301	42139	277697	F0064	OK	66	1	1	
99791	CP	BR9601	42146	36047	OK	M0030	165	9	8	104000 (M0030) [BR9601] 914492 (M0049) [BR9614] 345688 (M0049) [BR9611] 178631 (M0030) [BR9615] 176131 (M0030) [BR9606] 913345 (M0049) [BR9614] 347547 (M0030) [BR9611]
170589	CP	BR9606	42146	36047	41707	M0039	165	9	1	
8702956	CP	BRGT1301	42146	41779	F0010	M0030	541	2	1	
8703439	CP	BRGT1301	42146	41998	OK	M0030	435	1	1	
101667	CP	BR9601	42146	42084	F0073	M0030	370	3	3	102767 (F0073 x M0030) [BR9601] 348101 (F0073 x M0030) [BR9611]
102071	CP	BR9601	42146	42174	OK	M0030	445	1	1	
175564	CP	BR9606	42146	42421	F0013	M0030	166	3	3	920025 (F0013 x M0030) [BR9613] 176047 (F0013 x M0030) [BR9606]
347898	CP	BR9611	42146	42516	OK	M0030	350	2	1	
545490	CP	BR9617	42146	42516	F0025	M0030	350	2	1	
545555	CP	BR9617	42146	42586	F0016	M0030	244	2	2	913726 (F0016 x M0030) [BR9614]
207038	CP	BR9705	42146	42590	42689	M0030	183	1	1	
8702954	CP	BRGT1301	42146	42772	F0069	M0030	187	1	1	
4129267	CP	BR0904	42146	345477	F0022	M0030	74	1	1	
188240	CP	BR9703	42149	36021	F0045	OK	318	1	1	
207361	CP	BR9705	42149	42139	OK	M0026	386	2	1	
8703407	CP	BRGT1301	42149	42586	F0016	OK	553	1	1	
910854	CP	BR9713	42149	42590	OK	42689	165	1	1	
4127239	CP	BR0904	42156	42194	42194	M0014	163	1	1	
8701805	CP	BRGT1301	42156	277821	OK	M0014	106	1	1	
205432	CP	BR9705	42174	42242	OK	M0028	181	1	1	
181387	CP	BR9615	42174	42516	OK	M0070	327	3	2	347085 (M0070) [BR9611]
170689	CP	BR9606	42174	42516	F0025	M0028	327	3	1	
505522	CP	VRC095	42179	42211	F0043	OK	40	1	1	
173656	CP	BR9606	42194	36047	OK	M0049	312	1	1	
102631	CP	BR9601	42194	42242	F0049	42194	246	2	2	102635 (F0049 x 42194) [BR9601]
99770	CP	BR9601	42194	42251	OK	M0013	496	2	2	99773 (M0013) [BR9601]
335176	CP	RAD238	42198	42083	F0033	M0056	60	1	1	
334091	CP	RAD238	42198	42566	42566	M0056	59	1	1	
102315	CP	BR9601	42199	41779	F0010	M0011	317	4	4	543573 (F0010 x M0011) [BR9617] 183363 (F0010 x M0011) [BR9615] 919366 (42218 x M0039) [BR9613]
8701804	CP	BRGT1301	42211	42773	F0043	OK	123	1	1	
102963	CP	BR9601	42215	42421	F0013	M0044	267	2	2	104380 (F0013 x M0044) [BR9601]
205663	CP	BR9705	42218	42146	OK	M0030	416	3	3	207451 (M0030) [BR9705] 8701921 (M0030) [BRGT1301]
100557	CP	BR9601	42251	42421	F0013	M0013	384	2	2	100558 (F0013 x M0013) [BR9601]
99410	CP	BR9601	42251	42586	F0016	M0013	337	2	2	100261 (F0016 x M0013) [BR9601]
99491	CP	BR9601	42254	41709	OK	M0006	346	2	2	102562 (M0006) [BR9601]
102474	CP	BR9601	42270	36042	F0040	OK	273	2	2	104901 (F0040) [BR9601]
8704477	CP	BRGT1301	42270	42146	OK	M0030	445	1	1	
207156	CP	BR9705	42270	42576	F0019	OK	426	1	1	
2411566	CP	GT0002	42354	42827	F0009	M0033	74	3	2	2412947 (F0009 x M0033) [GT0002]
4123661	CP	BR0903	42360	103222	42120	205494	101	1	1	

Genotype id	Progeny type	Trial	Mum id	Dad id	Inferred mum	Inferred dad	Number sibs	Number assayed	Number mis matched	Details of other sibs with mismatches
4124898	CP	BR0903	42360	207548	OK	M0058	317	1	1	
4126608	CP	BR0903	42360	812842	345019	105265	156	1	1	
4109496	CP	BR0901	42362	42123	OK	M0033	80	1	1	
912214	CP	BR9713	42516	36021	F0045	M0070	141	1	1	
188421	CP	BR9703	42516	36047	F0035	41895	221	2	2	207278 (36047 x M0039) [BR9705]
8702944	CP	BRGT1301	42571	42174	F0053	M0028	61	1	1	
8703415	CP	BRGT1301	42575	41839	F0070	M0060	58	1	1	
544225	CP	BR9617	42576	41779	F0019	OK	701	4	4	8703314 (F0010) [BRGT1301] 912699 (F0019) [BR9614] 8703429 (F0018) [BRGT1301]
8701438	CP	BRGT1301	42576	207548	F0019	OK	48	1	1	
8701924	CP	BRGT1301	42582	36044	OK	M0052	150	1	1	
347386	CP	BR9611	42588	42586	F0016	42588	180	1	1	
545420	CP	BR9617	42588	42721	F0009	M0047	379	1	1	
347874	CP	BR9611	42589	42576	F0019	M0017	233	3	3	914570 (F0018 x M0017) [BR9614] 3563759 (F0019 x M0017) [BR9609]
4663319	CP	RES1295	42589	277817	F0015	M0017	20	3	3	4664095 (F0015 x M0017) [RES1295] 4665042 (F0015 x M0017) [RES1295]
4664390	CP	RES1295	42589	277825	F0027	M0017	20	2	2	4665105 (F0027 x M0017) [RES1295]
100082	CP	BR9601	42590	36021	36021	42689	326	2	2	344370 (36021 x 42689) [BR9611]
188073	CP	BR9703	42590	36047	42689	M0049	184	3	3	188076 (36047 x 42689) [BR9703] 681034 (36047 x 42689) [BR9701]
8703294	CP	BRGT1301	42590	42083	42689	OK	184	1	1	
99292	CP	BR9601	42590	42194	42689	OK	245	2	2	100311 (42689) [BR9601]
543327	CP	BR9617	42590	42374	42689	OK	230	1	1	
8703293	CP	BRGT1301	42590	179475	42689	OK	132	1	1	
8704236	CP	BRGT1301	42590	185891	42689	OK	111	1	1	
8703404	CP	BRGT1301	42590	186152	42689	OK	88	1	1	
183897	CP	BR9615	42591	42174	OK	M0039	188	1	1	
102444	CP	BR9601	42591	42218	42218	M0039	251	2	2	103634 (F0009 x M0064) [BR9601]
4124056	CP	BR0903	42658	41776	42012	M0008	564	4	1	
206733	CP	BR9705	42658	41779	F0010	M0008	468	2	2	544620 (F0010 x M0008) [BR9617]
8703444	CP	BRGT1301	42658	41996	F0021	M0008	382	3	1	
8733987	CP	BRGT1303	42658	42146	OK	M0030	181	2	1	
8703097	CP	BRGT1301	42658	42146	42146	42827	181	2	1	
8703081	CP	BRGT1301	42658	42571	F0053	M0028	59	1	1	
8703324	CP	BRGT1301	42658	99281	OK	M0064	177	1	1	
176247	CP	BR9606	42661	41779	F0014	OK	460	1	1	
101808	CP	BR9601	42661	42001	F0014	OK	168	1	1	
182738	CP	BR9615	42721	36069	OK	M0064	363	4	3	345230 (M0064) [BR9611] 915252 (M0064) [BR9614]
8703299	CP	BRGT1301	42721	36069	F0039	M0047	363	4	1	
187830	CP	BR9703	42721	42270	F0040	M0018	373	3	3	205067 (F0040 x M0018) [BR9705] 8702161 (F0040 x M0018) [BRGT1301]
188426	CP	BR9703	42721	42586	F0016	M0018	514	4	4	911353 (F0016 x M0018) [BR9713] 211089 (F0016 x M0018) [BR9707] 8703061 (F0016 x M0018) [BRGT1301]
104412	CP	BR9601	42731	42001	F0037	OK	344	1	1	
176369	CP	BR9606	42731	42661	F0014	OK	276	1	1	
8703082	CP	BRGT1301	42827	104685	F0068	42827	113	1	1	
8704235	CP	BRGT1301	42827	345843	F0042	OK	129	2	1	
8702264	CP	BRGT1301	104412	101591	OK	345359	112	1	1	
8701944	CP	BRGT1301	206074	343875	OK	M0067	152	1	1	
8702289	CP	BRGT1301	206074	345843	F0042	M0067	142	1	1	
1386158	CP	GT0001	276753	10389	10226	OK	153	2	1	
1388660	CP	GT0001	276753	10389	10226	36074	153	2	1	
1474503	CP	VRC070	276755	277850	OK	M0008	48	1	1	
1474164	CP	VRC070	276755	277853	OK	M0051	51	1	1	
8698923	CP	BRGT1301	277685	42194	F0065	OK	42	2	2	8703083 (42194) [BRGT1301]
1475145	CP	VRC071	277733	277839	OK	M0057	194	2	1	

Genotype id	Progeny type	Trial	Mum id	Dad id	Inferred mum	Inferred dad	Number sibs	Number assayed	Number mis matched	Details of other sibs with mismatches
4665013	CP	RES1295	277782	277817	F0015	M0037	20	2	2	4665371 (F0015 x M0037) [RES1295]
4663531	CP	RES1295	277783	277812	F0057	OK	19	2	1	
8702979	CP	BRGT1301	277787	277812	F0057	OK	90	1	1	
4663324	CP	RES1295	277815	277804	F0038	OK	19	2	2	4665127 (F0038) [RES1295]
4665452	CP	RES1295	277817	277781	F0015	OK	69	2	2	8701826 (F0015) [BRGT1301]
4664308	CP	RES1295	277817	277783	F0015	277812	18	1	1	
8701819	CP	BRGT1301	343569	180297	OK	99436	112	1	1	
6207419	CP	BRGT1201	345019	175880	42012	OK	13	1	1	
8698986	CP	BRGT1301	345230	183244	F0059	345230	133	3	3	8703406 (F0059 x 345230) [BRGT1301] 8702948 (F0059 x 345230) [BRGT1301]
8702255	CP	BRGT1301	345417	345359	345359	345219	97	2	1	
4108731	CP	BR0901	345843	170242	OK	M0068	34	2	1	
2770135	OP	Q14/1.38	1136432	0	F0004	0	10	1	1	
2770225	OP	Q14/1.38	1136436	0	F0004	0	38	1	1	
2770140	OP	Q14/1.38	1136438	0	F0004	0	37	1	1	
2770409	OP	Q14/1.38	1136439	0	F0004	0	11	1	1	
2770039	OP	Q14/1.38	1136442	0	F0004	0	5	1	1	
2769464	OP	Q14/1.38	1686709	0	F0065	0	62	1	1	
2769604	OP	Q14/1.38	1686716	0	F0005	0	43	1	1	
2769554	OP	Q14/1.38	1686722	0	F0005	0	50	1	1	
2769767	OP	Q14/1.38	1686724	0	F0005	0	47	1	1	
2769699	OP	Q14/1.38	2686112	0	F0005	0	17	1	1	