Final Report
Project NV072

NATIONAL INSTITUTE FOR
FOREST PRODUCTS INNOVATION

Sustained productivity gains in softwood plantations through enablement of single-step genomic selection

2025

Gippsland Centre

Funded by the Australian Government, Victorian Government & Industry Partners.

nifpi.org.au

NATIONAL INSTITUTE FOR
FOREST PRODUCTS INNOVATION
GIPPSLAND

Sustained productivity gains in softwood plantations through enablement of single-step genomic selection

Prepared for

National Institute for Forest Products Innovation

Gippsland

by

Dr Richard Kerr, Dr Josquin Tibbits, Dr Dominic Kain & Dr Tony McRae

Publication: Sustained productivity gains in softwood plantations through enablement of single-step genomic selection

Project No: NIF163-2122 [NV072]

IMPORTANT NOTICE

ISBN: 978-1-922718-57-0

Researcher:

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

Dr Dominic Kain
HQPlantations

Dr Josquin Tibbits
DEECA

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 Victorian Government.

Australian Government
Department of Agriculture, Fisheries and Forestry

Executive Summary

This project aimed to enhance the efficiency and accuracy of genetic selection in *Pinus radiata* breeding programs by integrating genomic data into traditional pedigree-based approaches. The primary focus was on applying single-step genomic selection (SSGS) methodology to improve the prediction accuracy of breeding values, thus accelerating genetic gain in Australian *P. radiata* populations. The project also explored methods for optimizing DNA sampling, processing, and genotyping for large-scale tree improvement efforts.

A key component of the project involved the use of the Axiom PRAD array, developed by an overseas consortium, with the capacity to call approximately 36,000 genetic variants. However, through rigorous filtering steps—based on Hardy-Weinberg Disequilibrium (HWD), Minor Allele Frequency (MAF), and custom tests using Opposing Homozygous Loci (OHL) analysis — the set was reduced to 8,871 variants. This highlighted the need for custom genotyping arrays tailored to Australian populations, and it reinforced the importance of obtaining a chromosome-level full genome assembly for *P. radiata*, which will enable more efficient genotyping in the future.

The project also aimed to verify the accuracy of pedigrees in breeding programs. The estimated pedigree error rate was between 4% and 5%, a low figure that reflects decades of meticulous breeding efforts. This result is particularly important for the project sponsors, including the Gippsland Centre of the National Institute for Forest Products Innovation (NIFPI) and HVP Plantations (HVP), whose orchards and archives were the focus of identity verification testing. Results from foliage sampling at Gelliondale showed a high level of accuracy in clonally replicated genotypes, which is critical for deployment programs.

An important milestone was the expansion of the genomic training population for single-step BLUP (Best Linear Unbiased Prediction). Previous analyses based on a smaller set of 2,000 trees did not yield significant improvements in prediction accuracy. However, with the training population now expanded to approximately 5,000 trees, meaningful gains in predictive accuracy have been observed. Furthermore, combining this expanded training population with enhanced pedigree accuracy has demonstrated a clear improvement in the reliability of breeding value predictions, marking a significant advancement in the application of SSGS methodology.

Although some experiments, such as those on DNA storage duration, were not completed by the project's end, the impact of correcting putative pedigree errors with high confidence has been assessed. This work confirmed the value of identifying and correcting pedigree errors for improving prediction accuracy and ensuring the reliability of genetic evaluations. Remaining experiments will continue under a sister project, "*Using genomics to double the rate of genetic gain in Australian forest tree improvement programs*" (VNC580-2122), ensuring continuity and comprehensive reporting of results.

In conclusion, the project has made significant progress in improving the genetic evaluation of *P. radiata* through genomic integration and pedigree validation. These advancements are expected to support faster, more accurate selection processes, ultimately accelerating genetic gain in Australian forestry.

NIFPI
Suite 6.03 , 36 Wellington St, Collingwood, Victoria, 3066
T +61 3 9927 3200
E info@nifpi.org.au
W www.nifpi.org.au

Table of Contents

Executive Summary ii
Introduction 1
Methodology 1
Field Sampling and DNA Processing 1
Array processing and SNP calling 2
Post-calling quality assurance checks 3
The pedigree and identity assurance pipeline 3
Single-Step Genetic Evaluation 4
Results 6
Field Sampling and DNA Processing 6
Array processing and SNP calling 6
Post-calling quality assurance checks 9
The pedigree and identity assurance pipeline 11
Determining the OHL rate cutoff 11
Duplicate checking 12
Sub-study 1 – across eras 13
Sub-study 3 – Emerging CP 17
Sub-study 4 – Individual Audits 17
Unintentional duplicates 17
Dyad error detection and first-instance recovery 18
Second-Round Recovery Pipeline 20
Final Thoughts: 21
Single-Step Genetic Evaluation 22
Pedigree Recovery in Action: Enhancing Single-Step Analysis 28
Uncovering Pedigree Errors: Key Case Studies 28
EBV Comparisons Pre- and Post-Pedigree Recovery 32
Discussion 35
Limitations and Future Work 36
Conclusions 37
Recommendations 38
References 39
Acknowledgements 40
Researcher's Disclaimer (if required) 40

Introduction

Genomic selection offers a transformative potential to accelerate the genetic improvement of radiata pine in Australia's national breeding program. By integrating genomic information, selection decisions become more accurate and are made earlier, significantly reducing the generation interval. This leads to a faster accumulation of genetic gains over time. Genomic data provide detailed insights into genetic potential, improving the reliability of breeding value predictions and the effectiveness of selection in breeding programs. Single-Step Genomic Selection (Legarra *et al*. 2009) is already well-implemented in *E. globulus*, thanks to the species' relatively simpler genome compared to that of radiata pine. The complexity and large size of the radiata pine genome have slowed the progress of genomic selection, as acquiring the necessary genomic resources is more demanding.

TBA is currently using SCION's Prad Axiom array, capable of generating genotype calls for up to 32,000 SNPs (Single Nucleotide Polymorphisms). This array-based approach allows for an efficient and relatively cost-effective method of genotyping, facilitating the initial stages of genomic selection in *P. radiata*. Alongside the short-term use of SNP arrays, efforts are underway to obtain whole genome sequencing data for *P. radiata*. This long-term approach will provide comprehensive genomic information, enhancing the accuracy of selection and long-term genetic improvement.

Breeding program operating funds and resources from previous FWPA projects — 'Tools, systems and enabling genetic technologies for pines and eucalypts' (VNC515-1920) and 'Quality assurance in the pedigree of radiata pine' (VNC561-2021) — supported the processing of 959 samples at an international laboratory in California and 1,920 samples at a local laboratory in Sydney, with the Prad Axiom array. The analysis of the raw data provided by these consignments was reported in the final report for VNC561-2021. In summary, the overall error rate in the national *P. radiata* pedigrees was found to be less than 5%. However, identifying the correct parent when erroneous parent-offspring pairs were detected was limited because too few historical parents had been assayed. Nonetheless, the low error rate was a positive outcome, as previous tests using a different methodology had suggested a significantly higher error rate.

The current project aimed to further validate the *P. radiata* national breeding program pedigree and confirm the identities of key selections in arboreta and orchards in the Gippsland and other regions. To complete this objective a further 7200 foliage samples were collected, with DNA extracted and assayed at the Ramaciotti centre in Sydney.

The total number of individuals assayed and passing quality control checks exceeded 9,000, approaching the critical threshold for the minimum size of a training population in single-step BLUP genetic evaluation. Therefore, a secondary aim of the project is to test the validity of single-step genomic selection in a conifer species using a "modest" size SNP array.

The project has also provided an opportunity to test sampling procedures and gain experience in submitting samples to various laboratories, as well as addressing challenges and issues that arise during the transfer of DNA between laboratories.

Methodology

Field Sampling and DNA Processing

Our field sampling and DNA processing protocols were designed to support large-scale tree improvement projects, focusing on efficiency and precision. Key optimizations included the use of specific envelope sizes and silica gel for sample integrity, a robust labelling system for tracking, and

careful selection of fascicles for consistency. The preparation of 96-well sample plates was refined to ensure standardized DNA extraction and compatibility with high-throughput sequencing.

Collaboration with key partners—TBA, AGRF, Ramaciotti Centre, and ThermoFisher Australia—enabled efficient communication and rapid resolution of issues, further enhancing workflow efficiency.

To test DNA yield and quality, a small experiment was planned using a cross-classification of storage treatments. Samples included fresh foliage (needles), and foliage stored for 2 weeks, 6 months, 2 years, and 4 years. Needle sample sizes for each storage treatment were set at 100 mg of wet tissue (for fresh samples) and 30 mg, 60 mg, and 90 mg of dried tissue for stored samples. This experiment will help refine the optimal conditions for long-term DNA storage and extraction.

Array processing and SNP calling

Raw SNP data were analyzed using Axiom Analysis Software (AxAS) Suite and the Best Practices Workflow (Affymetrix, Santa Clara, CA). The default protocol established for conifer species was used initially. This entails filtering:

- SNP using a SNP call rate cut-off (snp_cr_cutoff) > 97%
- samples using a Dish-QC threshold (dish_qc_cutoff) > 82%
- samples using a sample call rate (sample_cr_cutoff) > 90%
- plates using a percent of passing samples (plate_qc_cutoff) > 95%.

Initially, the AxAS was used to analyze individual projects as separate batches. The following table lists the projects along with the number of samples in each:

Table 1 Consignments relevant to this project

Project	**Laboratory**	Number of array plates	Number samples
a551114	ThermoFisher, Santa Clara	10	959[1]
CUN11425	Ramaciotti Centre	5	1920
CUN13262	Ramaciotti Centre	5	1920
CUN13548	Ramaciotti Centre	4	1536
CUN13900	Ramaciotti Centre	5	1920
CUN13931	Ramaciotti Centre	3	1152
CUN14278	Ramaciotti Centre	2	768
TOTAL			**10,175**

[1] The number of samples in project a551114 is not a multiple of 384, as the array plates contained mixtures of TBA and RPBC material

The Axiom Batch SSP Tool is a standalone software tool used for generating SNP Specific Priors (SSPs) for Axiom arrays. These SNP-specific priors enhance genotyping accuracy and consistency. During genotyping, prior models for a probeset provide the algorithm with information about the expected position and size of clusters for that probeset.

Initially, the tool used the best-performing array plates from each project batch, analyzed as a batch by AxAS (named “Best”), as input. The output is a models file, which serves as a priors template for subsequent AxAS analysis batches. All project batches were then combined into a single batch (named “Combined”) and analyzed using this new models file.

Upon completing a batch run, AxAS classifies SNPs into six categories: OTV, Other, CallRateBelow-Threshold, NoMinorHom, MonoHighResolution, and PolyHighResolution. The most reliable SNPs fall into the NoMinorHom, MonoHighResolution, and PolyHighResolution categories.

It was observed that the “Combined” batch had significantly fewer PolyHighResolution SNPs and more MonoHighResolution SNPs compared to the separate batches. This resulted in a lower overall number of the best and recommended SNPs. The differences in call rates between the “Best” batch, in which most data is from the a551114 project, and the “Combined” batch appear to be due to variations in DQC signal strength. The “Combined” batch seems to have a lower overall signal, causing the allele clusters to be closer together. This makes it harder for the software to accurately call genotypes, leading to a reduction in the best and recommended call rates.

It was decided to create a separate models file for the a551114 batch and another for a run that combined all Ramaciotti data into a single batch (named “Combined_Ramaciotti”). In each case, the best plates from a551114 and Ramaciotti data were selected to create the respective models files.

SNP calls for each sample were exported from the AxAS suite in variant call format (VCF). VCF files were obtained from the a551115 and Combined_Ramaciotti runs. A merged VCF file was then created by joining the data on common probe set IDs.

Although VCF files have become the standard for transferring genotype call data among researchers and organizations, they are not well-suited for downstream processing. Therefore, the VCF file was converted to a tabular format for use in the identity and pedigree assurance pipeline.

Post-calling quality assurance checks

Post-calling quality assurance checks were performed to ensure the accuracy and reliability of the genetic data following SNP calling, specifically addressing any discrepancies or errors arising from the inadequate design of probe sets. These probe sets can sometimes mark non-unique sites in the genome, leading to potential inaccuracies.

Check 1 MAF and Missingness. Remove SNP with a minor allele frequency (MAF) < 0.01 and to remove samples with a missingness value (percentage of SNP not called) > 10%. This check can be performed using popular software packages such as BCFTOOLS or PLINK, via simple command-line instructions.

Check 2 HWD. Filter SNPs based on Hardy-Weinberg disequilibrium (HWD). When a population meets Hardy-Weinberg proportions, the disequilibrium coefficient (D_A) is zero. Following the approach of Dodds et al. (2015), plotting the disequilibrium coefficient against the MAF for each SNP helps identify problematic SNPs. The plot's points are color-coded to indicate the strength of the alternative hypothesis, which posits that the SNP is not in Hardy-Weinberg equilibrium (HWE). The plot's upper boundary represents SNPs with maximum homozygosity at a given MAF, while the lower boundary represents SNPs with maximum heterozygosity. These boundaries create a "fin-like" appearance, hence the term FIN plots.

Check 3 High OHL fraction across dyads. Remove SNPs that exhibit an excessively high number of parent-offspring (dyad) misassignments. While a certain level of dyad errors is expected due to the imperfections in both field pedigree records and assay chemistry, an unusually high number of dyad errors may indicate issues with the SNP itself. Therefore, a SNP-by-SNP calculation of the fraction of dyads displaying opposing homozgyous loci (OHL) occurrences serves as an additional quality assurance measure. This check also ensures that the remaining SNPs post filtering were not retained solely because of a low MAF. SNPs with very low MAF have a lower likelihood of displaying OHL, as they usually present only one type of homozygote. To confirm that this is not the case, a complete distribution of MAFs ranging from 0.0 to 0.5 should be observed for SNPs that show zero to a small fraction of dyads with OHL.

The pedigree and identity assurance pipeline

The aim of the pipeline is to implement robust, “industrial” strength algorithms that require minimal human invention. The principal researcher intentionally avoided likelihood-based methods, such as those used by the R package Sequoia (Huisman 2017), because these methods (A) struggle to handle large sample sizes and high SNP counts, and (B) are not well-suited to the specific needs of forest tree data. The likelihood-based methods are typically designed for human genetics and assume limited prior knowledge about the samples. For instance, Sequoia lacks the capability to recognize tree planting dates, which prevents it from ruling out individuals as potential true parents based on age.

In-house software has been developed to efficiently handle large SNP datasets and sample sizes, while also incorporating all available data on planting dates, ortet locations, and other relevant information. Furthermore, the underlying hypothesis is that the field pedigrees are generally accurate, with the software’s role being to pinpoint any errors. In contrast, maximum likelihood programs take the opposite approach: they assume no prior knowledge, calculate all possible parent-offspring pairings based solely on genomic data, and then compare these findings with existing records.

The underlying methodology is Opposing Homozygous Locus (OHL) and is preferred due to its simplicity. OHL focuses on detecting instances where an individual has opposing homozygous alleles at the same locus between samples (e.g., parent-offspring pairs). OHL efficiently flags errors, such as sample contamination or mislabelling, without extensive calculations and requires less computational power, making it suitable for large-scale datasets and routine checks. While maximum-likelihood methods provide more detailed analysis, OHL is favoured for initial screening due to its speed and simplicity. It allows for rapid issue identification before conducting more complex analyses if needed.

The pipeline, which uses OHL as its core methodology, has been in continuous development since the completion of VNC561-2021. It involves running the following steps.

1. **Determining OHL rate cutoff**: Selects a suitable OHL rate threshold to identify false parent-offspring relationships, using a visual inspection of the OHL fractions for all dyads.

2. **Duplicate checking**: Evaluates the concordance rates of intentionally duplicated samples and detects unintentional duplicates. In either case the sample most likely to contain an error is flagged. This is done by assessing the OHL rates between the genotypes the samples are supposed to represent and their respective parents and offspring.

3. **Dyad error detection and first-instance recovery**: Identifies cases of incorrect parentage through OHL analysis and suggests alternative candidate parents. Incorporates information on the planting date and location of each genotype’s ortet, if available, to improve accuracy.

4. **Second-instance recovery**: A second-round recovery pipeline to complement the initial OHL-based analysis, addressing cases where parentage ambiguity remains. This pipeline curates a specialized subset of data for use in programs like Sequoia, focusing on the extended family of the focal progeny. By crafting inputs to fit Sequoia’s human genetics-oriented format, the software can handle smaller, targeted datasets while resolving complex parentage issues. This second-round analysis is expected to be needed in a minority of cases, such as when two or more putative true parents are proposed or when one parent is unassayed

Single-Step Genetic Evaluation

Single-step BLUP (Best Linear Unbiased Prediction) is an advanced method used in genetic evaluations that combines pedigree, phenotypic, and genomic information into a single comprehensive analysis. This approach is especially valuable in both animal and plant breeding programs, as it

enables breeders to generate more accurate predictions of an individual's genetic merit (or breeding value) by integrating data from multiple sources.

In the context of Australian tree breeding programs, single-step BLUP has been implemented through the enhancement of TBA's TREEPLAN software. These enhancements were developed through previous FWPA projects, and detailed information can be found in the final reports of those projects.

In the current project, field pedigrees recorded in DATAPLAN will be adjusted to account for the findings from pedigree error checking and recovery work. A revised Genomic Relationship Matrix (GRM) will be constructed using the latest SNP genotype call data, which will then be incorporated into a new TREEPLAN analysis. This analysis will integrate both the revised pedigree information and genomic data, providing a more accurate genetic evaluation.

By comparing the results from this updated analysis with previous evaluations, we can quantify the improvements in breeding value prediction accuracy attributed to the corrected pedigrees and the incorporation of additional genomic data.

Results

Field Sampling and DNA Processing

Although the experiment on DNA yield and quality across different storage durations and sample amounts was initiated, it is not yet complete at the time of writing this report. Due to the timing constraints of the current project, the results will be carried over to a sister project, "Using genomics to double the rate of genetic gain in Australian forest tree improvement programs" (project number VNC580-2122). This continuation ensures that the experiment will be fully completed and reported without interruption, maintaining the integrity of the work and its contribution to both projects.

Array processing and SNP calling

The array plate QC summaries are displayed in Table 3. At this stage, the AxAS analyses were conducted separately for each project. Following advice from ThermoFisher, an average call rate of over 98% for passing samples was used to select the plates for defining the SNP-Specific Priors (SSP) models files. Plates used in preparing the models file for analyzing the Ramaciotti data are highlighted in yellow, while those used for preparing the models file for the Santa Clara data are highlighted in green.

The probeset metrics summaries are displayed in Table 2. The Ramaciotti projects, referred to as Rama#2 and Rama#3, produced noticeably lower numbers of Best and Recommended probesets compared to others. Most projects generated over 23,000 probesets, with at least 11,000 of the PolyHighRes type. The goal is to maximize the number of PolyHighRes probesets. A combined run, which included data from all projects and used a models file based on the best plates across all projects, resulted in a consistent number of Best and Recommended probesets but yielded a slightly lower count of PolyHighRes probesets. Conducting separate analyses for each laboratory (Santa Clara and Ramaciotti) with models files tailored to each resulted in slightly improved outcomes.

Table 2 Probset metrics summaries for each project analysed as separate batches and for various combined batches

AxAS analysis	Best and Reco	Poly High Res	Other	No Minor Hom	Mono High Res	Call Rate Below Thresh	OTV
a55111	24,746	12,279	7,453	7,625	5,202	3,879	207
Rama#1	25,196	11,224	7,032	7,517	6,455	3,959	98
Rama#2	21,419	8,449	10,888	6,825	6,145	3,571	407
Rama#3	22,848	10,670	9,077	6,368	5,810	4,071	289
Rama#4	24,426	11,125	7,959	7,135	6,166	3,727	173
Rama#5	26,198	12,163	6,952	7,484	6,551	2,941	194
Rama#6	18,574	5,312	12,853	6,299	6,963	4,812	46
Combined using SSP based on “Combined Best”	23,700	8,705	9,596	7,369	7,626	2,895	94
a551114 using SSP based on “a551114 best”	27,490	11,134	7,855	8,630	7,726	824	116
Combined Rama using SSP based on “Rama best”	23,688	8,987	9,496	6,887	7,814	3,033	68

Table 4 shows the number of probesets shared between the analyses of the data derived from the Santa Clara and Ramaciotti laboratories, categorized by probeset type. There were 21,267 probesets common to both analyses, with most probeset categories showing agreement between the two.

A probeset list was imported into both the a551114 and Combined_Ramaciotti batches and used to export genotype call data into two separate VCF files. These files were then merged into a single VCF, containing genotype call data for 9,240 samples across 21,267 variants.

Table 3 Array plate QC Summaries (AxAS analyses done separately for each project at this point)

Plate Barcode	Result	Number of files in a batch	Number of files failing dish QC	Number of files failing QC Call rate	Number of samples that passed	Percent of passing samples	Average call rate for passing samples	Filtered Call Rate	Project	Alias
551114439136604182133 1	PASSED	17	0	0	17	100	99.138	99.728	a55111	SantaClara
551114439665807202191 5	PASSED	287	0	0	287	100	98.95	99.704	a55111	SantaClara
551114446913208102404 6	PASSED	384	0	1	383	99.74	98.874	99.723	CUN11425	Rama#1
551114439709707242180 2	PASSED	49	0	0	49	100	98.56	99.577	a55111	SantaClara
551114439665807202191 1	PASSED	289	0	0	289	100	98.53	99.532	a55111	SantaClara
551114439136604182133 3	PASSED	15	0	0	15	100	98.496	99.37	a55111	SantaClara
551114439064104102196 5	PASSED	159	0	0	159	100	98.335	99.351	a55111	SantaClara
551114440066708152100 6	PASSED	41	0	0	41	100	98.134	99.343	a55111	SantaClara
551114450350504152660 1	PASSED	384	0	0	384	100	98.065	99.432	CUN13931	Rama#5
551114446913208102404 5	PASSED	384	0	1	383	99.74	98.009	99.394	CUN11425	Rama#1
551114446913208102404 4	PASSED	384	0	1	383	99.74	97.954	99.365	CUN11425	Rama#1
551114439064104102196 7	PASSED	1	0	0	1	100	97.892	99.168	a55111	SantaClara
551114450350504152660 3	PASSED	384	0	0	384	100	97.717	99.51	CUN13900	Rama#4
551114441668404182245 8	PASSED	56	0	0	56	100	97.183	98.577	a55111	SantaClara
551114440066708152100 5	PASSED	45	0	0	45	100	96.937	98.56	a55111	SantaClara
551114448216105282596 5	PASSED	384	0	1	383	99.74	96.768	99.564	CUN13262	Rama#2
551114450350504152660 8	PASSED	384	1	0	383	99.74	98.662	99.686	CUN13931	Rama#5
551114450350504152661 0	PASSED	384	1	0	383	99.74	98.367	99.652	CUN13900	Rama#4
551114450402604152604 9	PASSED	384	1	0	383	99.74	98.109	99.51	CUN13931	Rama#5
551114450350504152660 5	PASSED	384	1	1	382	99.479	97.678	99.581	CUN13548	Rama#3
551114446913208102404 3	PASSED	384	1	1	382	99.479	97.46	99.252	CUN11425	Rama#1
551114450783404152664 0	PASSED	384	1	0	383	99.74	97.264	99.125	CUN13900	Rama#4
551114448446907152598 5	PASSED	384	1	0	383	99.74	96.688	99.476	CUN13262	Rama#2
551114450350504152660 7	PASSED	384	3	0	381	99.219	98.031	99.607	CUN13900	Rama#4
551114450350504152661 1	PASSED	384	4	7	373	97.135	97.669	99.547	CUN13548	Rama#3
551114448446907152598 4	PASSED	384	4	3	377	98.177	97.31	99.626	CUN13262	Rama#2
551114448446907152598 3	PASSED	384	6	0	378	98.438	95.852	99.291	CUN13262	Rama#2
551114450350504152660 4	PASSED	384	14	0	370	96.354	97.651	99.388	CUN13900	Rama#4
551114450350504152660 6	PASSED	384	19	0	365	95.052	96.596	99.349	CUN13548	Rama#3
551114450350504152660 2	PASSED	384	58	2	324	84.375	96.475	99.179	CUN13548	Rama#3
551114449185408132547 5	PASSED	384	109	0	275	71.615	96.444	99.51	CUN14278	Rama#6
551114446913208102404 7	FAILED	384	133	32	219	57.031	94.679		CUN11425	Rama#1
551114449185408132547 8	FAILED	384	253	2	129	33.594	92.941		CUN13262	Rama#2

Table 4 Numbers in common by probeset categories between the Santa Clara and Ramaciotti centre laboratories

		Ramaciotti Centre			Totals
		PolyHighRes	MonoHighRes	NoMinorHom	
Santa Clara	PolyHighRes	6546	205	1212	7963
	MonoHighRes	69	5330	1019	6418
	NoMinorHom	1184	1518	4184	6886
Totals		7799	7053	6415	21267

Post-calling quality assurance checks

Check 1 MAF and Missingness. PLINK was used to remove 10,792 SNP due to MAF less than 0.01 (leaving 10,475). No samples were removed as all had a call rate greater than 90%.

Check 2 HWE. Figure 1 displays the FIN plot, highlighting SNPs that exhibit either maximal homozygosity (upper boundary) or maximal heterozygosity (lower boundary) for a given MAF. In this analysis, 570 SNPs were identified along the upper boundary with a disequilibrium coefficient greater than 0.05, and these were removed, leaving a total of 9,905 SNPs.

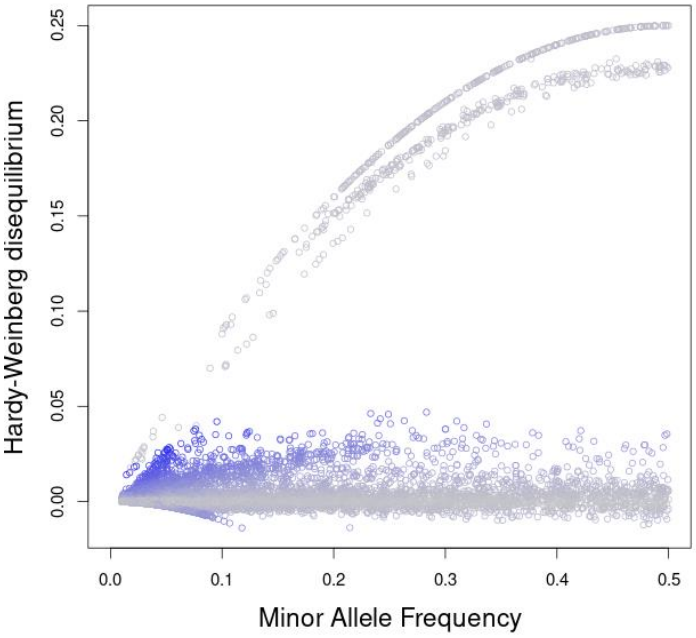

Figure 1 FIN plot showing the HW disequilibrium coefficients for the 10,475 SNP remaining after removal of SNP in Check 1.

Check 3 High OHL fraction across dyads. Figure 2 displays the results of computing the fraction of dyads that are OHL SNP by SNP. The x-axis represents the distribution of these fractions, ranging from 0 to 0.06, while the y-axis indicates the frequency of SNPs with each specific fraction. The data includes 13,747 dyads, with 6% equating to 825. While it's possible for this many errors to occur in the pedigree, it's unlikely that all 825 progeny would consistently display the opposing homozygote to the parent. In many cases, the progeny would be heterozygous. Before applying a more precise

mathematical method to determine the appropriate threshold for excluding SNPs, a preliminary estimate of 0.01 has been used. Based on this threshold 1,034 SNP were removed leaving 8,871.

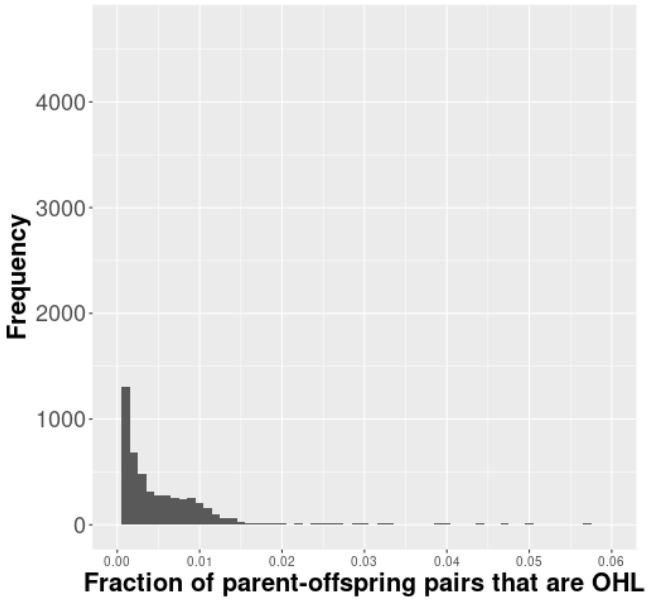

Figure 2

Figure 3 shows the distributions of MAF for SNPs categorized into bins based on the fractions of dyads that are OHL. Each bin represents a specific range of OHL fractions, illustrating how the MAF varies across different levels of OHL occurrence. While there is a clear trend where SNPs with low MAF tend to have zero OHL occurrence, it is reassuring to see an almost complete range of MAFs for SNPs with only small amounts of OHL occurrence.

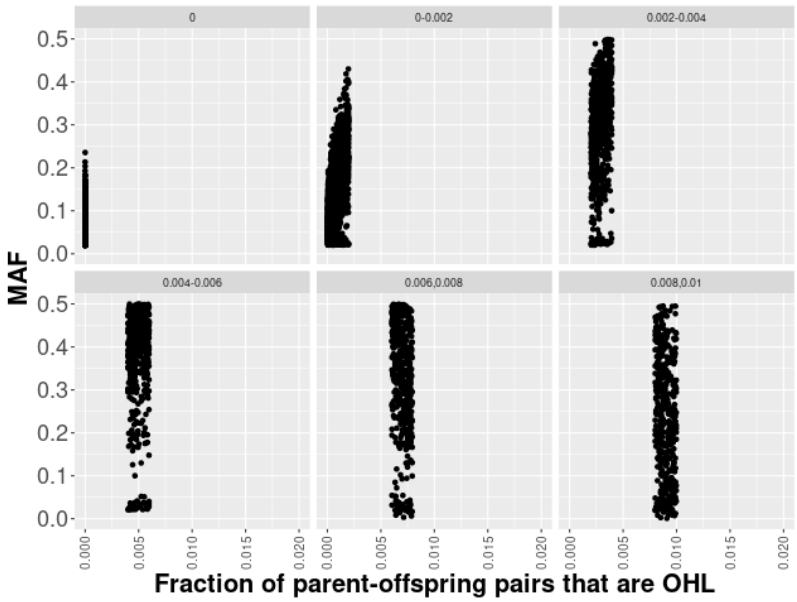

Figure 3 Distributions of MAF for SNPs categorized into bins based on the fractions of dyads that are OHL.

The pedigree and identity assurance pipeline

Determining the OHL rate cutoff

In this step we are computing the fraction of SNP that are OHL for every dyad. Figure 4 displays the results of computing the fraction of SNP that are OHL dyad by dyad. The x-axis represents the distribution of these fractions, ranging from 0 to 0.06, while the y-axis indicates the frequency of dyads with each specific fraction. There is a high frequency of dyads with very low OHL fractions, where any OHL occurrences are likely due to assay failures. In contrast, dyads with an OHL fraction greater than 0.005 are rare, which suggests potential errors in parent assignments

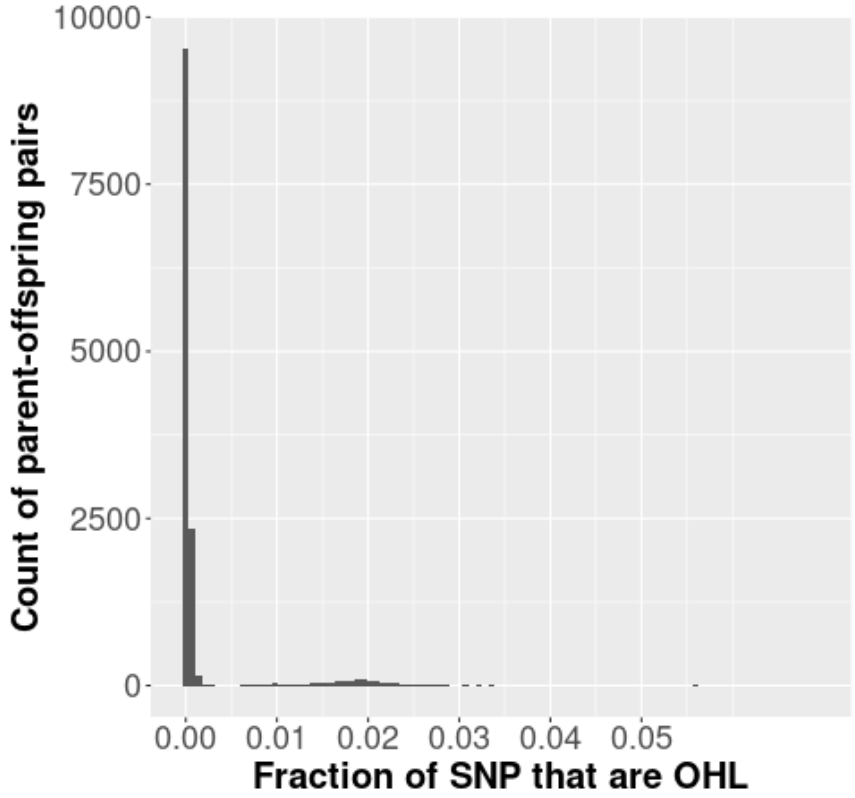

Figure 4 The results of computing the fraction of SNP that are OHL dyad by dyad.

Figure 5 is a close-up of Figure 4 obtained by limiting the x-axis values to between 0.002 and 0.08. It was decided that the appropriate cutoff to use was 0.005.

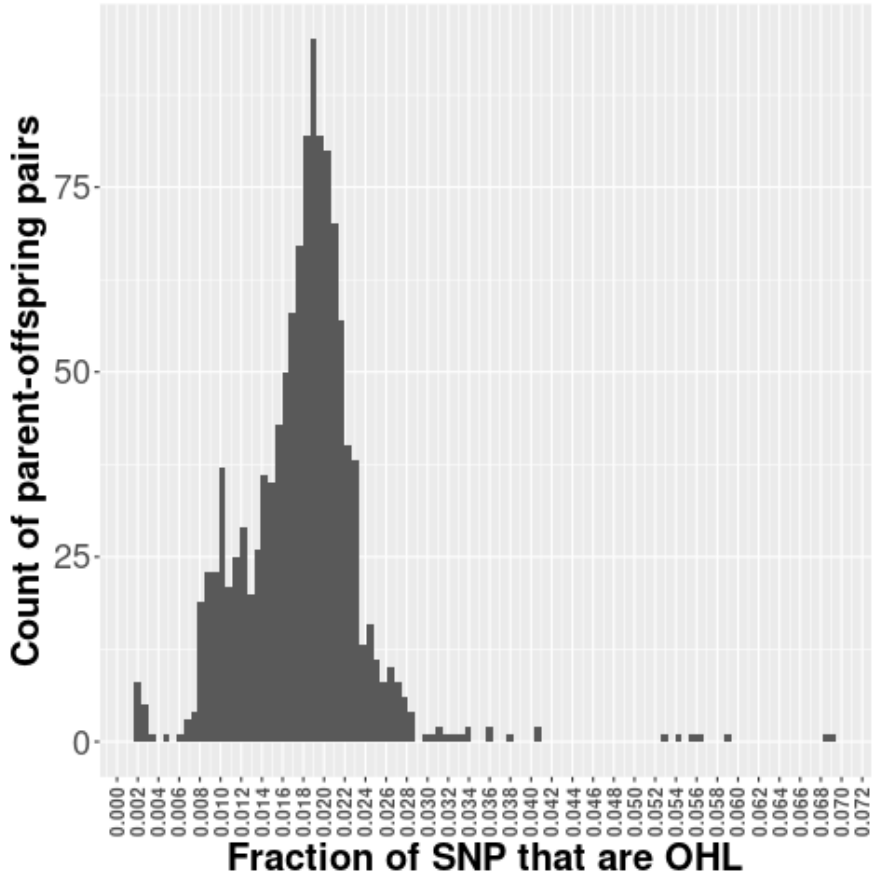

Figure 5 Close up of Figure 4 obtained by removing the SNP with zero occurrence of OHL, which results in a y-axis with a much lower maximum value, and by limiting the maximum x-axis value to 0.08.

.

Duplicate checking

The duplicate-checking phase of this project was largely driven by the specific needs of HVP Plantations (HVP), particularly for its nursery in Gelliondale, Victoria. Through planning discussions with Dr Jo Sasse, a strategy was developed to verify the identities of genotypes across HVP's orchards and archives. This strategy had two key elements: first, to verify genotypes planted in different eras based on the year of establishment, and second, to assess the integrity of the material. The material was classified by its pollination method—whether cross-pollinated (current or emerging), open-pollinated, or identified by the orchard itself.

Table 5 below summarizes the number of genotypes planted across various eras.

Table 5 Numbers of genotypes in HVP orchards by eras

Orchard/Archive	<2000	2000-2004	2005-2012	2013-2020	>2020
Archive orchard 2	11		8		
Archive orchard 3	1	26	93	9	9
Archive orchard 4			10		
OP orchard 1		4	8	1	

OP orchard 2			524	191	
OP orchard 3			119	191	
OP orchard 6				59	214
Seed orchard 1	49	6	6	4	
Seed orchard 5			46		
Seed orchard 6		24			
Christensens	21	38	74	8	
Scrubby Lane	25				

The *across-era* sub-study involved 222 ramets from 16 genotypes, with 5-7 ramets tested per era. The *cross-pollinated* sub-study included 19 genotypes from current cross-pollinated (CP) female parents, with 4 ramets per genotype, and 29 genotypes from emerging CP female parents, with 1-4 ramets per genotype.

For the *audits* of four of the three current open-pollinated (OP) orchards, 10% of ramets per orchard were tested, amounting to 222 genotypes and a total of 952 ramets, with 1-28 ramets tested per genotype. Where individual ramets appeared in multiple sub-studies, they were cross-referenced accordingly. Each orchard was annotated with the range of years in which its ramets were established.

At the Christensens and Scrubby Lane orchards, both situated outside the Gelliondale complex, 140 and 23 genotypes were tested, respectively. Of the 140 genotypes at Christensens, 67 had ramets located in the same or other orchards and were included in duplication checks. At Scrubby Lane, 5 of the 23 genotypes had ramets within the same orchard, allowing for similar verification.

In total, foliage samples were collected from 1,526 ramets representing 445 genotypes. Initially, it was decided to limit the DNA testing to no more than two ramets per genotype within the same orchard. This approach ensured a broader representation across genotypes while avoiding excessive duplication within individual orchards. The testing of the remaining foliage samples will depend on the results from the first round of analysis, allowing for a more targeted approach in subsequent testing phases.

Before discussing each sub-study, it is important to note that there were 689 instances where two samples were intentionally matched to the same genotype. In 41 cases, the concordance rate was below 0.97, indicating a mismatch between the samples. This suggests a general error rate of 6% in clonal replication.

Sub-study 1 – across eras

Fourteen genotypes were tested in this sub-study. The following is an example of the results for the genotype ‘A31057’. The year of the planting of the sample has been placed in parentheses.

Table 6 Genotype: A31057 Summary

Sample1 id	Sample1 location	Sample2 id	Sample2 location	Concordance rate
TBA-7634 (2012)	Seed orchard1	TBA-7607 (1991)	Seed orchard1	0.998748
TBA-7634 (2012)	Seed orchard1	TBA-7608 (1991)	Seed orchard1	0.998862
TBA-7634 (2012)	Seed orchard1	TBA-7565 (2004)	Seed orchard1	0.998633

TBA-7634 (2012)	Seed orchard1	TBA-7566 (2004)	Seed orchard1	0.99407
TBA-7634 (2012)	Seed orchard1	TBA-7633 (2012)	Seed orchard1	0.998633
TBA-7634 (2012)	Seed orchard1	TBA-7838 (2012)	Seed orchard5	0.998634
TBA-7634 (2012)	Seed orchard1	TBA-7840 (2012)	Seed orchard5	0.99725
TBA-7634 (2012)	Seed orchard1	TBA-6127	Youralla Rd	0.99383

- The sample TBA-7634 (2012) has high consistency
    - With other samples from the same orchard, such as TBA-7607 (1991), TBA-7608 (1991), TBA-7565 (2004), TBA-7566 (2004), and TBA-7633 (2012).
    - And with other samples from different orchards, such as TBA-7838 (2012) and TBA-7840 (2012) from Seed Orchard 5, and TBA-6127 from Youralla Rd
- All samples show a concordance rate above 0.99
- The samples from Seed Orchard 1, especially those from different planting years (1991, 2004, 2012), are highly consistent.
- The concordance rates between Seed Orchard 1 and Seed Orchard 5 also show strong consistency, indicating that the samples from both orchards represent the same genotype well.

Out of the 14 genotypes tested, 11 had straightforward 'good news' stories like the above. However, three cases presented some challenges. Let's take a closer look at some.

Genotype A31086 Analysis

Table 7 Genotype: A31086 Summary

Sample1 id	Sample1 location	Sample2 id	Sample2 location	Concordance rate
TBA-7579 (1997)	Seed Orchard 1	TBA-7670 (2004)	Seed Orchard 6	0.984134
TBA-7579 (1997)	Seed Orchard 1	TBA-7580 (1997)	Seed Orchard 1	0.984082
TBA-7579 (1997)	Seed Orchard 1	TBA-7671 (2004)	Seed Orchard 6	0.984219
TBA-7579 (1997)	Seed Orchard 1	TBA-7754 (2012)	Seed Orchard 5	0.984839
TBA-7579 (1997)	Seed Orchard 1	TBA-7581 (2006)	Seed Orchard 1	0.774109
TBA-7579 (1997)	Seed Orchard 1	TBA-7755 (2012)	Seed Orchard 5	0.771084
TBA-7581 (2006)	Seed Orchard 1	TBA-7755 (2012)	Seed Orchard 5	0.994105
TBA-7579 (1997)	Seed Orchard 1	TBA-6132	Youralla Rd	0.984439

Consistency Overview

- High Consistency:
    - TBA-7579 (1997) is highly consistent with:
        - Another sample from the same orchard: TBA-7580 (1997).
        - Samples from other orchards: TBA-7670 (2004) and TBA-7671 (2004) from Seed Orchard 6, and TBA-7754 (2012) from Seed Orchard 5.
    - OHL Rates: The DNA from TBA-7579 and its consistent samples shows no Opposing Homozygous Loci (OHL), meaning they are consistent with the DNA extracted from the assumed parents of Genotype A31086, confirming that the pedigree is correct.
- Low Consistency:
    - TBA-7579 (1997) has low concordance with:
        - TBA-7581 (2006) from Seed Orchard 1.
        - TBA-7755 (2012) from Seed Orchard 5.

Problematic Samples:

- TBA-7581 (2006) and TBA-7755 (2012) are consistent with each other (concordance rate 0.994105), but both are inconsistent with the main sample, TBA-7579.
- OHL Discrepancy: These two samples likely have misassigned genotype IDs, as their DNA contains Opposing Homozygous Loci (OHL), where their genotypes cannot logically match the parents of Genotype A31086.
- Further Investigation Needed: The next step is to identify the correct parents for TBA-7581 and TBA-7755, as they likely do not belong to Genotype A31086.

Genotype 00R3012 Analysis

Table 8 Genotype: 00R3012 Summary

Sample1 id	Sample1 location	Sample2 id	Sample2 location	Concordance_rate
TBA-7557	Seed Orchard 1	TBA-7556	Seed Orchard 1	0.996453
TBA-7557	Seed Orchard 1	TBA-7897	Archive Orchard 3	0.997022
TBA-7557	Seed Orchard 1	TBA-7898	Archive Orchard 3	0.996911
TBA-662	NGRC	TBA-7557	Seed Orchard 1	0.783434
TBA-662	NGRC	TBA-7556	Seed Orchard 1	0.779348
TBA-662	NGRC	TBA-7897	Archive Orchard 3	0.780093
TBA-662	NGRC	TBA-7898	Archive Orchard 3	0.780518

The genotype 00R3012 shows consistency across orchards at Gelliondale, Victoria (Seed Orchard 1 and Archive Orchard 3), with high concordance rates between these locations. However, significant discrepancies arise when comparing these samples to those at the National Genetic Resource Center (NGRC) in Mount Gambier:

- Gelliondale (Seed Orchard 1 & Archive Orchard 3): High concordance rates between samples indicate consistency at this location.
- NGRC (Mount Gambier): Samples from NGRC show low concordance when compared to Gelliondale samples, indicating these are not the same genotype.

Additionally:

- The DNA from all samples, regardless of location, exhibits Opposing Homozygous Loci (OHL), which means they are inconsistent with the DNA extracted from the assumed parents of 00R3012.
- This suggests that the samples at both Mount Gambier and Gelliondale do not represent the true genotype 00R3012.

Conclusion:

The samples at Gelliondale and Mount Gambier represent different genotypes, and both are inconsistent with 00R3012 due to discrepancies with the DNA from its assumed parents. Further investigation is needed to determine the identity of these samples.

Sub-study conclusion: The findings of this sub-study indicate that most genotypes remain consistent across different orchards and planting years. There is minimal evidence of problems linked to specific time periods or orchards. However, a few isolated cases deviate from the trend and merit further investigation.

Sub-study 2 – Current CP

In this sub-study, 19 genotypes commonly used as female parents over the past 4 to 5 years were targeted. Each genotype was sampled two to four times, depending on the number of locations where the ramets were planted. Concordance rates exceeded 0.98 across the board, confirming the consistency of all 19 genotypes and verifying their accuracy.

Sub-study 3 – Emerging CP

In this sub-study, 25 genotypes, highly ranked at the national level and selected for integration into HVP's CP program, were analyzed. Concordance rates were above 0.98 in all but one case—genotype A31086, which was discussed in the emerging eras sub-study (see above). In that instance, this genotype was used in both sub-studies.

Sub-study 4 – Individual Audits

Table 9 presents the results of consistency checks for individual orchards. No instances were found where a genotype was consistently replicated in error. For example, even when a genotype was replicated multiple times within an orchard, there were no cases where all three replicates disagreed.

Table 9 Results of the analyses for individual orchards

Orchard/Archive	Number of genotypes tested for consistency	Instances of Inconsistency Between Two Samples	Instances of Inconsistency Between HVP and Non-HVP Samples (e.g., NGRC)
OP1	6	1	0
OP2	100	4	4
OP3	8	0	0
OP6	104	3	3
Christensens	67	2	0
Scrubby Lane	5	0	0

The table demonstrates that, while minor inconsistencies were observed, particularly in OP1, OP2, and OP6, no major or widespread errors were detected across the orchards. Based on this data, HVP's overall error rate for clonally replicating genotypes for deployment is around 3 to 4%, which is within an acceptable range for this type of operation. Detailed summaries of individual genotype analyses, specifically for genotypes where samples were not consistent, are provided in Appendix A.

Unintentional duplicates

There were 144 instances where two samples representing different genotypes had concordance rates above 0.97, indicating they are the same genotype. In many cases, multiple ramets of one genotype matched multiple ramets of another. Based on this, 68 genotypes were found to have identity issues, meaning they were not what they were originally thought to be.

To put this in context, 446 genotypes were included in this pipeline, with a total of 1,526 samples processed. Out of these, 68 genotypes were misidentified, resulting in a mistaken identity rate of 15%. Notably, nearly all of these 68 genotypes (all but one) had progeny or parents assayed, which allowed us to identify the incorrect samples.

Table 10 provides two examples of how parent/offspring data was used to resolve these discrepancies.

Table 10 Examples of Unintentional Duplicates and the Use of Parent/Offspring Data to Determine the Incorrect Sample

Genotype 1 name	Sample 1 id	Genotype 2 name	Sample 2 id	Concordance rate	G1 evidence	G2 evidence

96R3508	TBA-8835	96R3608	TBA-8636	0.998866	P1:41709:0.018 P2:42577:0.018 C1:8896012:0.020	P1:36020:0.000 C1:10185204:0.0000
A30014	TBA-186	A33014	TBA-8278	0.988313	C1:9035913:0.0003	P1:41707:0.020 P2:277833:0.0270

In these examples, G1 evidence shows the Opposing Homozygous Loci (OHL) fractions between Genotype 1 and its assumed parents (P1 and P2) and children (C1, C2 etc), while G2 evidence shows the OHL fractions for Genotype 2 in relation to its assumed parents and offspring.

- In the first case, the higher OHL fractions for Genotype 1 (greater than 0.005) indicate that it is not 96R3508 but rather 96R3608. The similarity in the genotype names suggests this was a simple recording error.
- Similarly, in the second case, the sample labeled as A33014 represents A30014, again due to a recording mistake.

These cases demonstrate how parent/offspring data and Opposing Homozygous Loci (OHL) fractions were instrumental in correcting mistaken genotype identities caused by clerical errors. Unfortunately, such clerical errors appear to be relatively common. Fortunately, 56 of the misidentified genotypes were successfully resolved. However, there were 6 cases where both samples seem to have been assigned to the wrong genotype, and further work will be needed to uncover their true identity.

On a brighter note, four genotypes included in this study had previously lost their identity. Fortunately, we were able to successfully match them to known genotypes, resolving their identity.

The next step in the pipeline can proceed only after all intentional and unintentional duplicates have been removed. A removal option is provided within the software used for duplicate checking. Before the removal process, 9,240 samples collapsed to 8,557 unique genotypes. After the removal of duplicates, 8,476 samples were reduced to 8,476 genotypes, ensuring a one-to-one correspondence between samples and genotypes.

Dyad error detection and first-instance recovery

In this step, all 13,405 dyads, where both parent and offspring have been assayed, are checked for consistency by ensuring they do not exceed a threshold for Opposing Homozygous Loci (OHL) occurrences. After the software completed its run, 641 progeny were flagged, with either one or both parents identified as potentially incorrect leading to a total of 918 dyads flagged as incorrect. Figure 6 and Figure 7 provide some basic information about the types and amounts of errors

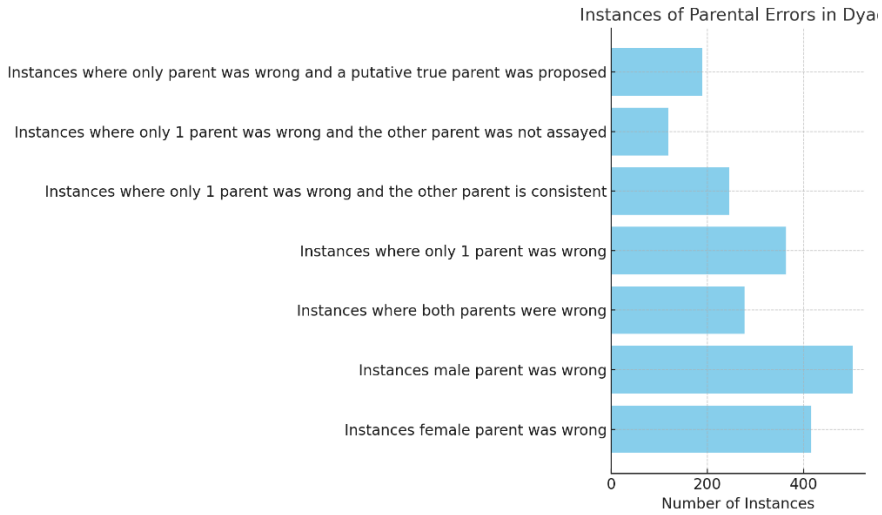

Figure 6 Instances of Parental Errors in Dyads

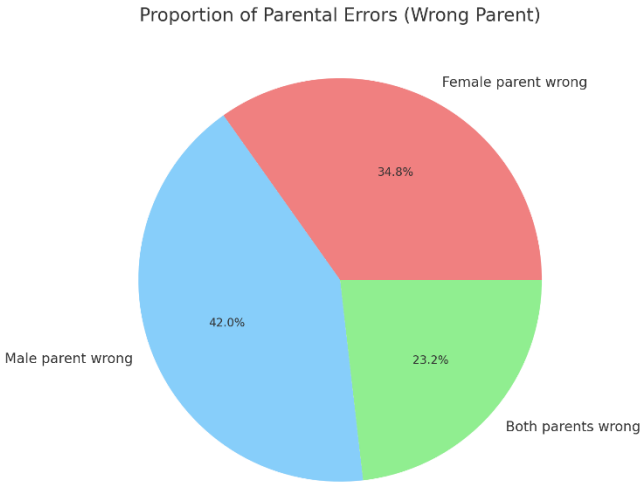

Figure 7 Proportions of Parental Errors (Wrong Parent)

The conclusions we can draw from this result are as follows:

1. **Higher Frequency of Incorrect Male Parent Assignments**: The analysis revealed that there are more instances where the male parent (502 cases) was incorrectly assigned compared to the female parent (416 cases). This suggests that there is a greater likelihood of error when assigning the correct pollen (male parent) than mistaking the seed parent (female parent). The difference in error rates might be attributed to the complexities involved in tracking pollen sources in breeding programs, where multiple potential male contributors could be present, increasing the possibility of misidentification.

2. **Improved Accuracy in Identifying True Parents**: Significant improvement in the identification of true parents has been achieved through the inclusion of a more extensive pool of assayed historical parents. In cases where only one parent was misidentified, a putative true parent was proposed in 54% of cases (198 out of 364 cases). In cases where two parents were misidentified, two putative true parents were proposed in 86% of cases (239 out of 277 cases). This advancement greatly increases the likelihood of accurately recovering the true maternal or paternal lineage. This marks a substantial increase from previous analyses, where the chance of finding the correct parent was only 5%. This improvement can likely be attributed to the concerted effort to expand the pool of assayed parents, allowing for a more comprehensive comparison and validation.
3. **Multiple Putative True Parents in Cases Where One Parent Was Not Assayed**: An interesting finding emerged from the 119 cases where one parent was wrong, and the other parent was not assayed. In 39 of these cases, two putative true parents were proposed. In total, two or more putative true parents were suggested in 58 instances. Notably, in most cases where two parents were proposed, one parent had not been assayed. The software ensures that the proposed putative true parent is not the parent listed as "not assayed." This finding may suggest that even though the non-assayed parent cannot be directly evaluated, the presence of two putative true parents indicates that there is a significant likelihood the non-assayed parent is also incorrect. In such cases, the software's ability to suggest two potential alternatives could be a sign that the family needs a reassessment, and both parents may have been incorrectly assigned, even when one remains unassayed.
4. **A Low Overall Error Rate**: Despite these findings, the overall dyad error rate in this analysis remains remarkably low at 6.8% (918 dyad errors out of 13,405 dyads). This is an encouraging result, especially when compared to error rates reported by tree breeding operations in other countries, where rates as high as 20% have been communicated informally. Our results underscore the success of the current methods and the diligence applied to parental verification in this breeding program.

Second-Round Recovery Pipeline

Building upon the first-round recovery package that employs OHL (Opposing Homozygous Loci), a second-round recovery pipeline is being developed to further resolve cases where initial parentage verification leaves some ambiguity. In a minority of cases — such as when two or more putative true parents are proposed, and the other parent is assayed and confirmed as consistent — this second round becomes necessary. Another scenario warranting second-round analysis arises when two parents are proposed, but only one has been assayed, leaving uncertainty around the unassayed parent.

To address these situations, the principal researcher proposes developing a streamlined approach that curates a specialized set of inputs for programs like Sequoia, which utilize maximum likelihood (ML) methods. The crafting process focuses on isolating the extended family of the focal progeny, reducing the scale of the problem while retaining the most relevant genetic relationships. This allows the ML algorithms to process a smaller, more manageable set of data while preserving the integrity of the analysis.

Given Sequoia’s origins in human genetics, the pedigree must be formatted in a way that aligns with its specific expectations around parameter files, life histories, and family structures. By carefully designing the inputs in this manner, Sequoia can be adapted to the tree breeding context, despite its original design being geared toward human genetic data.

This second-round recovery process is expected to occur in a relatively small number of cases, but it provides a crucial follow-up to the first-round OHL analysis. As the pipeline continues to be refined, it will offer an additional layer of verification, ensuring that even complex parentage cases can be resolved with greater confidence.

Final Thoughts:

The low error rate we have ascertained is a testament to the advancements made in expanding our genotyped parent pool, improving the accuracy of parentage assignments, and identifying true parents. As tree breeding programs continue to grow in complexity, maintaining this high level of accuracy becomes even more critical. By leveraging modern tools and focusing on thorough parentage verification, we are positioning ourselves as leaders in the field, demonstrating that high-quality, pedigreed material can be effectively tracked even in large-scale breeding operations.

With a 4.8% error rate in the face of potentially much higher error rates reported elsewhere, our tree breeding program is setting a high standard for parentage accuracy, ensuring the best possible foundation for future genetic improvement and commercial success.

Single-Step Genetic Evaluation

This section focuses on the ability of the genomic data to predict the Mendelian sampling term, which represents the genetic value unique to an individual and distinguishes it from its siblings. Early and accurate prediction of this term allows juvenile trees to be cloned and established in breeding arboreta, ensuring their availability as reproductive parents years before traditional phenotypic evaluation is completed.

Background

An individual's estimated breeding value (EBV) is the sum of two components:

- **Mid-parental EBV**: The average genetic value of the male and female parents.
- **Mendelian Sampling Term**: The genetic contribution unique to the individual, making it distinct from its siblings. Traditionally the greek letter Phi ϕ has been used to represent this term when the equation for the EBV is expressed mathematically.

Genomic selection focuses on predicting ϕ early in an individual's life. By reducing the generation interval, genomic selection enhances the rate of genetic gain as defined by the breeder's equation. In practical terms, reducing the generation interval contributes more to genetic gain than increased prediction accuracy of genetic value.

Integrated Analyses in Breeding Programs

Our breeding programs employ integrated analyses to maximize data utility. These analyses include:

- **Single-Step Best Linear Unbiased Prediction (BLUP):** Combines phenotypic data, pedigree information, and genomic data across generations and trial sites in a multivariate framework. TREEPLAN is our current evaluation tool that implements this technology.
- **H-matrix Blending**: Integrates the genomic relationship matrix (**GRM**) with the pedigree relationship matrix (**NRM**) to improve the accuracy of genetic evaluations.
- **Multi-Trait Modelling**: Treats measurements of traits (e.g., growth) across different sites and ages as distinct but genetically correlated traits.

Methodology for Evaluating Genomic Selection

To evaluate the effectiveness of genomic selection, we gauge the ability to predict ϕ for DNA-assayed juveniles. The methodology involves:

1. Incremental Training Set Size

A critical component of this study was assessing the effect of training set size on the effectiveness of genomic selection. The training set consists of individuals with both phenotypic measurements and DNA assays. Based on theoretical work by Dr Hans Daetwyler, the number required for a training set size to enable effective genomic selection in forest tree breeding is approximately 20,000 individuals.

Currently, we have assayed 8476 trees using the radiata SNP chip. Of these, 5104 have phenotypic measurements and constitute the training set. The remaining trees (3372) consist mostly of juveniles, which currently lack measurements, and antecedents. Over time, as the juveniles are assessed, the training set size will increase, but this is not the case as of December 2024. To investigate the impact of training set size, we incrementally increased the training set size:

- 0 (juveniles and antecedents only)
- 1582
- 3319

- 5104 (current full training set)

This experimental design enabled us to examine trends in the effectiveness of genomic selection by analyzing the correlation between genomic ϕ (predicted using genomics) and the phenotypic ϕ (observed ϕ) across varying training set sizes. Additionally, these trends provide a projection of the expected performance when the training set size reaches 20,000.

Table 11 summarizes the composition of the training set by trial, illustrating its incremental expansion across stages.

A smaller training set size does not mean fewer observations in the BLUP analysis. All phenotypic measurements (from all assayed trees) are always included in the analysis. However, a smaller training set size results in a smaller genomic relationship matrix (**GRM**).

2. Knockout Single-Step BLUP Analysis

For each training set size, a cohort of trees was selected, and their phenotypes were removed from the analysis. The analysis then predicted their ϕ values (genomic Phi's). These genomic Phi's were compared to their phenotypic Phi's, which are computed from:

- Rescaled phenotypes (such that a standardize additive variance is established across trials).
- Phenotypes corrected for spatial effects within trials.

3. Phenotypic Rescaling

Phenotypic measurements from different trials are rescaled by the additive variance predicted within each trial. This ensures comparability by removing scale effects arising from differences in site productivity. For example, a highly productive site will naturally have higher mean growth, which must be accounted for during analysis.

4. Correlation Analysis

The correlation between genomic Phi's and phenotypic Phi's serves as a metric for the efficacy of genomic selection. Observing how this correlation changes with incremental increases in training set size provides valuable insights into the relationship between data availability and genomic prediction accuracy.

Figure 8 illustrates the composition of plantings by year in our rolling front breeding program, which gained momentum in the early 2000s. The figure highlights the gradual turnover of generations within each planting year, rather than discrete generational shifts. This underscores the importance of ensuring genomic selection is effective within the dynamic framework of a rolling front breeding strategy. The overlaid box on the left shows the total number of phenotypic measurements included in all BLUP analyses, while the overlaid boxes on the right display the sizes of the training sets. Together with the 3372 remaining individuals, these totals determine the size of the genomic relationship matrix (**GRM**). Each training set size varies in the quantity of data available for each trait, reflecting the evolving nature of the program.

The **GRMs** are agnostic in that the SNPs used to generate genotype calls and construct the **G** are not necessarily associated with specific traits. A good policy, therefore, is to ensure the training set includes individuals measured for all traits. This ensures the **GRM** — and consequently genomic selection — can be effective across all traits.

Table 11 The composition of the training set by trial. The size of the training set was increased in a series of three incremental steps: all trials up to and including year 2013; all trials up to and including year 2018; and all trials up to and including year 2021. Trials with fewer than 20 genotypes represented in the training set are not shown.

Trial title	Year	SC Trait	Number of genotypes
BR9601 Airport Progeny Trial	1996	GROWTH_GTR	117
BR9606 Mumbannar Progeny Trial	1996	GROWTH_GTR	23
BR9615 Koomeela Progeny Trial	1996	GROWTH_TAS	60
BR9705 Kromelite Progeny Trial	1997	GROWTH_GTR	60
BR9611 Flynn Progeny Trial	1996	GROWTH_CGIPP	99
BR9701 Bussells Progeny Trial = RS43	1997	GROWTH_WA	23
BR9713 McFarlanes Block Stockdale	1997	GROWTH_CGIPP	35
BR0801 Connorville Progeny Trial	2008	GROWTH_TAS	124
VRC028 Salicki	1980	GROWTH_CGIPP	29
BR0901 Hexham Progeny and Realised Gain Trial	2009	GROWTH_CVIC	27
BR0903 Shelley Progeny and Realised Gain Trial	2009	GROWTH_DOTHI	25
RES1295 Rennick 4G CP & 3G x Guadalupe Progeny Trial	2005	GROWTH_GTR	32
BRGT1301 Caroline Aus/NZ Collaborative Progeny Trial	2013	GROWTH_GTR	243
BRGT1303 Heywood's Aus/NZ Collaborative Progeny Trial	2013	GROWTH_CGIPP	214
BRGT1304 Moogara Progeny/Realised Gain Trial	2013	GROWTH_TAS	84
BRGT1404 Jarrahwood Progeny/Realised Gains Trial	2014	GROWTH_WA	160
BRGT1403 Bundaleer Aus/NZ Progeny/Realised Gain Trial	2014	GROWTH_MVAL	158
BRGT1501 Johnson's Lane	2015	GROWTH_GTR	296
BR1602 Westerway	2016	GROWTH_TAS	269
BRGT1603 Upper Blessington	2016	GROWTH_TAS	20
BRGT1701 Dicksons Caroline Progeny and Embedded Gain	2017	GROWTH_GTR	174
BRGT1702 Dorodong Progeny and Embedded Gain	2017	GROWTH_GTR	91
BRGT1703 Strathbogie Progeny and Embedded Gain	2017	GROWTH_DOTHI	171
BR1804 Noolook (Saltmarsh) Progeny trial	2018	GROWTH_GTR	252
BRGT1803 Durham Progeny wth Embedded Gain	2018	GROWTH_CVIC	137
BRGT1902 Lowan Lane	2019	GROWTH_GTR	300
BRGT1903 Kentbruck	2019	GROWTH_GTR	227
BRGT2001 Mount Burr	2020	GROWTH_GTR	134
BRGT2002 Rennick	2020	GROWTH_GTR	264
BRGT2104 Powers Creek (Hinkley)	2021	GROWTH_GTR	464
BRGT2103 South Patchells	2021	GROWTH_GTR	258

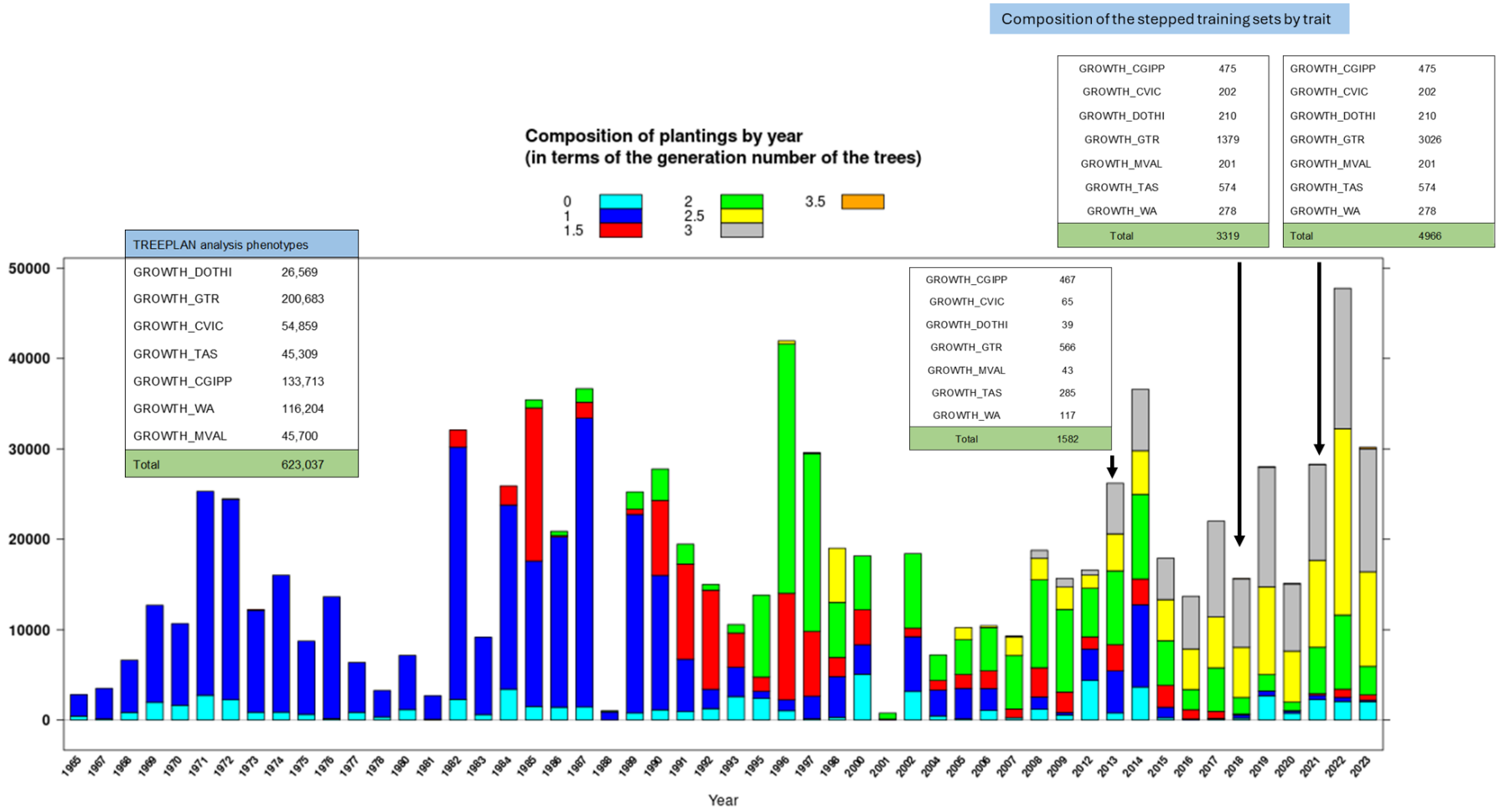

Figure 8 Composition of plantings by year in terms of the generation number of the progeny .

Results

Figure 9 illustrates the correlation between the additive Mendelian sampling term (ϕ_{add}) and the phenotypic Mendelian sampling term (ϕ_{phen}) as training set sizes increase to 20,000, across five traits (GROWTH_DOTHI, GROWTH_GTR, GROWTH_MVAL, GROWTH_TAS, etc.). It highlights the predictive performance of the genomic relationship matrix (GRM) versus the numerator relationship matrix (NRM) under different conditions.

Key Observations

1. Highlighting GROWTH_GTR (Top Right Panel):

- The top-right panel, representing GROWTH_GTR, demonstrates the most notable lift in correlation as the training set size increases from 3319 to 4906. This is consistent with the fact that most of the additional observations in this step were for GROWTH_GTR, bringing the total number of individuals measured for this trait to 3206—more than six times greater than for any other trait.
- This concentration of phenotypes illustrates the importance of having sufficiently large trait-specific training sets. Encouragingly, this suggests that once the phenotypic data for all traits exceed 3000–5000 individuals, we can expect acceptable correlations across the board.

2. NRM with Knockout Phenotypes Excluded (Red Line):

- The red line remains fixed at zero for all traits and training set sizes, reflecting that the NRM cannot predict ϕ_{add} without phenotypic information. This equates to random within-family selection, treating all siblings as having equal ϕ_{add}.

3. NRM with Knockout Phenotypes Included (Black Line):

- Once phenotypes are made available, the NRM achieves a stable correlation, which does not change with increasing training set size. This indicates that the NRM provides an upper bound on the predictive ability for ϕ_{add} reflecting the maximum possible correlation achievable through phenotypic data alone.

4. GRM with Knockout Phenotypes Excluded (Black Dots):

- The GRM predictions, represented by the black dots, improve with increasing training set size. For GROWTH_GTR in particular, we observe significant gains as the training set expands, reflecting the large number of phenotypes available for this trait.
- As training set sizes grow for other traits, the GRM's predictive performance for these traits is also expected to improve, approaching the upper bound established by the NRM with phenotypes included.

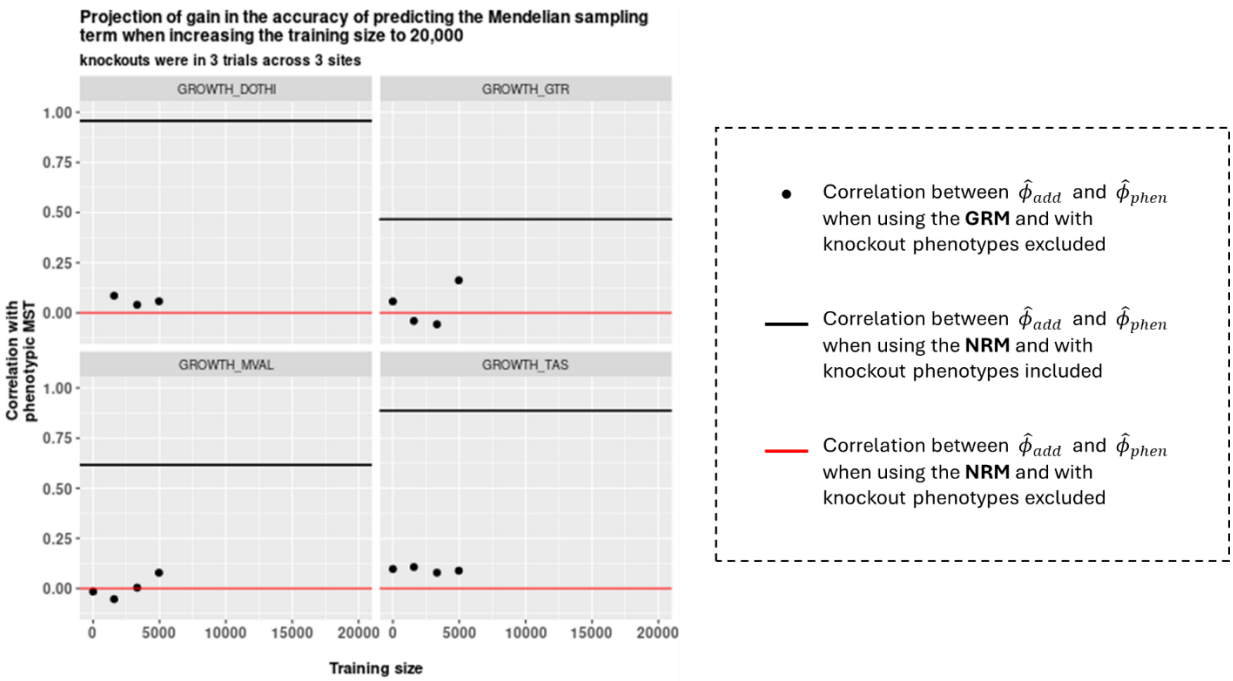

Figure 9 The results of the single-step analyses showing the changes in correlation between $\hat{\phi}_{add}$ and $\hat{\phi}_{phen}$ when training size is increased.

Future Considerations

The results presented here are based on a SNP chip with limited utility for genomic selection. Although the chip initially contained over 30,000 probesets, fewer than 9,000 SNPs passed quality control filters. This constraint impacts the effectiveness of genomic selection. However, ongoing efforts in a sister project aim to significantly improve genomic resources for *Pinus radiata*. These improvements include:

- A new high-density SNP array containing hundreds of thousands of high-quality, reliable SNPs.
- A low-density assay with tens of thousands of SNPs, enabling accurate imputation to high-density data.

The current assay was designed as a short-term solution. With the anticipated genomic resources, we expect substantial improvements in correlation values, further enhancing the effectiveness of genomic selection for *P. radiata* breeding programs.

Pedigree Recovery in Action: Enhancing Single-Step Analysis

Accurate pedigree information forms the backbone of genetic evaluations, directly influencing the precision of breeding value estimates. In a previous section, we discussed how the Scion DNA assay was used to detect dyad errors and propose putative true parents. These corrections necessitate updates to the field pedigree database (DATAPLAN), often involving reallocation of individuals to existing or newly created family IDs. Unlike human pedigrees, where individuals are typically recorded with their mother and father, forestry pedigrees map individuals to families characterized by parental identities and types.

Uncovering Pedigree Errors: Key Case Studies

The examples below highlight key challenges in pedigree recovery and their implications for genetic evaluation, particularly in the context of single-step analysis.

1. **Pedigee Errors Across Generations:** Table 12 highlights a case of systemic error across different epochs, where parent 10258 was misidentified as a true parent in multiple progeny. This suggests a broader issue, likely stemming from historic data inaccuracies or sampling errors. To validate these findings, DNA from a second ramet archived in Tasmania should be tested. If the results confirm misidentification, the parent should be reassigned to a genetic group. Addressing systemic errors like this improves the reliability of genetic evaluations by ensuring accurate family structure.

Table 12 Systematic Pedigree Errors (Parent 10258 Example)

Genotyp ID	Family ID	Ortet location	Year planted	FP	MP	FP status	MP status	Dyads Tested / Erroneous	NPTP
41687	3702	RAD114	1967	10086	10258	NA	ERR	10/10	0
42098	3867	RAD120	1968	10218	10258	OK	ERR	10/10	0
42119	3876	RAD124	1969	10223	10258	OK	ERR	10/10	0
42145	3888	RAD117	1968	36023	10258	OK	ERR	10/10	0
42709	4199	RAD114	1967	11102	10258	OK	ERR	10/10	0
302940	9329	RAD114	1967	11106	10258	OK	ERR	10/10	0
42207	3912	RAD120	1968	10258	11097	ERR	OK	10/10	0
205952	5898	BR9705	1997	10258	0	ERR	NA	10/10	0
2148481	5898	BR0801	2008	10258	0	ERR	NA	10/10	0
8704592	5898	BRGT1301	2013	10258	0	ERR	NA	10/10	0

Table 13 Complex Cases Involving Multiple True Parents (Parent 41776 Example)

Genotype ID	Family ID	Ortet location	Year Planted	FP	MP	FP status	MP status	Dyads Tested / Erroneous (FP)	Dyads Tested / Erroneous (MP)	NPTP	PTP1	PTP2
175880	5910	BR9606	1996	42658	41776	OK	ERR		200/22	1	36027	
1169859	5910	BR0502	2005	42658	41776	OK	ERR		200/22	0		
4124056	5910	BR0903	2009	42658	41776	OK	ERR		200/22	1	42012	
8702162	5910	BRGT1301	2013	42658	41776	ERR	ERR	59/5	200/22	1	36015	
8736750	5910	BRGT1303	2013	42658	41776	OK	ERR		200/22	2	102655	42207
9063080	117385	BR1804	2018	104000	41776	ERR	ERR	19/1	200/22	2	680354	41707
9060976	117397	BR1804	2018	169212	41776	ERR	ERR	53/2	200/22	1	2447903	
9063049	117525	BR1804	2018	335383	41776	ERR	ERR	13/3	200/22	1	912699	
9063829	117545	BR1804	2018	344731	41776	ERR	ERR	82/6	200/22	2	42199	100784
9061630	117565	BR1804	2018	346397	41776	ERR	ERR	120/4	200/22	2	347340	169109
9063821	117607	BR1804	2018	517123	41776	ERR	ERR	81/8	200/22	2	347340	99799
9060982	117614	BR1804	2018	517412	41776	ERR	ERR	8/3	200/22	1	679912	
9063925	117647	BR1804	2018	909836	41776	ERR	ERR	47/3	200/22	2	2447903	101224
9061688	117650	BR1804	2018	910059	41776	ERR	ERR	57/2	200/22	2	207213	516936
9060965	117710	BR1804	2018	2449089	41776	ERR	ERR	159/8	200/22	2	347340	517123
10213412	124358	Penola	2022	100785	41776	ERR	OK	28/4		2	36044	8702139
9804807	126951	BRGT2102	2021	2954971	41776	NA	ERR		200/22	1	277826	
10197488	126951	Niggl	2022	2954971	41776	NA	ERR		200/22	1	277826	
10196903	130475	Niggl	2022	4663324	41776	ERR	ERR	8/1	200/22	1	176047	
10963956	136770	Penola	2023	182509	41776	ERR	ERR	3/1	200/22	2	4664965	42576
10804200	136798	McGillivrays	2023	2412316	41776	OK	ERR		200/22	1	277826	
10803653	136854	McGillivrays	2023	4663319	41776	OK	ERR		200/22	1	277826	

2. Complex Errors Involving Multiple Putative Parents: Table 13 illustrates the complexities arising from genotype 41776, misassigned as a male parent in family 5910 and other CP families. Five full-sibs within family 5910 were misassigned, with multiple putative true parents (PTPs) identified, likely due to polymix pollination. These cases demonstrate the challenges of historic breeding practices and highlight the importance of nuanced interpretation, cross-referencing historical records, and genomic validation to identify the most likely parents. Resolving such errors ensures that breeding value estimates reflect the true genetic relationships among progeny.

3. Clear-Cut Error Cases: Table 14 highlights a clear-cut case with genotype 8701919, where one putative true parent (PTP1) was proposed to replace the incorrect parent. These straightforward errors are easier to address but still require cross-validation with auxiliary records (e.g., pollen and flower records) to ensure the change aligns with historical data. This reinforces the need for a centralized auxiliary database to support pedigree recovery.

Table 14 Clear-Cut Case (Genotype 8701919 Example)

Genotype ID	Family ID	Ortet location	Year planted	FP	MP	FP status	MP status	Dyads Tested / Erroneous	N PTP	PTP1
8701919	4014	BRGT1301	2013	36044	36015	OK	ERR	84/1	1	42098

4. Sibship Inconsistencies: Table 15 presents the case of alleged CP family 25055, where genomics suggests that two individuals assumed to be full-sibs share a common true parent (PTP1) while the third belongs to a different family. This highlights the challenges of verifying sibship in controlled crosses and the need to validate genomic findings against auxiliary records. Addressing such inconsistencies is essential for accurate family-based genetic evaluations

Table 15 Sibship verification (Family 25055 Example)

Genotype ID	Family ID	Ortet location	Year planted	FP	MP	FP status	MP status	Tested/ Erroneous (FP)	Tested/ Erroneous (MP)	PTP1	PTP2
2147595	25055	BR0801	2008	42658	100541	ERR	ERR	59/5	42/5	10218	101778
8097083	25055	BRGT1501	2015	42658	100541	OK	ERR		42/5	99912	
8739129	25055	BRGT1303	2013	42658	100541	OK	ERR		42/5	99912	

The following is a proposed, structured framework for validating putative true parents (PTP)

1. Cross-Validation with Historical Records

Why: To confirm that the proposed PTP is plausible based on historical data about the breeding program.

How:

- Check pollen and flower records for the relevant time-period to ensure the proposed PTP was available and used in the cross.
- Verify geographical proximity: Confirm whether the proposed parent and the seed parent were in the same location during the breeding event.

- Review breeding logs, crossing records, or any documented mating designs (if available).

Challenges:

- Historic records may be incomplete, inconsistent, or not digitized.
- This underscores the need for a centralized auxiliary database that captures such information for validation purposes.

2. Re-Test Using DNA Assays

Why: To confirm that the genetic match between the progeny and the proposed PTP holds up under further scrutiny.

How:

- Resample DNA from the proposed PTP (if available) and rerun the DNA assay to validate the genetic relationship.
- Test DNA from multiple ramets of the proposed PTP to rule out potential errors due to contamination or mislabelling.

Challenges:

- If a PTP is not available or its clones (ramets) are no longer accessible, this step may not be feasible.

3. Genomic Likelihood Analysis

Why: To assess the statistical likelihood that the proposed PTP is the true parent based on genomic data.

How:

- Use likelihood-based programs such as Sequoia to calculate the likelihood of the proposed PTP being a true parent.
- Compare the genomic similarity of the progeny to the PTP and other potential parents to ensure the proposed PTP is the most plausible candidate.

Challenges:

- Requires a robust and well-curated DNA dataset with all plausible parents genotyped.

4. Validation Through Cross-Progeny Analysis

Why: To look for consistency across siblings and other progeny associated with the PTP.

How:

- Validate whether the PTP matches as the parent for multiple progeny within the same family.
- If the PTP appears consistently across many progeny, it strengthens the case for its validity.

Challenges:

- Requires comprehensive genomic data for all progeny in the family.

5. Assigning Genetic Groups When Validation Fails

Why: When a PTP cannot be conclusively validated, it’s important to assign a plausible genetic group to the progeny.

How:

- Assign the progeny to a landrace group or an epoch-based genetic group based on the breeding history (e.g., "Australia Pre-1970 Selected").
- Document the uncertainty and the reason for assigning a group rather than an individual parent.

6. Centralized Auxiliary Database for Validation

Why: To streamline validation and avoid reliance solely on genomics.

What to Include:

- Pollen and flower usage records.
- Mating design plans and field crossing histories.
- Geographic and temporal availability of parents.
- Metadata about ramet locations and cloning events.

Implementation:

- Begin digitizing and centralizing historical records.
- Link the auxiliary database to the pedigree and DNA assay database for cross-referencing.

7. Protocol for Final Validation

Assign a confidence score to the PTP proposal based on the results of the above steps:

- High confidence: PTP matches genomic data, historical records, and biological plausibility.
- Moderate confidence: PTP matches genomic data but has incomplete historical support.
- Low confidence: PTP is plausible based on genomic data but lacks historical evidence and is biologically ambiguous.

If confidence is low, assign the progeny to a genetic group instead of a specific parent.

EBV Comparisons Pre- and Post-Pedigree Recovery

Following the identification of moderate to high-confidence putative true parents (PTPs), a total of **326 progeny** underwent parentage corrections. These updates resulted in **243 new PTP assignments**, replacing **211 previously recorded parents**. While 89 new families were created to accommodate unique parental combinations, most corrections (235 instances) mapped individuals to existing family IDs within the pedigree database. To assess the impact of these changes, estimated breeding values (EBVs) were compared for both progeny and parents across key traits.

1. Progeny EBV Comparisons:

As shown in Figure 10, the comparison of progeny EBVs pre- and post-pedigree recovery reveals more dramatic shifts, particularly for traits heavily influenced by parentage (e.g., growth traits). These shifts emphasize the strong dependence of progeny EBVs on the breeding values of their assigned parents. Parentage corrections can result in notable recalculations when progeny are reassigned to new parents with higher or lower genetic merit.

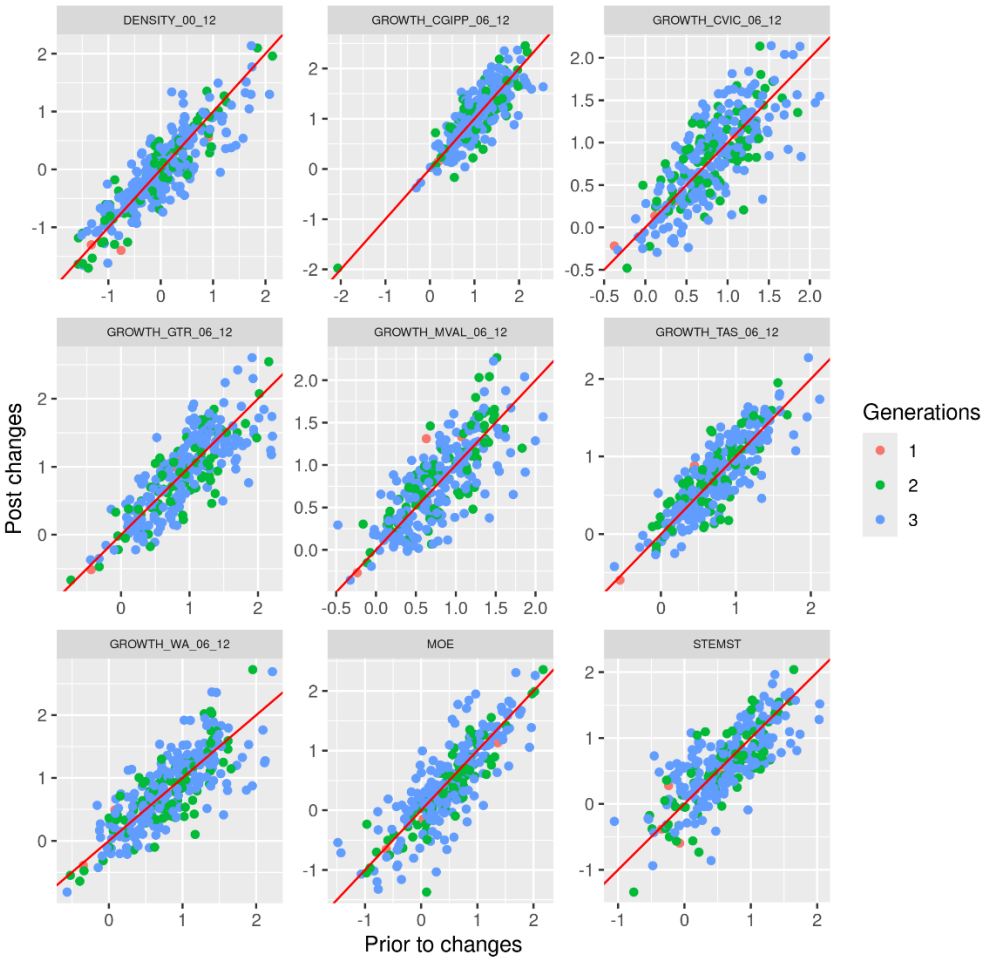

Figure 10 Comparison of EBVs computed prior to (x-axis) and post-pedigree recovery changes (y-axis) for the 326 progeny. The observed shifts in EBV highlight the strong dependence of progeny EBVs on parental EBVs and the cascading impact of pedigree corrections.

2. Parent EBV Comparisons:

Figure 11 highlights the impact of these corrections on the 211 parents replaced by PTPs. Compared to the progeny, the changes in parental EBVs are generally smaller and more constrained, clustering closely along the 1:1 line. This is expected because parental EBVs are informed by a larger number of progeny records, which stabilize the estimates. However, subtle shifts still occur when:

- Alleged progeny are removed from a parent's pedigree, reducing its influence.
- New progeny are added to the parent, increasing its contribution and potentially influencing its EBV.

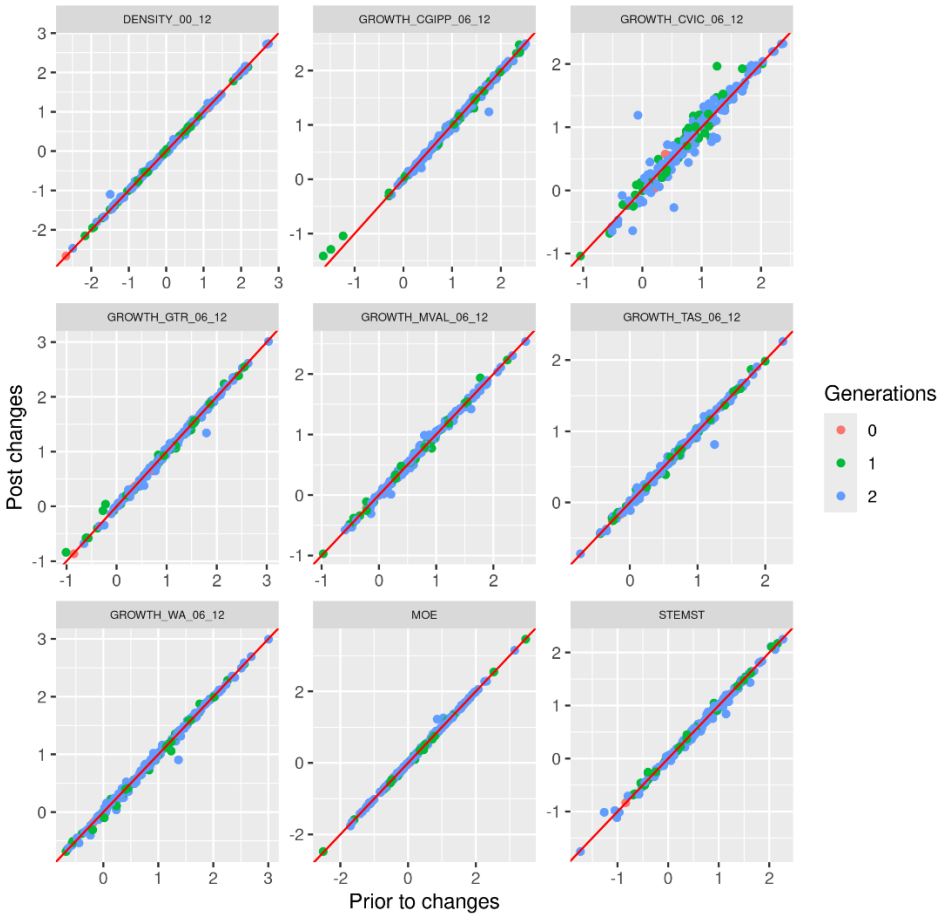

Figure 11 Comparison of EBVs computed prior to (x-axis) and post-pedigree recovery changes (y-axis) for the 211 replaced parents. Smaller, more constrained shifts reflect the stabilizing effect of multiple progeny on parental EBVs.

Discussion

This project demonstrates the ongoing evolution of genomic selection techniques in *Pinus radiata* breeding programs. Significant progress has been made across key areas, including field sampling, DNA processing, pedigree verification, and evaluating the effects of training set size on the accuracy of genomic selection. Additionally, we examined the impact of pedigree recovery on EBV predictions, highlighting the critical role of accurate parentage in genetic evaluations. The only facet of the project not completed was the experiment to determine the boundaries of DNA storage conditions; this will be carried forward into the sister project, "*Using genomics to double the rate of genetic gain in Australian forest tree improvement programs*" (VNC580-2122), ensuring continuity and a successful outcome.

A critical component of this project was the use of the Axiom PRAD array, developed by an overseas consortium, which has the capacity to call approximately 36,000 variants. The raw CEL data obtained from two laboratories (which involved unique challenges) was analyzed using the Axiom Array Software (AxAS), which recommended a best and recommended set of 21,267 variants. Further filtering based on Hardy-Weinberg Disequilibrium (HWD) and Minor Allele Frequency (MAF) reduced the number to 9,905 variants. Our custom filtering test, which focused on high OHL fractions across dyads, further whittled down the set to 8,871 variants.

This considerable reduction from an initial set of 36,000 variants underscores the importance of optimizing genotyping arrays. The high rate of filtering highlights the inefficiency of using generic arrays where many variants are not segregating or prove to be unreliable in Australian populations. This validates our efforts to pursue a custom-designed array, where chip real estate is optimized, and the initiative to obtain a chromosome-level full genome assembly for *P. radiata*. This will provide the foundation for designing a more efficient chip specifically tailored to Australian populations, ensuring that valuable resources are not wasted on variants that do not contribute to genetic differentiation.

The evaluation of single-step genetic evaluation and the prediction of Mendelian sampling terms (ϕ) further emphasize the transformative impact of genomic selection in P. radiata breeding programs. By accurately predicting ϕ early in a juvenile's life, breeding programs can significantly reduce the generation interval, which has a greater impact on genetic gain than increasing prediction accuracy. Our results demonstrate that incremental increases in training set size improve the correlation between genomic predictions (ϕ_add) and observed phenotypes (ϕ_phen). This underscores the importance of expanding the training set to the critical threshold of 20,000 individuals, as predicted by theoretical models. While progress is ongoing, the results for traits such as GROWTH_GTR highlight the substantial gains achievable when sufficient trait-specific phenotypic data are available.

Equally significant has been the work around pedigree error checking and recovery, which has direct implications for genetic evaluations. Through rigorous error detection and correction facilitated by the Scion DNA assay, parentage for 326 progeny was updated, involving the reassignment of 243 putative true parents (PTPs) and the creation of 89 new families. Comparative analyses of EBVs pre- and post-pedigree recovery demonstrate that progeny EBVs are highly sensitive to parentage changes, reflecting their dependence on parental breeding values. In contrast, changes to parental EBVs were generally smaller and more constrained due to the stabilizing influence of multiple progeny. These results highlight the cascading effects of pedigree recovery across generations and reinforce the importance of robust validation protocols and auxiliary databases to ensure confidence in parentage corrections.

The successful implementation of single-step BLUP methods, combined with rigorous pedigree recovery, marks a significant advancement in breeding value estimation for *P. radiata*. Future work, including the development of high-density SNP arrays and improved genomic resources, will further

enhance the precision and utility of genomic selection, supporting continued genetic gain across breeding programs.

Limitations and Future Work

While progress has been made in increasing the size of the training population, reaching the target of 20,000 genotyped individuals will be critical for maximizing the benefits of genomic selection. Expanding the training population will remain a key focus in future efforts.

Conclusions

- This project has made significant strides toward enhancing the efficiency and accuracy of genetic selection in *Pinus radiata* breeding programs. By integrating genomic data with traditional pedigree-based approaches, the project demonstrated that Single-Step Genomic Selection (SSGS) methodology is an effective tool for improving the prediction of breeding values and accelerating genetic gain in tree improvement programs.
- Key milestones include the validation of pedigrees, which showed a very respectable 4-5% error rate, highlighting the accuracy and rigor of breeding records, particularly in orchards managed by HVP Plantations (HVP). The collaboration between industry partners such as the Gippsland Centre of the National Institute for Forest Products Innovation (NIFPI) and other stakeholders has been instrumental in achieving these results.
- Although the current training population has grown to 9,000 individuals, more substantial gains in predictive accuracy are anticipated once the target population size of 20,000 genotyped trees is reached.
- While the Axiom PRAD array served as a valuable tool for this project, the considerable reduction of usable variants after filtering indicates inefficiencies in using a generic chip not specifically tailored to Australian *P. radiata* populations. The high rate of variant exclusion highlights the need for the development of a custom-designed genotyping chip optimized for local conditions. This would ensure more effective use of resources and significantly enhance the accuracy and reliability of genetic evaluations moving forward.
- Additionally, the project's efforts to address DNA sampling, storage conditions, and optimization of SNP arrays have laid a solid foundation for future genomic evaluations. Ongoing work in the sister project (VNC580-2122) will ensure that the experiments not fully completed within the current timeline will be carried forward.
- The project has advanced the understanding and implementation of genomic selection in *P. radiata* breeding programs. With continued efforts to expand genomic datasets and refine genetic evaluation models, the prospects for achieving faster genetic gains in Australian forestry are promising.

Recommendations

1. **Develop a Custom Genotyping Chip**
   Given the high rate of variant exclusion with the Axiom PRAD array, prioritize the development of a custom genotyping array tailored to Australian *P. radiata* populations. This will enhance accuracy and efficiency in genomic selection processes by ensuring that the chip captures more relevant and segregating variants for local conditions.
2. **Expand the Genomic Training Population**
   Continue efforts to expand the genomic training population to the target of 20,000 genotyped individuals. This is critical to achieving measurable gains in prediction accuracy for breeding value estimations, especially in single-step genomic selection.
3. **Optimize DNA Sampling and Storage Protocols**
   Refine protocols for long-term DNA storage and extraction, ensuring optimal DNA yield and quality across varying storage conditions. This is particularly important for maintaining sample integrity over time, which supports the scaling of genotyping and genomic selection efforts.
4. **Increase Collaboration with Research and Industry Partners**
   Foster deeper collaborations with organizations such as Tree Breeding Australia (TBA), AGRF, Ramaciotti Centre, and ThermoFisher Australia to streamline DNA processing workflows and enhance data sharing capabilities. Strengthening these partnerships will be vital for scaling future genomic selection activities and reducing project timelines.
5. **Enhance Pedigree Validation and Identity Assurance**
   Further develop and apply robust pipelines for pedigree validation and identity assurance, such as the Opposing Homozygous Locus (OHL) method. This ensures the continued integrity of breeding programs, particularly in mitigating pedigree errors, which were found to be low but significant.
6. **Continue Single-Step Genomic Evaluation**
   Complete the single-step genetic evaluation for *P. radiata* using the expanded genomic training population and revised Genomic Relationship Matrices (GRMs). The integration of corrected pedigree data with genomic information will likely yield more accurate predictions and should be prioritized in ongoing projects.
7. **Monitor Bulk Seed Use in Breeding Programs**
   Implement more stringent checks on the use of bulk seed in planned field trials to prevent identity errors. Bulk seed should be clearly segregated and labelled to avoid confusion with pedigreed material, ensuring that true genetic progress is accurately tracked.

References

Dodds, K.G., McEwan, J.C., Brauning, R., Anderson, R.M., van Stijn, T.C., & Clarke, S.M. (2015). Construction of relatedness matrices using genotyping-by-sequencing data. BMC Genomics, 16, 1047. https://doi.org/10.1186/s12864-015-2252-3.

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*.

Weir, B.S. (1996). Genetic Data Analysis II: Methods for Discrete Population Genetic Data. Sunderland, MA: Sinauer Associates.

Acknowledgements

This project would not have been possible without the invaluable contributions and support from several individuals and organizations. We would like to express our deepest gratitude to the following individuals and organizations for their invaluable contributions to this project:

- **National Institute for Forest Products Innovation (NIFPI)** for its support and funding.
- **HVP Plantations (HVP)** for their ongoing collaboration and provision of resources.
- **Kinneret Hemo** and her dedicated team at the **HVP Gelliondale Nursery** for their dedicated assistance with field operations and logistics.
- **Dr Jo Sasse** for her assistance with the planning of the sampling effort at Gelliondale.
- **David McKersie** and his team at Mount Gambier, including **John**, **Caitlin**, **Toneya** and **Duncan**, for their essential work in foliage sampling and technical support.
- **Nicole Burtt** and the team at the **Australian Genome Research Foundation (AGRF)** in Adelaide for their excellent work in DNA extraction.
- **Firoozeh Salehzadeh** and **Christie Foster** at the **Ramaciotti Centre** in Sydney for their invaluable assistance in processing our samples using the PRAD chip.
- **David Rayner** at Thermo Fisher Australia for his technical expertise in the analysis of Axiom arrays.
- **Dr Shaieste Gholami** for her help in processing the DNA extraction plates.
- **Katerina Viduka** for her assistance in undertaking the experiment to optimize DNA Sampling and storage protocols.
- Our colleagues at **Tree Breeding Australia** for their ongoing collaboration and support throughout this project.

Researcher’s Disclaimer (if required)